Bioinformatics and its relation to data and computer science

Jaap Heringa
Department of Computer Science
Faculty of Science
http://ibi.vu.nl, j.heringa@vu.nl

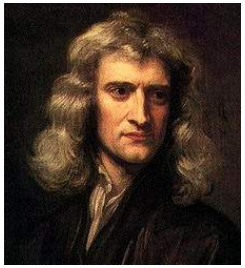
History of Science

- Anatomy, architecture

1632

- Dynamics, mechanics

- Informatics



1726

(Cybernetics – Wiener, 1948)

(Cybernetics has been defined as the science of control in machines and animals, and hence it applies to technological, animal and environmental systems)

1948

- Genomics, bioinformatics, systems biology

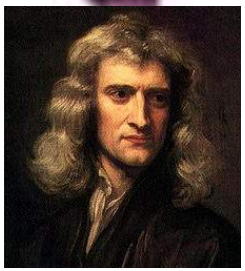
time

History of Science

1632

- Anatomy, architecture →
- Dynamics, mechanics →

- Informatics

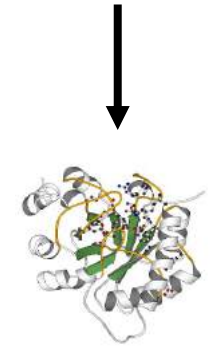


1726

(Cybernetics – Wiener, 1948)

(Cybernetics has been defined as the science of control in machines and animals, and hence it applies to technological, animal and environmental systems)

1948



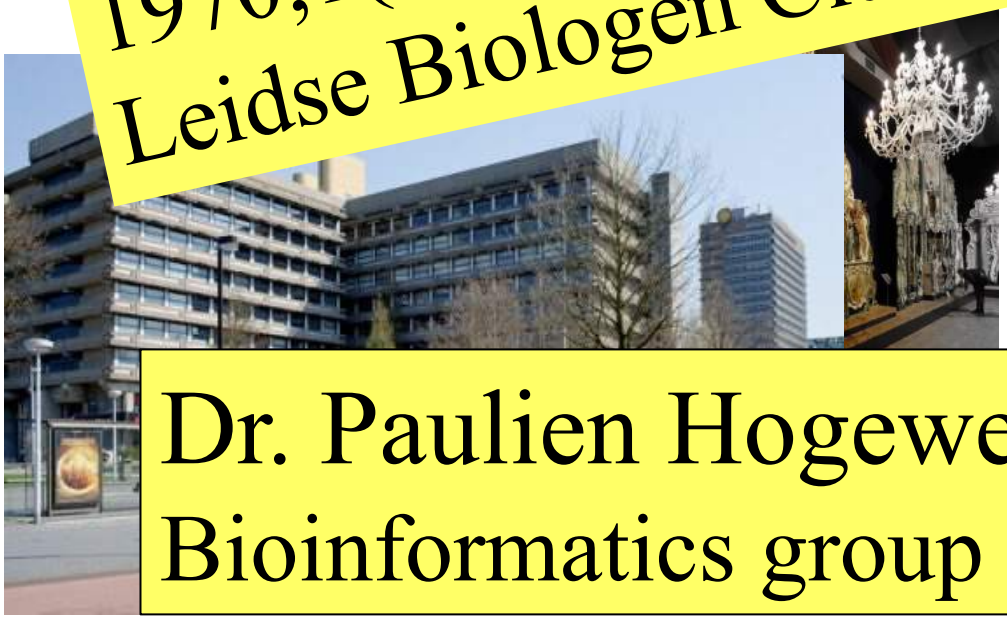
  - Genomics, bioinformatics, systems biology →

time

Bioinformatics originated in Utrecht

Bioinformatics originated in Utrecht

Hesper B, Hogeweg P. Bioinformatica: een werkconcept. Kameleon. 1970;1(6):28–29. (In Dutch.) Leiden: Leidse Biologen Club



Dr. Paulien Hogeweg
Bioinformatics group

Bioinformatics originated in Utrecht

Hesper B, Hogeweg P. Bioinformatica: een werkconcept. Kameleon. 1970;1(6):28–29. (In Dutch.) Leiden: Leidse Biologen Club

"… studying informatic processes in biotic systems"

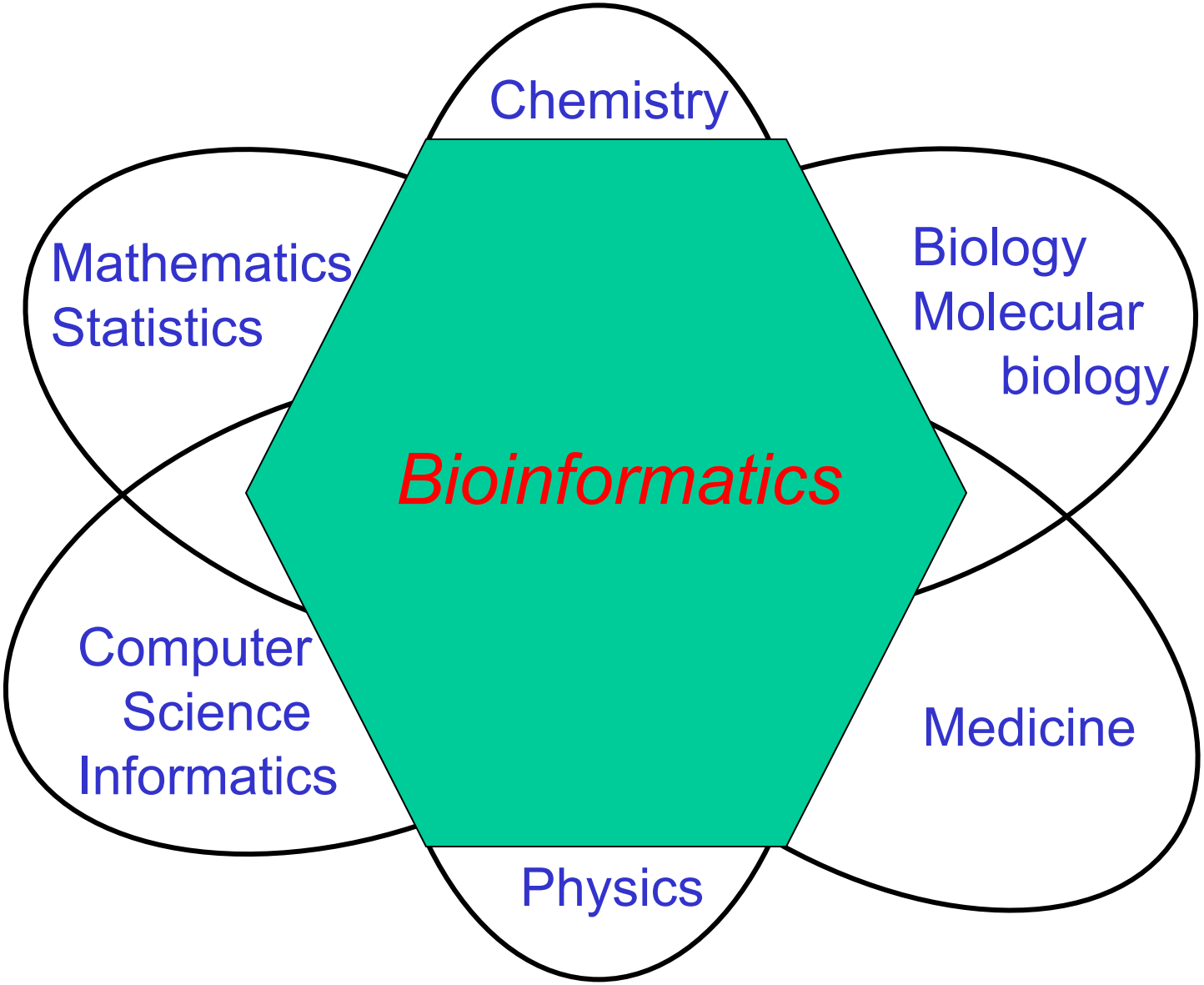
Hogeweg P (2011) The Roots of Bioinformatics in Theoretical Biology. PLoS Comput Biol 7(3): e1002021. https://doi.org/10.1371/journal.pcbi.1002021

Anatomy, dynamics, informatics

Modern life sciences are data sciences..

..and are becoming ever more inter-disciplinary

Bioinformatics



What is driving Life Sciences

Technology/high-throughput measurements

- Bio-sciences
  - Genomics: HTP measurements; e.g. Sequencing (NGS), Chip-seq, RNA-seq
  - Proteomics, metabolomics
    - X-ray, NMR, Mass Spectrometry
  - Imaging, optical measurement techniques, single cell measurements, single molecule tracking
  - Lots of new stuff coming up…

Data generating technologies enabled by IT

It’s a nervous field....

- Changes all the time
  - New measurement techniques
  - New data time and again
  - New technology, formats, standards, hypes
  - New insights

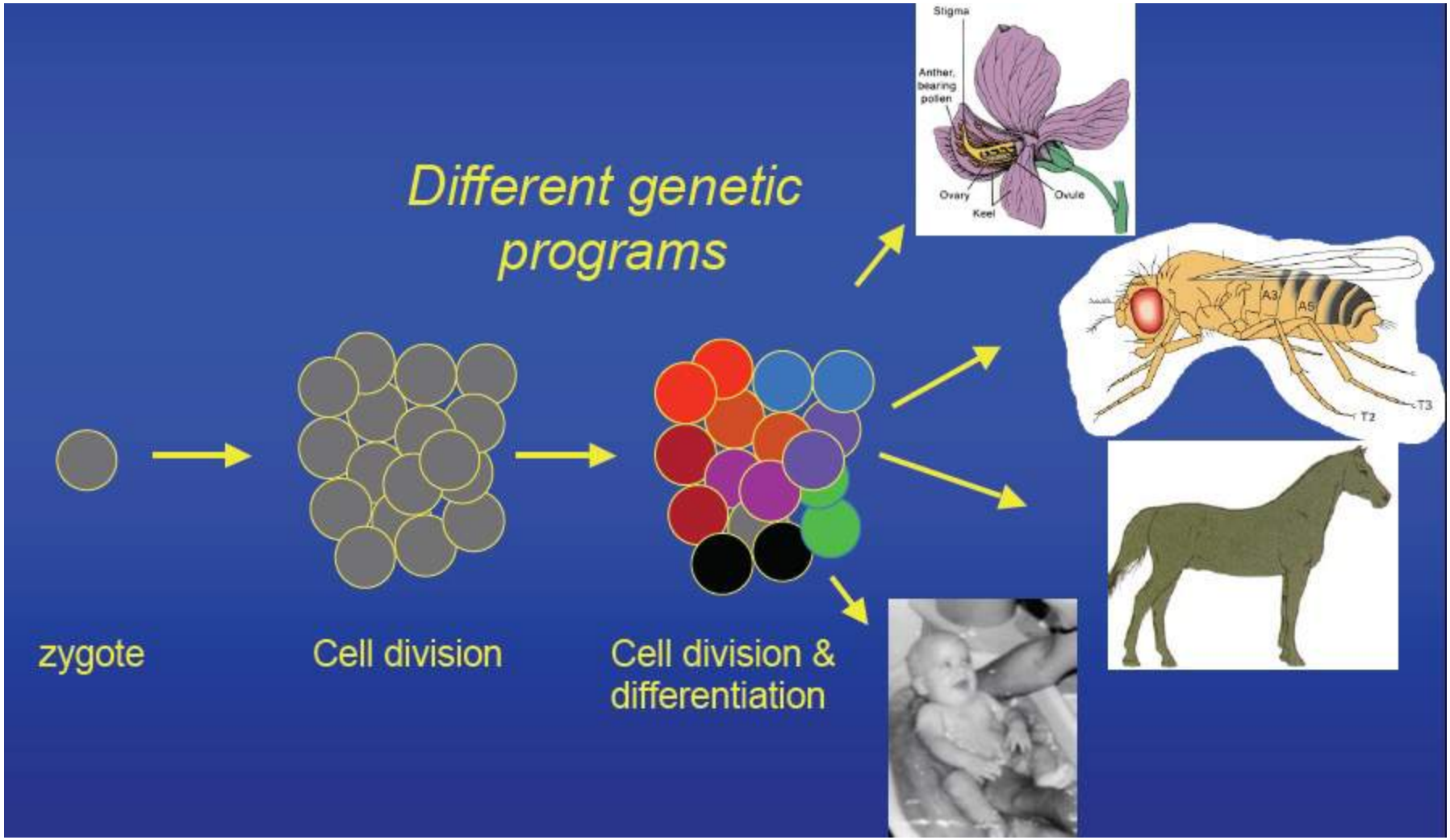
It's a nervous field....

- Changes all the time
  - New measurement techniques
  - New data time and again
  - New technology, formats, standards, hypes
  - New insights

- Compare this to studying ancient Greek philosophy
  - Not many new data
  - Perhaps insights develop

Multicellular organisms:

Development of a zygote into a mature organism: many questions remain!



What makes a biological species: how are differences generated and what are the consequences of these differences?

- What is causing the difference between species? How do species arise?

- What is causing the difference between members of a population?

Diversity in complexity and size

- Enormous diversity in scope:
  - Part of organism – virus
  - Single cell – bacterium, unicellular organisms
  - Multicellular organisms (C. elegans 1000 cells, blue whale)
- Science of big numbers: about 42 trillion (~$5*10^{13}$) cells in human organism, divided over 210 different types of tissue.
- A human cell holds about 42 million proteins.
- Almost all cells contain DNA and many (shorter) RNA molecules
- In addition to the genetical machinery, there is the gut and oral microbiome having profound influences on health

Important questions in biology and medicine are dealing with the decoding of the ‘information’ that resides in the genetic material.

How can this……

…lead to this?

DNA: Genotype → Phenotype

Bioinformatics in the olden days

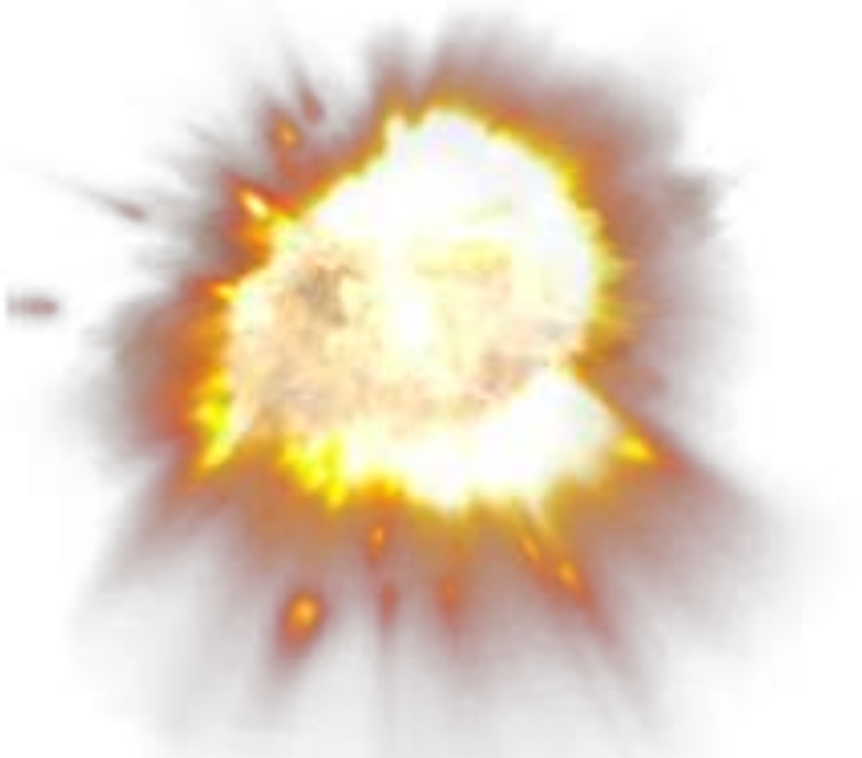
- Close to Molecular Biology:
  - (Statistical) analysis of protein and nucleotide structure
  - Protein folding problem
  - Protein-protein and protein-nucleotide interaction
- Many essential methods were created early on (1970s - ..)
  - Protein sequence analysis (pairwise and multiple alignment)
  - Protein structure prediction (secondary, tertiary structure)
  - Protein interaction (docking) prediction

Bioinformatics in the olden days

- Evolution was studied and methods created
  - Phylogeny: evolutionary ancestry
  - Phylogenetic reconstruction (clustering – e.g., Neighbour Joining (NJ) method)

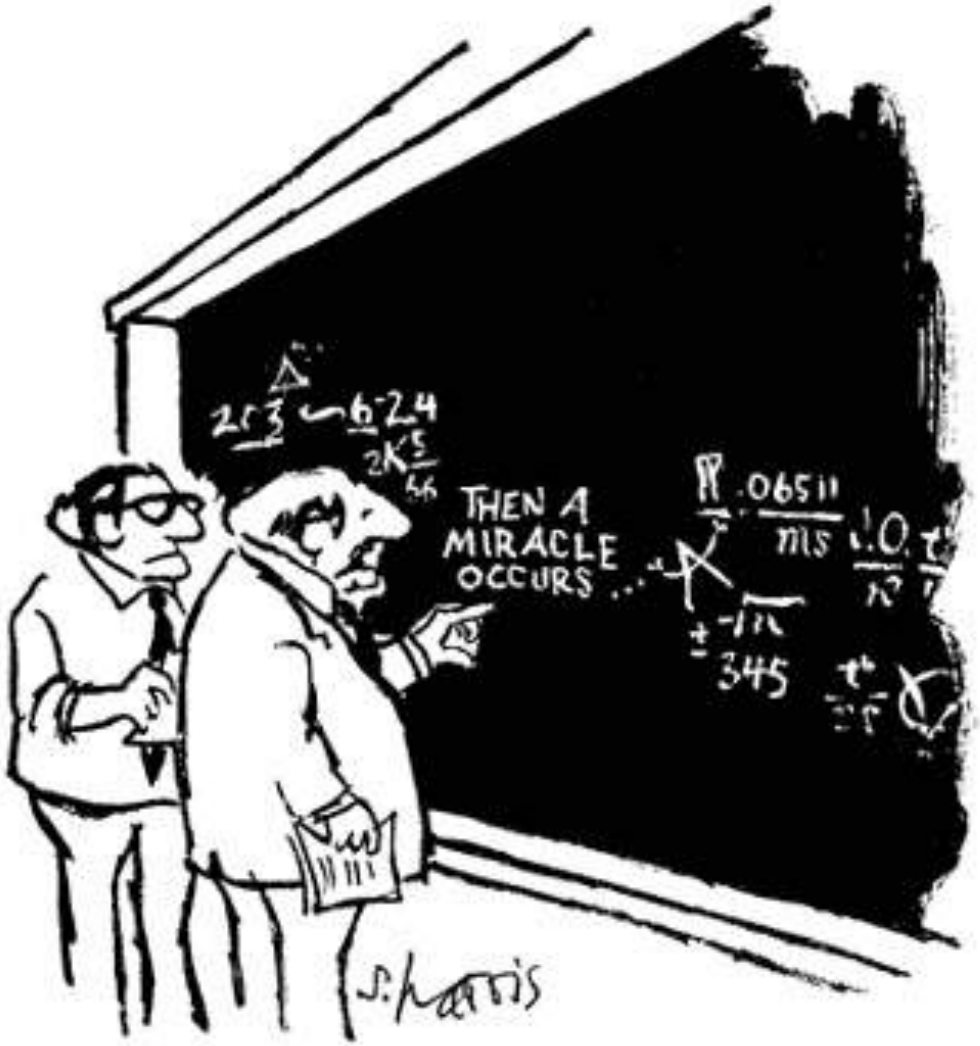
We were making methods..
We were analysing data..
Trying to become important

But then….



… the bioinformatics big bang

The Human Genome Project (HGP)



"I think you should be more explicit here in step two."

The Human Genome Project

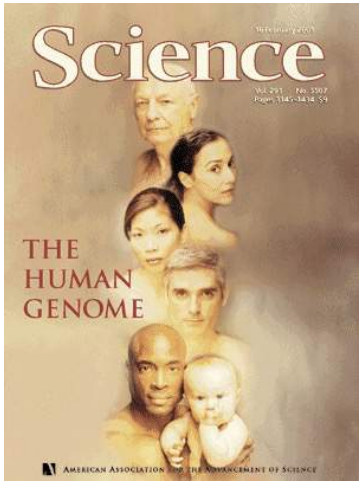
The first global collaborative and inter-disciplinary life science project with big data exchange via the internet

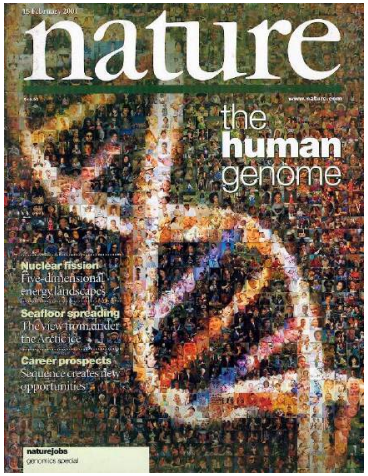
The Human Genome Project

The first global collaborative and inter-disciplinary life science project with big data exchange via the internet

… Although "collaborative"should perhaps be taken with a grain of salt..

The Human Genome Project





A nervous race between academy (HGC) and industry (Celera).

- At stake were patenting issues and the prospect of formidable impediment of progress in biomedical sciences
- The main character: Dr. Craig Venter (Celera)

Human genome project (1990 – 2003)

- ‘a milestone for humanity’

- performed using traditional sequencing techniques

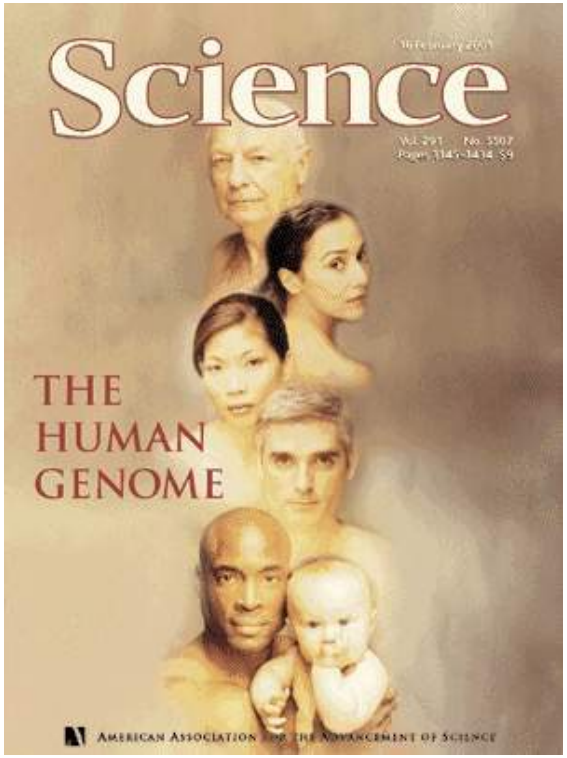
Human genome project (1990 – 2003)

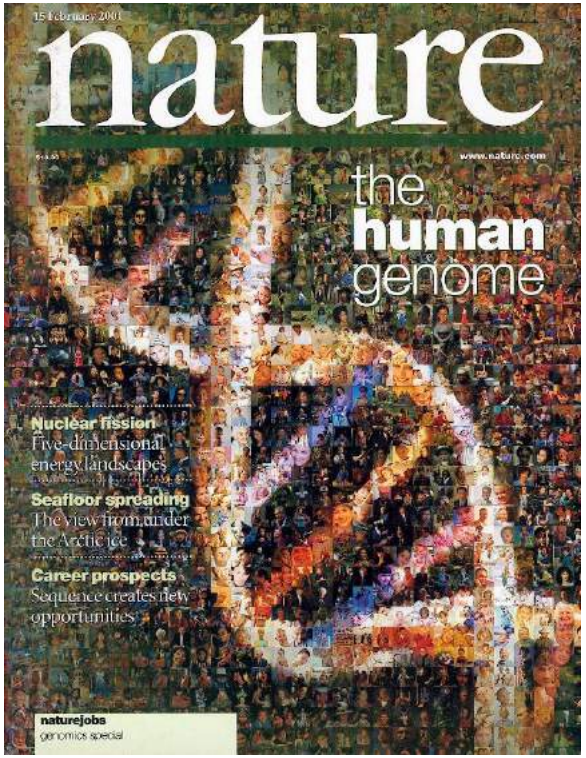
- ‘a milestone for humanity’

- performed using traditional sequencing techniques

Craig Venter’s thread: human genome data might be made proprietary via patents by Celera Genomics

The Human Genome -- 26 June 2000





Dr. Craig Venter

Celera Genomics

-- Shotgun method

Francis Collins (USA) /

Sir John Sulston (UK)

Human Genome Project

The Human Genome -- 26 June 2000

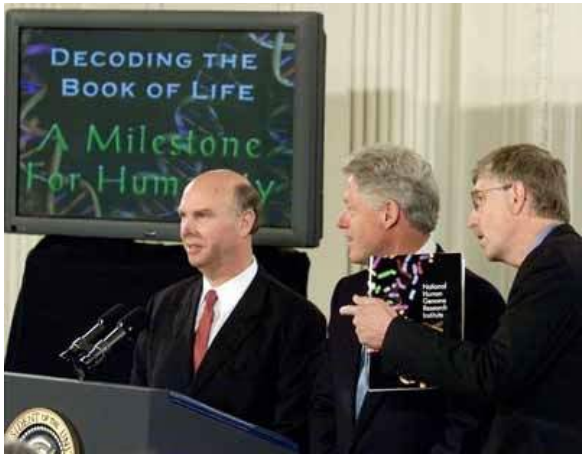
"Without a doubt, this is the most important, most wondrous map ever produced by humankind."

U.S. President Bill Clinton on 26 June 2000 during a press conference at the White House.

26th June 2000

On 26 June 2000, leaders of the public project and Celera announce completion of a working draft of the human genome sequence. Collins and Venter are seen here on television with Ari Patrinos of the DoE, who cut through the animosity between the rival projects to broker the joint announcement at the White House in Washington.

Outside, celebrations continue with Eric Lander of the Whitehead Institute, Baylor's Richard Gibbs, and Waterston and Richard Wilson from Washington University.



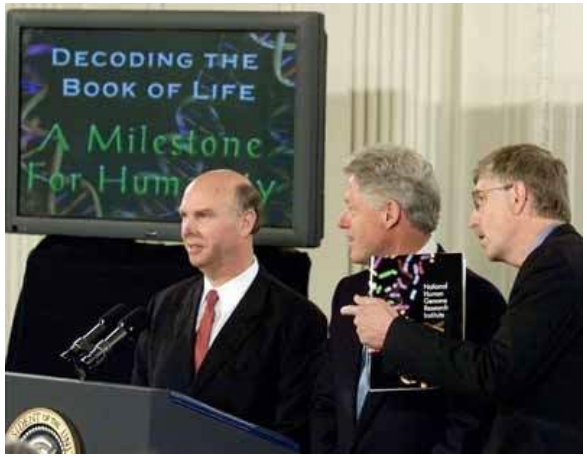
The press conference at the white house, hosted by President Bill Clinton

On hand at a press conference that followed the White House genome announcement are (from l) Dr. Craig Venter, Celera; Dr. Ari Patrinos, U.S. Department of Energy, and Dr. Francis Collins, director, NHGRI. DOE and NIH are the two federal agencies involved in the Human Genome Project.

26th June 2000

On 26 June 2000, leaders of the public project and Celera announce completion of a working draft of the human genome sequence. Collins and Venter are seen here on television with Ari Patrinos of the DoE, who cut through the animosity between the rival projects to broker the joint announcement at the White House in Washington

Outside, celebrations continue with Eric Lander of the Whitehead Institute, Baylor's Richard Gibbs, and Waterston and Richard Wilson from Washington University.



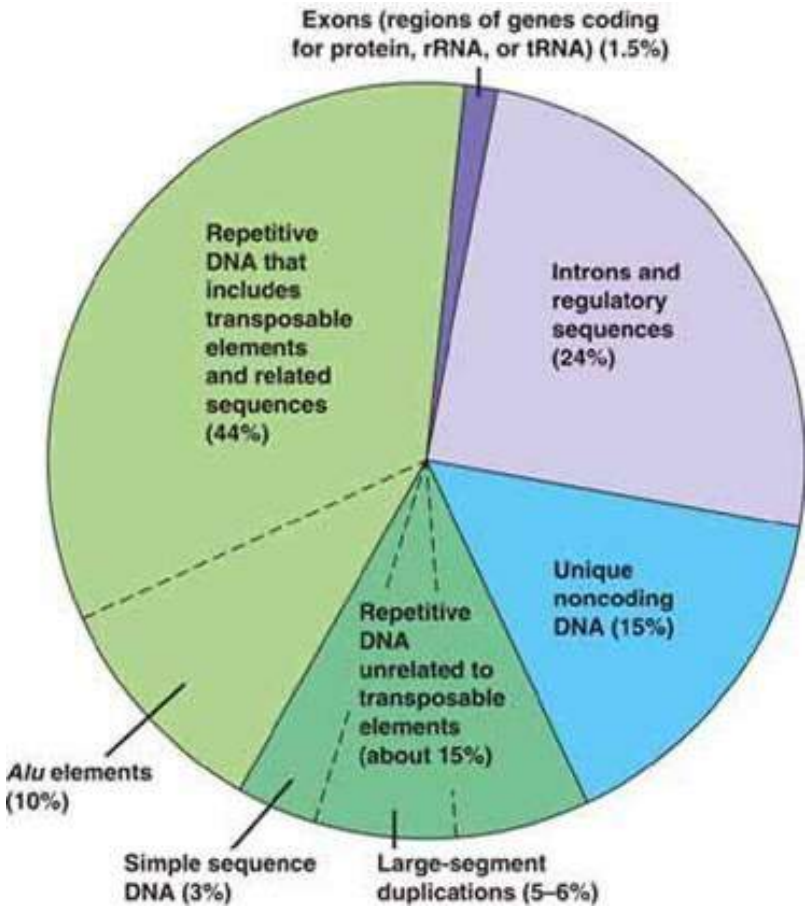
The press conference at the white house, hosted by President Bill Clinton

All is well that ends well…

Onlookers at a press conference that followed the White House genome announcement are (from l) Dr. Craig Venter, Celera; Dr. Ari Patrinos, U.S. Department of Energy, and Dr. Francis Collins, director, NHGRI. DOE and NIH are the two federal agencies involved in the Human Genome Project.

Human genome project - in numbers

- 23 chromosome pairs
- 20.000 genes
- 2.9 billion base pairs (out of 3.3 billion)



Sequencing

Reconstructing a DNA sequence from many randomly selected short fragments (reads)

Reads may contain (experimental) **E**rrors…

Shotgun Method - Pros and Cons

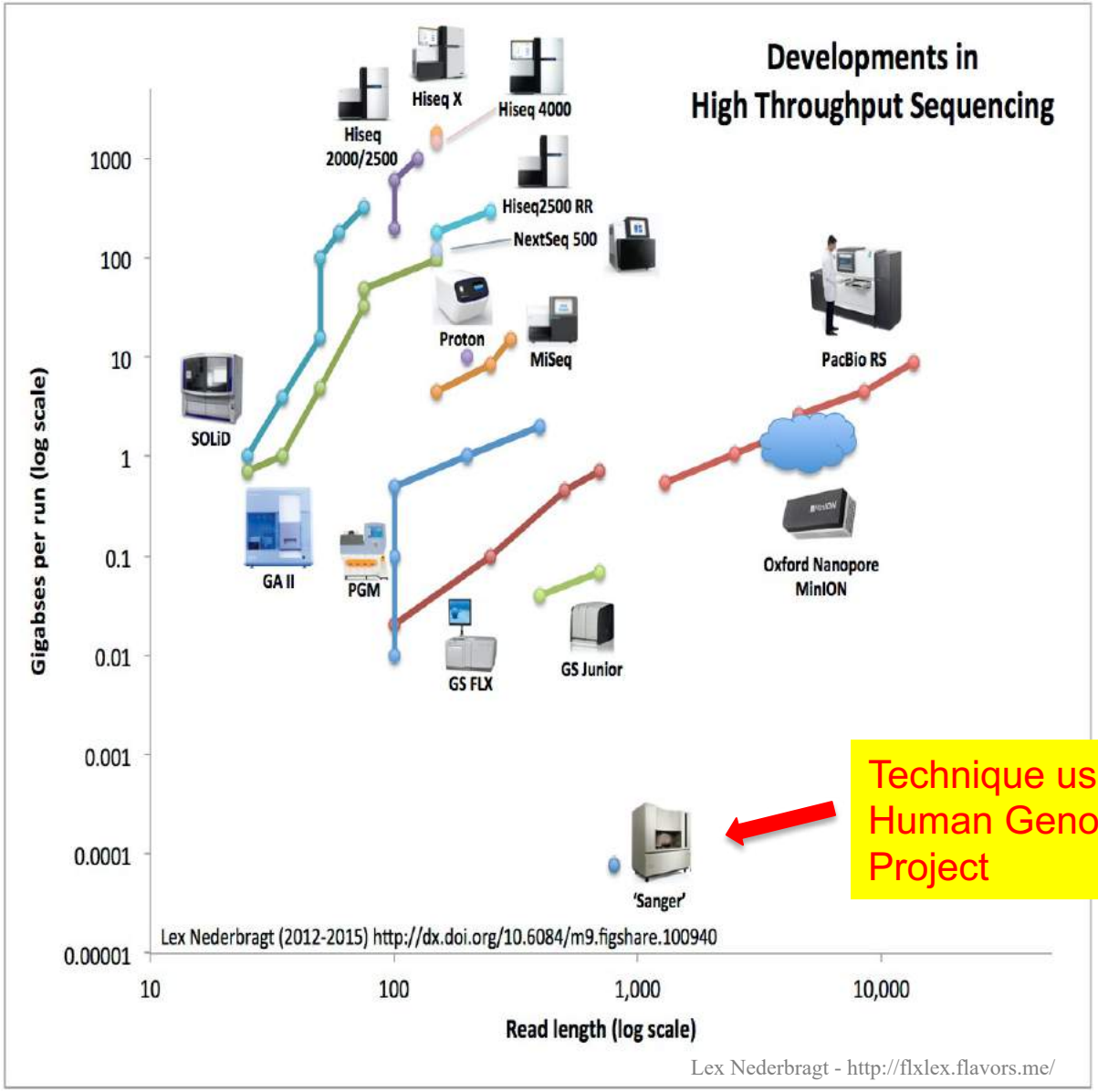
Celera versus Human Genome Project

- *Pros*
  - Human labour reduced to minimum
- *Cons*
  - Computationally demanding – $O(n^2)$ comparisons
  - High error rate in contig construction
    - Repeats as the main problem
    - The human genome is very repetitious (~50%)

Next Generation Sequencing (NGS)

- Massively parallel sequencing of **millions to billions** of short fragments
- Very fast
  - (Sanger sequencing max 384 DNA samples in a single batch (run) in up to 24 runs a day)
- Huge amounts of data generated in single sequencing experiment (many TBs)
- Much reduced cost (1 human genome: HGP 3 billion $ *versus* NGS ~10,000 $)
- Shorter fragments (reads) than with Sanger sequencing
  - Many different techniques exist but based on approx. same principle. Differences reside mainly in chemical usage and the way fragments are stuck to the surface

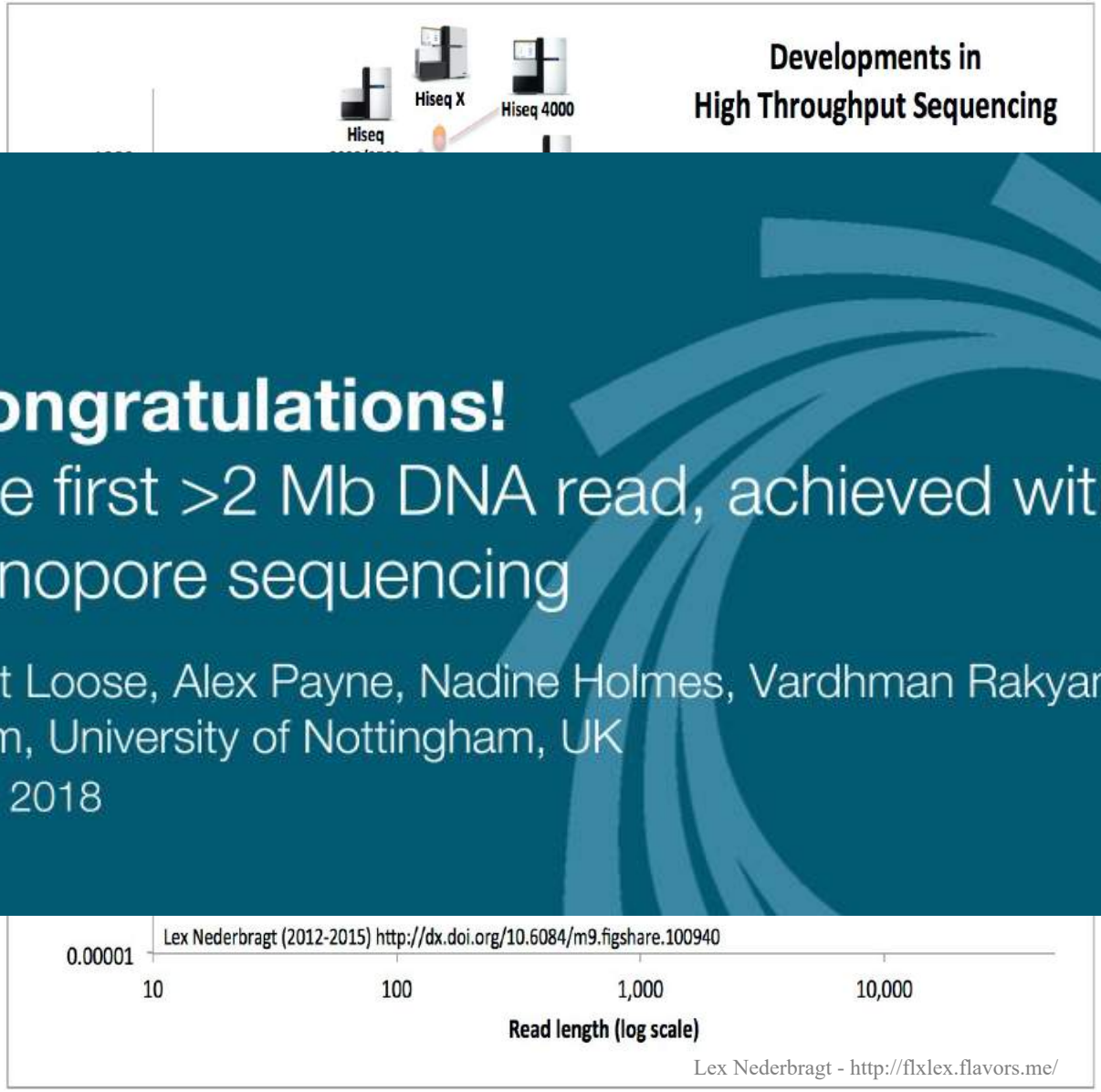
Next Generation Sequencing



Technique used for Human Genome Project

Source: Walter Pirovano, BaseClear

Next Generation Sequencing



Congratulations!

The first >2 Mb DNA read, achieved with nanopore sequencing

Matt Loose, Alex Payne, Nadine Holmes, Vardhman Rakyan & team, University of Nottingham, UK

May 2018

Source: Walter Pirovano, BaseClear

illumina®

HiSeq 2500 System

MinION (2015)

Source: Walter Pirovano, BaseClear

Oxford NANOPORE Technologies

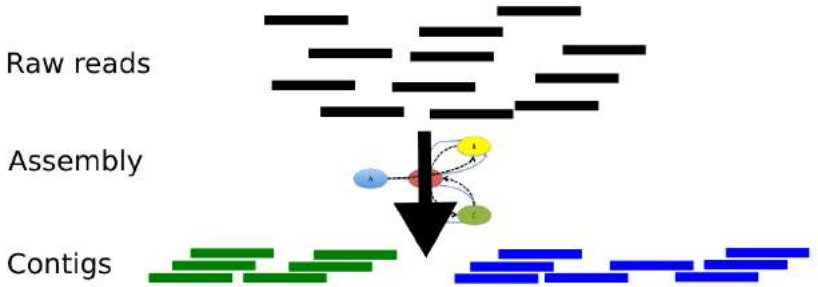
NGS output

- Millions to a billion of sequenced short fragments
  (data handling not easy)
- Can sequence either DNA or RNA sequences
  - Abundance may be estimated (deep sequencing)
- What to do next? *BIOINFORMATICS*

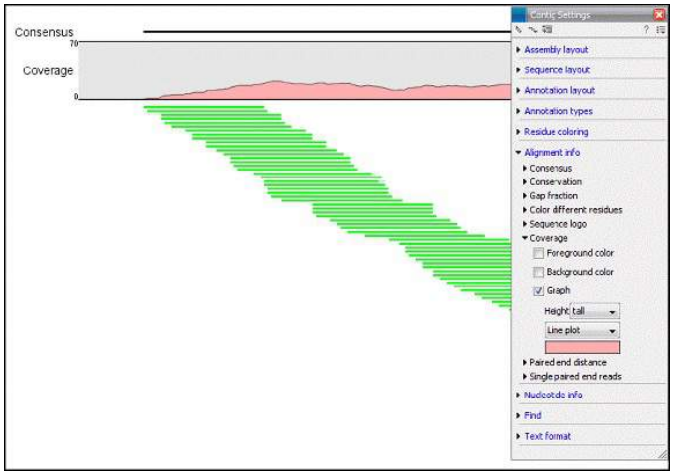
Putting the reads together using bioinformatics

Two main ways of stringing together the many short reads into a complete genome sequence

– *De novo* assembly of a genome



– Assembly using alignment onto a **reference genome**



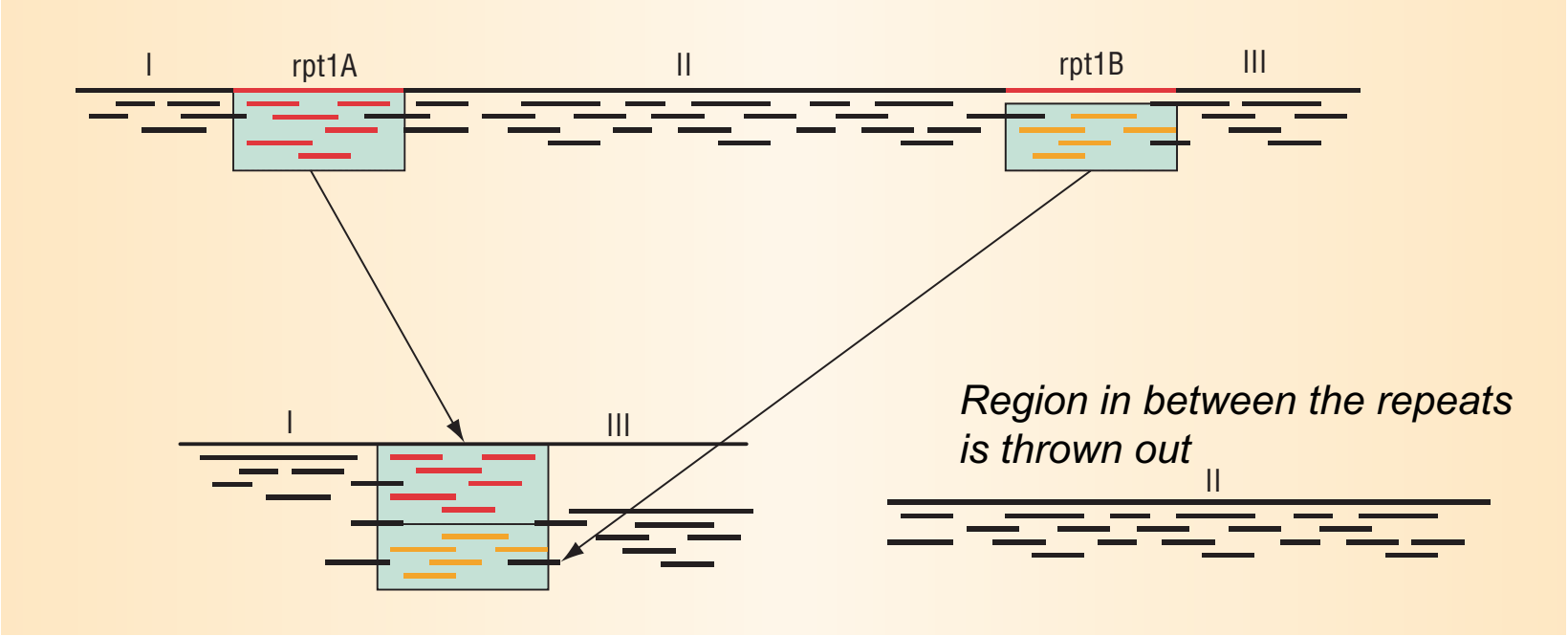
De novo sequencing - a contig

- Reconstructing a complete genome *de novo* requires testing possible overlaps between all possible pairs of reads and then building the whole genome together according to some criterion:

  - A known and related problem in Computer Science is the **Shortest Superstring Problem** (SSP), where all fragments are strung up to produce the shortest overall string *(i.e.* genome).
    - However, the shortest possible string is not an ideal criterion because genomes have many repeating fragments (human DNA >50% repetitious)

Repetitive elements

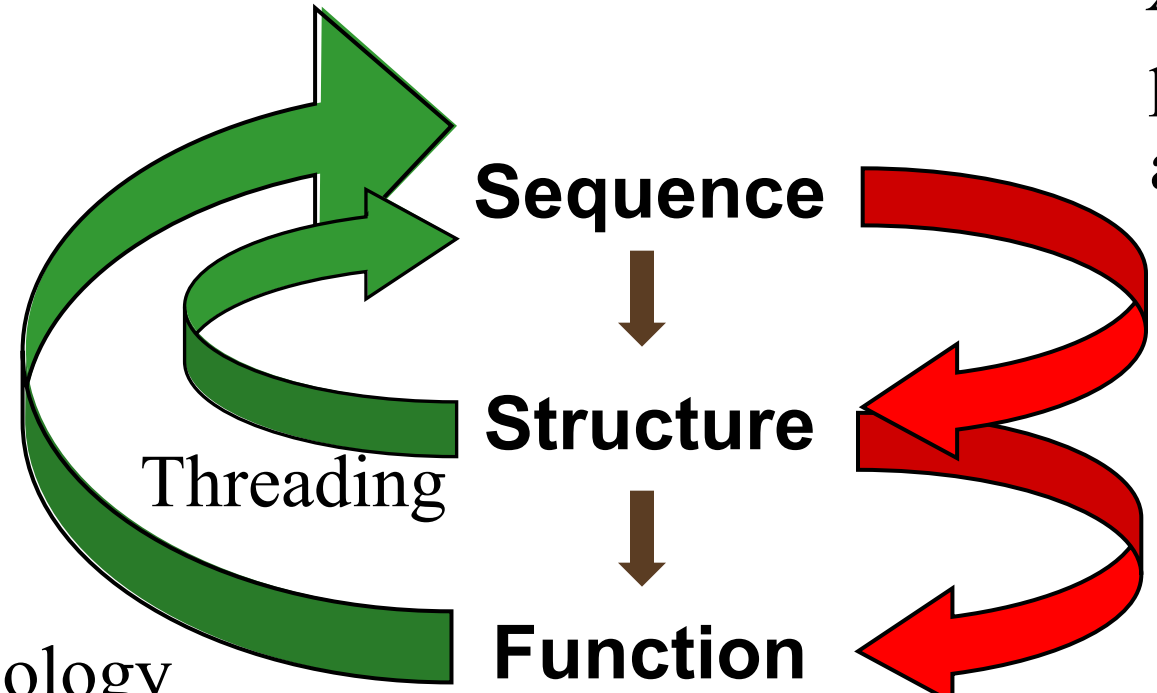
- Repeats can cause major problems to the assembler;
  - Reads corresponding to two separate repeats may be collapsed in a single contig



Why bother with genomics?

- Human DNA contains ~20k genes, encoding for proteins
  - Many genes may encode multiple forms of protein (e.g. through alternative splicing)
- DNA also encodes many different types of functional RNA molecules
- The big challenge is finding out the function of these components in the cell and how they interact.
- Cells and organisms are information processing entities
  - Understanding how they work will give us clues for avoiding or treating diseases.

Sequence-Structure-Function



Ab initio prediction and folding

impossible but for the smallest structures

Homology searching (BLAST)

Function prediction from structure

very difficult

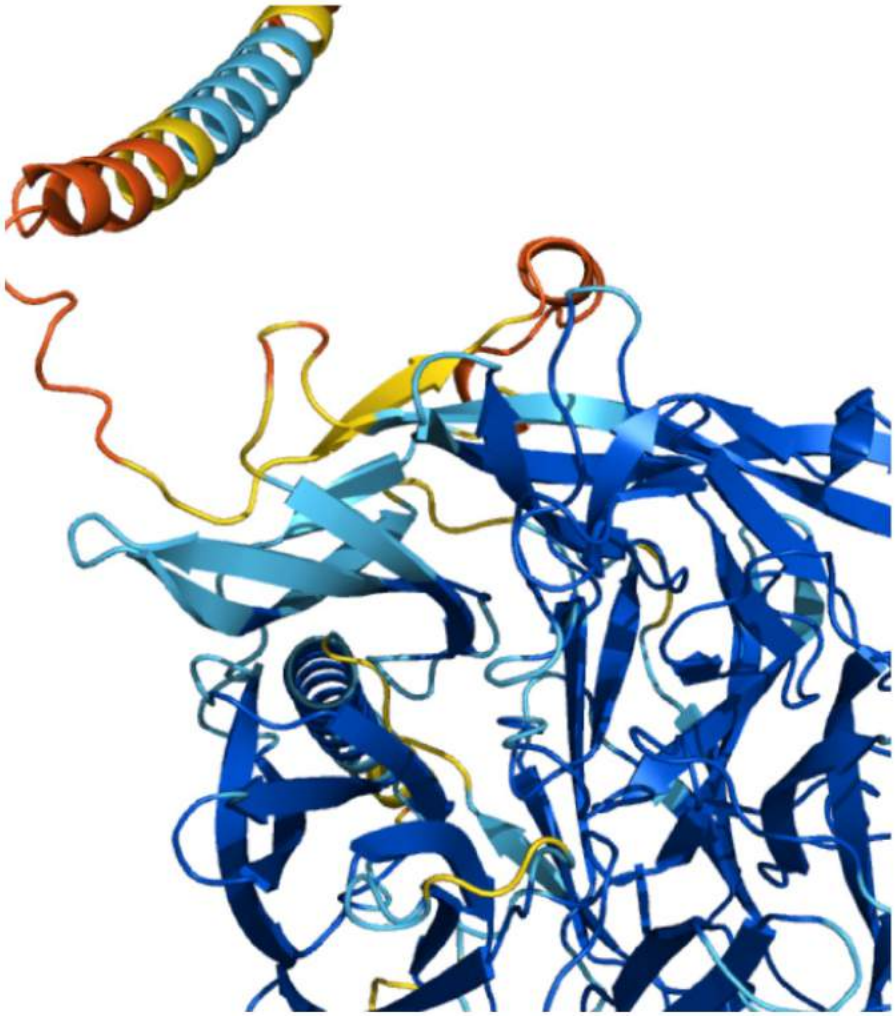
We can neither predict structure from sequence ('folding problem'), nor predict function from structure. However, we can do the knowledge-based activities designated by the green arrows based on the homology principle (see earlier slides) thanks to the availability of curated and annotated databases

AlphaFold

Deep learning ‘solving’ protein folding problem

AlphaFold is an AI system developed by DeepMind that predicts a protein’s 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.

DeepMind and EMBL’s European Bioinformatics Institute (EMBL-EBI) have partnered to create AlphaFold DB to make these predictions freely available to the scientific community. The database covers the complete human proteome (including fragments for long proteins) and the proteomes of 47 other key organisms (e.g. mouse), as well as the majority of manually curated UniProt entries (Swiss-Prot). In 2022 we plan to expand the database to cover a large proportion of all catalogued proteins (the over 100 million in UniRef90).



Q8I3H7: May protect the malaria parasite against attack by the immune system.
Mean pLDDT 85.57.

Searching for similarities

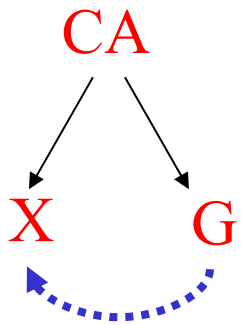
- The main question: what is the function of the new gene?

- The “lazy” investigation without doing experiments:
  - Find a set of similar proteins
  - Identify similarities and differences
  - For long proteins it is often good to identify domains first and then compare the corresponding (sub)sequences separately
    - A domain is a unit of function
    - Multi-domain proteins have a compound function

Inferring homology from similarity

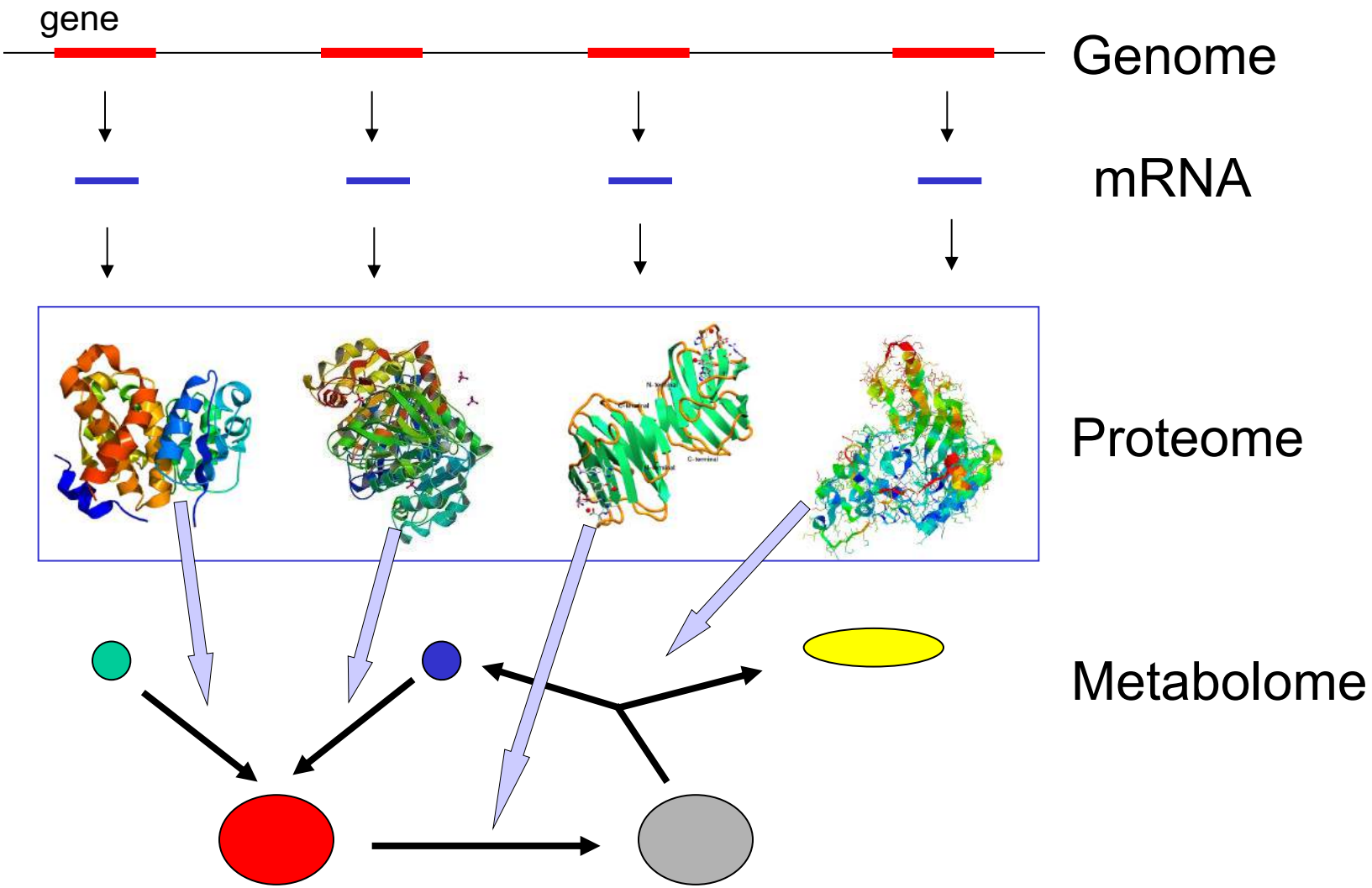
- Homology: sharing a common ancestor
  - a binary property (yes/no)
- Common ancestry makes it more likely that genes share the same function
  - It's a nice tool:

    When (a known gene) G is *homologous* to (an unknown gene) X, we gain a lot of information on X by transferring what we know about G

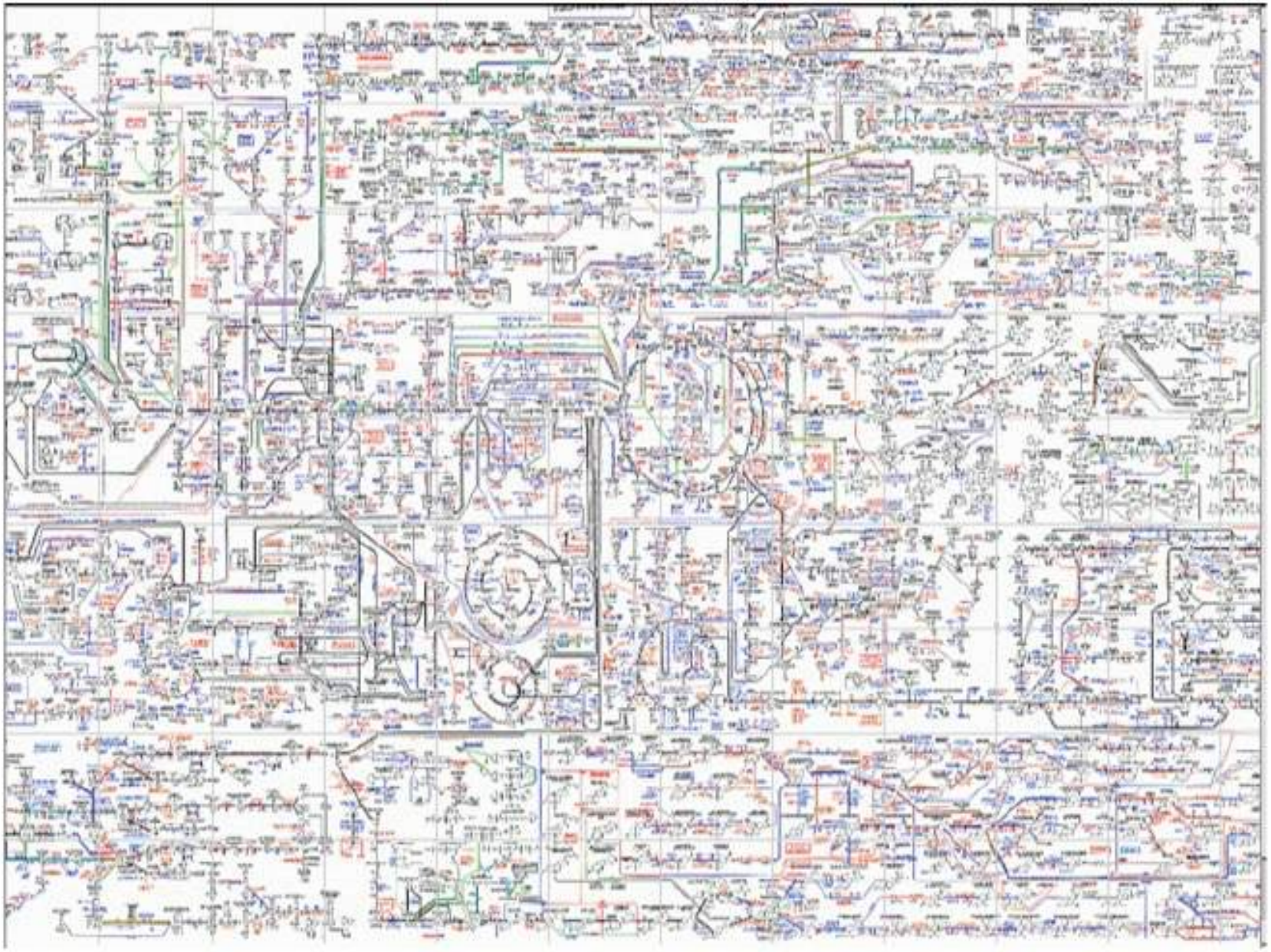


DNA makes RNA makes Protein

From gene to function



The functional network level

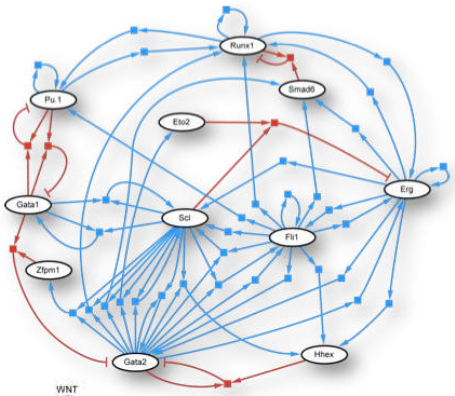


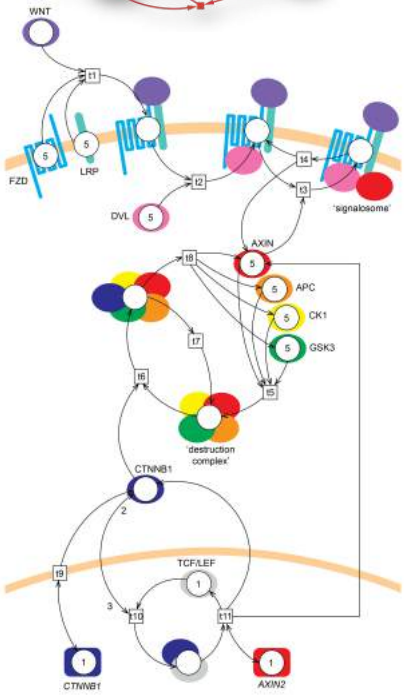
There are various networks in the cell

- Gene regulation
- Protein-protein interaction
- Signalling
- Metabolomic
- Other

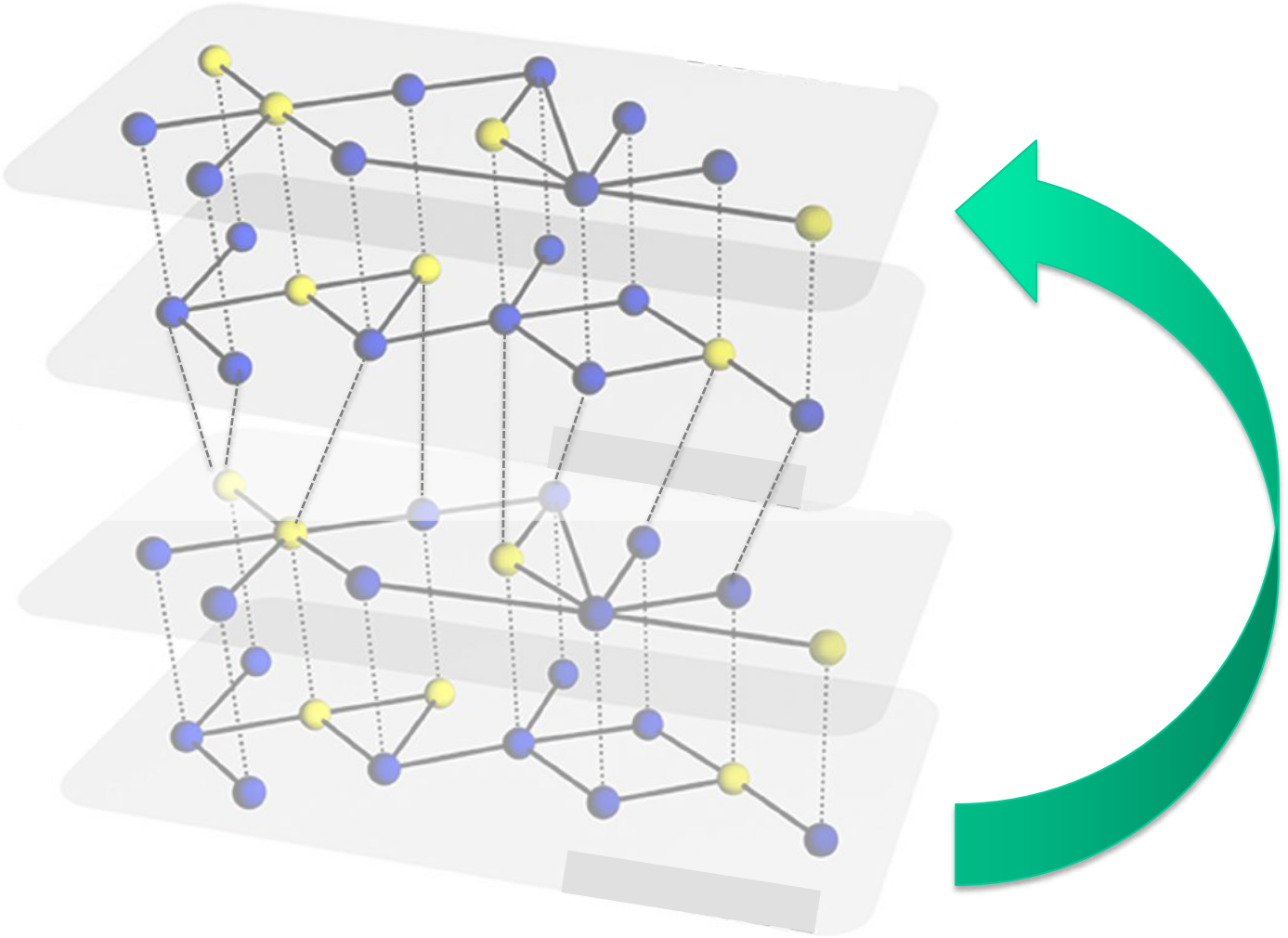
–These networks are interconnected and function in a multi-level way – should function adequately

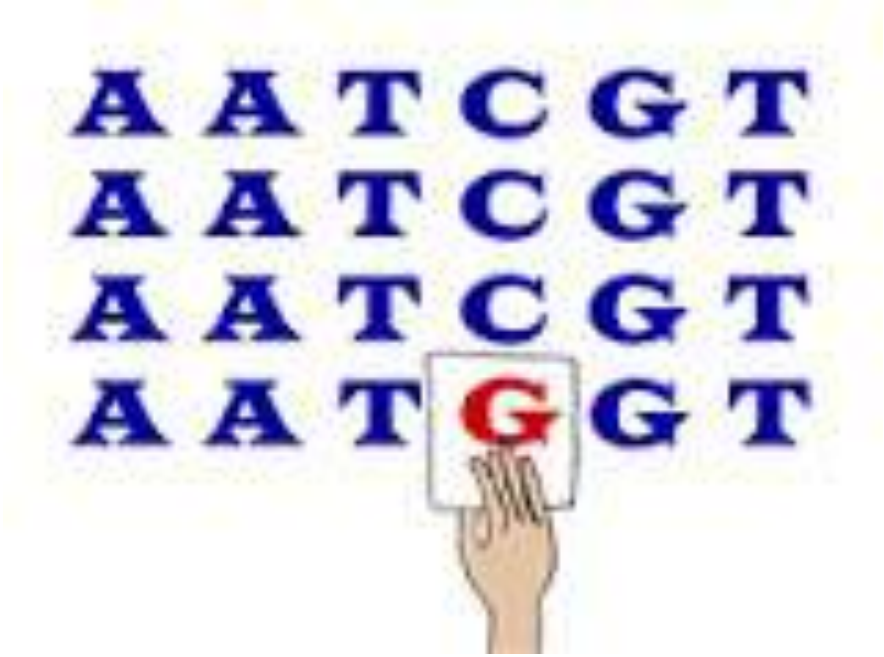
(*note that they are not really there*)





Heterarchically-connected network layers in the cell



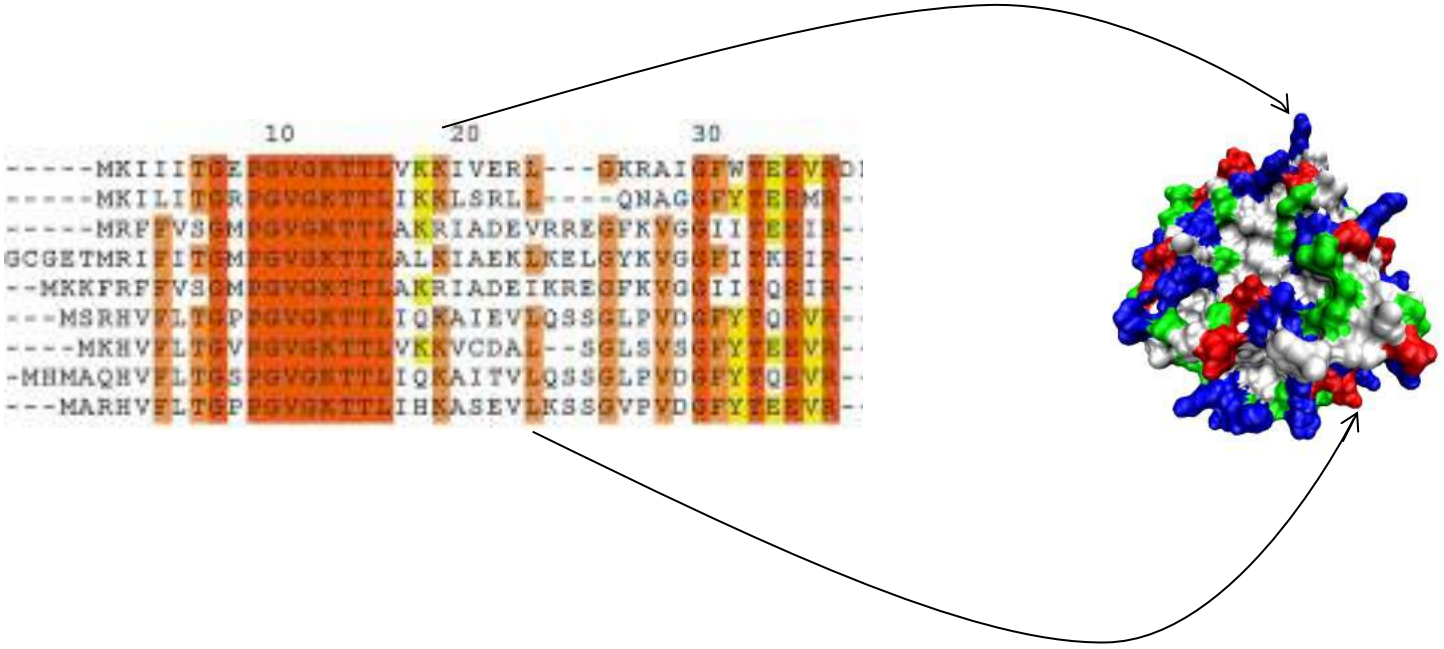


Individual sequence differences may lead to different cellular behaviour at the network level…

Sequences become different during evolution

Protein multiple sequence alignment

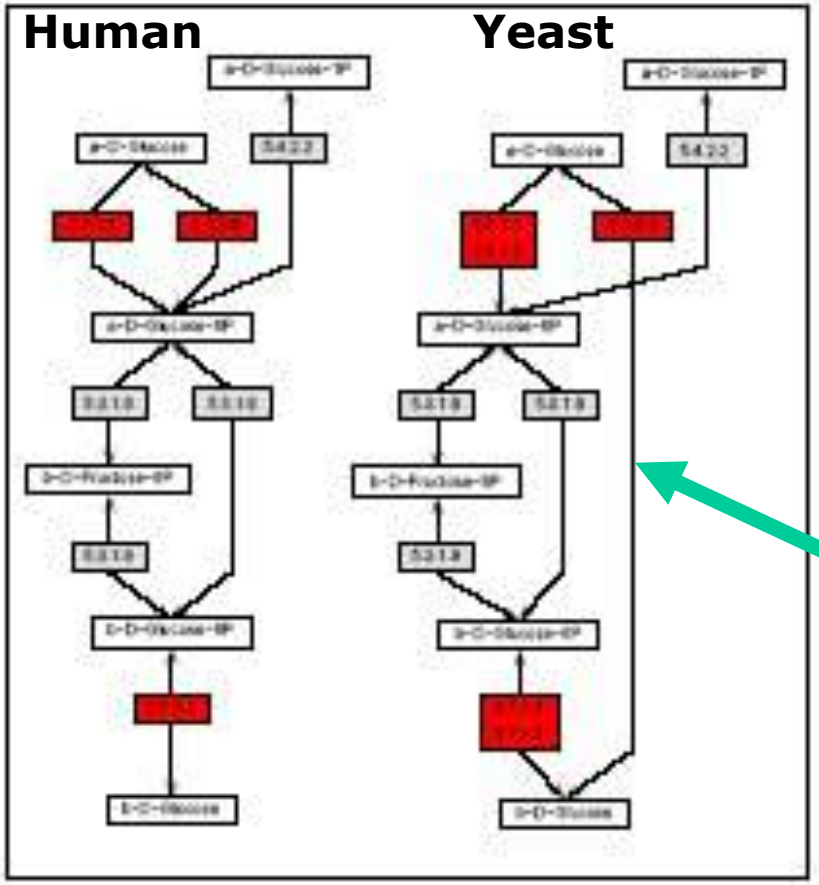
Evolution and three-dimensional protein structure information



What do we see if we colour code the space-filling (CPK) protein model?

- E.g., red for conserved alignment positions to blue for variable (unconserved) positions.

Network Evolution



Networks become different during evolution

- *Homo sapiens* (human) and (right) *Saccharomyces cerevisiae* (baker's yeast).
- Changes in controlling enzymes (boxes in red) and the pathway itself have occurred

Modelling vulval development in C. elegans

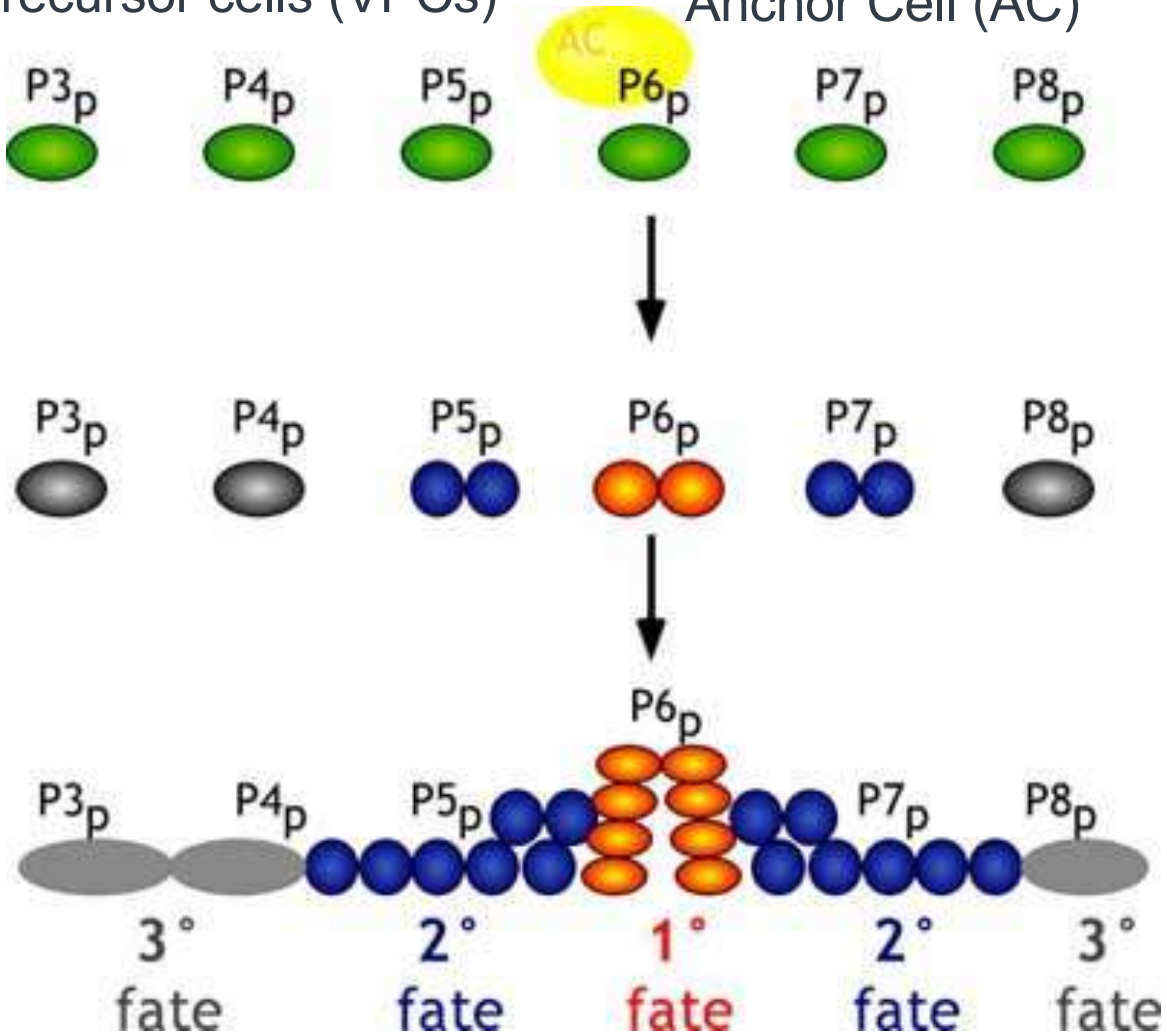
Caenorhabditis elegans

- 1mm long
- 1000 cells
- Intensively studied (Sydney Brenner started research in the 1960s)

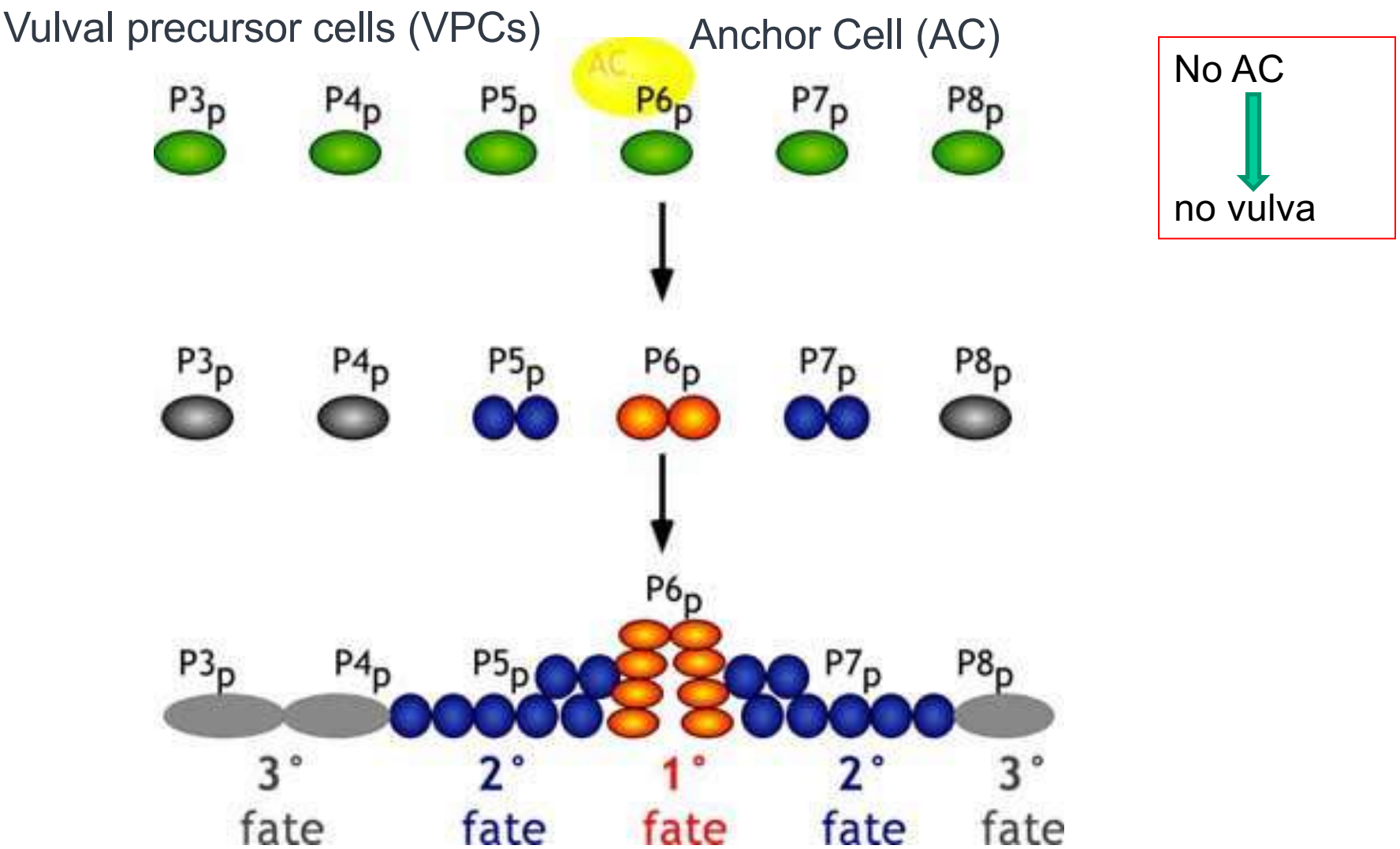
Cell fates and the onset of the vulva

Vulval precursor cells (VPCs)

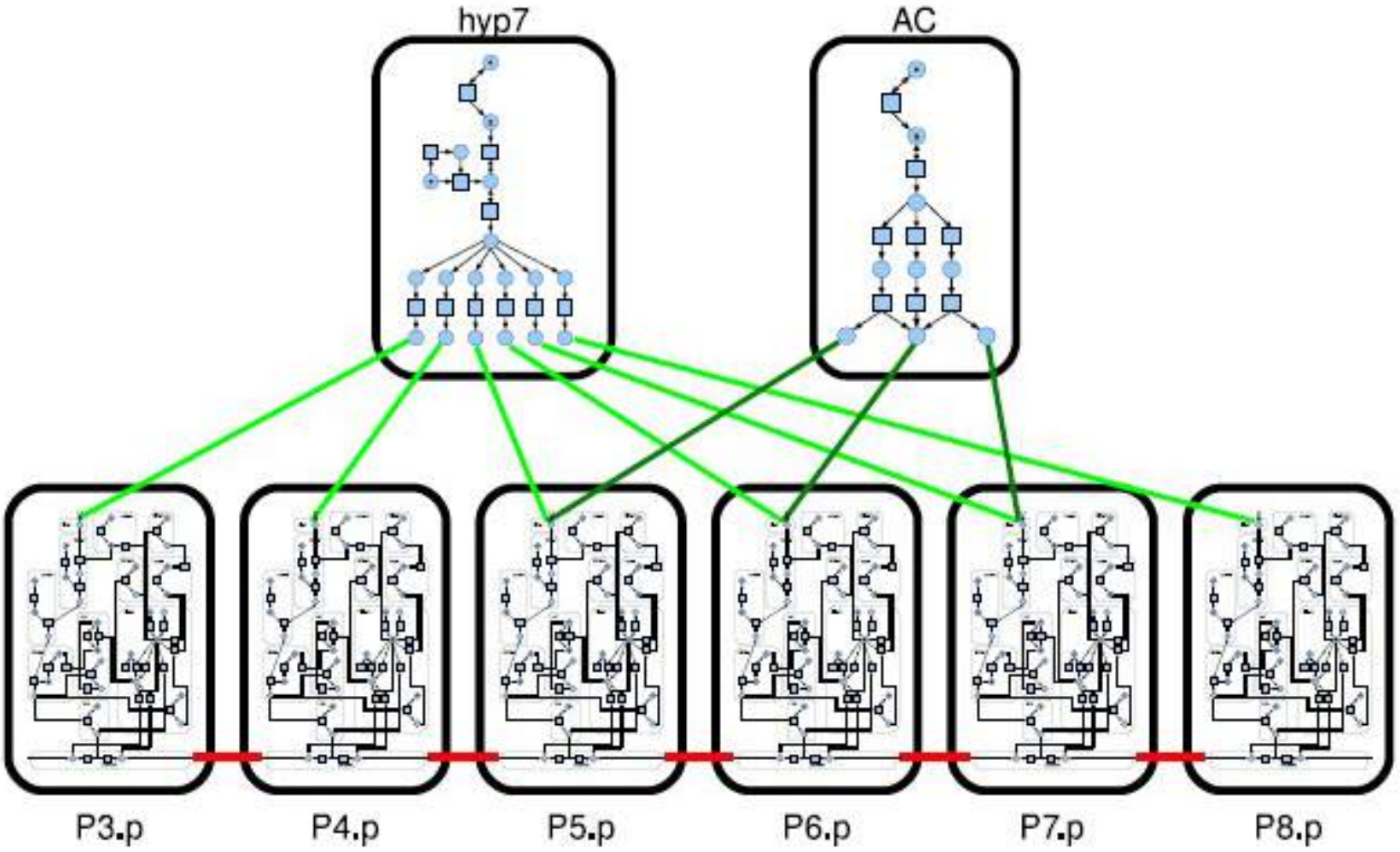
Anchor Cell (AC)



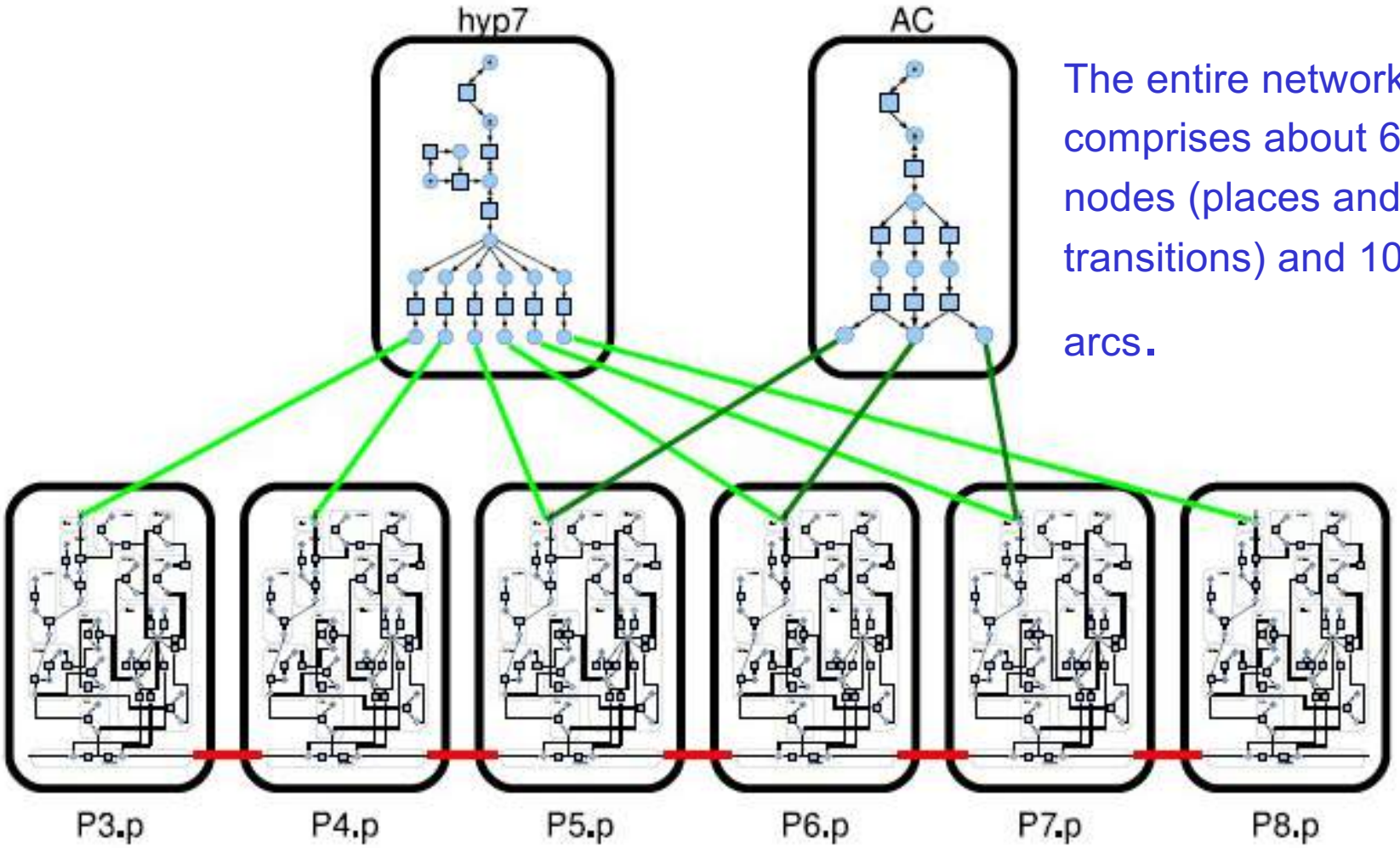
Cell fates and the onset of the vulva



Petri Net Model of *C. elegans* Vulval Development



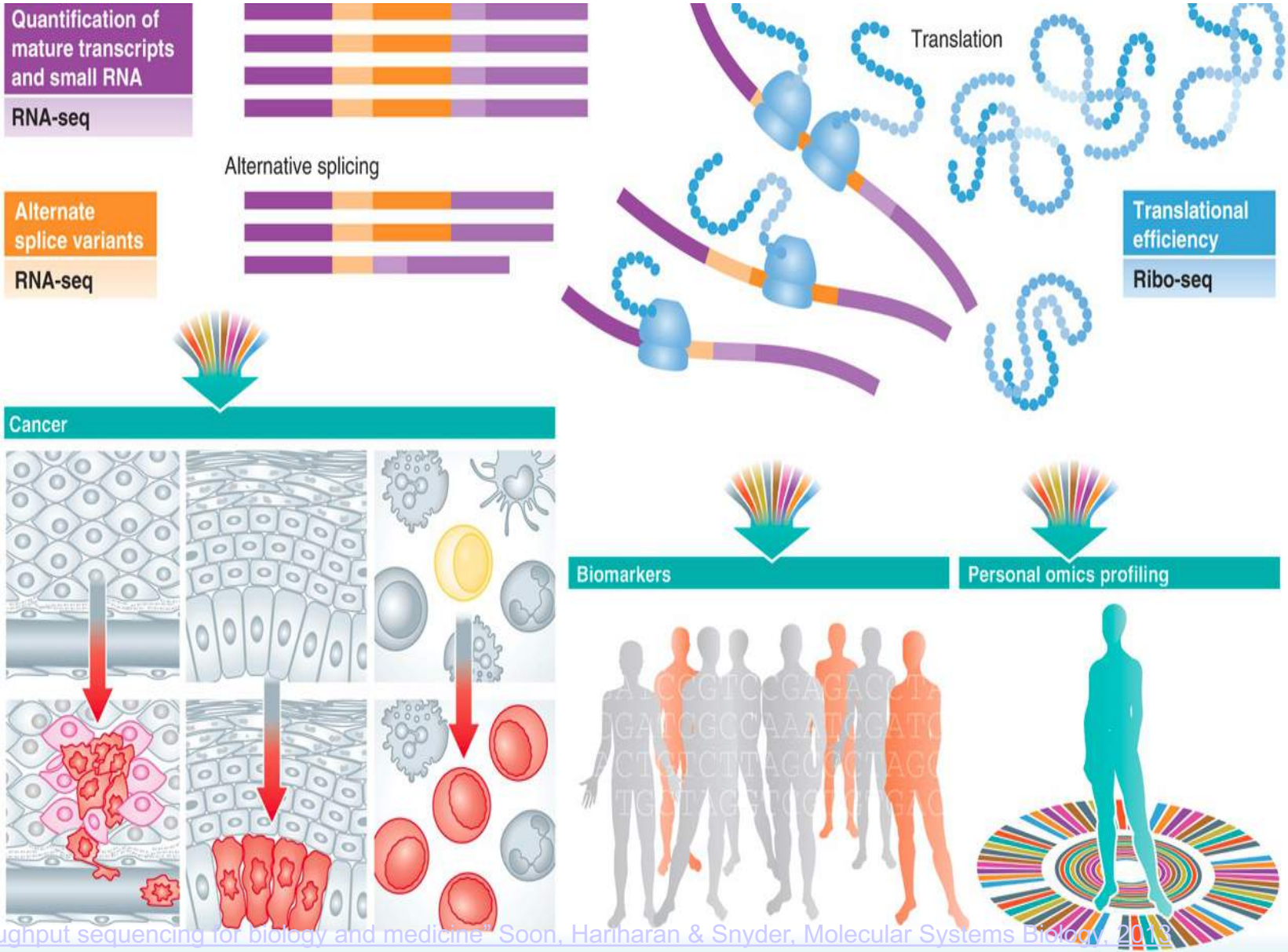
Petri Net Model of *C. elegans* Vulval Development



The entire network comprises about 600 nodes (places and transitions) and 1000 arcs.

The multicellular model contains genes, proteins and mi-RNA, and modelled functionalities such as protein production, interaction, downregulation, degradation and signalling (transport) through time.

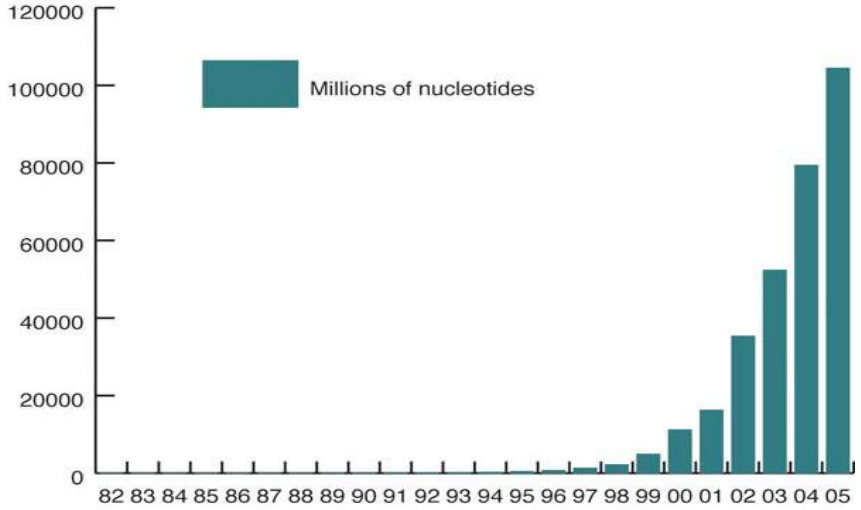
NGS and cancer: Which genes cause it

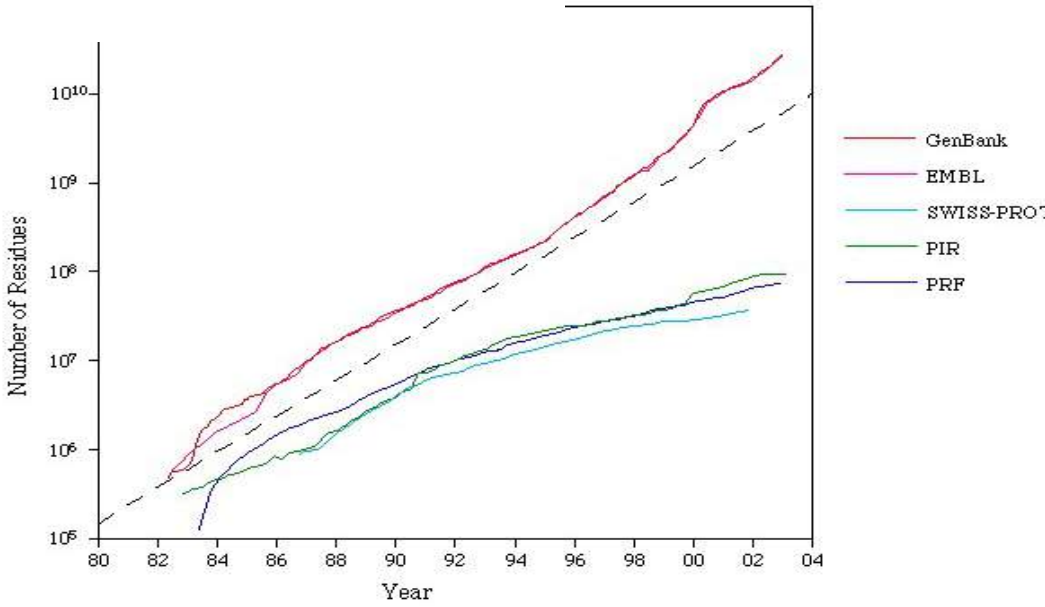


"High-throughput sequencing for biology and medicine" Soon, Hariharan & Snyder, Molecular Systems Biology, 2013

The data tsunami

- Exponential growth of databases



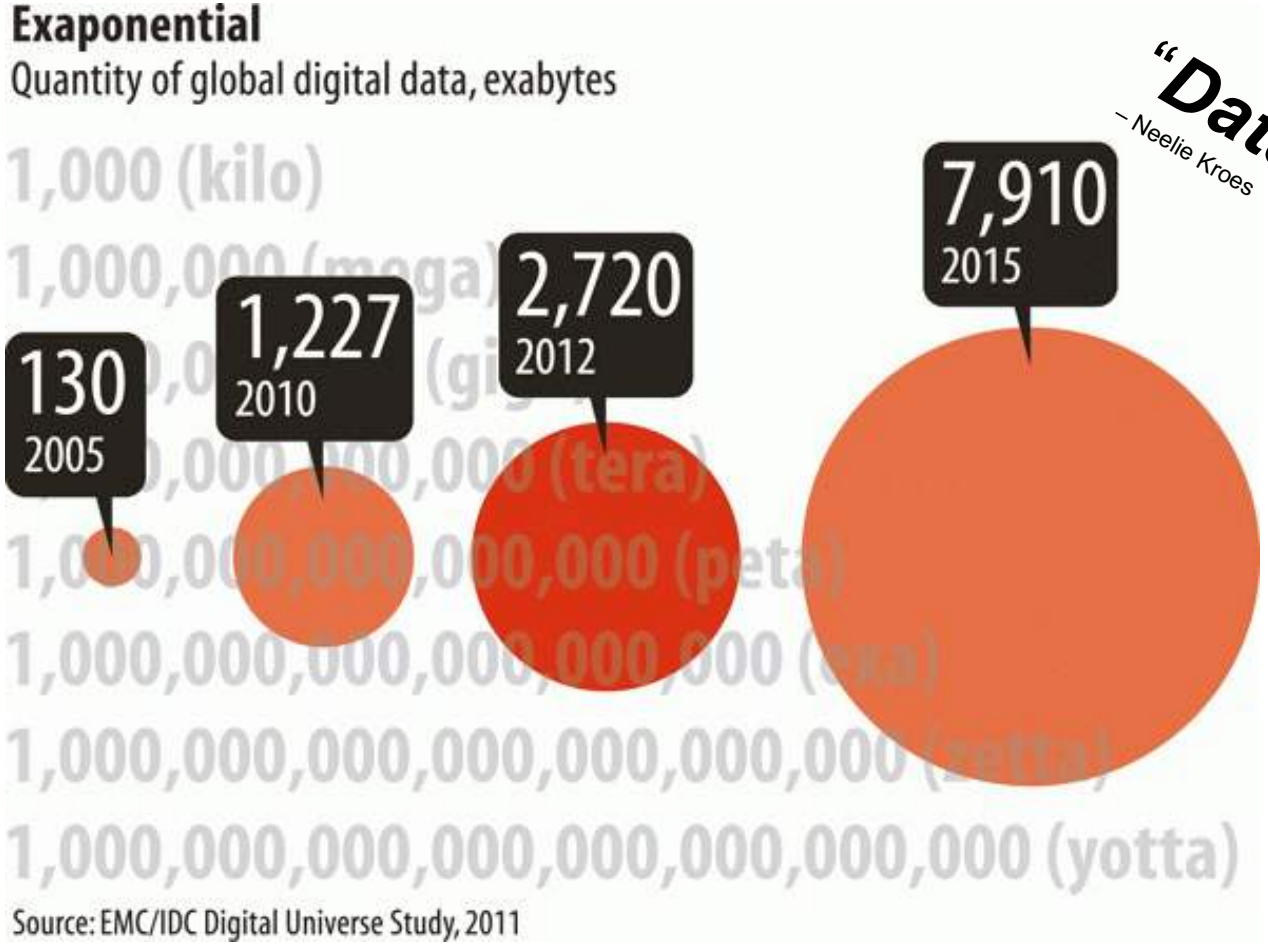


Straight line implies exponential growth

The Economist on the data tsunami..

Welcome to the yotta world

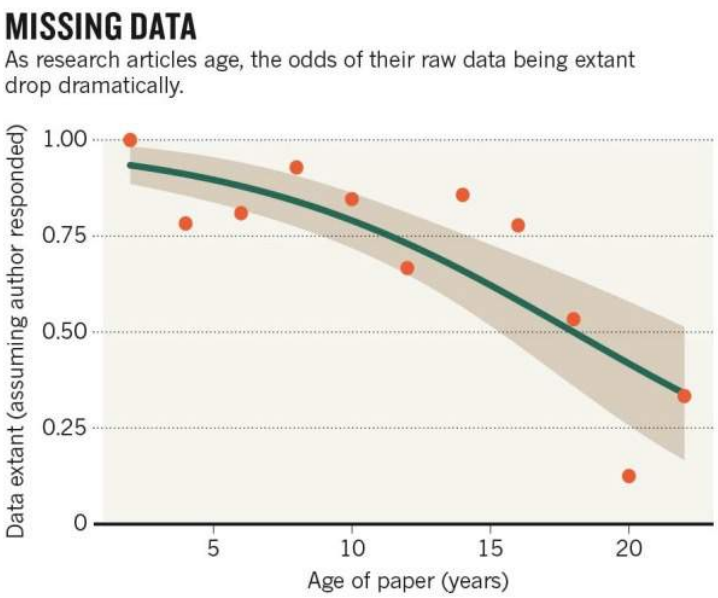
Big Data will flood the planet



"Data is the new oil"
– Neelie Kroes

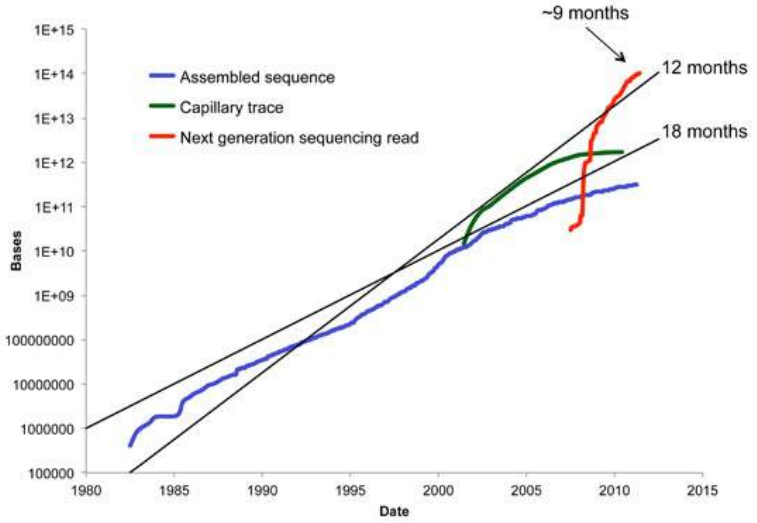
http://www.economist.com/node/21537922

BIG DATA: TWO PROBLEMS - DATA LOSS AND DATA GROWTH



Nature news, 19 December 2013

'Oops, that link was the laptop of my former PhD student'



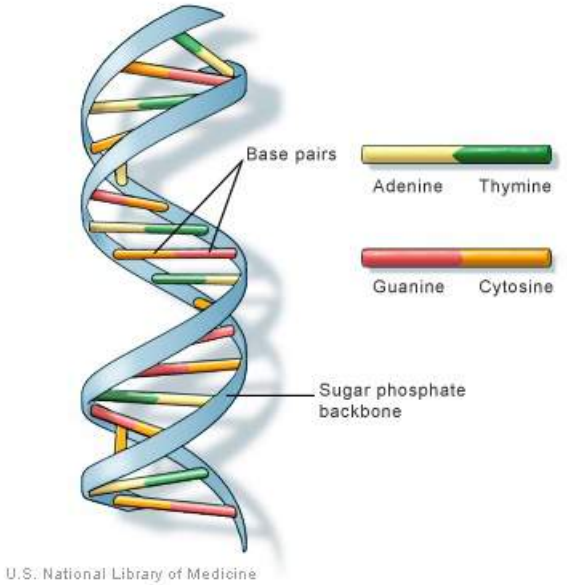
- Computer speed and storage capacity is doubling every 18 months and this rate is steady (Moore's law)
- DNA sequence data is doubling every 5-6 months over the last 3 years and looks to continue for this decade

The champion of data storage?

- Storing all data of 2020 (50 zettabytes)?
- Ultra modern disk technology?
- Or a molecule that evolved over about 4.2 billion years…

The champion of data storage?

- Storing all data of 2020 (50 zettabytes)?
- Ultra modern disk technology?
- Or a molecule that evolved over about 4.2 billion years…



DNA can store 1 yottabyte of data on roughly 1 gram!

George Church, Harvard Univ.

1 zettabyte = 10^{21} bytes, 1 yottabyte = 10^{24} bytes

The champion of data storage?

- Storing all data of 2020 (50 zettabytes)?
- Ultra modern disk technology?
- Or a molecule that evolved over about 4.2 billion years…

Reading out the information is getting better and better (sequencing), but 'writing' DNA is still problematic

DNA can store 1 yottabyte of data on roughly 1 gram!

George Church, Harvard Univ.

1 zettabyte = 10^{21} bytes, 1 yottabyte = 10^{24} bytes

Information sciences are fundamentally changing the world

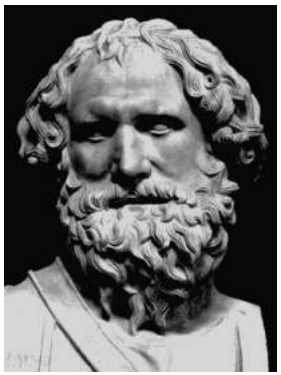
- Through (information) technology
  - Political, societal (technology application)
  - Life sciences (bio-based economy)
  - Health and quality of life

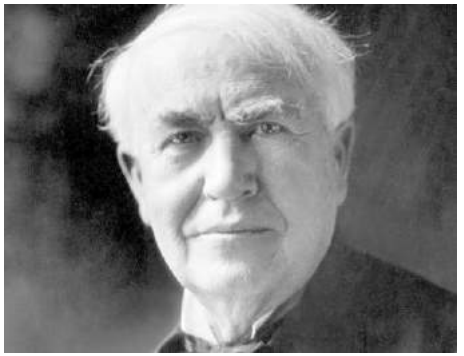
Where are we heading?

- Finding subatomic particles (Higgs boson)
- Large-scale surveillance
- Predicting longer term weather, landslide, earthquake (e.g. DeepMind)
- Predicting spread of disease (Google can already do flu)
- Social trends
  - Rapper Jay-Z in 2015 moved concert from Stockholm to Gothenburg as Spotify's big data analysis proved a larger fan base there

Data

"Too often we forget that genius, too, depends upon the data within its reach, that even Archimedes could not have devised Edison's inventions." Ernest Dimnet.





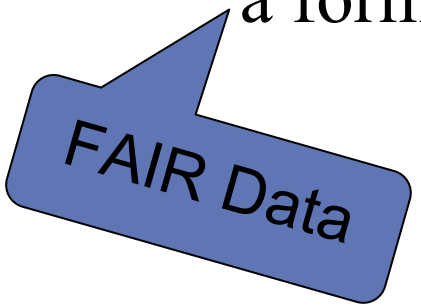
THE FAIR DATA PRINCIPLE

- Data should be
  - **F**indable, **A**ccessible, **I**nteroperable, **R**eusable

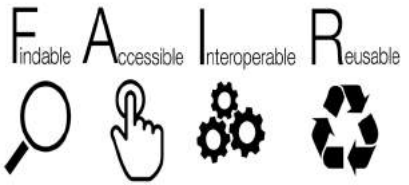
  (Barend Mons, LUMC & DTL)
- Publishing existing and new datasets in **semantically interoperable format** that can be understood by computer systems.
- By **semantically annotating data items and metadata**, we can use computer systems to (semi) **automatically combine different data sources**, resulting in richer knowledge discovery.
- **Metadata** all important

FAIR DATA STEWARDSHIP

- Combination of all expertise to treat data well and durable in a project and beyond:
  - Experiment design **and data-design**;
  - **Re-use** of existing data where possible;
  - Planning of the **storage**, **networking** and **computing** infrastructure;
  - **Data acquisition and processing**;
  - **Data publishing in a format that allows functional interlinking of data(sets)** as well as in a format suitable for long-term preservation.



FAIR DATA



- 2014: FAIR (Findable, Accessible, Interoperable, Reusable) data principles launched at Leiden Lorentz meeting (DTL driven)
- 2016: G20 adopt FAIR Principles
- 2017: Open European Science Cloud (EOSC) stipulates FAIR principles
- 2017: G7 adopt FAIR principles
- 2017: ELIXIR ESFRI bases its platforms on FAIR principles
- 2017: Science funders (e.g. NWO in The Netherlands) stipulate adherence
- 2017: GO-FAIR initiative endorsed by Dutch, German and French Governments
- 2019: Open Science rolled out across Europe

WHAT IS FAIR DATA?

FAIR Data aims to support existing communities in enabling valuable scientific data and knowledge to be published and utilised in a 'FAIR' manner.

Findable- (meta)data is uniquely and persistently identifiable. Should have basic machine readable descriptive metadata.

Accessible - data is reachable and accessible by humans and machines using standard formats and protocols.

Interoperable - (meta)data is machine readable and annotated with resolvable vocabularies/ontologies.

Reusable - (meta)data is sufficiently well-described to allow (semi)automated integration with other compatible data sources.

Machines should be able to understand the data!

WHAT IS FAIR DATA?

FAIR Data aims to support existing communities in enabling valuable scientific data and knowledge to be published and utilised i

th

sources.

Wilkinson et al,
Nature Scientific Data, 2016

15 FAIR DATA PRINCIPLES SINCE 2016

To be Findable:

F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

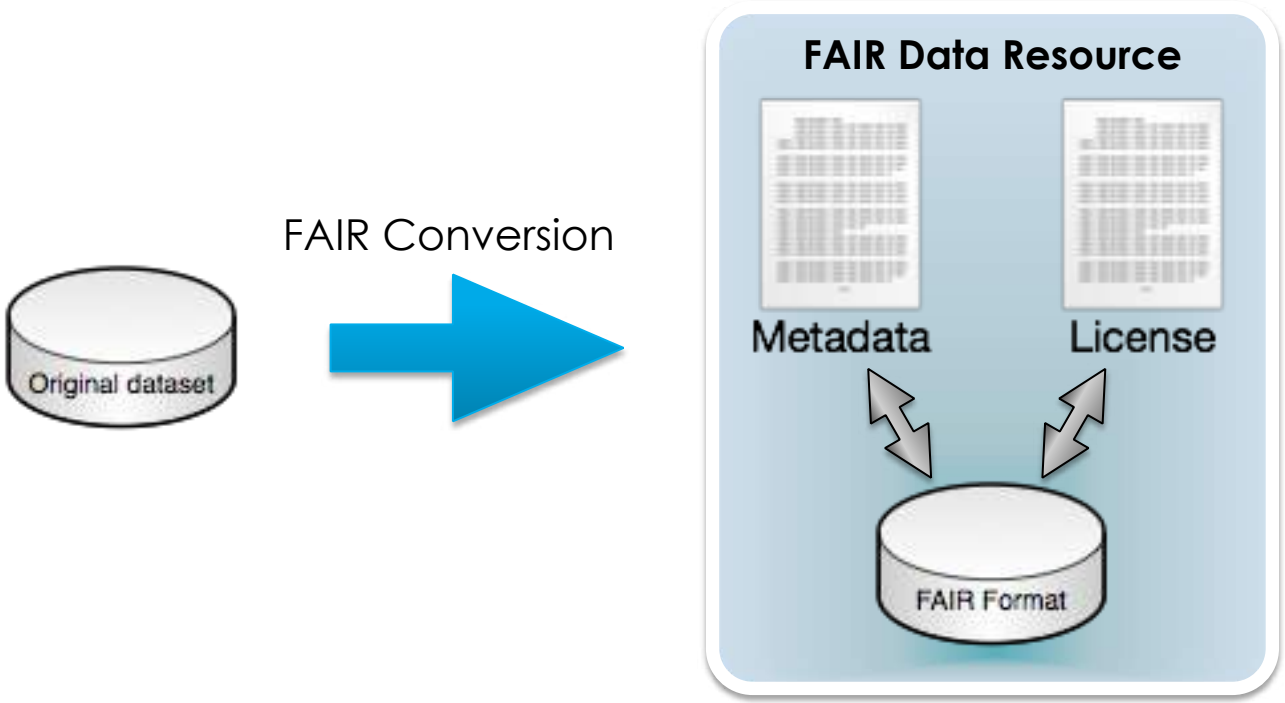
To be Reusable:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

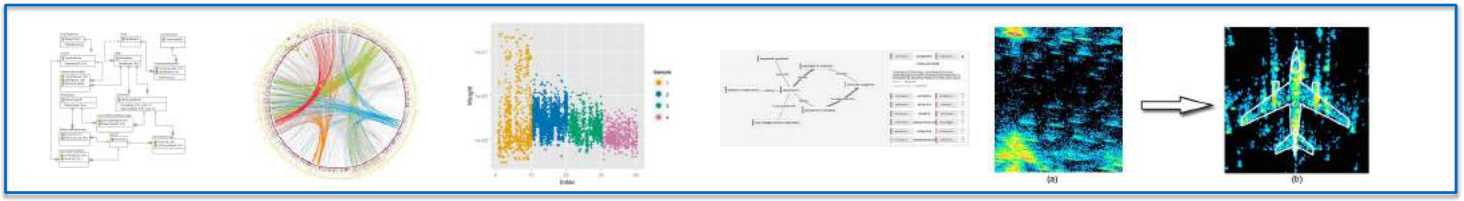
FAIR DATA RESOURCE

Datasets expressed using one of the prescribed standards of the FAIR Data Protocol.

The original dataset is transformed into a **FAIR format** and proper **metadata** and **license** are added to produce a FAIR Data Resource. Original and the FAIR version can co-exist, each one fulfilling its own purpose.

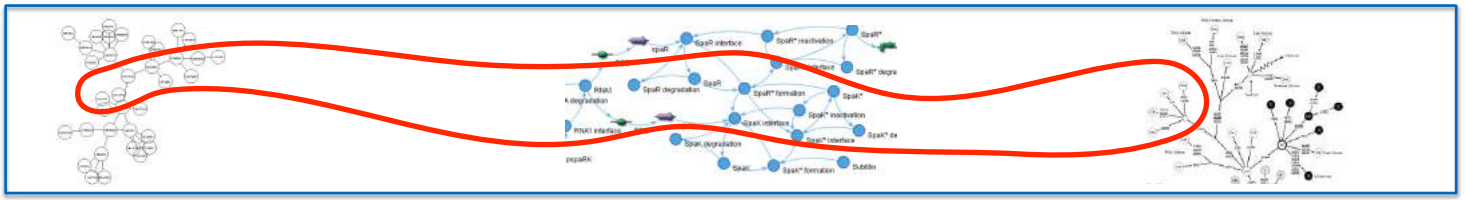


High-Performance Analysis



Analysis transformation

FAIR (meta)data
(RDF,XML etc.)



FAIR transformation

FAIR download
(in local format)

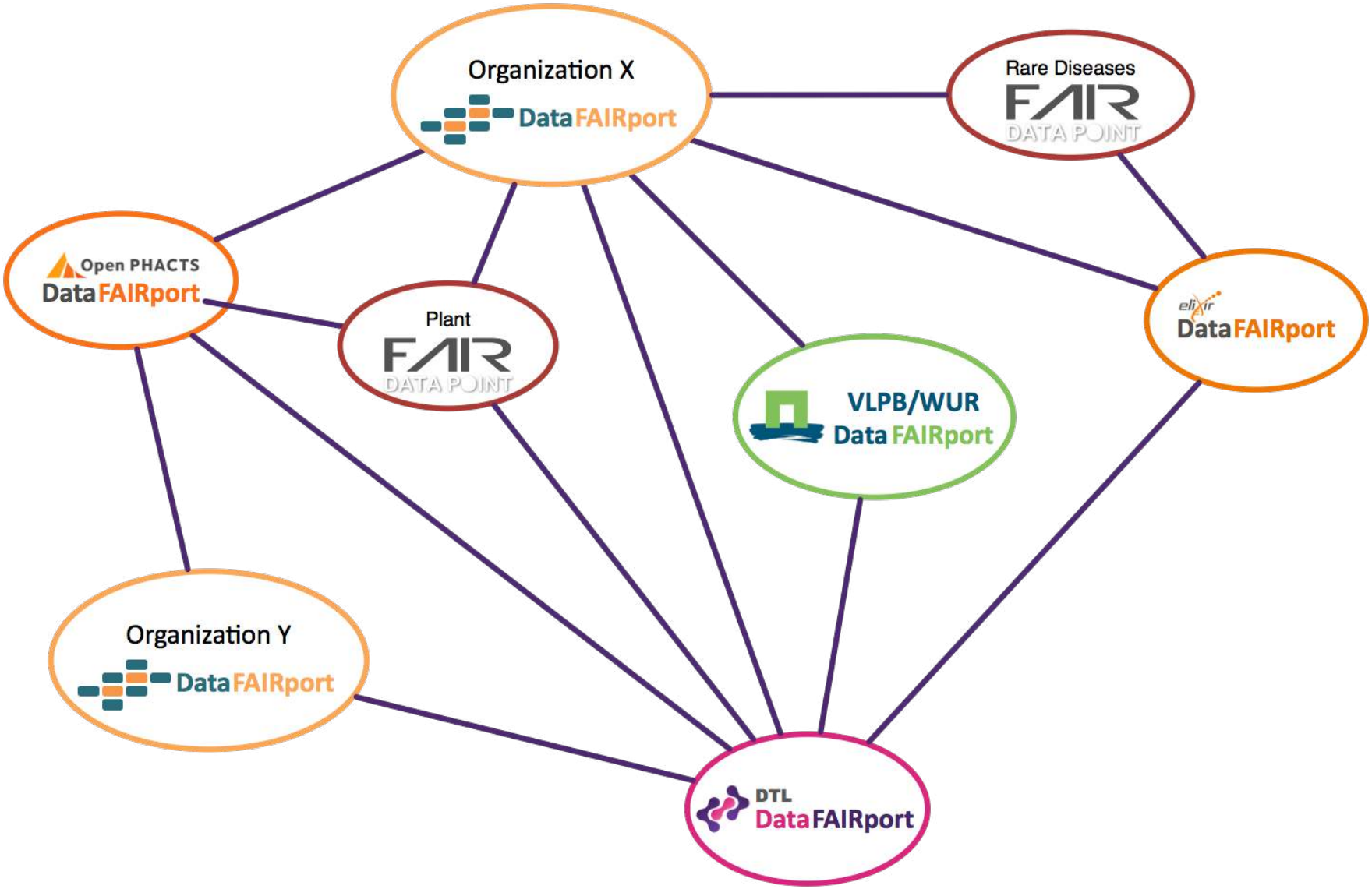
Processed data
(primary storage format)

Initial transformation

Provenance

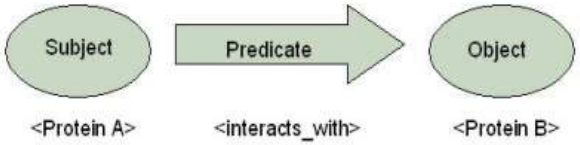
Raw data
(many formats)

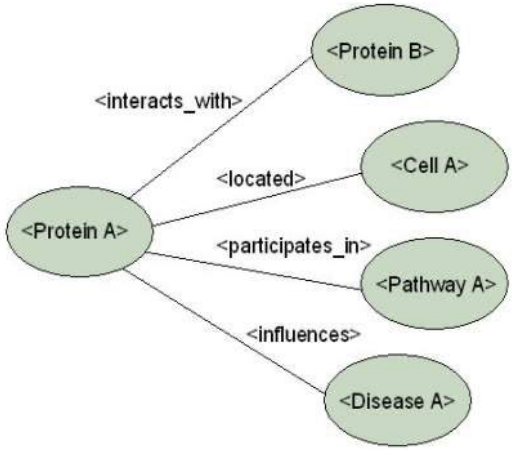
DISTRIBUTED ARCHITECTURE OF FAIR DATA POINTS



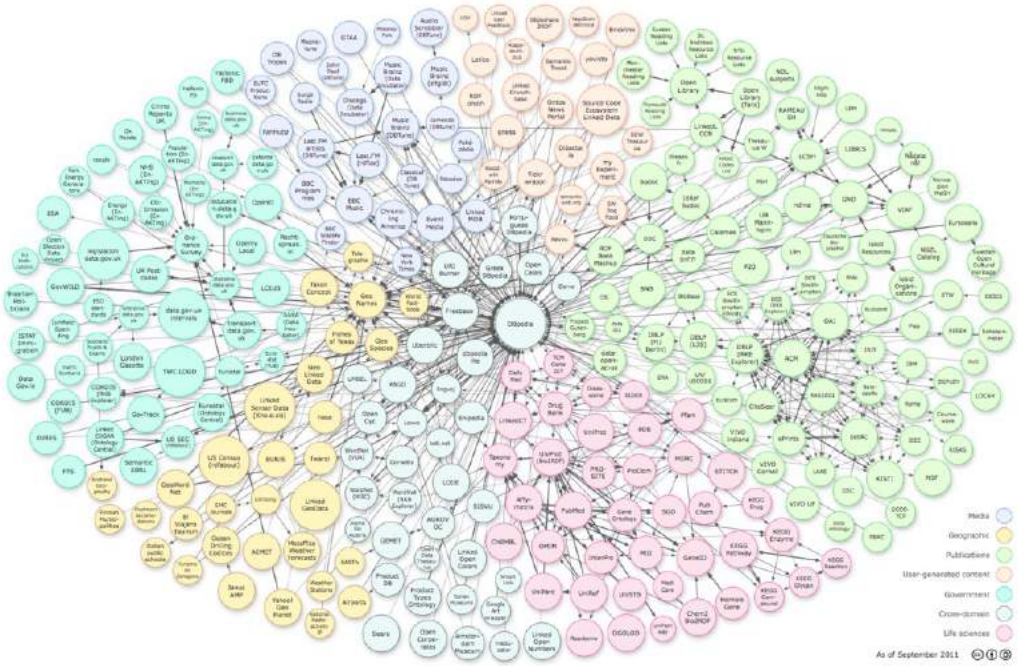
SEMANTIC INTEROPERABILITY THROUGH RDF (RESOURCE DESCRIPTION FRAMEWORK)

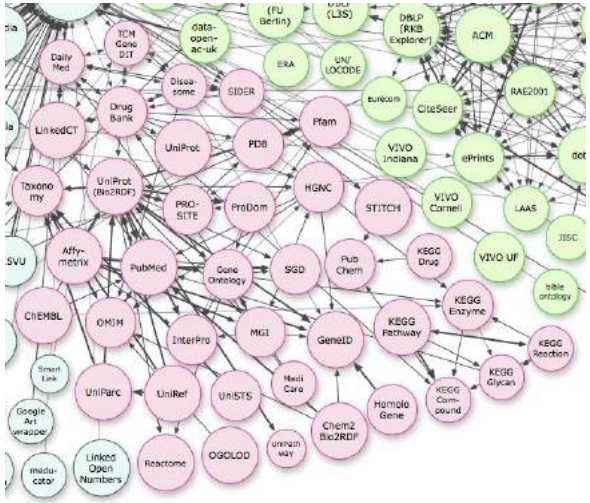
Unit of information: the data triple





Linked data cloud

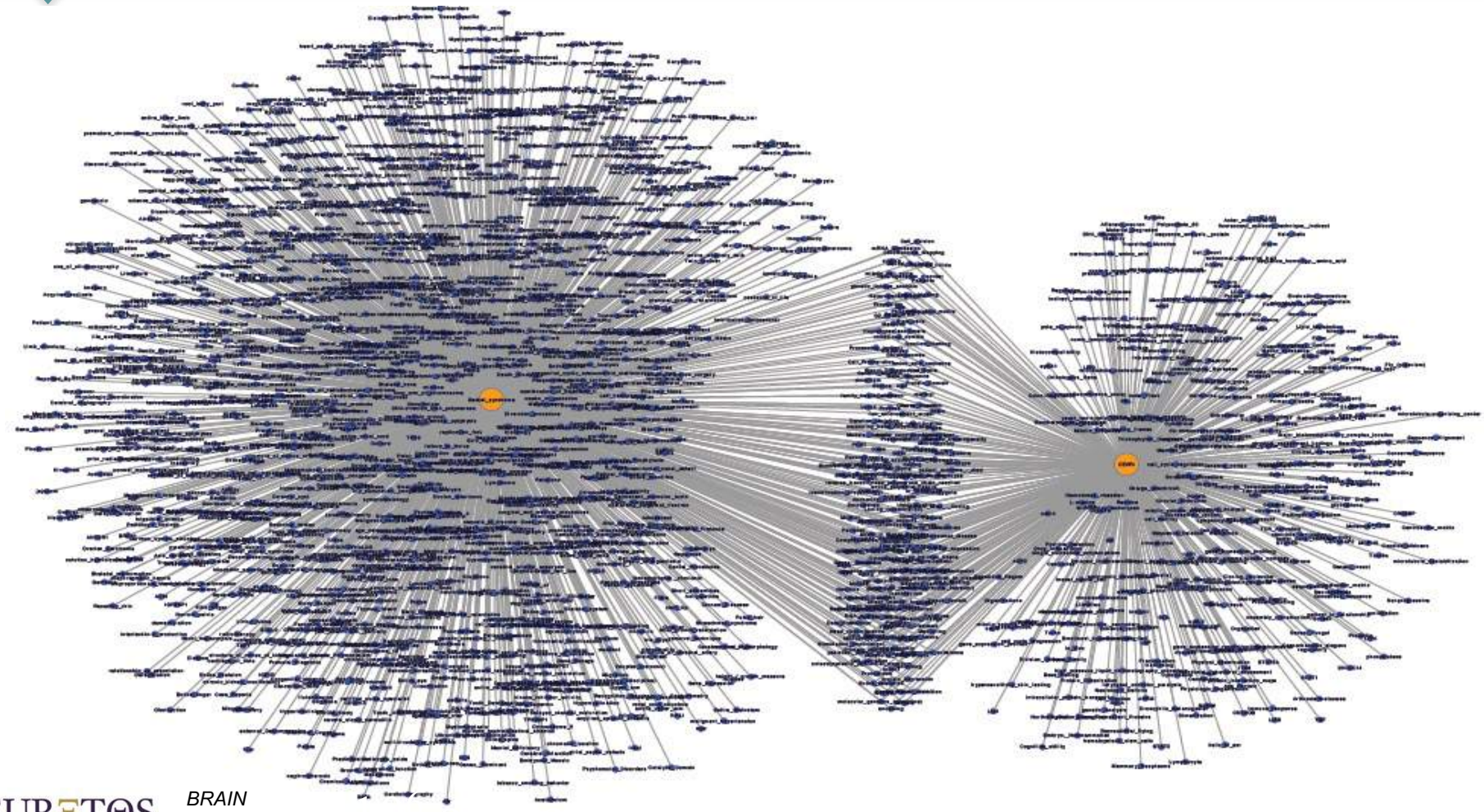




Data interoperability

VU UNIVERSITY AMSTERDAM Faculty of Sciences

TRAFFICKING THE DATA HIGHWAY: THE POWER OF INTEROPERABILITY



BRAIN

EURETOS

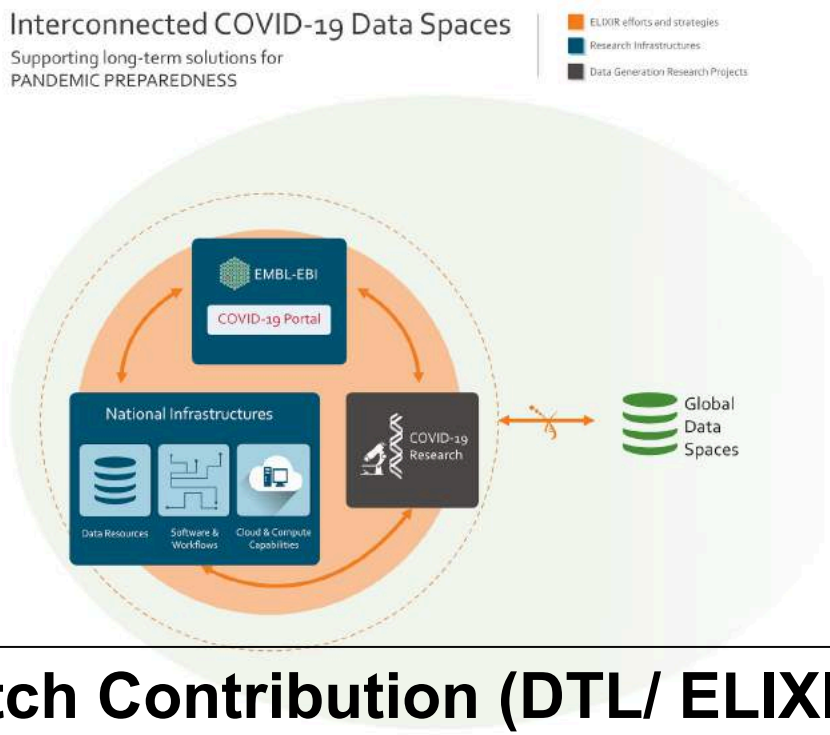
http://www.euretos.com/

DTL DUTCH TECHCENTRE FOR LIFE SCIENCES

VU UNIVERSITY AMSTERDAM Faculty of Sciences

EU data infrastructure for Covid-19

ELIXIR, EMBL-EBI and EOSC-Life response to COVID-19



Priority is to drive open and rapid access to data, tools and workflows for the European COVID-19 response and research

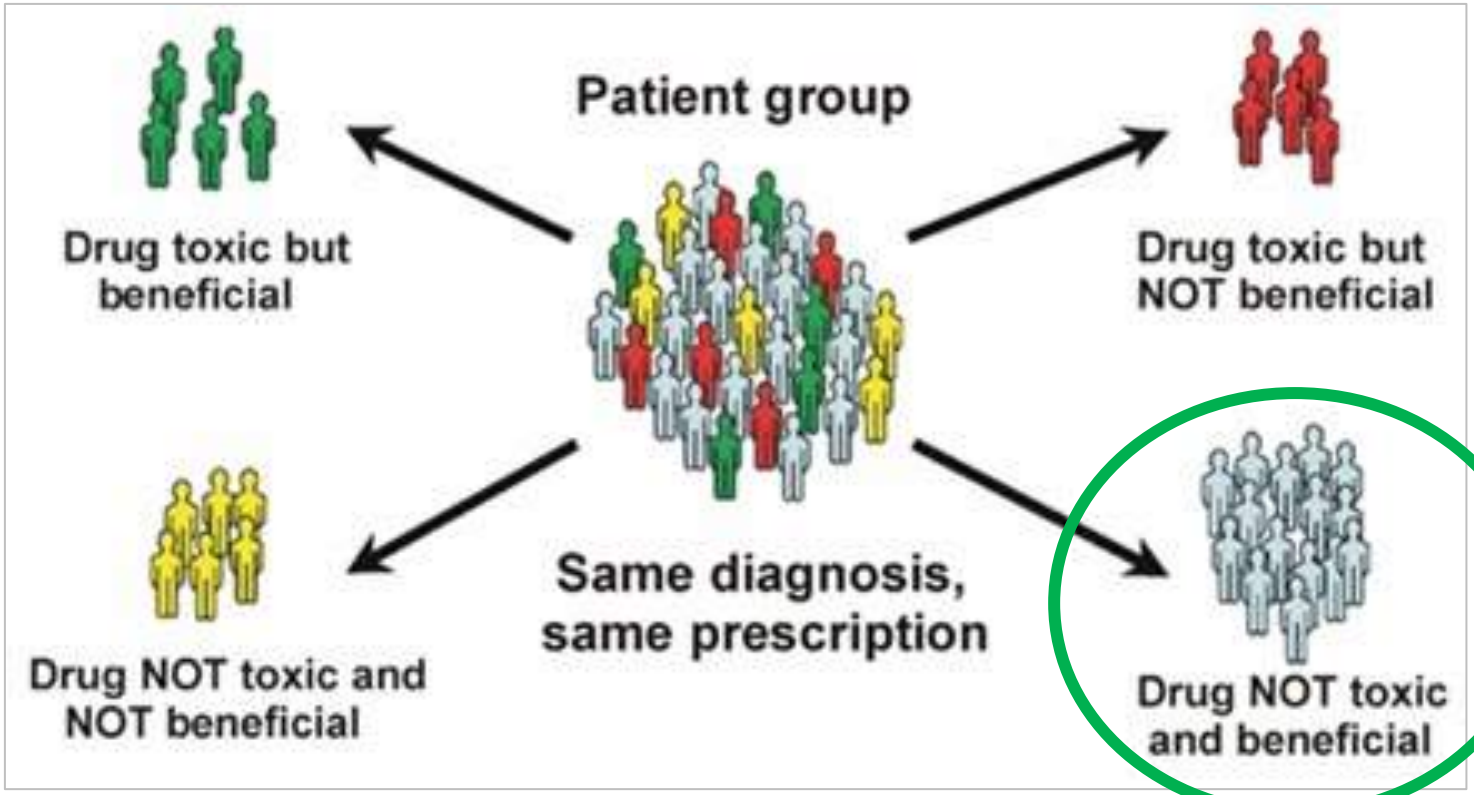
We will achieve this via alignment of national infrastructures, European research infrastructures (e.g. EMBL-EBI) and H2020 projects

Long-term sustainable solutions, build on open standards and aligned with EOSC

Dutch Contribution (DTL/ ELIXIR-NL and GO FAIR):

- semantic data model based on the Case Report Form (CRF) model following the WHO standards.
- VODAN-in-a-box: FAIR Data Points (6 African countries have FDPs installed)
- Dutch UMCs connected via FDPs

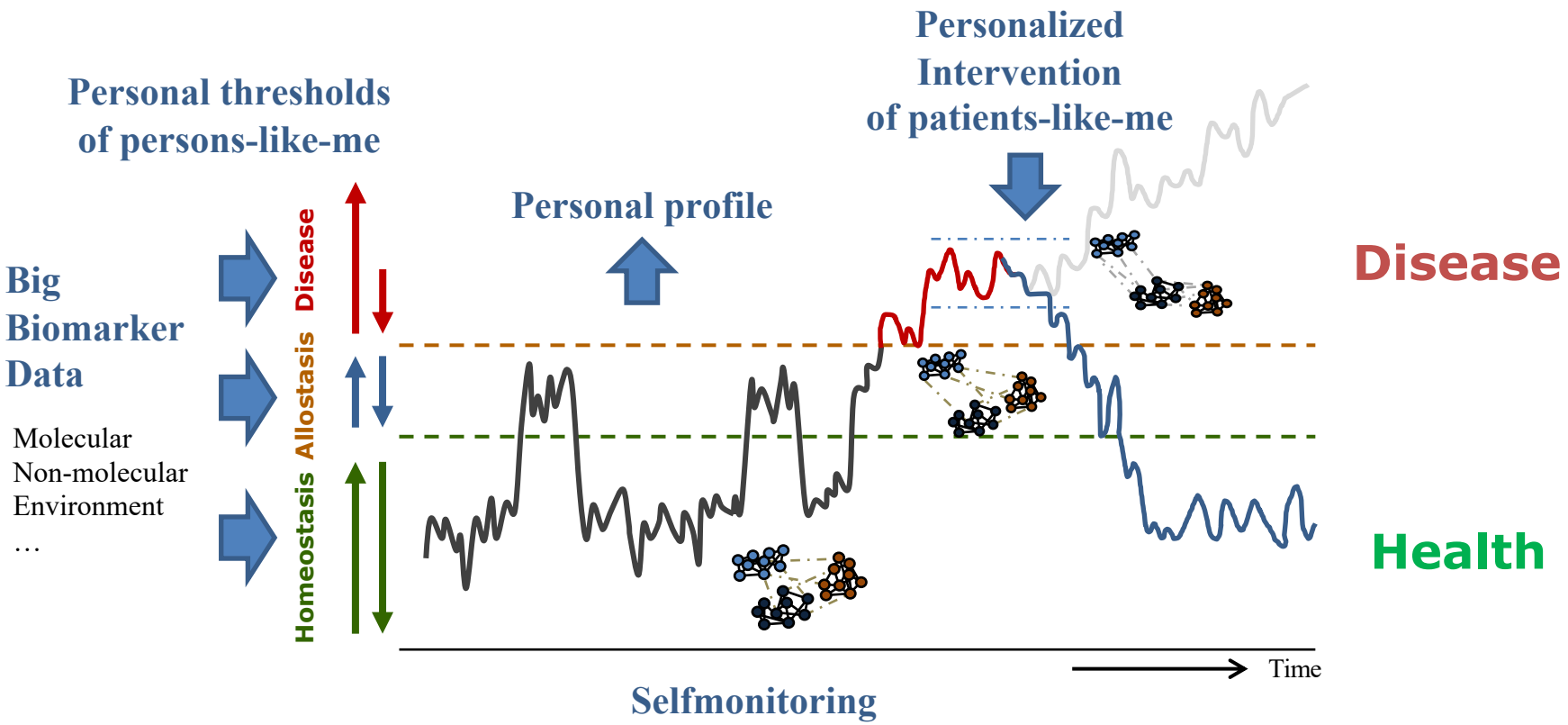
Principle of Personalized Medicine



Source: Chakma, Journal of Young Investigators, 16, 2009

- The **right drug** for **right patient** at **right dose** at **right time**
- Molecular biomarkers as key drivers of patient selection
- = Precision medicine or Targeted medicine

Personalized health(care) model



Adapted from Jan van der Greef, TNO

{See eg Chen … Snyder, Cell 2012, 148: 1293}

Personalised Medicine

- Based on genotyping (genomics)
- Combining Medicine (combination drugs), Nutrition and Lifestyle
- Big data generated by –omics technologies, patient data, wearables
- Increasingly based on exposome
  - defined as “the measure of all exposures of an individual during a lifetime and how these relate to health”
- Deal with (data) privacy issues and GDPR

What do others say?

Prof. Peter Coveney
Physical chemist and director of the
Centre for Computational Science at UCL

“In such a forward-looking field as this, you can only make advances if you know both material science and computer science. **You can’t get away with being an expert in just one area anymore.** Old-fashioned chemistry cooking is over.”

Wrapping up

- Bioinformatics has a history already
- A wide scope
- Science of big numbers
- Algorithms have lots of scalability problems
- Modelling as a crucial analytics tool (systems biology, metabolomics)
- Data stewardship is crucial
  - Long term preservation of public data
- FAIR data principles (Findable, Accessible, Interoperable, Reusable)
- Crucially important societal application of big data interoperability: **Personalised/precision medicine**

health RI research infrastructure

Radboudumc TNO innovation for life DTL eatris

Health-RI 4-minute movie at

https://www.youtube.com/watch?v=FoWqSZeaOxs

Personal Health Train (PHT) 3-minute movie at
https://www.youtube.com/watch?v=mktAtHmy-FM