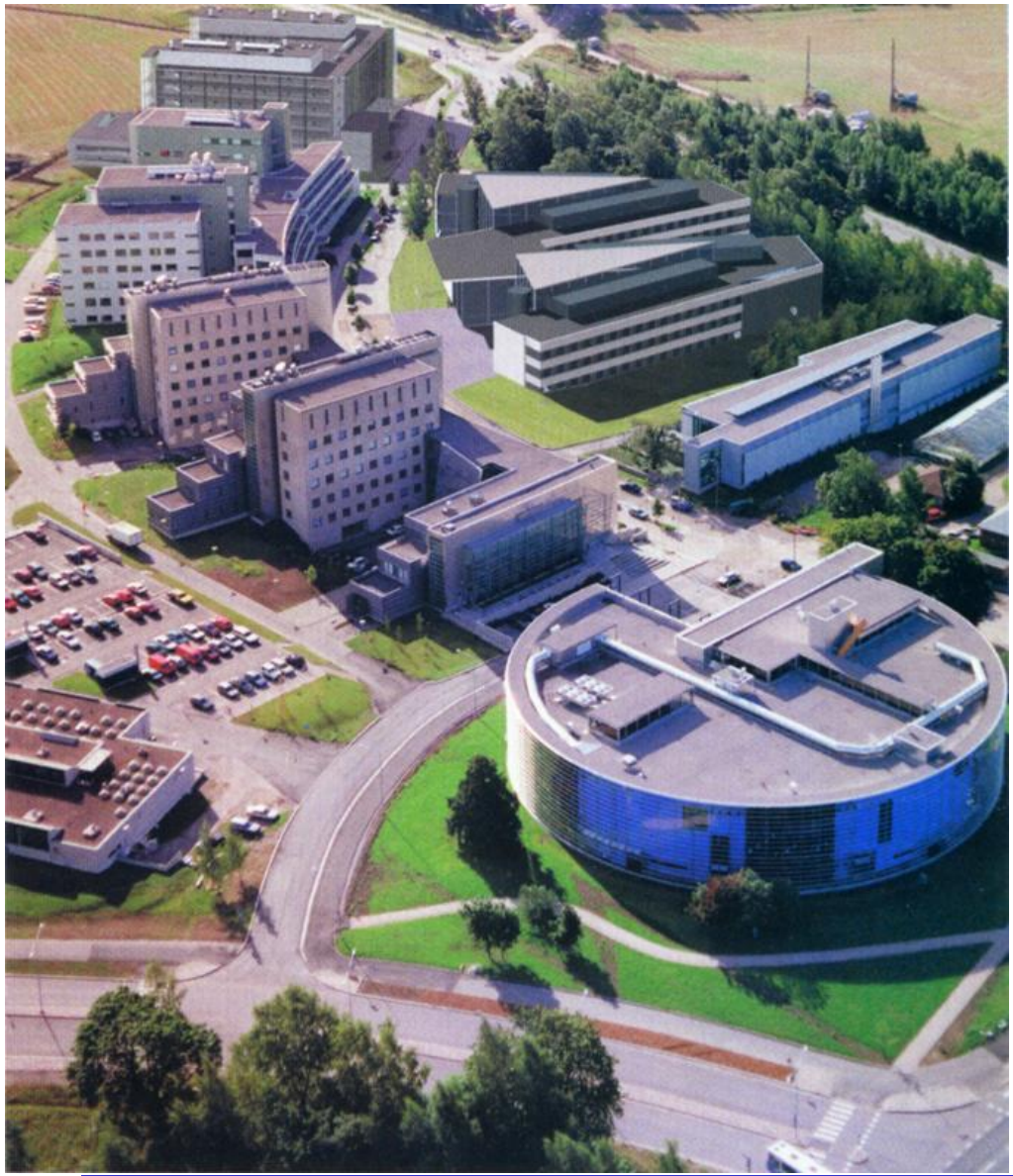
VIIKKI BIOCENTER
University of Helsinki



Mitä uudet DNA sekvensointimenetelmät voivat paljastaa luonnonjärjestelmästä?

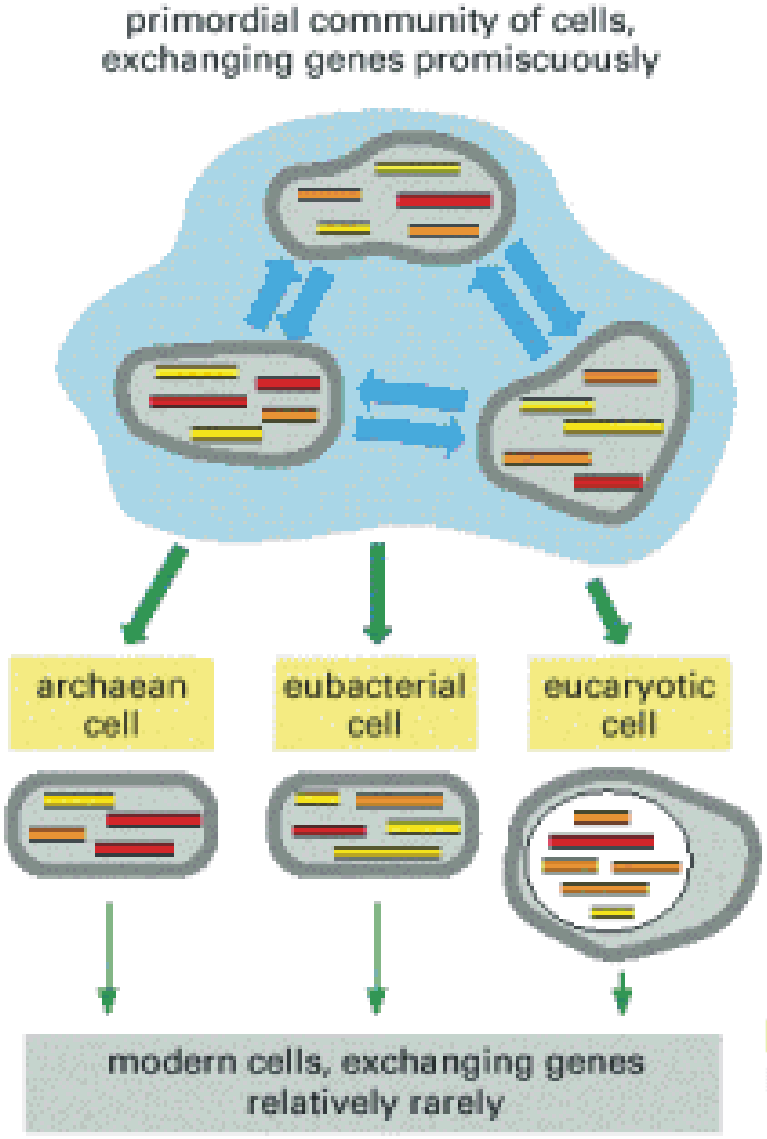
Petri Auvinen

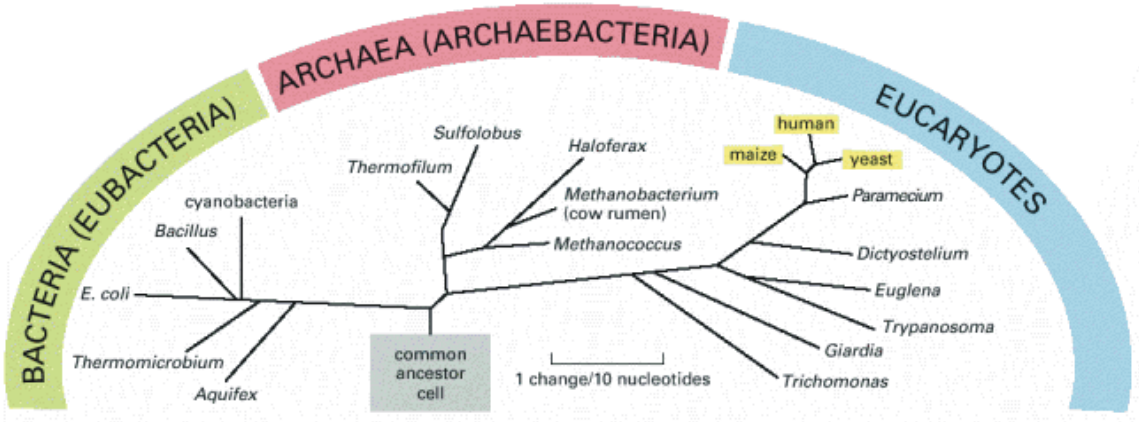
DNA Sequencing and Genomics Laboratory

Institute of Biotechnology

Kuinka solut kehittyivät?

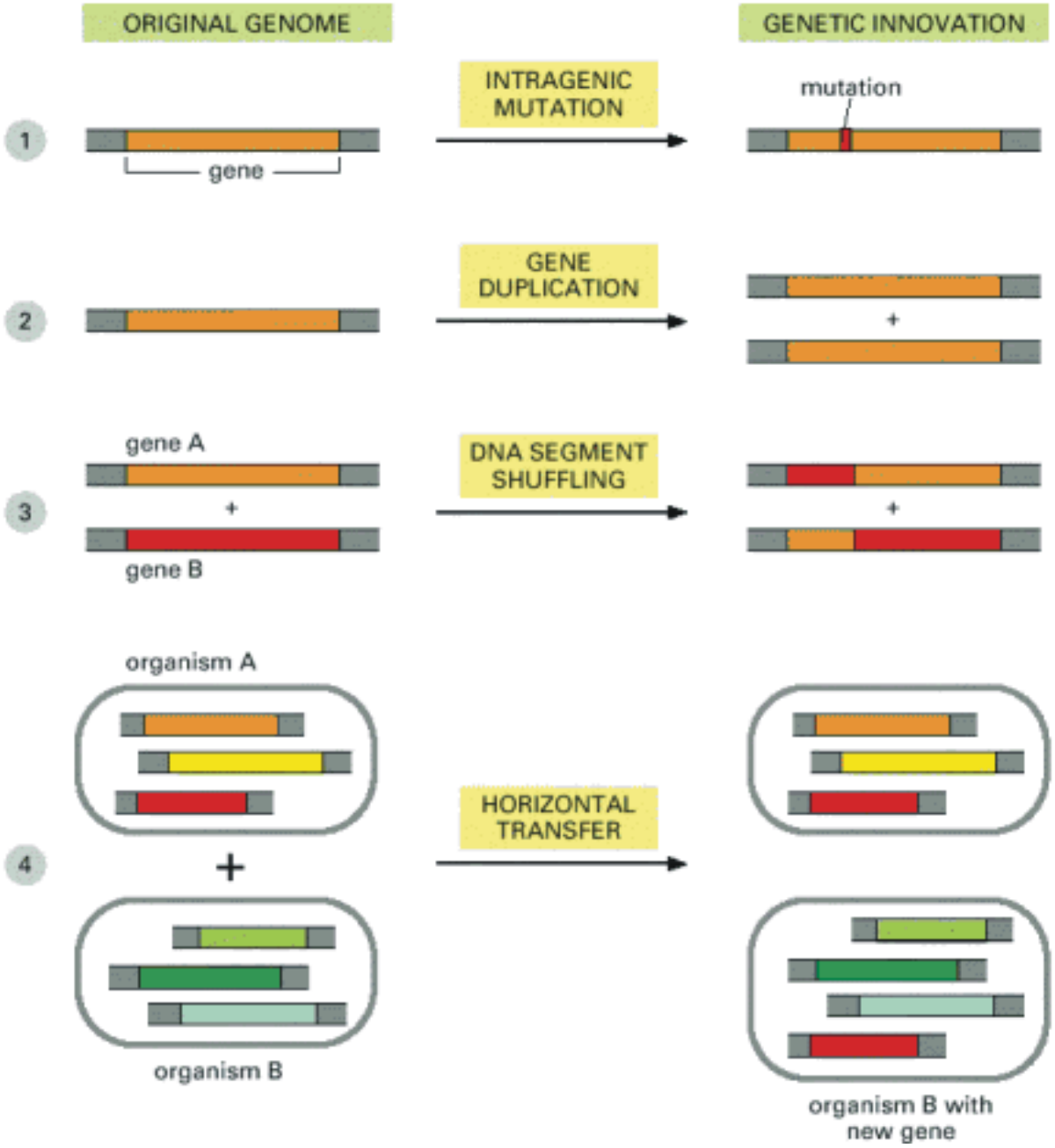
Kolmenlaisia soluja



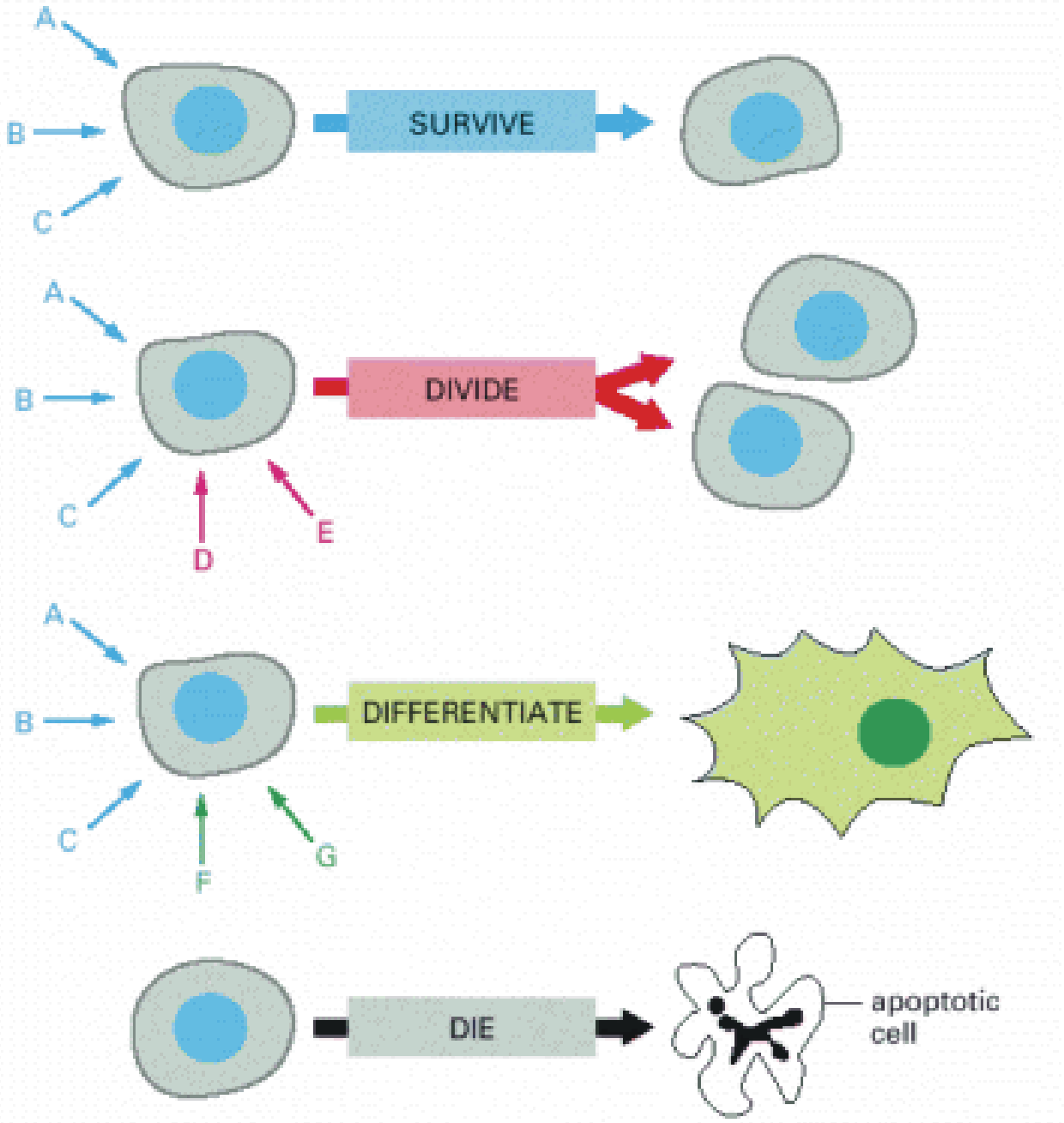


Petri Auvinen , DNA Sequencing and Genomics Laboratory,
Institute of Biotechnology, University of Helsinki

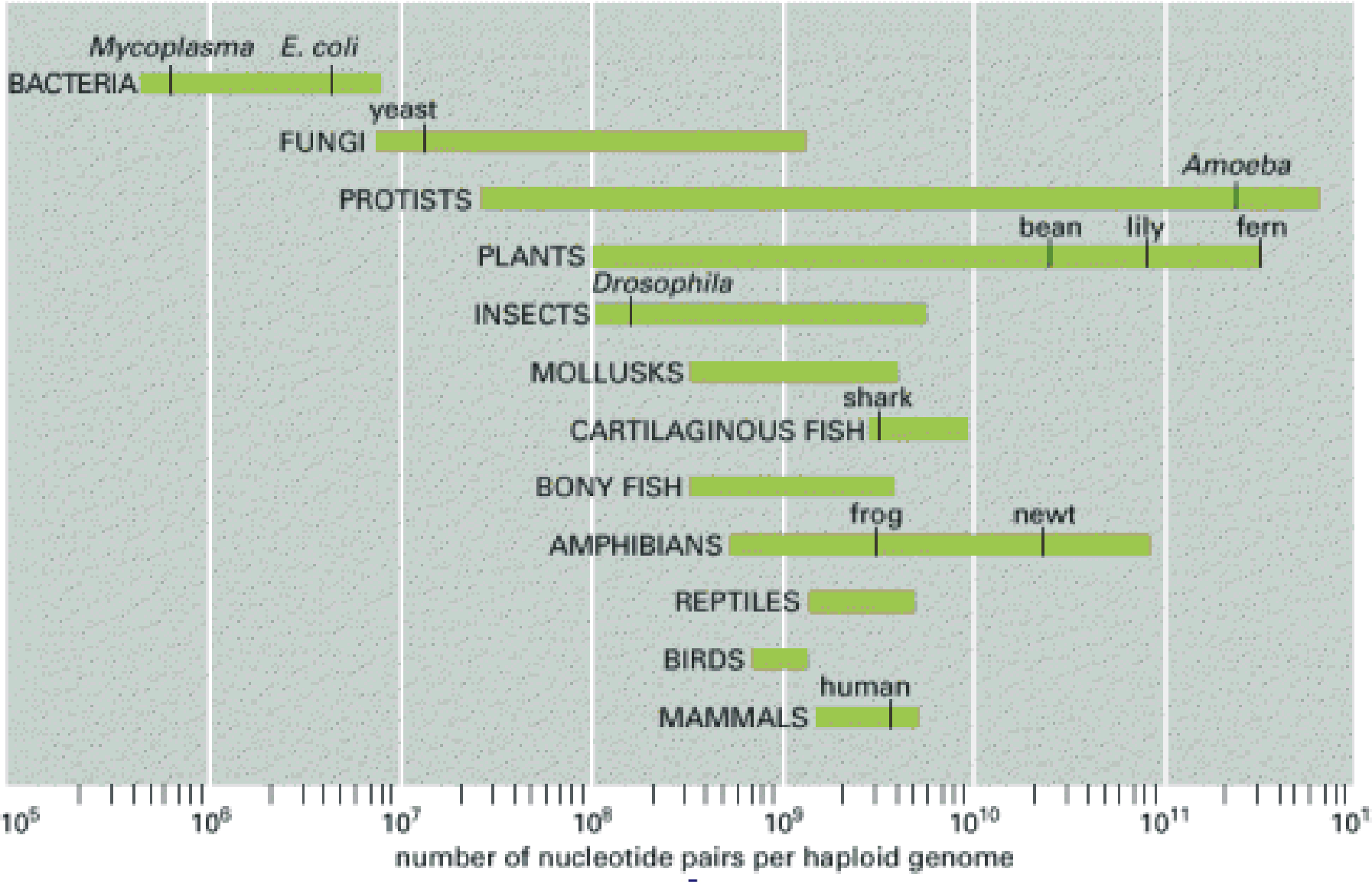
Innovaatiojärjestelmät



Solujen kohtalot



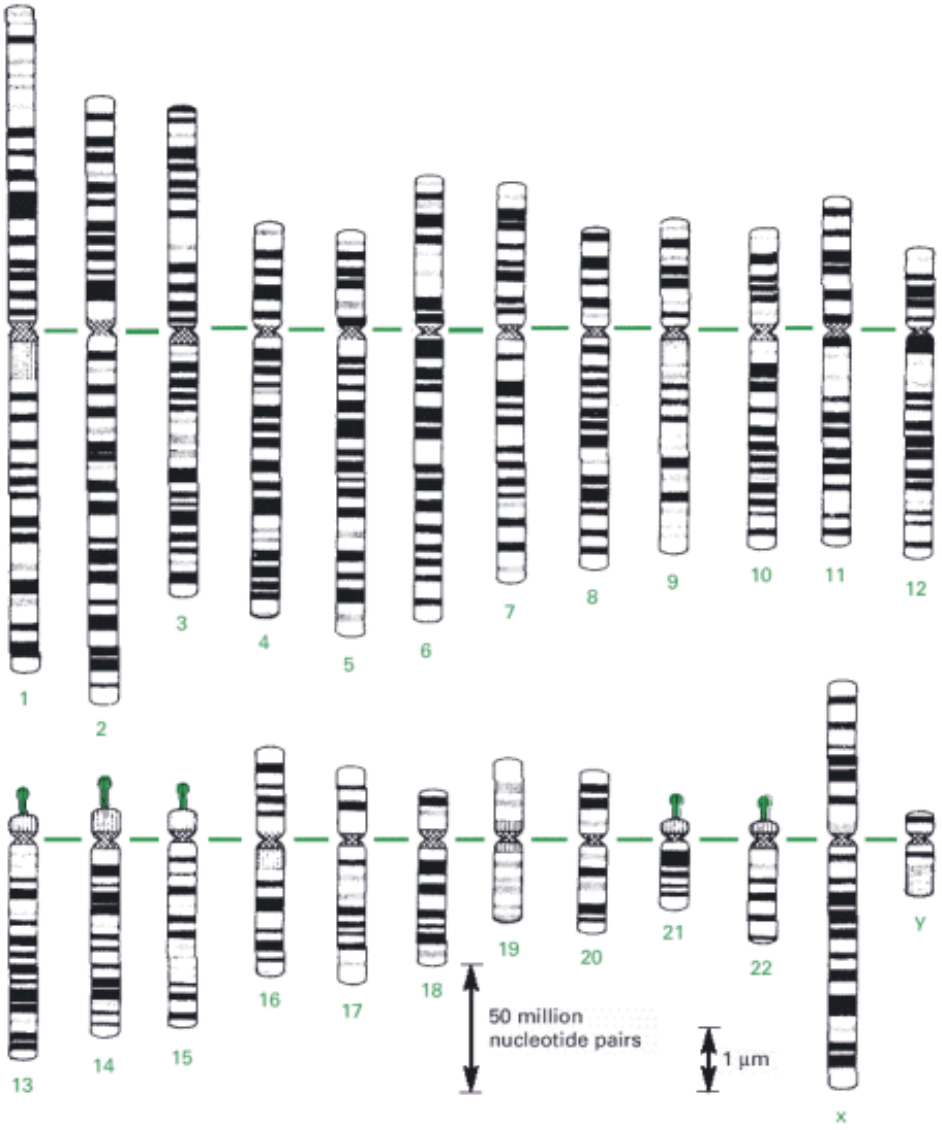
Genomien koko



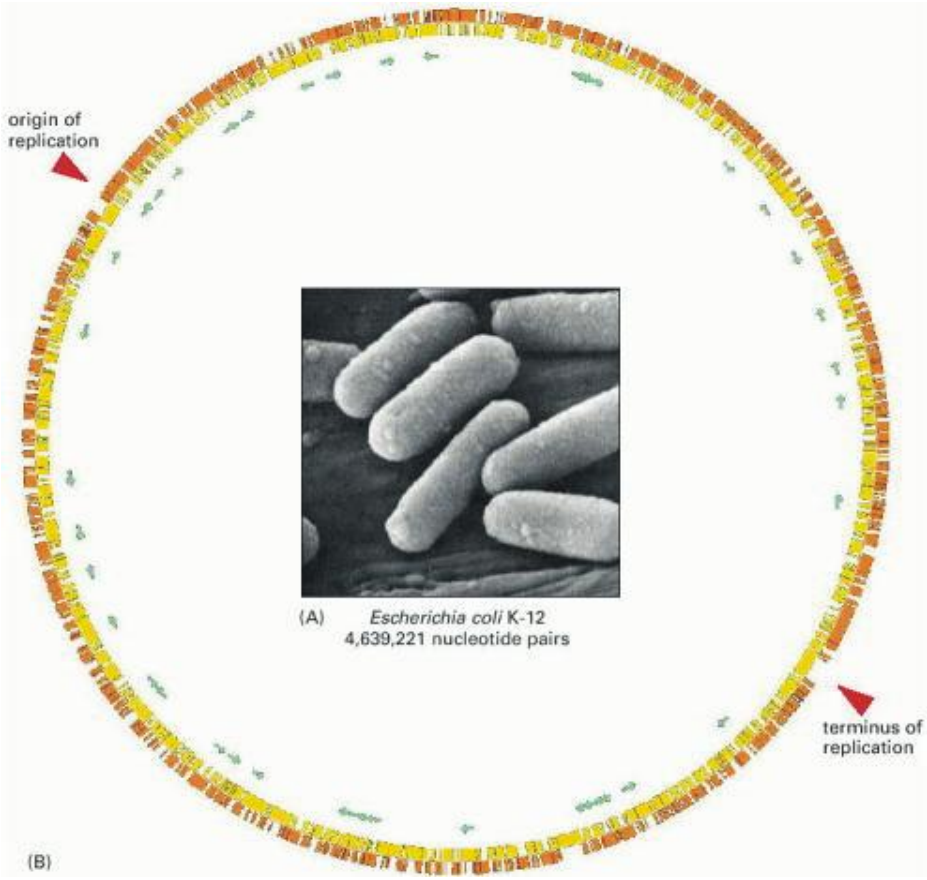
Kuinka monta geeniä tarvitaan?

Species	Size of genome (Mb)	Approximate number of genes	References
Eukaryotes			
Arabidopsis thaliana (plant)	125	25 500	AGI (2000)
Caenorhabditis elegans (nematode worm)	97	19 000	CESC (1998)
Drosophila melanogaster (fruit fly)	180	13 600	Adams *et al.* (2000)
Homo sapiens (human)	3200	30 000–40 000	IHGSC (2001); Venter *et al.* (2001)
Saccharomyces cerevisiae (yeast)	12.1	5800	Goffeau *et al.* (1996)
Bacteria			
Escherichia coli K12	4.64	4400	Blattner *et al.* (1997)
Mycobacterium tuberculosis H37Rv	4.41	4000	Cole *et al.* (1998)
Mycoplasma genitalium	0.58	500	Fraser *et al.* (1995)
Pseudomonas aeruginosa PA01	6.26	5700	Stover *et al.* (2000)
Streptococcus pneumoniae	2.16	2300	Tettelin *et al.* (2001)
Vibrio cholerae El Tor N16961	4.03	4000	Heidelberg *et al.* (2000)
Yersinia pestis CO92	4.65	4100	Parkhill *et al.* (2001)
Archaea			
Archaeoglobus fulgidus	2.18	2500	Klenk *et al.* (1997)
Methanococcus jannaschii	1.66	1750	Bult *et al.* (1996)

Ihmisen kromosomit



Kolibakteerin kromosomi



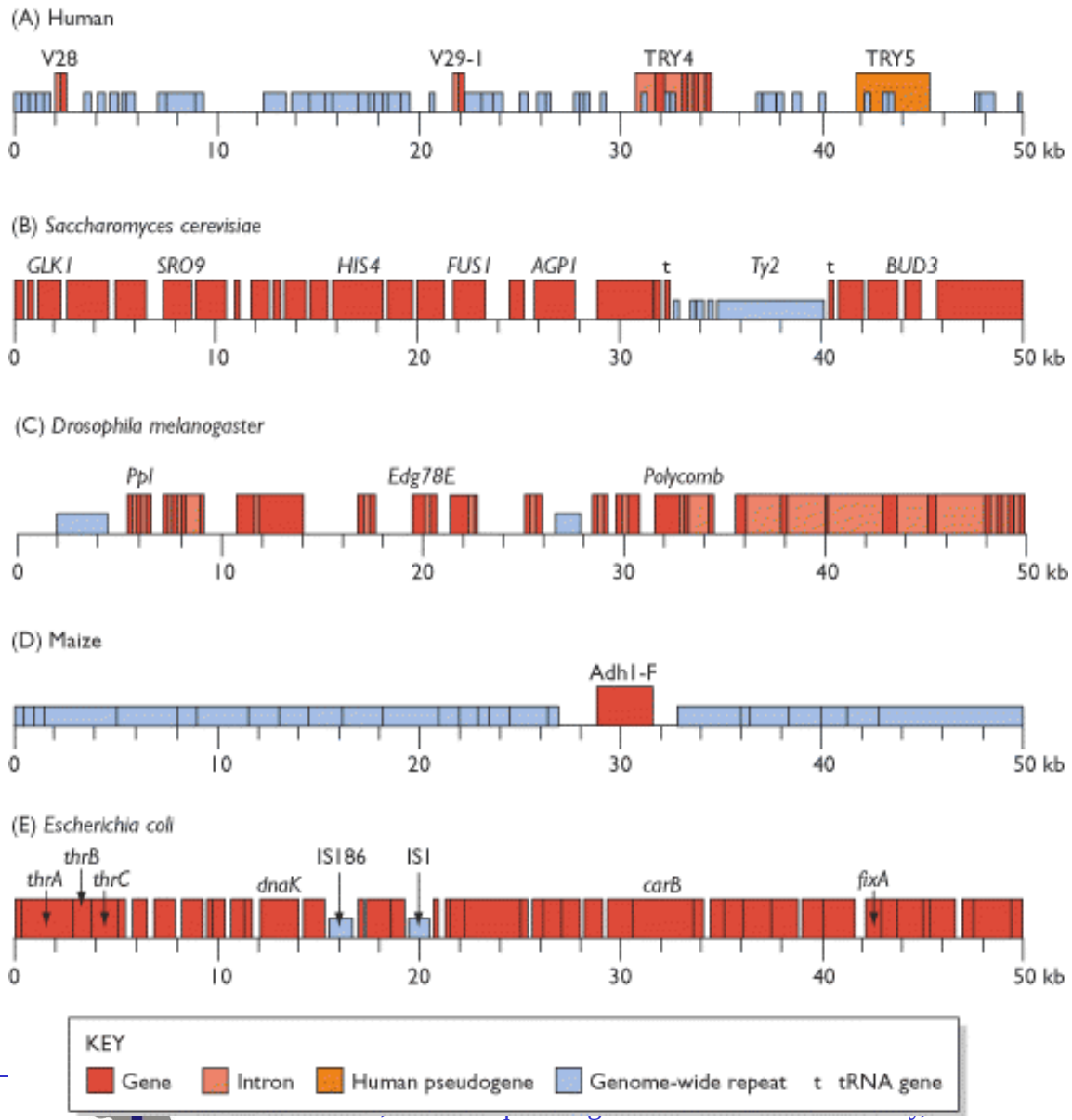
Petri Auvinen , DNA Sequencing and Genomics Laboratory,
Institute of Biotechnology, University of Helsinki

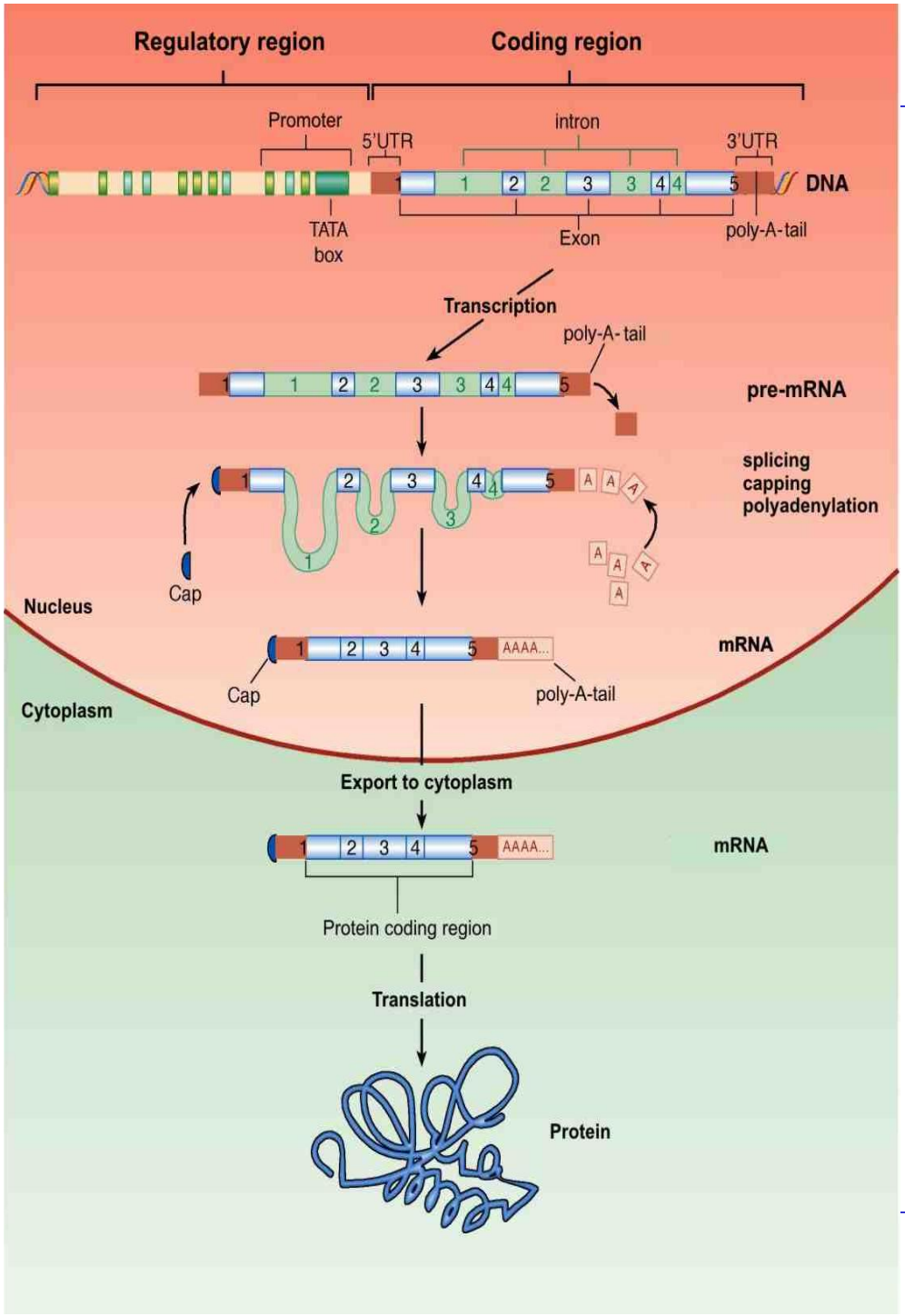
Genome

-2-3 % of the human genome is coding for proteins and rest is non coding
-non coding consists of is introns, intergenic regions

-but 5% of the human genome is under purifying selection by analysis of the sequence
-what is this 2-3% which is not coding for proteins?

-small RNAs (miRNA, siRNA, long noncoding RNAs, regulatory sequences
-long RNA might not be conserved by sequence still having function?





Eukaryoottien geenien ilmentyminen

DNA

↓ **Transcription**

pre-mRNA

↓ **Processing**

mRNA

↓ **Export**

↓ **Translation**

protein

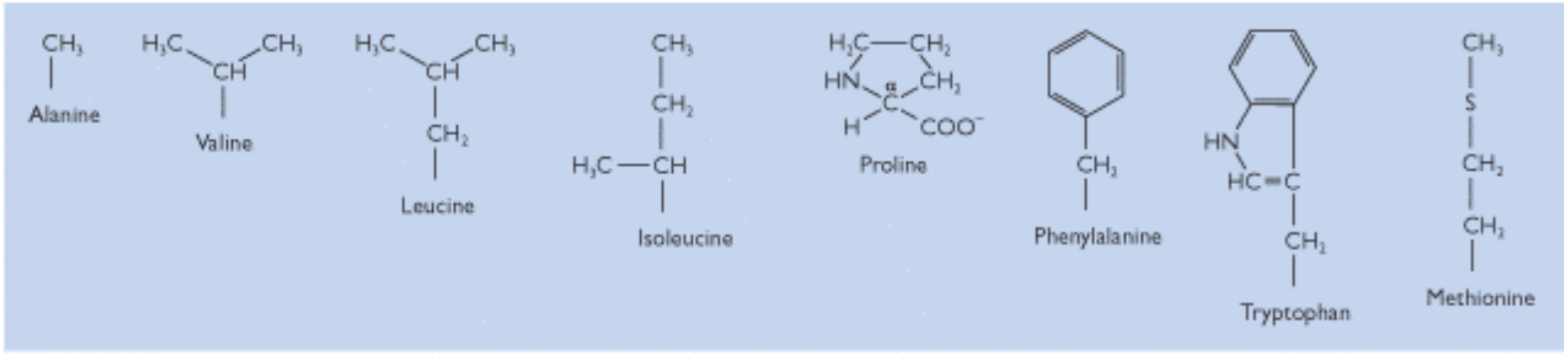
Petri Auvinen , DNA Sequencing and Genomics Laboratory,
Institute of Biotechnology, University of Helsinki

THE CODE

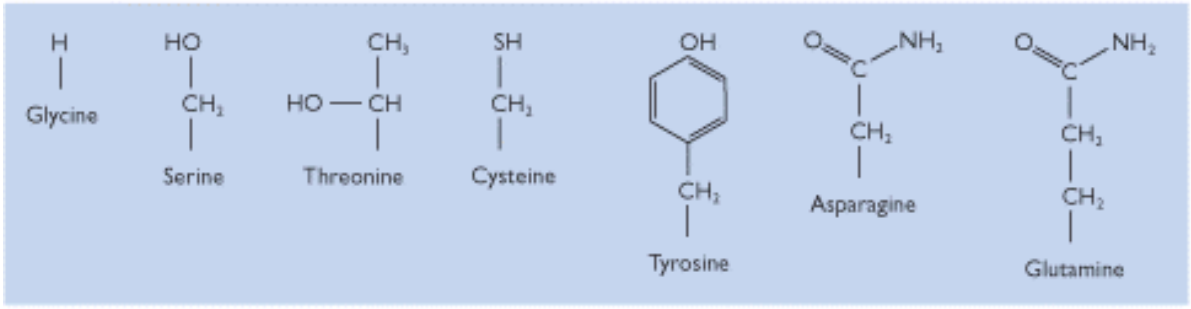
UUU, UUC: phe UUA, UUG: leu	UCU, UCC, UCA, UCG: ser	UAU, UAC: tyr UAA, UAG: stop	UGU, UGC: cys UGA: stop UGG: trp
CUU, CUC, CUA, CUG: leu	CCU, CCC, CCA, CCG: pro	CAU, CAC: his CAA, CAG: gln	CGU, CGC, CGA, CGG: arg
AUU, AUC, AUA: ile AUG: met	ACU, ACC, ACA, ACG: thr	AAU, AAC: asn AAA, AAG: lys	AGU, AGC: ser AGA, AGG: arg
GUU, GUC, GUA, GUG: val	GCU, GCC, GCA, GCG: ala	GAU, GAC: asp GAA, GAG: glu	GGU, GGC, GGA, GGG: gly

Amino acid side groups

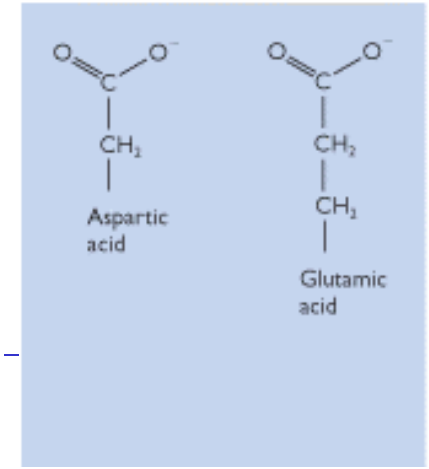
(A) Non-polar R groups



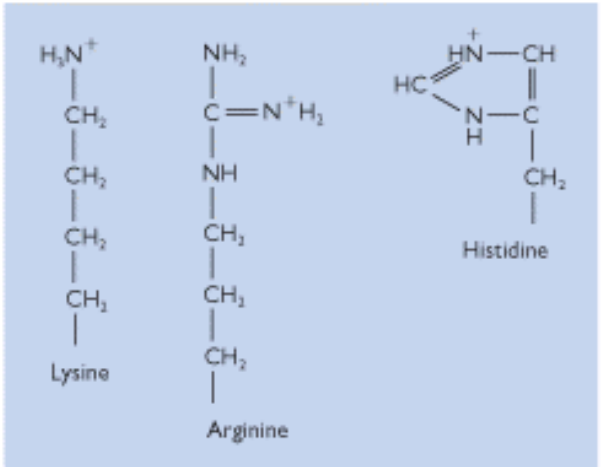
(B) Polar R groups



(C) Negatively charged R groups

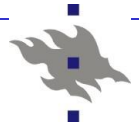


(D) Positively charged R groups

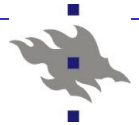


Genomitutkimuksen historian lyhyt oppimäärä

- 1865 Mendel genetiikan perusteet
- 1953 Watson and Crick kaksoiskierre
- 1966 Nirenberg, Khorana, Holley geneettinen koodi
- 1977 Maxam ja Gilbert sekä Sanger sekvensointi
- 1982 Genebank perustettiin
- 1990 Ihmisen genomiprojekti aloitettiin
- 1996 Hiivan genomisekvenssi valmistui
- 1998 *C. elegans* sekvenssi valmistui
- 1997 *E. coli* sekvenssi valmistui
- 2000 *D. melanogaster*
- 2001 Ihmisen genomin draft versio

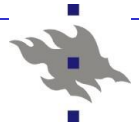


- 2002 Hiiren genomi
- 2003 Ihmisen genomiprojekti loppui
- 2004 Ihmisen genomin valmisversio, rotta, kana
- 2005 Koira, simpanssi, ensimmäinen GWAS, HapMap 1
- 2006 ensimmäinen koko genomiin perustuva testi, merisiili, mehiläinen, NCBI genotyyppi ja fenotyyppi tietokanta
- 2007 Han kiinalainen genomi, ensimmäinen henkilökohtainen genomi (Venter), ENCODE pilotti, Wellcome trust case. vs control tutkimus, ihmisen geneettinen variaatio, reesusapina
- 2008 Glioblastooman geneettinen tutkimus, syöpägenomi AML), nokkaeläin, Yoruba (Afrikka) genomi, NGS genomi (Watson),



- 2009 Korealaisen genomi, ihmisen metylaatiokartta, naudan genomi, MGC mammalian gene collection,
- 2010 Nuffield council Bioetiikka julkaisu henkilökohtaisesta terveydenhuollosta, Miller syndrome geeni paikannettu eksonisekvensointia käyttäen, 1000 genomia projektin pilottiosuus valmis, Neadenrthal genomi, modENCODE julkaisu, yli 1000 hiiren KO kantaa, UK Biobank 500 000 osallistujaa, 5000 GWAS tutkimusta julkaistu, Etelä-Afrikkalaisia ihimsen genomeita

- Nature , 2011:470;204

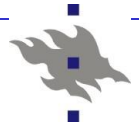


nature Vol 437|15 September 2005|doi:10.1038/nature03959

ARTICLES

Genome sequencing in microfabricated high-density picolitre reactors

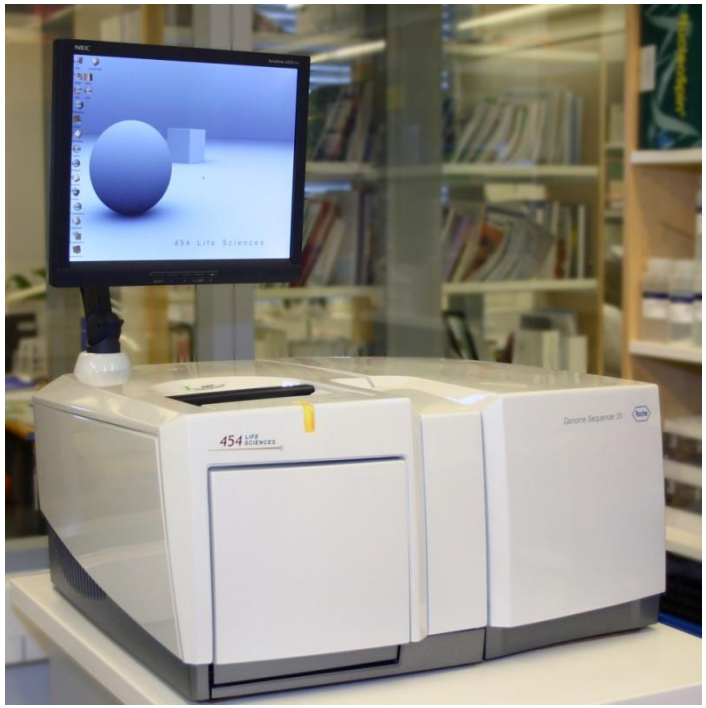
Marcel Margulies[1]*, Michael Egholm[1]*, William E. Altman[1], Said Attiya[1], Joel S. Bader[1], Lisa A. Bemben[1], Jan Berka[1], Michael S. Braverman[1], Yi-Ju Chen[1], Zhoutao Chen[1], Scott B. Dewell[1], Lei Du[1], Joseph M. Fierro[1], Xavier V. Gomes[1], Brian C. Godwin[1], Wen He[1], Scott Helgesen[1], Chun He Ho[1], Gerard P. Irzyk[1], Szilveszter C. Jando[1], Maria L. I. Alenquer[1], Thomas P. Jarvie[1], Kshama B. Jirage[1], Jong-Bum Kim[1], James R. Knight[1], Janna R. Lanza[1], John H. Leamon[1], Steven M. Lefkowitz[1], Ming Lei[1], Jing Li[1], Kenton L. Lohman[1], Hong Lu[1], Vinod B. Makhijani[1], Keith E. McDade[1], Michael P. McKenna[1], Eugene W. Myers[2], Elizabeth Nickerson[1], John R. Nobile[1], Ramona Plant[1], Bernard P. Puc[1], Michael T. Ronan[1], George T. Roth[1], Gary J. Sarkis[1], Jan Fredrik Simons[1], John W. Simpson[1], Maithreyan Srinivasan[1], Karrie R. Tartaro[1], Alexander Tomasz[3], Kari A. Vogt[1], Greg A. Volkmer[1], Shally H. Wang[1], Yong Wang[1], Michael P. Weiner[4], Pengguang Yu[1], Richard F. Begley[1] & Jonathan M. Rothberg[1]

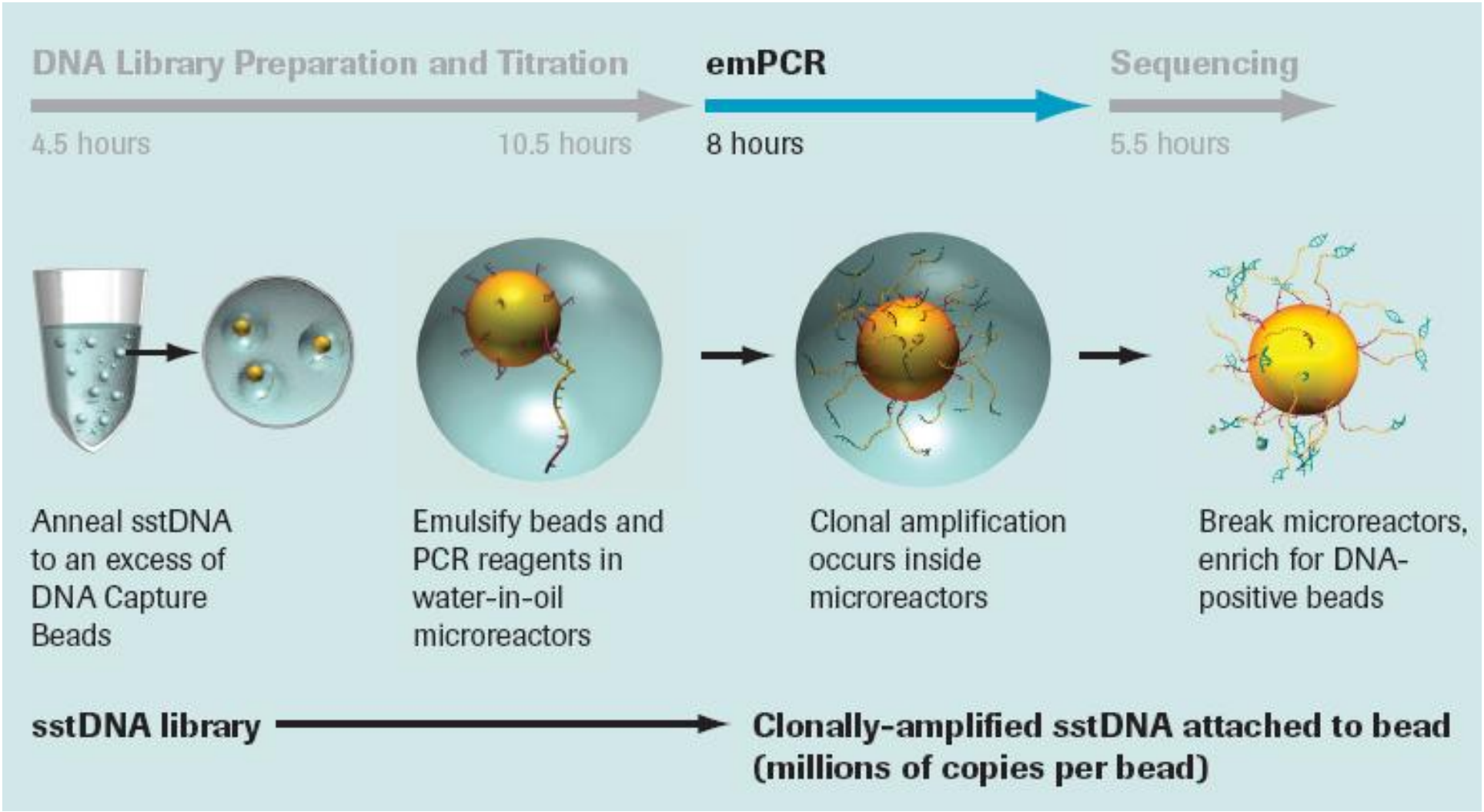


Genome Sequencer

(http://www.454.com/,http://www.roche.com)
(Margulies, *et al*. Nature 2005,437, 376)

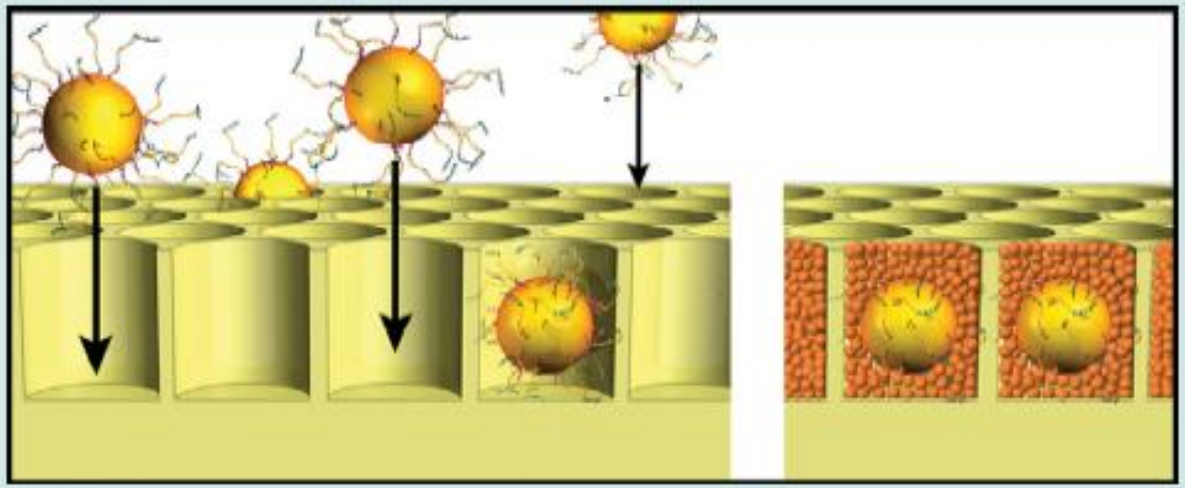
- Genome Sequencer GS20;FLX; Titanium
  - 454 Life Science, Roche
- Parallel Sequencing
  - Shotgun sequencing
    - No plasmid libraries
    - Linkers ligated to fragments
    - Emulsion PCR
    - Picotiter plate, 3 400 000 wells
  - Pyrosequencing
    - Detection with sensitive CCD camera
    - Read length 400+ bp
    - Run time ca. 10h
    - Raw sequence ca. 400-600 Mb/run




DNA Library Preparation and Titration
4.5 hours
10.5 hours
emPCR
8 hours
Sequencing
5.5 hours
Anneal sstDNA to an excess of DNA Capture Beads
Emulsify beads and PCR reagents in water-in-oil microreactors
Clonal amplification occurs inside microreactors
Break microreactors, enrich for DNA-positive beads
sstDNA library
Clonally-amplified sstDNA attached to bead (millions of copies per bead)

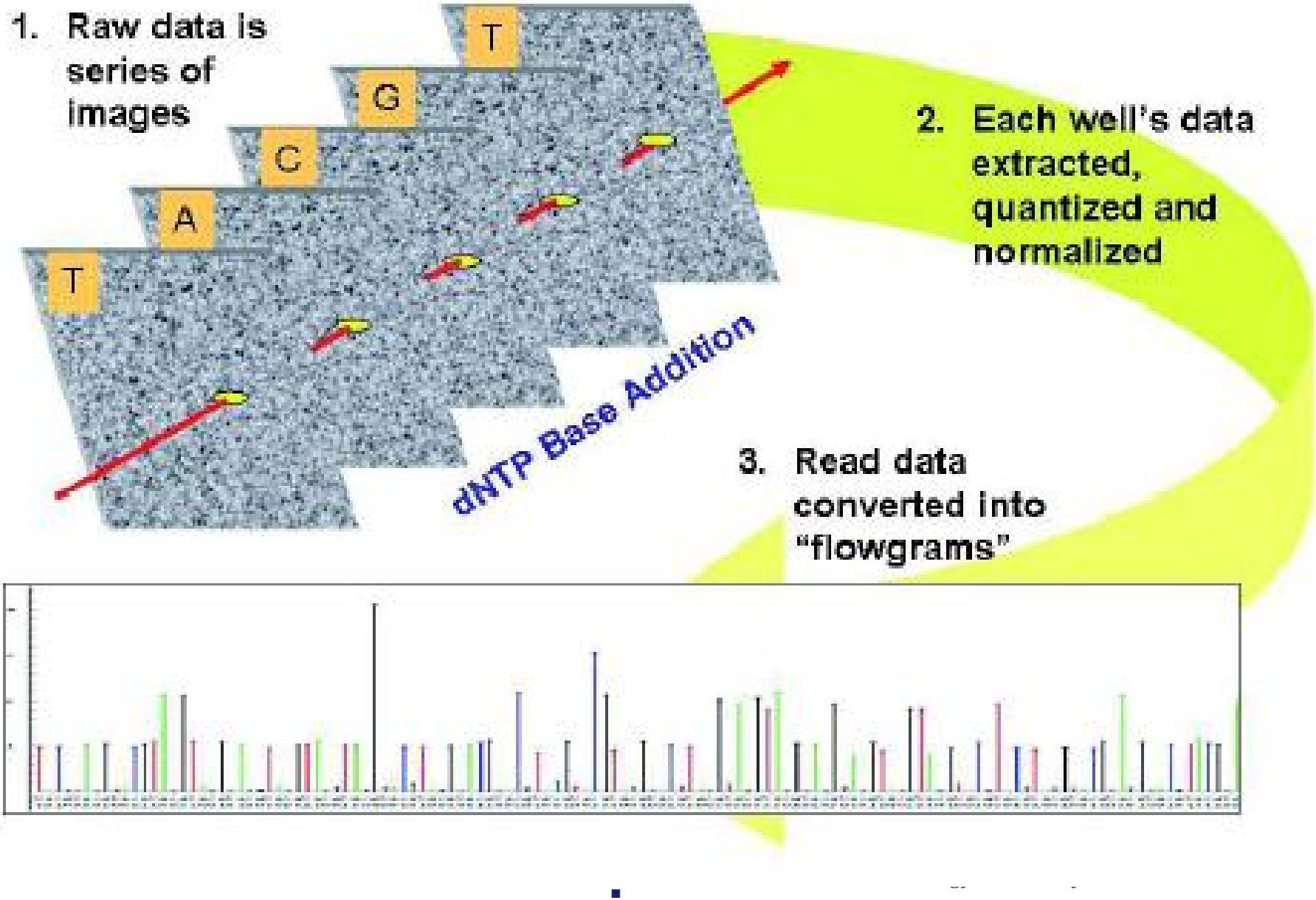
DNA Library Preparation and Titration → emPCR → **Sequencing**

4.5 hours — 10.5 hours — 8 hours — 5.5 hours



- Well diameter: average of 44 µm
- A single clonally amplified sstDNA bead is deposited per well
- 200,000 reads obtained in parallel on large-format PicoTiterPlate device

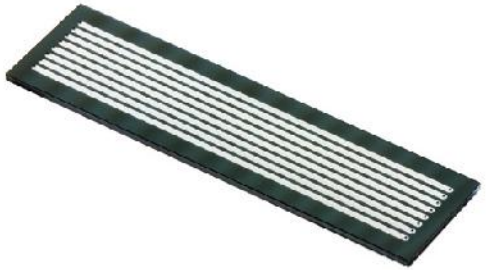
Amplified sstDNA library beads → **Quality reads**


1. Raw data is series of images
T
G
C
A
T
dNTP Base Addition
2. Each well's data extracted, quantized and normalized
3. Read data converted into "flowgrams"

Illumina/Solexa Genome Analyzer (www.illumina.com ; Bentley,DR Curr Opinion Genet Dev 2006, 16,545-552)

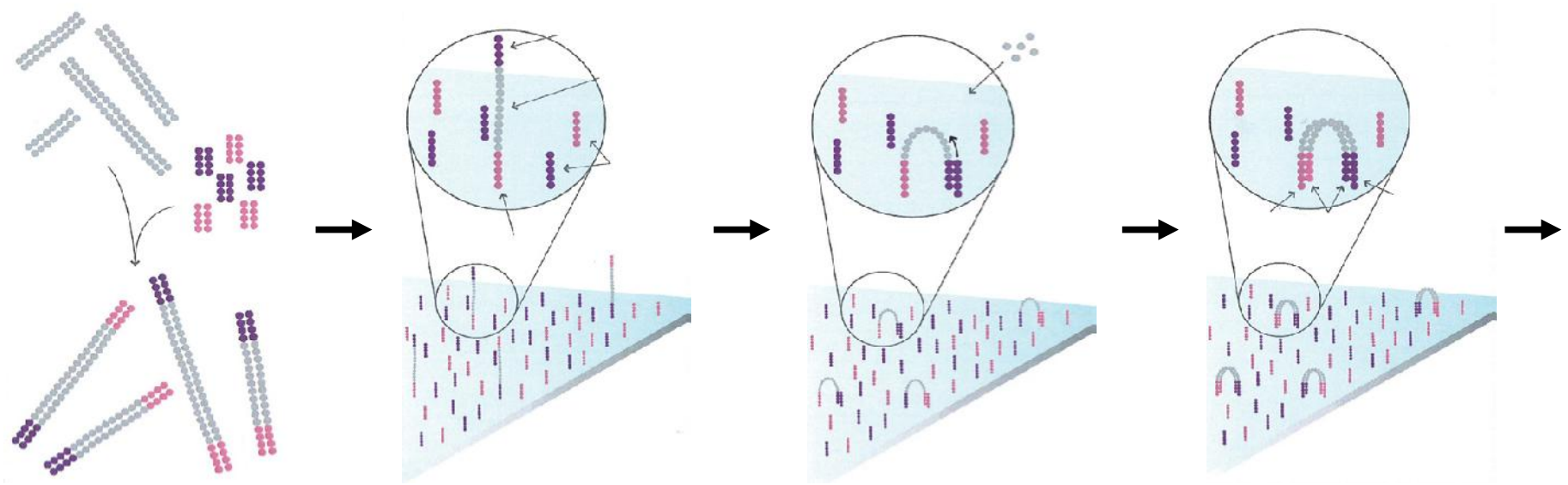
- Clonal Single Molecule Array technology
  - Sequencing-by-synthesis technology
  - Reversible terminator-based sequencing
    - removable fluorescence
  - Flow cell with > 10 million clusters
  - 1–8 samples / run
  - 3 laser system (660, 635, and 532 nm)
  - Read length 35 - 100 bp, up to 30 Gb / run
    - Run time 3 – 10 days,

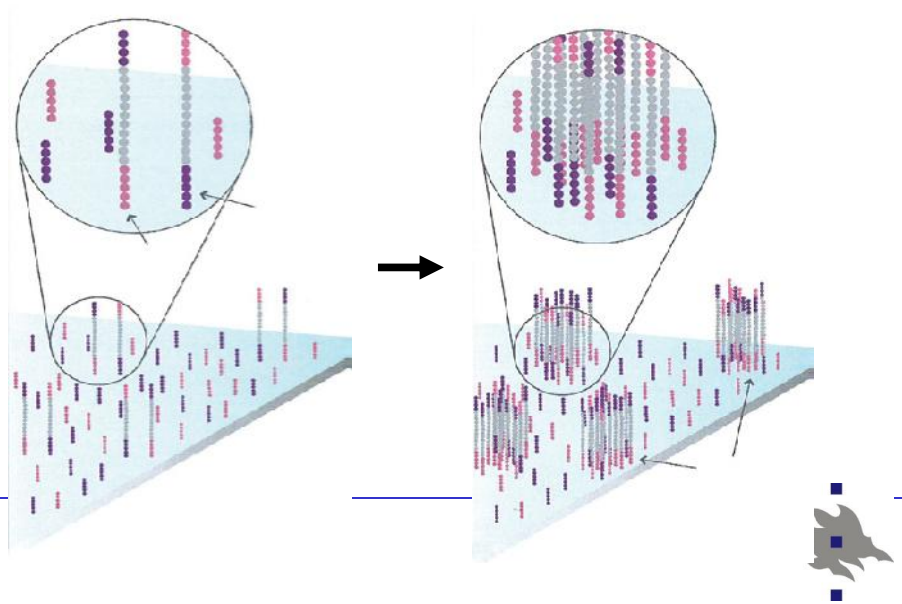
Cluster Station



Flow cell

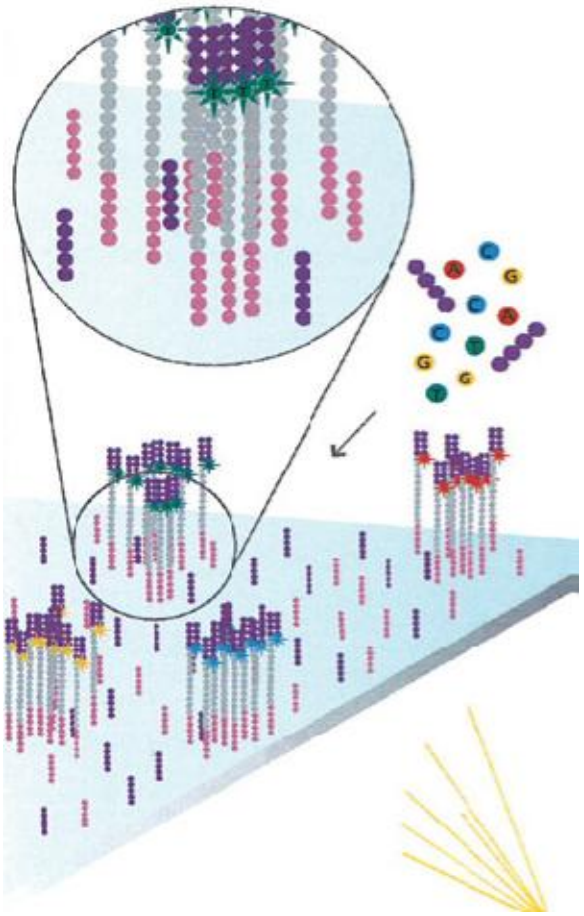
Illumina/Solexa

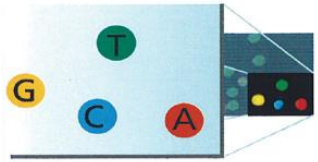




- **Sample preparation**
  - **100ng–1µg**
  - **Attaching to Flow cell**
  - **Bridging**
  - **PCR**
    - **Elongation**
    - **Denaturation**
    - **Clonal amplification**

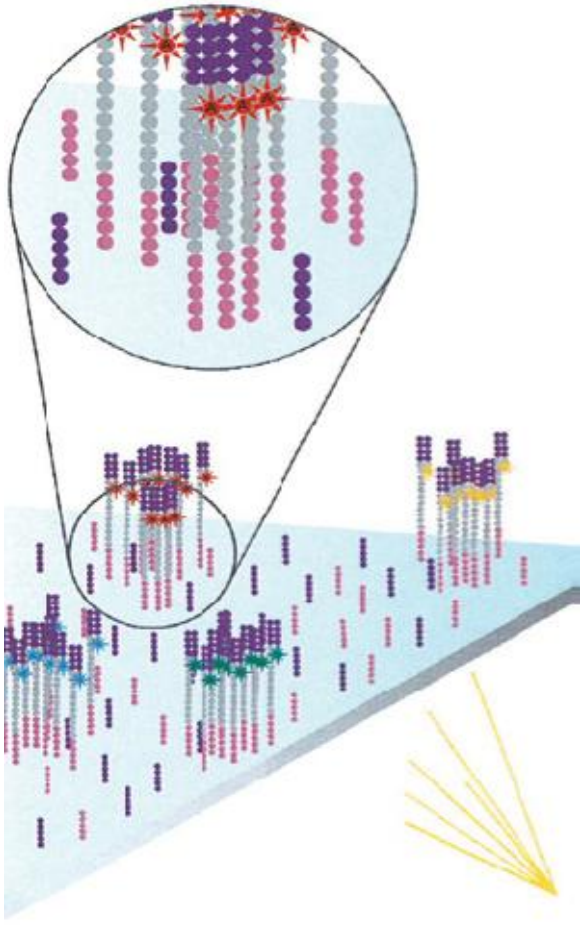
Illumina/Solexa sequencing

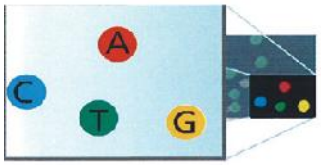




Sequencing

- First bases
- Fluorescent reversible terminators
- Detection with laser and CCD camera





Sequencing

- Second bases detected after removal of label and blocking

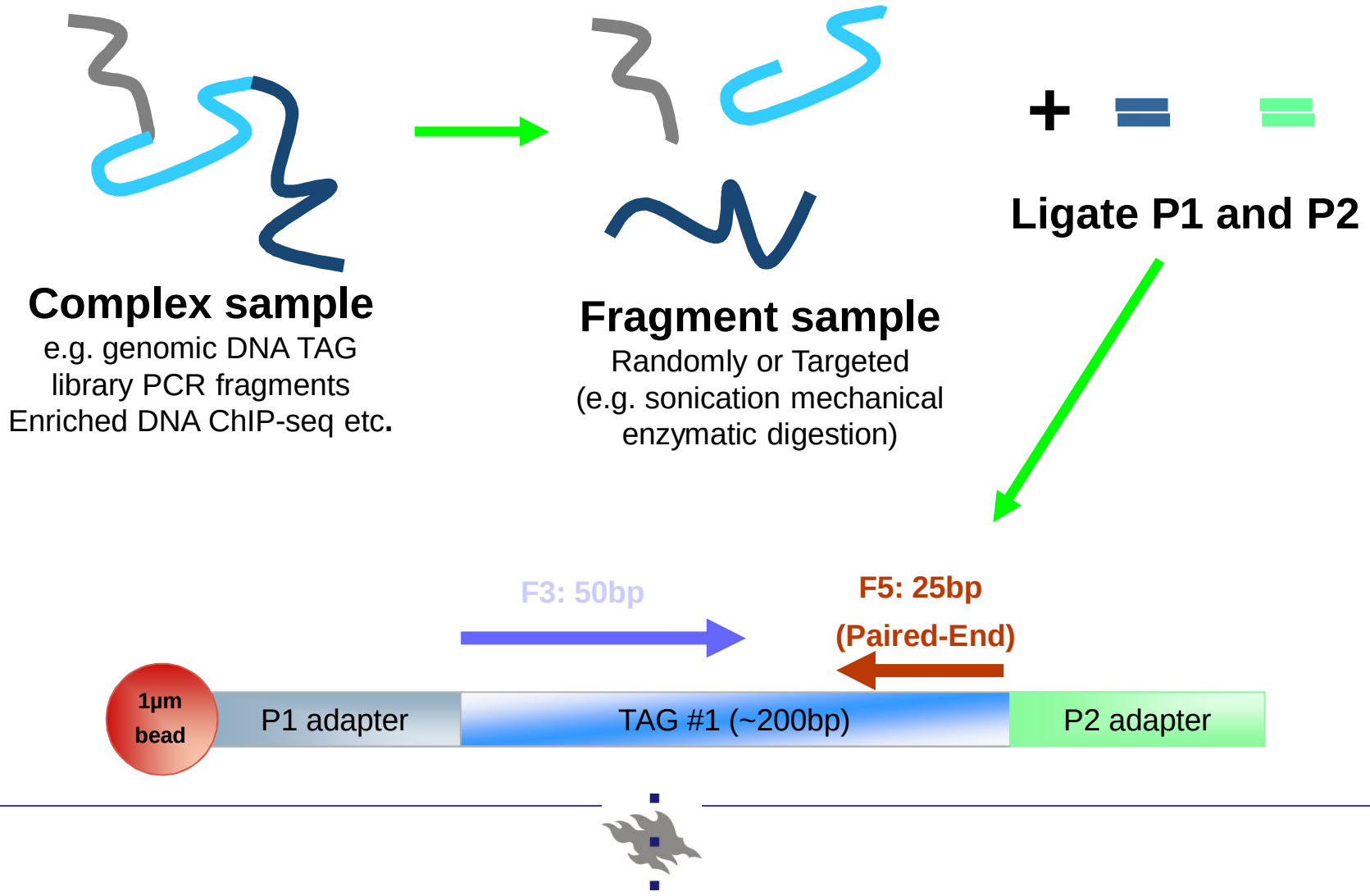
SOLiD 4, Applied Biosystems

(http://www.appliedbiosystems.com)

- Sequencing by Ligation
  - emPCR
  - Attaching to glass slides
  - Labelled probes
    - Four colours
    - 2 base encoding system
  - Repeated ligation steps
  - Detection with 4 Mpixel camera
  - Read lenght 35-50 bp
  - 1-2 slides / run
  - 30 Gb / run/slide
  - 500x10 ^{6} reads/slide
  - Run time 3.5 -14 days

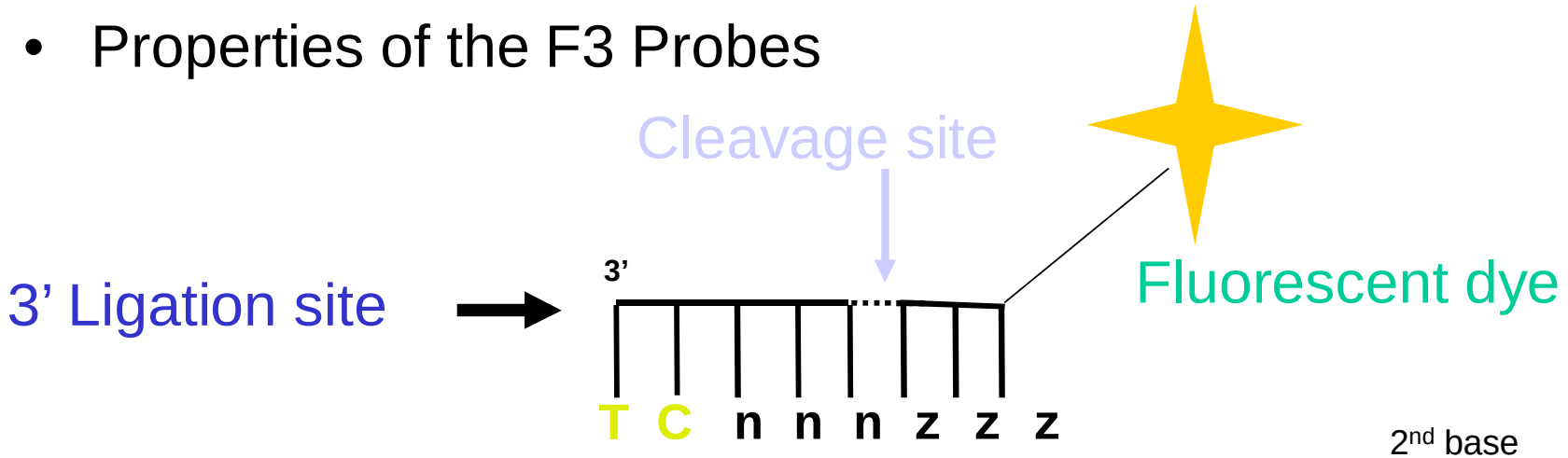
Shendure, J. *et. al.*, Science 2005, 309, 1728-1732

Fragment Library



The F3 SOLiD chemistry

- Properties of the F3 Probes

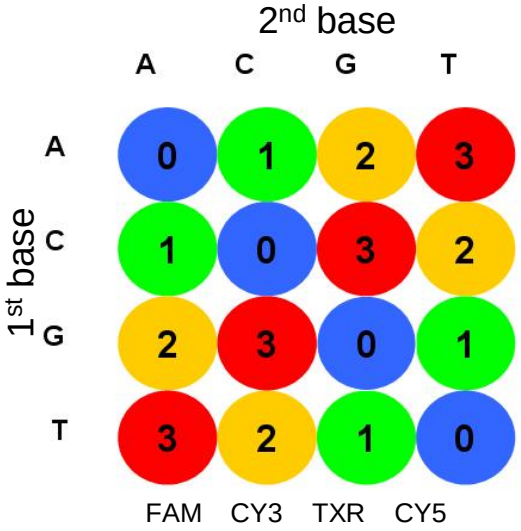


1024 Octamer Probes (4^5)

4 Dyes 4 dinucleotides 256 probes per dye

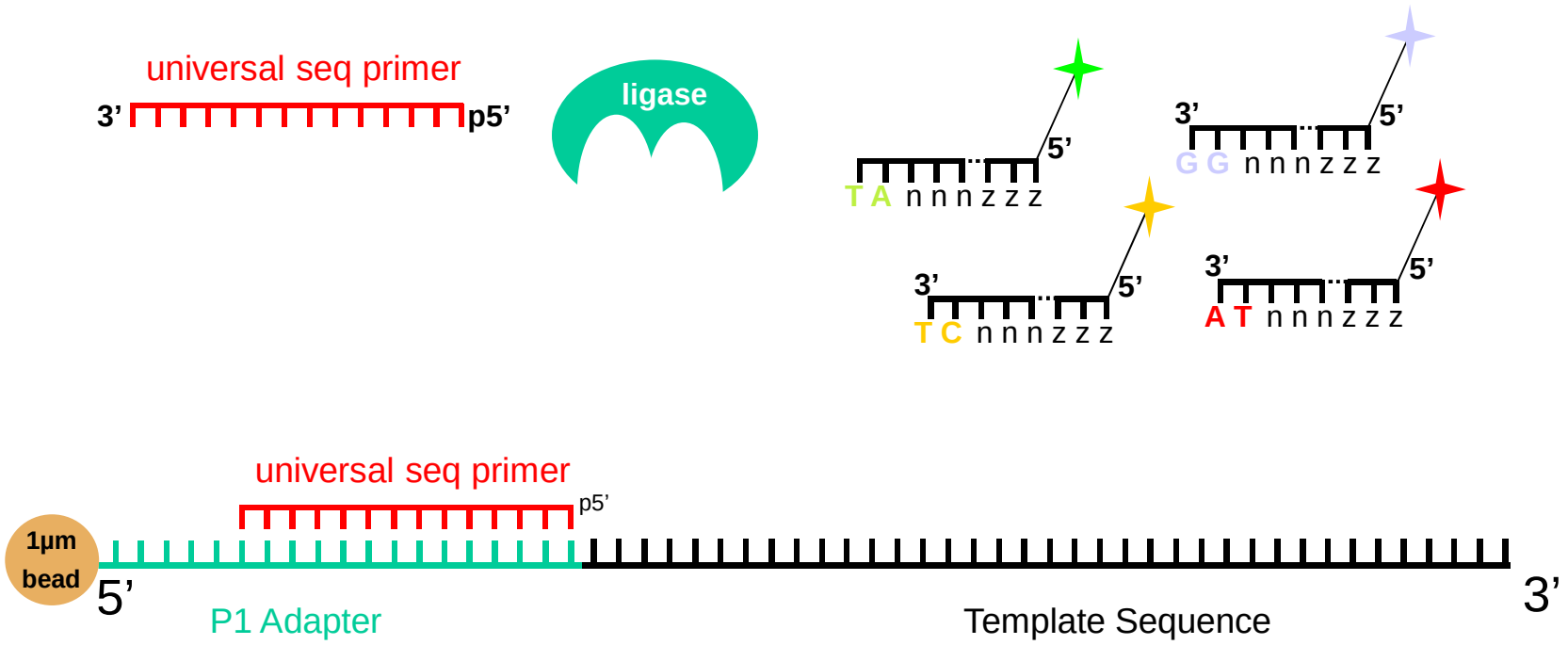
Each dimer is encoded by a color

N= Degenerate bases Z= Universal bases



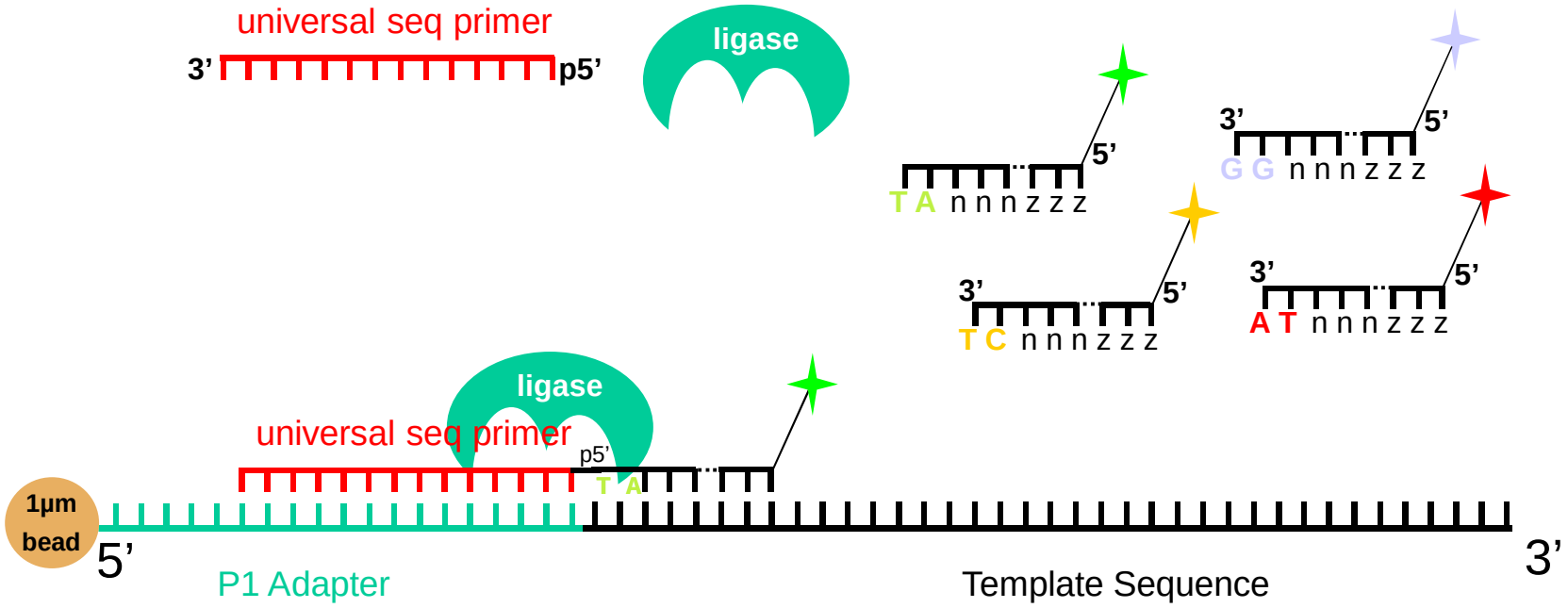
SOLiD Chemistry System 4-color ligation

Step 2: the ligation reaction process



SOLiD Chemistry System 4-color ligation

Step 2: the ligation reaction process

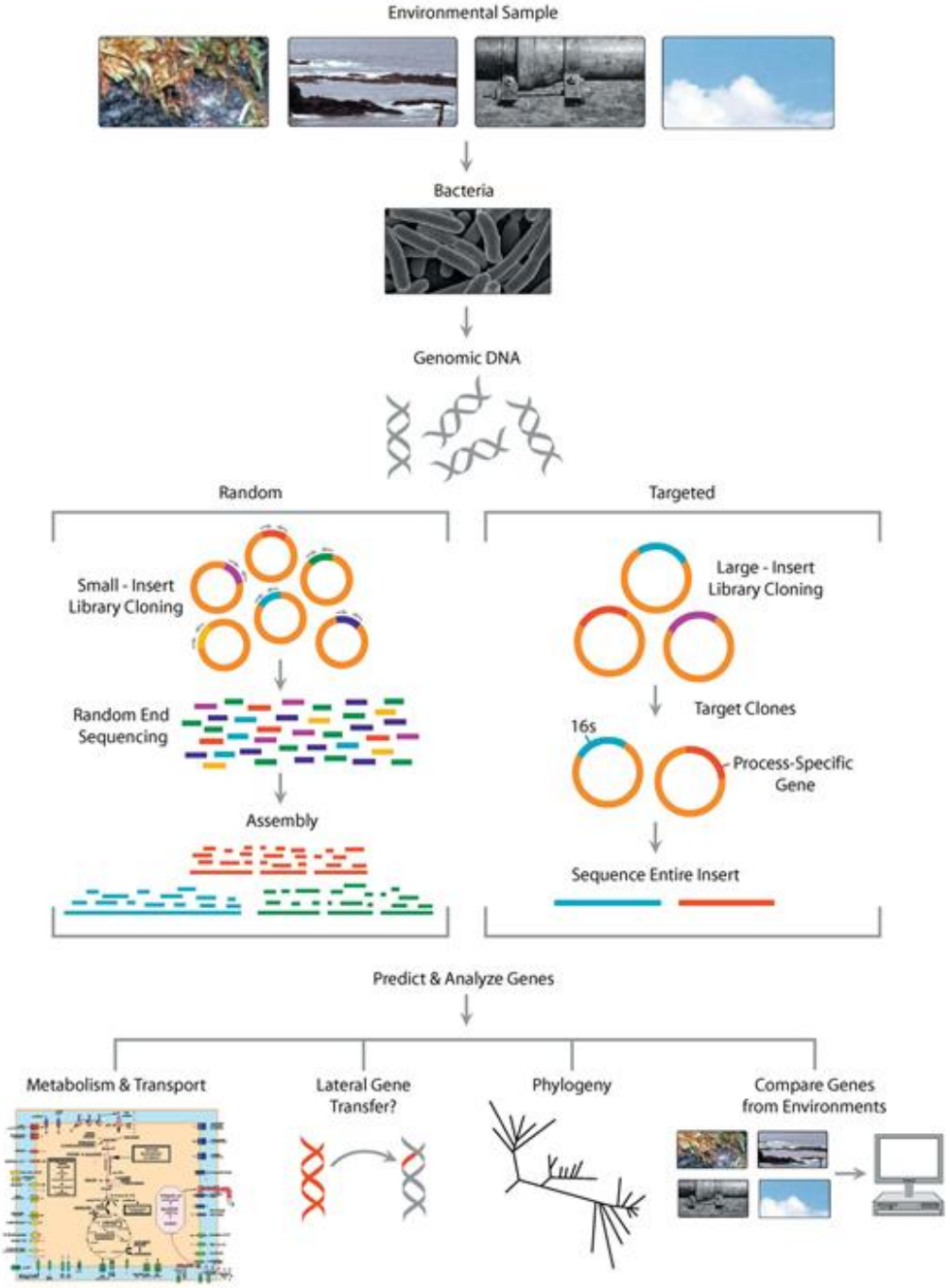


A decade's perspective on DNA sequencing technology

Elaine R. Mardis[1]

etri Auvinen , DNA Sequencing and Genomics Laboratory,
Institute of Biotechnology, University of Helsinki

Metagenomiikka



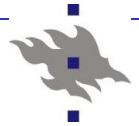
Spatially differing bacterial communities in water columns of the northern Baltic Sea

Kaisa Koskinen[1], Jenni Hultman[1], Lars Paulin[1], Petri Auvinen[1] & Harri Kankaanpää[2]

[1]DNA Sequencing and Genomics Laboratory, Institute of Biotechnology, University of Helsinki, Helsinki, Finland; and [2]Finnish Environment Institute, Marine Research Centre, Helsinki, Finland

Aims of the study

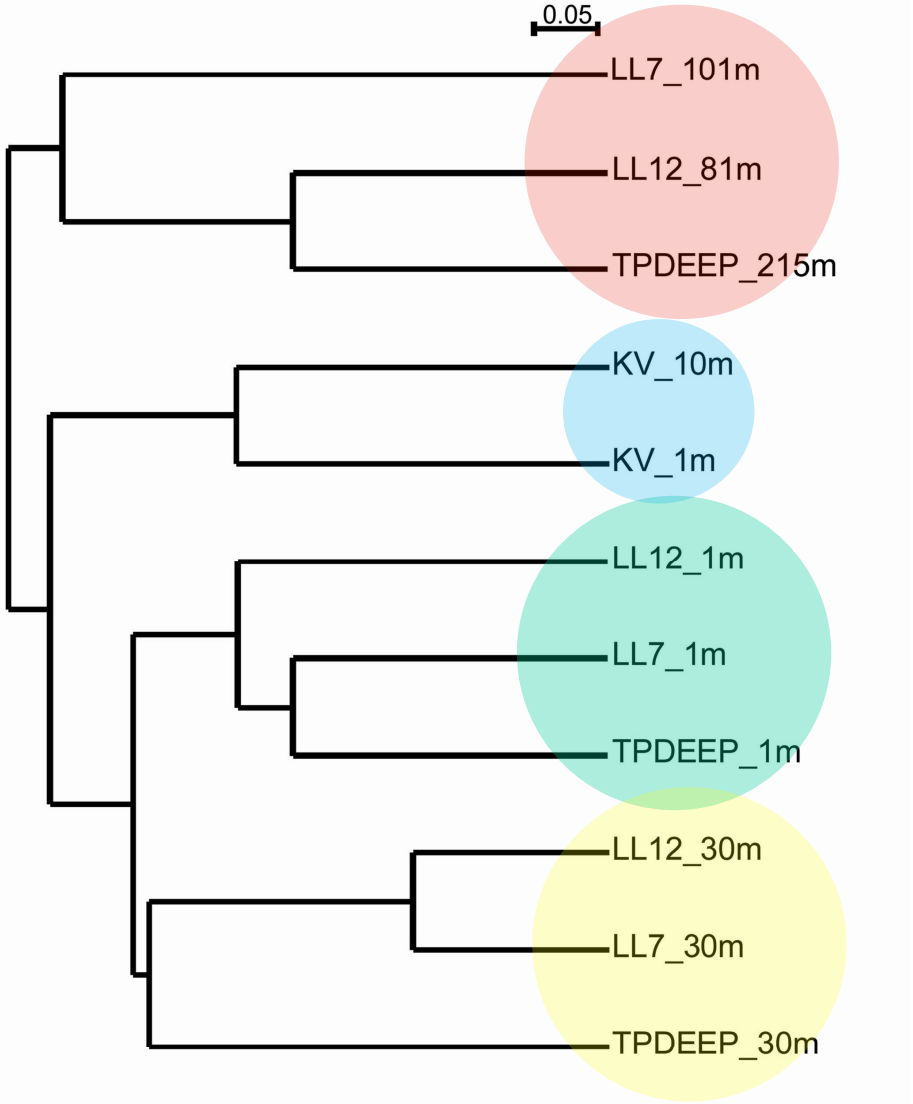
- to characterise the bacterial communities of the northern Baltic Sea by 454 sequencing
- to determine the horizontal and vertical distribution and diversity of the bacterial communities
- to determine how the bacterial community structure is regulated by environmental factors such as depth (pressure), oxygen concentration, pH, salinity, temperature, and nutrient concentrations



Patchiness

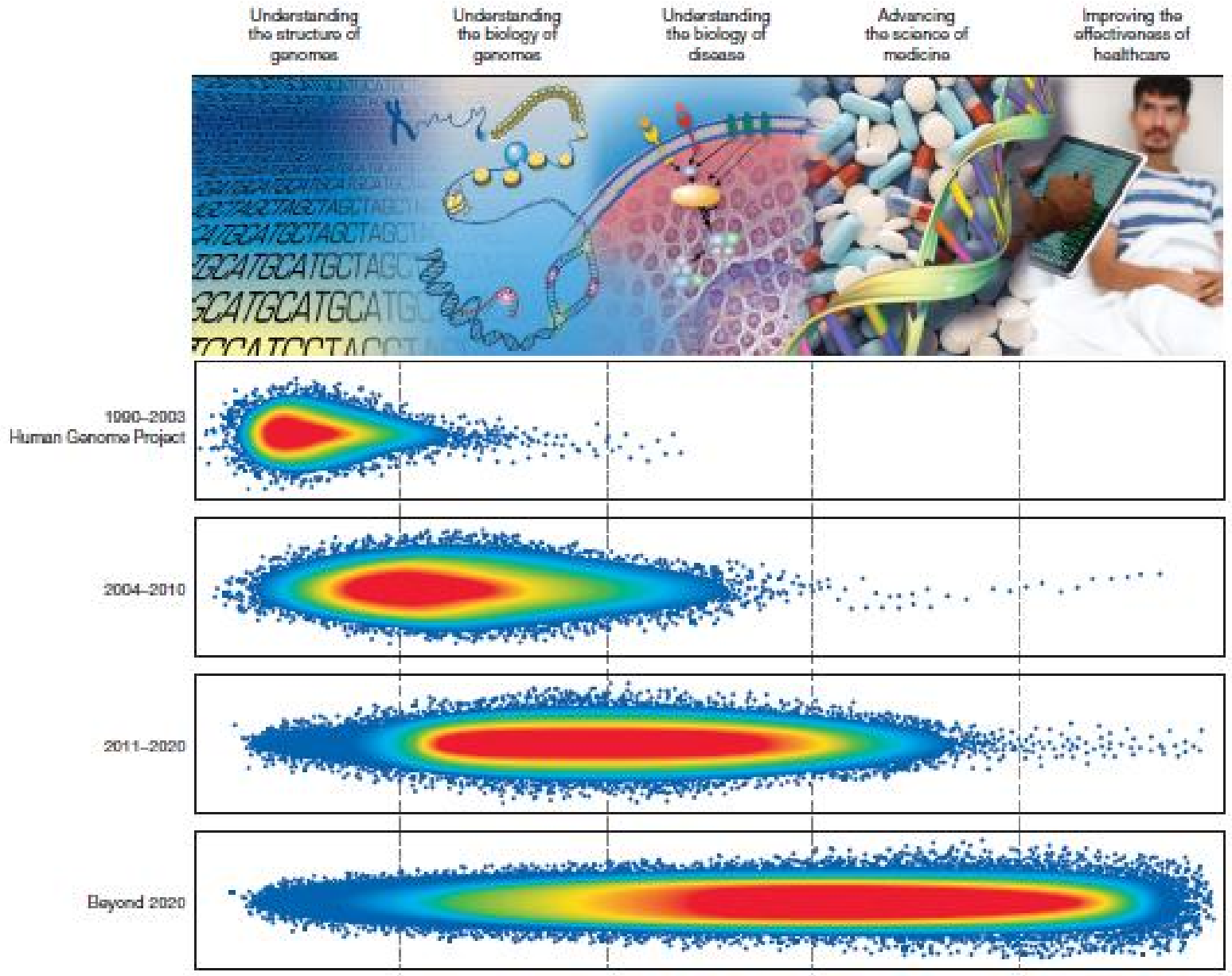
- **All the samples were different: from 1390 different OTUs only 11 were present in all samples → less than 1% of observed OTUs present in all samples → substantial patchiness in distribution of the bacterial communities in the study area**

- **Samples taken from same depth but different sites were more similar than samples from different depth but same location**



Charting a course for genomic medicine from base pairs to bedside

Eric D. Green[1], Mark S. Guyer[1] & National Human Genome Research Institute*



aboratory,
lsinki

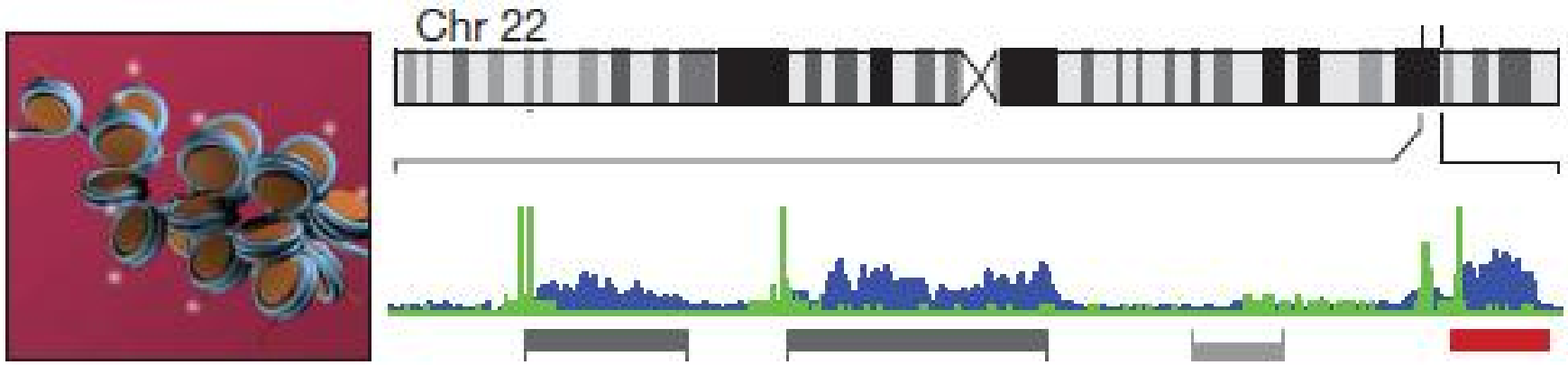


Figure 2 | Chromatin state maps. The genomic sites of chromatin modifications or protein binding can be mapped, using chromatin immunoprecipitation (ChIP) and massively parallel sequencing. The figure highlights chromatin marks associated with the active promoters (green) and actively transcribed regions (blue), in a region on chromosome 22. The four features shown correspond to two active protein-coding (dark grey), one inactive protein-coding (light grey) and one long intergenic non-coding RNA (maroon). Image courtesy of B. Wong (ClearScience).

Initial impact of the sequencing of the human genome

Eric S. Lander[1]

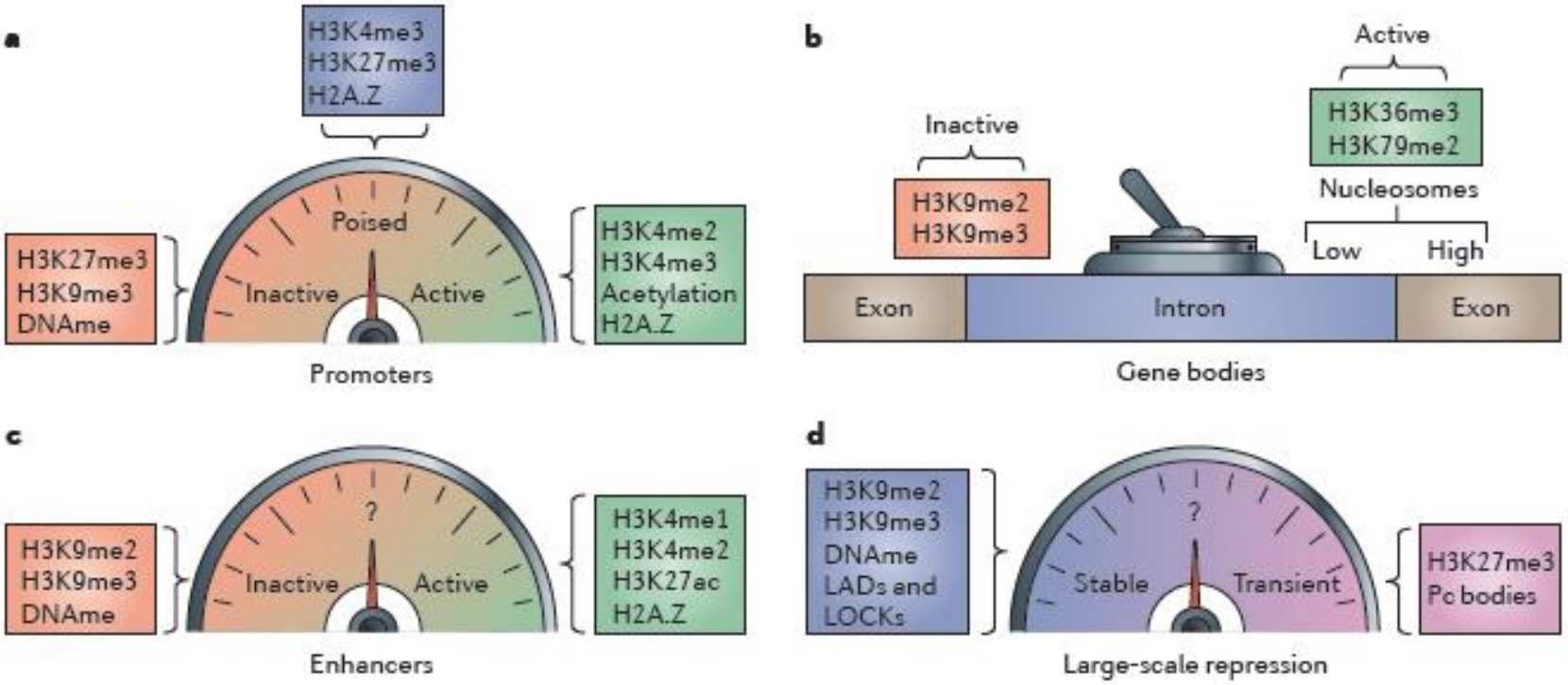
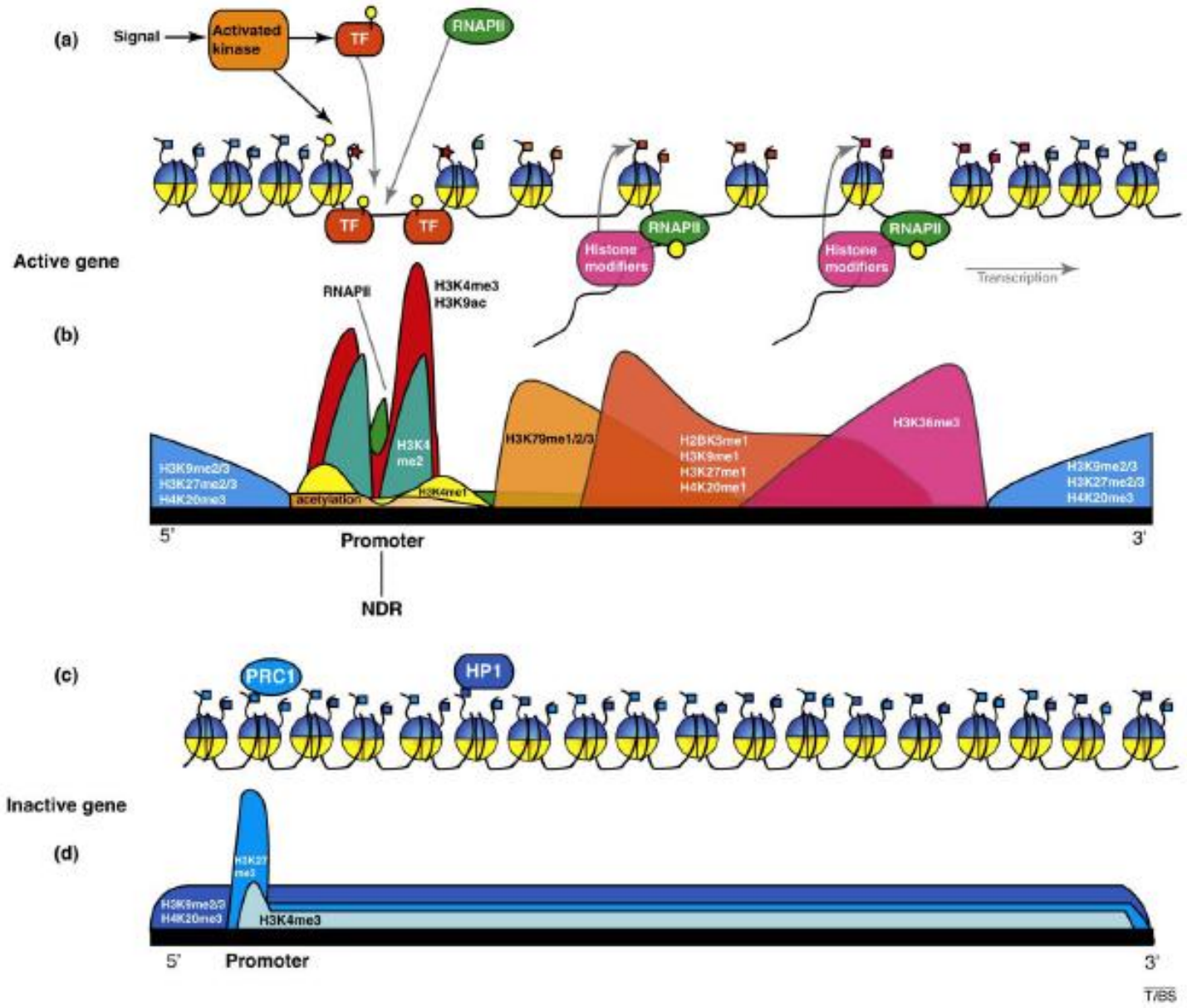


Figure 4 | **'Dashboard' of histone modifications for fine-tuning genomic elements.** In addition to enabling

Charting histone modifications and the functional organization of mammalian genomes

Vicky W. Zhou *†§||¶, *Alon Goren* *†§¶ *and Bradley E. Bernstein* *†§



Fast signals and slow marks: the dynamics of histone modifications

Teresa K. Barth and Axel Imhof

Petri Auvinen , DNA Sequencing and Genomics Laboratory,
Institute of Biotechnology, University of Helsinki

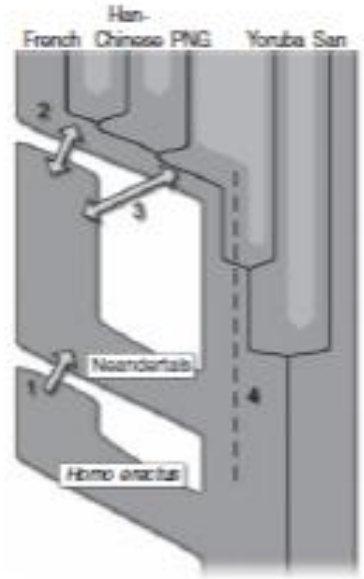


Fig. 6. Four possible scenarios of genetic mixture involving Neandertals. Scenario 1 represents gene flow into Neandertal from other archaic hominins, here collectively referred to as *Homo erectus*. This would manifest itself as segments of the Neandertal genome with unexpectedly high divergence from present-day humans. Scenario 2 represents gene flow between late Neandertals and early modern humans in Europe and/or western Asia. We see no evidence of this because Neandertals are equally distantly related to all non-Africans. However, such gene flow may have taken place without leaving traces in the present-day gene pool. Scenario 3 represents gene flow between Neandertals and the ancestors of all non-Africans. This is the most parsimonious explanation of our observation. Although we detect gene flow only from Neandertals into modern humans, gene flow in the reverse direction may also have occurred. Scenario 4 represents old substructure in Africa that persisted from the origin of Neandertals until the ancestors of non-Africans left Africa. This scenario is also compatible with the current data.

A Draft Sequence of the Neandertal Genome

Richard E. Green,[1]*†‡ Johannes Krause,[1]†§ Adrian W. Briggs,[1]†§ Tomislav Maricic,[1]†§ Udo Stenzel,[1]†§ Martin Kircher,[1]†§ Nick Patterson,[2]†§ Heng Li,[2]† Weiwei Zhai,[3]†‖ Markus Hsi-Yang Fritz,[4]† Nancy F. Hansen,[5]† Eric Y. Durand,[3]† Anna-Sapfo Malaspinas,[3]† Jeffrey D. Jensen,[6]† Tomas Marques-Bonet,[7,13]† Can Alkan,[7]† Kay Prüfer,[1]† Matthias Meyer,[1]† Hernán A. Burbano,[1]† Jeffrey M. Good,[1,8]† Rigo Schultz,[1] Ayinuer Aximu-Petri,[1] Anne Butthof,[1] Barbara Höber,[1] Barbara Höffner,[1] Madlen Siegemund,[1] Antje Weihmann,[1] Chad Nusbaum,[2] Eric S. Lander,[2] Carsten Russ,[2] Nathaniel Novod,[2] Jason Affourtit,[9] Michael Egholm,[9] Christine Verna,[21] Pavao Rudan,[10] Dejana Brajkovic,[11] Željko Kucan,[10] Ivan Gušic,[10] Vladimir B. Doronichev,[12] Liubov V. Golovanova,[12] Carles Lalueza-Fox,[13] Marco de la Rasilla,[14] Javier Fortea,[14]¶ Antonio Rosas,[15] Ralf W. Schmitz,[16,17] Philip L. F. Johnson,[18]† Evan E. Eichler,[7]† Daniel Falush,[19]† Ewan Birney,[4]† James C. Mullikin,[5]† Montgomery Slatkin,[3]† Rasmus Nielsen,[3]† Janet Kelso,[1]† Michael Lachmann,[1]† David Reich,[2,20]*† Svante Pääbo[1]*†

Neandertals, the closest evolutionary relatives of present-day humans, lived in large parts of Europe and western Asia before disappearing 30,000 years ago. We present a draft sequence of the Neandertal genome composed of more than 4 billion nucleotides from three individuals. Comparisons of the Neandertal genome to the genomes of five present-day humans from different parts of the world identify a number of genomic regions that may have been affected by positive selection in ancestral modern humans, including genes involved in metabolism and in cognitive and skeletal development. We show that Neandertals shared more genetic variants with present-day humans in Eurasia than with present-day humans in sub-Saharan Africa, suggesting that gene flow from Neandertals into the ancestors of non-Africans occurred before the divergence of Eurasian groups from each other.

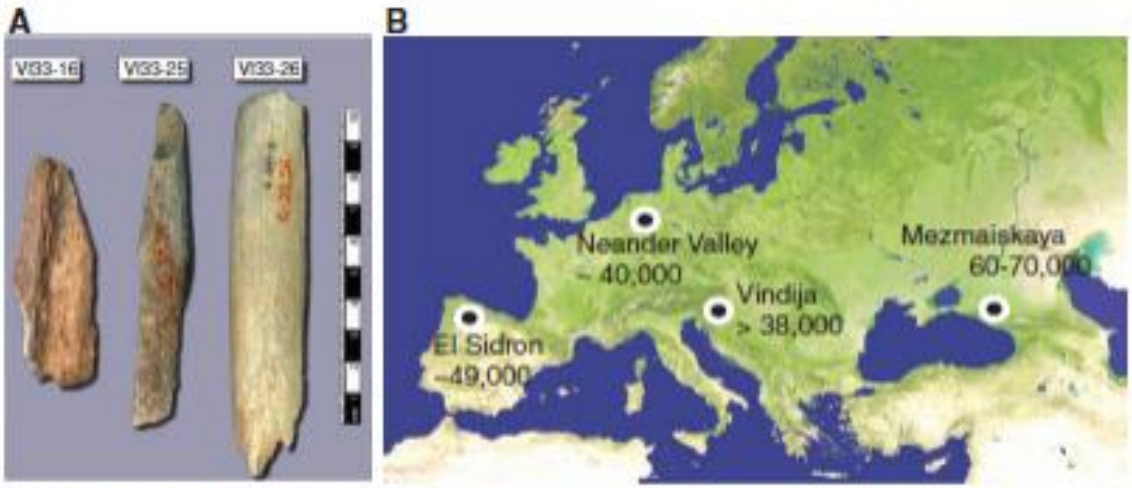


Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

Genetic history of an archaic hominin group from Denisova Cave in Siberia

David Reich[1,2]*, Richard E. Green[3,4]*, Martin Kircher[3]*, Johannes Krause[3,5]*, Nick Patterson[2]*, Eric Y. Durand[6]*, Bence Viola[3,7]*, Adrian W. Briggs[1,3], Udo Stenzel[3], Philip L. F. Johnson[8], Tomislav Maricic[3], Jeffrey M. Good[9], Tomas Marques-Bonet[10,11], Can Alkan[10], Qiaomei Fu[3,12], Swapan Mallick[1,2], Heng Li[2], Matthias Meyer[3], Evan E. Eichler[10], Mark Stoneking[3], Michael Richards[7,13], Sahra Talamo[7], Michael V. Shunkov[14], Anatoli P. Derevianko[14], Jean-Jacques Hublin[7], Janet Kelso[3], Montgomery Slatkin[6] & Svante Pääbo[3]

Using DNA extracted from a finger bone found in Denisova Cave in southern Siberia, we have sequenced the genome of an archaic hominin to about 1.9-fold coverage. This individual is from a group that shares a common origin with Neanderthals. This population was not involved in the putative gene flow from Neanderthals into Eurasians; however, the data suggest that it contributed 4–6% of its genetic material to the genomes of present-day Melanesians. We designate this hominin population 'Denisovans' and suggest that it may have been widespread in Asia during the Late Pleistocene epoch. A tooth found in Denisova Cave carries a mitochondrial genome highly similar to that of the finger bone. This tooth shares no derived morphological features with Neanderthals or modern humans, further indicating that Denisovans have an evolutionary history distinct from Neanderthals and modern humans.

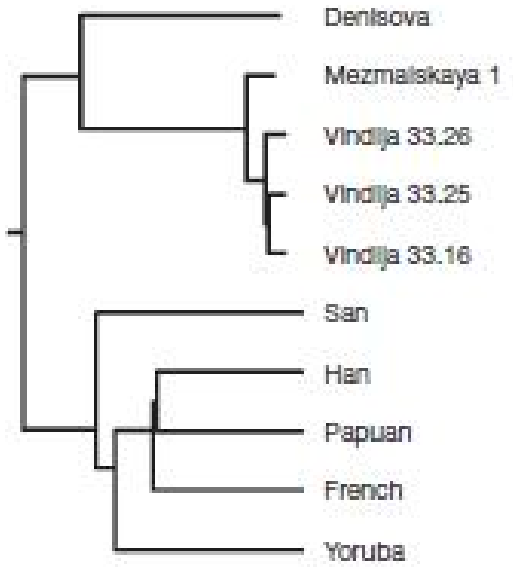


Figure 1 | A neighbour-joining tree based on pairwise autosomal DNA sequence divergences for five ancient and five present-day hominins. Vindija 33.16, Vindija 33.25 and Vindija 33.26 refer to the catalogue numbers of the Neanderthal bones.

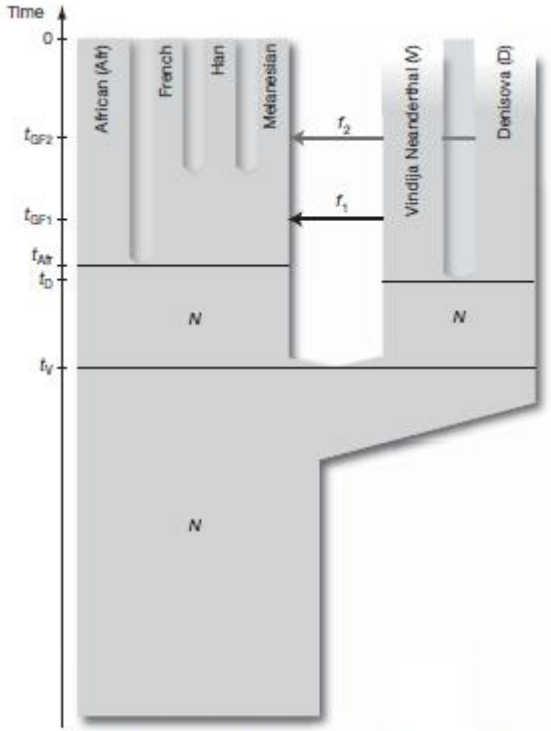


Figure 3 | A model of population history compatible with the data. N denotes effective population size, t denotes time of population separation, f denotes amount of gene flow and t_{GF} denotes time of gene flow.

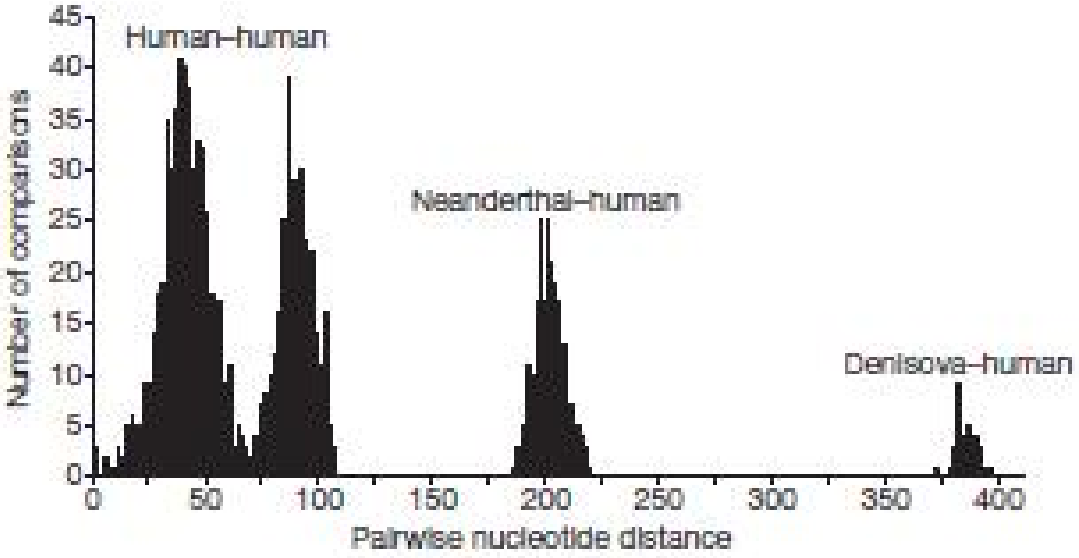


Figure 2 | Distribution of pairwise nucleotide differences. Pairwise nucleotide differences from all pairs of complete mtDNAs from 54 present-day and one Pleistocene modern human, six Neanderthals and the Denisova hominin are shown.

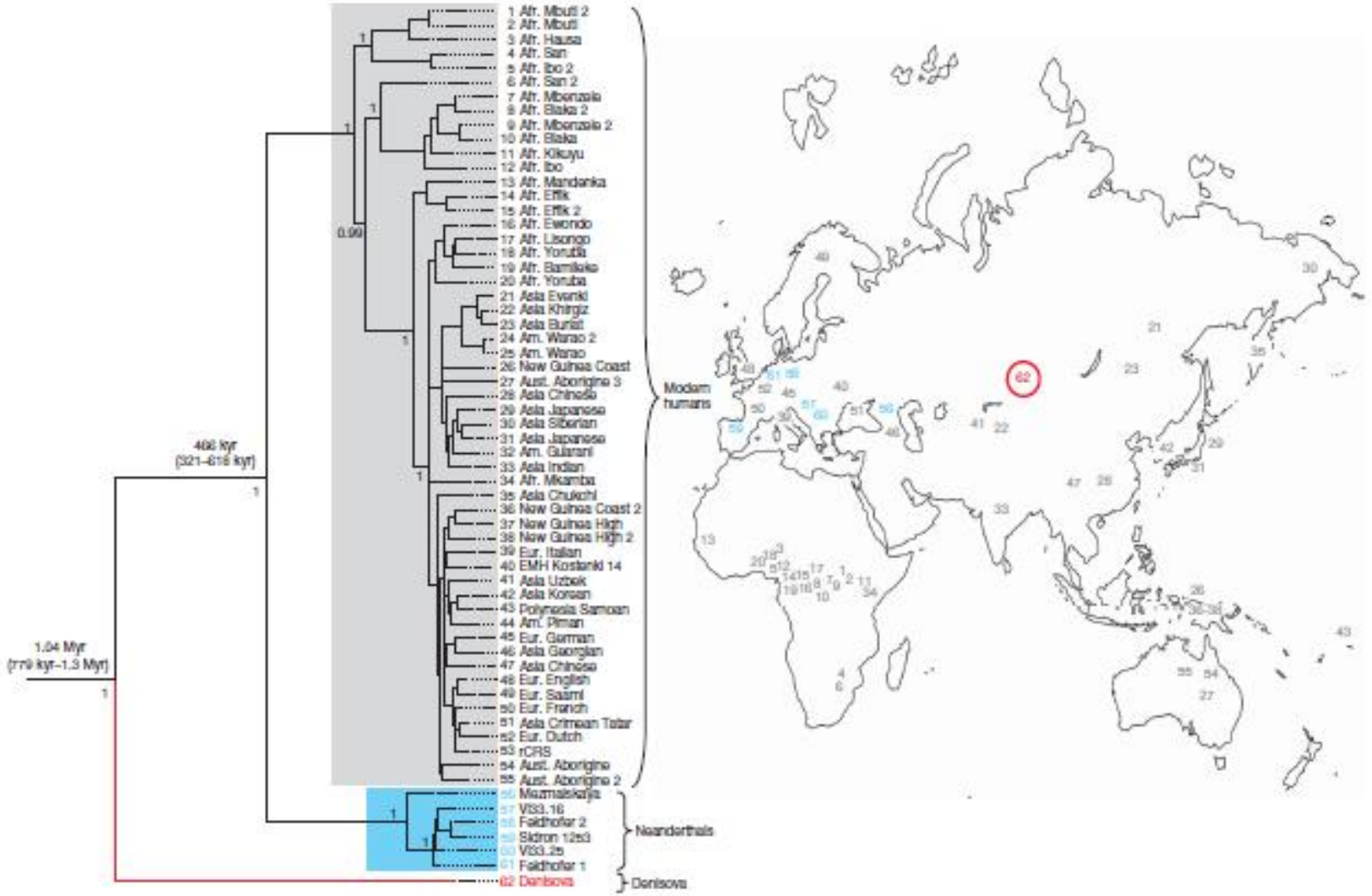
The complete mitochondrial DNA genome of an unknown hominin from southern Siberia

Johannes Krause[1], Qiaomei Fu[1], Jeffrey M. Good[2], Bence Viola[1,3], Michael V. Shunkov[4], Anatoli P. Derevianko[4] & Svante Pääbo[1]

equencing and Genomics Laboratory,
Institute of Biotechnology, University of Helsinki

The complete mitochondrial DNA genome of an unknown hominin from southern Siberia

Johannes Krause[1], Qiaomei Fu[1], Jeffrey M. Good[2], Bence Viola[1,3], Michael V. Shunkov[4], Anatoli P. Derevianko[4] & Svante Pääbo[1]



ory,

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], Kerstin Lindblad-Toh[5], Eric S. Lander[5], James R. Knight[6], Gerard P. Irzyk[6], Karin M. Fredrikson[7], Timothy T. Harkins[7], Sharon Sheridan[7], Tom Pringle[8] & Stephan C. Schuster[1]

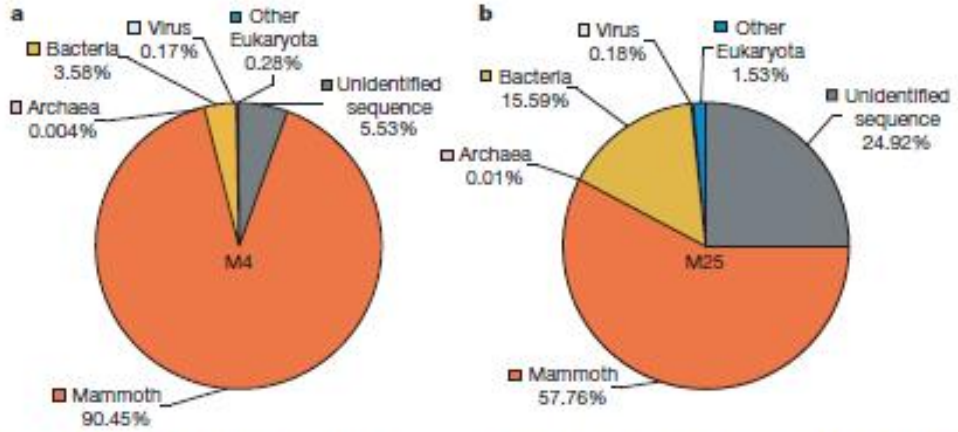


Figure 1 | Species composition of metagenomic DNA extracted from mammoth hair. a, b, Pie charts for the M4 (**a**) and M25 (**b**) data sets show the percentage of sequencing reads assigned to taxa for mammoth, Archaea, Bacteria, virus, as well as the two unspecified categories 'other Eukaryota' and 'unidentified sequence'. The taxon distribution exemplifies the variability of the endogenous DNA content of ancient specimens.

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], Kerstin Lindblad-Toh[5], Eric S. Lander[5], James R. Knight[6], Gerard P. Irzyk[6], Karin M. Fredrikson[7], Timothy T. Harkins[7], Sharon Sheridan[7], Tom Pringle[8] & Stephan C. Schuster[1]

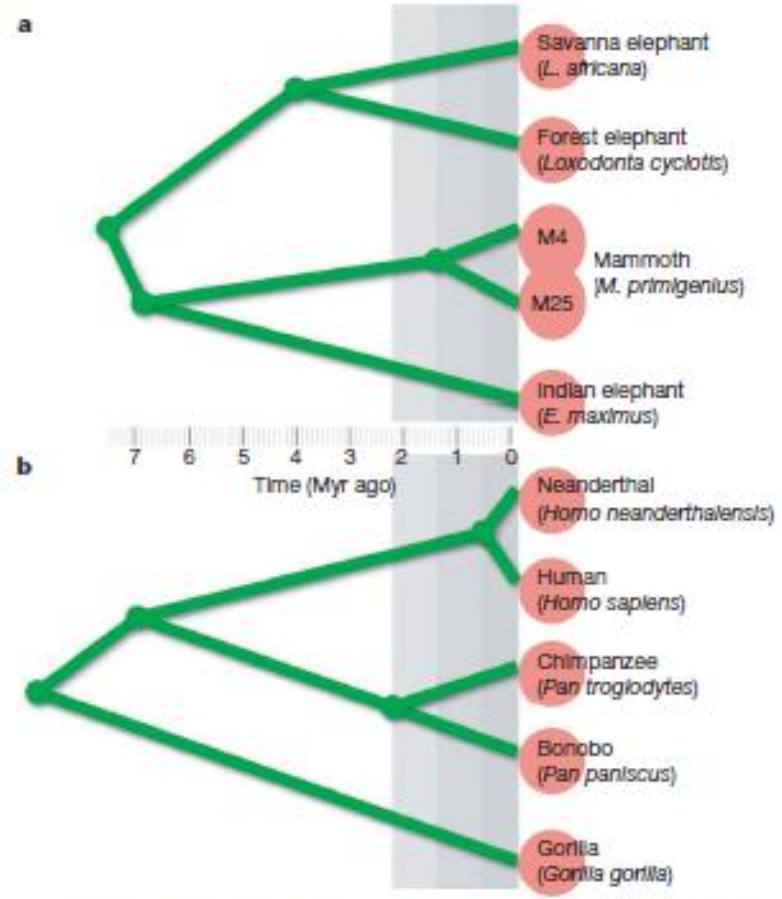
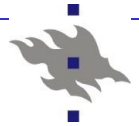


Figure 3 | Comparison of phylogenies. **a**, Elephantids; **b**, hominoids. We show estimated divergence times, that is, times to the common ancestor averaged across autosomes (see Methods). Red circles at the leaves of the phylogenetic tree indicate discernable species. This distinction was not made for the two clades of mammoth (M4 and M25) based on the fossil record (merged red circles).

Mitä voidaan analysoida genominlaajuisesti?

- Genomi- ja metagenomi-sekvensointi
- RNA sekvensointi (mRNA, miRNA, lncRNA etc)
- RNA rakenneanalyysit
- Epigeneettiset muutokset (metylaatio, histonit, ei-histoni proteiinit
- Rakenteellisetmuutokset (SNP)
- Genomien variaatiot (pan genomi)
- Tuman proteiinien ja DNA:n suuremmat organisaatiot
- Evoluutio
- Systeemibiologia



- Kiitokset mielenkiinnosta