Genetics and Genomics

Outline for the second part of this Course

Lectures 1-2:

Introduction to the course // Sequencing technologies // What is a genome? // Genome variation // Genome-wide association studies

Lecture 2-3:

Regulatory networks: genes, gene expression, regulatory elements, transcription factors

Lectures 3-4:

Disease Genomics and therapeutics – the era of precision medicine

Genetics and Genomics

Exam + Python exercise(s)

Written exam: 3h (date so far unknown)

For a good grade:
be able to solve list of questions covering the lectures
(if you pay attention in class and participate in the polls, you'll be more than fine)

For an excellent grade:
A few integrative questions

Note: questions both in English and French
Answering in English is recommended, but French is OK

We will use online polling

A QR code will be displayed for each lecture.

The polling is anonymous, but you will still be requested to fill in your name / email….
(you can fill in whatever….)

Whose music do you like the most?

A. Taylor Swift
B. The Weeknd
C. GIMS
D. Zaz

60

Genomics – a short history

The big four….



James Watson
- DNA
- First NHGRI director
- HGP panel

Walter Gilbert
- Sequencing
- HGP panel member
- Genome Corp
- «one dollar per base»

David Botstein
- HGP panel member
- Genetic linkage map using restriction fragment length polymorphisms

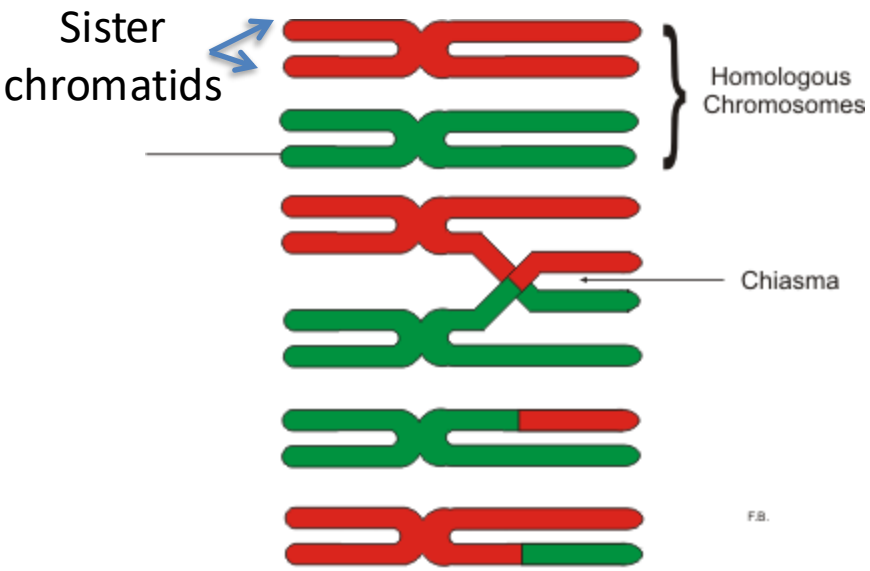
Lee Hood
- Dye termination sequencing (Smith and Hunkapillar)

Botstein and his human genetic linkage map

Q1a

A genetic map shows the linear order of the genes along a chromosome with distance proportional to the frequency of recombination.
(i.e. the closer two genes are, the greater the probability that they will crossover together)

Crossover: rearrangement of genetic information



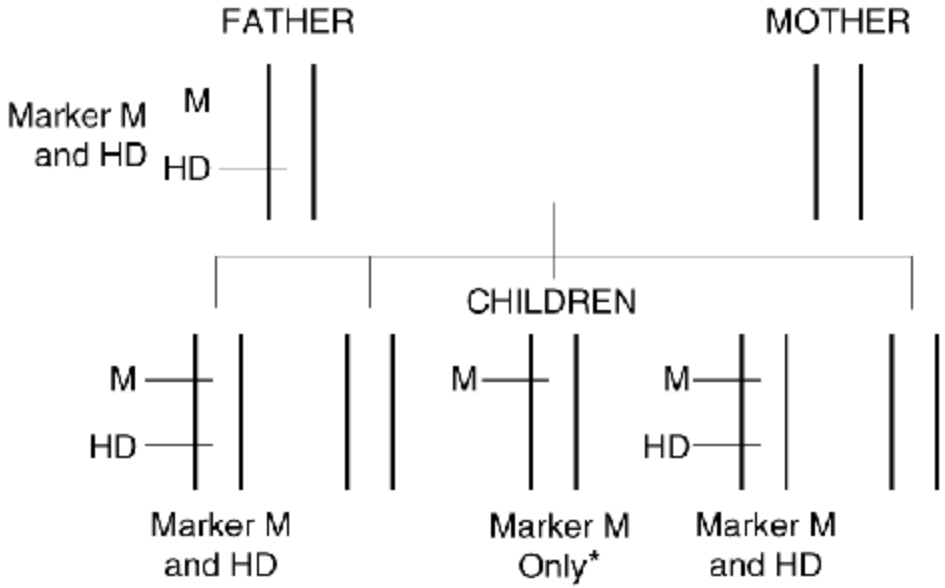
We can use specific genomic signatures (e.g. polymorphisms or microsatellites (see later)) as markers

→ If a specific marker correlates with a specific phenotype or disease, then you know that the marker is linked to a gene and you can "genetically " map this gene

Botstein and his human genetic linkage map

A genetic map shows the linear order of the genes along a chromosome with distance proportional to the frequency of recombination.
(i.e. the closer two genes are, the greater the probability that they will crossover together)

Prof. Nancy Wrexler, driving force behind locating affected gene (Neuropsychologist, Columbia U)



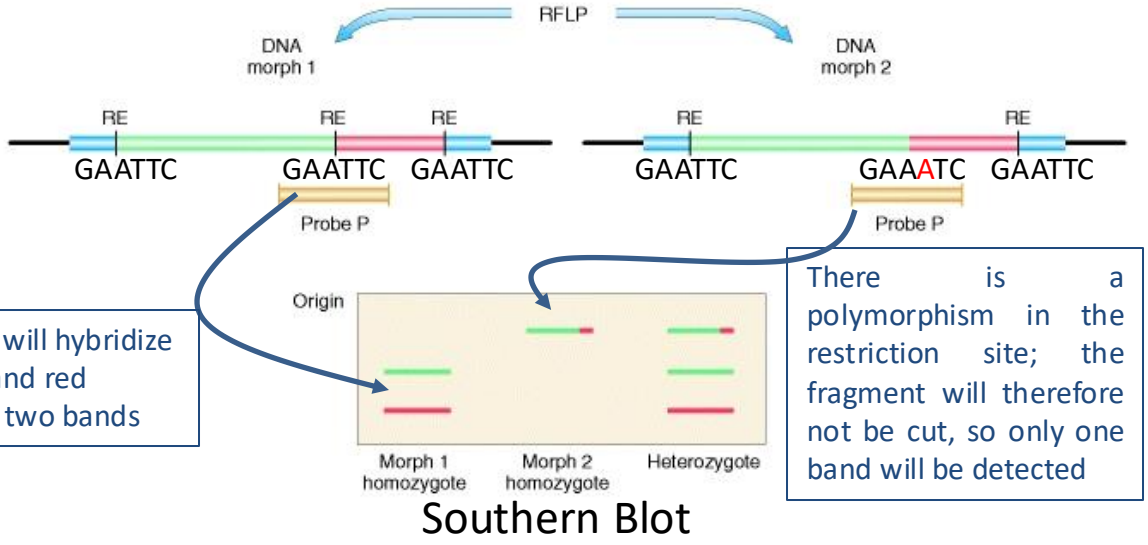
Used to "map" Huntington Disease gene in 1983!

So how do we know whether M and HD are linked in an individual?

Botstein and his human genetic linkage map: RFLP method

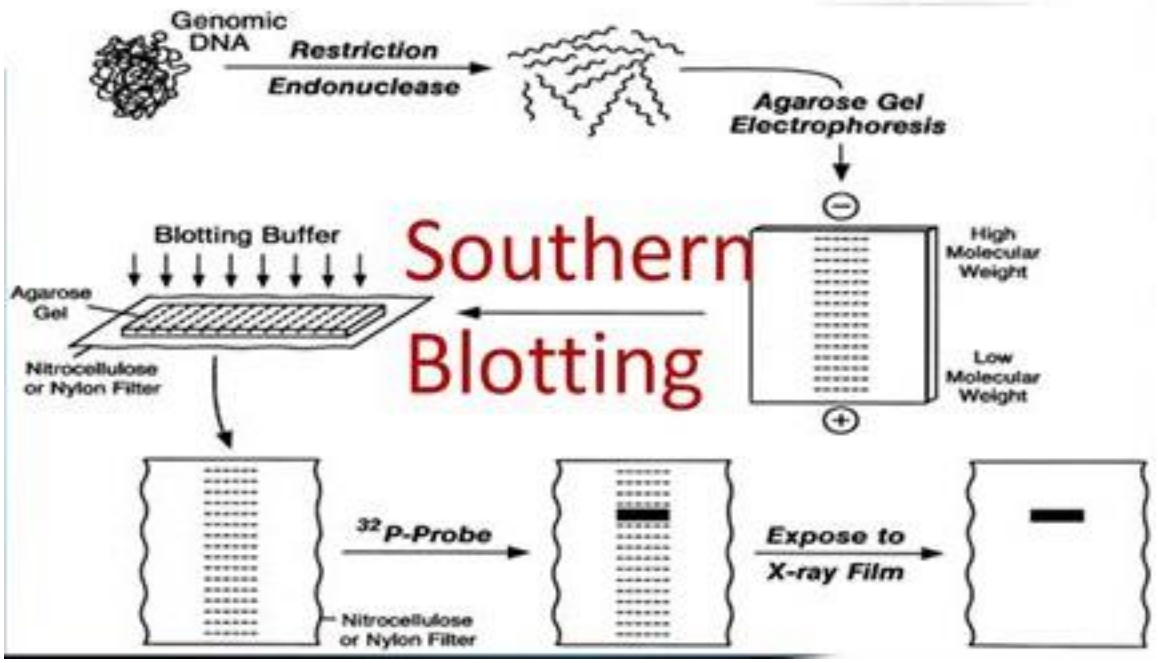
Q1b

RFLP = restriction fragment length polymorphism (e.g. EcoRI cuts GAATTC)



When cut, probe P will hybridize both to the green and red fragment and yield two bands

There is a polymorphism in the restriction site; the fragment will therefore not be cut, so only one band will be detected



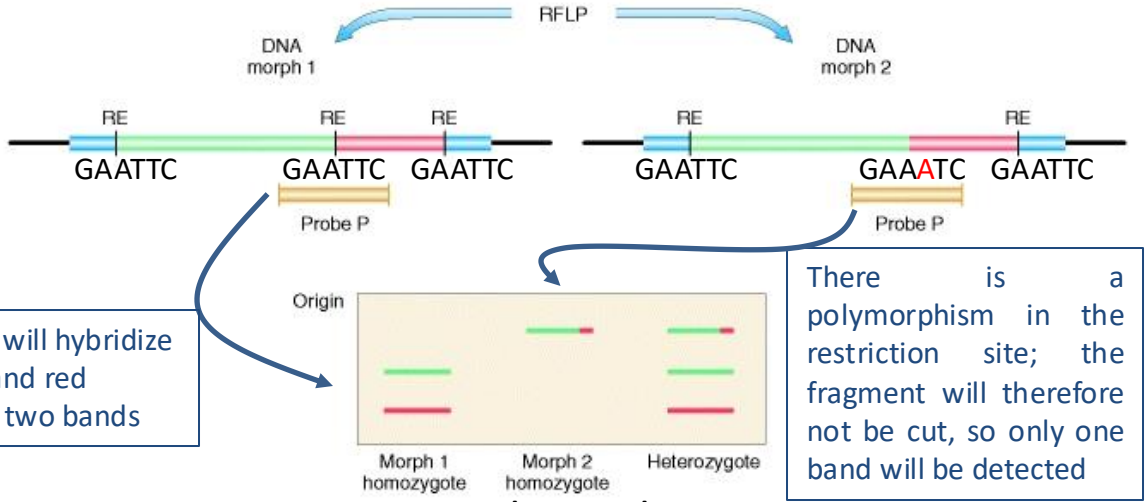
A **Southern blot** is a method used in molecular biology for the detection of a specific DNA sequence in DNA samples. Southern blotting combines transfer of electrophoresis-separated DNA fragments to a filter membrane and subsequent fragment detection by probe hybridization.

The method is named after British biologist Edwin **Southern** (first published in 1975).

Botstein and his human genetic linkage map: RFLP method

Q1b

RFLP = restriction fragment length polymorphism (e.g. EcoRI cuts GAATTC)



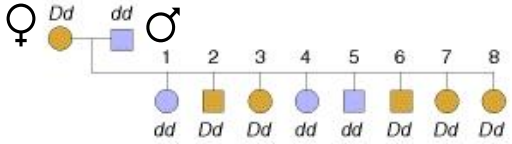
When cut, probe P will hybridize both to the green and red fragment and yield two bands

There is a polymorphism in the restriction site; the fragment will therefore not be cut, so only one band will be detected

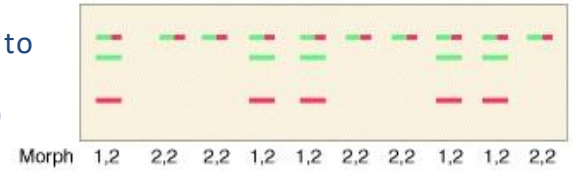
Southern Blot

Linkage to *D* locus

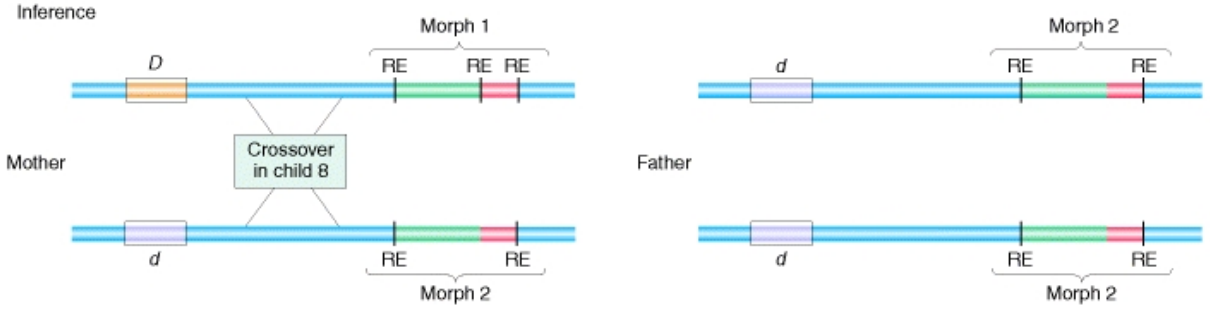
D = disease



Clear linkage of the *D* locus to the RFLP locus (only child 8 is recombinant)



If an individual is heterozygous for two morphs of an RFLP, this heterozygous "locus" can be used as a marker in chromosomal mapping



How was the human genetic linkage map constructed?

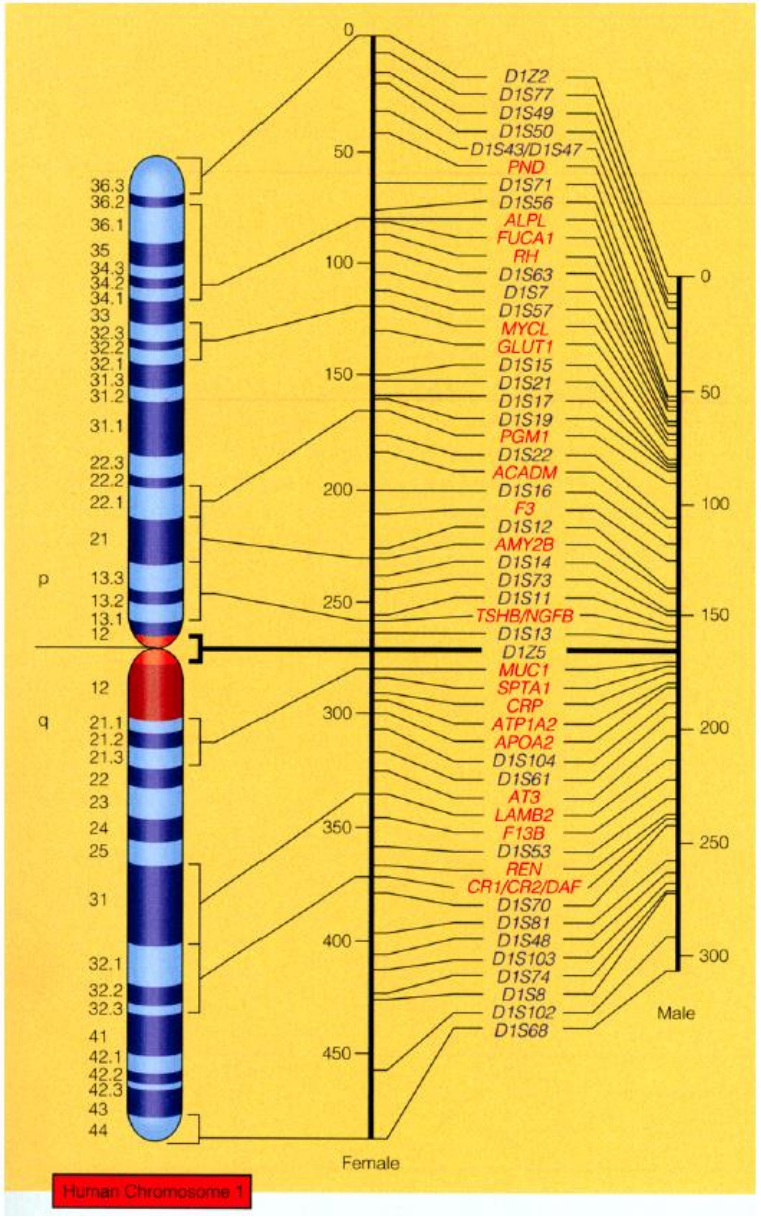
Q1b

A. We target a marker / RFLP (e.g. “P”) and start sequencing from this site
B. We focus on other genetic diseases and find a linked marker
C. We find other markers that link with “P” and examine the order of these markers
D. We use genomes from better characterized organisms (e.g. mouse) as a “genetic” guide

Botstein and his human genetic linkage map: RFLP method

petit

region 2, band 1, sub-band 3



- 30,000 markers in the human genome (100-500 bp)

Q2a

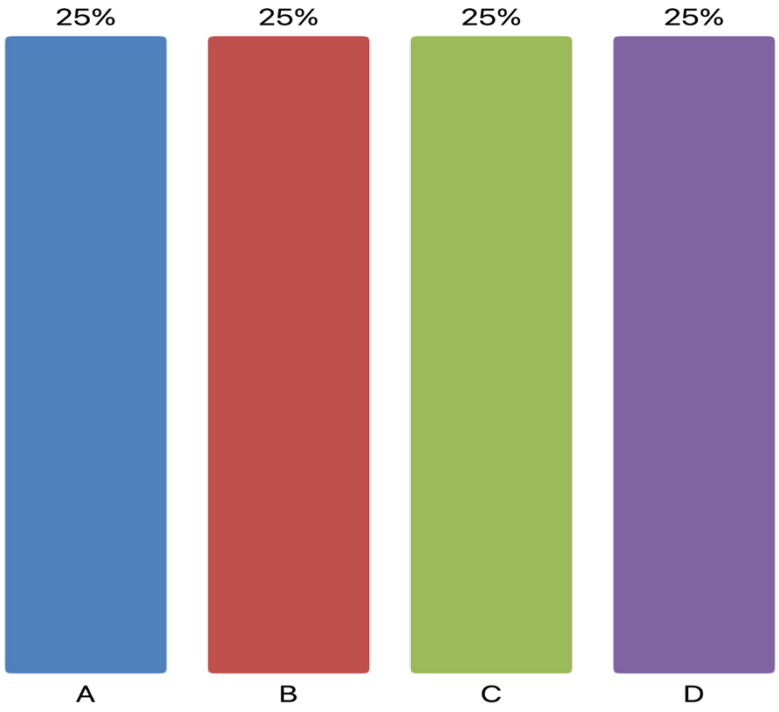
- 1 genetic map unit (cM) = 1% chance (=1 out of 100 meioses resulting in a recombination event) that a marker at one genetic locus on a chromosome will be separated from a marker at a second locus due to crossing over in a single generation = about 1Mb in humans (physical distance)

- cM = centimorgan

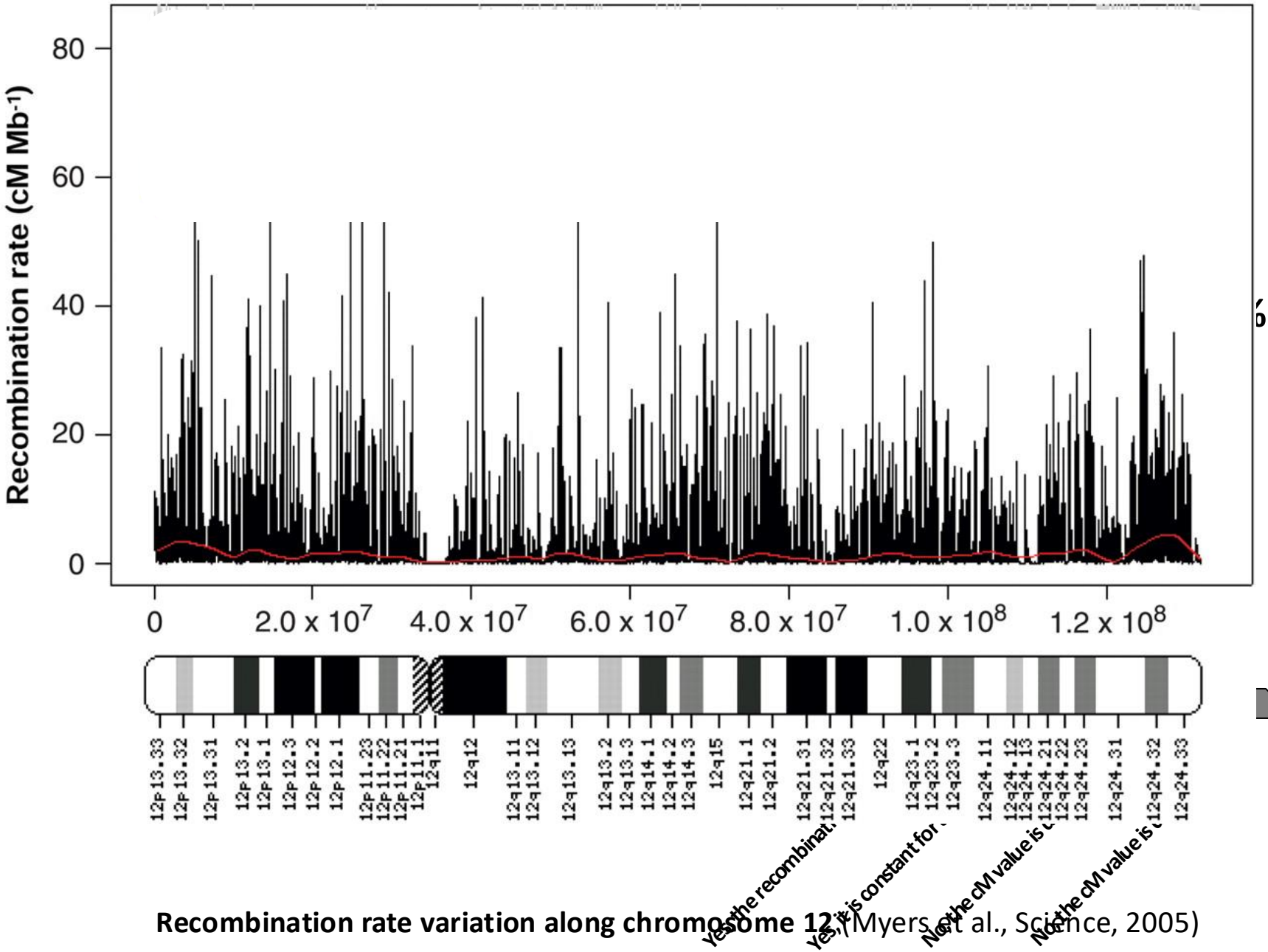
Does cM represent a constant value accros organisms?

Q2b

A. Yes, the recombination frequency along chromosomes is constant independent of organism
B. Yes, it is constant for all multicellular organisms, but differs from that of bacteria
C. No, the cM value is different in each organism, but is constant along chromosomes
D. No, the cM value is dependent on the organism and on the chromosomal region



80



Recombination rate variation along chromosome 12 (Myers et al., Science, 2005)

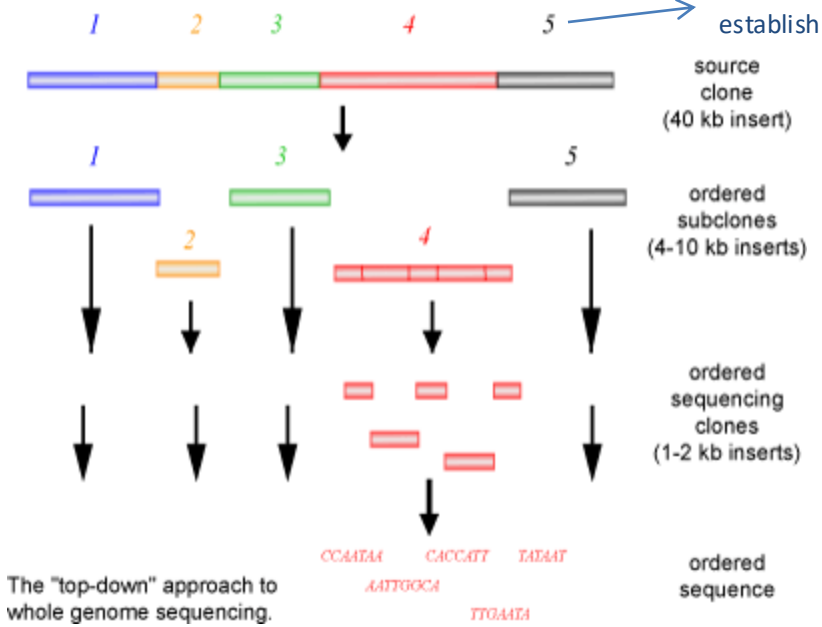
Genomics – a short history

The human genome sequence race (1990): public versus private

Chromosome walking

Q3a

Genetic map helped to establish the order!



The "top-down" approach to whole genome sequencing.

NHGRI & Sanger Center

E. coli

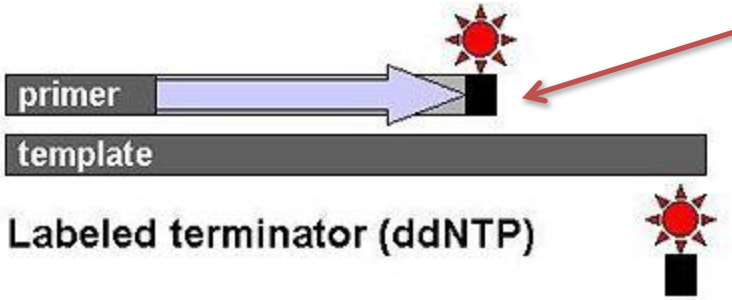
S. cerevisiae

Mycoplasma capricolum

Caenorhabditis elegans

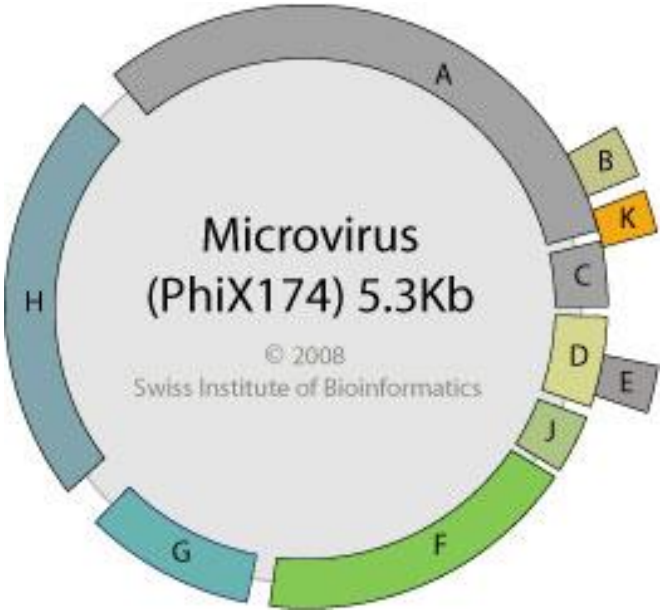
Genomics – a short history

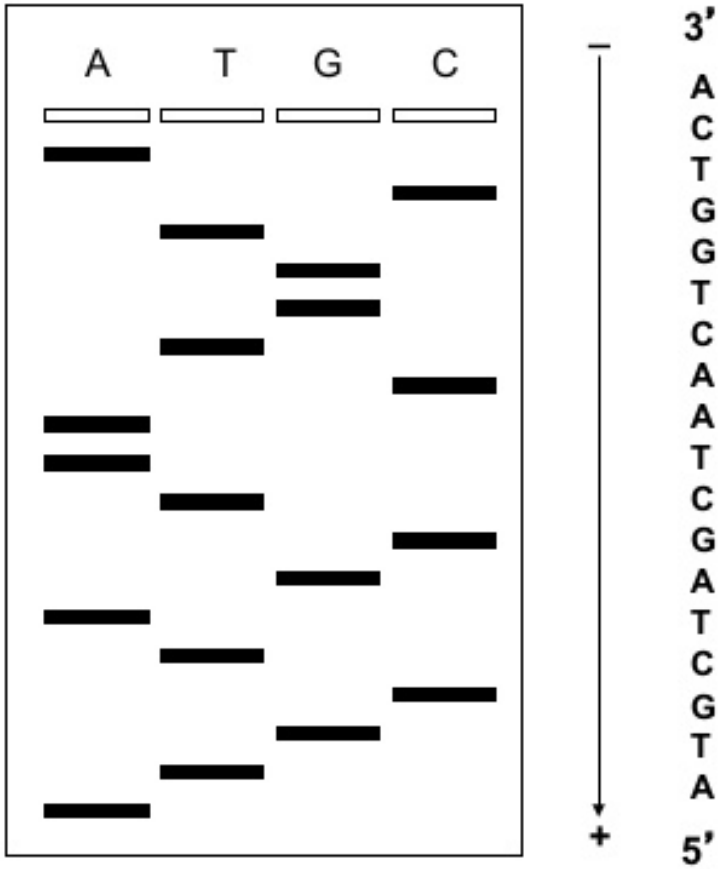
Q3b 1975: Sanger sequencing (chain termination sequencing)



Use of radioactivity to visualize bands
(4 reactions, each involving one kind of ddNTP)

Sanger et al., Nature, 1977

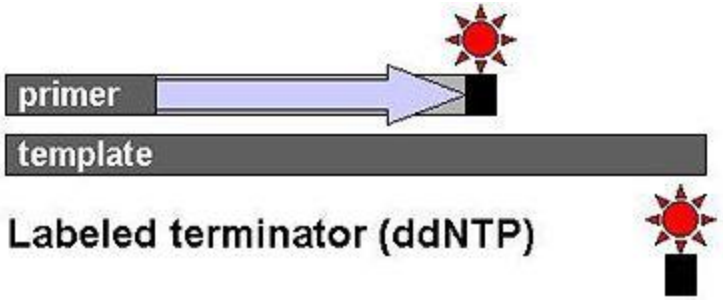


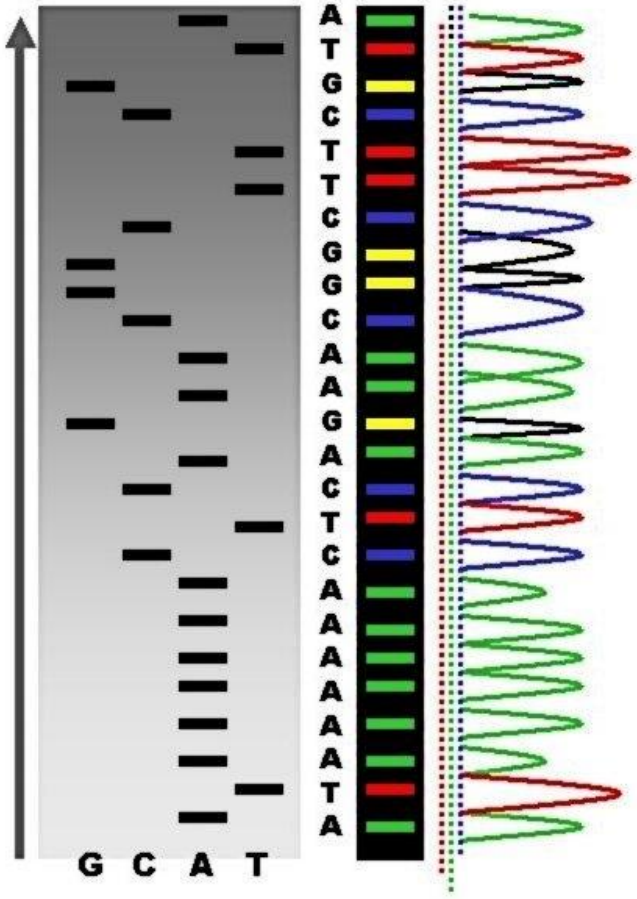


Genomics – a short history

Q3b Smith and Hunkapillar introduced fluorescent dyes plus use of a capillary column

Only 1 reaction and automatable!





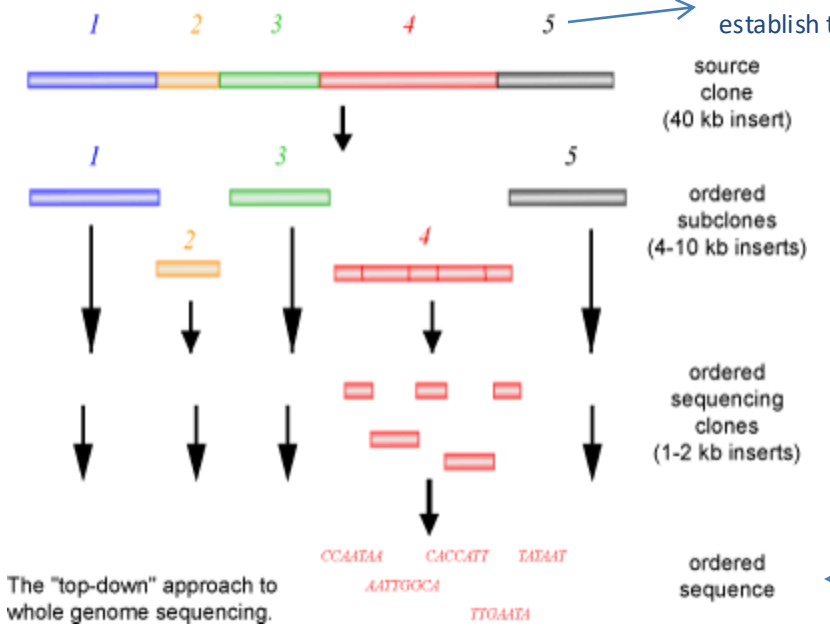
Genomics – a short history

The human genome sequence race (1990): public versus private

Chromosome walking

Q3a

Genetic map helped to establish the order!



The "top-down" approach to whole genome sequencing.

Sanger sequencing
(Smith and Hunkapillar method)

NHGRI & Sanger Center

E. coli

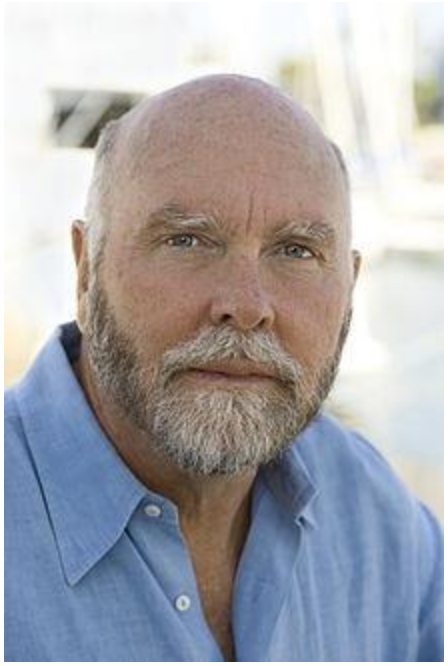
S. cerevisiae

Mycoplasma capricolum

Caenorhabditis elegans

Genomics – a short history

A “new” kid on the block: Craig Venter

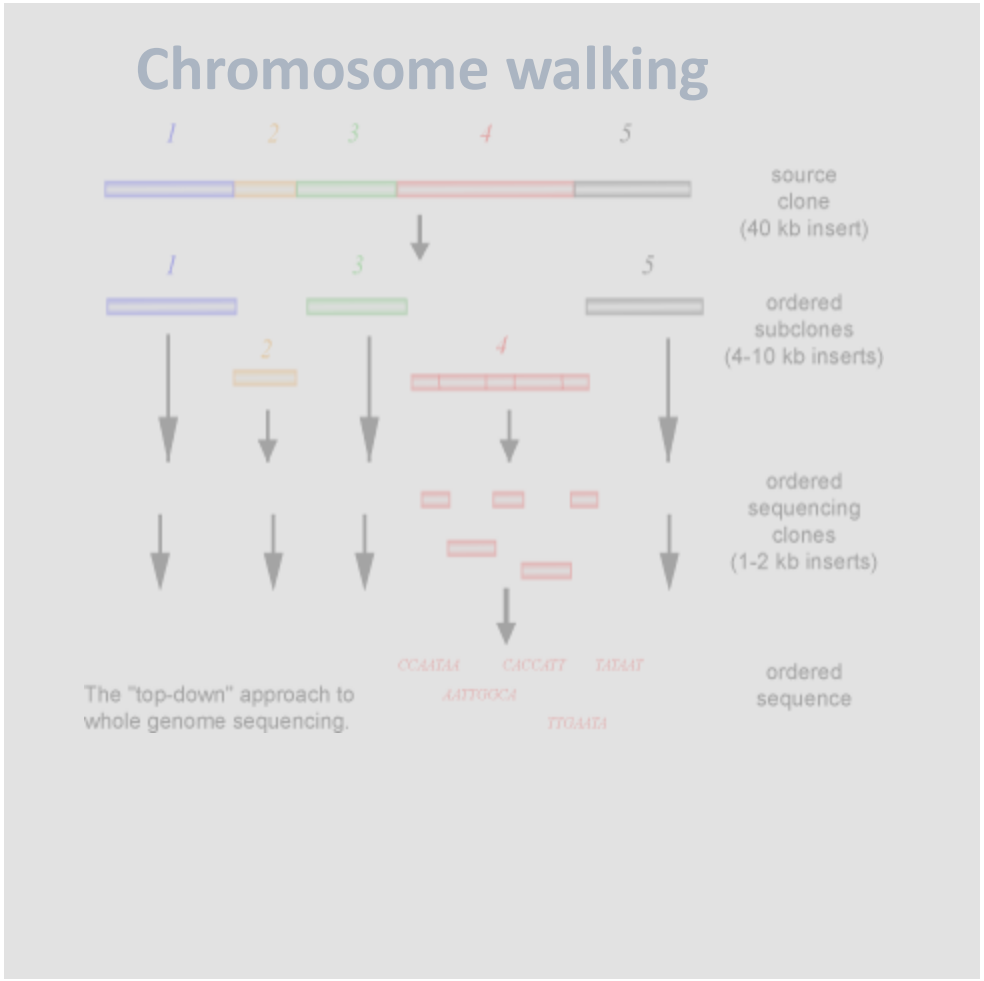


- NIH staff scientist
- Sequences cDNA → patents
- Wallace Steinberg provides funds for TIGR (non-profit) and Human Genome Sciences (profit)
- Mid-nineties: SmithKline Beecham → $110 million for exclusive rights of patents

J. Craig Venter™
INSTITUTE

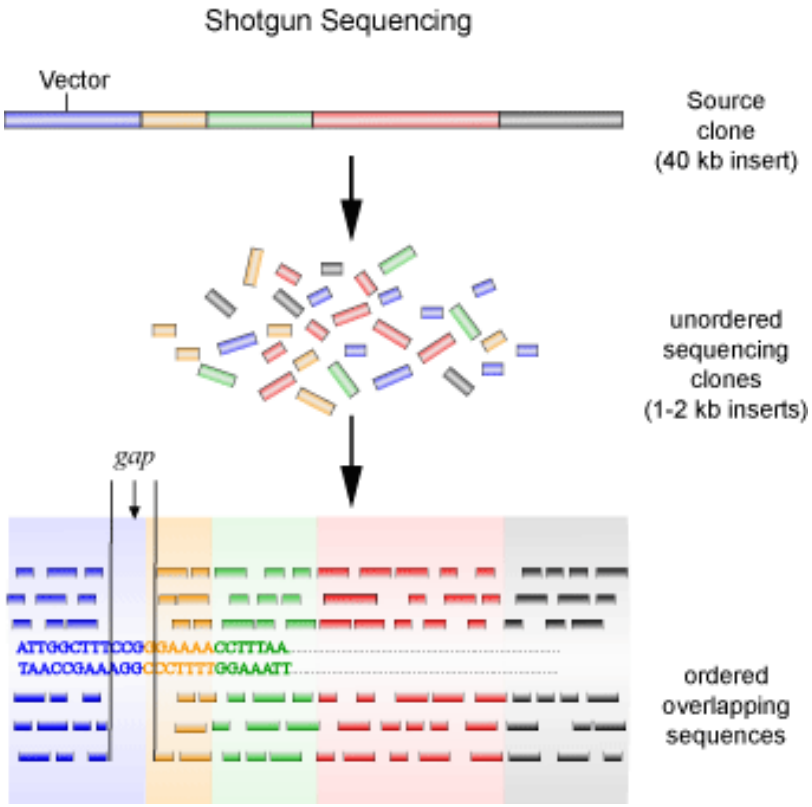
Genomics – a short history

The human genome sequence race (1990): public versus private



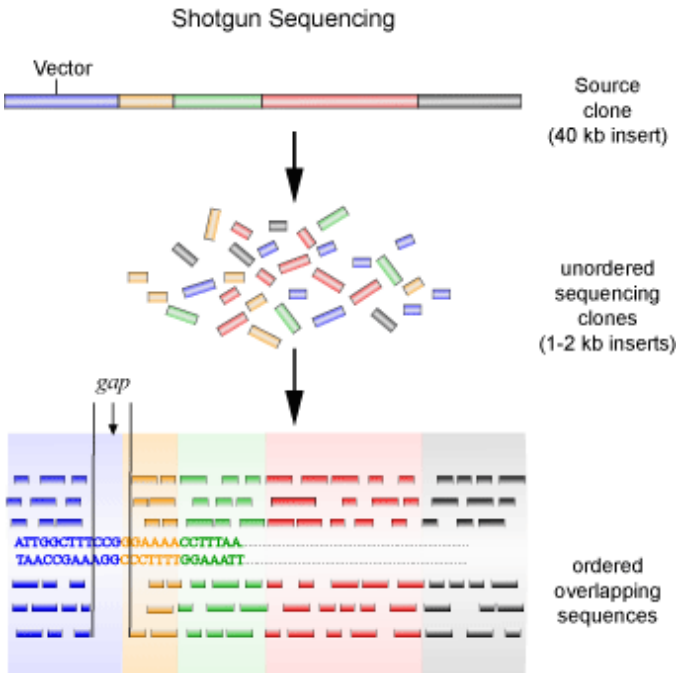
Q3a

Shot-gun sequencing



Genomics – a short history

The assembly problem



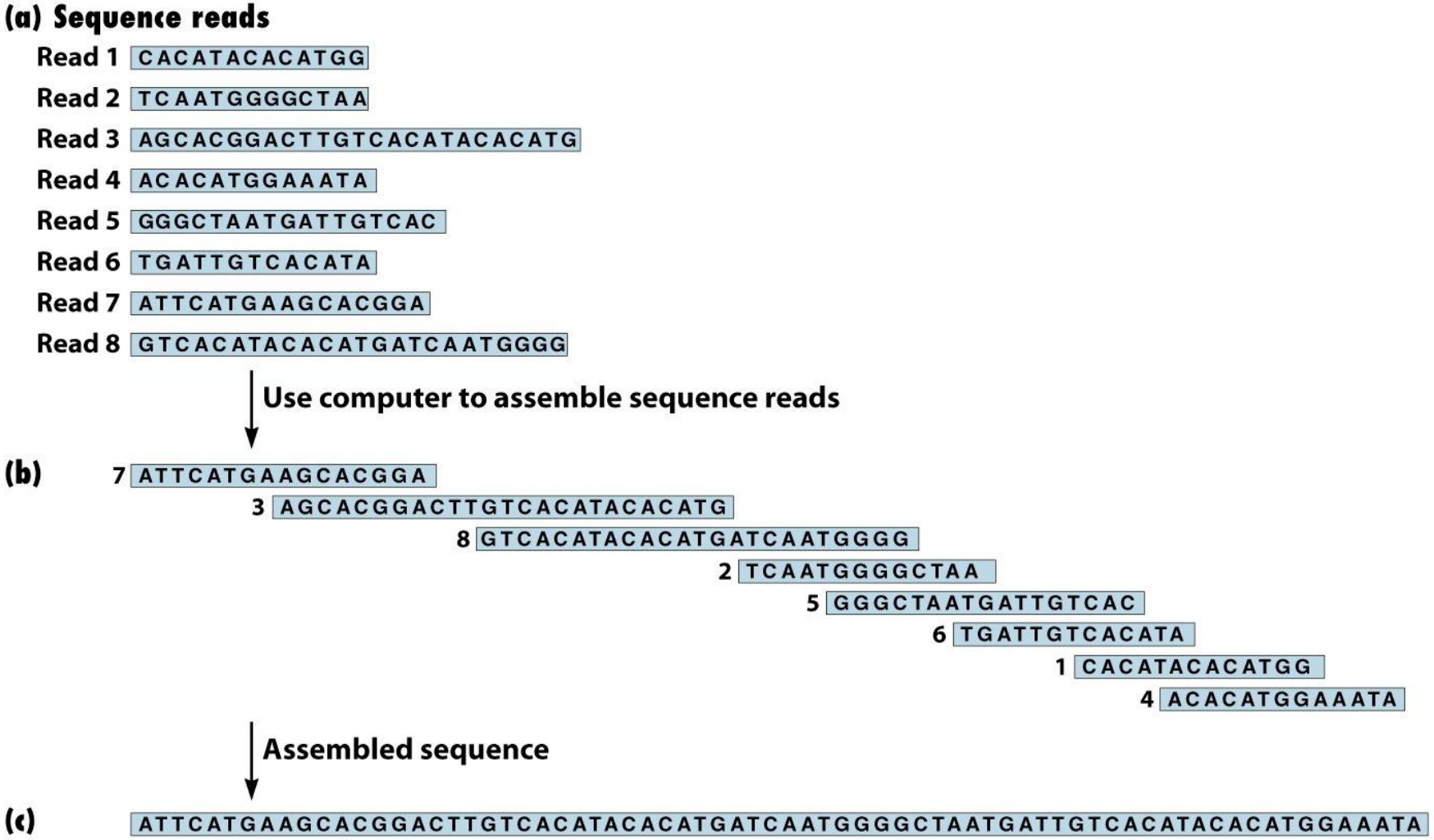
Fragments is a fundamental part of
Part of every sequencing method and
Sequencing method and every sequencing project
Sequence from fragments is a fundamental
Assembly of a sequence from fragments

But when more complicated (i.e. more like the actual human genome):

Q3a *That he is mad, 'tis true: 'tis true, 'tis pity;*
And pity 'tis 'tis true
(Hamlet)

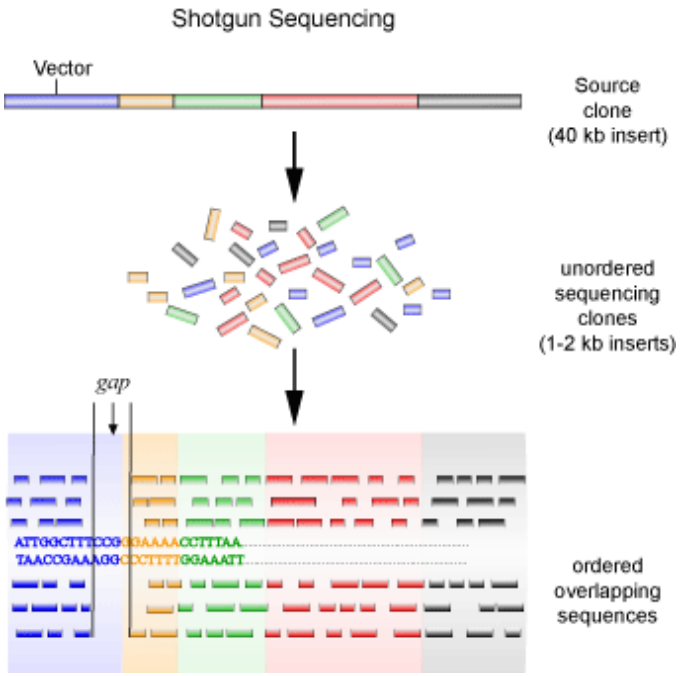
Genomics – a short history

The assembly problem



Genomics – a short history

The assembly problem



Fragments is a fundamental part of
Part of every sequencing method and
Sequencing method and every sequencing project
Sequence from fragments is a fundamental
Assembly of a sequence from fragments

But when more complicated:

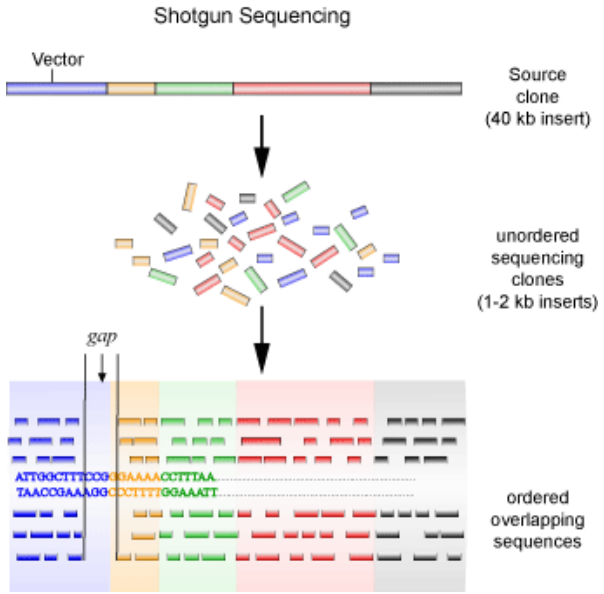
That he is mad, 'tis true: 'tis true, 'tis pity;
And pity 'tis 'tis true
(Hamlet)

Sutton, G. G., White, O., Adams, M. D., Kerlavage, A. R.
TIGR Assembler: A new tool for assembling large shotgun sequencing projects
Genome Science and Technology. 1995 Jan 01; 1(1): 9-19.

- pairing oligonucleotide content (set of n-mers of length 10 -12) instead of actual sequences → reducing search time
- repeat regions are dealt with by increasing the match criteria stringency and by assembling these regions last
- constraints such as clone length (wet bench)

Genomics – a short history

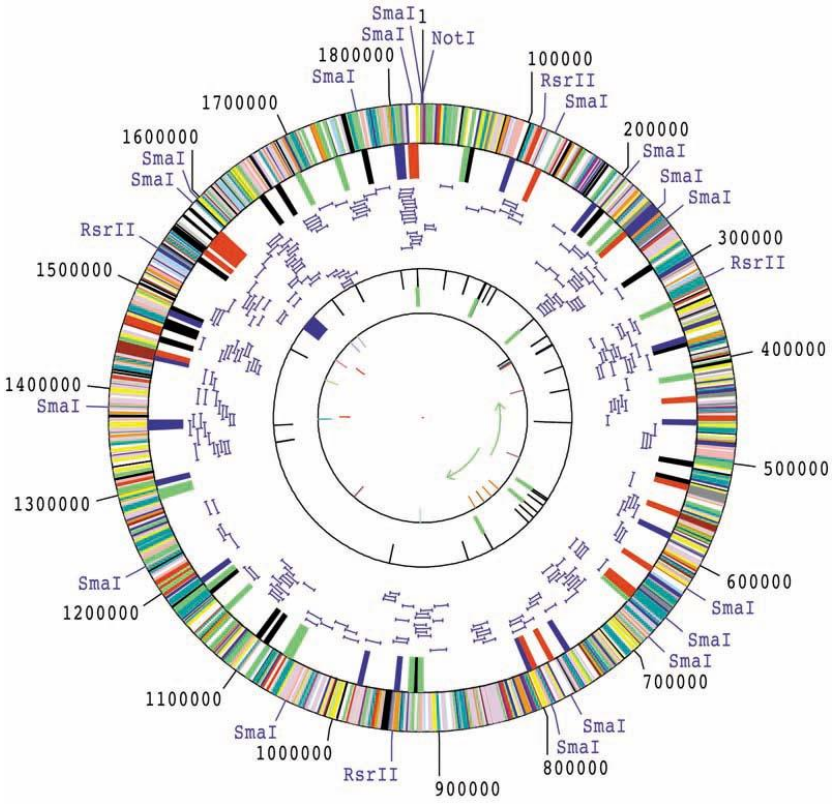
TIGR shocks the world



30 h on a SPARCCenter 2000 (500 Mb RAM)

1995: the first bacterial genome

Haemophilus influenzae (1.8 Mb)

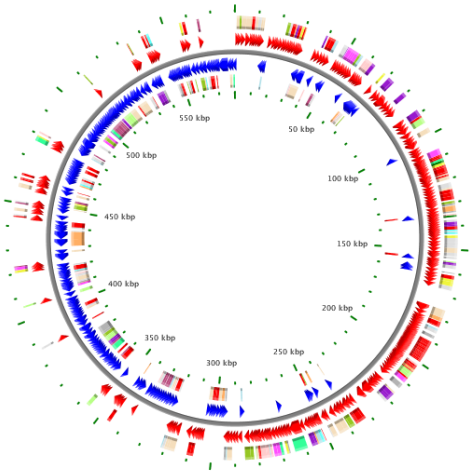


Sutton, G. G., White, O., Adams, M. D., Kerlavage, A. R. TIGR Assembler: A new tool for assembling large shotgun sequencing projects
Genome Science and Technology. 1995 Jan 01; 1(1): 9-19.

Genomics – a short history

TIGR shocks the world

1995: *Mycoplasma genitalium* (580 kb, shortest bacterial genome) took 8 months



Other genomes arrived quickly:
Methanococcus jannaschii
Archaeoglobus fulgidus
Helicobacter pylori
Borrelia burgdorferi

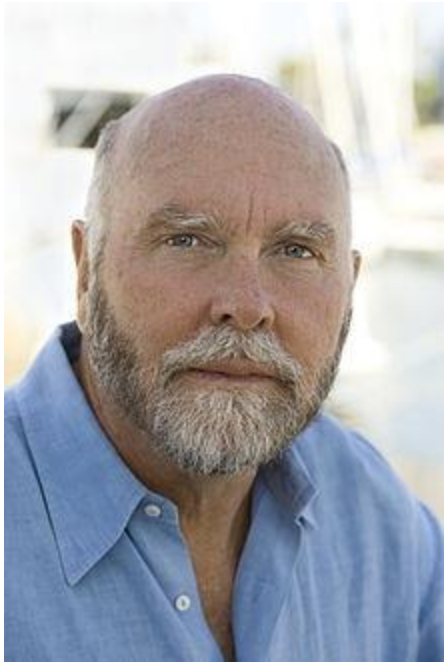
What about academia?

In 1997: both *E. coli* K12 and *S. cerevisiae* were finished

- Yeast required >100 laboratories involving >600 scientists
- Only 3.4% of sequencing duplication

Genomics – a short history

A "new" kid on the block: Craig Venter



- NIH staff scientist
- Sequences cDNA → patents
- Wallace Steinberg provides funds for TIGR (non-profit) and Human Genome Sciences (profit)
- Mid-nineties: SmithKline Beecham → $110 million for exclusive rights of patents
- First genome success of TIGR: *Haemophilus influenzae* --> shotgun sequencing!

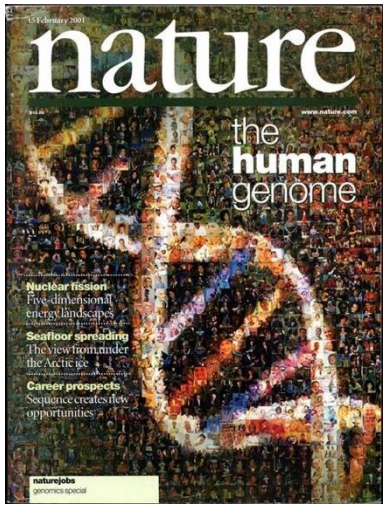
J. Craig Venter™
INSTITUTE

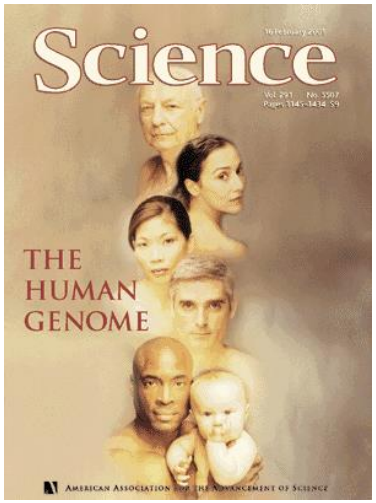
- 1997: Venter founds Celera Genomics with investment from Perkin Elmer (Applied Biosystems → Hunkapillar).
- Wellcome Trust and NIH step it up, the race is on.

The human genome

First announcement

In June 2000: first announcement of a working draft with the Nature and Science papers in February 2001





James Kent (UCSC)

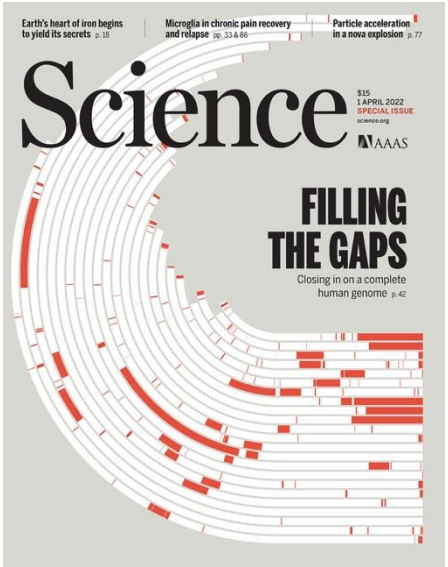
Eugene Myers (Celera)

International Human Genome Sequencing Consortium (2001) Nature 409:860-921; Venter et al. (2001) Science 291:1304-1351.

In June 2001: finished chromosome 20 and chromosome 1 in May 2006

Gregory et al. (2006), Nature, 441, 315-321

Q4

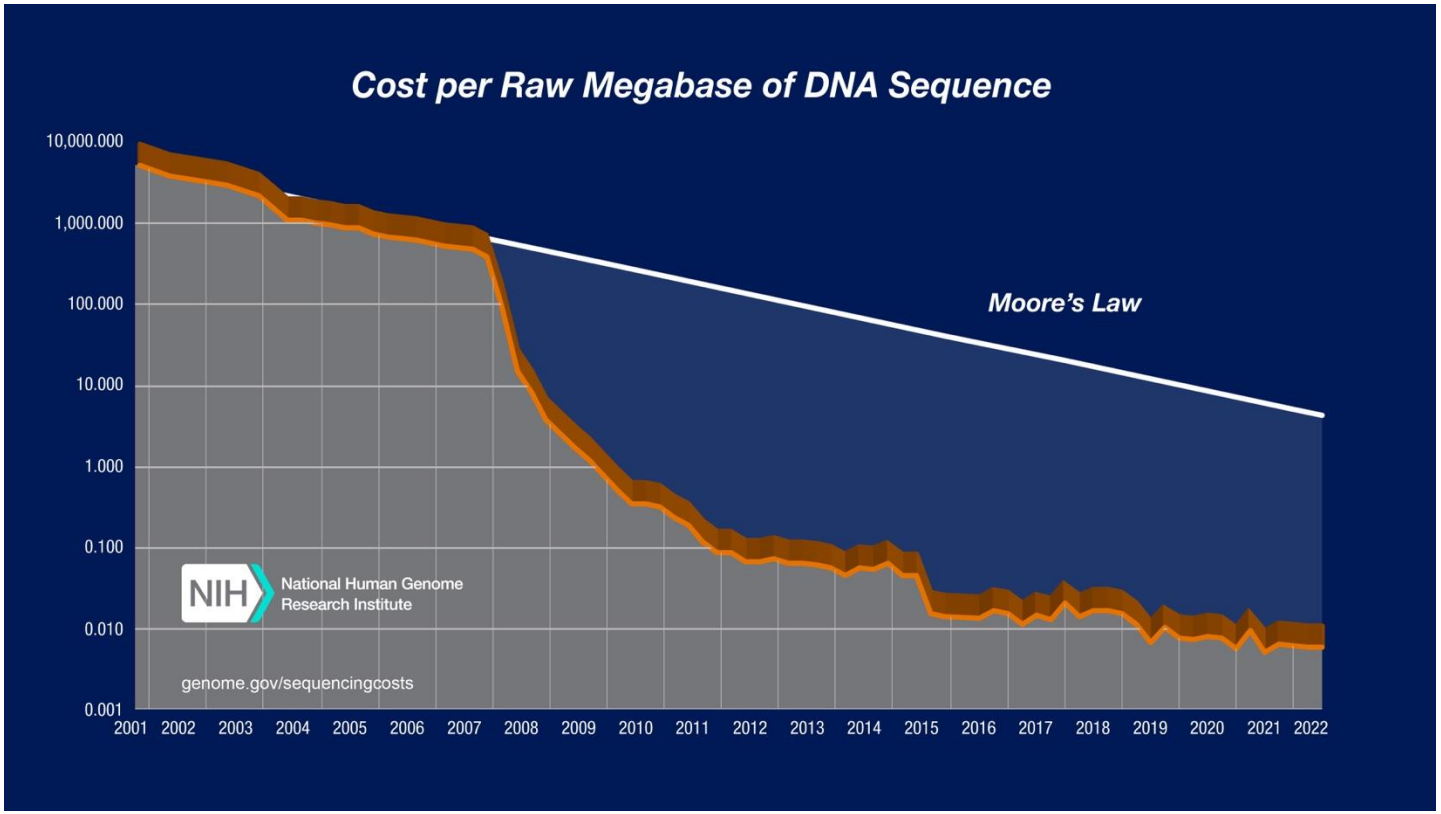


April 2022: From 200 million bases of DNA untouched to only 10 million bases and the Y chromosome only "roughly" known

Genomics – a short history

Sequencing cost

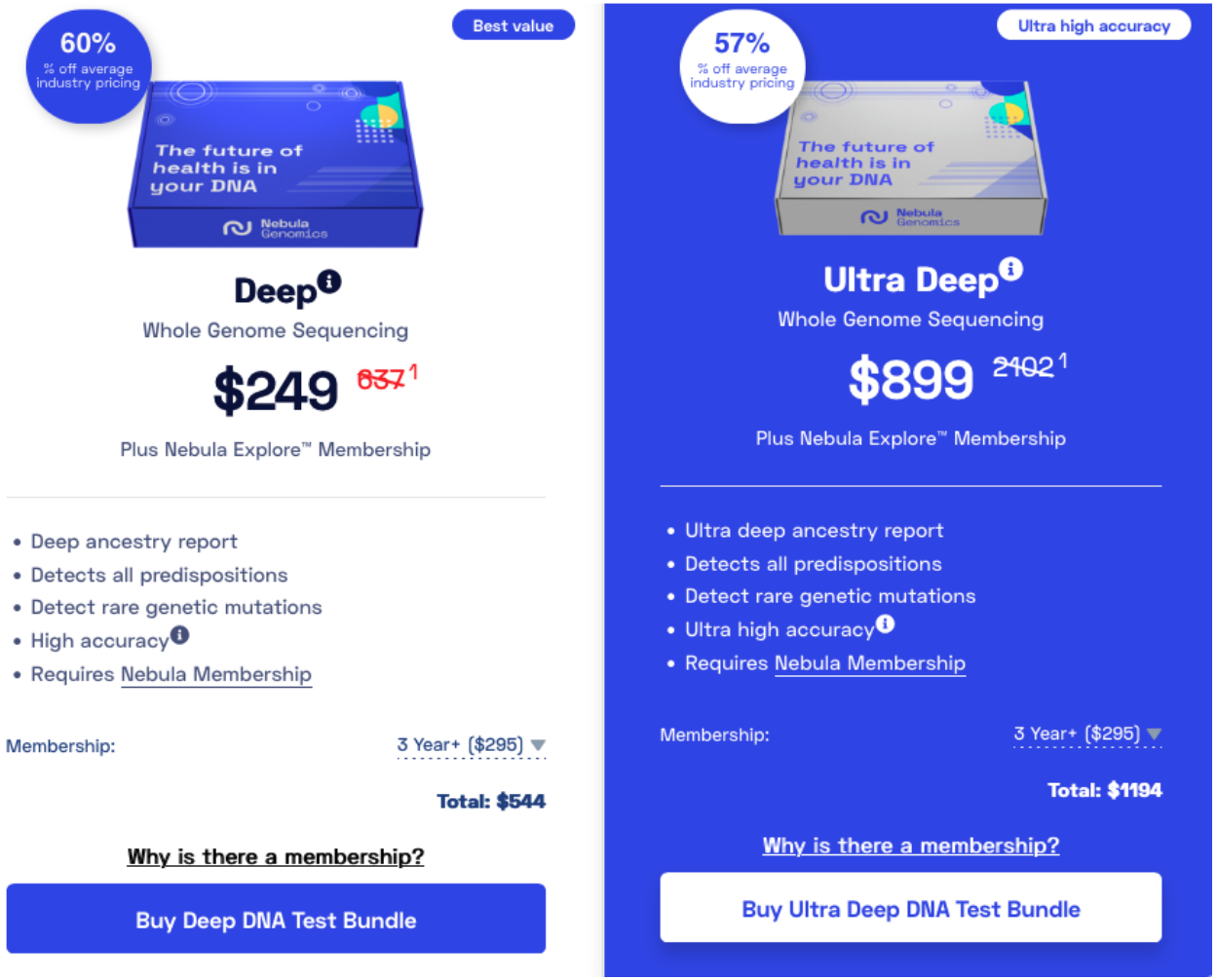
Year	Cost per base
1985	10 dollars
1991	One dollar
1993	50 cents
2019	1 x 10^{-8} cents



Whole sale!

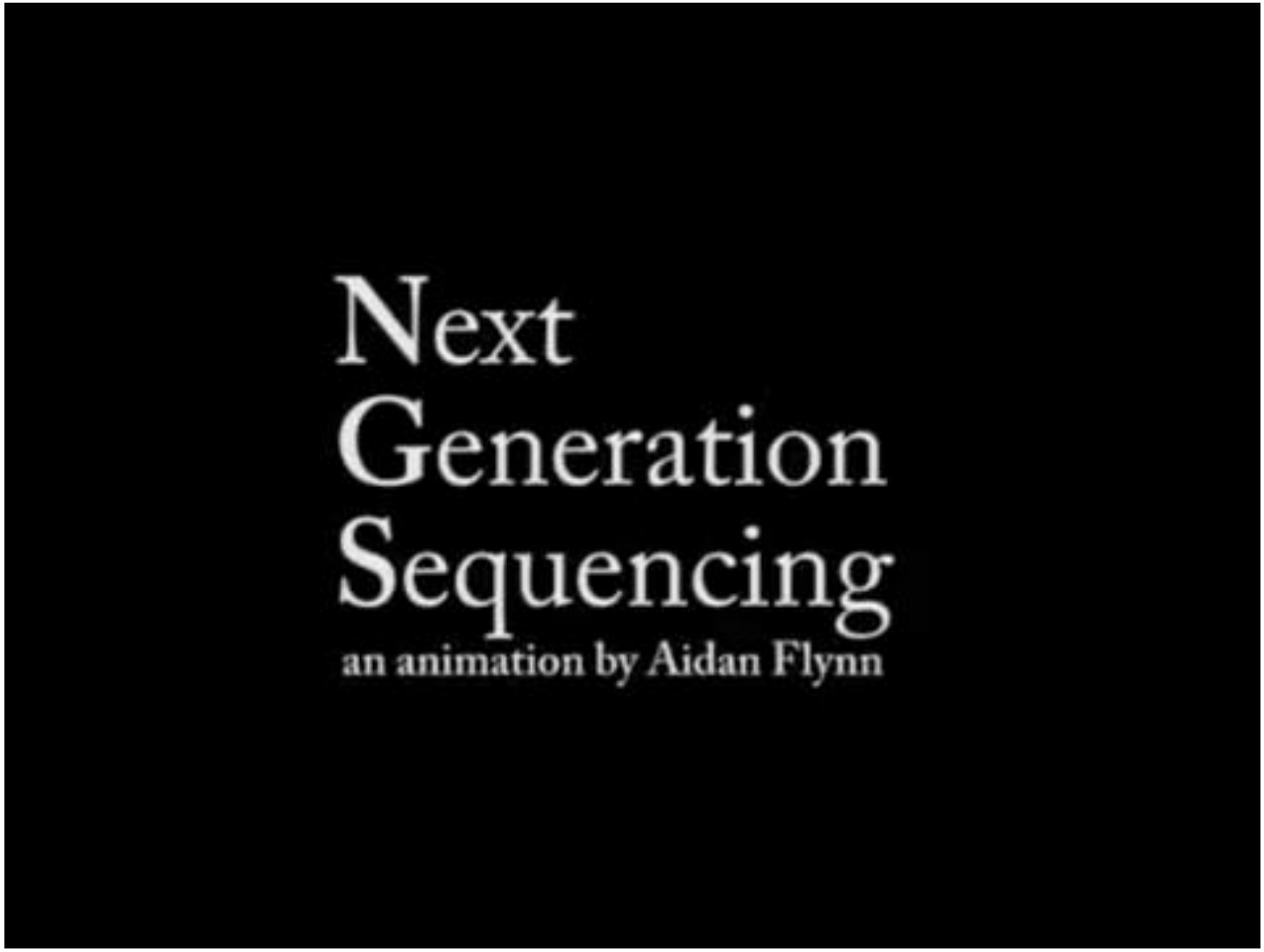
Nebula Genomics (a George Church lab spin-off)

October 11th 2024



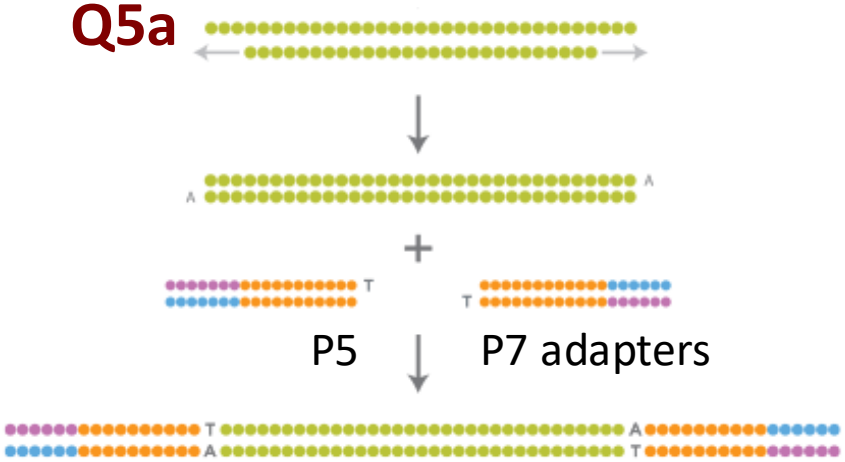
The revolution of high-throughput sequencing: Illumina

http://www.youtube.com/watch?v=77r5p8IBwJk&noredirect=1



The revolution of high-throughput sequencing: Illumina

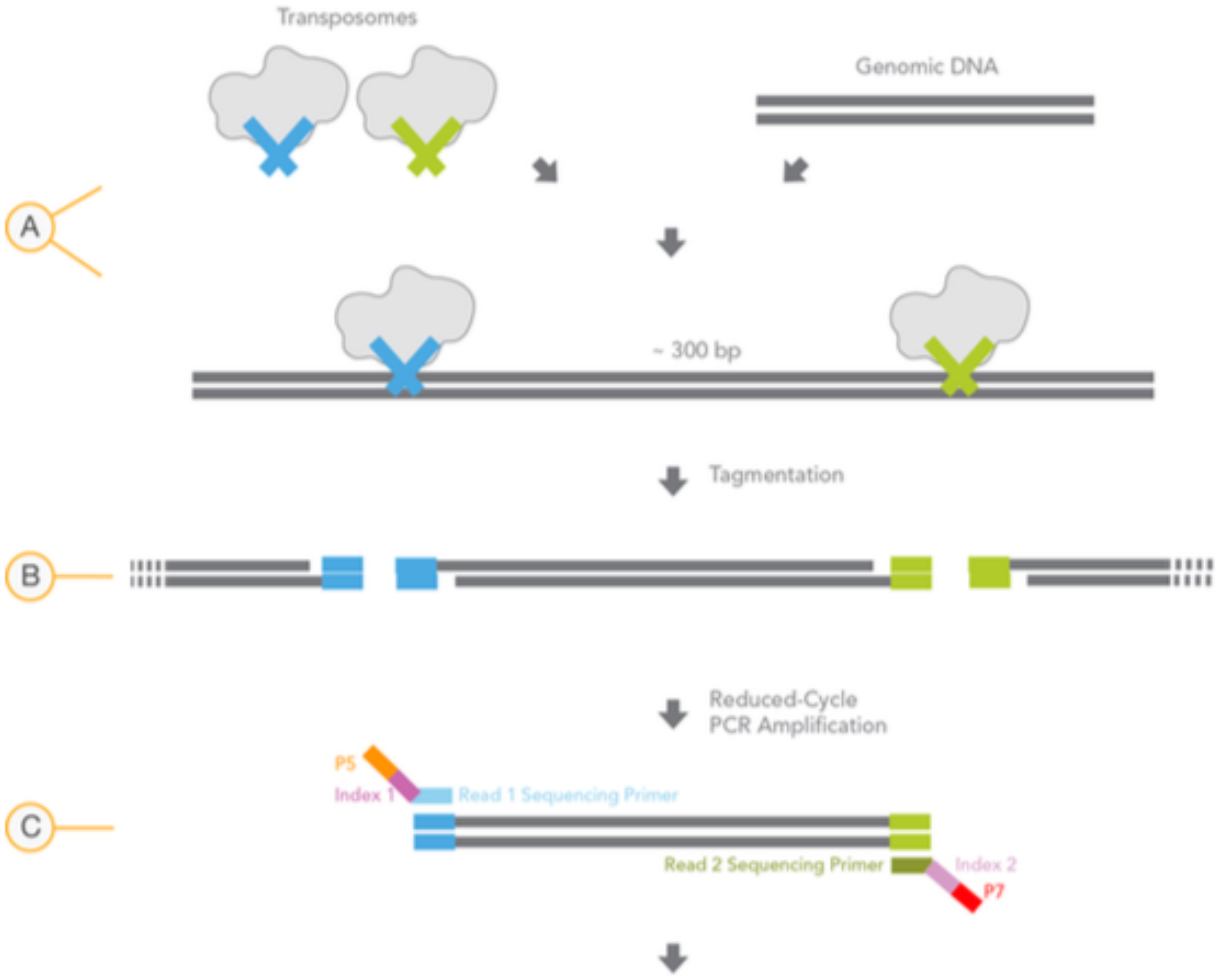
Q5a



1) Samples consisting of longer fragments are first sheared into a random library of 100-300 base-pair long fragments.
2) After fragmentation the ends of the obtained DNA-fragments are repaired and an A-overhang is added at the 3'-end of each strand.
3) Afterwards, adaptors which are necessary for amplification and sequencing are ligated to both ends of the DNA-fragments (inefficient → transposases!)
4) These fragments are then size selected and purified.

The revolution of high-throughput sequencing: Illumina

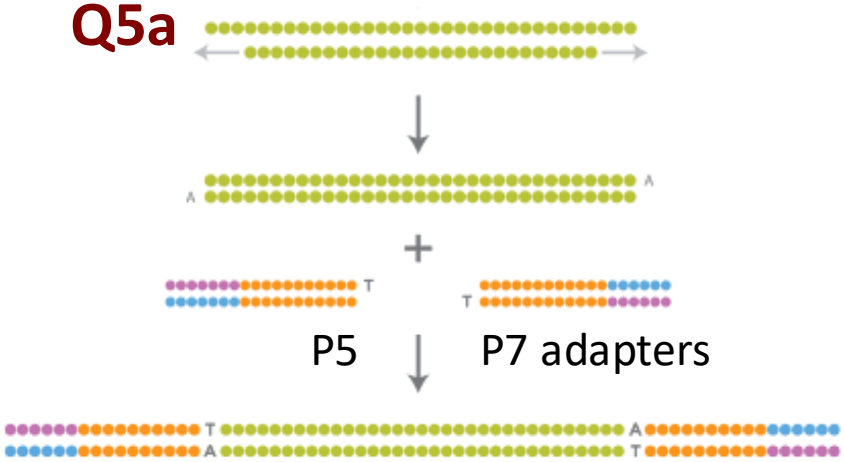
Because of inefficient ligation, “shearing” of DNA is performed with transposase (Tn5) loaded with P5 – P7-compatible adapters



Blot, M.; Heitman, J.; Arber,W. Tn5-mediated bleomycin resistance in Escherichia coli requires the expression of host genes. Mol. Microbiol. 1993, 8, 1017–1024

The revolution of high-throughput sequencing: Illumina

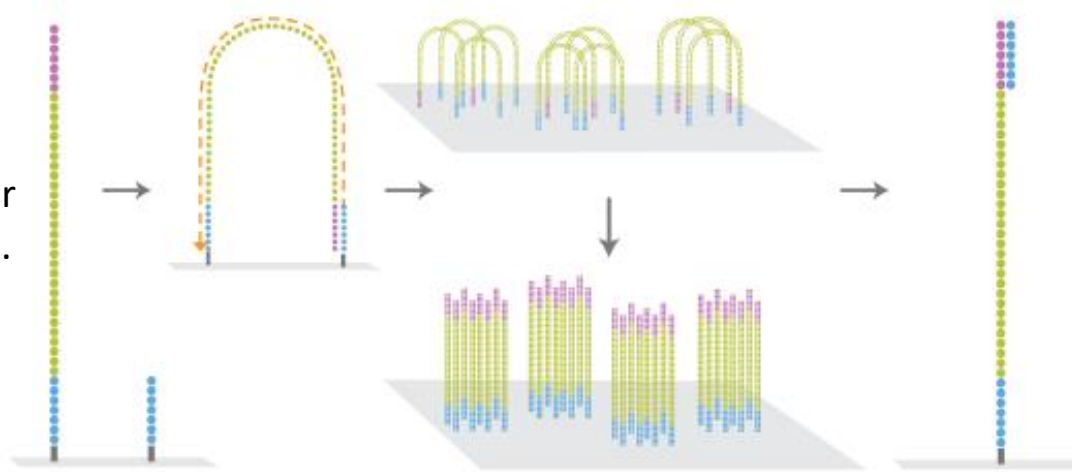
Q5a



1) Samples consisting of longer fragments are first sheared into a random library of 100-300 base-pair long fragments.
2) After fragmentation the ends of the obtained DNA-fragments are repaired and an A-overhang is added at the 3'-end of each strand.
3) Afterwards, adaptors which are necessary for amplification and sequencing are ligated to both ends of the DNA-fragments (inefficient → transposases!)
4) These fragments are then size selected and purified.

The Cluster Generation is performed on the Illumina cBot:

1) Single DNA-fragments are attached to the flow cell by **hybridizing** to oligos on its surface that are complementary to the ligated adaptors, followed by extension after which original DNA templates are discarded.
2) The DNA-molecules are then amplified by **bridge amplification** which results in hundreds of millions of unique clusters.
3) Finally, the reverse strands are cleaved and washed away and the sequencing primer is hybridized to the DNA-templates.



Source: http://www.dkfz.de/gpcf/hiseq_technology.html

Appendix: Bridge amplification (2-3)

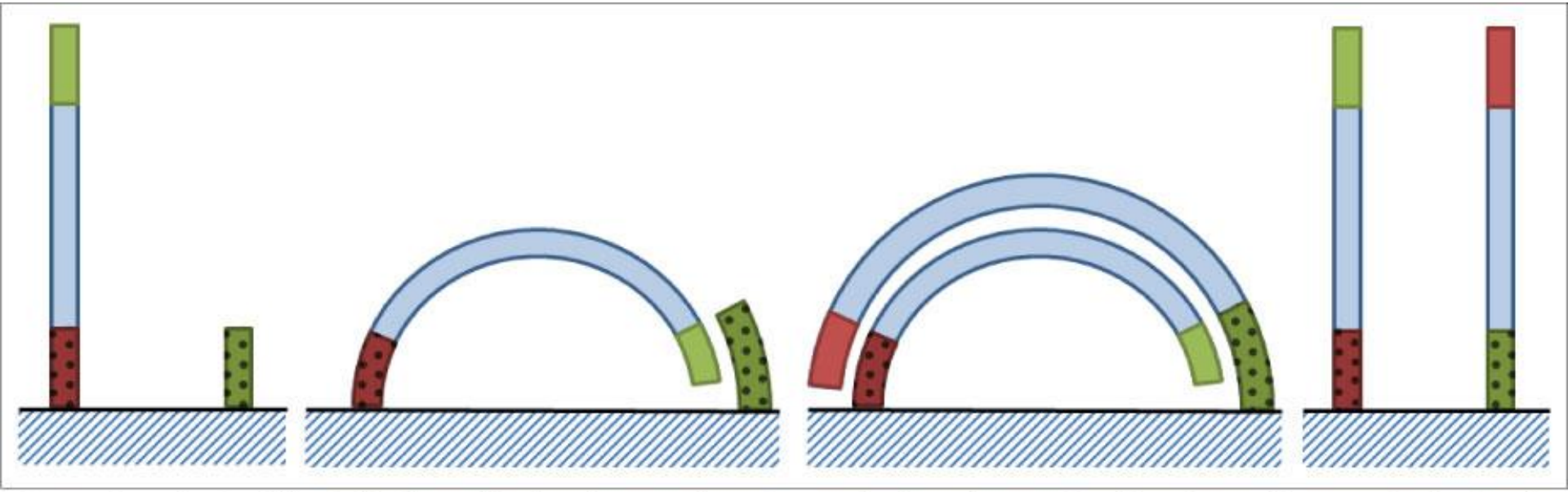
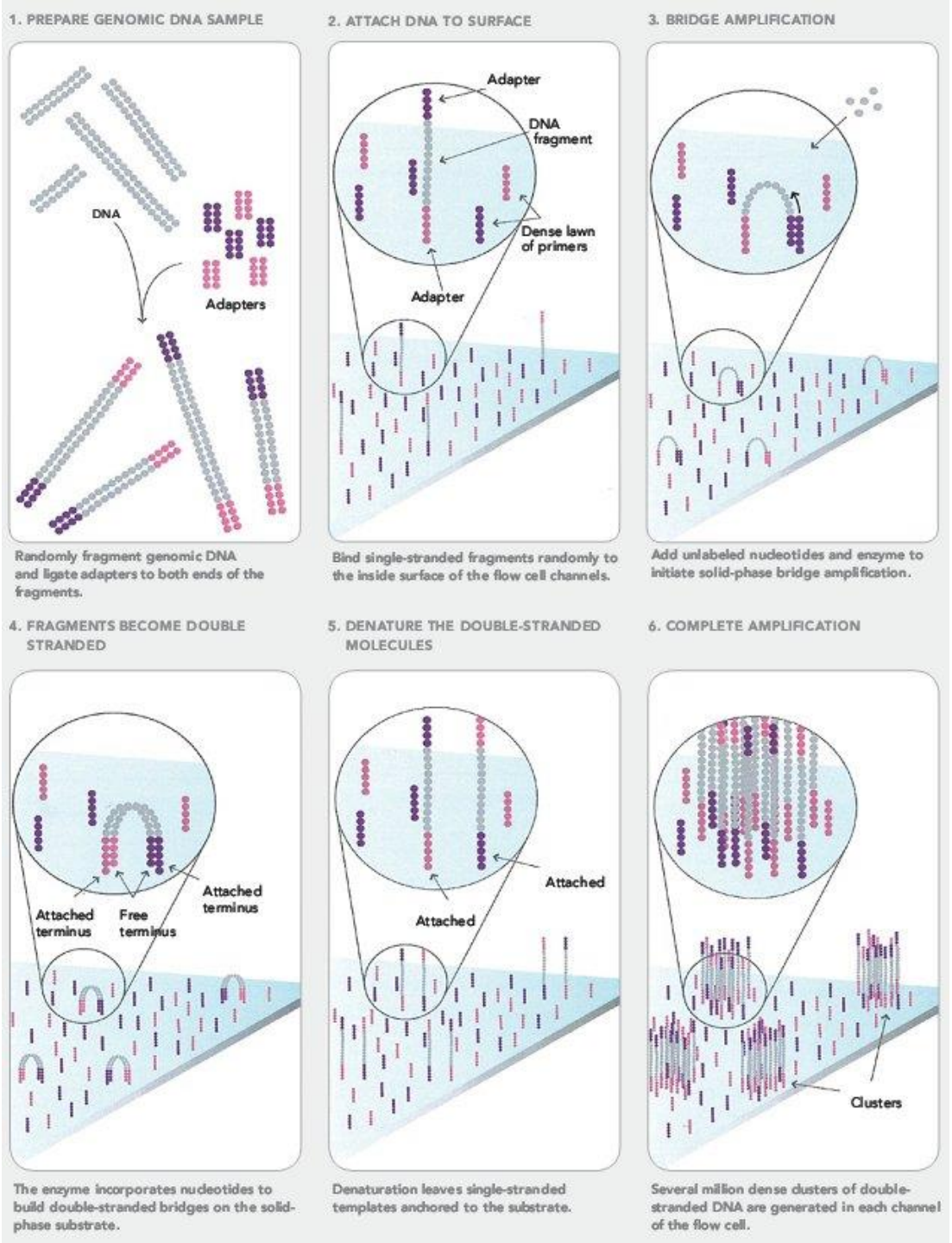


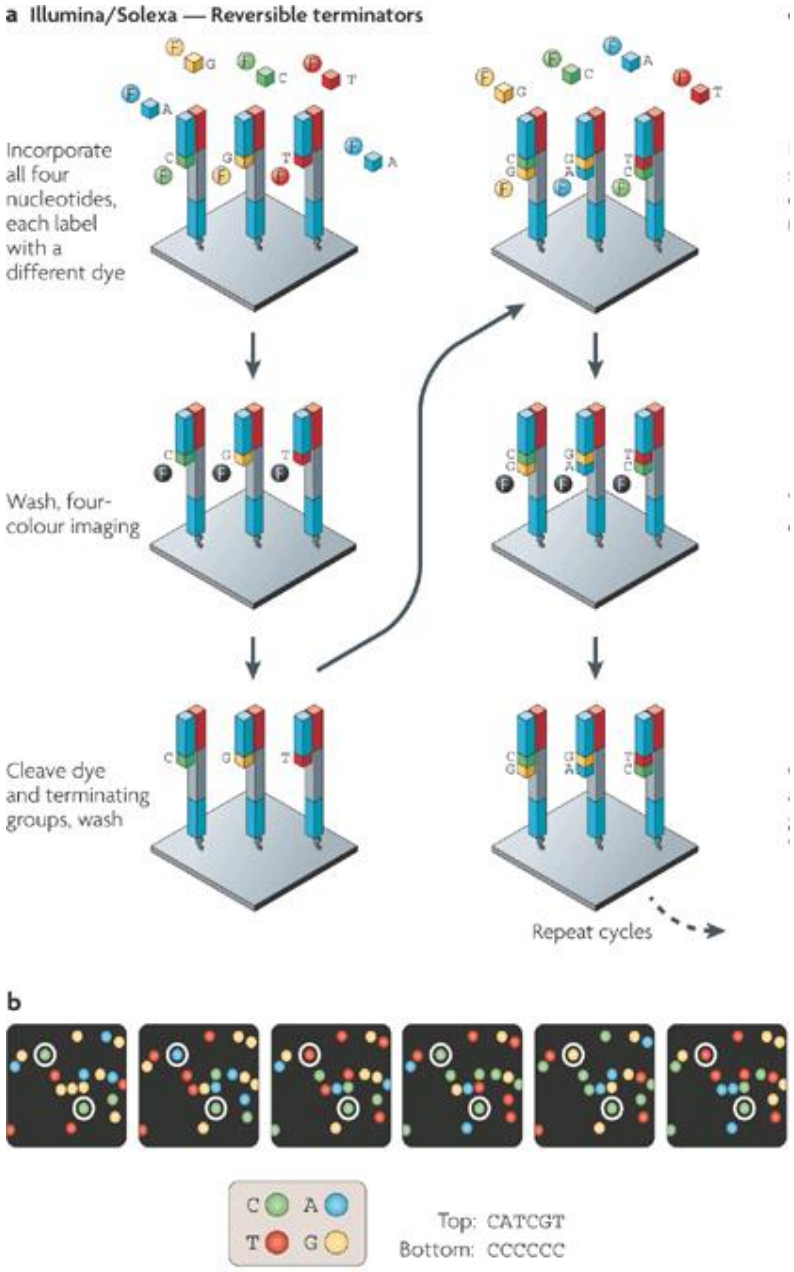
FIGURE 1: The mechanism of Illumina's bridge amplification. The process shown is repeated until DNA fragments form dense clusters. The green and red regions show the two primer sequences ligated to the blue DNA fragments. Primers that are complementary to these regions, and which are attached to the platform, are denoted in the cognate colour with spots. Once amplification is complete, the fragments can be sequenced from either of the two platform primers (i.e. the primers indicated by the spotted regions) or from both to produce paired-end reads.

From http://www.scielo.org.za/img/revistas/sajs/v108n11-12/16f01.jpg

Appendix: Bridge amplification (1-3)



The revolution of high-throughput sequencing: Illumina



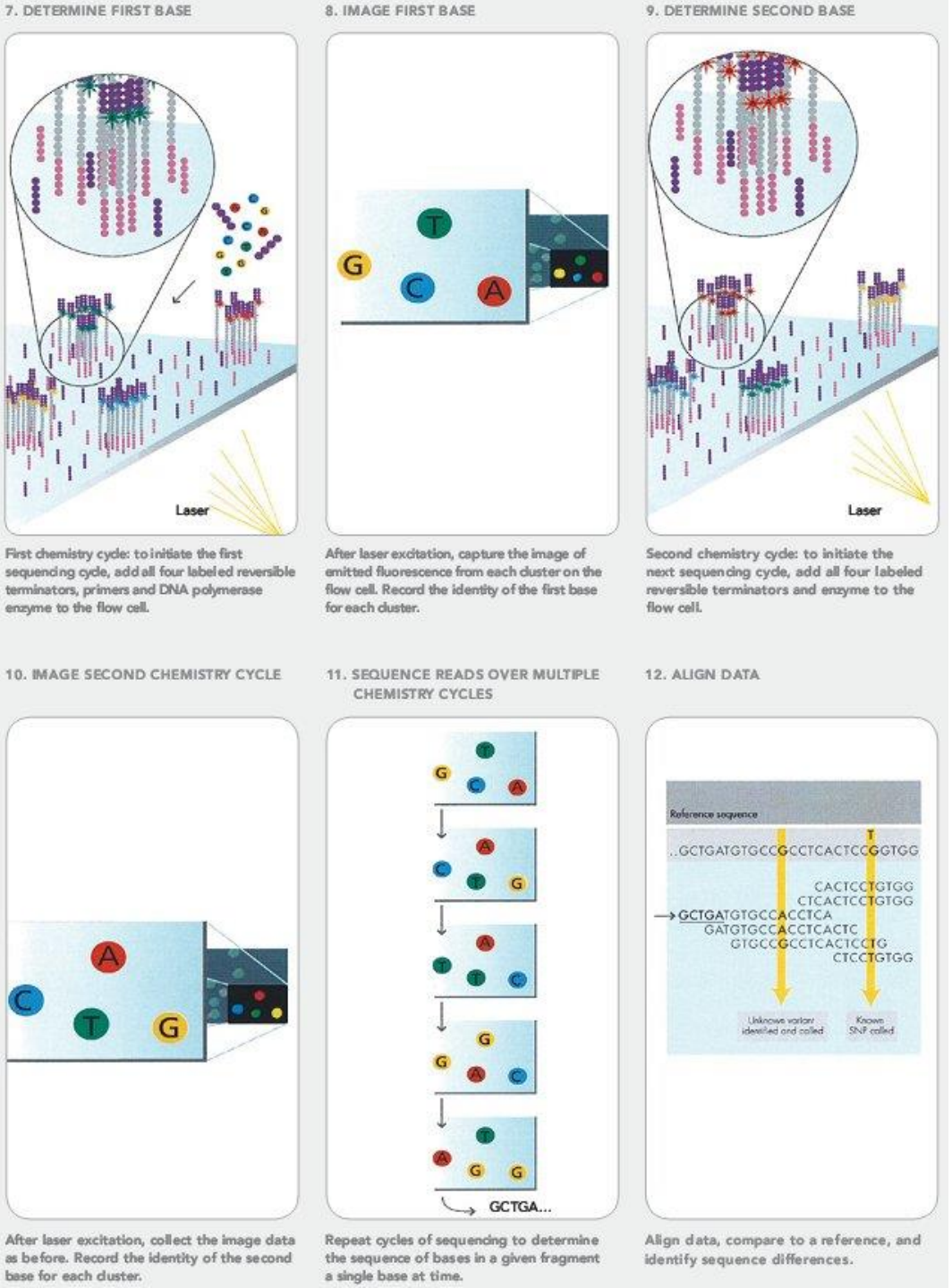
1. During sequencing the huge amount of generated clusters are sequenced simultaneously.
2. The DNA-templates are copied base by base using the four nucleotides (ACGT) which are fluorescently-labeled and reversibly terminated.
3. After each synthesis step, the clusters are excited by a laser which causes fluorescence of the last incorporated base.
4. After that, the fluorescence label and the blocking group are removed allowing the addition of the next base.
5. The fluorescence signal after each incorporation step is captured by a built-in camera, producing images of the flow cell.

Advantages

- all amplification occurs on the solid surface
- all amplified product remains covalently bound in specific pixels of an ordered array

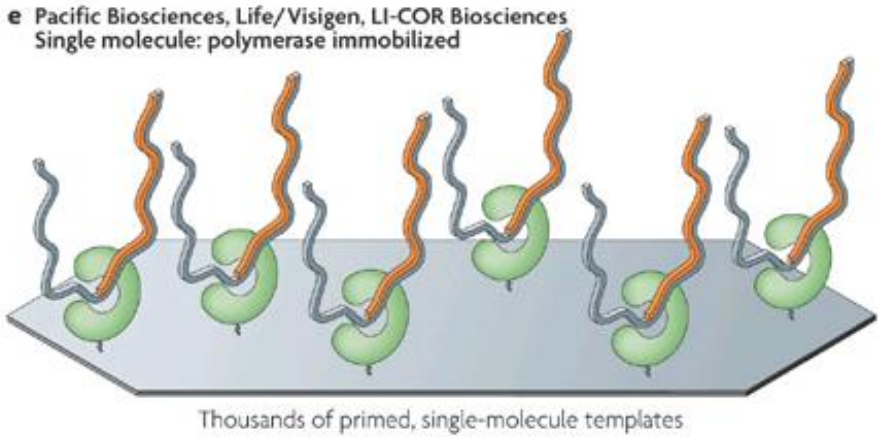
Metzker et al., Nat. Rev. Genet., 2010

Appendix: Bridge amplification (3-3)

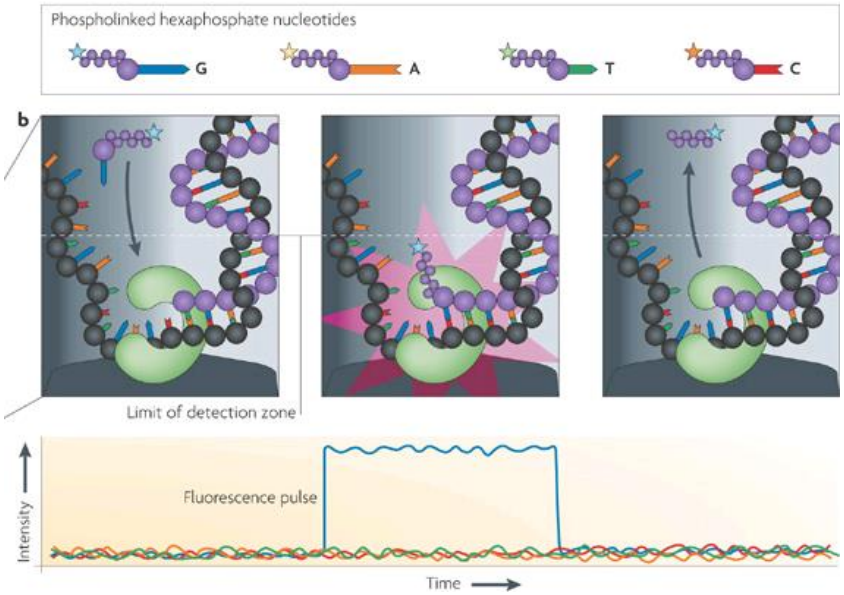


The revolution of high-throughput sequencing: 3rd and 4th generation technologies: *Pacific Biosciences*

Q5b



Green bean = immobilized DNA polymerase
Gray strand = unsequenced DNA, orange = sequenced



Real-time sequencing:
fluorescence pulse
(Unlike reversible terminators,
real-time nucleotides do not halt
the process of DNA synthesis)

Advantages:

- Single molecule sequencing
- (very) long reads possible
- Not too expensive

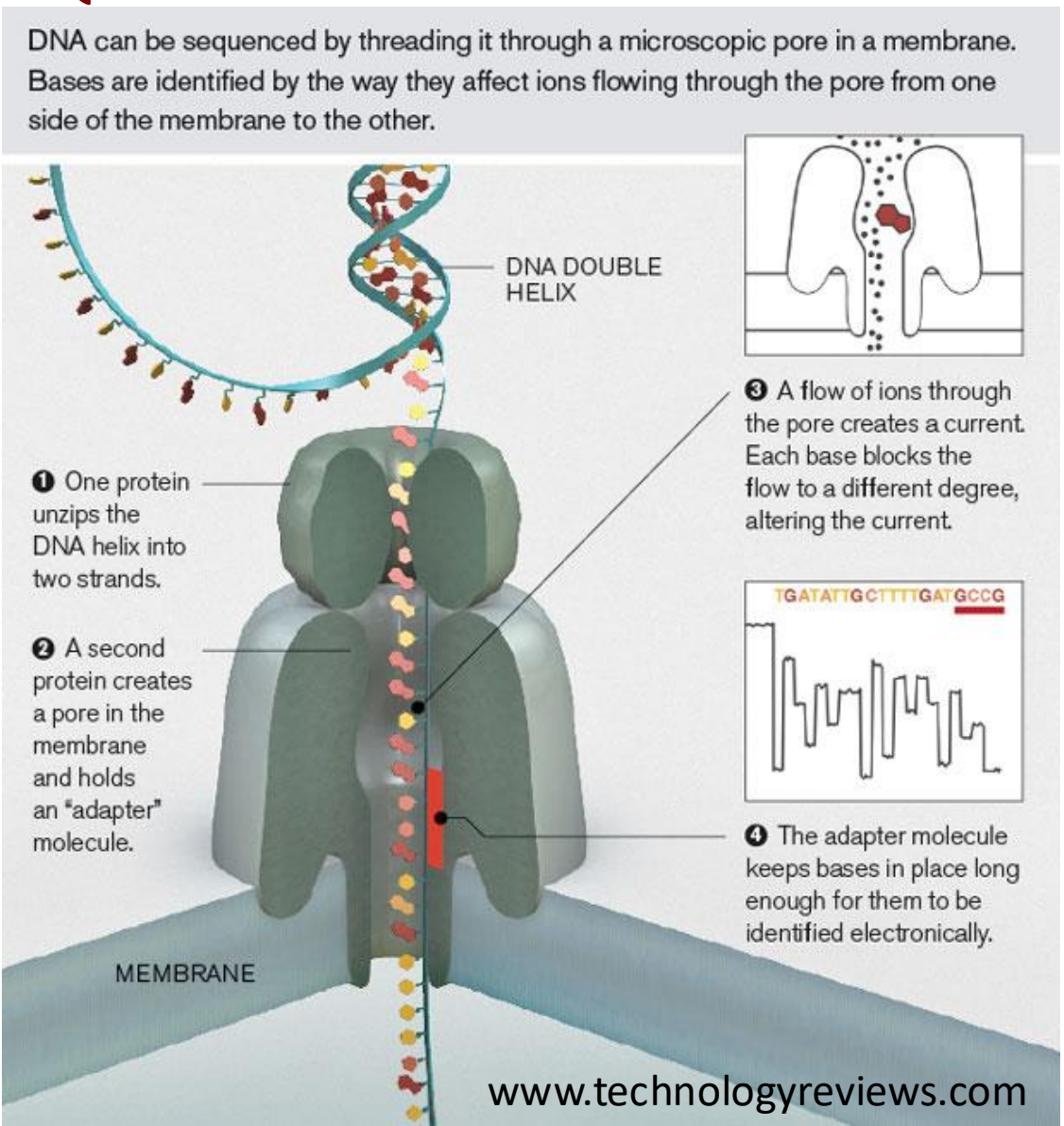
Disadvantages:

- Robustness
- More error-prone
- High machine cost

Metzker et al., Nat. Rev. Genet., 2010

The revolution of high-throughput sequencing: 3rd and 4th generation technologies: *Nanopore Sequencing*

Q5b



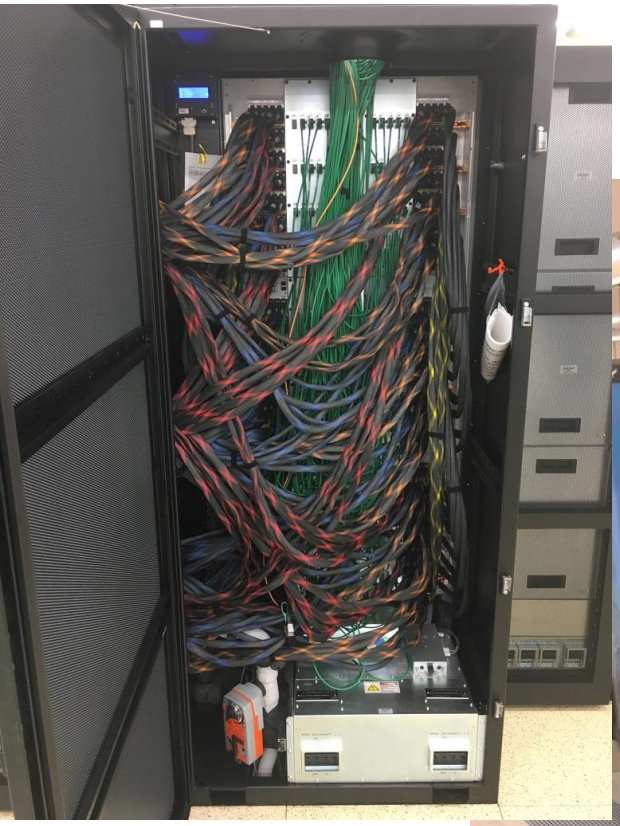
Nanopore Minion

The sequencing revolution – Reality
”Biology is (becoming) a data science

The sequencing revolution – Reality

Swiss National SuperComputing Center
(October 2021; Piz Daint)

8th on the TOP500 ranking of supercomputers until the end of 2015, higher than any other supercomputer in Europe



Human genome is largest genome so far sequenced

A. Yes

B. Yes, but others are almost as large

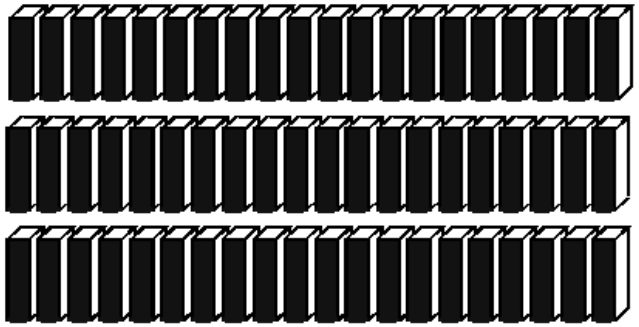
C. No, we have found genomes that are 10-fold larger or more

D. Don't know

60

Genomics by the numbers

Human Genome
Mouse Genome



???

Fruit Fly Genome

~160,000,000 bp

Nematode Genome

~100,000,000 bp

Yeast Genome

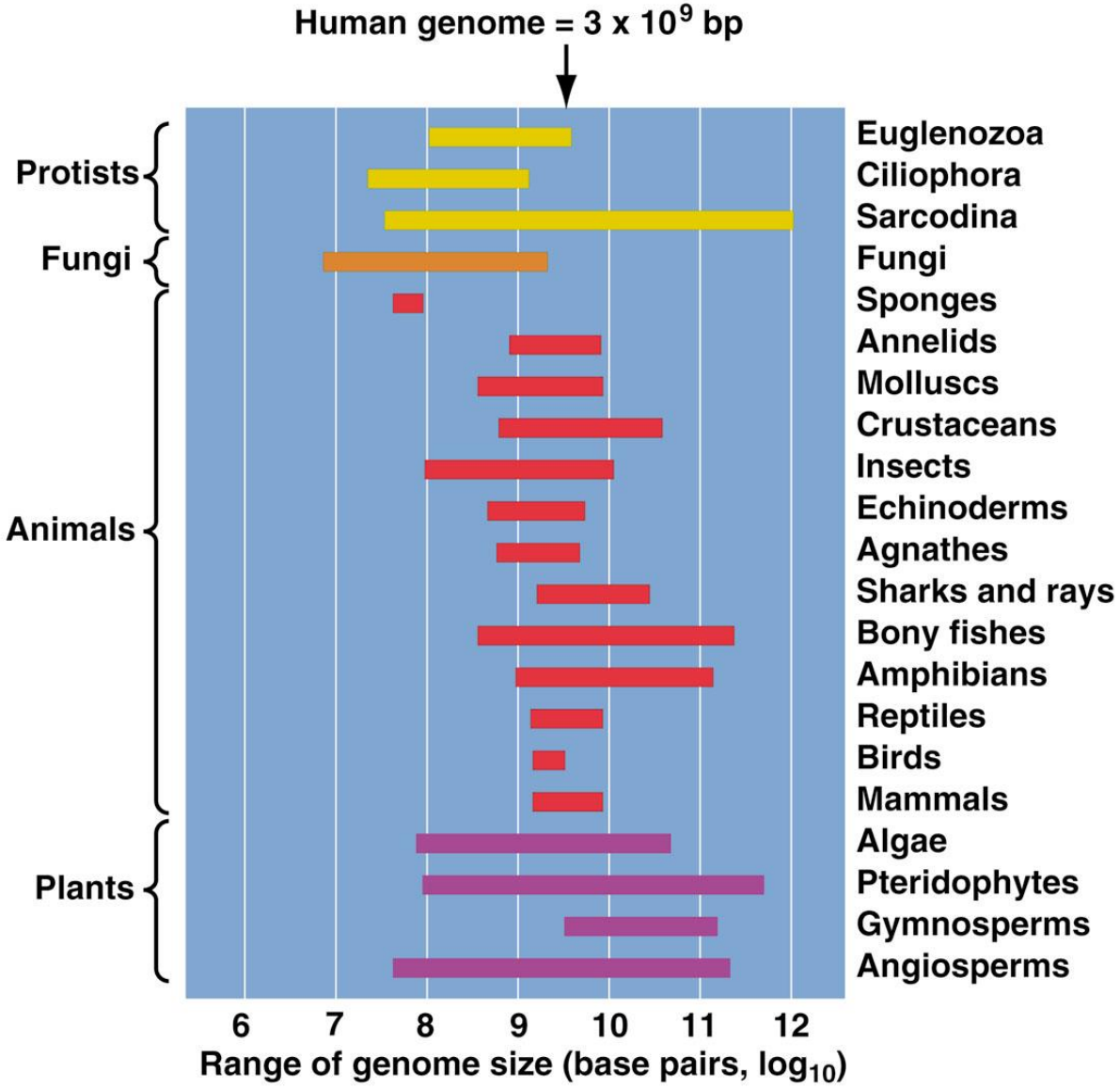
~15,000,000 bp

E. coli Genome

~5,000,000 bp

Genomics by the numbers

Q6a



No correlation between complexity and genome size

Genomics by the numbers

~3,000 bp (0.0001%) of Human Genome Sequence

TGCCGCGGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAAATGCACCTTGCAAACGTAACAGGAACCCGA
CTAGGATCATCGGGAAAAGGAGGAGGAGGAGGAAGGCAGGCTCCGGGGAAGCTGGTGGCAGCGGGTCCTGGGTCTGGCGGACCCTGACGCGAAGGA
GGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCCTCCAGCGTTGCCAACTGGACCTAAAGAGAGGCCGCGACTGTCGCCC
ACCTGCGGGATGGGCCTGGTGCTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGCGAGGGAGGGAGGCTGGGAGTCAGAATCGGGAAAGGGA
GGTGCGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCGTAGTGGGTGGAGAAAGCCGCTAGAGCAA
ATTTGGGGCCGGACCAGGCAGCACTCGGCTTTTAACCTGGGCAGTGAAGGCGGGGGAAAGAGCAAAAGGAAGGGGTGGTGTGCGGAGTAGGGGTGGG
TGGGGGGAATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAGGTCTTTGGCATTAGGAGCTTGA
GCCCAGACGGCCCTAGCAGGGACCCCAGCGCCCGAGAGACCATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGTTGTCTCCAAACTTTTTTCAGGTGAGA
AGGTGGCCAACCGAGCTTCGGAAAGACACGTGCCCACGAAAGAGGAGGGCGTGTGTATGGGTTGGGTTTGGGGTAAAGGAATAAGCAGTTTTTAAAAA
GATGCGCTATCATTCATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTAAGAAGAGATGGAAGAATGAACTGAAGCTGATTGAAT
AGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACTATGCACTTGTTTTATTTCATTTTTCTAAGAAACTAAAAATACTT
GTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCCCTTCATGCCTTGGACACTTGATTGTCTTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTG
CTCAGAAAACATTTCTTGACTGAATTCAGCCAACAAAAATTTTGGGGTAGGTAGAAAATATATGCTTAAAGTATTTATTGTTATGAGACTGGATATATCTAGTA
TTTGTCACAGGTAAATGATTCTTCAAAAATTGAAAGCAAATTTGTTGAAATATTTATTTTGAAAAAAGTTACTTCACAAGCTATAAATTTTAAAAGCCATAGG
AATAGATACCGAAGTTATATCCAACTGACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACTTTAATGAGATTTTTTAAAAT
AGCATCTAAGTTCGGAATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCCTGTTATTATACTGG
TGTGAATGAATGAATAGGTACTGCTCTCTCTTGGGACATTACTTGACACATAATTACCCAATGAATAAGCATACTGAGGTATCAAAAAAGTCAAATATGTTATA
AATAGCTCATATATGTGTGTAGGGGGGAAGGAATTTAGCTTTCACATCTCTCTTATGTTTAGTTCTCTGCATGTGCAGTTAATCCTGGAACTCCGGTGCTAAGG
AGAGACTGTTGGCCCTTGAAGGAGAGCTCCTCCCTGTGGATGAGAGAGAAGGACTTTACTCTTTGGAATTATCTTTTTGTGTTGATGTTATCCACCTTTTGTT
ACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCCTAGTCCTTATAAAGTCAAAATGTTAATTGGCATAAATTATAGACTTTTTTTAGCAGAGAACTTT
GAGGAACCTAAATGCCAACCAGTCTAAAAATGCAGTTTTCAGAAGAATGAATATTTCATGGATAGTTCTAAATACTAATGAACTTTAAAATAGCTTACTATTG
ATCTGTCAAAGTGGGTTTTTATATAATTTTCTTTTTACAAATCACCTGACACATTTAATATAGGTTAAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTAT
TACAAAGGCTGCTAAGTGTGTTAAGAGCATACTCATTTCTGTTCTCCAAAATATTTCATAAGGTGCTTTAAGAATAGGTATGTTTTTAAAAGTTAAGTTCCTAC
TATTTATAGGAACTGACAATCACCTAAAATACCAATGATTACAAACTTCCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGTAAAAAACATATTTTCTGCATT
AAGTTAGGCAGTATTGCTTAGTTTTCAAAGTGGTAGGCTTTGGAGTCAGATTATTTTGATTCAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGC
AGGTCCCTTAACATCTCTGTGTGTGACTTGACCTTTAAAATTTGGAGACTGTCATAGGGGTTAATCCCTTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTT
AACTGCTATTATTATGGATTACCATATTTTCACATTCATCACAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTTGAGTATAATTTTGTGACTCTCAA
TCTGGATATGCAATGAGTGGGCCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTCACATGGCCTTACCAGATATACAGGAAACACGTCACATGTTTCTATT
GTATGTTGTTAAATGCCTTAGAATTTAACTTTCTGAATAGGATCCCTTCAGTTTGAGAGTCATAAAAGAGTAAAATTATTATGGTAT

Genomics by the numbers

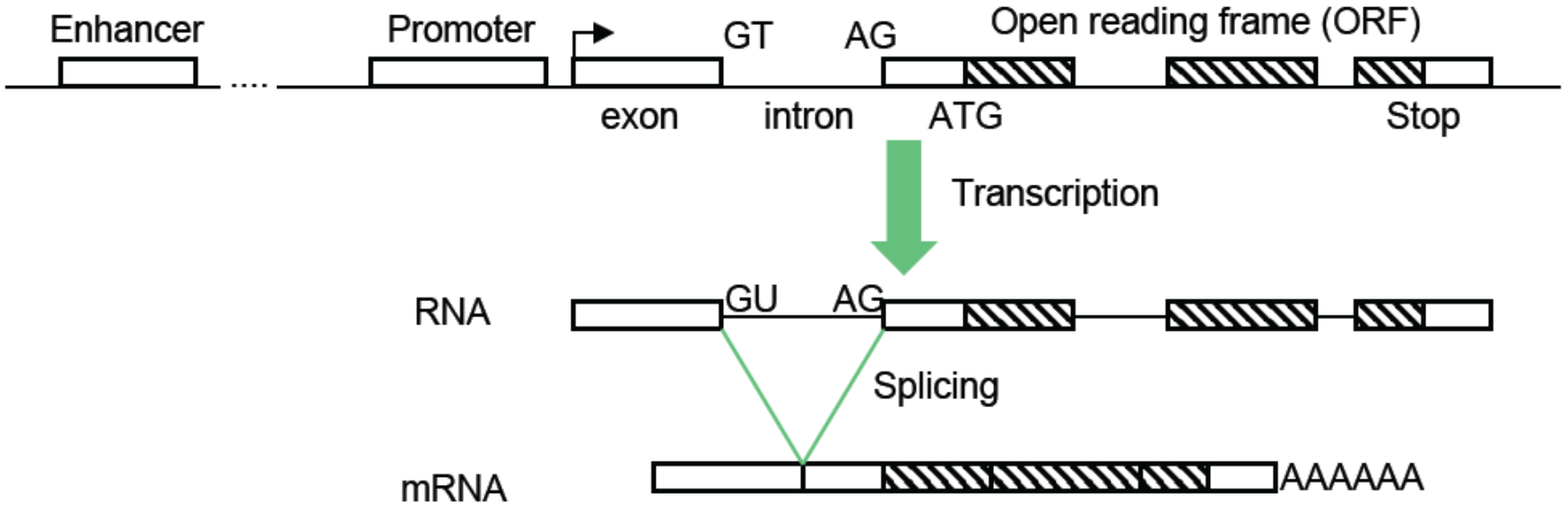
~3,000 bp (0.0001%) of Human Genome Sequence

TGCCGCGGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAAATGCACCTTGCAAACGTAACAGGAACCCGA
CTAGGATCATCGGGAAAAGGAGGAGGAGGAGGAAGGCAGGCTCCGGGGAAGCTGGTGGCAGCGGGTCCTGGGTCTGGCGGACCCTGACGCGAAGGA
GGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCCTCCAGCGTTGCCAACTGGACCTAAAGAGAGGCCGCGACTGTCGCCC
ACCTGCGGGATGGGCCTGGTGCTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGCGAGGGAGGGAGGCTGGGAGTCAGAATCGGGAAAGGGA
GGTGCGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCGTAGTGGGTGGAGAAAGCCGCTAGAGCAA
ATTTGGGGCCGGACCAGGCAGCACTCGGCTTTTAACCTGGGCAGTGAAGGCGGGGGAAAGAGCAAAAGGAAGGGGTGGTGTGCGGAGTAGGGGTGGG
TGGGGGGAATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAGGTCTTTGGCATTAGGAGCTTGA
GCCCAGACGGCCCTAGCAGGGACCCCAGCGCCCGAGAGACCATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGTTGTCTCCAAACTTTTTTTCAGGTGAGA
AGGTGGCCAACCGAGCTTCGGAAAGACACGTGCCCACGAAAGAGGAGGGCGTGTGTATGGGTTGGGTTTGGGGTAAAGGAATAAGCAGTTTTTAAAAA
GATGCGCTATCATTCATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTAAGAAGAGATGGAAGAATGAACTGAAGCTGATTGAAT
AGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACTATGCACTTGTTTTATTTCATTTTTCTAAGAAACTAAAAATACTT
GTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCCCTTCATGCCTTGGACACTTGATTGTCTTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTG
CTCAGAAAACATTTCTTGACTGAATTCAGCCAACAAAAATTTTGGGGTAGGTAGAAAATATATGCTTAAAGTATTTATTGTTATGAGACTGGATATATCTAGTA
TTTGTCACAGGTAAATGATTCTTCAAAAATTGAAAGCAAATTTGTTGAAATATTTATTTTGAAAAAAGTTACTTCACAAGCTATAAATTTTAAAAGCCATAGG
AATAGATACCGAAGTTATATCCAACTGACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACTTTAATGAGATTTTTTAAAAT
AGCATCTAAGTTCGGAATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCCTGTTATTATACTGG
TGTGAATGAATGAATAGGTACTGCTCTCTCTTGGGACATTACTTGACACATAATTACCCAATGAATAAGCATACTGAGGTATCAAAAAAGTCAAATATGTTATA
AATAGCTCATATATGTGTGTAGGGGGGAAGGAATTTAGCTTTCACATCTCTCTTATGTTTAGTTCTCTGCATGTGCAGTTAATCCTGGAACTCCGGTGCTAAGG
AGAGACTGTTGGCCCTTGAAGGAGAGCTCCTCCCTGTGGATGAGAGAGAAGGACTTTACTCTTTGGAATTATCTTTTTGTGTTGATGTTATCCACCTTTTGTT
ACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCCTAGTCCTTATAAAGTCAAAATGTTAATTGGCATAAATTATAGACTTTTTTTAGCAGAGAACTTT
GAGGAACCTAAATGCCAACCAGTCTAAAAATGCAGTTTTCAGAAGAATGAATATTTCATGGATAGTTCTAAATACTAATGAACTTTAAAATAGCTTACTATTG
ATCTGTCAAAGTGGGTTTTTATATAATTTTCTTTTTACAAATCACCTGACACATTTAATATAGGTTAAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTAT
TACAAAGGCTGCTAAGTGTGTTAAGAGCATACTCATTTCTGTTCTCCAAAATATTTCATAAGGTGCTTTAAGAATAGGTATGTTTTTAAAAGTTAAGTTCCTAC
TATTTATAGGAACTGACAATCACCTAAAATACCAATGATTACAAACTTCCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGTAAAAAACATATTTTCTGCATT
AAGTTAGGCAGTATTGCTTAGTTTTCAAAGTGGTAGGCTTTGGAGTCAGATTATTTTGATTCAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGC
AGGTCCCTTAACATCTCTGTGTGTGACTTGACCTTTAAAATTTGGAGACTGTCATAGGGGTTAATCCCTTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTT
AACTGCTATTATTATGGATTACCATATTTTCACATTCATCACAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTTGAGTATAATTTTGTGACTCTCAA
TCTGGATATGCAATGAGTGGGCCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTCACATGGCCTTACCAGATATACAGGAAACACGTCACATGTTTCTATT
GTATGTTGTTAAATGCCTTAGAATTTAACTTTCTGAATAGGATCCCTTCAGTTTGAGAGTCATAAAAGAGTAAAATTATTATGGTAT

~5% of Human Genome Sequence is Constrained Across Mammals (and Presumed Functional) ; 5% of 3B Bases = ~150M Bases

Genomics by the numbers

The signature of a gene



Genomics by the numbers

~3,000 bp (0.0001%) of Human Genome Sequence

TGCCGCGGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAAATGCACCTTGCAAACGTAACAGGAACCCGA
CTAGGATCATCGGGAAAAGGAGGAGGAGGAGGAAGGCAGGCTCCGGGGAAGCTGGTGGCAGCGGGTCCTGGGTCTGGCGGACCCTGACGCGAAGGA
GGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCCTCCAGCGTTGCCAACTGGACCTAAAGAGAGGCCGCGACTGTCGCCC
ACCTGCGGGATGGGCCTGGTGCTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGCGAGGGAGGGAGGCTGGGAGTCAGAATCGGGAAAGGGA
GGTGCGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCGTAGTGGGTGGAGAAAGCCGCTAGAGCAA
ATTTGGGGCCGGACCAGGCAGCACTCGGCTTTTAACCTGGGCAGTGAAGGCGGGGGAAAGAGCAAAAGGAAGGGGTGGTGTGCGGAGTAGGGGTGGG
TGGGGGGAATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAGGTCTTTGGCATTAGGAGCTTGA
GCCCAGACGGCCCTAGCAGGGACCCCAGCGCCCGAGAGACCATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGTTGTCTCCAAACTTTTTTTCAGGTGAGA
AGGTGGCCAACCGAGCTTCGGAAAGACACGTGCCCACGAAAGAGGAGGGCGTGTGTATGGGTTGGGTTTGGGGTAAAGGAATAAGCAGTTTTTAAAAA
GATGCGCTATCATTCATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTAAGAAGAGATGGAAGAATGAACTGAAGCTGATTGAAT
AGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACTATGCACTTGTTTTATTTCATTTTTCTAAGAAACTAAAAATACTT
GTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCCCTTCATGCCTTGGACACTTGATTGTCTTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTG
CTCAGAAAACATTTCTTGACTGAATTCAGCCAACAAAAATTTTGGGGTAGGTAGAAAATATATGCTTAAAGTATTTATTGTTATGAGACTGGATATATCTAGTA
TTTGTCACAGGTAAATGATTCTTCAAAAATTGAAAGCAAATTTGTTGAAATATTTATTTTGAAAAAAGTTACTTCACAAGCTATAAATTTTAAAAGCCATAGG
AATAGATACCGAAGTTATATCCAACTGACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACTTTAATGAGATTTTTTAAAAT
AGCATCTAAGTTCGGAATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCCTGTTATTATACTGG
TGTGAATGAATGAATAGGTACTGCTCTCTCTTGGGACATTACTTGACACATAATTACCCAATGAATAAGCATACTGAGGTATCAAAAAAGTCAAATATGTTATA
AATAGCTCATATATGTGTGTAGGGGGGAAGGAATTTAGCTTTCACATCTCTCTTATGTTTAGTTCTCTGCATGTGCAGTTAATCCTGGAACTCCGGTGCTAAGG
AGAGACTGTTGGCCCTTGAAGGAGAGCTCCTCCCTGTGGATGAGAGAGAAGGACTTTACTCTTTGGAATTATCTTTTTGTGTTGATGTTATCCACCTTTTGTT
ACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCCTAGTCCTTATAAAGTCAAAATGTTAATTGGCATAAATTATAGACTTTTTTTAGCAGAGAACTTT
GAGGAACCTAAATGCCAACCAGTCTAAAAATGCAGTTTTCAGAAGAATGAATATTTCATGGATAGTTCTAAATACTAATGAACTTTAAAATAGCTTACTATTG
ATCTGTCAAAGTGGGTTTTTATATAATTTTCTTTTTACAAATCACCTGACACATTTAATATAGGTTAAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTAT
TACAAAGGCTGCTAAGTGTGTTAAGAGCATACTCATTTCTGTTCTCCAAAATATTTCATAAGGTGCTTTAAGAATAGGTATGTTTTTAAAAGTTAAGTTCCTAC
TATTTATAGGAACTGACAATCACCTAAAATACCAATGATTACAAACTTCCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGTAAAAAACATATTTTCTGCATT
AAGTTAGGCAGTATTGCTTAGTTTTCAAAGTGGTAGGCTTTGGAGTCAGATTATTTGATTCAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGC
AGGTCCCTTAACATCTCTGTGTGTGACTTGACCTTTAAAATTTGGAGACTGTCATAGGGGTTAATCCCTTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTT
AACTGCTATTATTATGGATTACCATATTTTCACATTCATCACAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTTGAGTATAATTTTGTGACTCTCAA
TCTGGATATGCAATGAGTGGGCCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTCACATGGCCTTACCAGATATACAGGAAACACGTCACATGTTTCTATT
GTATGTTGTTAAATGCCTTAGAATTTAACTTTCTGAATAGGATCCCTTCAGTTTGAGAGTCATAAAAGAGTAAAATTATTATGGTAT

~1.5% Encodes for Protein (Genes) ; Corresponds to ~18-22K Genes ; Many More than ~22K Different Proteins ; Good Inventory at Present

No correlation between complexity and number of genes

Q6b

Species and Common Name	Estimated Total Size of Genome (bp)*	Estimated Number of Protein-Encoding Genes*
Saccharomyces cerevisiae (unicellular budding yeast)	12 million	6,000
Trichomonas vaginalis	160 million	60,000
Plasmodium falciparum (unicellular malaria parasite)	23 million	5,000
Caenorhabditis elegans (nematode)	95.5 million	18,000
Drosophila melanogaster (fruit fly)	170 million	14,000
Arabidopsis thaliana (mustard; thale cress)	125 million	25,000
Oryza sativa (rice)	470 million	51,000
Gallus gallus (chicken)	1 billion	20,000-23,000
Canis familiaris (domestic dog)	2.4 billion	19,000
Mus musculus (laboratory mouse)	2.5 billion	30,000
Homo sapiens (human)	2.9 billion	20,000-25,000

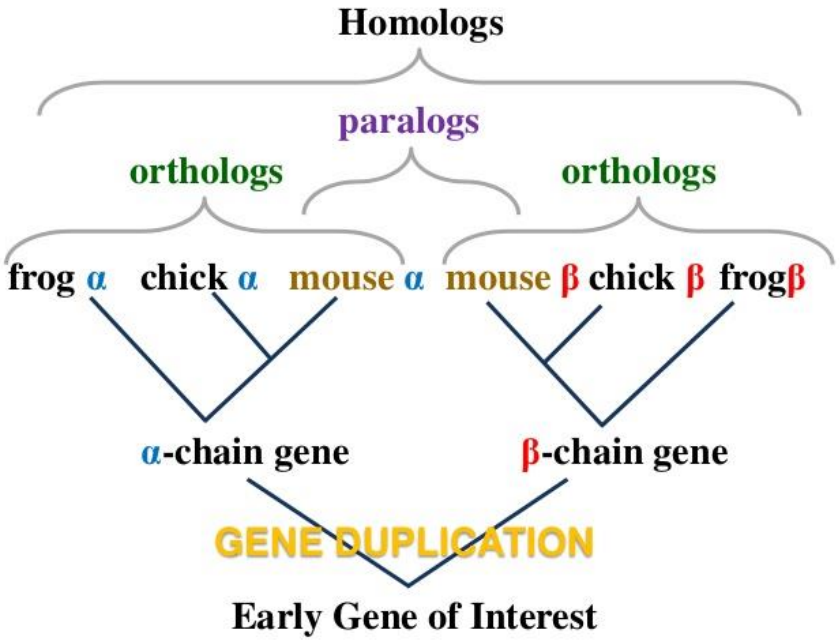
Pray, L. (2008) Eukaryotic genome complexity. Nature Education 1(1)

Comparative Genomics

Q7

Homolog (qualitative term → % identity, not % homology)

A gene related to a second gene by descent from a common ancestral DNA sequence. The term, homolog, may apply to the relationship between genes separated by the event of **speciation** (see ortholog) or to the relationship between genes separated by the event of genetic **duplication** (see paralog).



Ortholog:

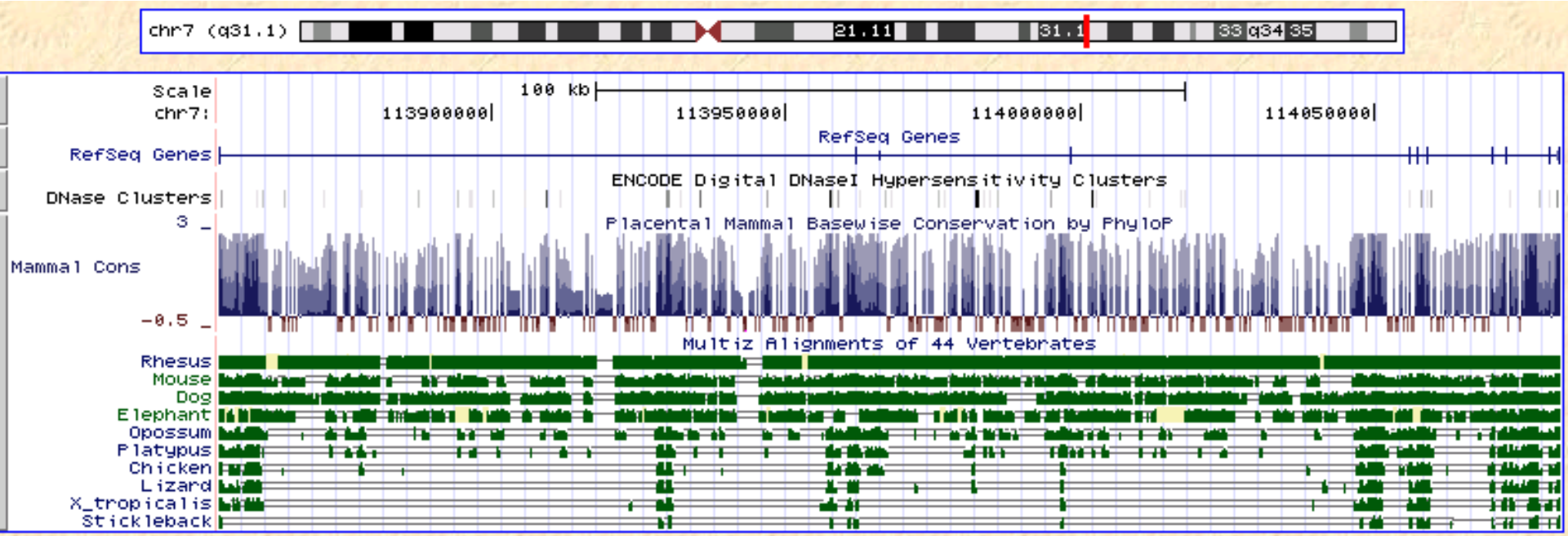
Genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes. (See also Paralogs.).

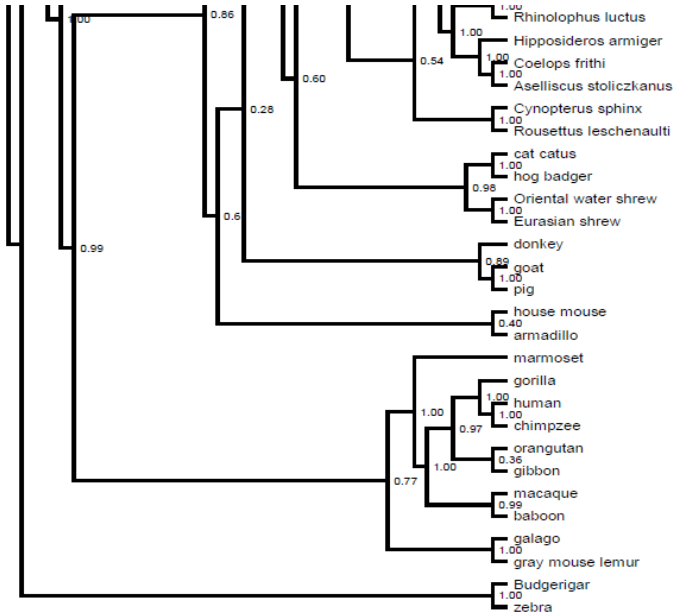
Paralog:

Genes related by duplication often (but not always) within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

Comparative Genomics: a test case → FOXP2

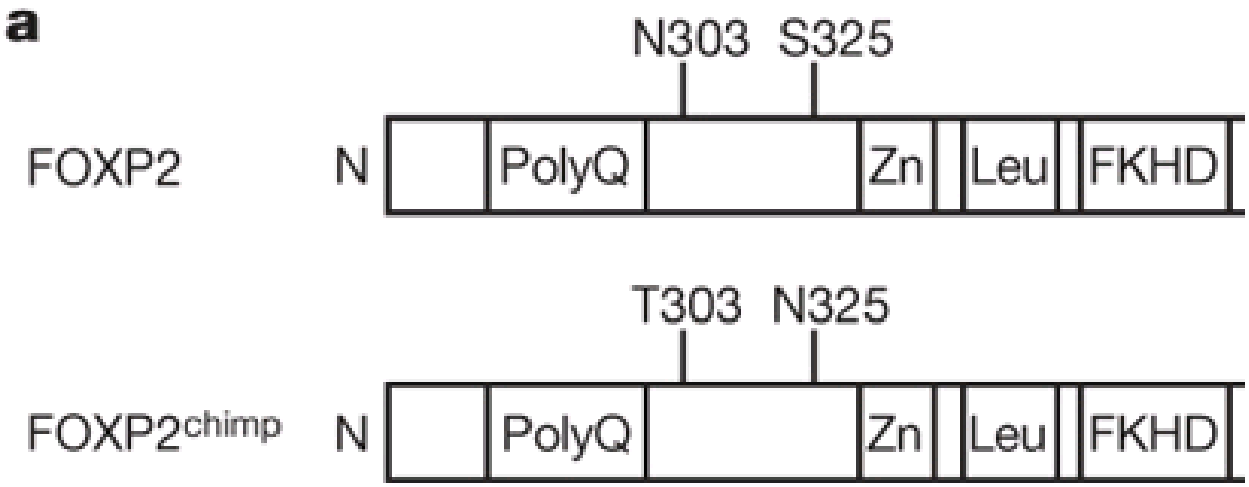
Gene alignments: FOXP2





Comparative Genomics: a test case → FOXP2

Bonkowsky and Chien, Dev. Dyn., 2005



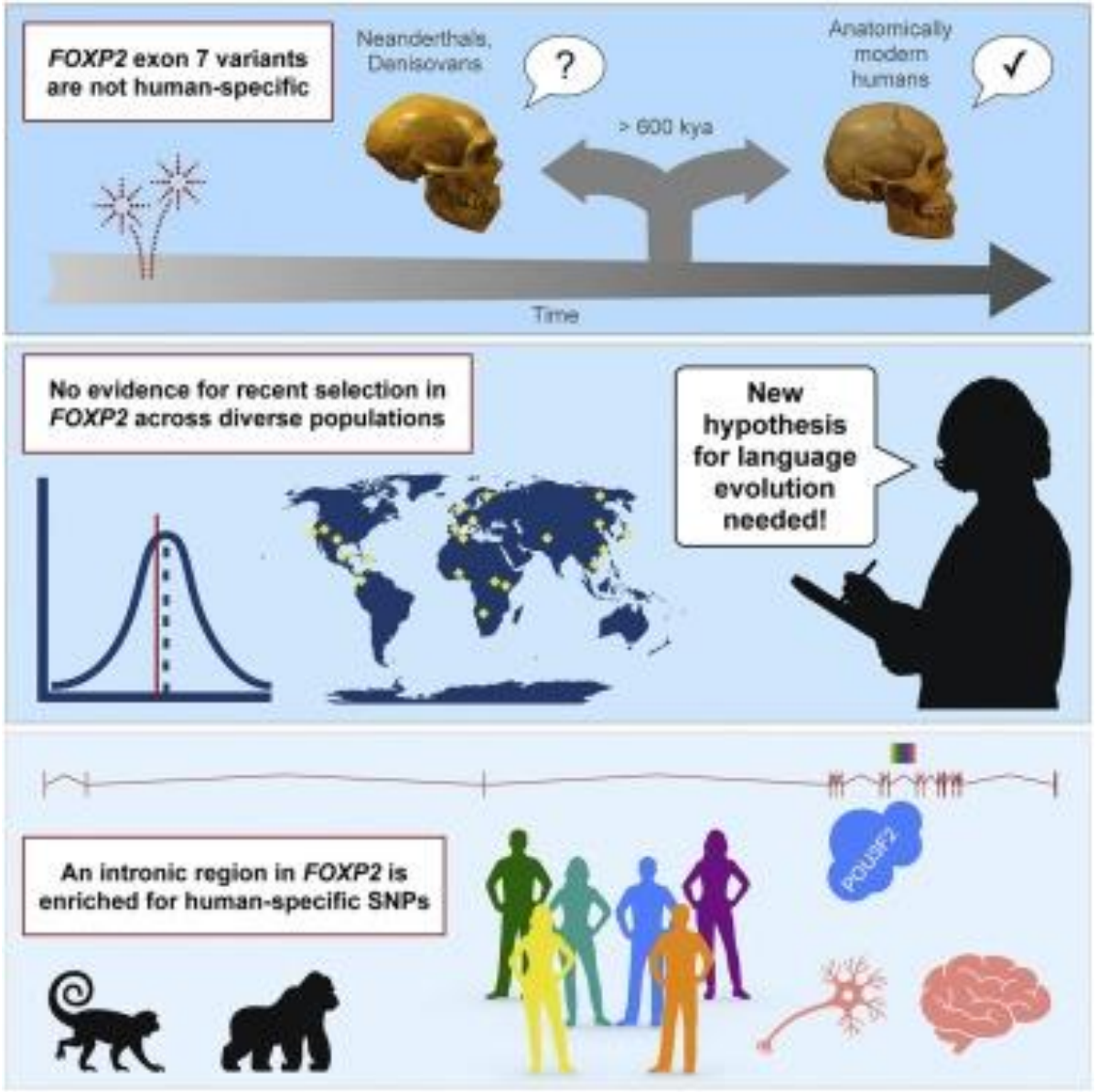
- Gene mutation in people with facial motor control and mental processing of language
- Proposal that aa composition of human FOXP2 has undergone accelerated evolution, with change occurring around the time of language emergence in human
- Insertion of both human mutations into mice causes changes in vocalizations as well as other behavioral changes, such as a reduction in exploratory tendencies

Q8

Konopka et al., Nature, 2009

Comparative Genomics: a test case → FOXP2

BUT: …..!!



New hypothesis: a link to differential gene expression?

Atkinson et al., Cell, 2018

Nice example of how research findings are not set in stone and evolve themselves based on new / more data

Genomics by the numbers

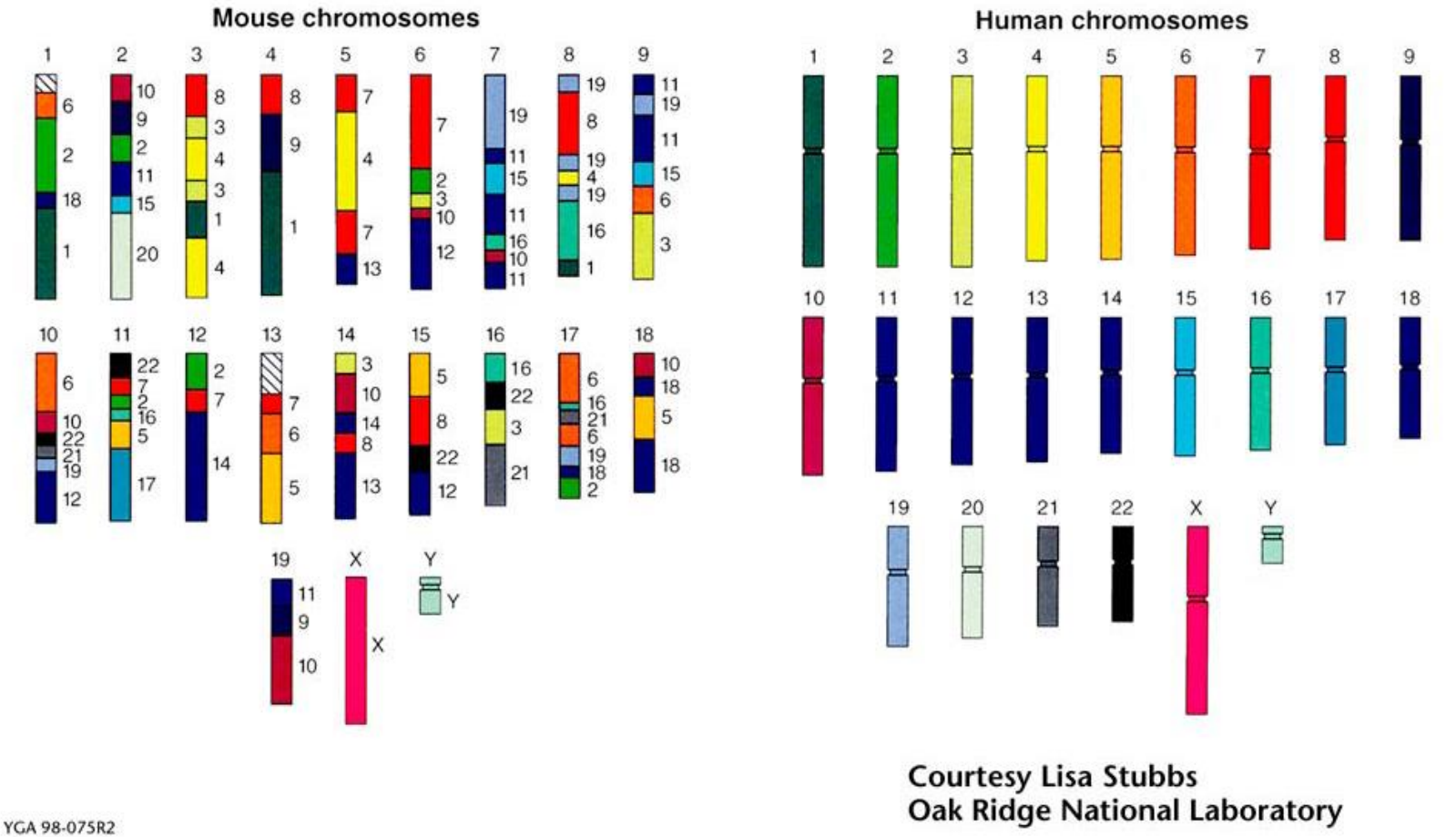
How unique are humans?

- Large numbers of genes are in common with other organisms "homologues"
  - ~50% of our genes are also found in *Drosophila*
  - ~40% of our genes are also found in *C. elegans*
  - ~30% of our genes are also found in yeast
  - ~80% of our genes are shared with the mouse
  - ~96% of our genes are shared with chimpanzees
  - ~100 of our genes are even shared with bacteria

Genomics by the numbers

Q9

Co-localization of genes on chromosomes of different species = “Syntenic regions”



The “humanzee”, can apes and humans (theoretically) interbreed?

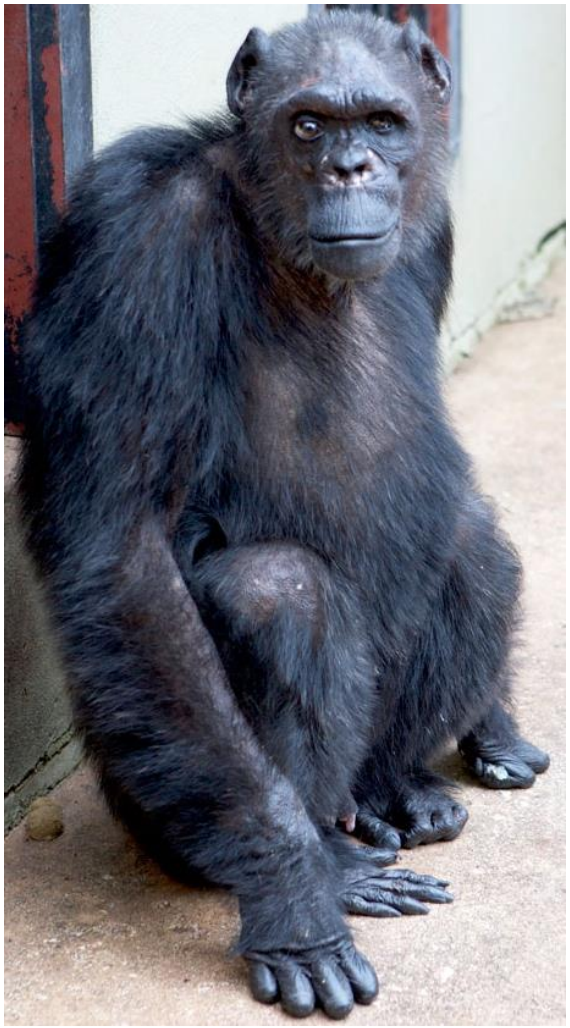
A. No, because they are genetically too distinct
B. No, because the sperm cannot penetrate the egg membrane
C. A humanzee could be generated, but would be sterile
D. Yes, both species are compatible

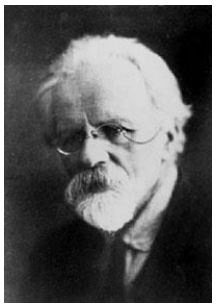
80

The humanzee

Q10

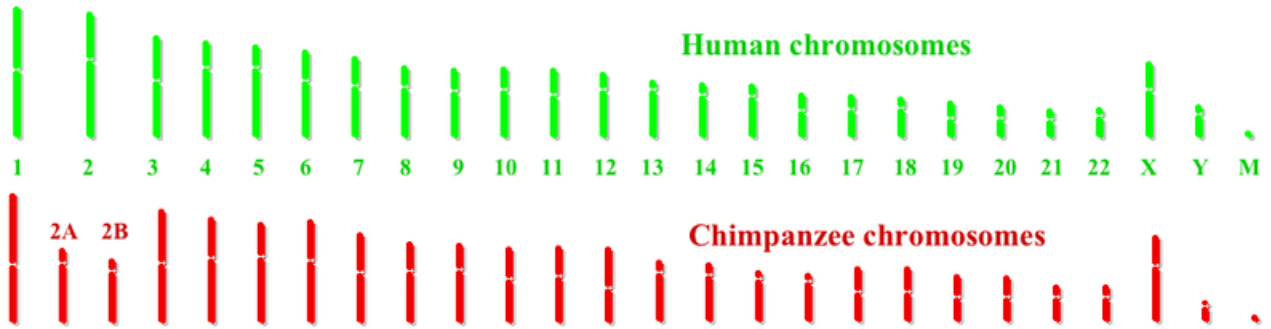
Oliver, the presumed humanzee





Ilya Ivanov attempted a chimp-human hybrid in 1920's

After humans and chimps diverged, interbreeding between the «proto»human and «proto»chimp occurred for ~1.2 million years (based of analysis of X versus automosomal chromosomes) (Patterson et al., Nature, 2006)

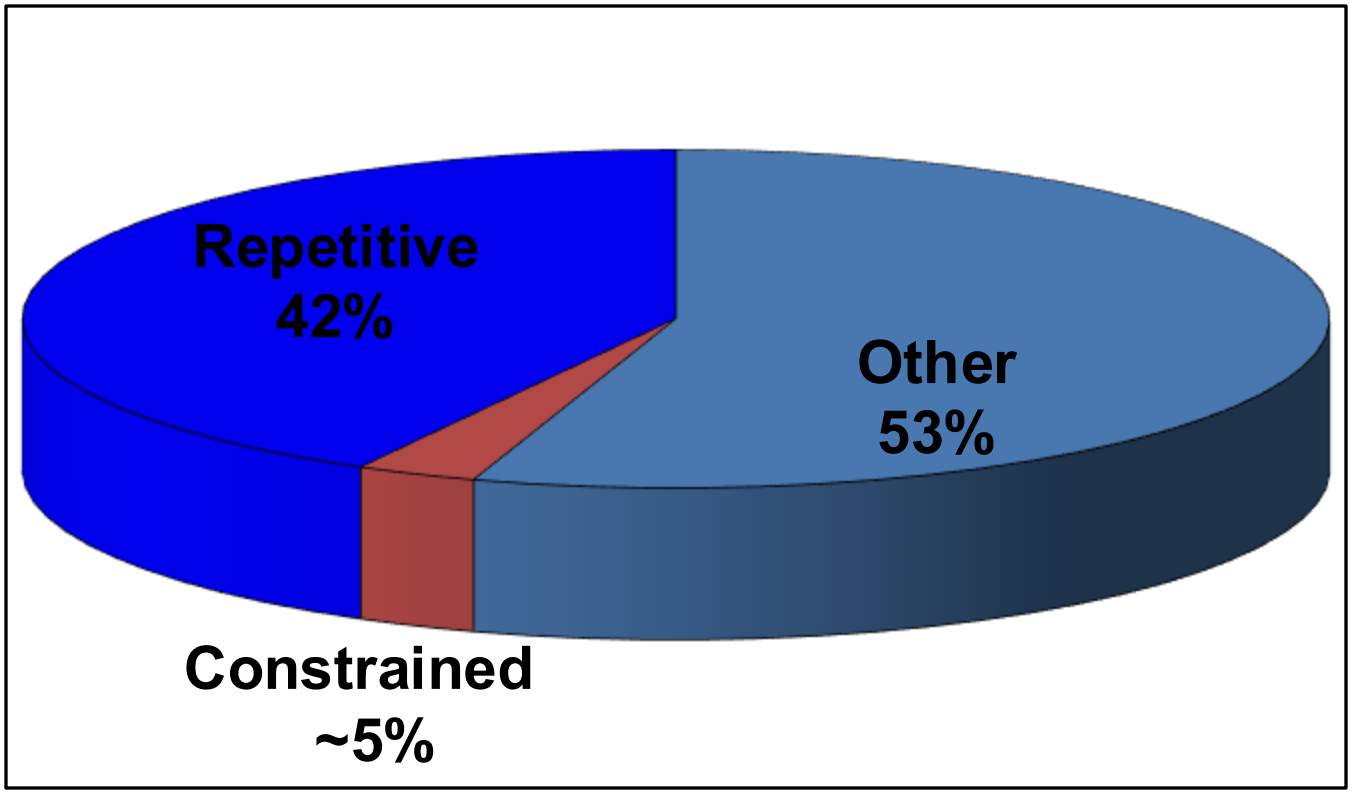


Humans have 23, apes have 24 chromosome pairs

This degree of chromosomal similarity is roughly equivalent to that found in equines (horse + donkey = mule)

What about the rest of the genome?

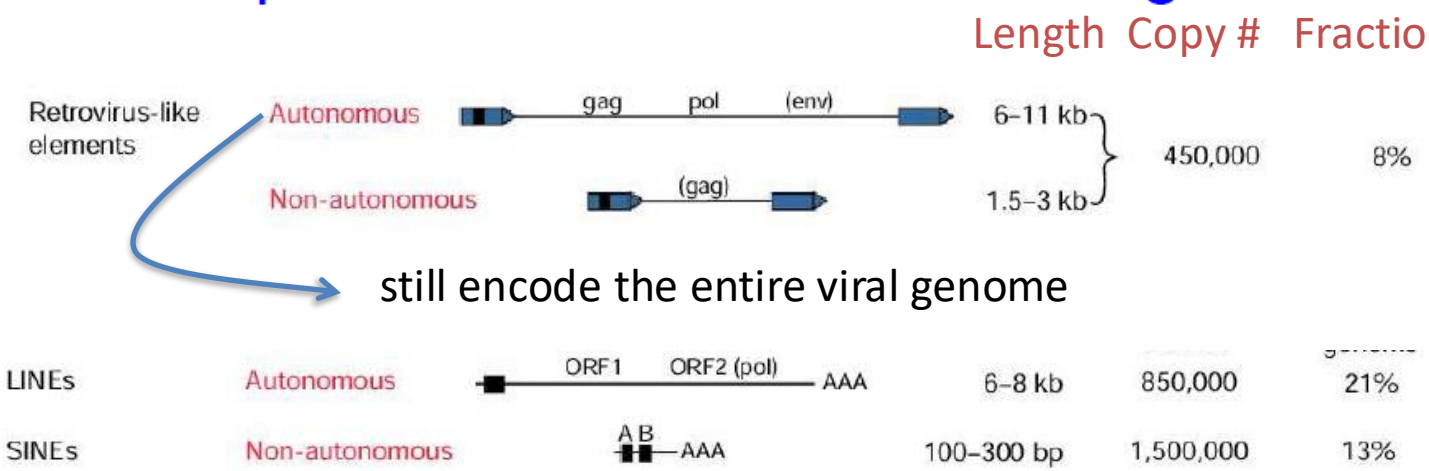
Q11



What about the rest of genome

Q12

Transposable elements in human genome



LINE elements have shed the viral body and act as selfish genes by simply staying in the host, whereas SINEs depend on LINEs to be transcribed (tiny bugs prey on small bugs who prey on big bugs etc.)

Gag is a polyprotein and is an acronym for Group Antigens (ag)
→ determines retroviral core
Pol is the reverse transcriptase
Env is the viral envelope protein

Source: Gerton Lunter, Oxford Lecture

What about the rest of genome

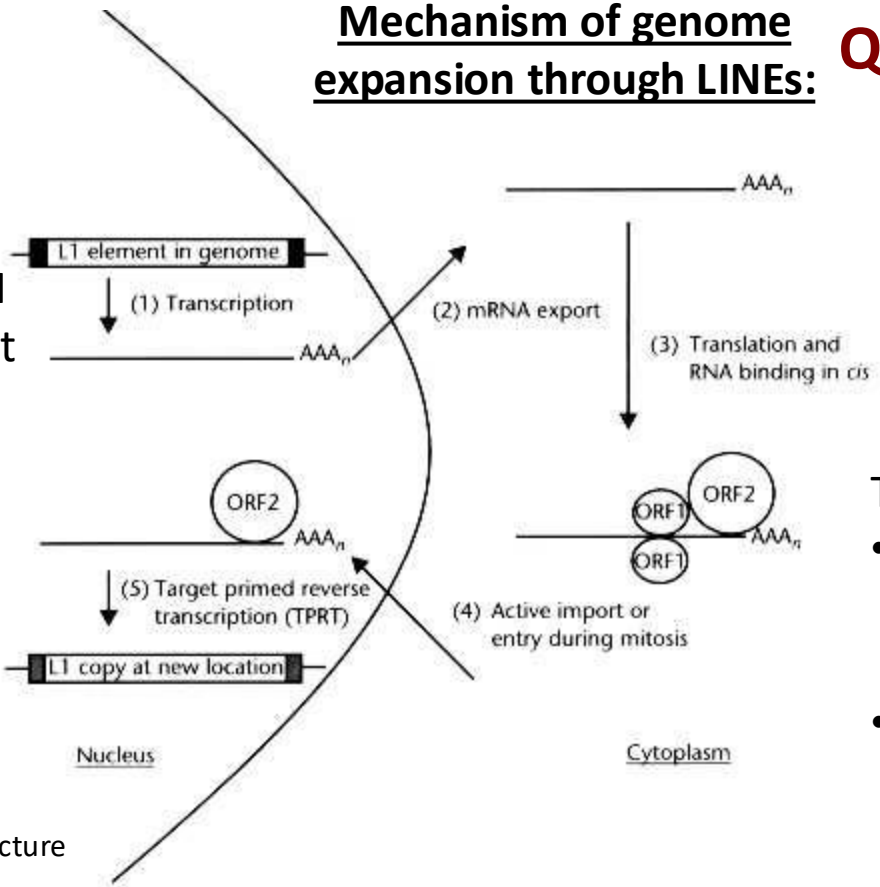
Activity of transposable elements

Activity varies greatly per organism:

- Humans: Rather quiet, ≈ 50 active LINEs, no or very few active DNA transposons, no LTRs through to be active.
- Mice: ≈ 3000 active LINEs, many active DNA transposons, many active LTRs.

Mechanism of genome expansion through LINEs: **Q13**

Through RNA pol II of the host



The ORFs code for:

- reverse transcriptase with high specificity for L1 mRNA
- Endonuclease, allowing integration in the genome

Source: Gerton Lunter, Oxford Lecture

Q12

What about the rest of genome

Transposable elements in human genome

			Length	Copy #	Fraction of genome
Retrovirus-like elements	Autonomous	gag pol (env)	6–11 kb	450,000	8%
	Non-autonomous	(gag)	1.5–3 kb		

still encode the entire viral genome

LINEs	Autonomous	ORF1 ORF2 (pol) AAA	6–8 kb	850,000	21%
SINEs	Non-autonomous	A B AAA	100–300 bp	1,500,000	13%

LINE elements have shed the viral body and act as selfish genes by simply staying in the host, whereas SINEs depend on LINEs to be transcribed
(tiny bugs prey on small bugs who prey on big bugs etc.)

DNA transposon fossils	Autonomous	transposase	2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

- DNA transposons and retrotransposons code for *transposase* (or related *integrase*). Insert double-stranded DNA into host genome.

Source: Gerton Lunter, Oxford Lecture

What about the rest of genome

In our genome:

- One LINE element (LINE1) is particularly abundant and active
- One SINE element (Alu) is particularly abundant and active
- DNA transposons have been active, but not active now.

Repeat class	Fraction of genome (%)	Copy number
LINEs	20.99	850,000
LINE1	(17.39)	
SINEs	13.64	1,500,000
Alu	(10.74)	
LTRs	8.55	450,000
DNA elements	3.03	300,000
Unclassified	0.15	
Total transposable elements	**46.36**	

Source: Gerton Lunter, Oxford Lecture

What about the rest of genome

Activity of transposable elements

Activity varies greatly per organism:

- Humans: Rather quiet, ≈ 50 active LINEs, no or very few active DNA transposons, no LTRs through to be active.
- Mice: ≈ 3000 active LINEs, many active DNA transposons, many active LTRs.

In mice, transposable elements are much more active: 10% of spontaneous mutations causing a noticeable effect are due to transposable elements.

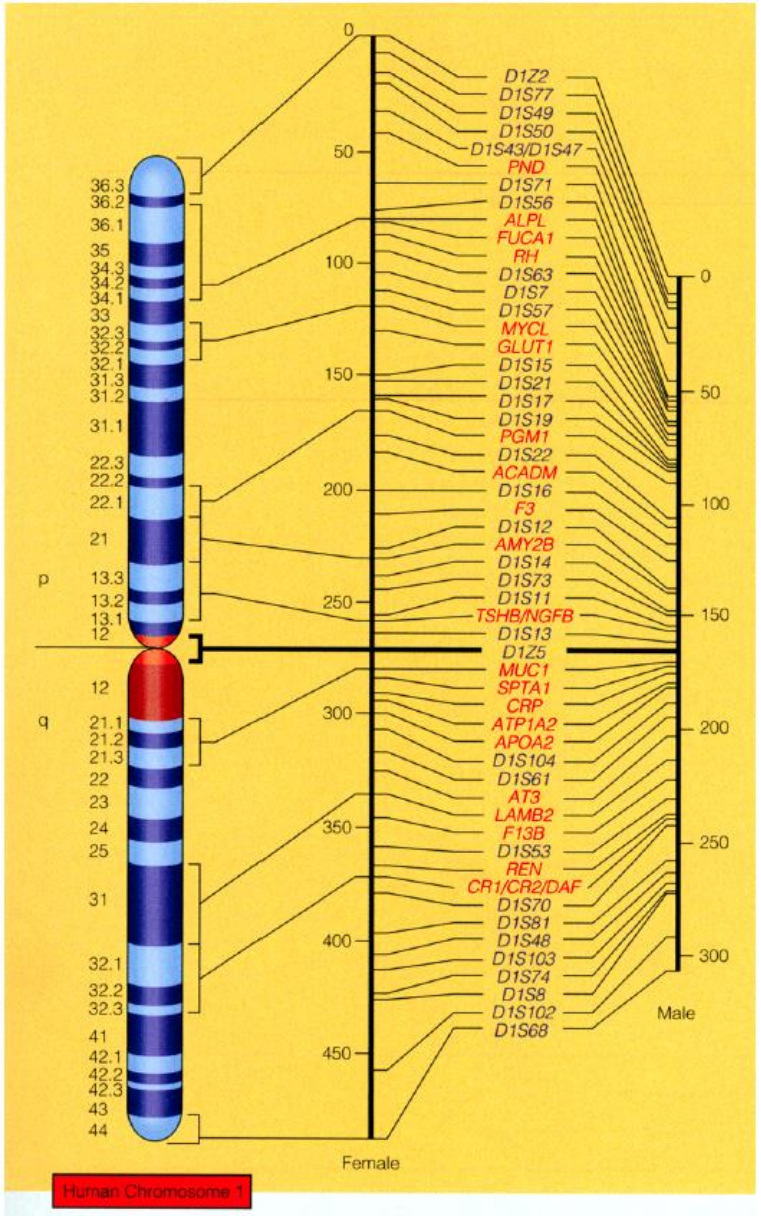
Say cheese: just one piece of 'junk' DNA can produce several coat colours in genetically identical mice.

Source: Gerton Lunter, Oxford Lecture

Botstein and his human genetic linkage map: RFLP method

petit

region 2, band 1, sub-band 3



- 30,000 markers in the human genome (100-500 bp)
- 1 genetic map unit (cM) = 1% chance that a marker at one genetic locus on a chromosome will be separated from a marker at a second locus due to crossing over in a single generation = about 1Mb in humans (physical distance)

What about the rest of genome

Repeat elements: satellites

Q14a

Satellites (micro and mini): section of repeated DNA stretches (e.g. GGGCAGG)

Class	Size of Locus	Number of Alleles	Number of Loci in Population	Rate of Mutation
Microsatellite	30–300 bp	2–10	200,000	10^{-3}
Multilocus Minisatellite	1–20 kb	2–10	30,000	10^{-3}

2-5 bp repeat unit

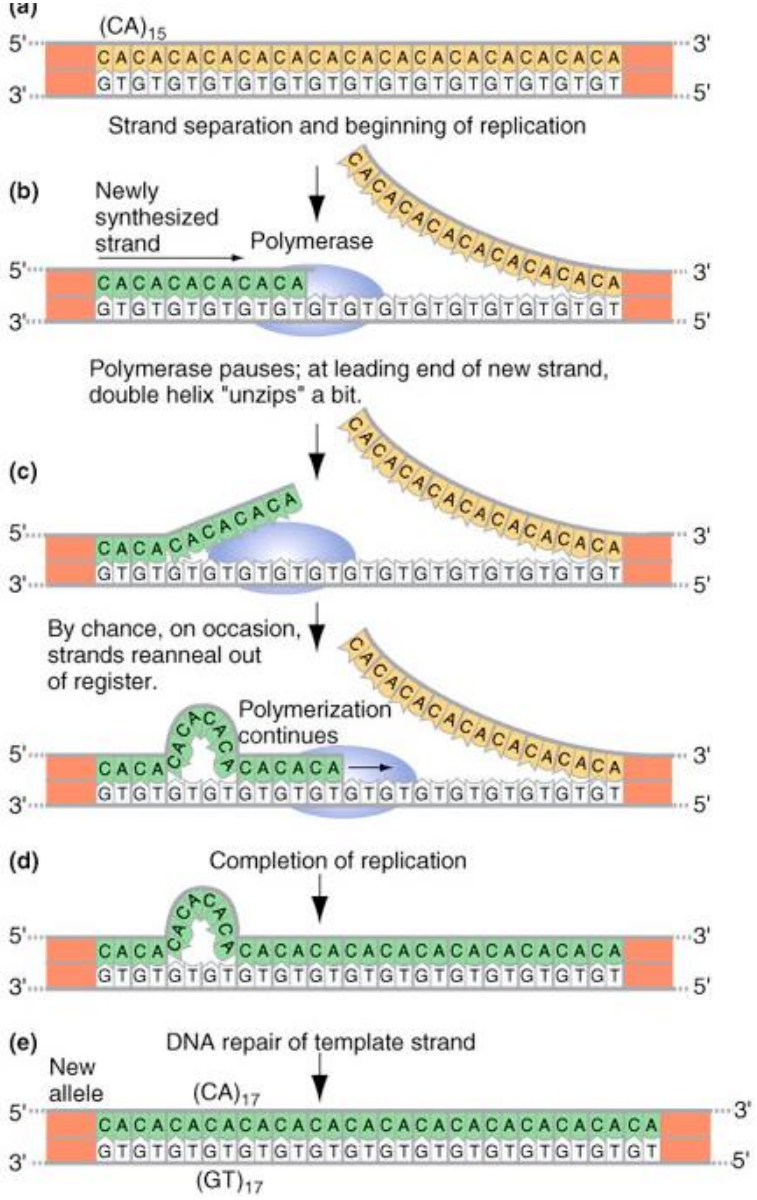
10-100 bp repeat unit

Q14b

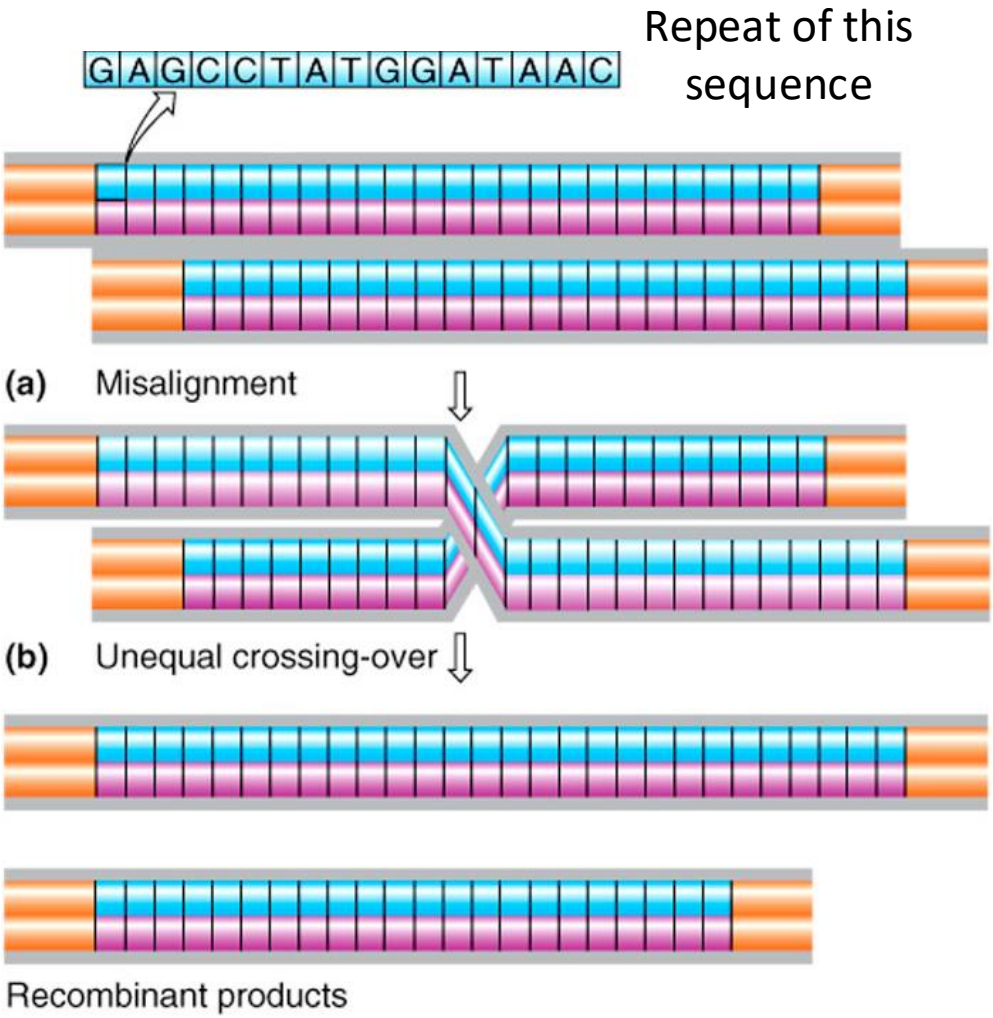
What about the rest of genome

Repeat elements

Origin of a microsatellite



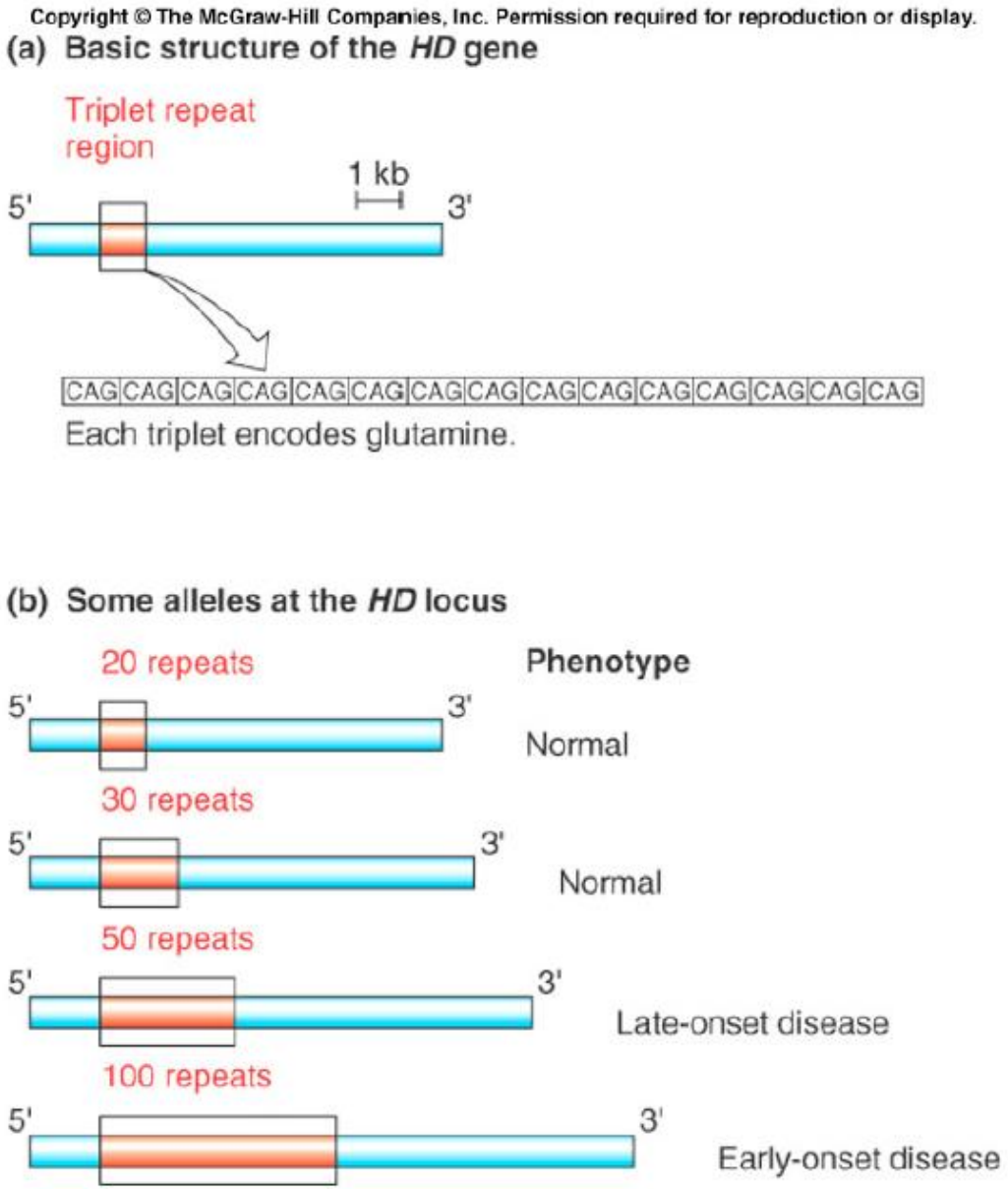
minisatellite



The two alleles have now a differentially sized minisatellite

What about the rest of genome

Repeat elements: disease



Huntington's disease is an example of a microsatellite triplet repeat in a coding region

What about the rest of genome

Repeat elements: DNA fingerprinting

Q15

1985 – Alec Jeffreys made two key findings

– Each minisatellite locus is highly polymorphic

– Most minisatellites occur at multiple sites around the genome

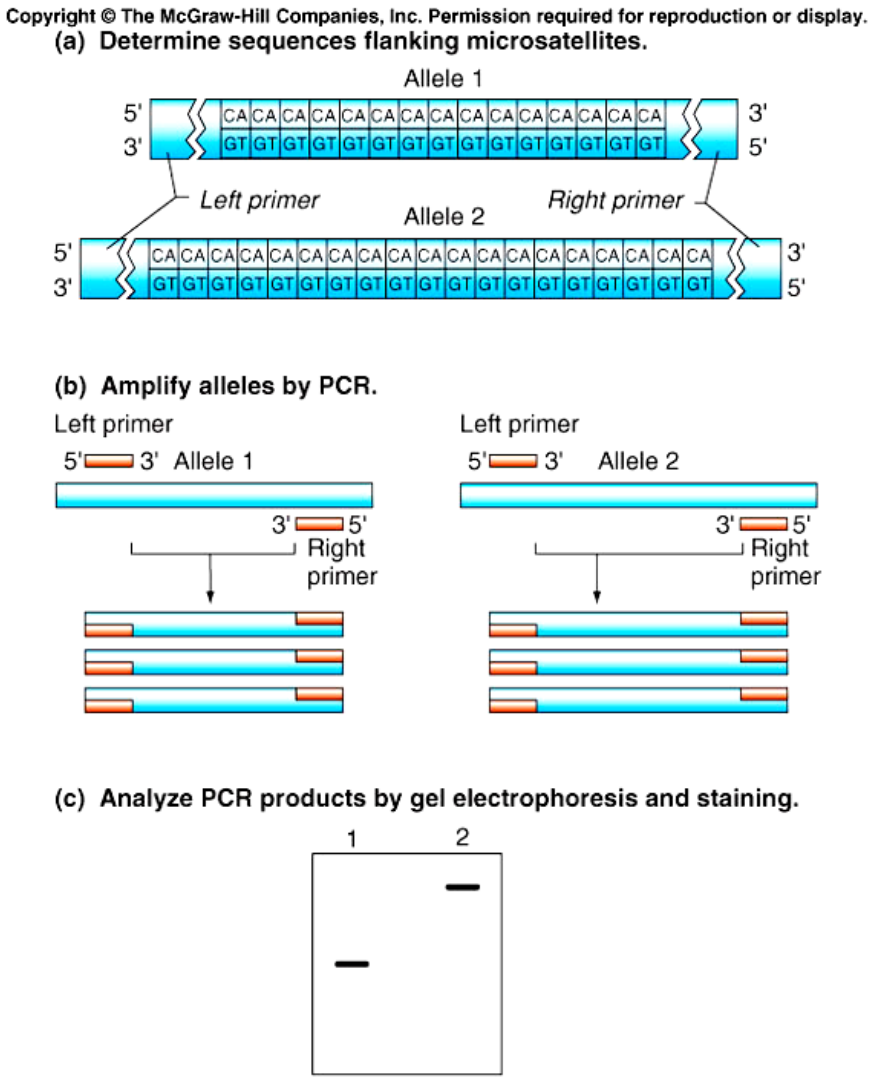
Thus, they could be used to generate a "DNA fingerprint", i.e. a pattern of simultaneous genotypes at a group of unlinked loci:

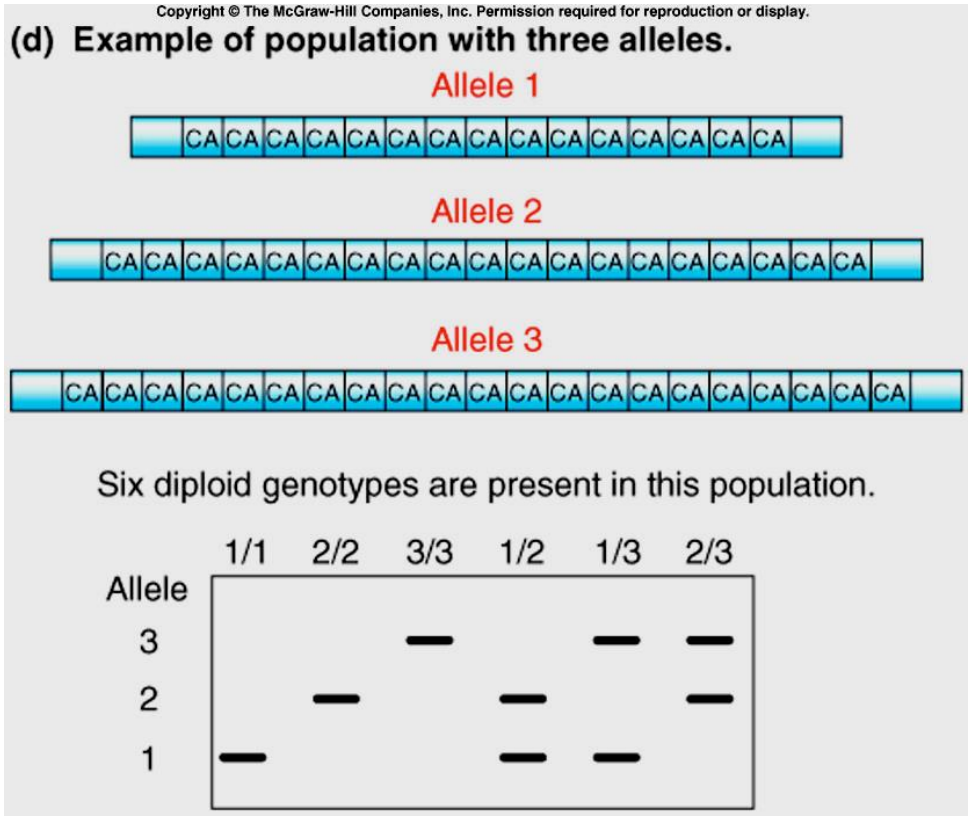
→ Most useful minisatellites have 10 – 20 sites around genome and can be analyzed on one gel

What about the rest of genome

Q15 Repeat elements: PCR-based DNA profiling

Increased sensitivity and allowed automation





What about the rest of genome
Repeat elements: DNA fingerprinting

1985 – Alec Jeffreys made two key findings
– Each minisatellite locus is highly polymorphic
– Most minisatellites occur at multiple sites around the genome

Thus, they could be used to generate a "DNA fingerprint", i.e. a pattern of simultaneous genotypes at a group of unlinked loci:

→ Most useful minisatellites have 10 – 20 sites around genome and can be analyzed on one gel

First application:
-Migration case of a Ghanese boy (son or nephew?)
- Forensics: Conviction of Colin Pitchfork (raped and murdered 2 girls)

What about the rest of genome
Summary

Genome:

- Small proportion (5%)
  - Genes
  - Functional (regulatory elements)
- Large proportion (~50%)
  - Repeat elements: micro- and minisatellites
  - Transposons
    - SINEs
    - LINEs
  - Important roles in evolution and forensic application
- Non-coding DNA (45%) – “junk”?