

Bioinformatics and the Cloud

QUESTnet 2013

2-5 July, Royal Pines Resort, Queensland

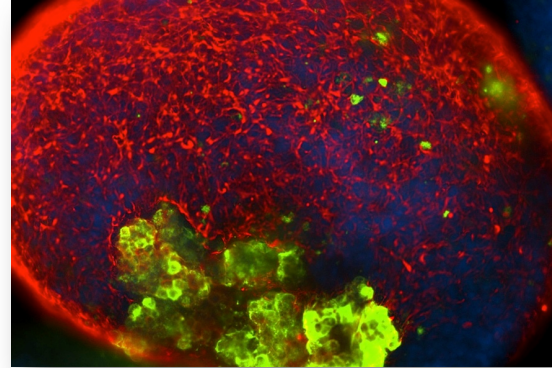
Andrew Yates
Ensembl Core Project Leader

European Molecular Biology Laboratory,
European Bioinformatics Institute

What is Bioinformatics?



© Martin Abegglen CC
<http://www.flickr.com/photos/twicepix/3332881889>



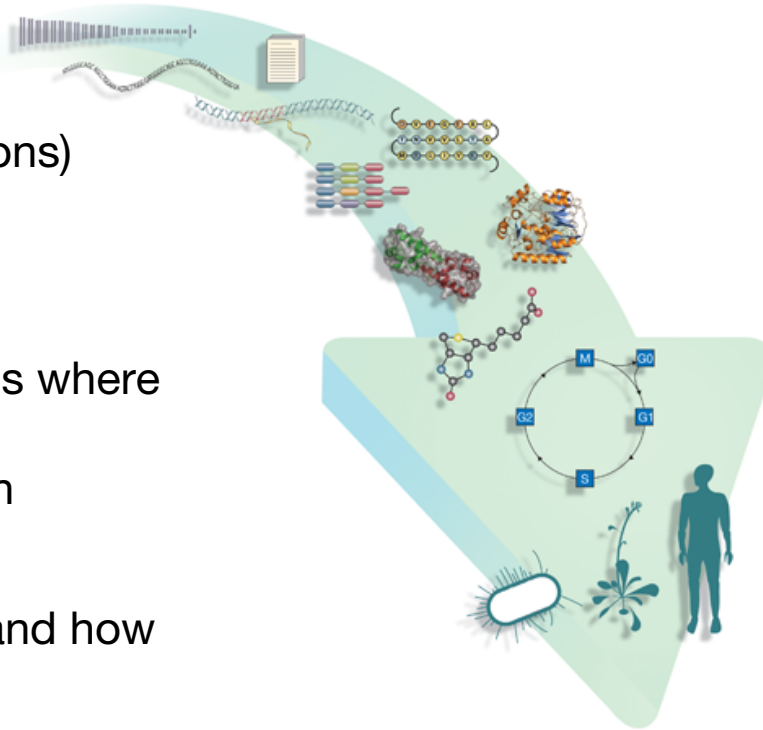
© UC Irvine CC
<http://www.flickr.com/photos/ucirvine/8126968874>



How can Computing and Biology enable each other?

Understanding Life – from molecules to systems

- Parts Dictionary
- Thesaurus (interactions)
- Complete Networks (Circuit Diagrams)
- Atlas – what happens where
- How it works & when (simulations)
- How it goes wrong and how to put it right



What is EMBL-EBI?

- Provides data and software support for Bioinformatics
- Part of the European Molecular Biology Laboratory
- International non-profit research institute
- 21 members states plus associate
- 1,500 members of staff across 5 outstations
- European Bioinformatics Institute is a UK outstation



What Does EMBL-EBI Provide?

- Archives
 - Records of scientific publications and output
 - Ensures stability future reproducibility
- Value-added resources
 - Usually built from archived data
 - Enables science
 - Data analysis by world leaders
- Research and special projects
 - Investigating biology, outreach & training to text mining

- 1st associate member of EMBL
- Joined in 2008
- Promotes excellence in molecular biology in Australia
- BRAEMBL (Bioinformatics Resources Australia – EMBL)
 - Institute for Molecular Bioscience (IMB) at University of Queensland
 - Currently mirrors 13TB of EBI data where beneficial & practical
 - Without local expertise Australian scientists cannot make novel insights

Bioinformatics – Understanding DNA

- 1st genome sequenced in 1976
 - bacteriophage MS2 – 3,569 base pairs long
- Human genome released in 2000*
 - 3.2bp (billion base pairs)
- 1,252 complex organism genomes now sequenced



*Human Genome Printout at the
Wellcome Collection, London.
© Russ London at en.wikipedia*

* Continually refined since then

DNA – How Can We Use It?

DNA extraction

Sequencing

ACTGTCGATCGATA ACTGTCGATCGATA
 ACTGTCGATCGATA ACTGTCGATCGATA
 ACTGTCGATCGATA ACTGTCGATCGATA

DNA reads

$$[1]$$


New Genomes

[2]



Diagnostic



Population Genetics

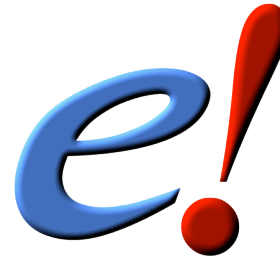
[1] © Alfredo Gutiérrez, [2] © Menna Jones taken from en.wikipedia

EMBL-EBI 

What am I presenting today?

- 3 case studies of how clouds are aiding bioinformatics
- Two as a cloud consumer
 - Ensembl and Amazon Web Services - Content Distribution
 - Helix Nebula – Cloud Pipelines
- One as a cloud provider
 - Embassy Cloud - IaaS
- A novel method of data archiving

Case Study One - Ensembl & AWS



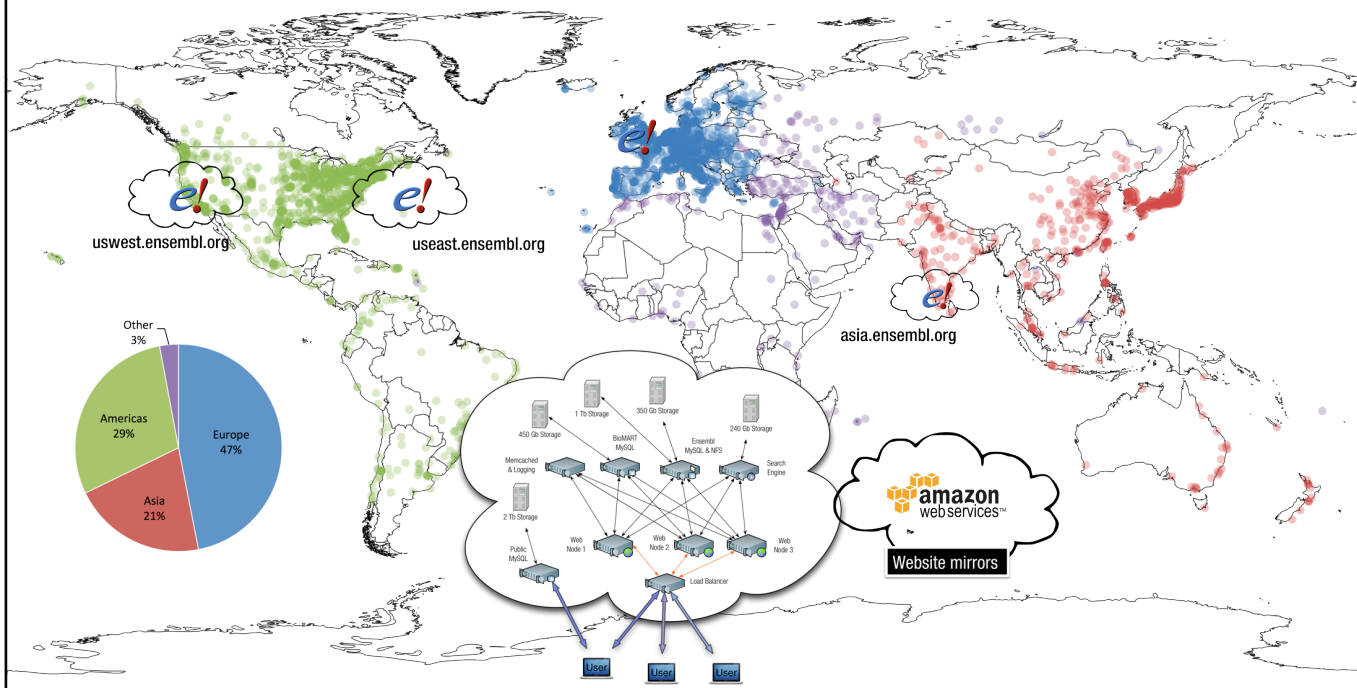
- Ensembl is a world leader in the provision of genome data and annotation
- Based at EMBL-EBI and the Wellcome Trust Sanger Institute, Cambridge
- Launched in 1999
- Approximately 5 data and code releases per year
- Highly accessed from around the world
 - Over 20 million hits per month
 - 2 million unique visitors per year



Ensembl – The Geolocation Problem

- A west coast USA user; ~11 seconds for 1 page
- A UK user; ~2 seconds for the same page
- 2009 we launched our 1st US West Coast mirror
 - Do-it-yourself content delivery network
 - Buy hardware
 - Ship servers and an engineer to California
 - Cut our LA load time down to ~3 seconds

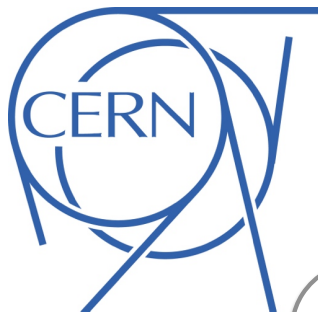
Ensembl – AWS Deployment



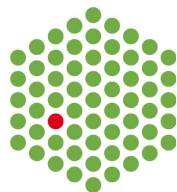
Ensembl – An AWS Success Story

- Cost
 - We can approx. host all 3 AWS mirrors for one DIY server
 - “Free” upgrades
- No need to travel
- Website redundancy
- Deployment
 - One SOP required to deploy on multiple sites
- Makes hosting an Ensembl Sydney mirror possible
- The sun never sets on Ensembl

Case Study Two – Helix Nebula



EMBL



Setup with the idea that in the future compute will be in the cloud

A European cloud computing partnership: big science teams up with big business



Strategic Plan

- ▶ Establish multi-tenant, multi-provider cloud infrastructure
- ▶ Identify and adopt policies for trust, security and privacy
- ▶ Create governance structure
- ▶ Define funding schemes



To support the computing capacity needs for the ATLAS experiment

EMBL



Setting up a new service to simplify analysis of large genomes, for a deeper insight into evolution and biodiversity



To create an Earth Observation platform, focusing on earthquake and volcano research

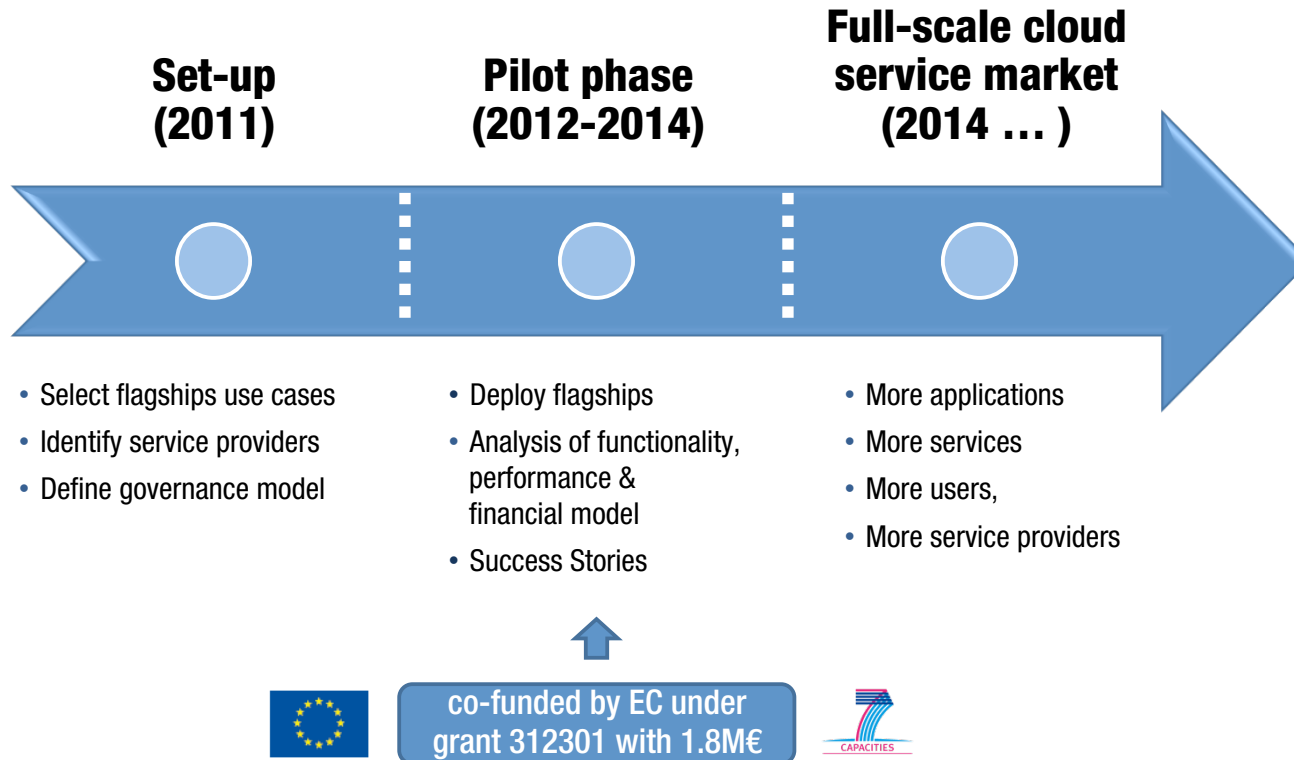


<http://www.helix-nebula.eu>
contact@helix-nebula.eu

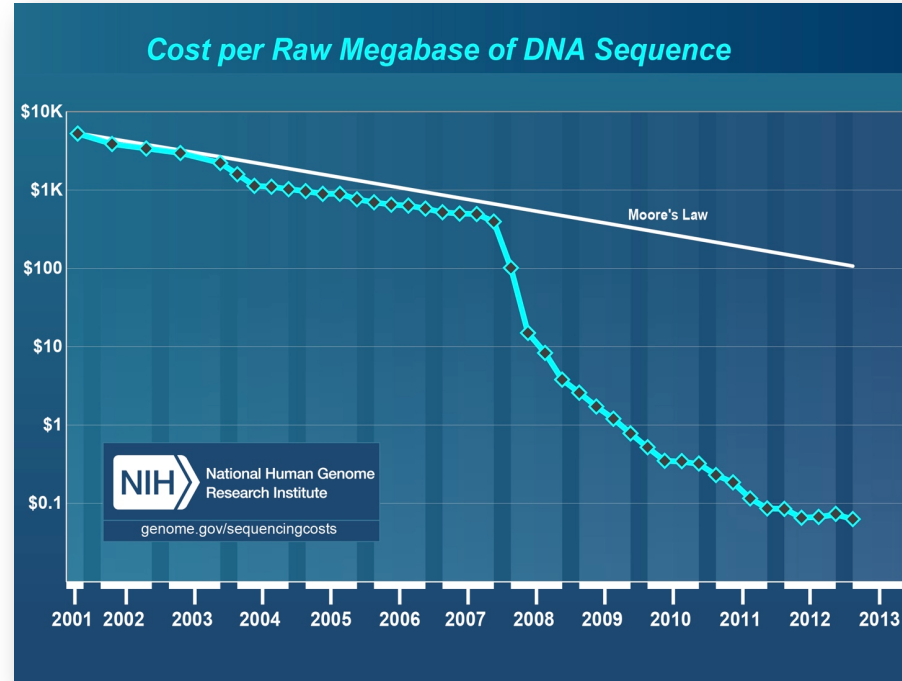


@HelixNebulaSC
HelixNebula.TheScienceCloud

Timeline



DNA Sequencing – Outcompetes Moore's Law



Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed 7th June 2013

DNA Sequencing – Data Volumes

- Population studies
 - ~7 billion *Homo sapiens*
- Sequencers are everywhere
 - ~8.7 million species in the world
 - Agriculture – disease resistant traits
 - Epidemiology – tracing infection
 - Deep sea explorations
- More than 1,500 high throughput sequencers in the world
 - 3.6PB a day
 - 1.1 – 2.2 ExaBytes a year

DNA Sequencing – Next Generation Technologies



Illumina MiSeq – a \$125K
bench-top sequencer available today

2.3 human genomes per day

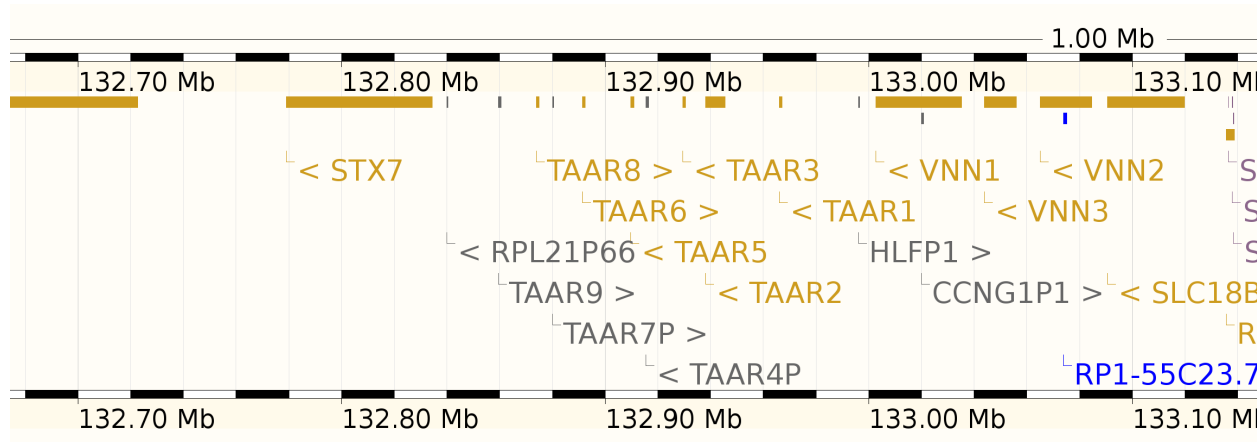


Oxford Nanopore MinION; a
\$900 laptop sequencer in
development

Tens of Gb per day

Helix Nebula – Genome Annotation

Whilst our capacity to read DNA has increased we still require a way to assemble a genome and bring context (primarily genes) to it



Ensembl has developed well respected pipelines to efficiently locate Genes on genomes

Helix Nebula –Genome Annotation as a Flagship Project

- Annotation requires expertise and compute
 - Makes it a good challenge for Helix Nebula
- Challenges to the cloud providers
 - Can they deliver the compute?
 - Can they deliver the IO?
 - Can they deliver the support?
- Can we provide pan-European genome analysis tools?

Helix Nebula – Conclusions

- All flagships have deployed scientific applications
- Each involving tens of thousands of jobs
- Running on 3 data centres **Atos** **CloudSigma** **T-Systems**
- Developing a model for future federated clouds
 - Building on lessons from the proof of concept
 - Interoperability
 - Cloud federation

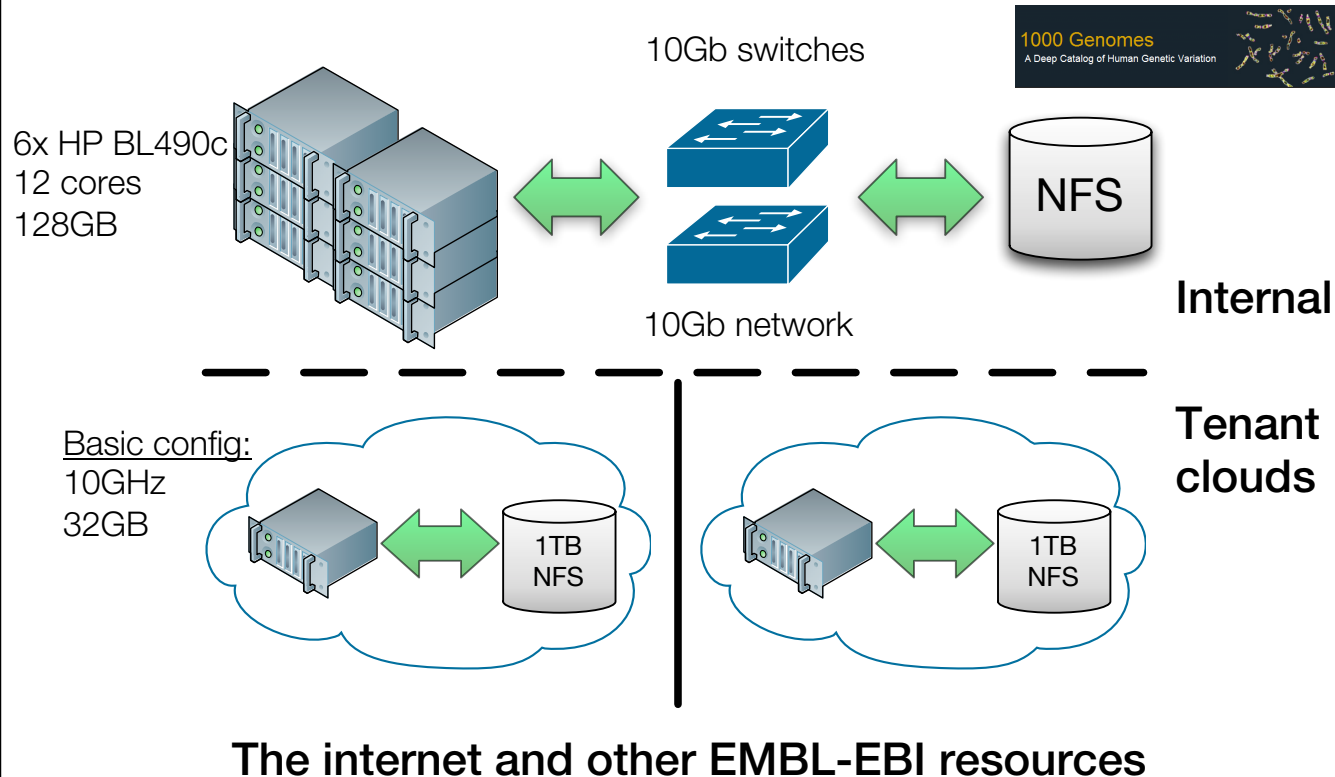
Case Study Three – Embassy Cloud

- We have very large data sets
 - 2PB of single copy data held in European Nucleotide Archive's (ENA) Sequence Read Archive (SRA)
 - 16-20PB of disk is spun at EMBL-EBI
- Geography and network still a limiting factor
 - Code is much smaller than data
- EMBL-EBI is piloting Infrastructure as a Service (IaaS)
 - Learning from other cloud providers

Embassy Cloud – Details

- Aims to provide secure, flexible infrastructure to tenant organisations close to EMBL-EBI's data
 - High bandwidth
 - Low latency
- Tenants are both academic & commercial
- Hosted at EMBL-EBI but is outside of our LAN
- Built on top of VMware's vCloud Director
 - Hypervisors are clustered meaning automatic VM restart
 - Machine maintenance without cloud downtime

Embassy Cloud - Infrastructure



Embassy Cloud – Results

- 8 organisations on Embassy Cloud for multiple uses
- One live mirror service (<http://europe.omim.org/>)
- 30 VMs are running at any one time
- Over 100 VMs have been deployed during the pilot phase
- Provides users with a viable mechanism to circumvent geography

Archiving Data in DNA

- We think about DNA a lot
- We think about long term storage a lot
- DNA is very robust
- DNA is data dense
 - 1g of DNA can store 2PB including error correction
- What if we can store data as DNA?



© Flying Puffin CC

<http://www.flickr.com/photos/flyingpuffin/5899473679/>

Archiving DNA – The Theory

LETTER

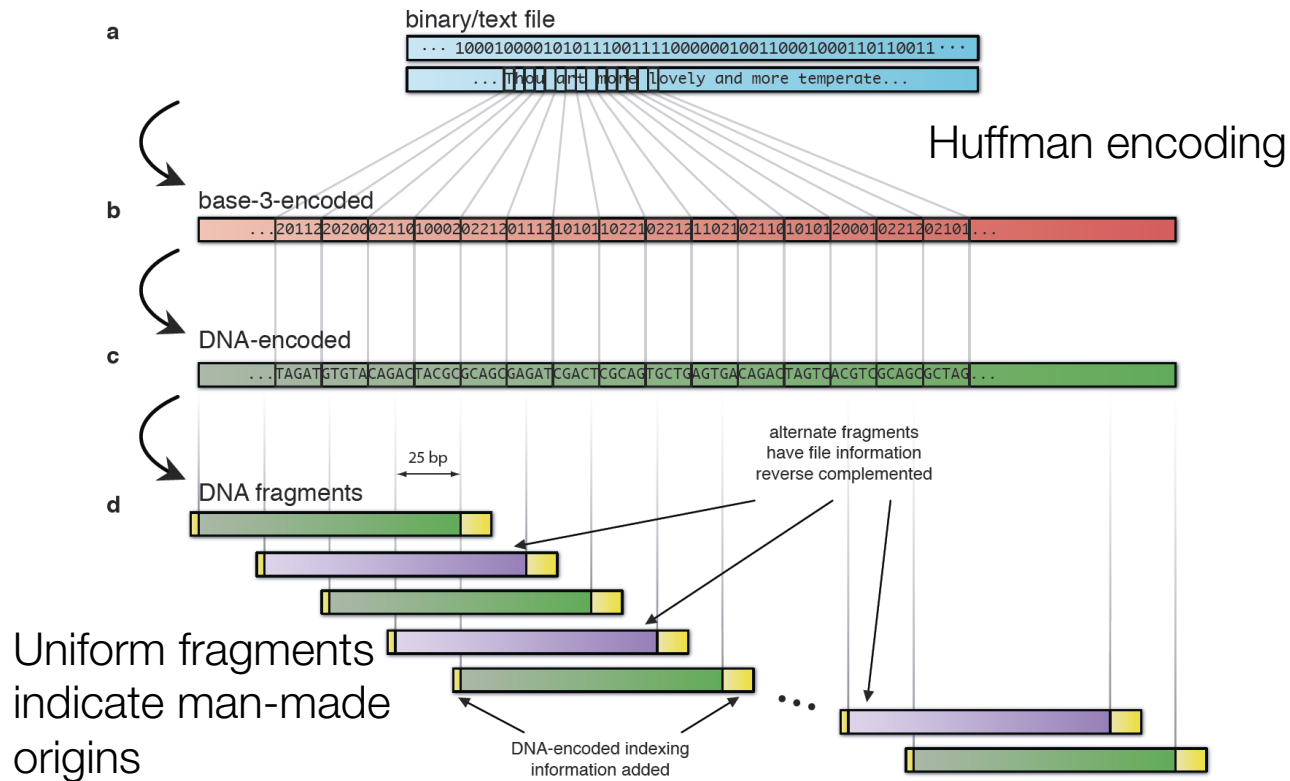
doi:10.1038/nature11875

Towards practical, high-capacity, low-maintenance information storage in synthesized DNA

Nick Goldman¹, Paul Bertone¹, Siyuan Chen², Christophe Dessimoz¹, Emily M. LeProust², Botond Sipos¹ & Ewan Birney¹

- Not the first attempt at encoding data in DNA
- A number of unique developments
 - Data representation in base-3 rather than binary
 - DNA encoding avoids runs of the same nucleotide
 - 4 fold redundancy

Archiving Data – Converting Binary to DNA



Archiving DNA – Results and the Future

- DNA synthesised by Agilent Technologies
- All files were recovered
- DNA synthesis costs \$12K per MB
 - Same as 1MB in 1980
 - A human genome would cost \$38,400,000 to make
- Technique has been patented
- A viable 1000 year data archive
- Within a decade we believe this will be cost effective for 50 year archives



Agilent Technologies

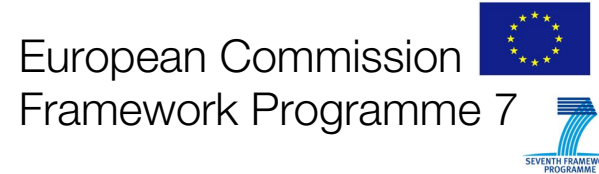
Conclusions

- Bioinformatics continues to push our capacity and ability to store, process and display data
- Science requires a global perspective
- Cloud infrastructures present unique opportunities
 - Augment our data processing
 - Bringing consumers & data together
- Bioinformatics can and will continue to help computing

Acknowledgments

- Ensembl
 - Stephen Keenan, Stephen Trevanian
- EMBL-EBI
 - Nick Goldman, Ewan Birney, Andy Cafferkey, Paul Flicek
- EMBL
 - Rupert Lueck
- Agilent Technologies
 - Siyuen Chen, Emily LeProust
- All staff at EMBL-EBI and EMBL

- Funding



Questions?