How feasible is DNA as a long-term digital storage system?

Nick Goldman

EMBL-European Bioinformatics Institute

EMBL-EBI

Cambridge, UK

@EBIgoldman & @Nick_Goldman

EMBL-EBI

155 petabytes

of storage capacity in our data centres

~38 million

requests to EMBL-EBI websites every day

EMBL-EBI delivered

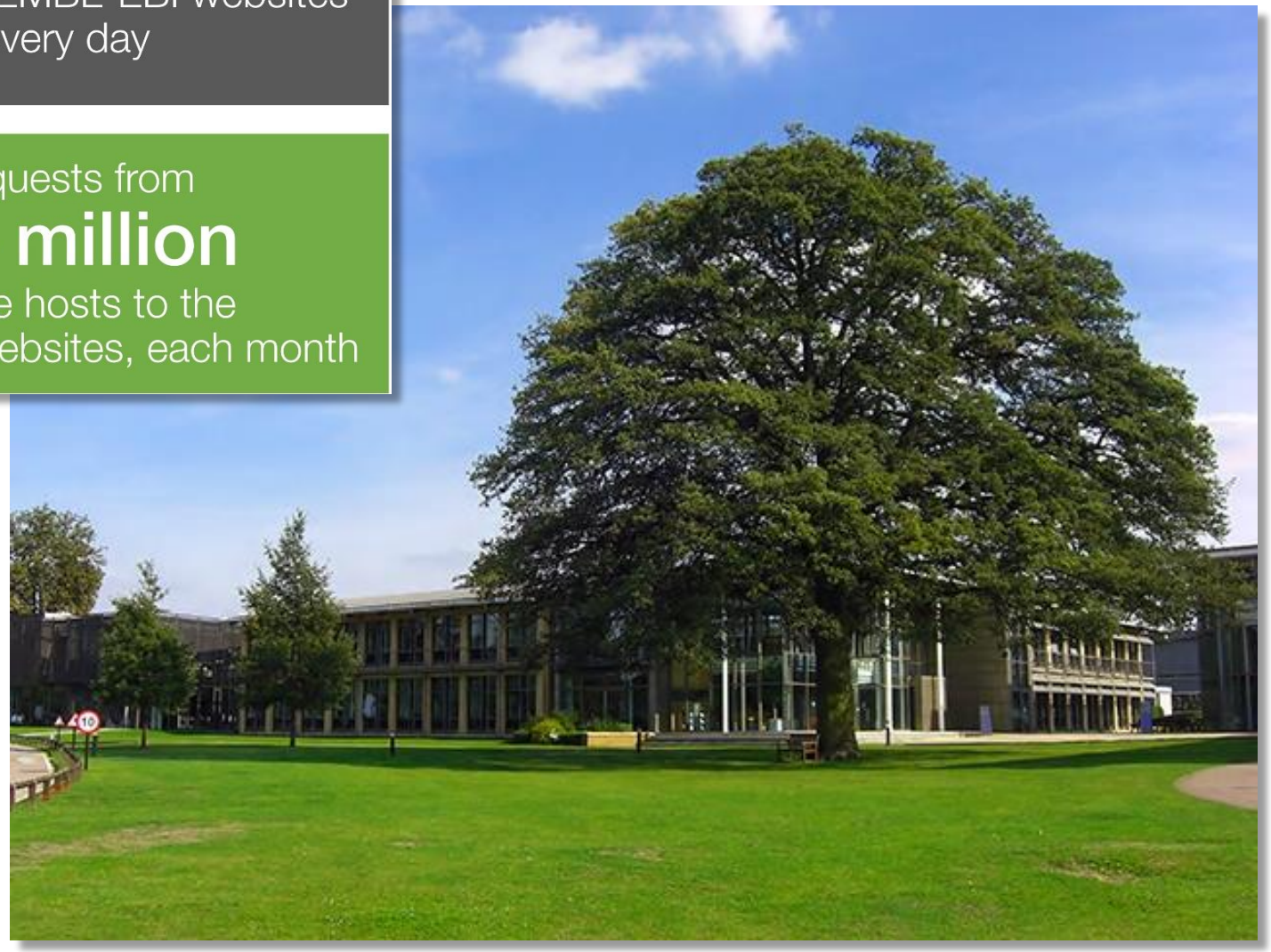
140 million

jobs to its users in 2017

Requests from

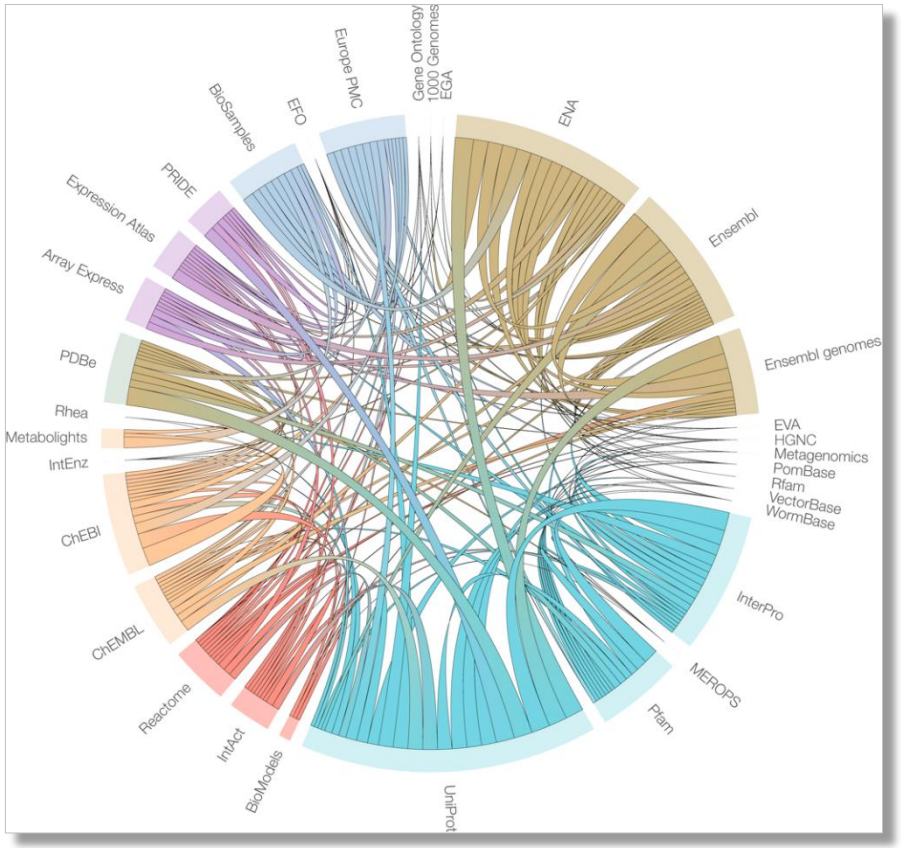
3.3 million

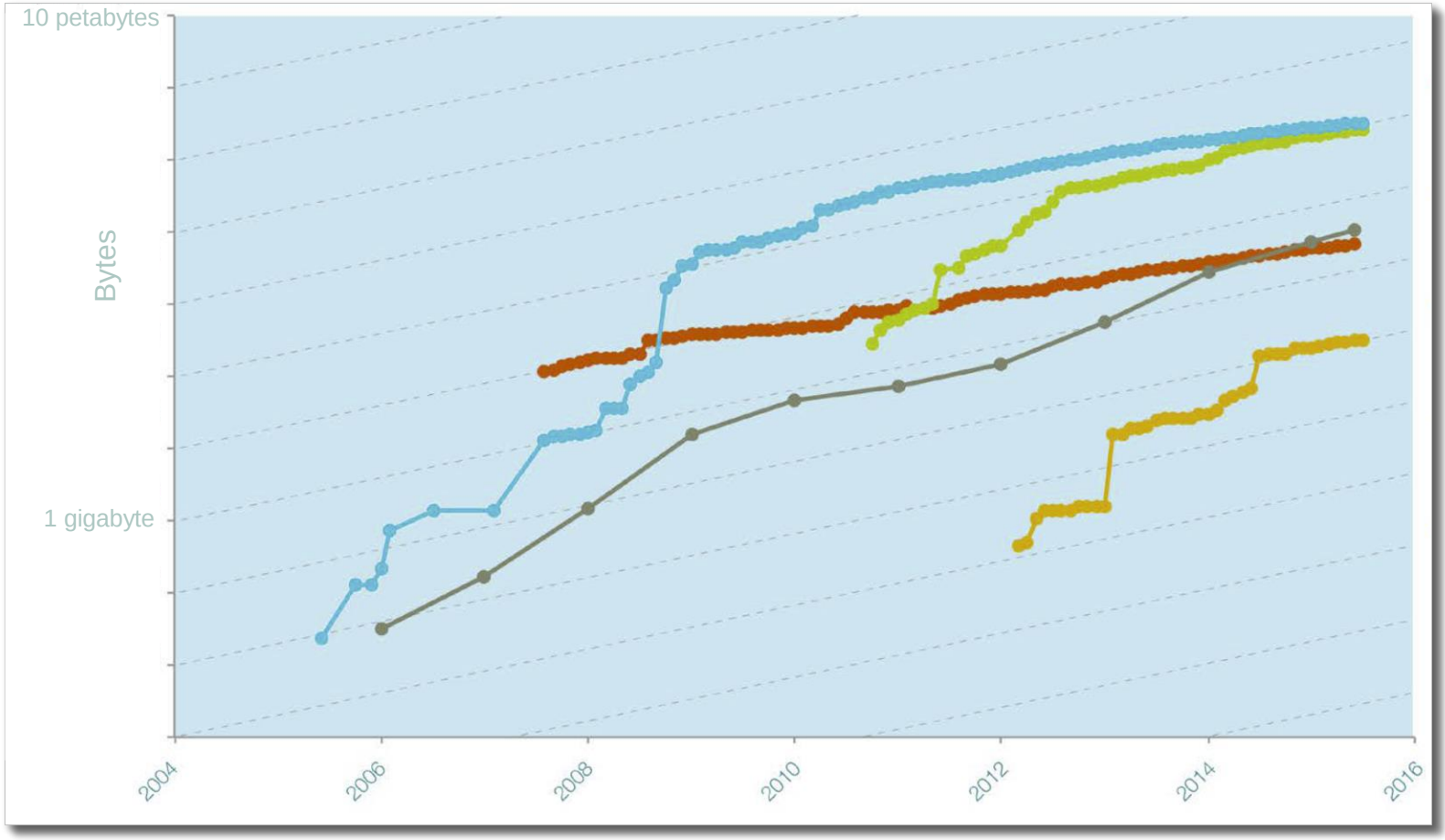
unique hosts to the EMBL-EBI websites, each month



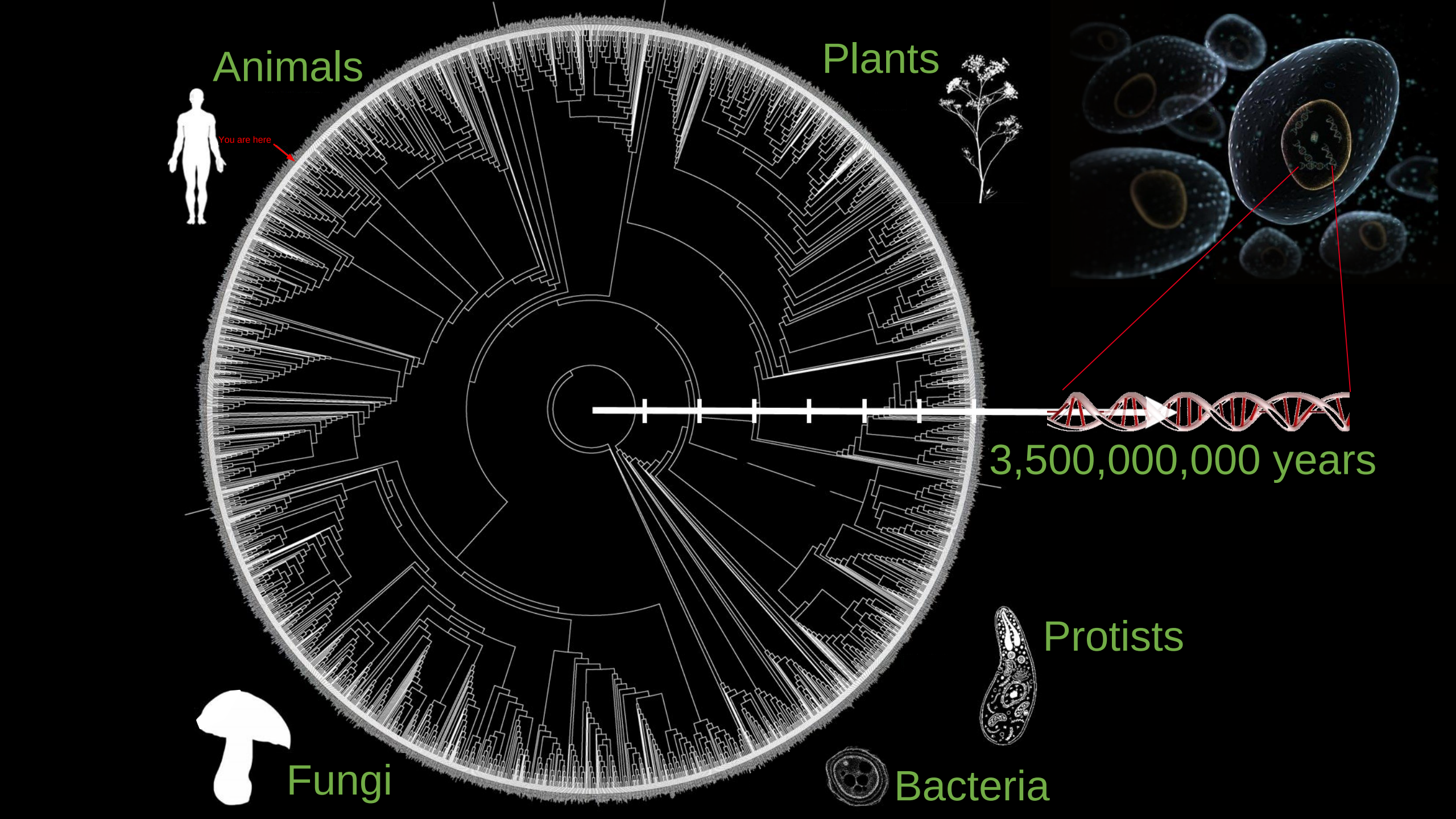
EBI is a major data provider to the life sciences research community

Databases at EBI



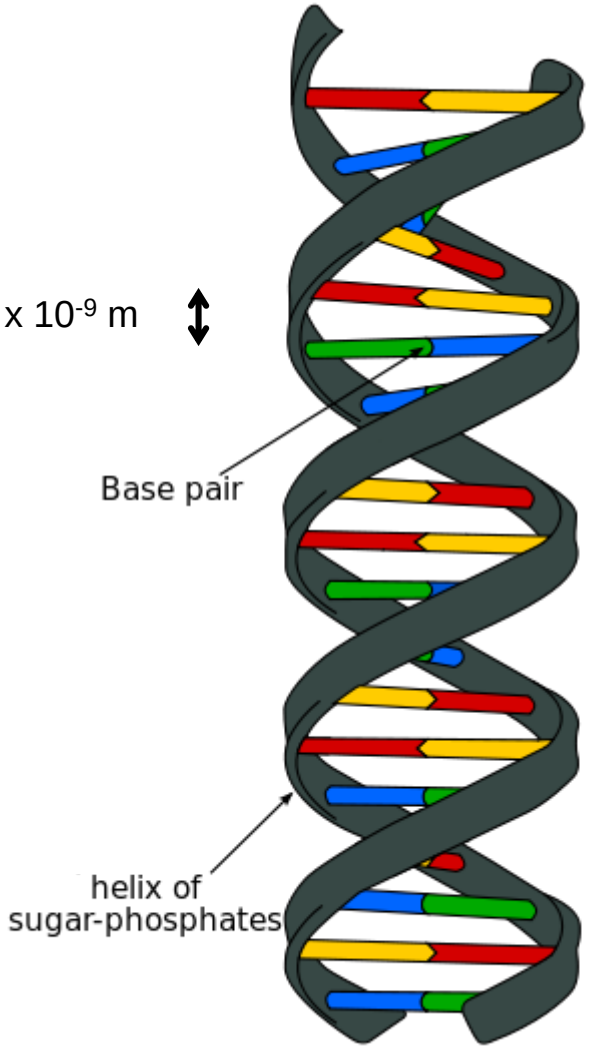


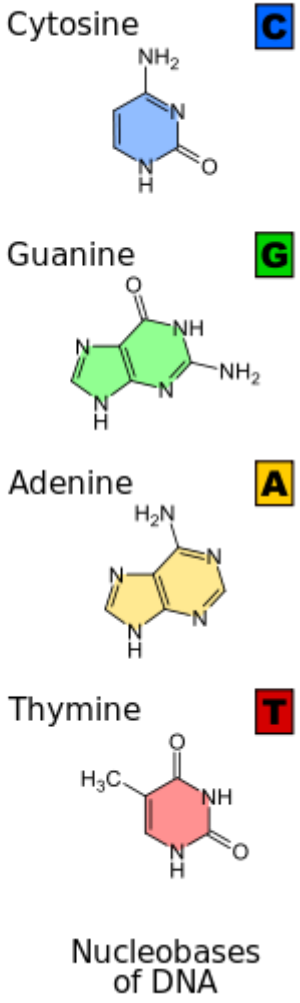

GASTWERK HOTEL/
HAMBURG, GERMANY_


Animals
You are here
Plants
3,500,000,000 years
Protists
Bacteria
Fungi

DNA

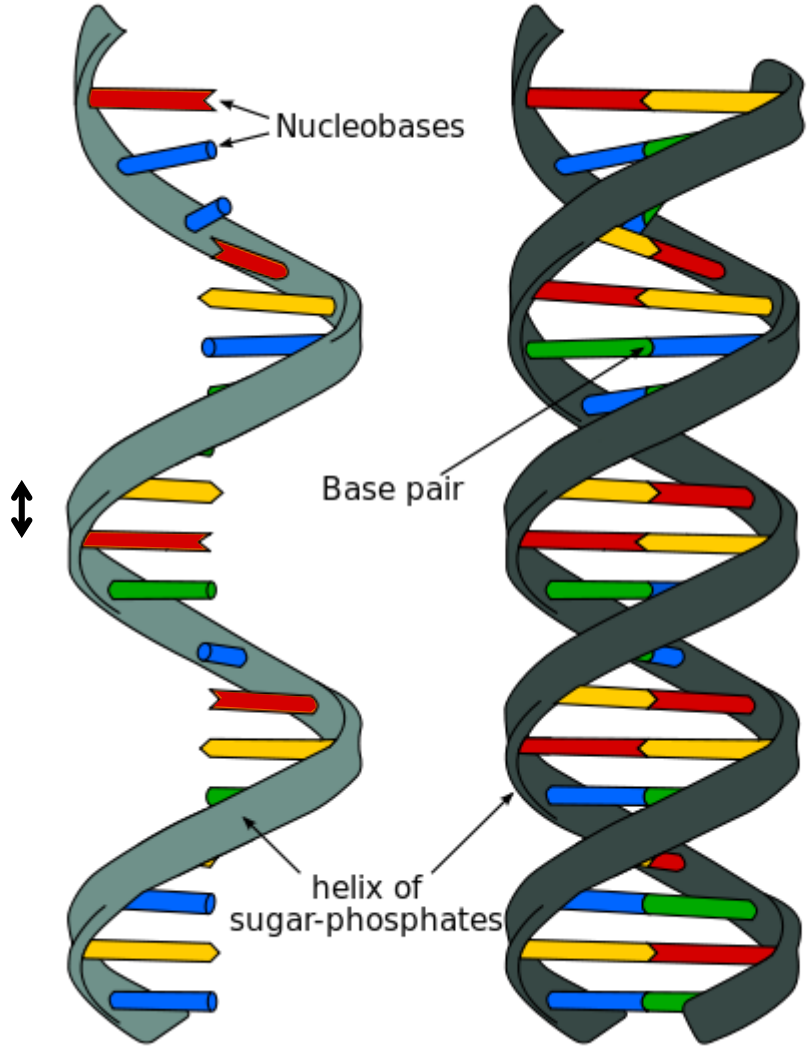
4 Å = 0.4 nm = 0.4 x 10^{-9} m

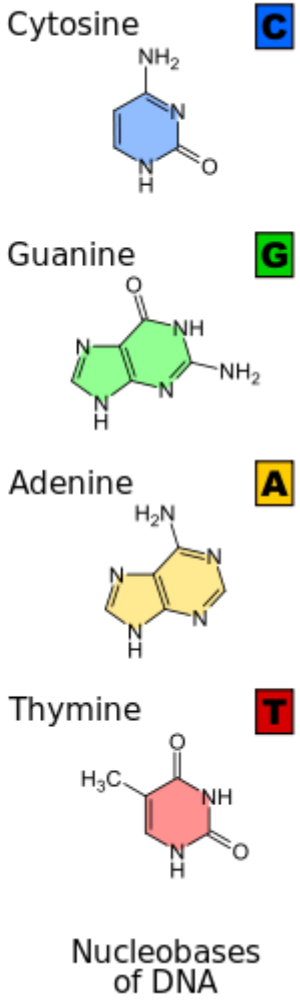




DNA

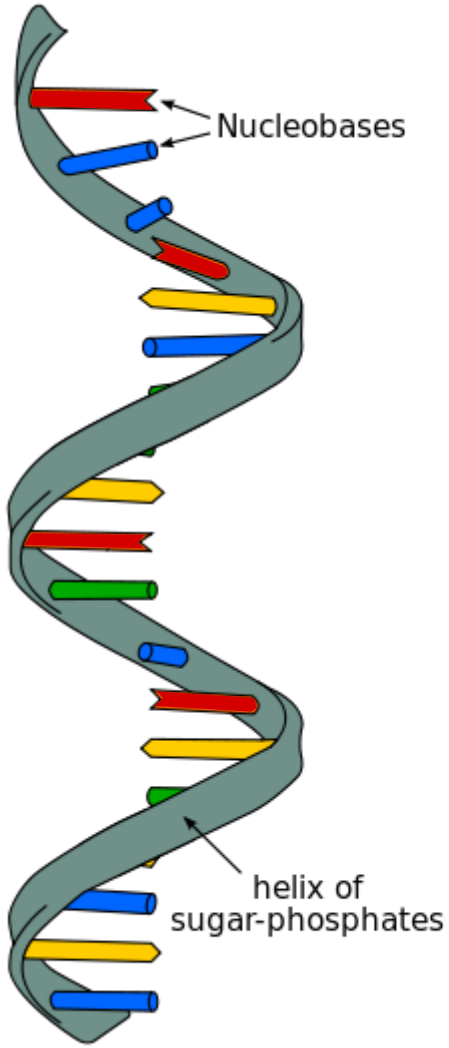
4 Å = 0.4 nm = 0.4 x 10^{-9} m

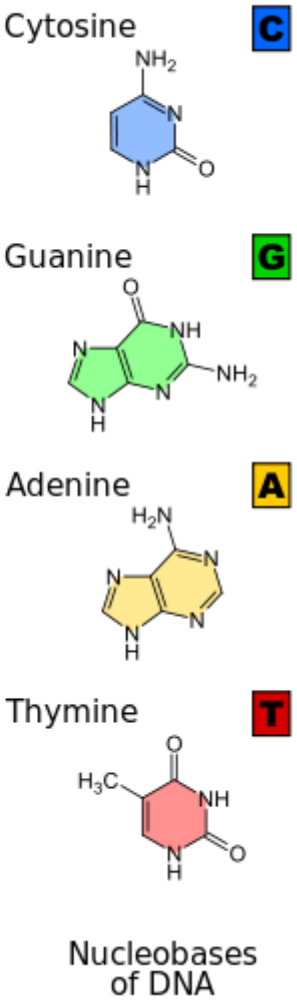




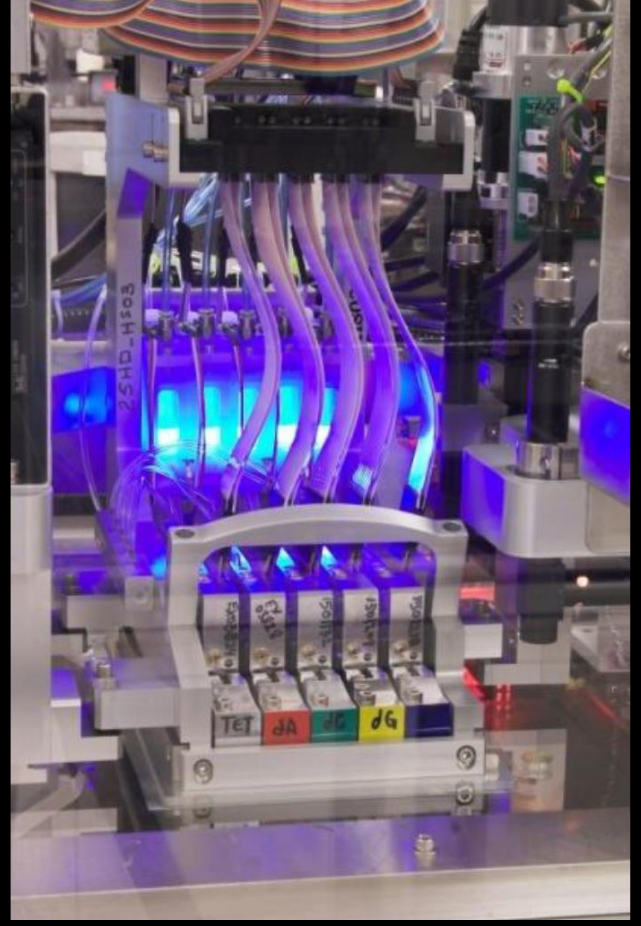
DNA

TCCTACGGATGCTAGACAC

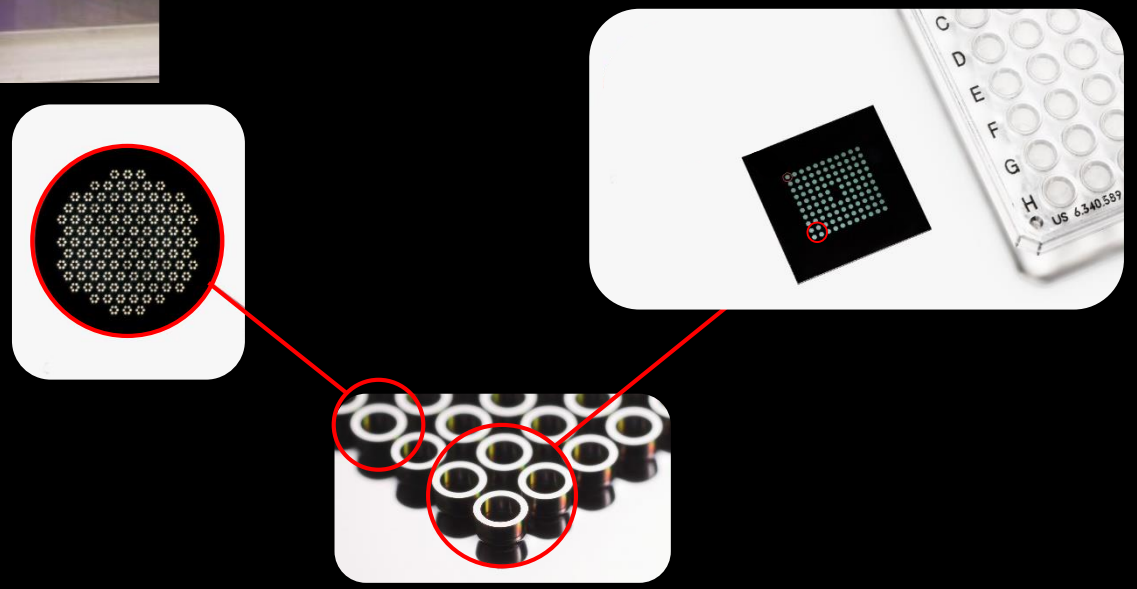




Nucleobases of DNA



High-throughput DNA sequence writers (DNA synthesis)

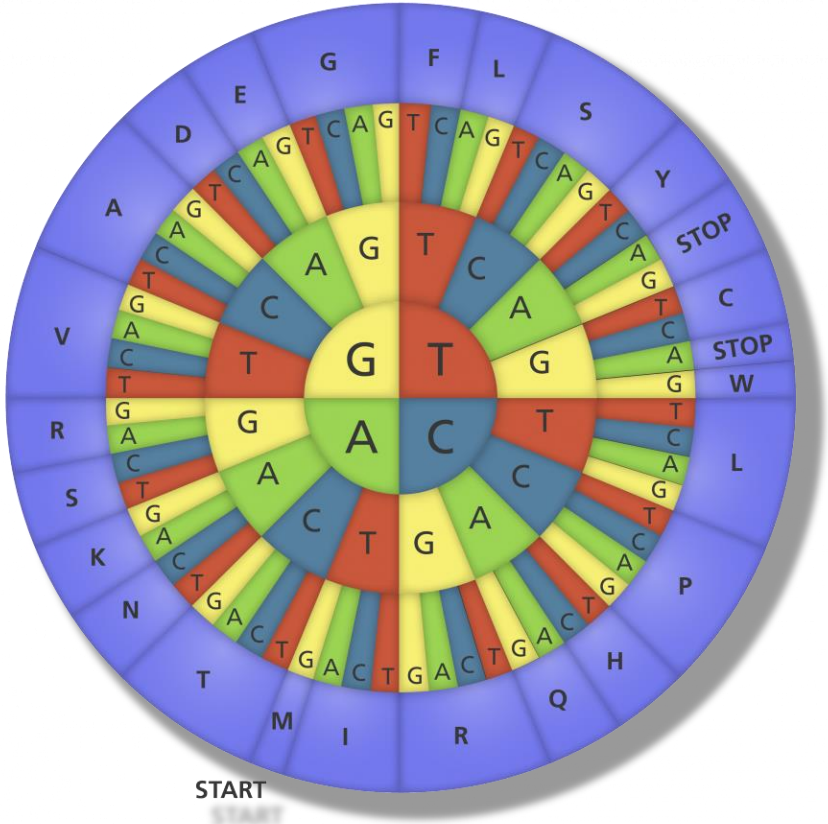


High-throughput DNA sequence readers (DNA sequencing)

	T			C			A			G		
T	TTT	Phe	F	TCT	Ser	S	TAT	Tyr	Y	TGT	Cys	C
	TTC			TCC			TAC			TGC		
	TTA	Leu	L	TCA			TAA	STOP		TGA	STOP	
	TTG			TCG			TAG			TGG	Trp	W
C	CTT	Leu	L	CCT	Pro	P	CAT	His	H	CGT	Arg	R
	CTC			CCC			CAC			CGC		
	CTA			CCA			CAA	Gln	Q	CGA		
	CTG			CCG			CAG			CGG		
A	ATT	Ile	I	ACT	Thr	T	AAT	Asn	N	AGT	Ser	S
	ATC			ACC			AAC			AGC		
	ATA			ACA			AAA	Lys	K	AGA	Arg	R
	ATG	Met	M	ACG			AAG			AGG		
G	GTT	Val	V	GCT	Ala	A	GAT	Asp	D	GGT	Gly	G
	GTC			GCC			GAC			GGC		
	GTA			GCA			GAA	Glu	E	GGA		
	GTG			GCG			GAG			GGG		

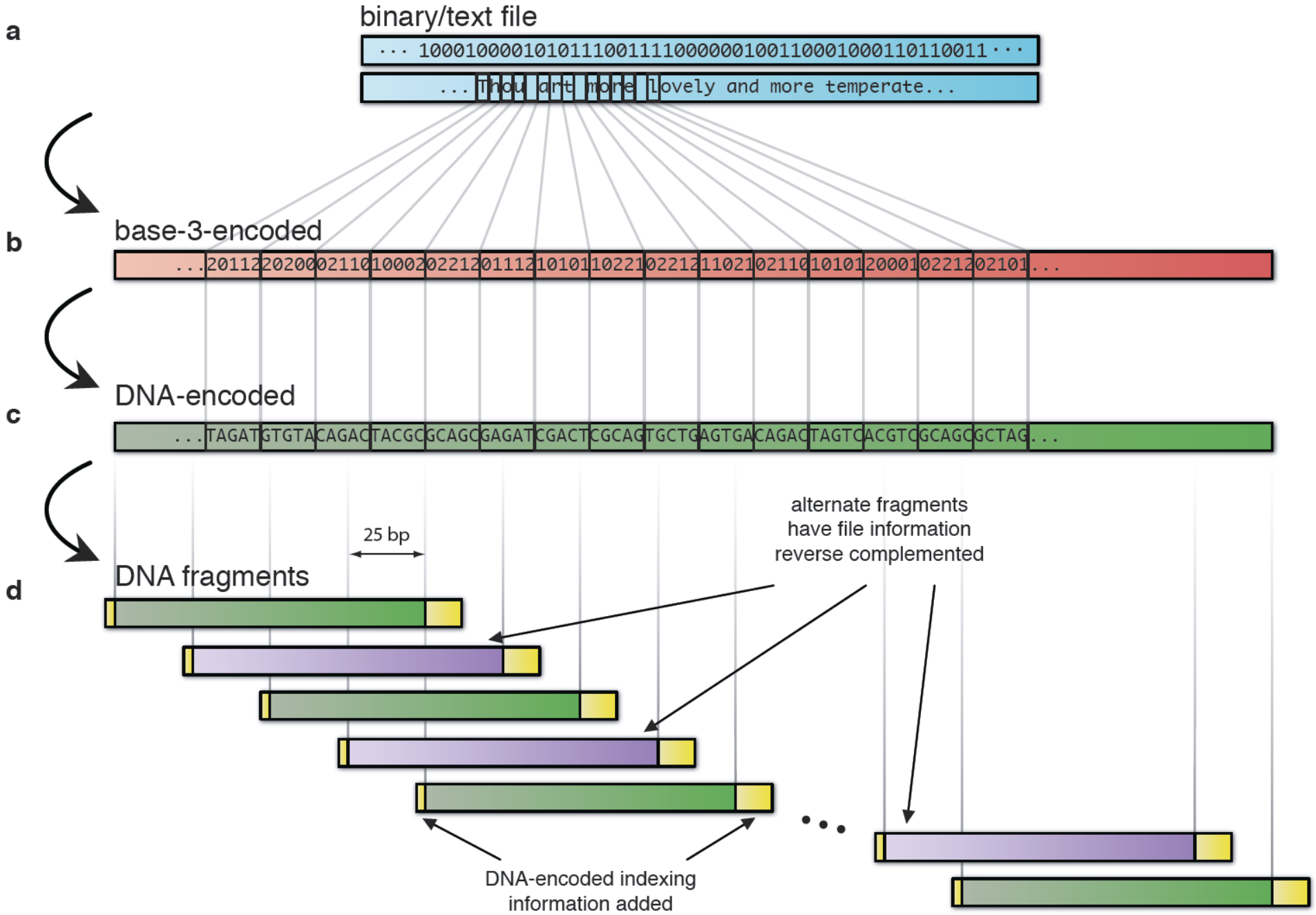
the Genetic code of life…

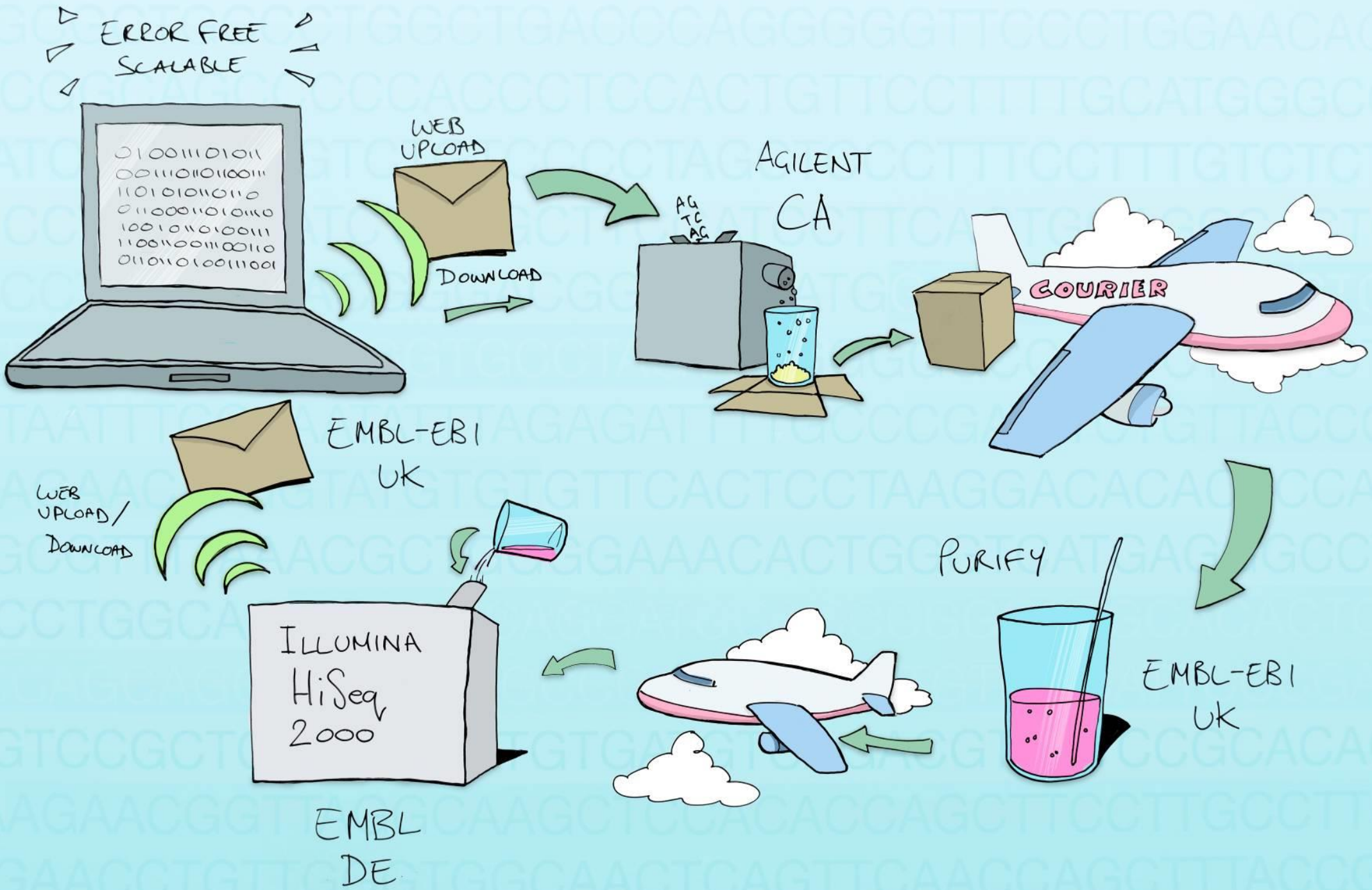
…codes for proteins,
but not mp3s or pdfs

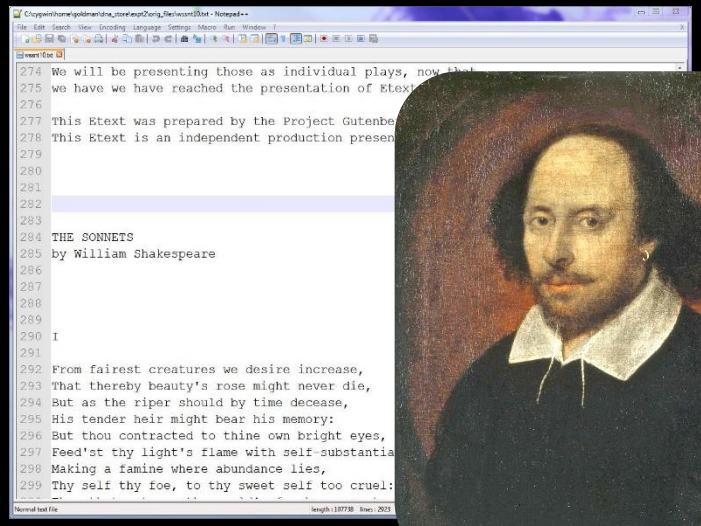


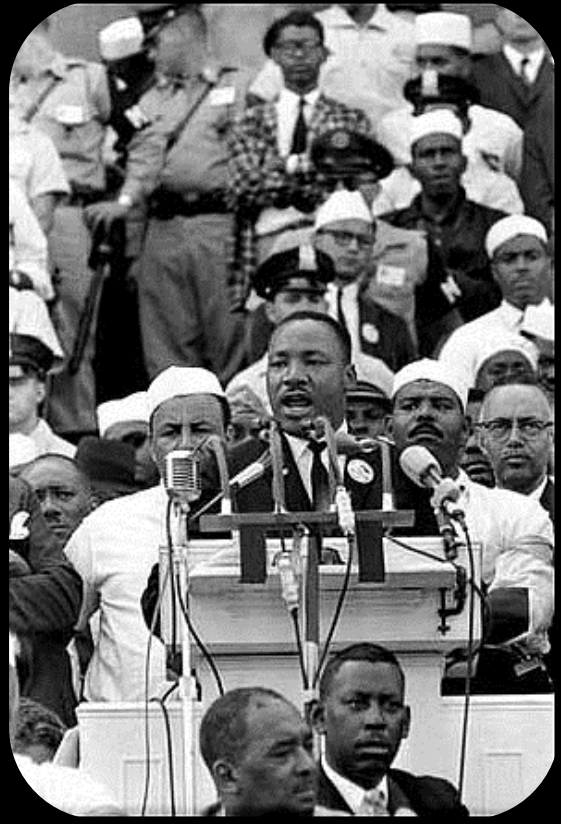
Amino acid code

A – Alanine	G – Glycine	M – Methionine	S – Serine
C – Cysteine	H – Histidine	N – Asparagine	T – Threonine
D – Aspartic acid	I – Isoleucine	P – Proline	V – Valine
E – Glutamic acid	K – Lysine	Q – Glutamine	W – Tryptophan
F – Phenylalanine	L – Leucine	R – Arginine	Y – Tyrosine


a
binary/text file
··· 10001000010101110011111000000100110001000110110011 ···
... Thou art more lovely and more temperate...
b
base-3-encoded
...201122020002110100020221201112101011022102212110210211010120001022120210 1...
c
DNA-encoded
...TAGATGTGTACAGACTACGCGCAGCGAGATCGACTCGCAGTGCTGAGTGACAGACTAGTCACGTCGCAGCGCTAG...
d
25 bp
DNA fragments
alternate fragments
have file information
reverse complemented
DNA-encoded indexing
information added


ERROR FREE
SCALABLE
WEB
UPLOAD
DOWNLOAD
AGILENT
CA
AG
TC
AG
COURIER
EMBL-EBI
UK
WEB
UPLOAD/
DOWNLOAD
PURIFY
ILLUMINA
HiSeq
2000
EMBL
DE
EMBL-EBI
UK





equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1950).
[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate–sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

1 MB of computer files = almost invisible dust of DNA


Genome Analyzer
illumina

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three inter-twined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate di-ester groups joining β-D-deoxy-ribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There

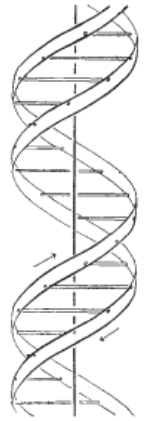
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three inter-twined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.



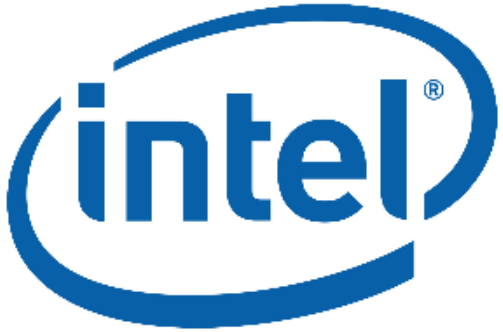
This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate di-ester groups joining β-D-deoxy-ribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The

What has happened since?

“NAM” (Nucleic Acid Memory) research established

Microsoft



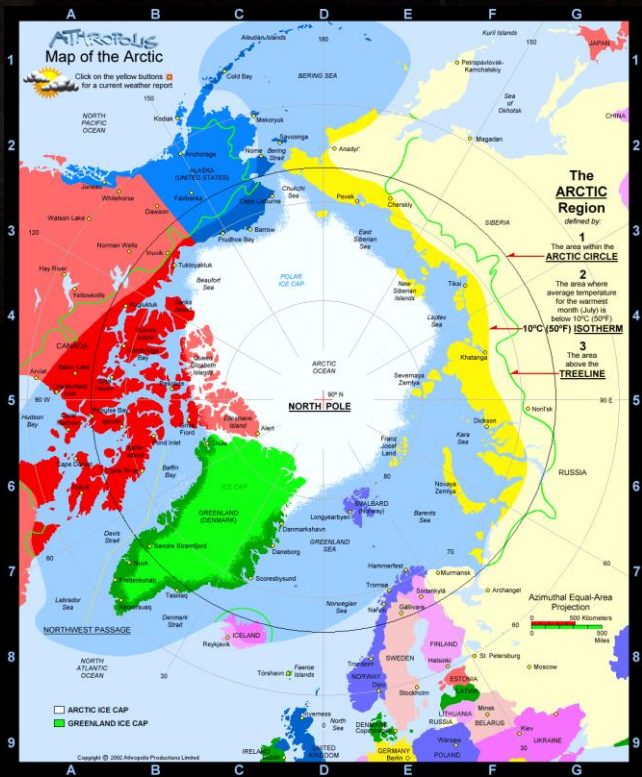
Charlotte Jarvis,
Music of the Spheres

www.artforeating.co.uk


Bitcoin Challenge
Nick Goldman
http://tinyurl.com/davosdna
http://www.ebi.ac.uk/research/goldman/dna-storage


Bitcoin Challenge Davos 2015
Sequence the DNA, decode the message, claim the prize of 1 Bitcoin
1 BITCOIN


PASTEUR
Universiteit Antwerpen

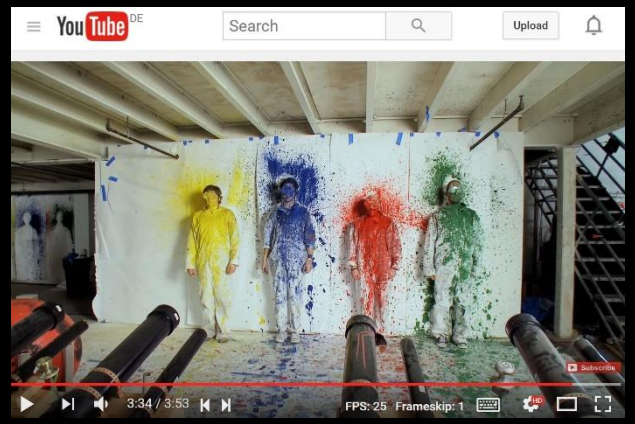

Map of the Arctic
Click on the yellow buttons for a current weather report
The ARCTIC Region
defined by:
1
The area within the ARCTIC CIRCLE
2
The area where average temperature for the warmest month (July) is below 10°C (50°F)
10°C (50°F) ISOTHERM
3
The area above the TREELINE
NORTH POLE
NORTHWEST PASSAGE
Azimuthal Equal-Area Projection
ARCTIC ICE CAP
GREENLAND ICE CAP

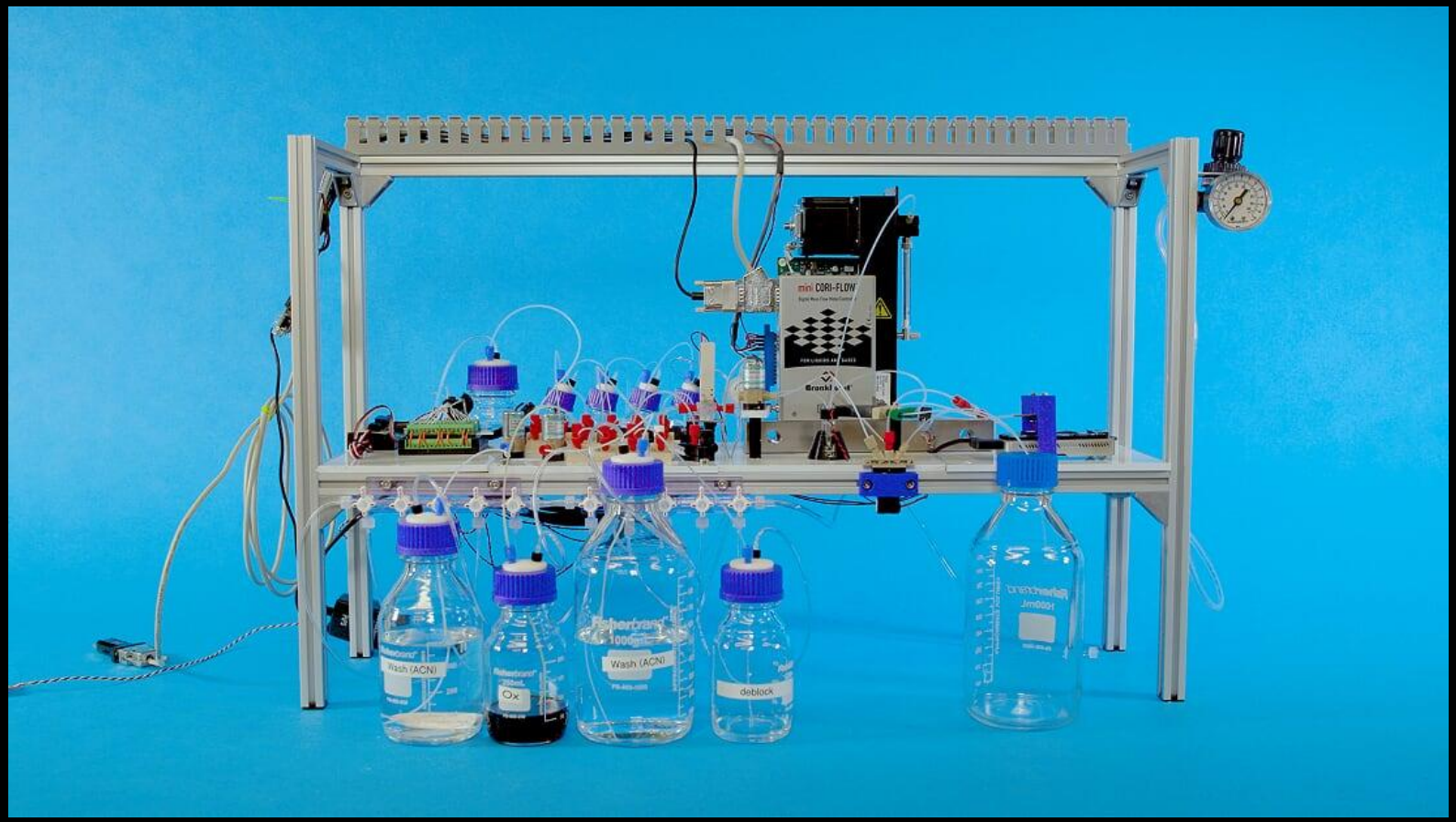

unicef
for hvert b
BARNERETTIGHETENE
FNs KONVENSJON OM BARNETS RETTIGHETER
ALLE BARN ER FØDT FRIE MED LIKE RETTIGHETER

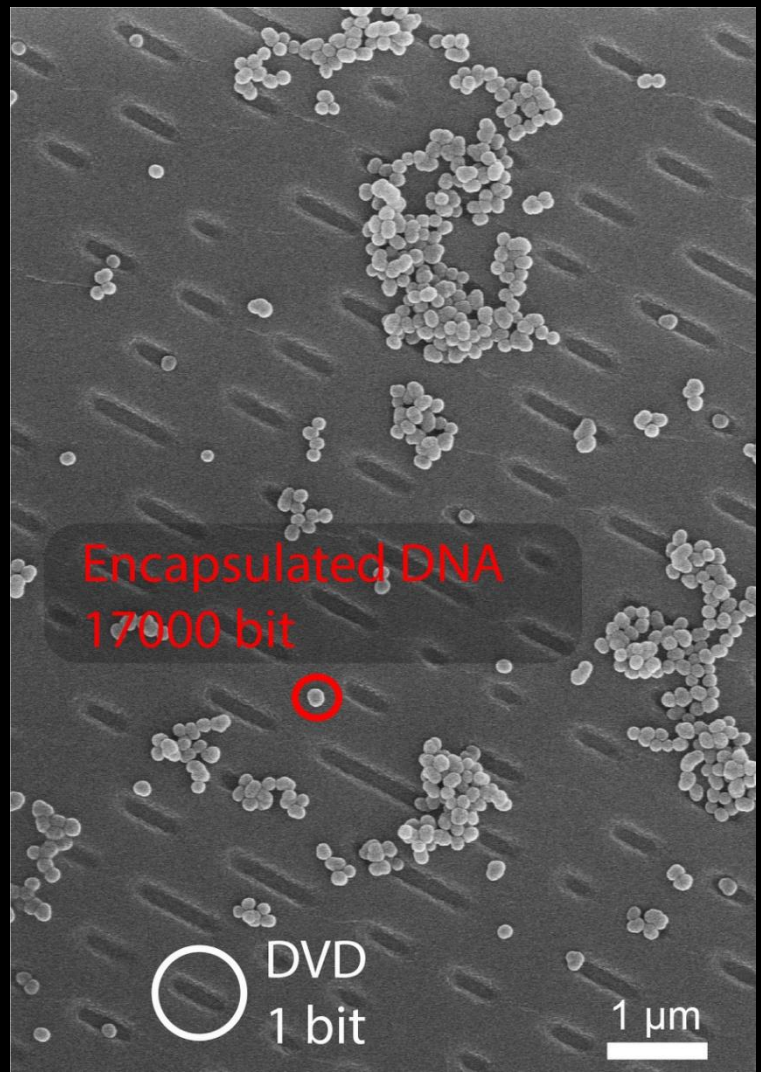

Gjelder hele Svalbard

Technicolor and Harvard University

University of Washington and Microsoft Research







Picture credit:
Robert Grass, Julian Koch
ETH Zürich

6 areas where progress is being made:

- phosphoramidite synthesis continuing to improve (accuracy, speed, cost);
  enzymatic synthesis companies appearing regularly (various companies)
- new data encoding techniques reduce the amount of DNA needed (multiple teams)
- DNA random access memory (RAM — University of Illinois)
- first ‘table-top’ end-to-end device prototype exists (University of Washington/Microsoft)
- can data encoding permit ‘computing in DNA’? (image search — University of Washington/Microsoft)
- integration of information in DNA into materials (ETH Zürich)

Why use DNA for storing information?

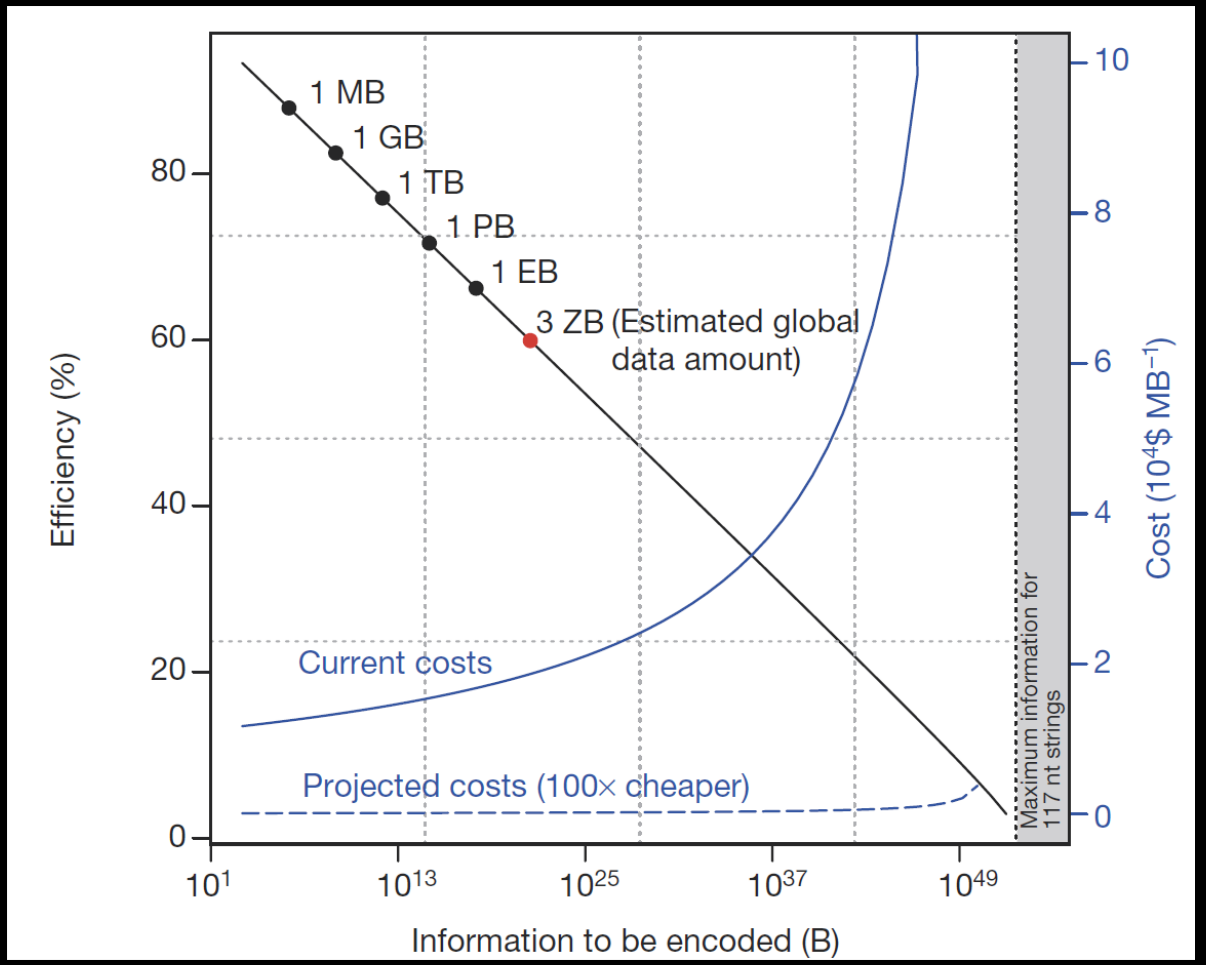
Why use DNA for storing information?

First: why not to use DNA for storing

Why use DNA for storing information?

1 test-tube full of DNA = 1 000 000 CD ROMs

x 10 000



Encoding technique scales up beyond global data volumes

Archives made of DNA would take up very little space

DNA lasts for a very long time

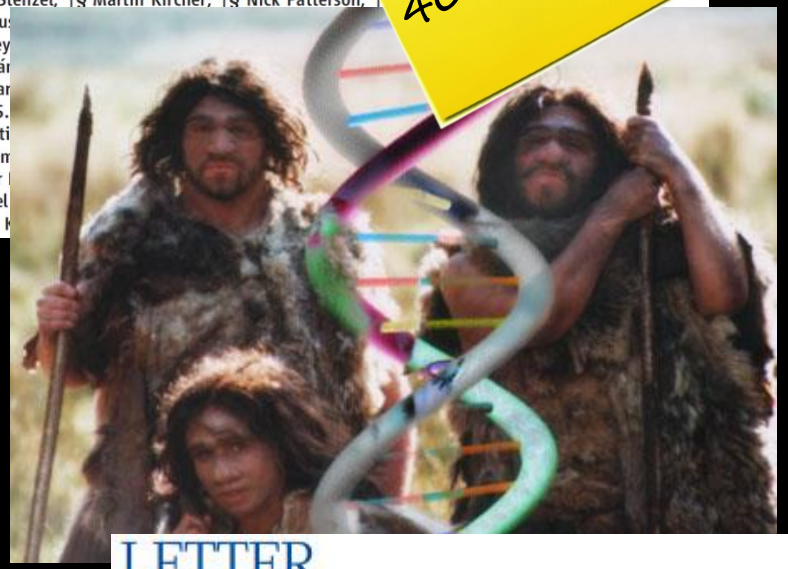
710 7 MAY 2010 VOL 328 SCIENCE www.sciencemag.org

RESEARCH ARTICLE

A Draft Sequence of the Neandertal Genome

Richard E. Green,[1]*†‡ Johannes Krause,[1]†§ Adrian
Udo Stenzel,[1]†§ Martin Kircher,[1]†§ Nick Patterson,[2]
Markus
Jeffrey
Hernán
Barbar
Eric S.
Christi
Vladim
Javier
Daniel
Janet

40 000 years old



Vol 456|20 November 2008|doi:10.1038/nature07446 nature

LETTERS

Sequencing the nuclear genome of the extinct woolly mammoth

Webb Miller[1], Daniela I. Drautz[1], Aakrosh Ratan[1], Barbara Pusey[1], Ji Qi[1], Arthur M. Lesk[1], Lynn P. Tomsho[1], Michael D. Packard[1], Fangqing Zhao[1], Andrei Sher[2]†, Alexei Tikhonov[3], Brian Raney[4], Nick Patterson[5]
Lindblad-Toh[5], Eric S. Lander[5], Ja
Tom Pringle[8] & Steph

20 000 years old

www.sciencemag.org SCIENCE VOL 306 26 NOVEMBER 2004 1561

Rise and Fall of the Beringian Steppe Bison

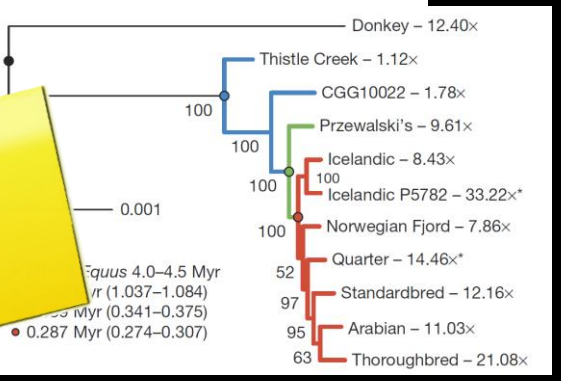
Beth Shapiro,[1,2] Alexei J. Drummond,[2] Andrew Rambaut,[2]
Michael C. Wilson,[3] Paul E. Matheus,[4] Andrei V. Sher,[5]
Oliver G. Pybus,[2] M. Thomas P. Gilbert,[1,2] Ian Barnes,[6]
Jonas Binladen,[7] Eske Willerslev,[1,7] Anders J. Hansen,[7]
Gennady F. Baryshnikov,[8] James A. Burns,[9] Sergei Davydov,[10]
Jonathan C. Driver,[11] Duane G. Fr
Grant Keddie,[14] Pavel Kosint
Larry D. Martin,[17] Robert O. St
Richard Tedford,[20] Sergei Z

LETTER 74 | NATURE | VOL 499 | 4 JULY 2013

Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse

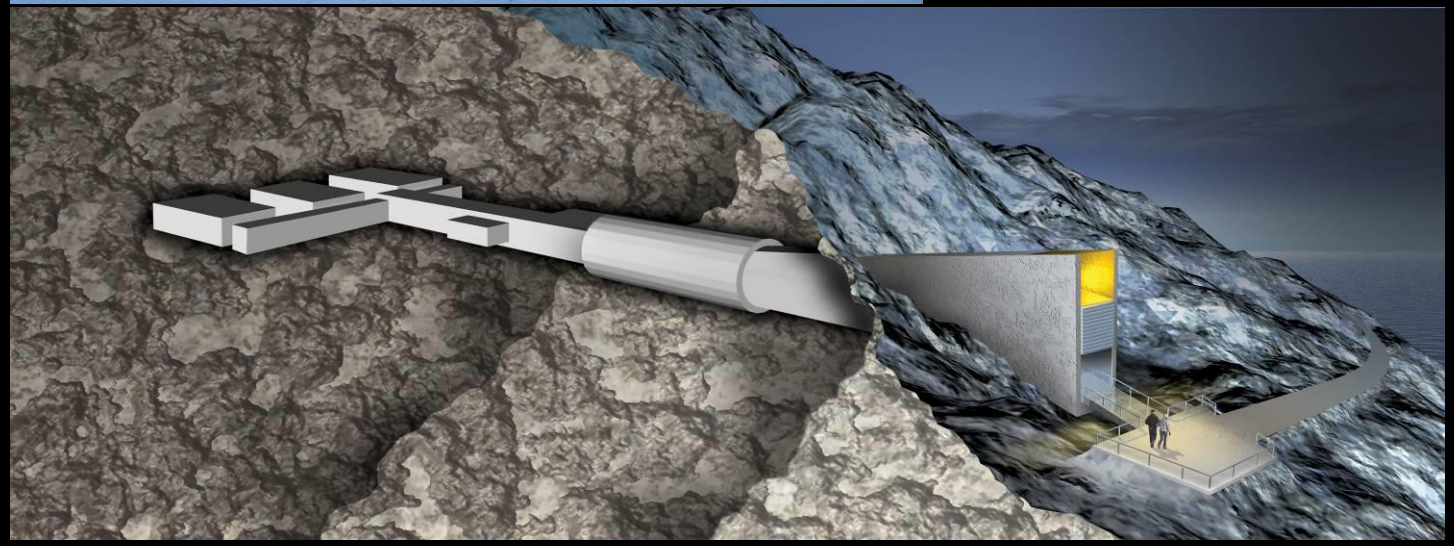
Ludovic Orlando[1]*, Aurélien Ginolhac[1]*, Guojie Zhang[2]*, Dua
Enrico Cappellini[3], Bent Petersen[3], Ida Moltke[4,7], Philip L. F.
Thorfinn Korneliussen[1], Anna-Sapfo Malaspinas[1], Josef Vogt[1]
Andrei Dolocan[12], Jesper Stenderup[1], Amhed M. V. Velazque
Grant D. Zazula[13], Andaine Seguin-Orlando[1,14], Cecilie Mort
Jacobo Weinstock[16], Kristian Gregersen[1,15], Knut H.
Douglas F. Antczak[18], Mads F. Bertelsen[22],
John Mundy[25], Anders Krogh[1,4],
Jesper V. Olsen[10],

560 000 – 780 000 years old



60 000 years old

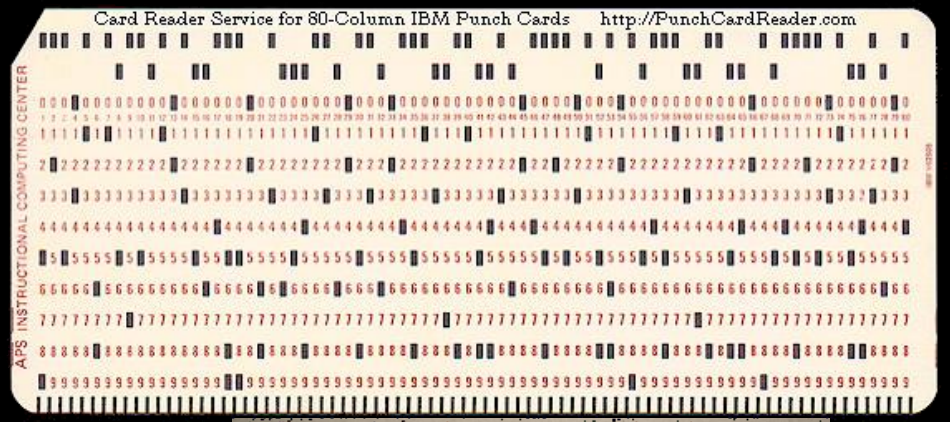
DNA lasts for a very long time in easy-to-create conditions



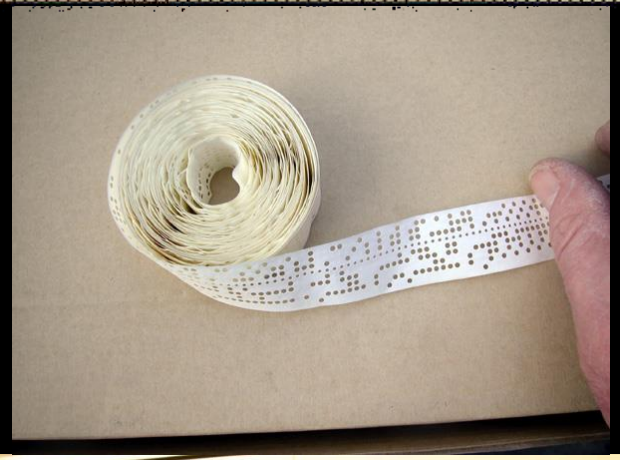
With DNA it is easy to make many copies of your data — exponentially!


boardem
FLOPPY DISK


BASF
Do not touch the video tape
Ne touchez pas la bande vidéo
BASF L 750
Betamax
The Future is Here


Card Reader Service for 80-Column IBM Punch Cards
http://PunchCardReader.com
APS INSTRUCTIONAL COMPUTING CENTER


DYSAN CE PACK 200 M/B

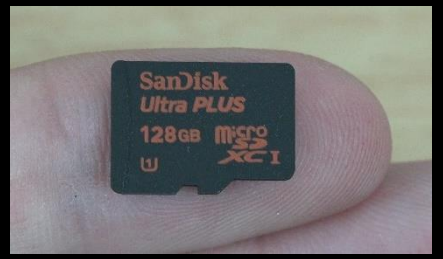


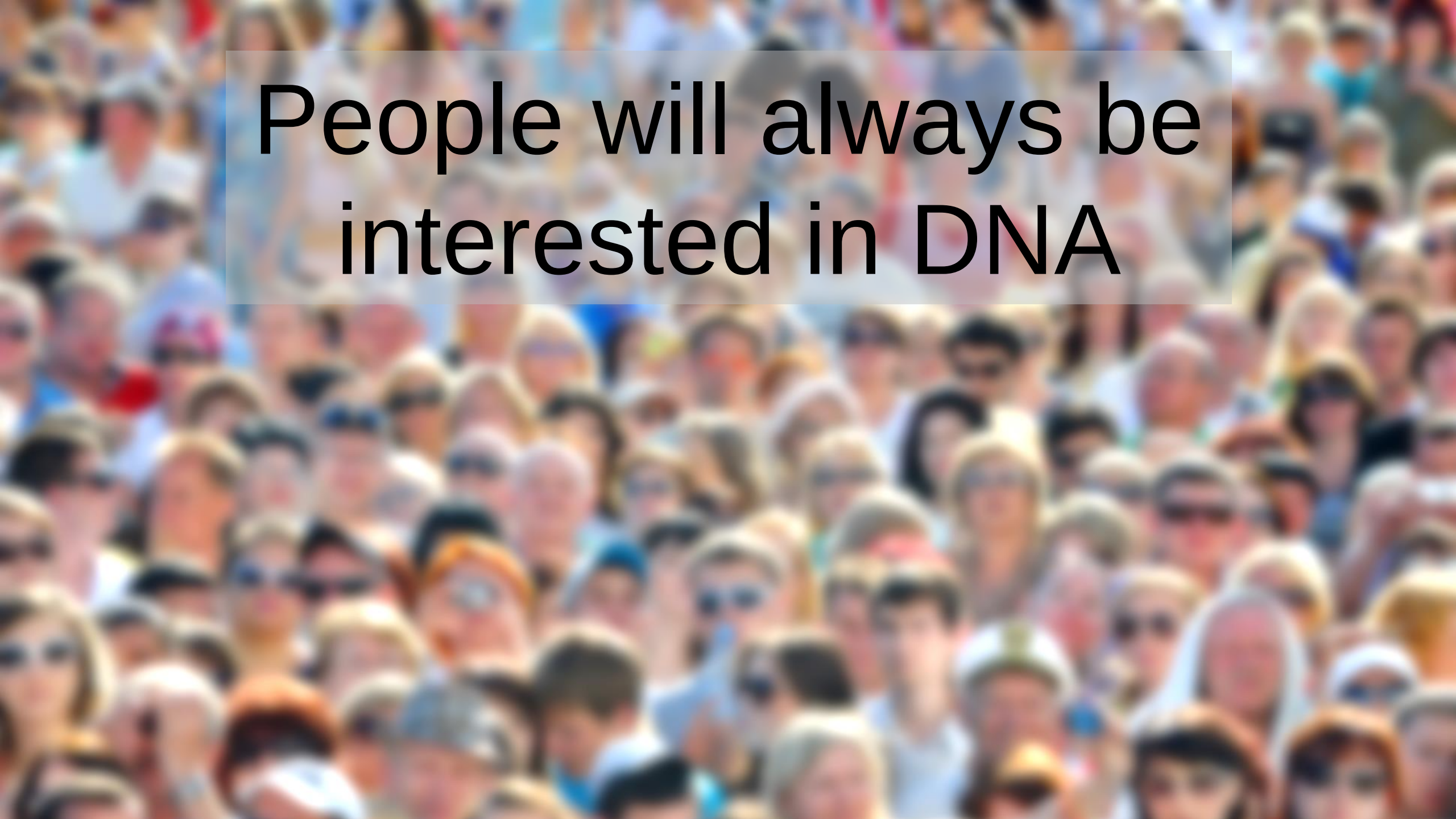

imation
CD-R
700MB - 80min


FOUR WEDDINGS AND A FUNERAL
dts HD
Blu-ray Disc


amazon cloud drive


maxell
TR-1
800 m
400 m
TDK
DC 2120
MINI DATA CARTRIDGE
QIC80 FORMATTED


SanDisk
Ultra PLUS
128GB
micro SD XC I


People will always be interested in DNA

Ewan Birney

Siyuan Chen

Paul Bertone*

Emily LeProust

Christophe Dessimoz*

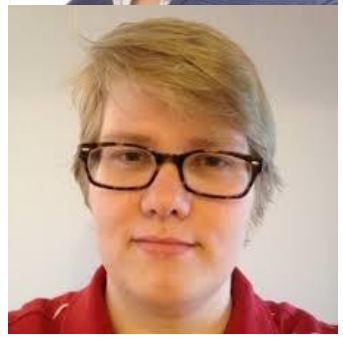
Agilent Technologies

TWIST BIOSCIENCE

Botond Sipos*

Jossy Sayir*

Emily Hesketh*



EMBL

wellcome trust sanger institute

and additional thanks to:

Cambridge: David MacKay†, Graeme Mitchison†

EBI: Kevin Gori*, Daniel Henk*, Remco Loos*, Ari Löytynoja*, Hazel Marsden*, Tim Massingham*, Sarah Parks*, Roland Schwarz*

EMBL-Heidelberg: Vladimir Benes, Dinko Pavlinić, Jonathon Blake*

EMBLEM: Birgit Kerber, Boris Bryk

Art for Eating: Charlotte Jarvis **Echidna:** Oliver Bradley

* = now moved to new positions