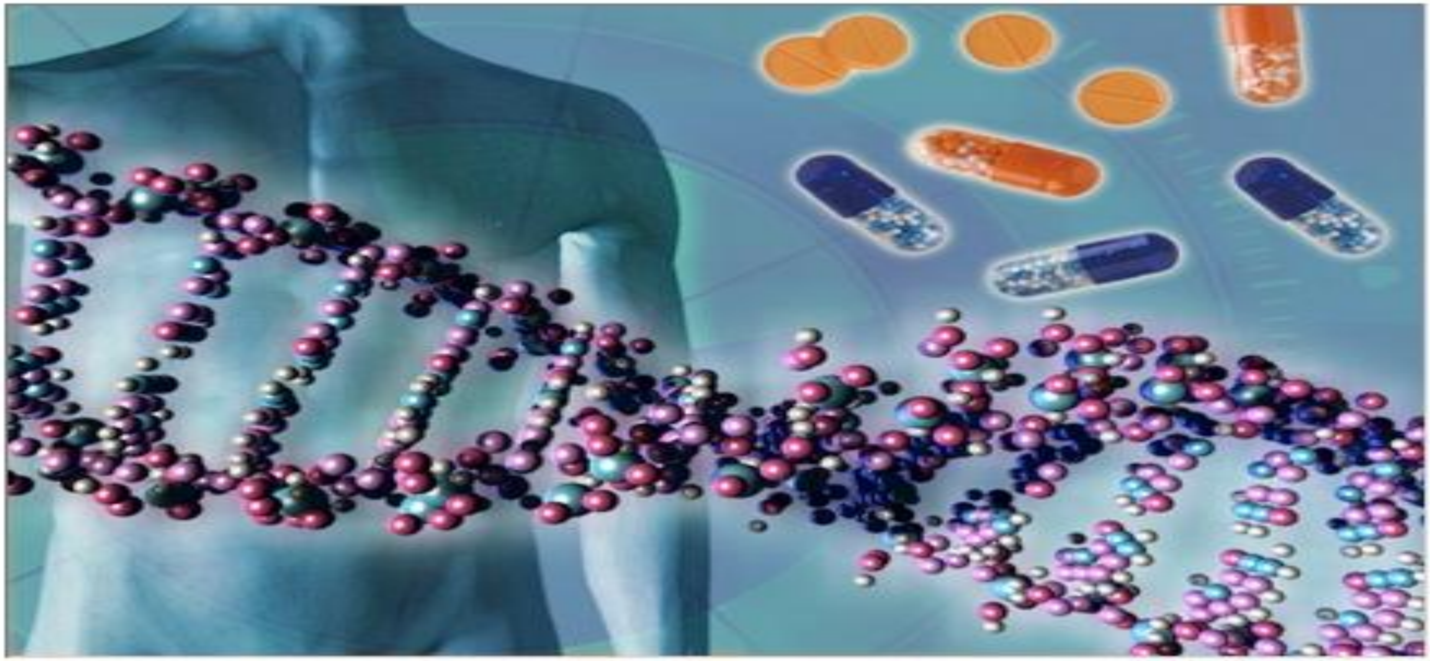
VirginiaTech
Invent the Future

Institute for
Critical Technology *and* Applied Science

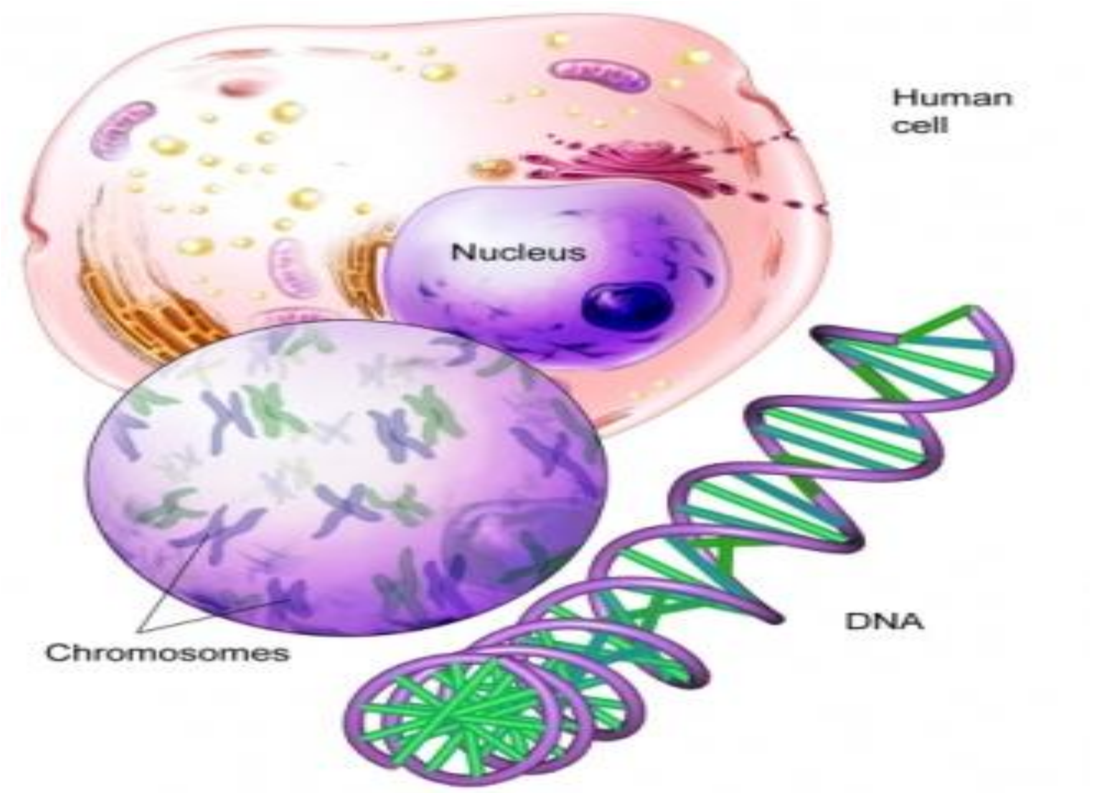
Genome Sequencing



Mohamed N. Seleem

VirginiaTech
Invent the Future

What is genetic material

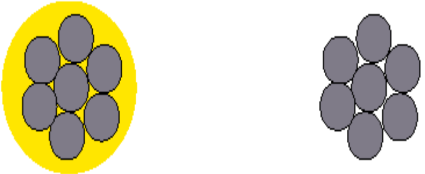


Frederick Griffith

transforming principle

1929

Streptococcus pneumoniae



Smooth colonies secrete a capsule and kill mice.

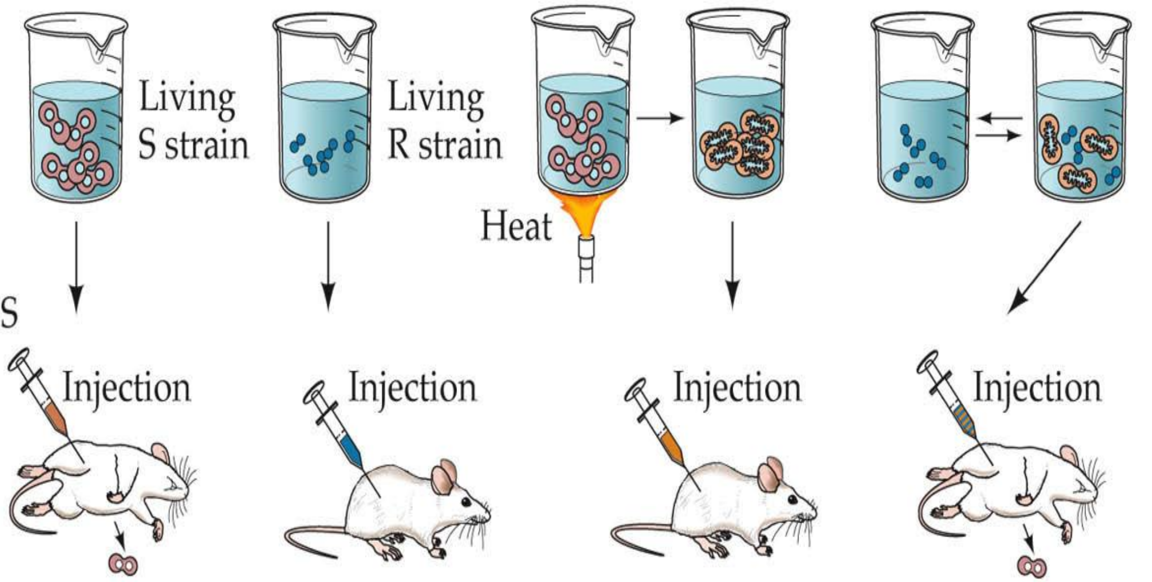
Rough colonies do not secrete a capsule and do not kill mice

EXPERIMENT

Question: Can the presence of dead bacterial cells genetically transform living bacterial cells?

METHOD

RESULTS



Living S strain

Living R strain

Heat

Injection

Injection

Injection

Injection

Mouse dies

Living S strain cells found in heart

Mouse healthy

No bacterial cells found in heart

Mouse healthy

No bacterial cells found in heart

Mouse dies

Living S strain cells found in heart

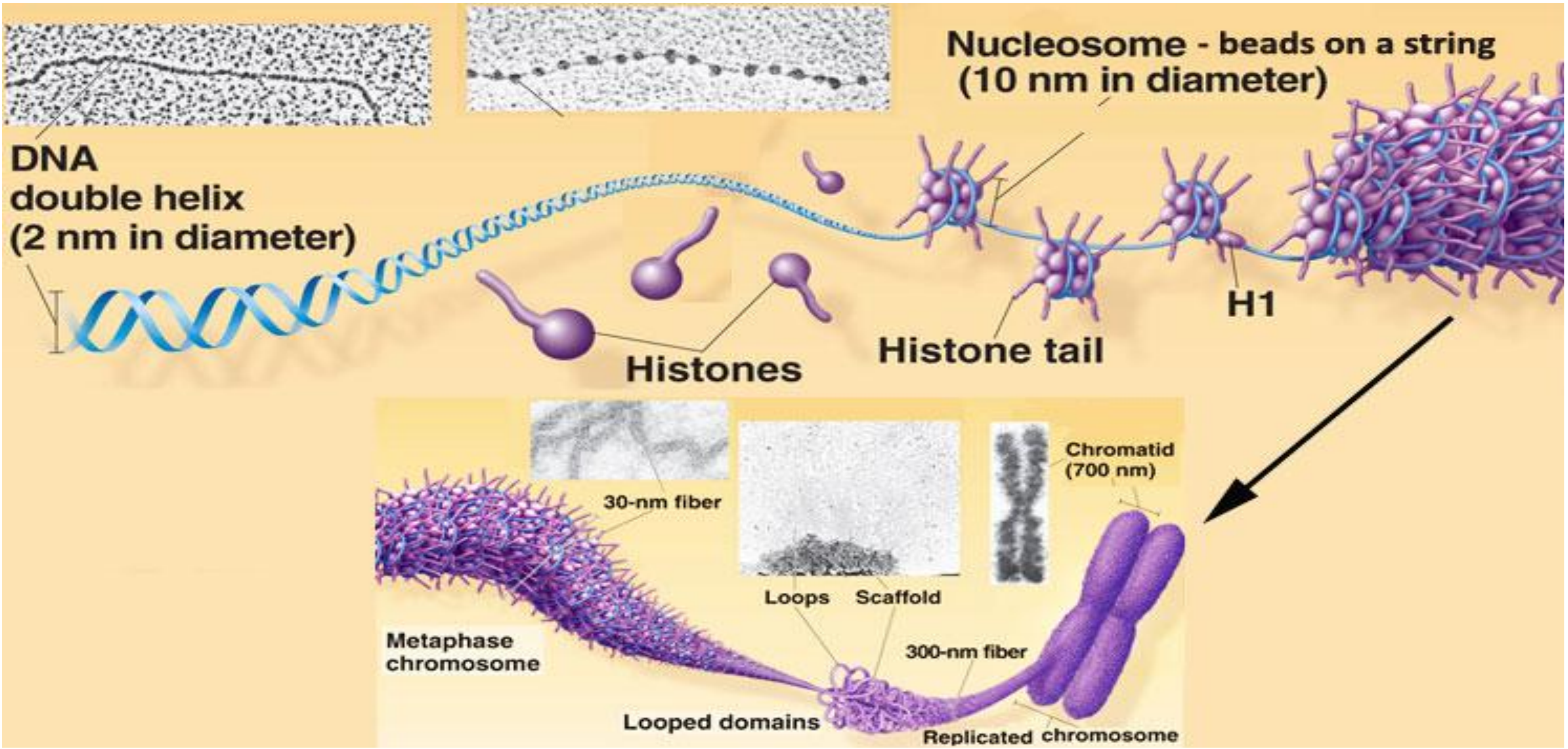
Conclusion: A chemical component from one cell is capable of genetically transforming another cell.

LIFE: THE SCIENCE OF BIOLOGY, *Seventh Edition*, Figure 11.1 Genetic Transformation of Nonvirulent Pneumococci

Genetic materials???

- Protein (chromosomes 90% protein)
- DNA
- Carbohydrate
- Lipids

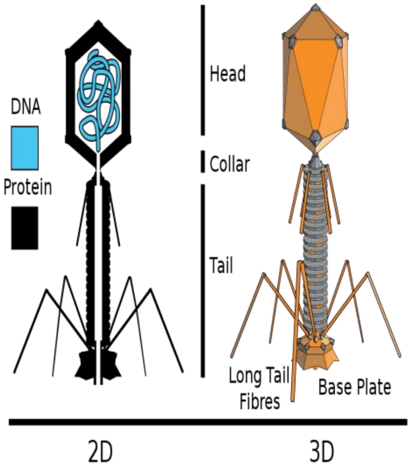
1944 Oswald Avery

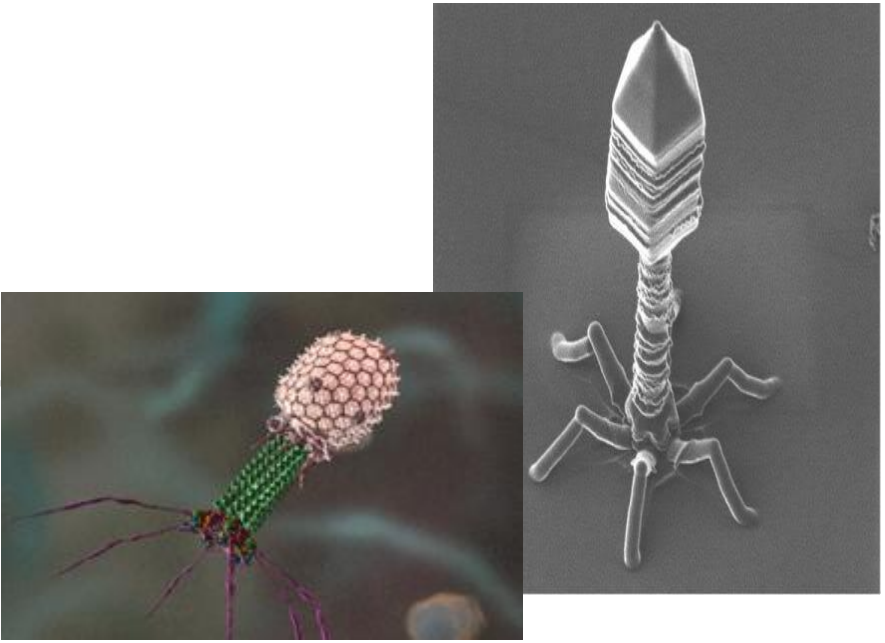


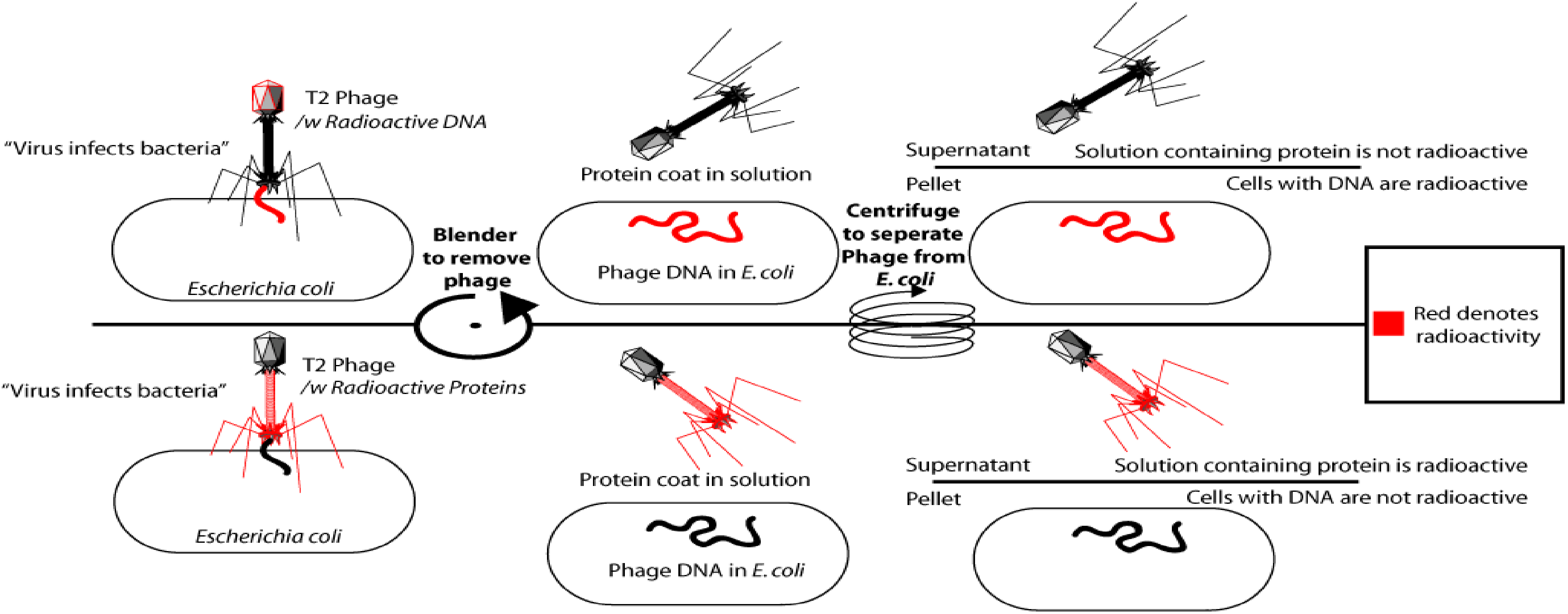
VirginiaTech
Invent the Future

Courtesy of Cold Spring Harbor Laboratory Archives. Noncommercial, educational use only.

Alfred Hershey and Martha Chase
1952







T2 Phage /w Radioactive DNA

"Virus infects bacteria"

Escherichia coli

Blender to remove phage

Protein coat in solution

Phage DNA in *E. coli*

Centrifuge to seperate Phage from *E. coli*

Supernatant — Solution containing protein is not radioactive

Pellet — Cells with DNA are radioactive

T2 Phage /w Radioactive Proteins

"Virus infects bacteria"

Escherichia coli

Protein coat in solution

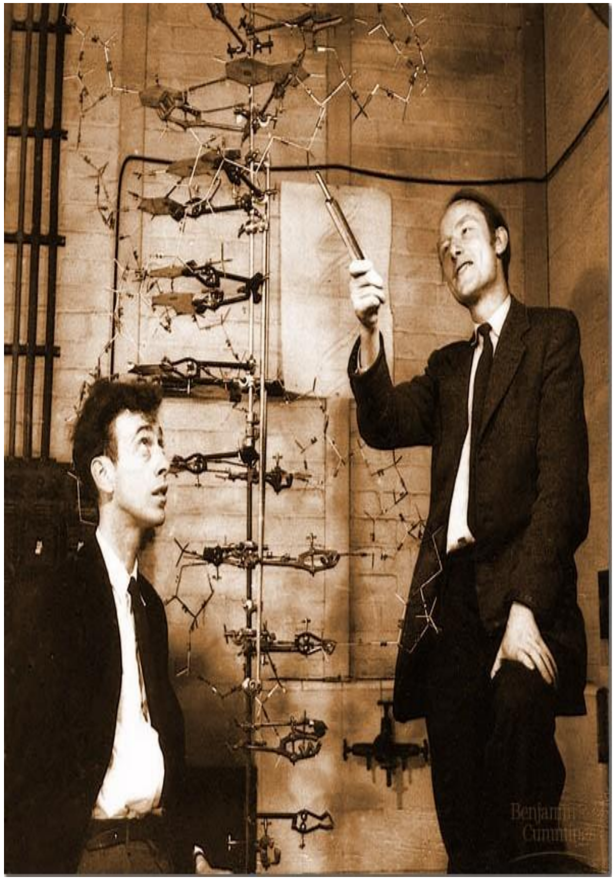
Phage DNA in *E. coli*

Supernatant — Solution containing protein is radioactive

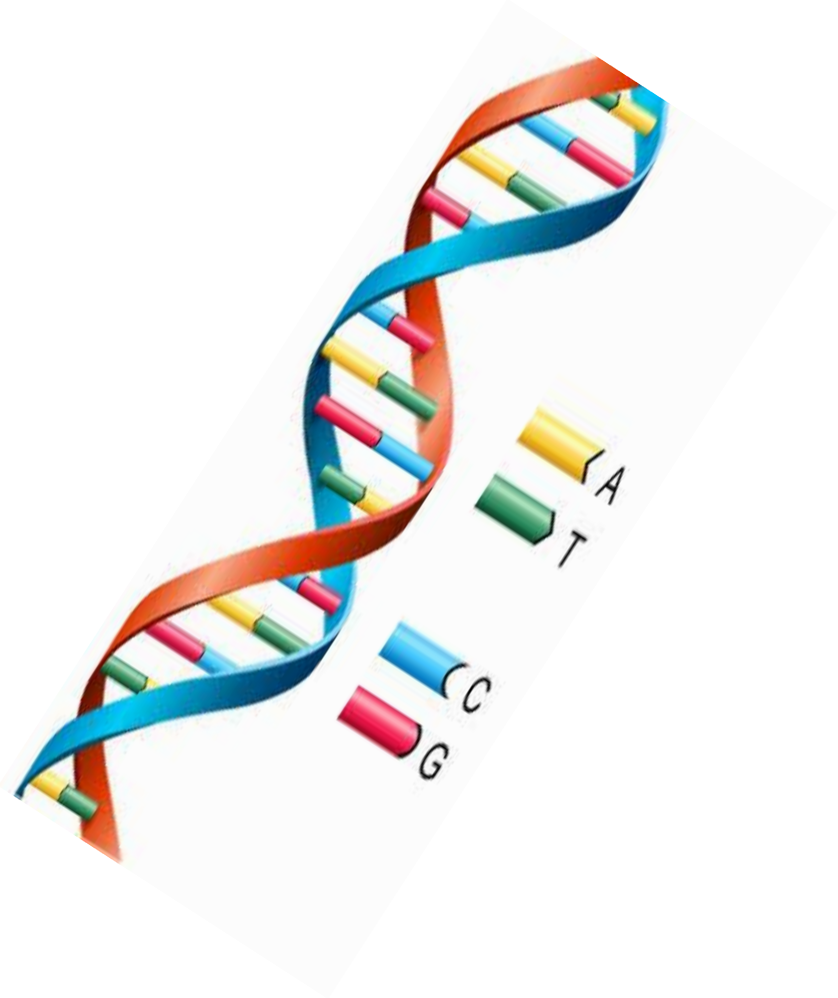
Pellet — Cells with DNA are not radioactive

🟥 Red denotes radioactivity

Cracking The Code

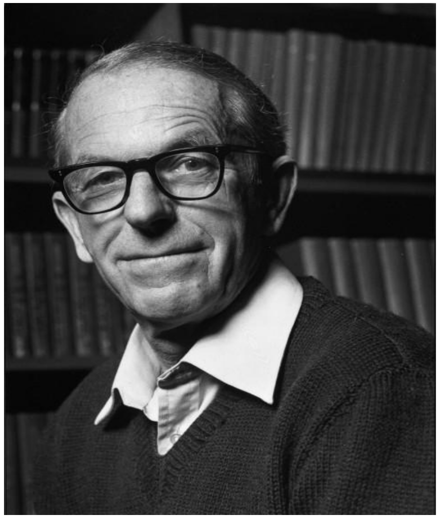


Watson and Crick
1953 DNA structure
Nobel Prize 1962



DNA Sequencing

Sanger Method
DNA sequencing by enzymatic synthesis



Frederick Sanger
Nobel Prize 1958, sequence of insulin
Nobel Prize 1980, DNA sequence

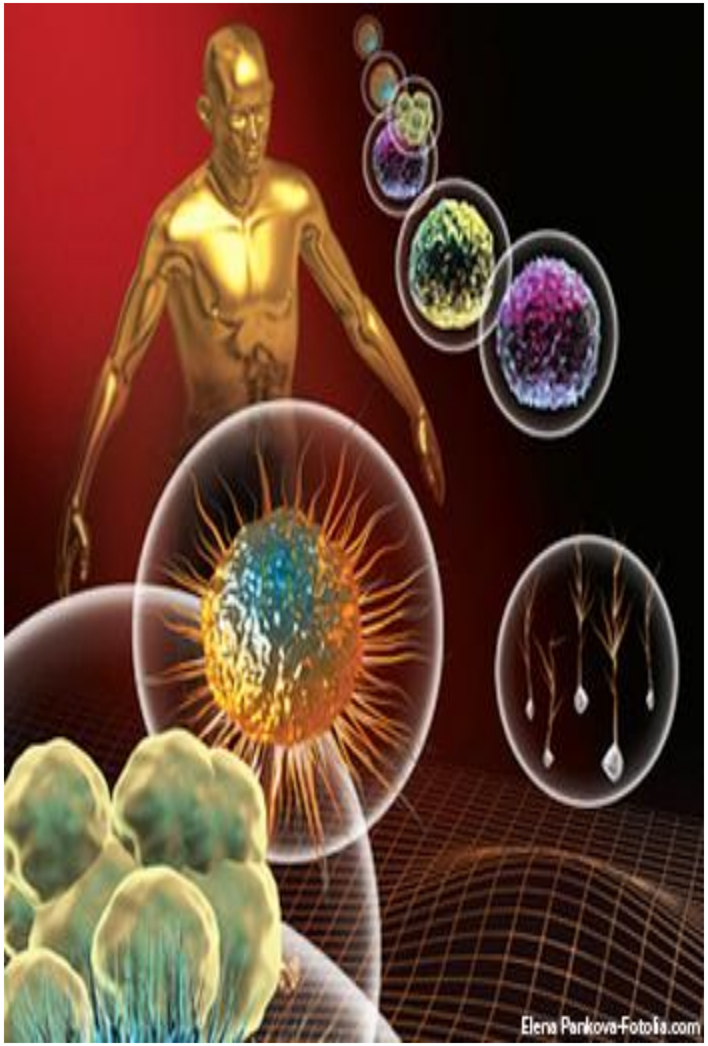
Maxam–Gilbert Method
DNA sequencing by chemical degradation

Walter Gilbert
Nobel Prize 1980, DNA sequence

What is DNA Sequencing?

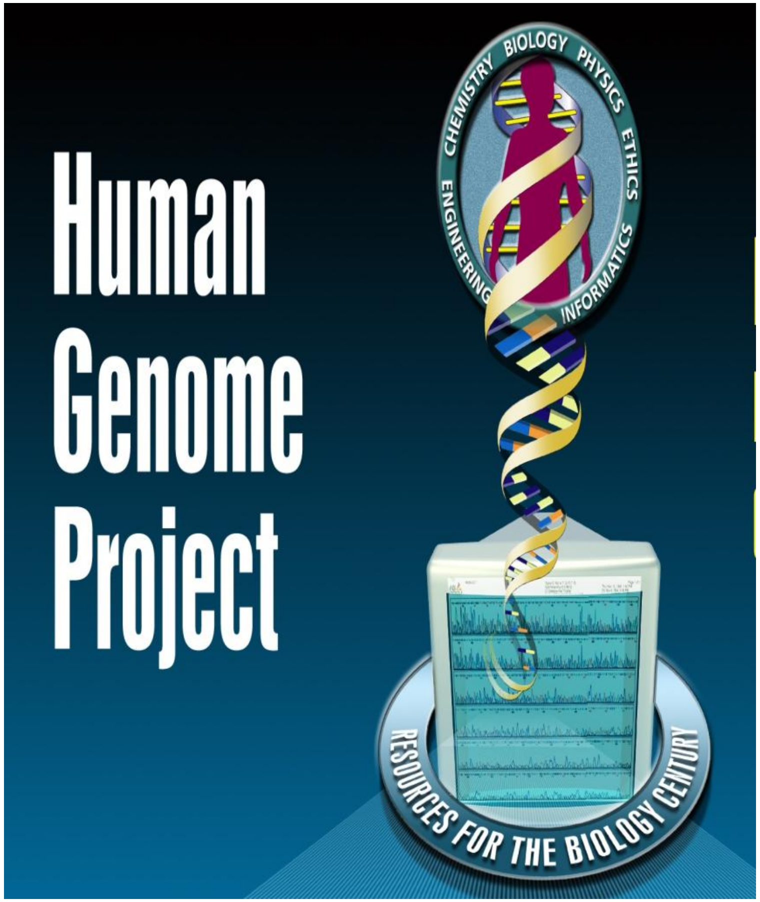
"**Sequencing**" means finding the order of nucleotides on a piece of DNA .

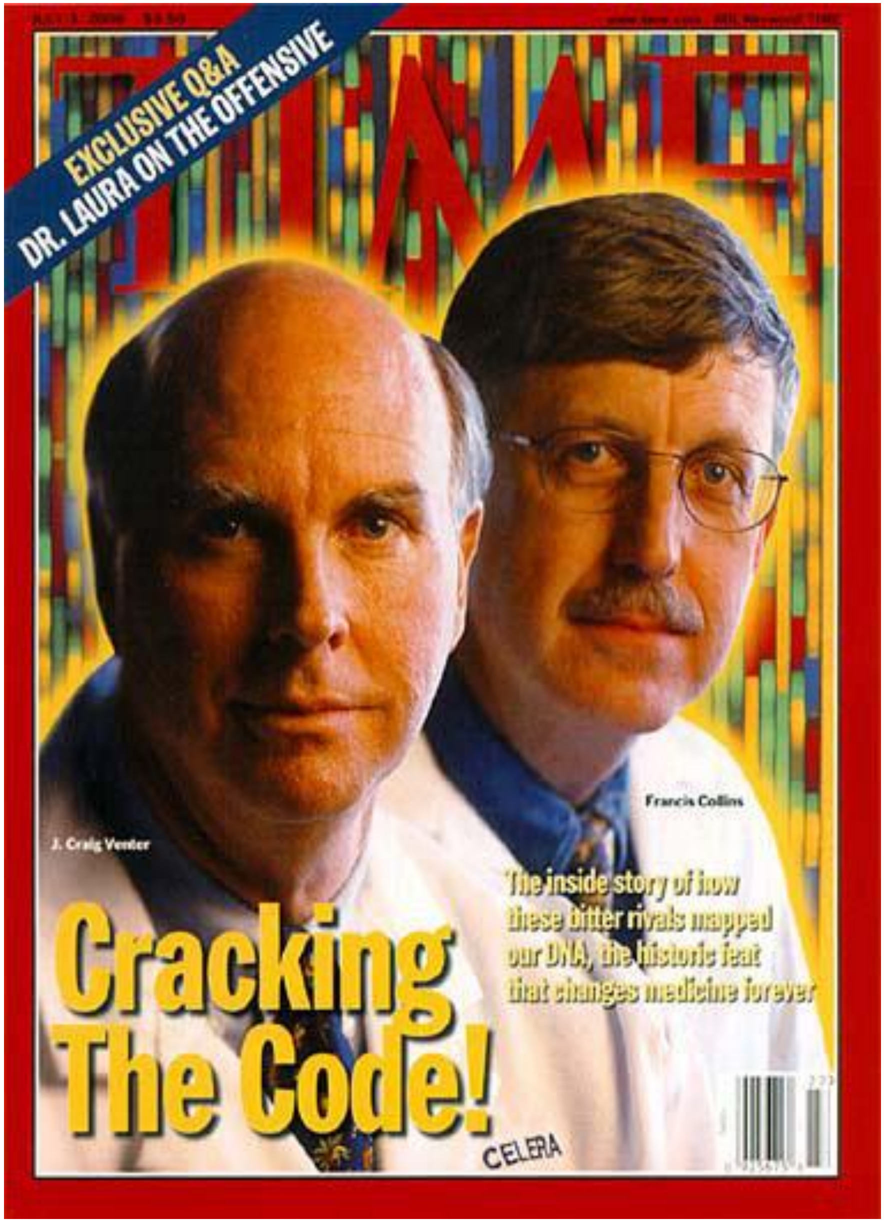
From Gene to Genome



Elena Pankova-Fotolia.com

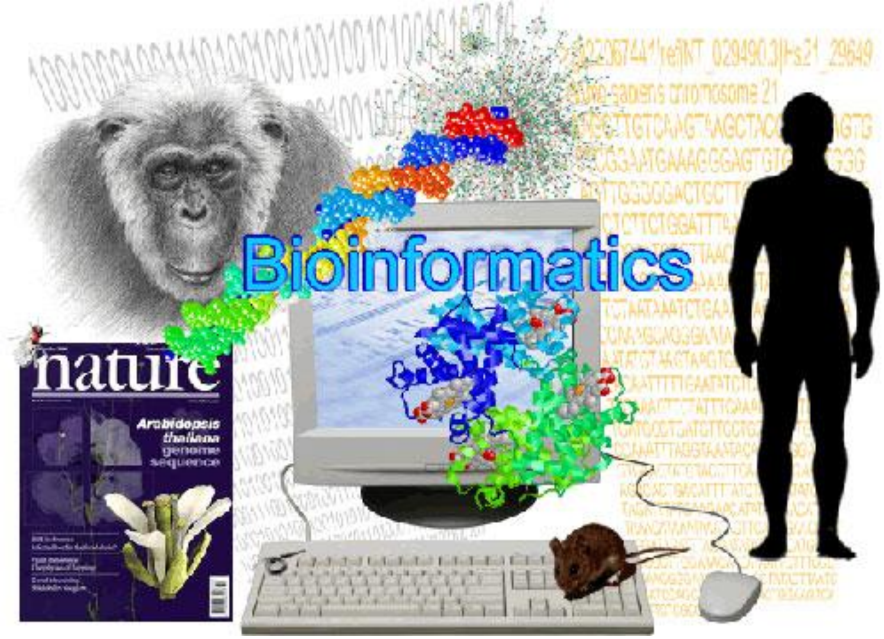
Human Genome Project
1990-2003
$3 billion




TIME
EXCLUSIVE Q&A
DR. LAURA ON THE OFFENSIVE
J. Craig Venter
Francis Collins
The inside story of how these bitter rivals mapped our DNA, the historic feat that changes medicine forever
Cracking The Code!
CELERA

Francis Collins
J. Craig Venter

Supercomputers &Bioinformatics

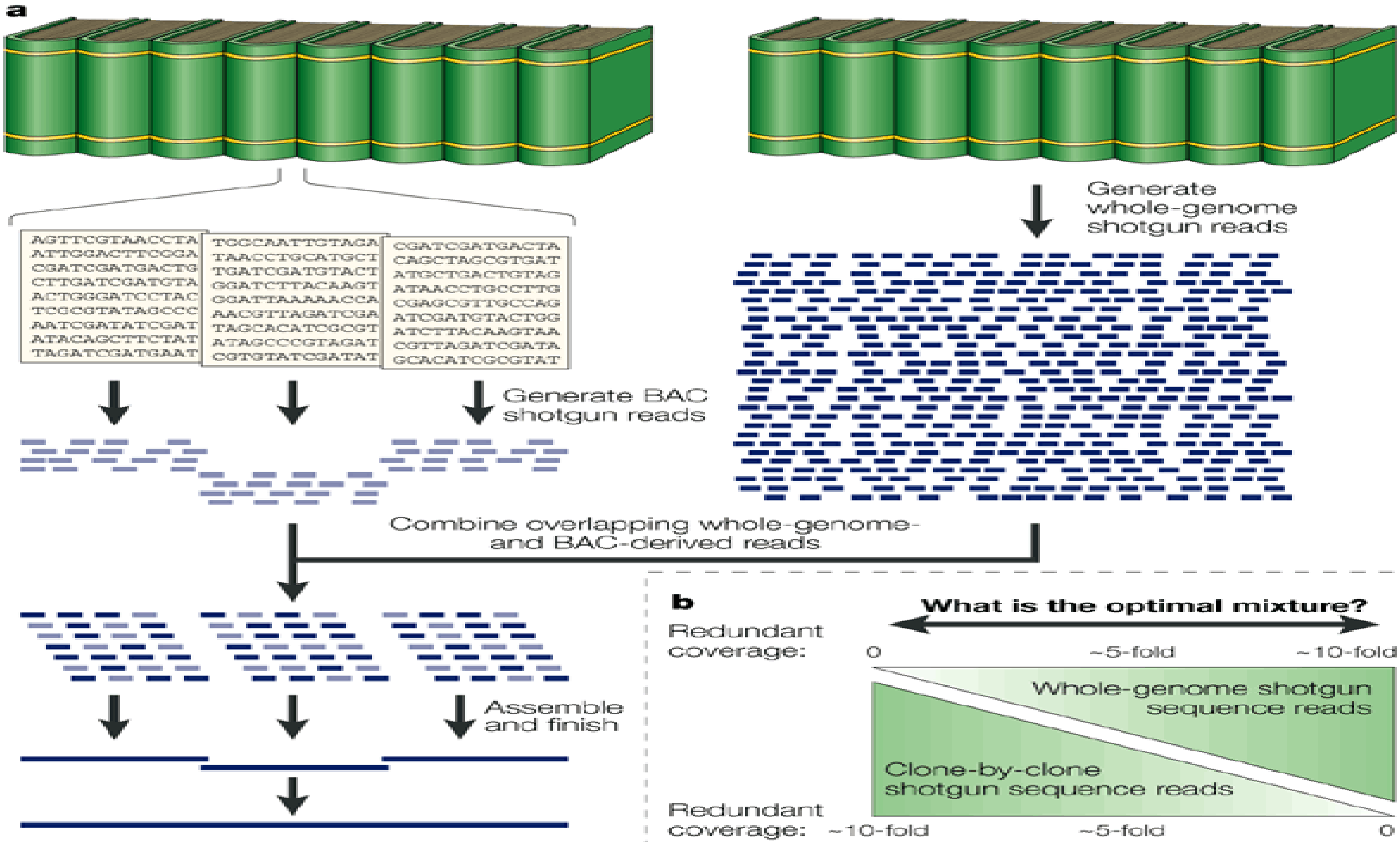

Bioinformatics
nature
Arabidopsis thaliana genome sequence
Homo sapiens chromosome 21

•Sequencing of the whole genome of the Organism

•Sequence must be annotated

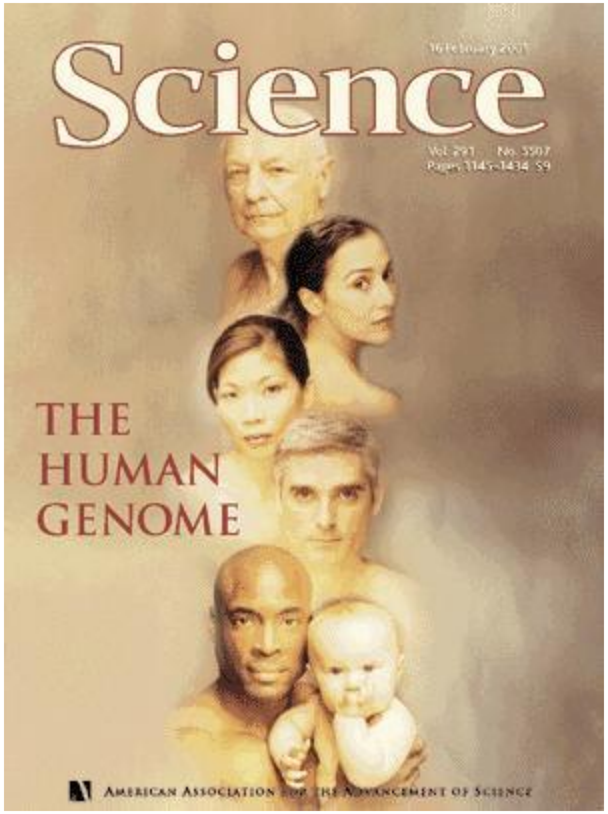
– Location of genes (locationofgenes)
– Location of transcribed regions (coding region)
– Location of promoters, start codons and terminators
– Function of other DNA sequences
– Translated Protein and assigned function

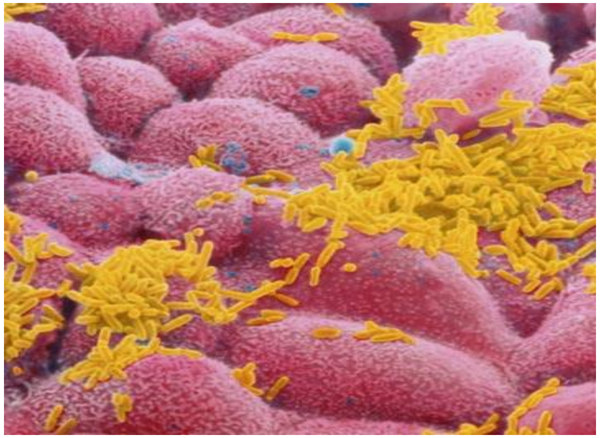
Celera



Nature Reviews | Genetics

Draft 2000
Complete 2003



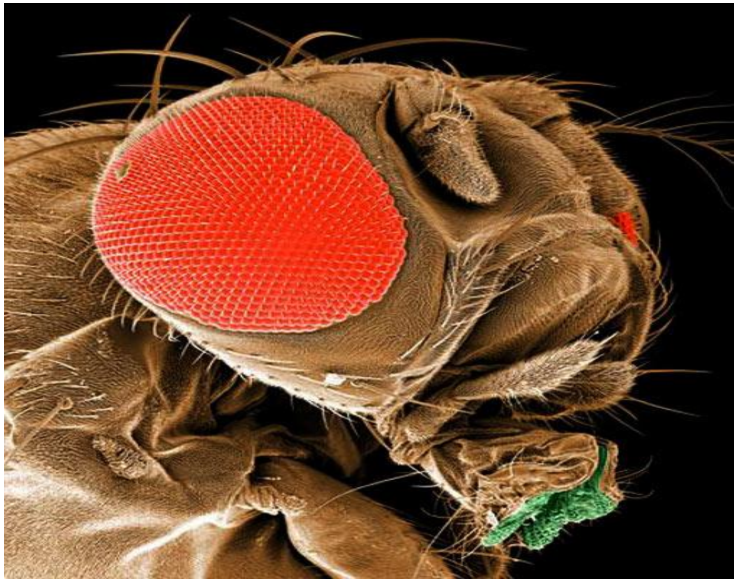


H. influenzae

1.8 million bases

1700 genes

First Bacteria by Celera

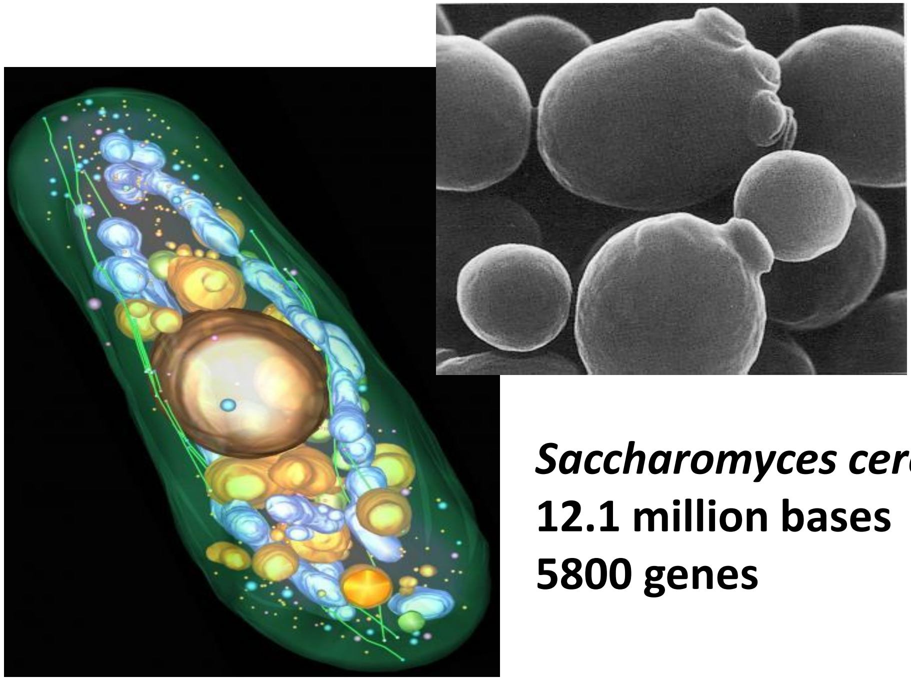


Drosophila melanogaster

137 million bases

13,700 genes

Celera



Saccharomyces cerevisiae

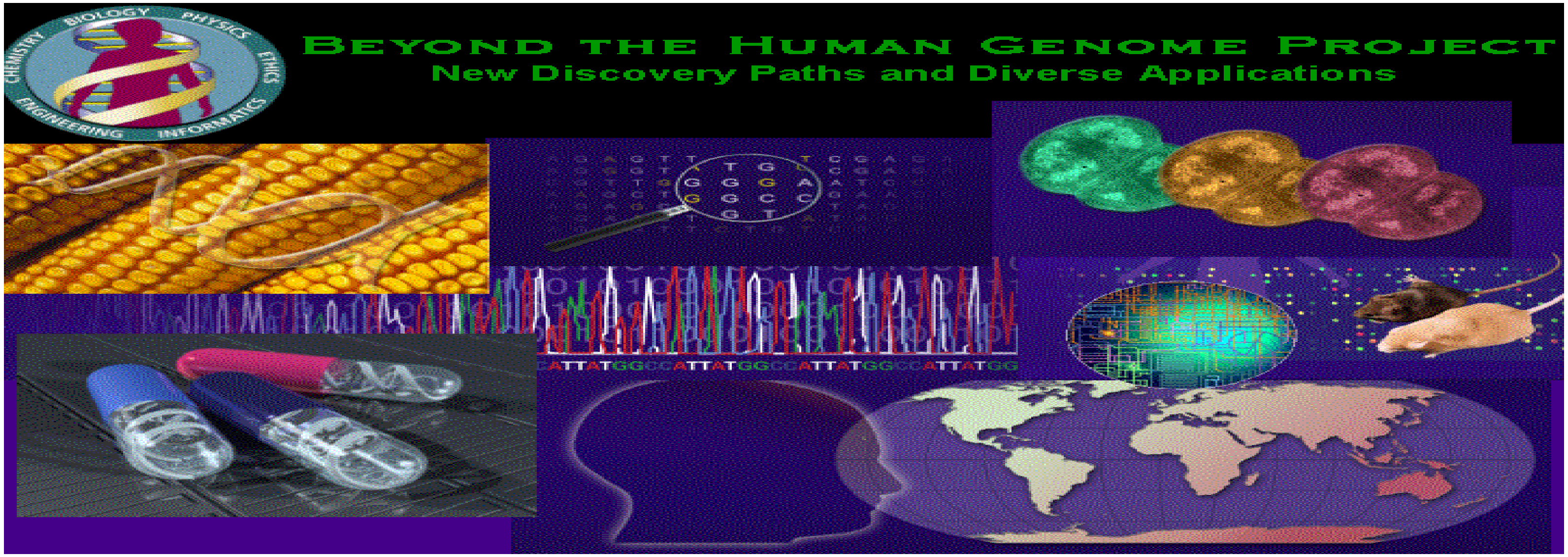
12.1 million bases

5800 genes

Caenorhabditis elegans
97 million bases
19,000 genes

***Oryza Sativa* (Rice)**
430 million base
60,000 genes

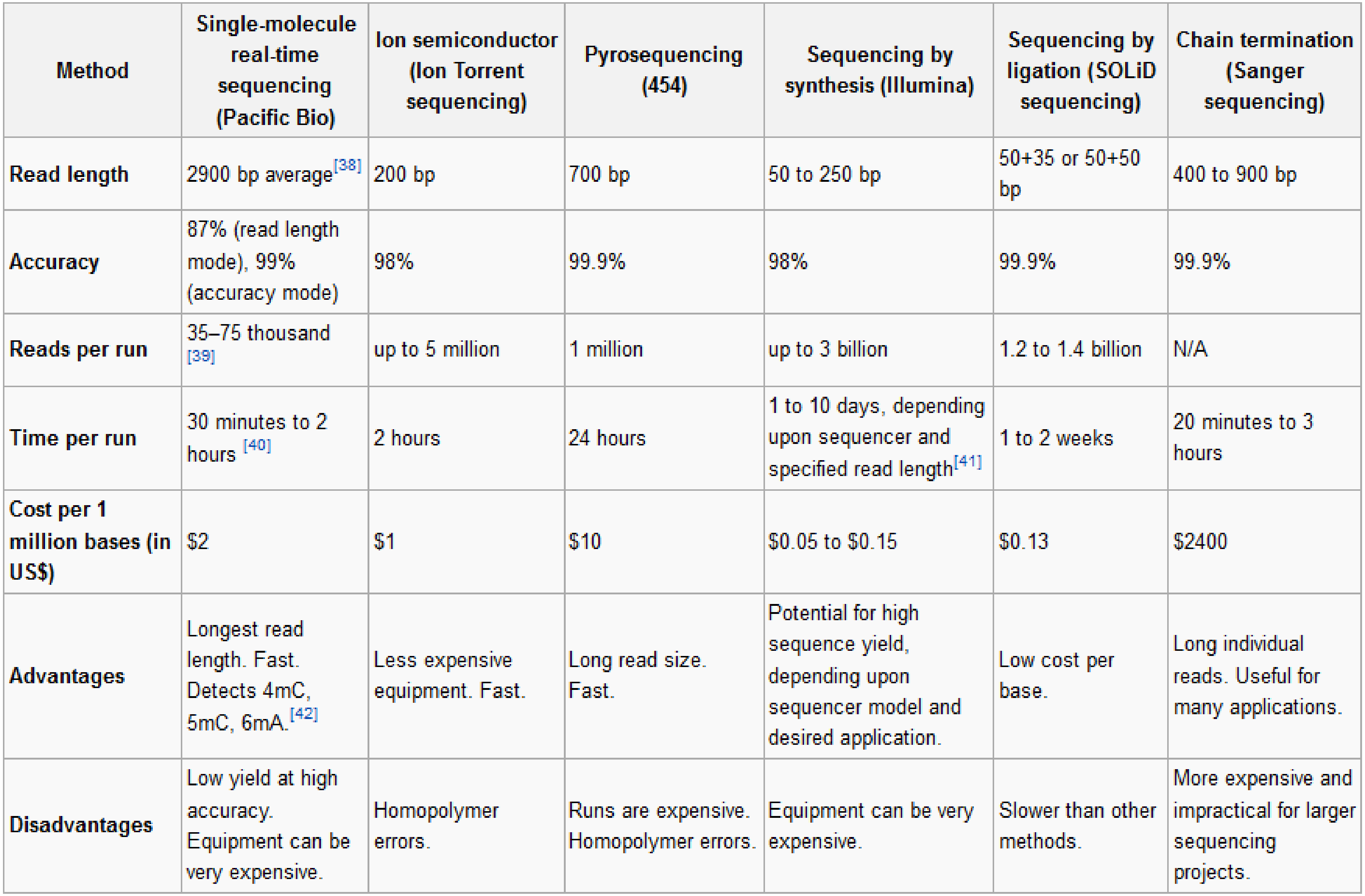
***Homo sapiens* (human)**
3.2 billion base
~25,000 genes
2% only code for protein
100,000 proposed earlier
40,000 after first draft


BEYOND THE HUMAN GENOME PROJECT
New Discovery Paths and Diverse Applications
BIOLOGY
PHYSICS
ETHICS
INFORMATICS
ENGINEERING
CHEMISTRY
CATTATGGCCATTATGGCCATTATGGCCATTATGG

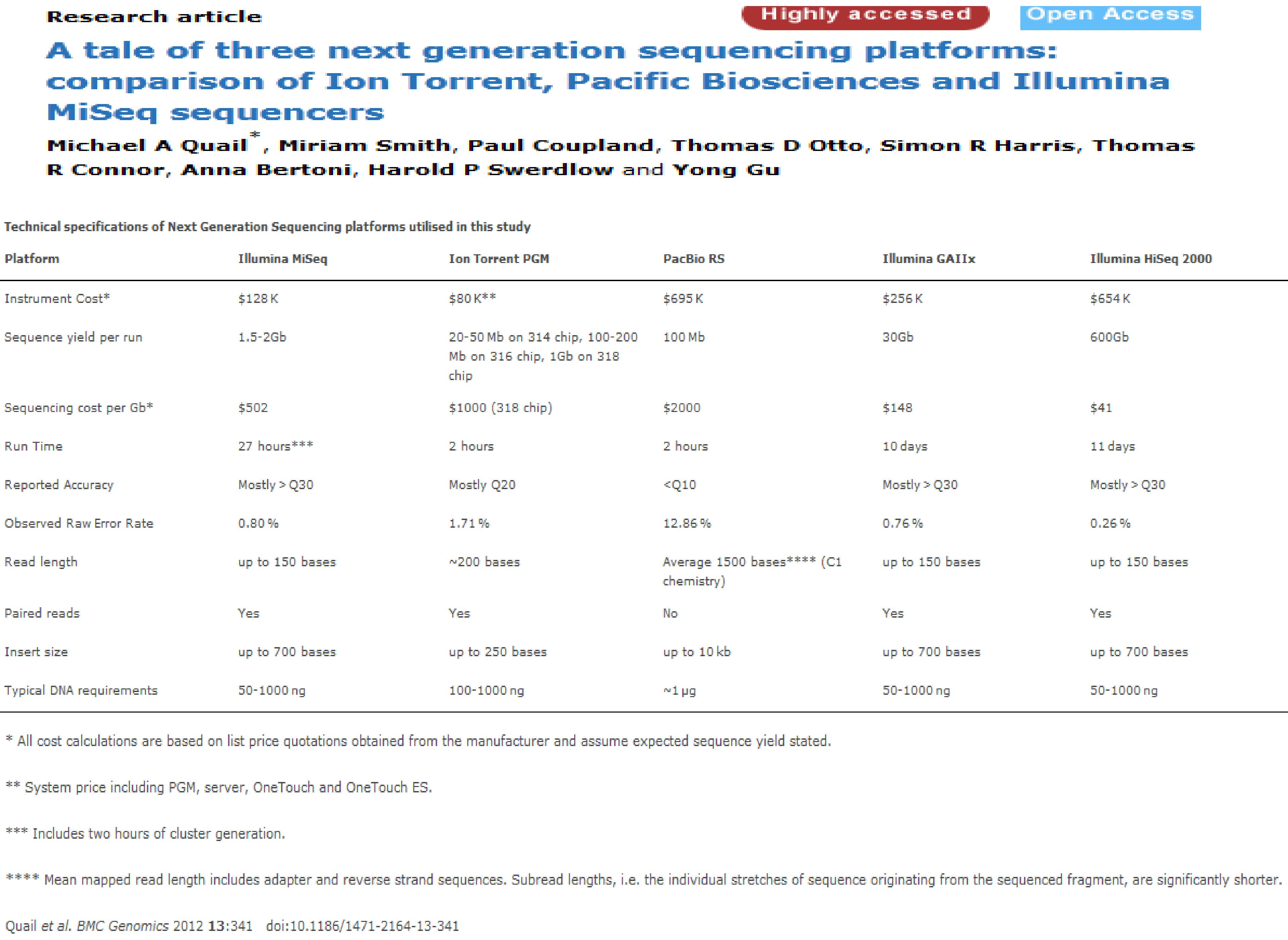
$1,000 Genome by 2015-2020

Every Child Genome

Comparison of next-generation sequencing methods [36][37]



Method	Single-molecule real-time sequencing (Pacific Bio)	Ion semiconductor (Ion Torrent sequencing)	Pyrosequencing (454)	Sequencing by synthesis (Illumina)	Sequencing by ligation (SOLiD sequencing)	Chain termination (Sanger sequencing)
Read length	2900 bp average[38]	200 bp	700 bp	50 to 250 bp	50+35 or 50+50 bp	400 to 900 bp
Accuracy	87% (read length mode), 99% (accuracy mode)	98%	99.9%	98%	99.9%	99.9%
Reads per run	35–75 thousand [39]	up to 5 million	1 million	up to 3 billion	1.2 to 1.4 billion	N/A
Time per run	30 minutes to 2 hours [40]	2 hours	24 hours	1 to 10 days, depending upon sequencer and specified read length[41]	1 to 2 weeks	20 minutes to 3 hours
Cost per 1 million bases (in US$)	$2	$1	$10	$0.05 to $0.15	$0.13	$2400
Advantages	Longest read length. Fast. Detects 4mC, 5mC, 6mA.[42]	Less expensive equipment. Fast.	Long read size. Fast.	Potential for high sequence yield, depending upon sequencer model and desired application.	Low cost per base.	Long individual reads. Useful for many applications.
Disadvantages	Low yield at high accuracy. Equipment can be very expensive.	Homopolymer errors.	Runs are expensive. Homopolymer errors.	Equipment can be very expensive.	Slower than other methods.	More expensive and impractical for larger sequencing projects.



Research article

Highly accessed | Open Access

A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers

Michael A Quail*, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow and Yong Gu

Technical specifications of Next Generation Sequencing platforms utilised in this study

Platform	Illumina MiSeq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000
Instrument Cost*	$128 K	$80 K**	$695 K	$256 K	$654 K
Sequence yield per run	1.5-2Gb	20-50 Mb on 314 chip, 100-200 Mb on 316 chip, 1Gb on 318 chip	100 Mb	30Gb	600Gb
Sequencing cost per Gb*	$502	$1000 (318 chip)	$2000	$148	$41
Run Time	27 hours***	2 hours	2 hours	10 days	11 days
Reported Accuracy	Mostly > Q30	Mostly Q20	<Q10	Mostly > Q30	Mostly > Q30
Observed Raw Error Rate	0.80 %	1.71 %	12.86 %	0.76 %	0.26 %
Read length	up to 150 bases	~200 bases	Average 1500 bases**** (C1 chemistry)	up to 150 bases	up to 150 bases
Paired reads	Yes	Yes	No	Yes	Yes
Insert size	up to 700 bases	up to 250 bases	up to 10 kb	up to 700 bases	up to 700 bases
Typical DNA requirements	50-1000 ng	100-1000 ng	~1 μg	50-1000 ng	50-1000 ng

* All cost calculations are based on list price quotations obtained from the manufacturer and assume expected sequence yield stated.

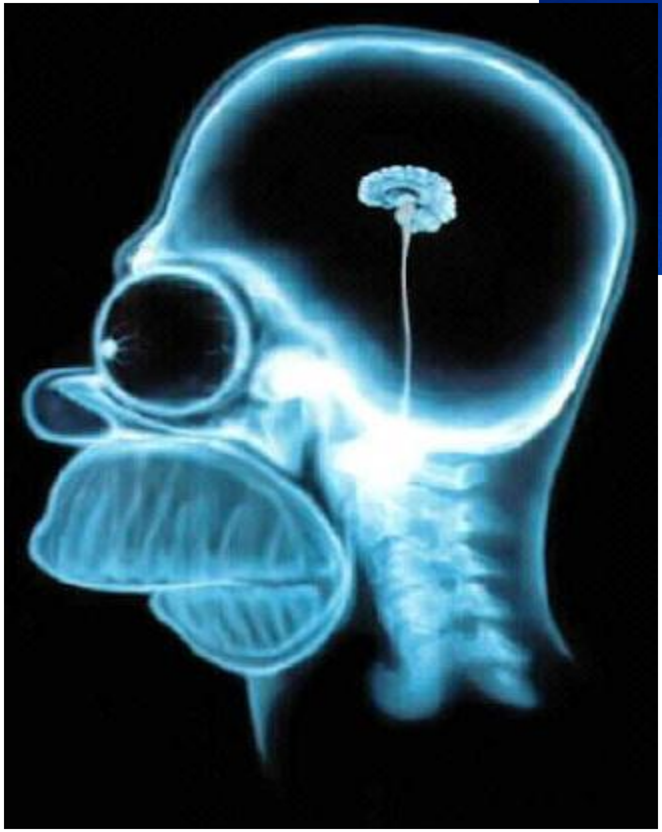
** System price including PGM, server, OneTouch and OneTouch ES.

*** Includes two hours of cluster generation.

**** Mean mapped read length includes adapter and reverse strand sequences. Subread lengths, i.e. the individual stretches of sequence originating from the sequenced fragment, are significantly shorter.

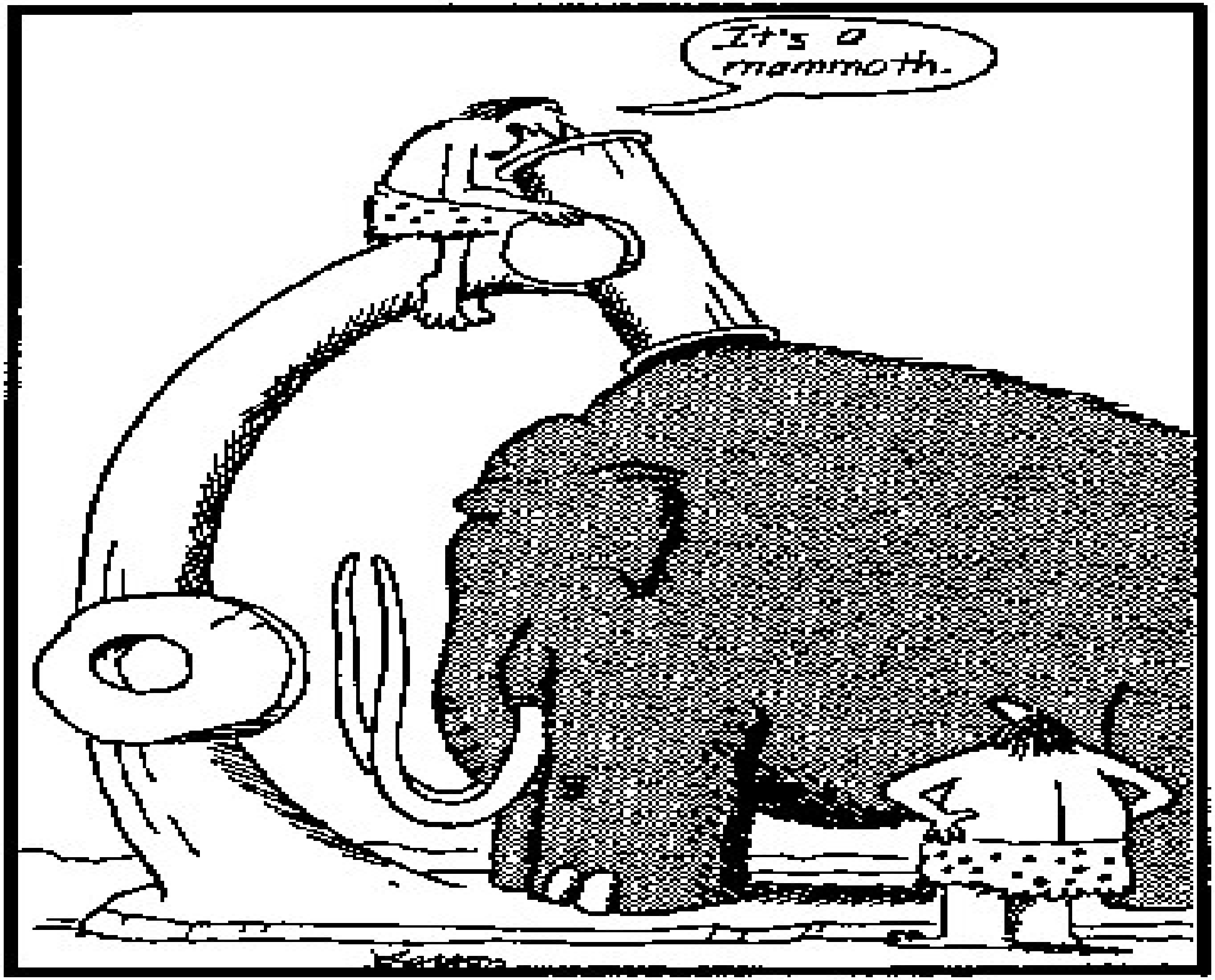
Quail *et al. BMC Genomics* 2012 **13**:341 doi:10.1186/1471-2164-13-341

Sequencing is just letters



How to use it?

Questions



Early microscope