

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 the clicker system:

<http://rwpoll.com>

For your account

(Check the box “I have not purchased a license, but would still like to create an account.” and complete the form with your EPFL email address.

Session ID: GG024

Note: it is possible that this only works with your smart phone

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)

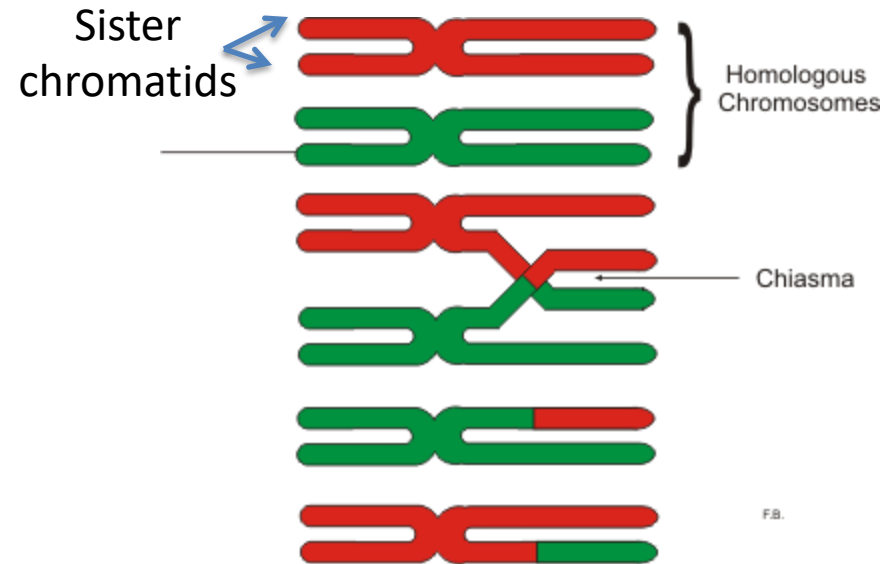
Q1a

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)

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

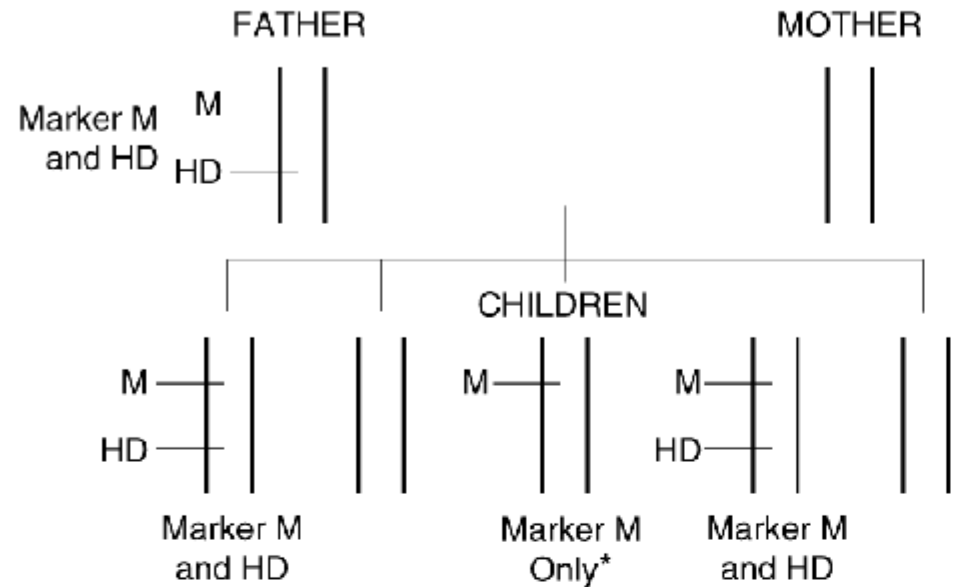
Botstein and his human genetic linkage map

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)



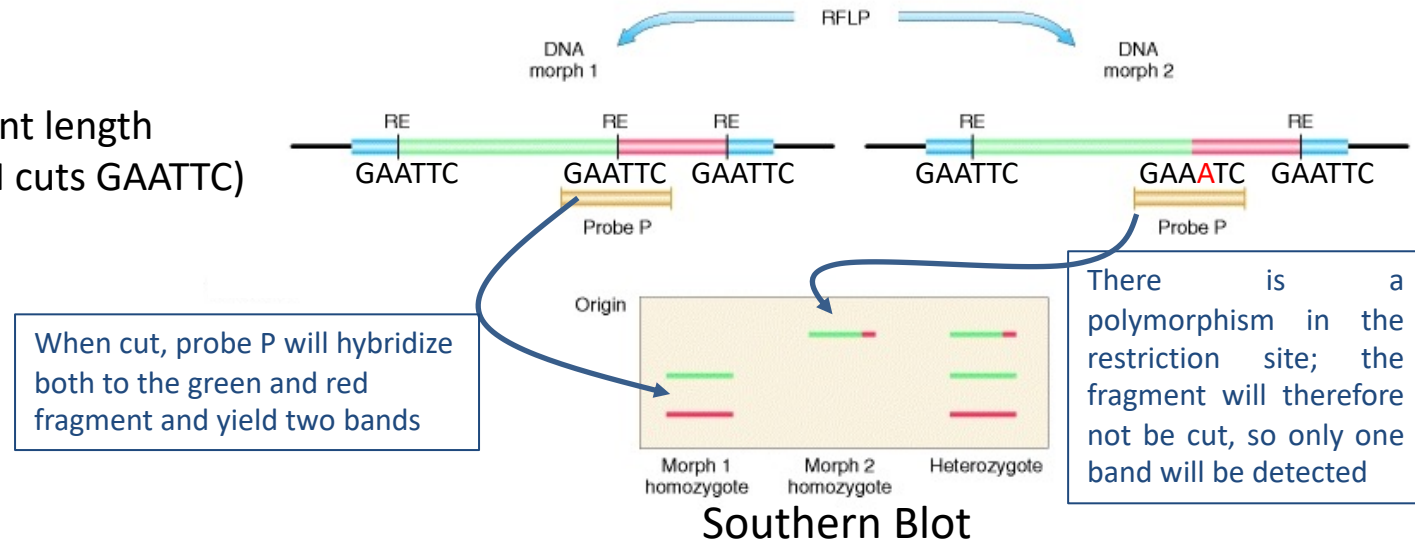
Used to “map” Huntington Disease gene in 1983!

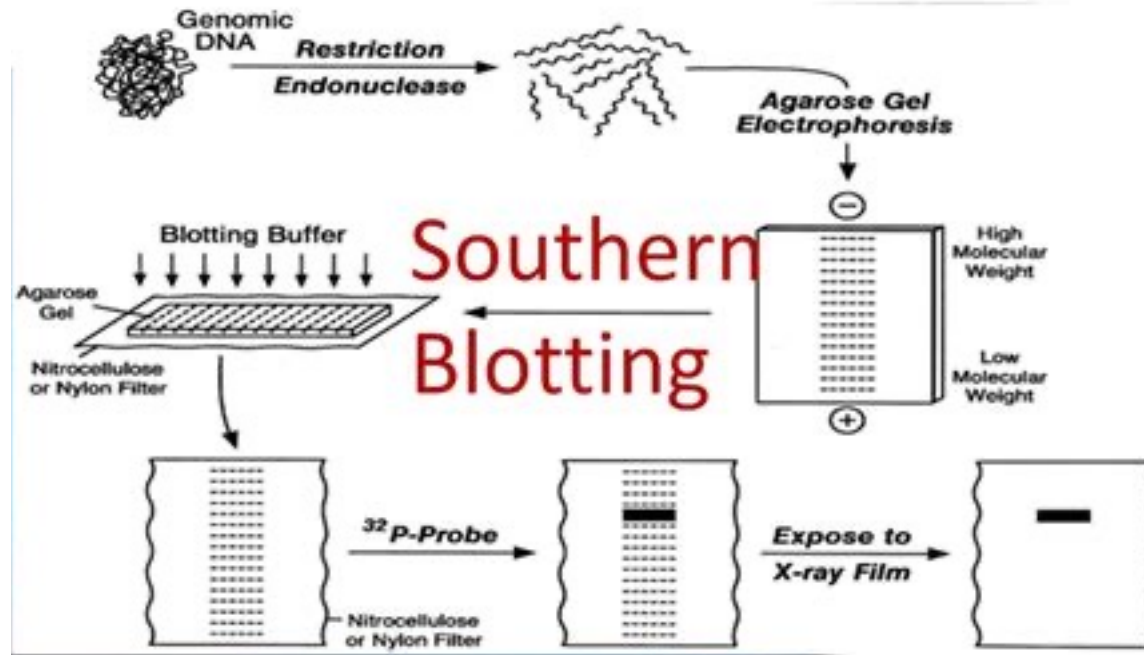
So how do we know whether M and HD are linked in an individual?

Botstein and his human genetic linkage map: RFLP method

Q1b

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





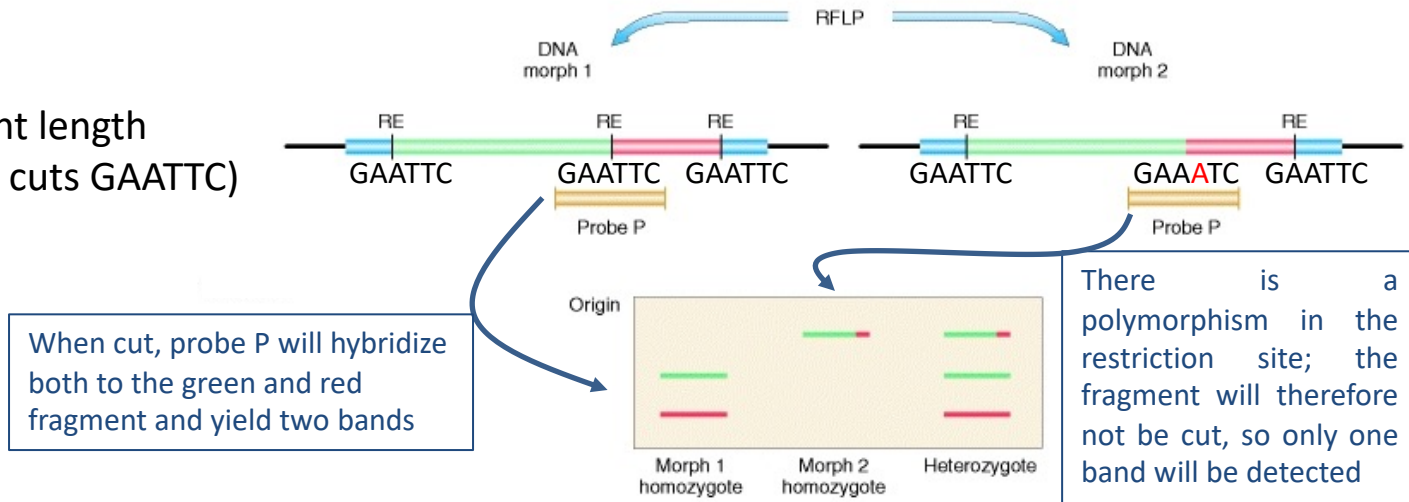
A **Southern blot** is a method used in molecular biology for the detection of a specific DNA sequence in DNA samples. Southern blotting combines transfer of electrophoresis-separated DNA fragments to a filter membrane and subsequent fragment detection by probe hybridization.

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

Botstein and his human genetic linkage map: RFLP method

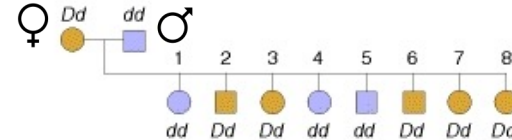
Q1b

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

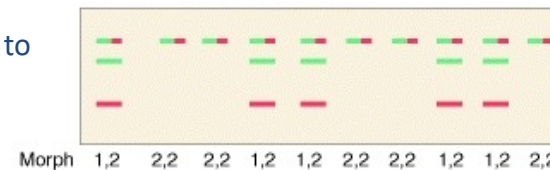


Southern Blot

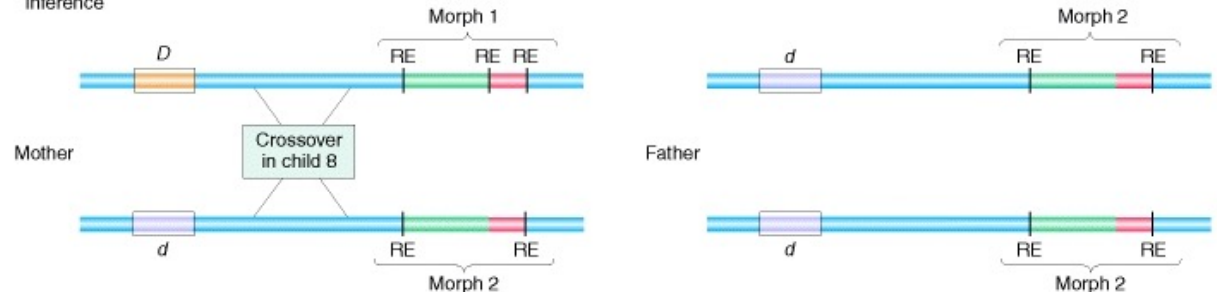
Linkage to *D* locus
D = disease



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



Inference

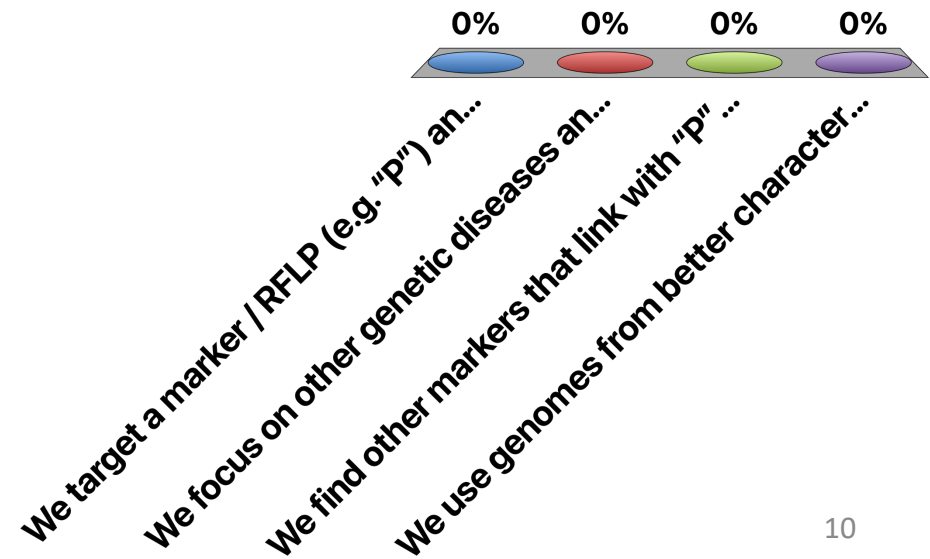


If an individual is heterozygous for two morphs of an RFLP, this heterozygous "locus" can be used as a marker in chromosomal mapping

How was the human genetic linkage map constructed?

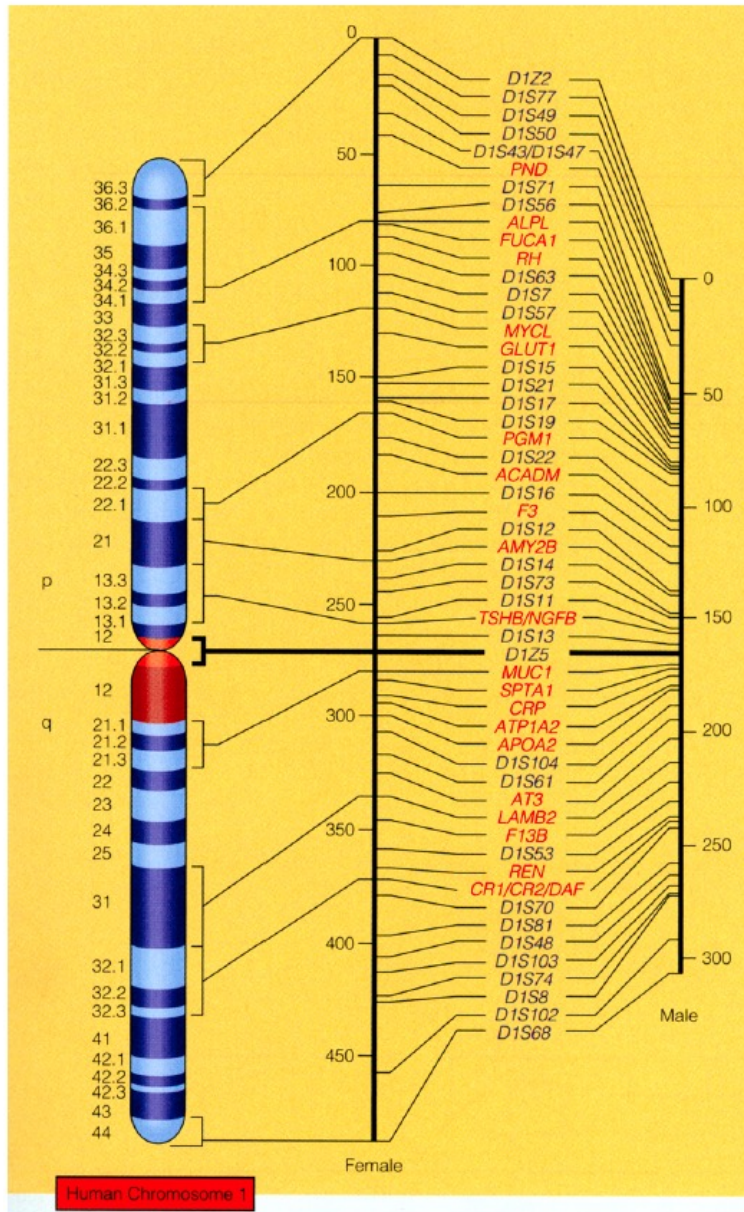
Q1b

- 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)

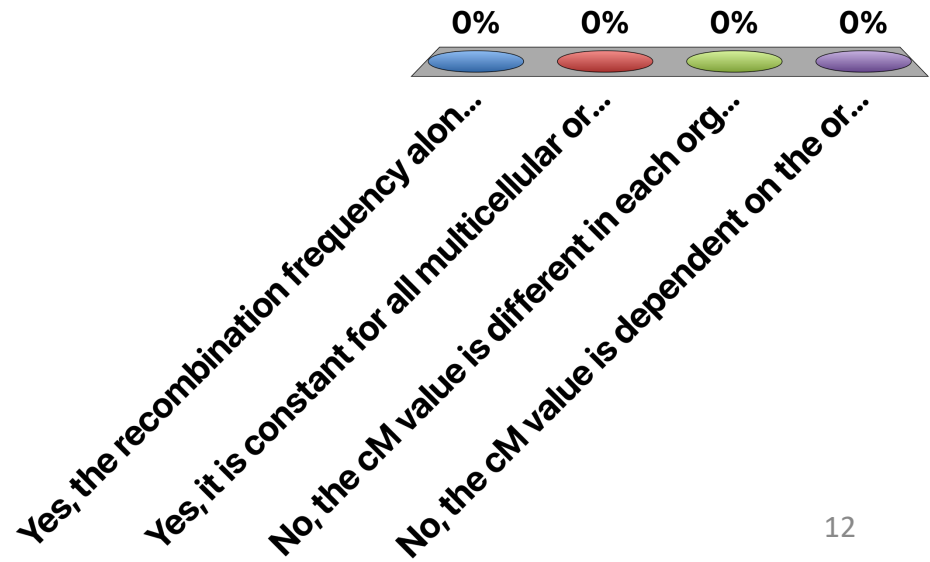
Q2a

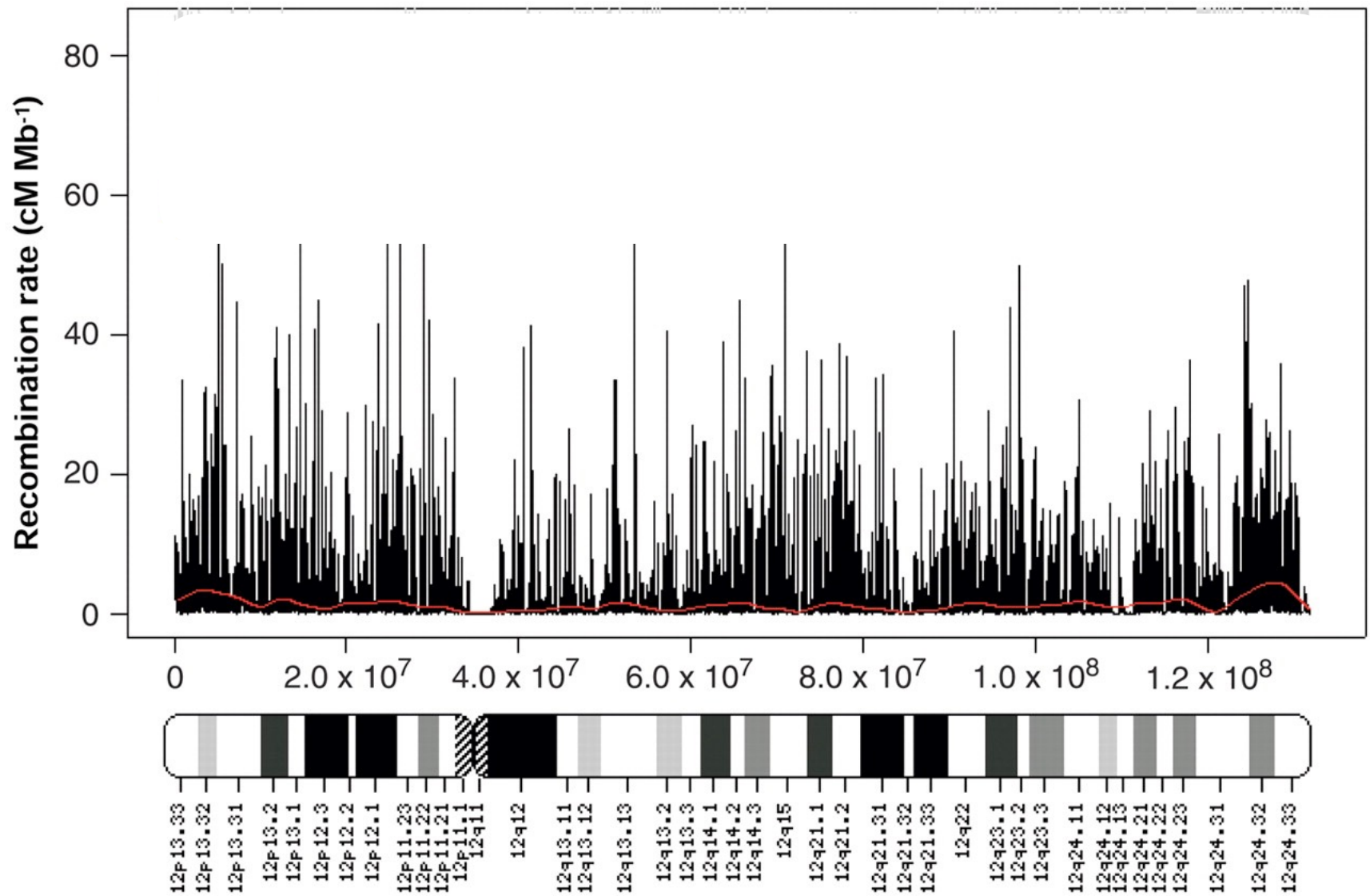
- 1 genetic map unit (cM) = 1% chance (=1 out of 100 meioses resulting in a recombination event) that a marker at one genetic locus on a chromosome will be separated from a marker at a second locus due to crossing over in a single generation = about 1Mb in humans (physical distance)
- cM = centimorgan

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





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

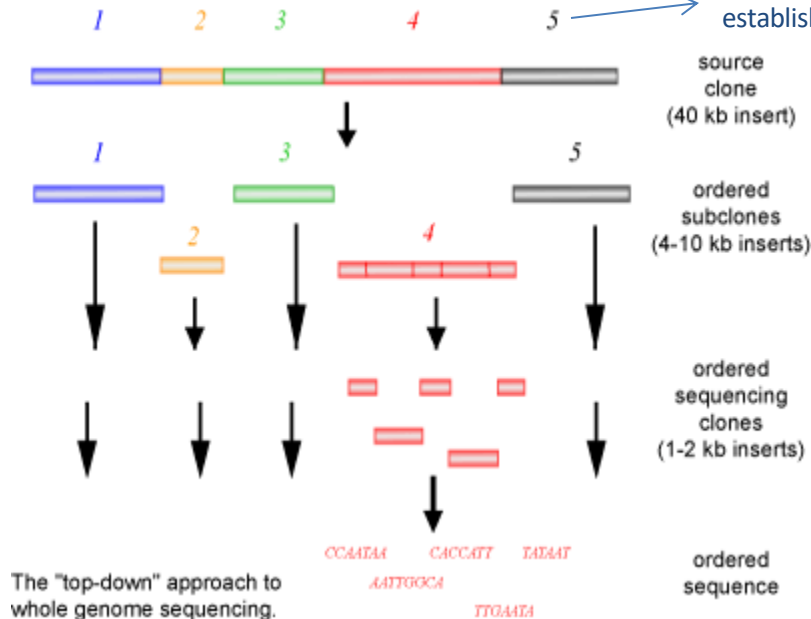
Genomics – a short history

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

Chromosome walking

Q3a

Genetic map helped to establish the order!



NHGRI & Sanger Center

E. coli

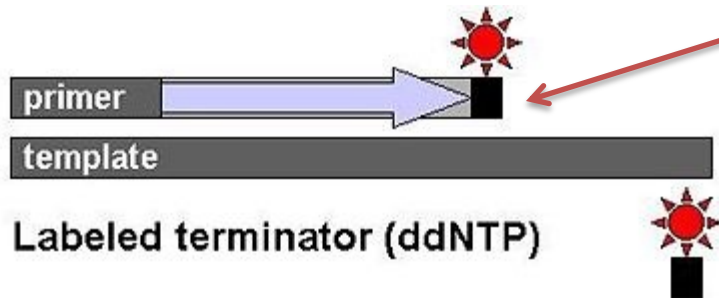
S. cerevisiae

Mycoplasma capricolum

Caenorhabditis elegans

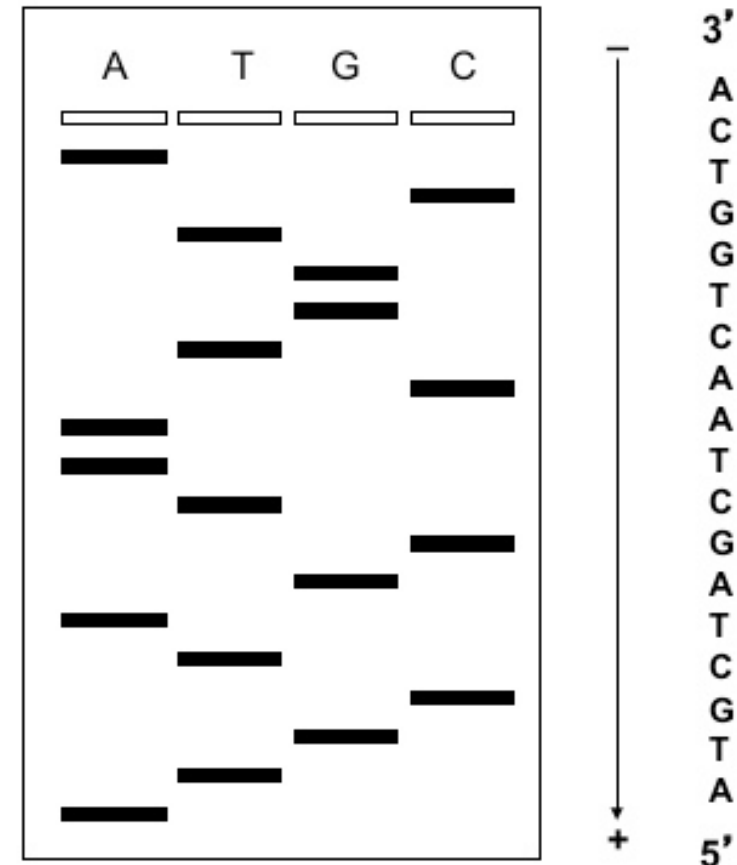
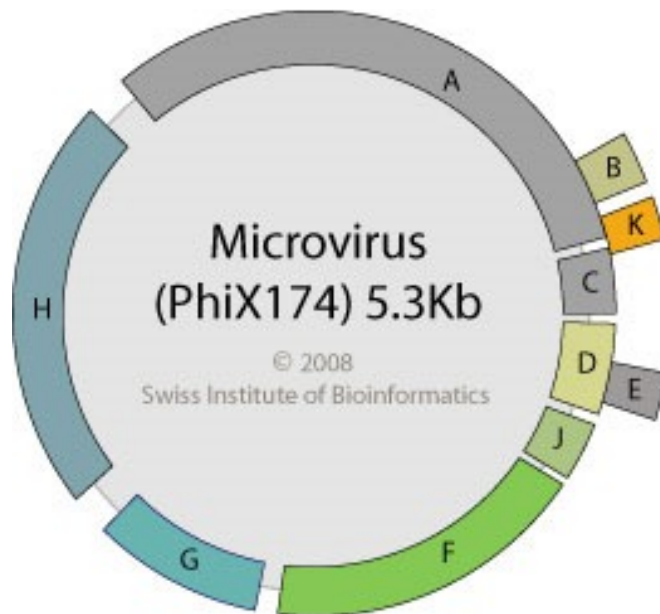
Genomics – a short history

Q3b 1975: Sanger sequencing (chain termination sequencing)



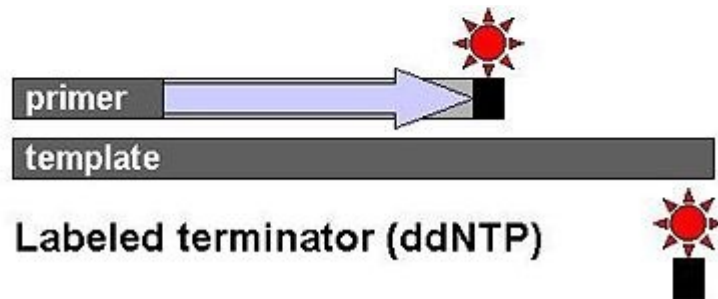
Use of radioactivity to visualize bands
(4 reactions, each involving one
kind of ddNTP)

Sanger et al., Nature, 1977

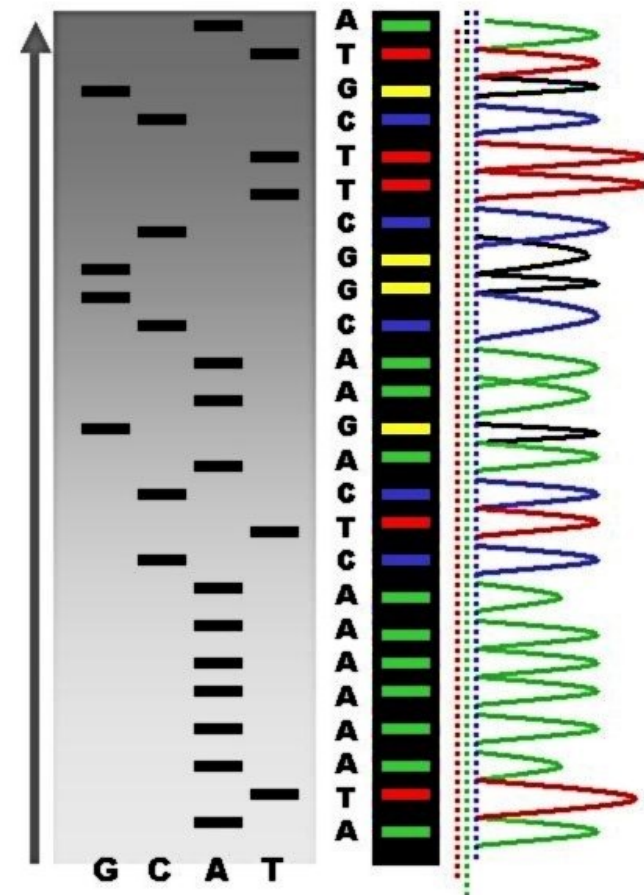


Genomics – a short history

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



Only 1 reaction and automatable!

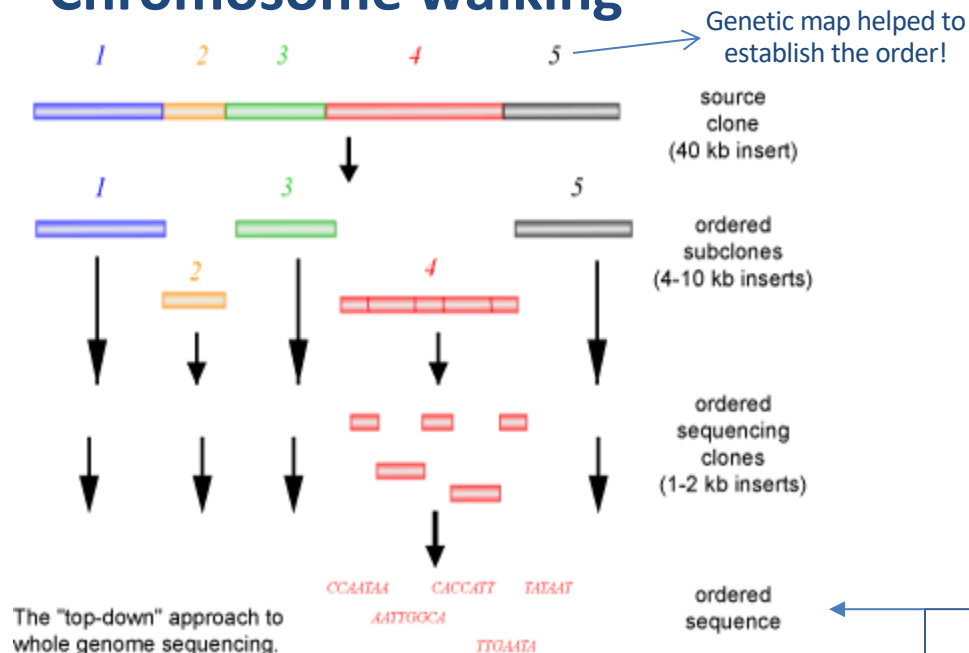


Genomics – a short history

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

Chromosome walking

Q3a



Sanger sequencing
(Smith and Hunkapillar method)

NHGRI & Sanger Center

E. coli

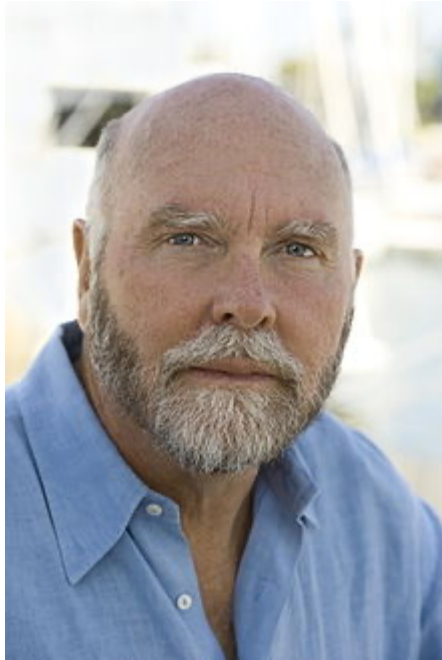
S. cerevisiae

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™
I N S T I T U T E

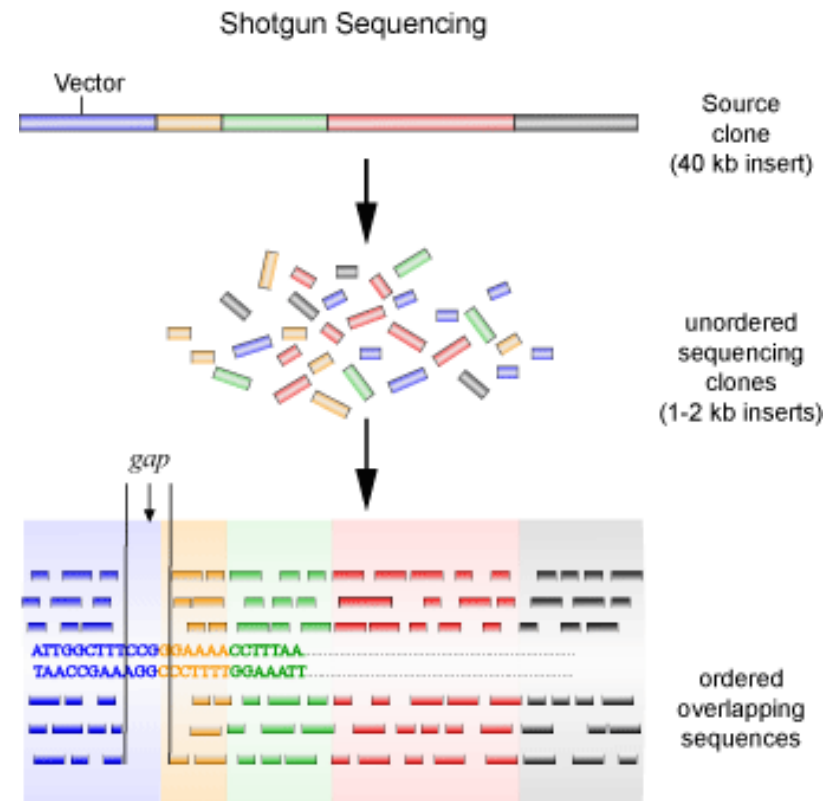
Genomics – a short history

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

Chromosome walking

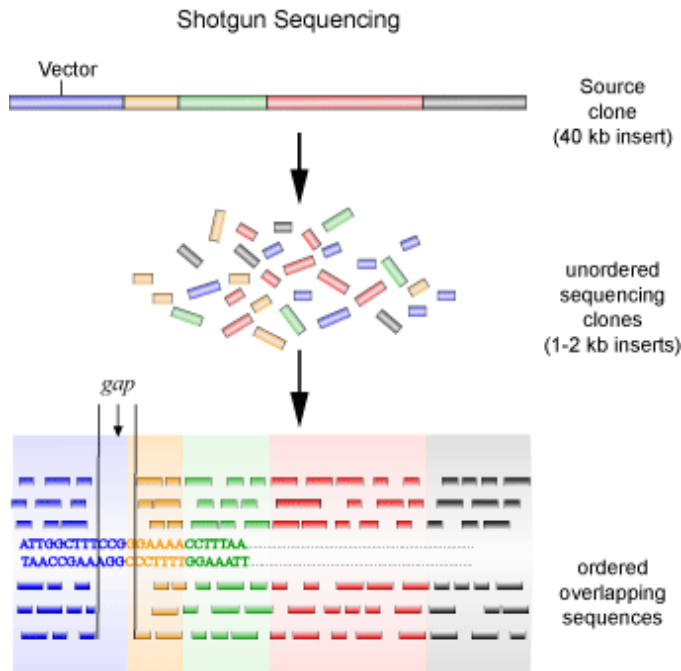


Q3a Shot-gun sequencing



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

(a) Sequence reads

Read 1 CACATACACATGG
Read 2 TCAATGGGGCTAA
Read 3 AGCACGGACTTGTCACATACACATG
Read 4 ACACATGGGAAATA
Read 5 GGGCTAATGATTGTCAC
Read 6 TGATTGTCACATA
Read 7 ATTCATGAAGCACGGA
Read 8 GTCACATACACATGATCAATGGGG

Use computer to assemble sequence reads

(b)

7 ATTCATGAAGCACGGA
3 AGCACGGACTTGTCACATACACATG
8 GTCACATACACATGATCAATGGGG
2 TCAATGGGGCTAA
5 GGGCTAATGATTGTCAC
6 TGATTGTCACATA
1 CACATACACATGG
4 ACACATGGGAAATA

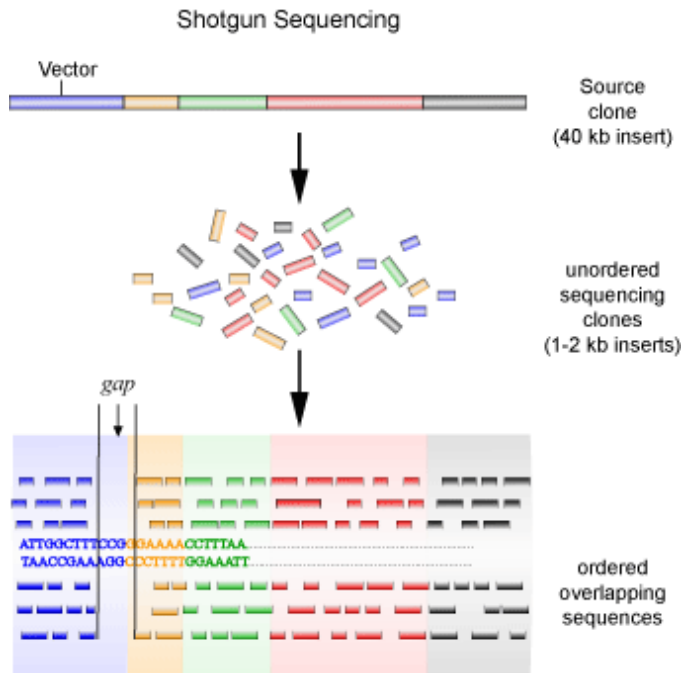
Assembled sequence

(c)

ATTCATGAAGCACGGACTTGTCACATACACATGATCAATGGGGCTAATGATTGTCACATACACATGGGAAATA

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.

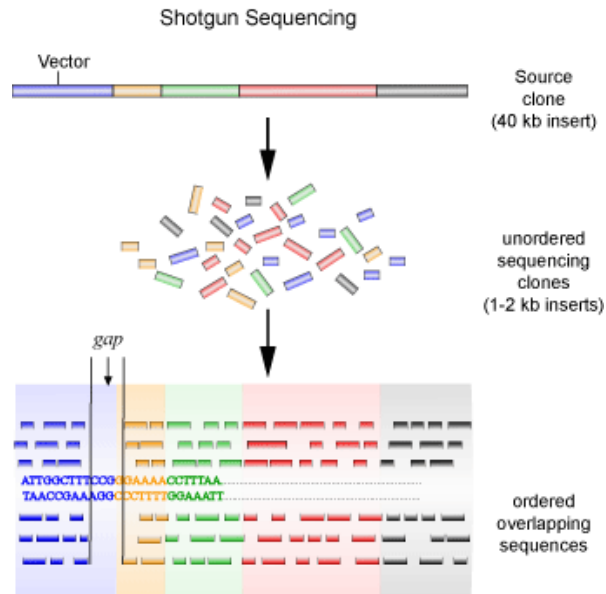
TIGR Assembler: A new tool for assembling large shotgun sequencing projects

Genome Science and Technology. 1995 Jan 01; 1(1): 9-19.

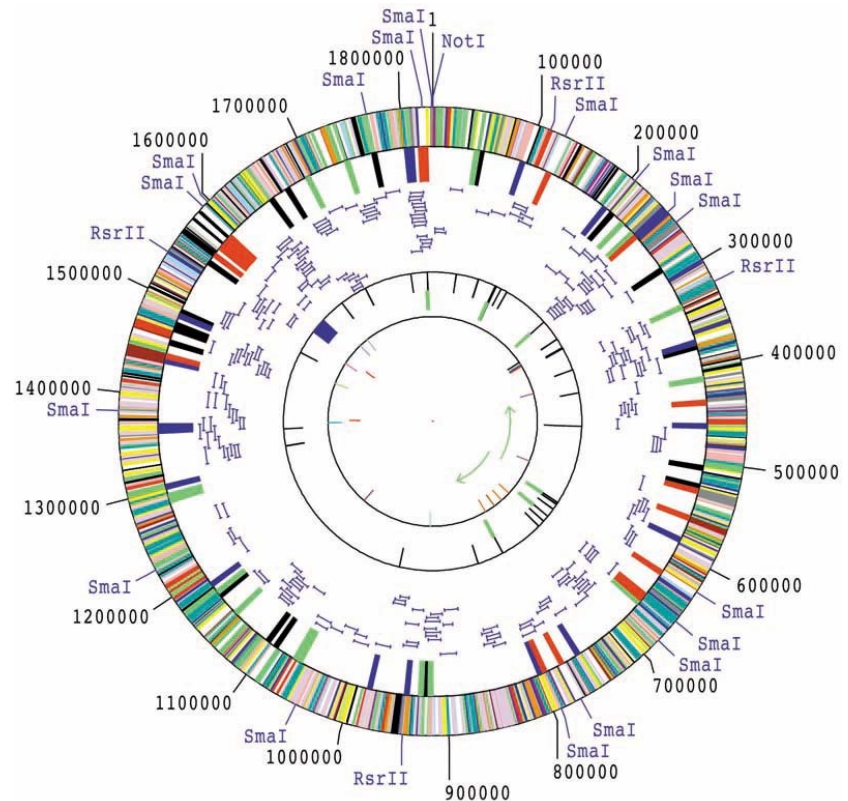
- pairing oligonucleotide content (set of n-mers of length 10 -12) instead of actual sequences → reducing search time
- repeat regions are dealt with by increasing the match criteria stringency and by assembling these regions last
- constraints such as clone length (wet bench)

Genomics – a short history

TIGR shocks the world



1995: the first bacterial genome
Haemophilus influenzae (1.8 Mb)

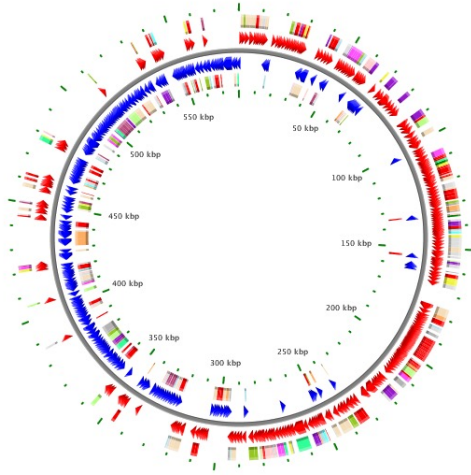


[Sutton, G. G.](#), White, O., Adams, M. D., Kerlavage, A. R.
TIGR Assembler: A new tool for assembling large shotgun sequencing projects
Genome Science and Technology. 1995 Jan 01; 1(1): 9-19.

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

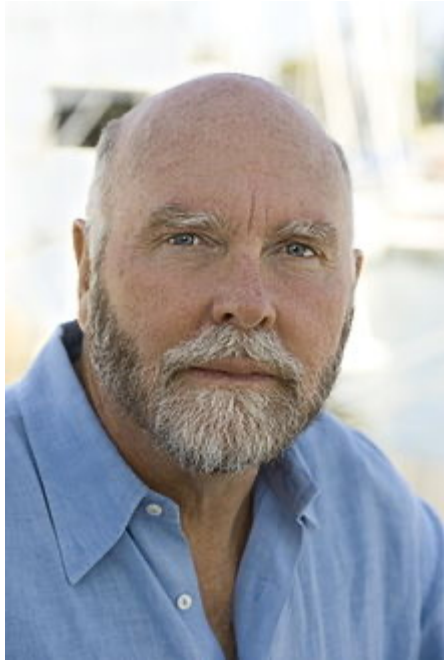
What about academia?

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



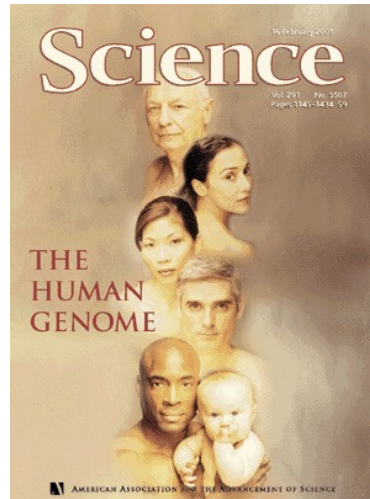
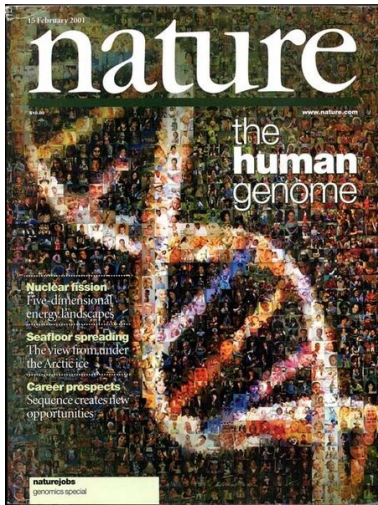
- NIH staff scientist
- Sequences cDNA → patents
- Wallace Steinberg provides funds for TIGR (non-profit) and Human Genome Sciences (profit)
- Mid-nineties: SmithKline Beecham → \$110 million for exclusive rights of patents
- First genome success of TIGR: *Haemophilus influenzae* --> shotgun sequencing!
- 1997: Venter founds Celera Genomics with investment from Perkin Elmer (Applied Biosystems → Hunkapillar).
- Wellcome Trust and NIH step it up, the race is on.

J. Craig Venter™
I N S T I T U T E

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



Q4

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

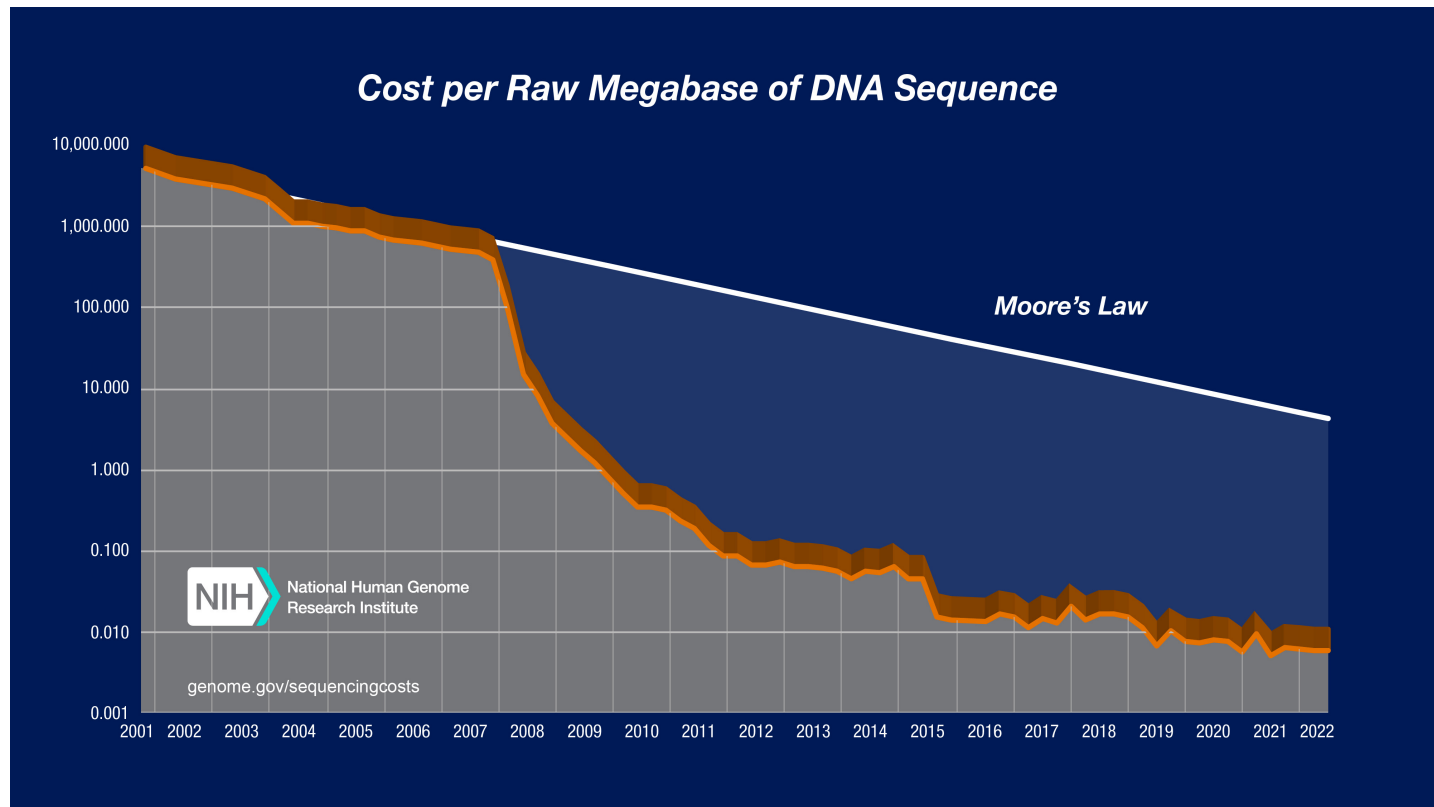


April 2022: From 200 million bases of DNA untouched to only 10 million bases and the Y chromosome only “roughly” known

Genomics – a short history

Sequencing cost

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



Whole sale!

Nebula Genomics (a George Church lab spin-off)

October 11th 2024

60%
% off average
industry pricing

Best value



The future of
health is in
your DNA

Nebula
Genomics

Deepⁱ
Whole Genome Sequencing

\$249 ~~637~~¹

Plus Nebula Explore™ Membership

- Deep ancestry report
- Detects all predispositions
- Detect rare genetic mutations
- High accuracyⁱ
- Requires [Nebula Membership](#)

Membership: 3 Year+ (\$295) ▼

Total: \$544

Why is there a membership?

Buy Deep DNA Test Bundle

57%
% off average
industry pricing

Ultra high accuracy



The future of
health is in
your DNA

Nebula
Genomics

Ultra Deepⁱ
Whole Genome Sequencing

\$899 ~~2402~~¹

Plus Nebula Explore™ Membership

- Ultra deep ancestry report
- Detects all predispositions
- Detect rare genetic mutations
- Ultra high accuracyⁱ
- Requires [Nebula Membership](#)

Membership: 3 Year+ (\$295) ▼

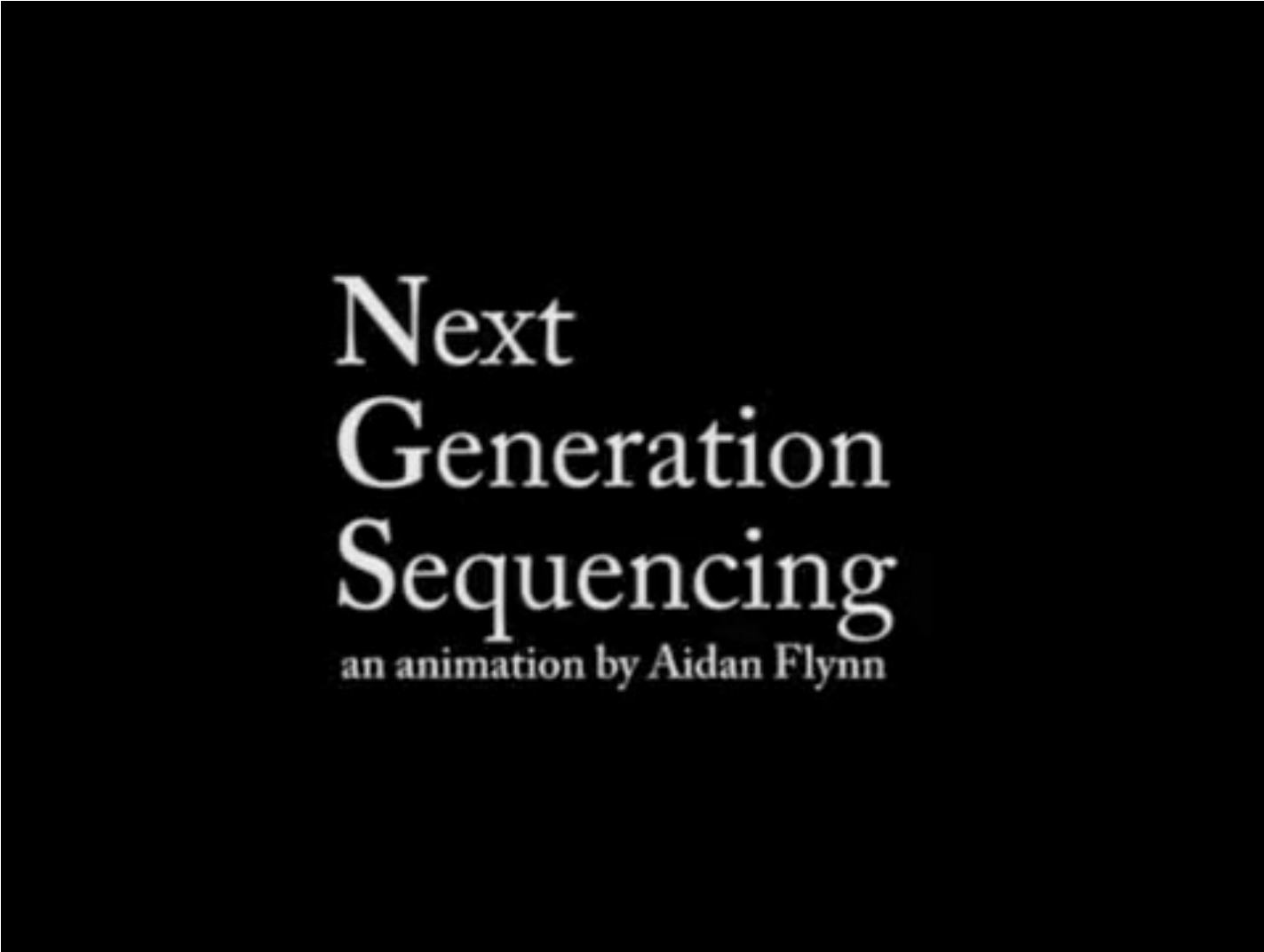
Total: \$1194

Why is there a membership?

Buy Ultra Deep DNA Test Bundle

The revolution of high-throughput sequencing: Illumina

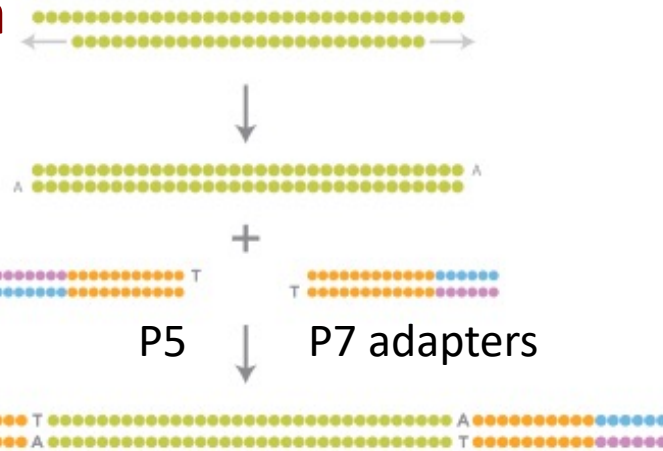
<http://www.youtube.com/watch?v=77r5p8IBwJk&noredirect=1>

A black rectangular frame containing white text. The text is centered and reads 'Next Generation Sequencing' in a large, serif font, with 'Next' on the first line, 'Generation' on the second, and 'Sequencing' on the third. Below this, in a smaller, lowercase serif font, is the text 'an animation by Aidan Flynn'.

Next
Generation
Sequencing
an animation by Aidan Flynn

The revolution of high-throughput sequencing: Illumina

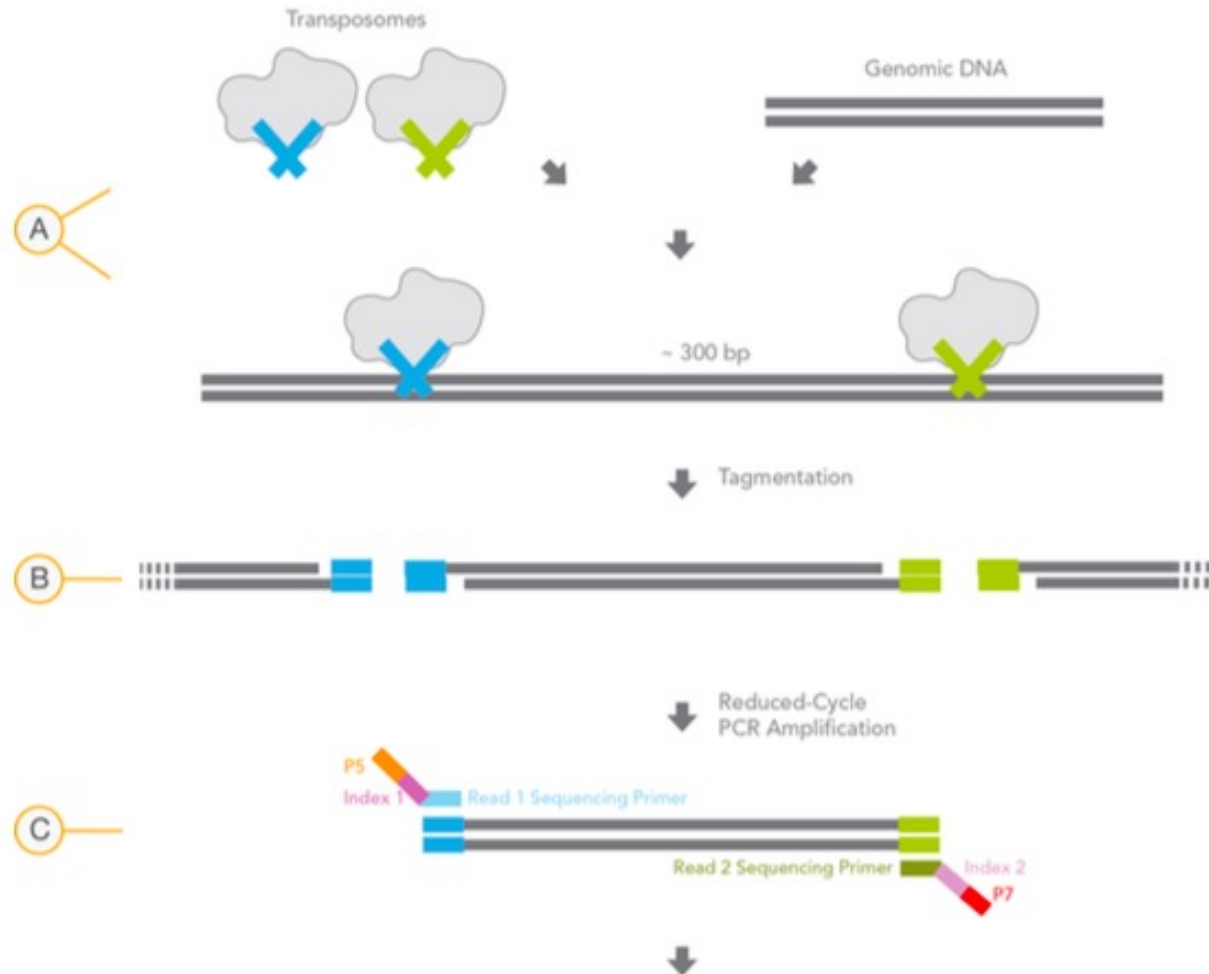
Q5a



- 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

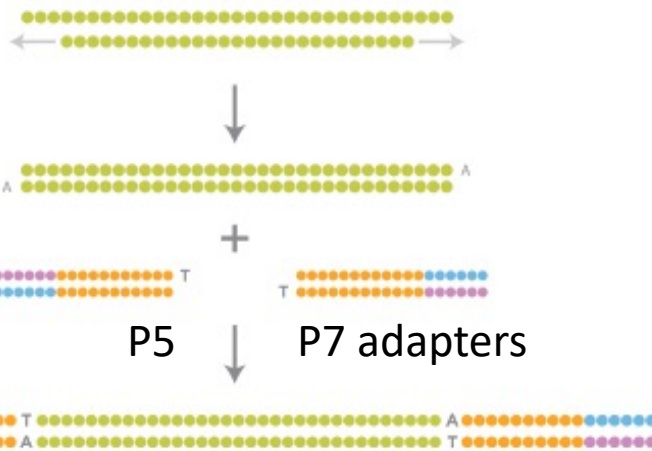
Because of inefficient ligation, “shearing” of DNA is performed with transposase (Tn5) loaded with P5 – P7-compatible adapters



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

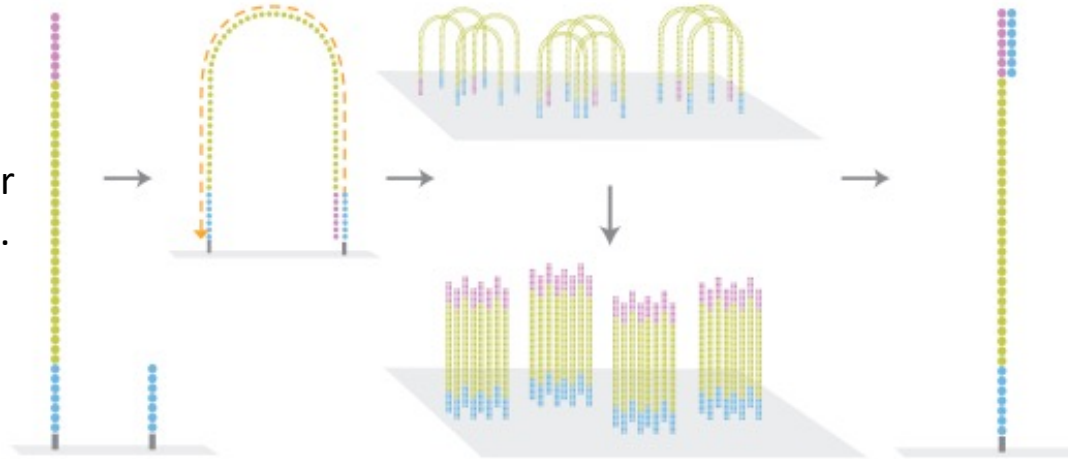
Q5a



- 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.



Appendix: Bridge amplification (2-3)

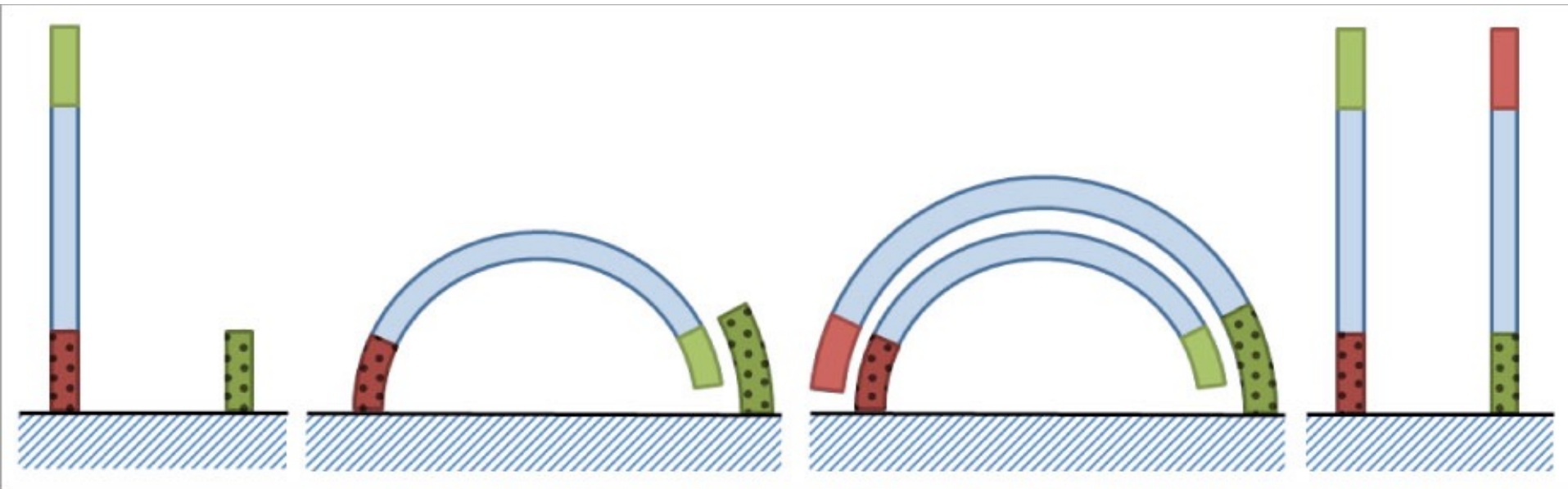
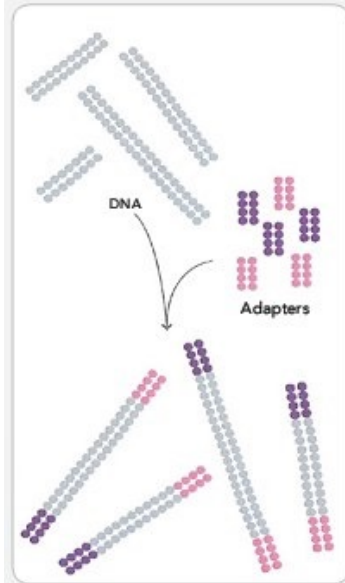


FIGURE 1: The mechanism of Illumina's bridge amplification. The process shown is repeated until DNA fragments form dense clusters. The green and red regions show the two primer sequences ligated to the blue DNA fragments. Primers that are complementary to these regions, and which are attached to the 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>

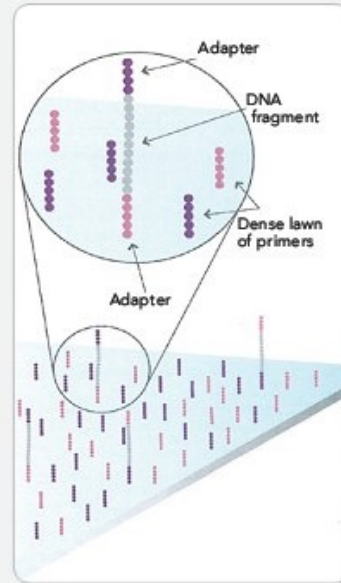
Appendix: Bridge amplification (1-3)

1. PREPARE GENOMIC DNA SAMPLE



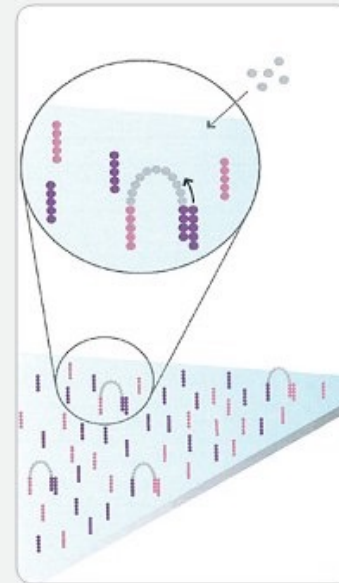
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE



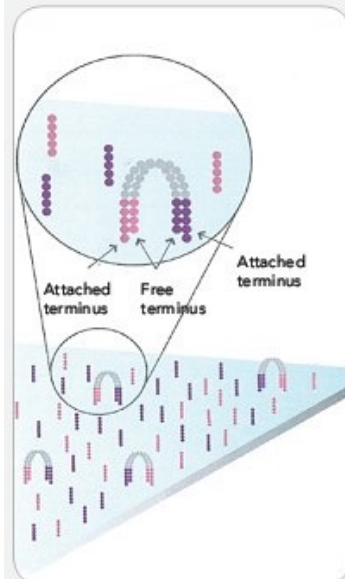
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION



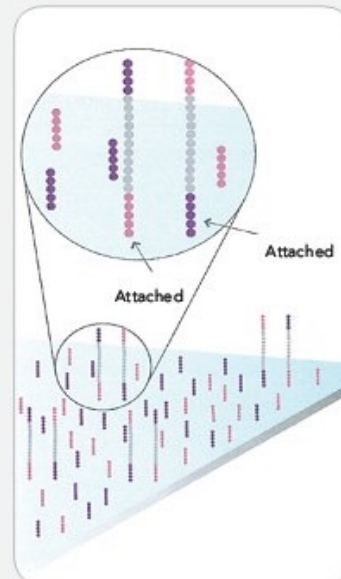
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED



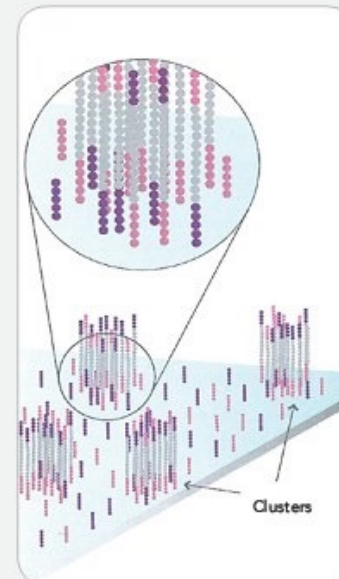
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



Denaturation leaves single-stranded templates anchored to the substrate.

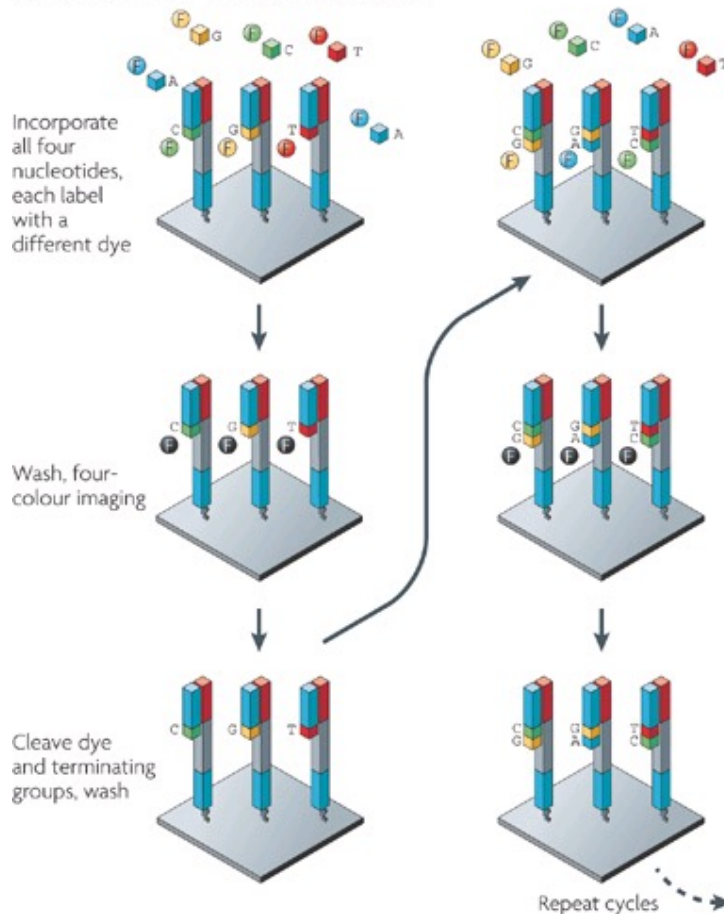
6. COMPLETE AMPLIFICATION



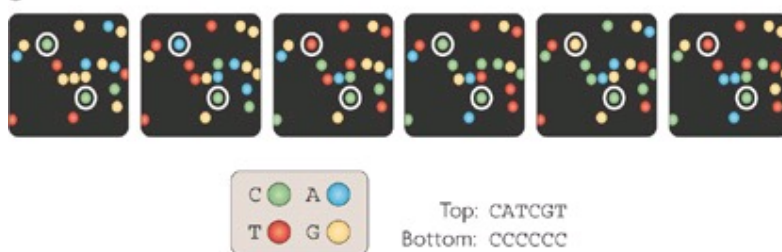
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

The revolution of high-throughput sequencing: Illumina

a Illumina/Solexa — Reversible terminators



b



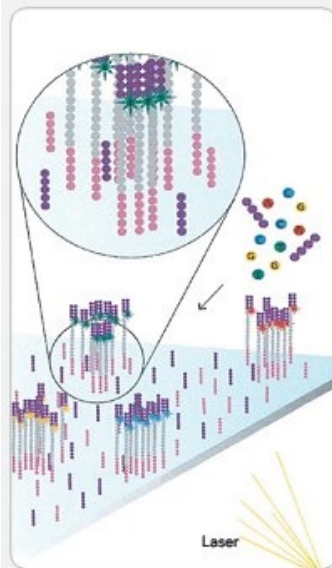
1. During sequencing the huge amount of generated clusters are sequenced simultaneously.
2. The DNA-templates are copied base by base using the four nucleotides (ACGT) which are fluorescently-labeled and reversibly terminated.
3. After each synthesis step, the clusters are excited by a laser which causes 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

Appendix: Bridge amplification (3-3)

7. DETERMINE FIRST BASE



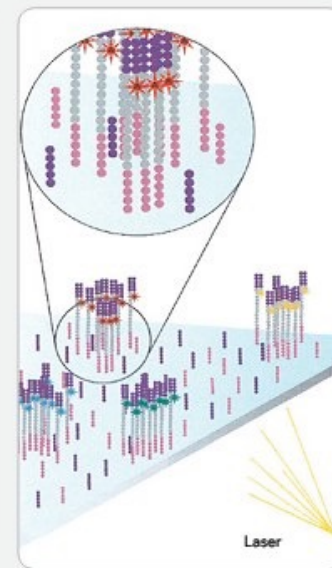
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE



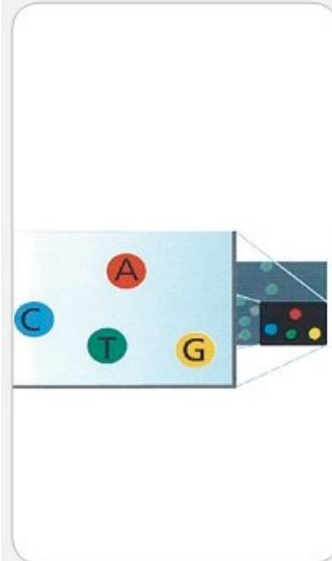
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE



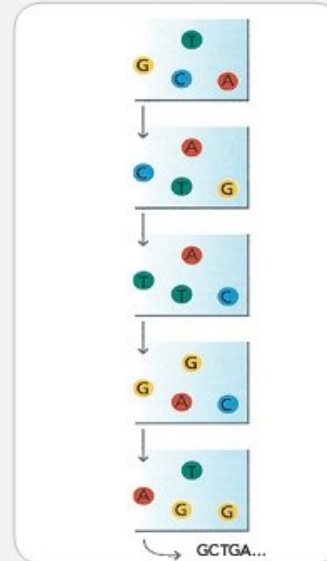
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE



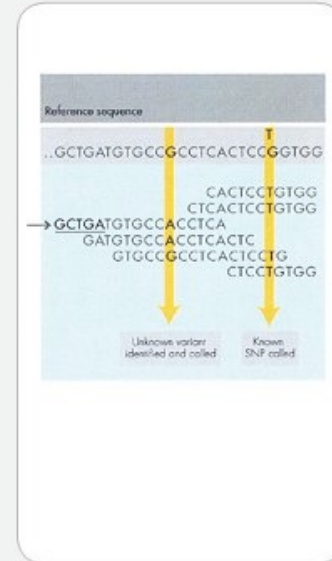
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES



Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

12. ALIGN DATA

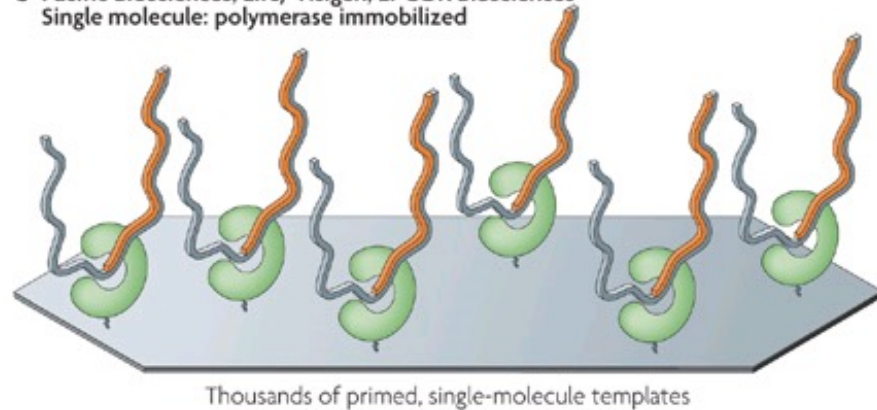


Align data, compare to a reference, and identify sequence differences.

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

Q5b

e Pacific Biosciences, Life/Visigen, LI-COR Biosciences
Single molecule: polymerase immobilized

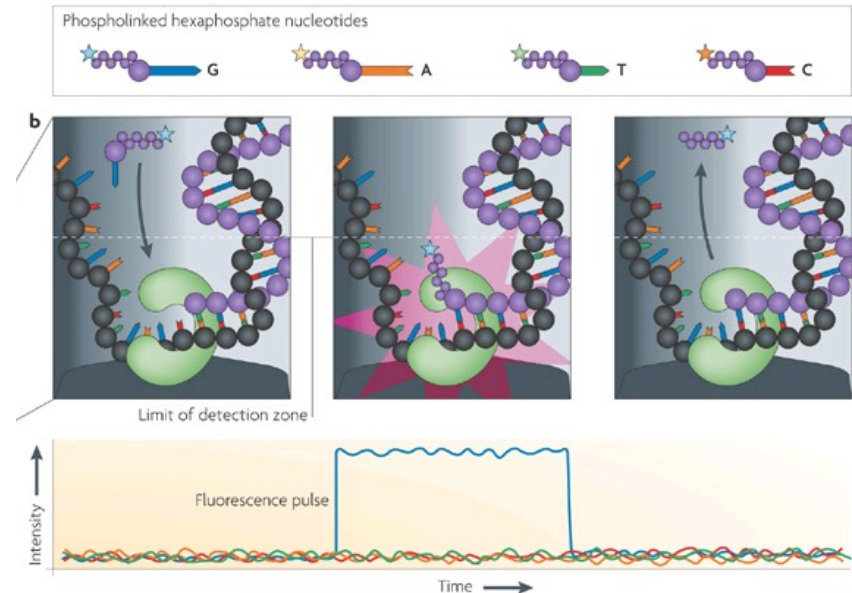


Green bean = immobilized DNA polymerase

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:

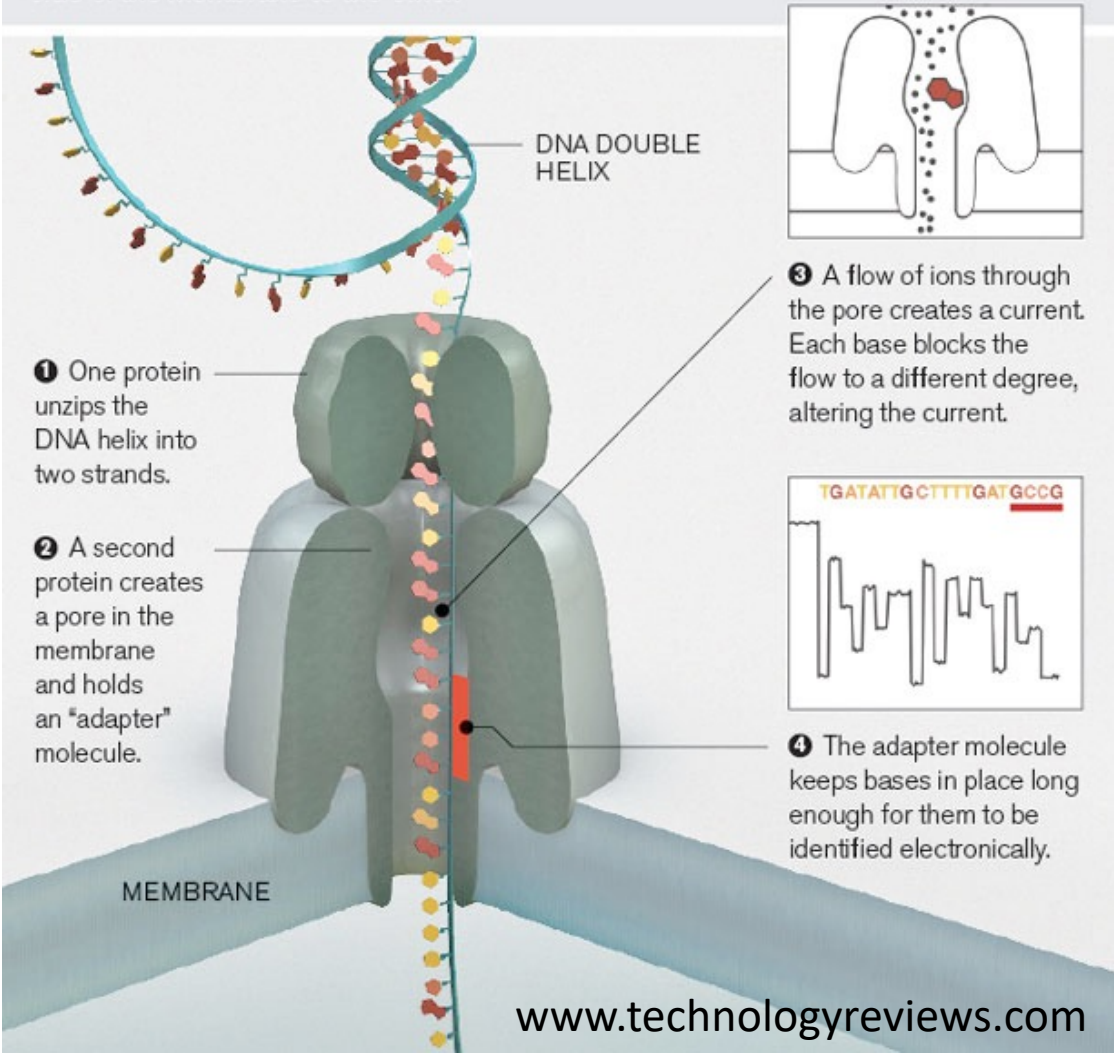
- Robustness
- More error-prone
- High machine cost

The revolution of high-throughput sequencing:

3rd and 4th generation technologies: *Nanopore Sequencing*

Q5b

DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.



Nanopore Minion



The sequencing revolution – Reality

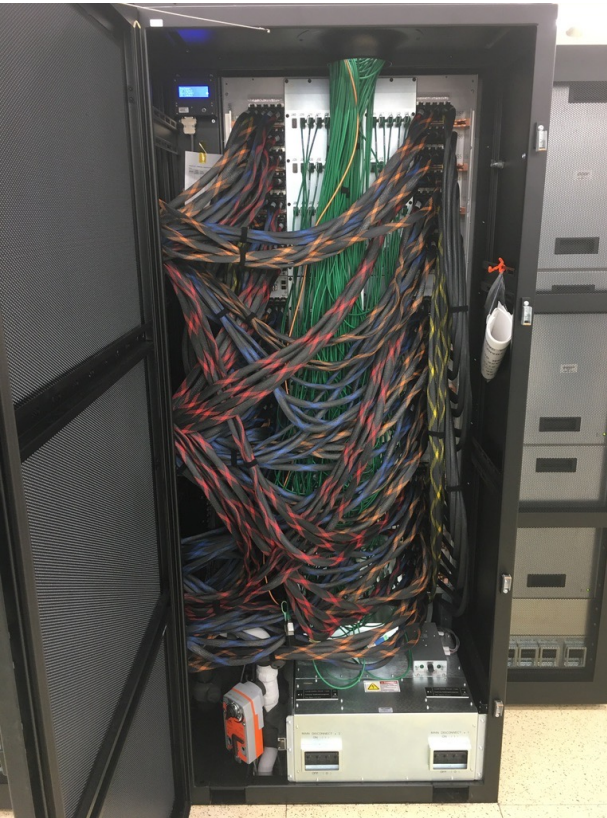
”Biology is (becoming) a data science



The sequencing revolution – Reality

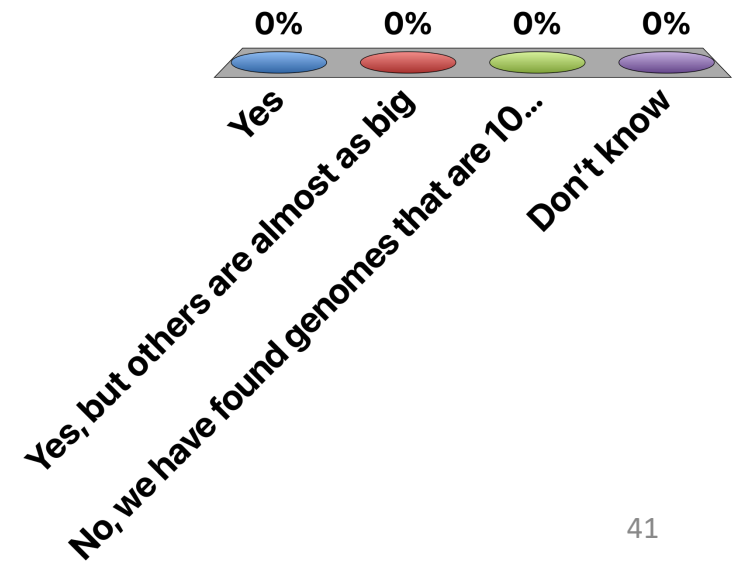
Swiss National SuperComputing Center
(October 2021; Piz Daint)

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

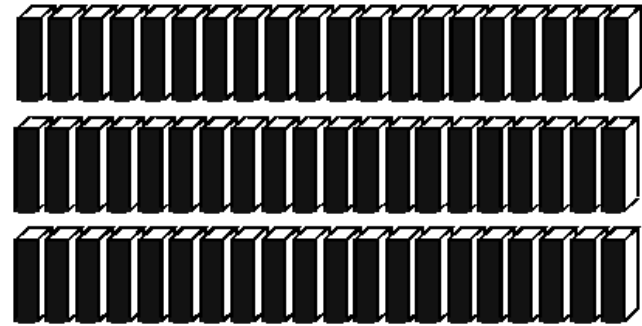
- A. Yes
- B. Yes, but others are almost as big
- C. No, we have found genomes that are 10-fold larger or more
- D. Don't know



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

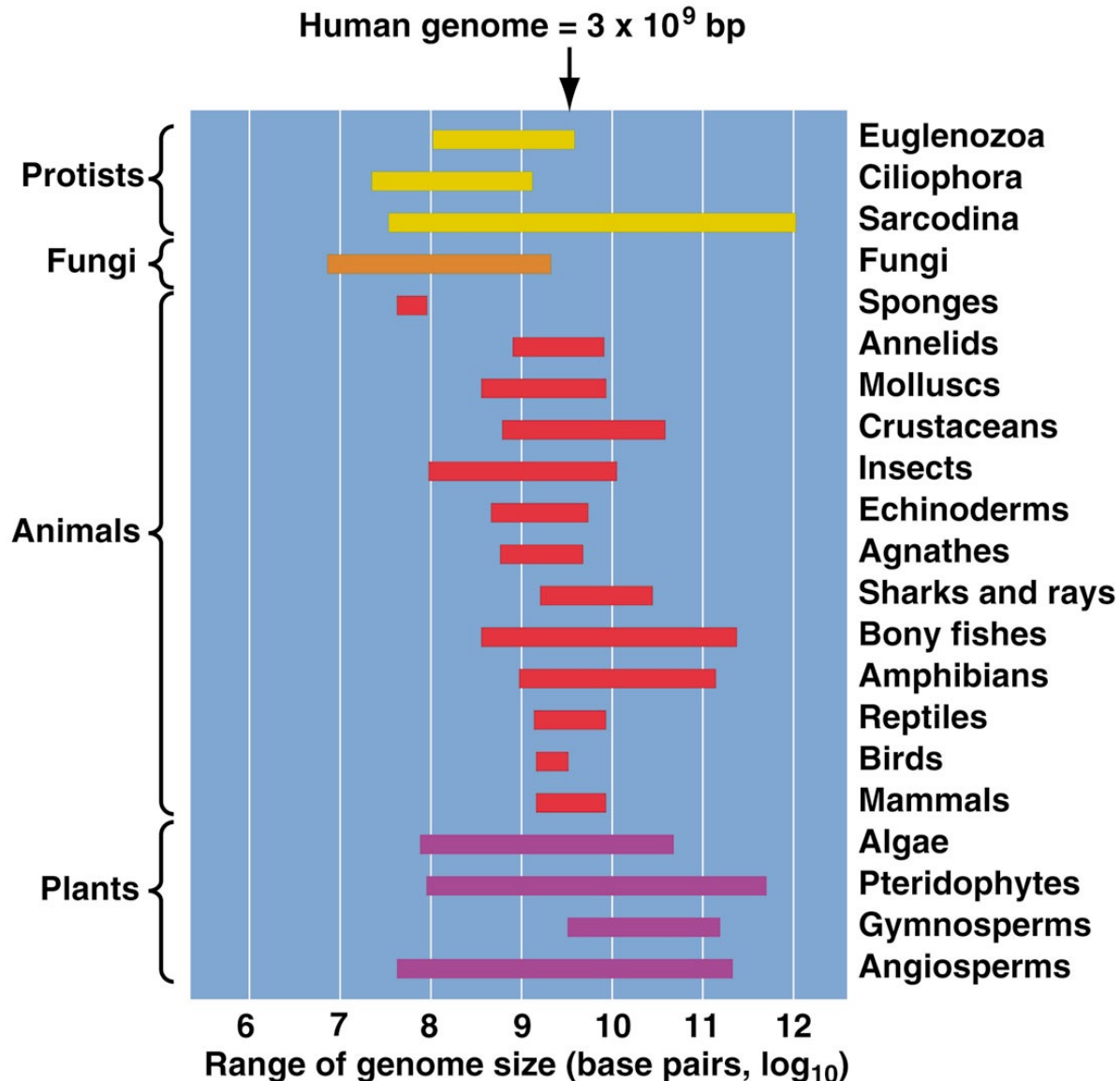
E. coli Genome



~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
CTAGGATCATCGGGAAAAGGAGGAGGAGGAGGAAGGCAGGCTCCGGGGAAAGCTGGTGGCAGCGGGTCTGGGTCTGGCGGACCTTGACGCGAAGGA
GGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCTCCAGCGTTGCCAACTGGACCTAAAGAGAGGGCCGCGACTGTGCCCC
ACCTGCGGGATGGGCTGTGCTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGGAGGGAGGGAGGCTGGGAGTCAGAATCGGGAAAGGGA
GGTGCGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCCTAGTGGGTGGAGAAAGCCGCTAGAGCAA
ATTTGGGGCCCGGACCAGGCAGCACTCGGCTTTTAACCTGGGCAGTGAAGGCGGGGGAAAGAGCAAAAGGAAGGGGTGGTGTGCGGAGTAGGGGTGGG
TGGGGGGAATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAGGTCTTTGGCATTAGGAGCTTGA
GCCCAGACGGCCCTAGCAGGGACCCAGCGCCGAGAGACCATGCAGAGGTGCGCTCTGGAAGAGGCCAGCGTTGTCTCCAAACTTTTTTTCAGGTGAGA
AGGTGGCCAACCGAGCTTCGGAAAGACACGTGCCACGAAAGAGGAGGGCGTGTGTATGGGTGGGGTTGGGGTAAAGGAATAAGCAGTTTTTAAAAA
GATGCGCTATCATTATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTAGAAGAGATGGAAGAATGAACTGAAGCTGATTGAAT
AGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACTATGCACTTGTTTTATTTCATTTTTCTAAGAACTAAAAATACTT
GTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCCCTTCATGCCTTGGACACTTGATTGTCTTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTG
CTCAGAAAAACATTTCTTGACTGAATTCAGCCAACAAAAATTTGGGGTAGGTAGAAAATATATGCTTAAAGTATTTATTGTTATGAGACTGGATATATCTAGTA
TTTGTACAGGTAAATGATTCTTCAAAAATTGAAAGCAAATTTGTTGAAATATTTATTTTGAAGAAAGTTACTTCACAAGCTATAAATTTTAAAGCCATAGG
AATAGATACCGAAGTTATATCCAACCTGACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACTTTAATGAGATTTTTTAAAT
AGCATCTAAGTTCGGAATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCTGTTATTATACTGG
TGTGAATGAATGAATAGGTACTGCTCTCTTGGGACATTACTTGACACATAATTACCAATGAATAAGCATACTGAGGTATCAAAAAAGTCAAATATGTTATA
AATAGCTCATATATGTGTGTAGGGGGGAAGGAATTTAGCTTTCACATCTCTCTTATGTTTAGTTCTCTGCATGTGCAGTTAATCCTGGAAGTCCGGTGCTAAGG
AGAGACTGTTGGCCCTTGAAGGAGAGCTCCTCCCTGTGGATGAGAGAGAAGGACTTTACTCTTTGGAATTATCTTTTGTGTTGATGTTATCCACCTTTTGT
ACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCTAGTCTTATAAAGTCAAATGTTAATTGGCATAAATTATAGACTTTTTTTAGCAGAGAACTTT
GAGGAACCTAAATGCCAACCAGTCTAAAAATGCAGTTTTCAGAAGAATGAATATTTTATGGATAGTTCTAAATACTAATGAACCTTTAAATAGCTTACTATTG
ATCTGTCAAAGTGGGTTTTTATATAATTTCTTTTACAAATCACCTGACACATTTAATATAGGTTAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTAT
TACAAAGGCTGCTAAGTGTGTTAAGAGCATACTCATTTCTGTTCTCCAAAATATTTTATAAGGTGCTTTAAGAATAGGTATGTTTTTAAAGTTAAGTTCCTAC
TATTTATAGGAAGTGAACATCACCTAAAAATACCAATGATTACAAACTTCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGAAAAACATATTTTCTGCATT
AAGTTAGGCAGTATTGCTTAGTTTTCAAAGTGGTAGGCTTTGGAGTCAGATTATTTTATTGATTGAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGC
AGGTCCCTTAACATCTCTGTGTGTGACTTGACCTTTAAATTTGGAGACTGTCATAGGGGTAAATCCCTTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTT
AACTGCTATTATTATGGATTACCATATTTTACATTCATCACAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTGAGTATAATTTTGTGACTCTCAA
TCTGGATATGCAATGAGTGGGCCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTACATGGCCTTACCAGATATACAGGAAACACGTACATGTTTCTATT
GTATGTTGTTAAATGCCTTAGAATTTAACTTTCTGAATAGGATCCCTCAGTTTGAGAGTCATAAAGAGTAAAATTATTATGGTAT

Genomics by the numbers

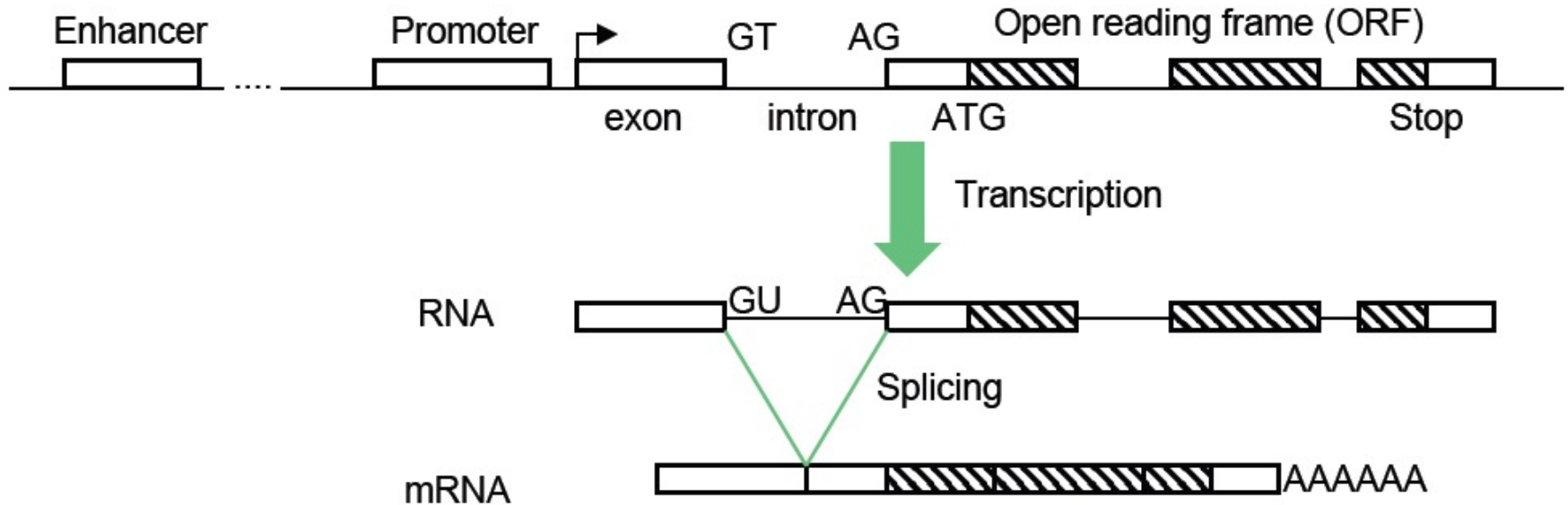
~3,000 bp (0.0001%) of Human Genome Sequence

TGCCGCGGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAATGCACCTTGCAAACGTAACAGGAACCCGA
CTAGGATCATCGGGAAAAGGAGGAGGAGGAGGAAGGCAGGCTCCGGGGAAGCTGGTGGCAGCGGGTCTGGGTCTGGCGGACCTGACGCGAAGGA
GGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCTCCAGCGTTGCCAACTGGACCTAAAGAGAGGGCCGCGACTGTGCCCC
ACCTGCGGGATGGGCTGTGCTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGAGAGGAGGGAGGCTGGGAGTCAGAATCGGGAAAGGGA
GGTGCGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCCTAGTGGGTGGAGAAAGCCGCTAGAGCAA
ATTTGGGGCCGGACCAGGCAGCACTCGGCTTTTAACCTGGGCAGTGAAG**GCGGGGGAAAGAGCAAAAGGAAGGGGTGGTGTGCGG**AGTAGGGGTGGG
TGGGGGGAATTGGAAGCAAATGACATCA**CAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAGGTCTTTGGCATTAGGAGCTTGA**
GCCCAGACGGCCCTAGCAGGGACCCAGCGCCGAGAGACCATGCAGAGGTGCGCTCTGGAAAAGGCCAGCGTTGTCTCCAAACTTTTTTTCAGGTGAGA
AGGTGGCCAACCGAGCTTCGGAAAGACACGTGCCACGAAAGAGGAGGGCGTGTGTATGGGTGGGTTGGGGTAAAGGAATAAGCAGTTTTTAAAAA
GATGCGCTATCATTATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTAGAAGAGATGGAAGAATGAAGTGAAGCTGATTGAAT
AGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACTATGCACTTGTTTTATTTCTATTTTCTAAGAACTAAAAATACTT
GTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCTTCATGCCTTGGACACTTGATTGTCTTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTG
CTCAGAAAACATTTCTTGACTGAATTCAGCCAACAAAAATTTGGGGTAGGTAGAAAATATATGCTTAAAGTATTTATTGTTATGAGACTGGATATATCTAGTA
TTTGTACAGGTAAATGATTCTTCAAAAATTGAAAGCAAATTTGTTGAAATATTTATTTTGAAGAAAGTTACTTCACAAGCTATAAATTTTAAAGCCATAGG
AATAGATACCGAAGTTATATCCAACCTGACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACCTTAAATGAGATTTTTTAAAT
AGCATCTAAGTTCGGAATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCTGTTATTATACTGG
TGTGAATGAATGAATAGGTACTGCTCTCTTGGGACATTACTTGACACATAATTACCCAATGAATAAGCATACTGAGGTATCAAAAAAGTCAAATATGTTATA
AATAGCTCATATATGTGTGTAGGGGGGAAGGAAT**TTAGCTTTCACATCTCTCTTATGTTTAGTTCTCTGCATGTGCAGTTAATCTCTGA**ACTCCGGTGCTAAGG
AGAGACTGTTGGCCCTTGAAGGAGAGCTCCTCCCTGTGGATGAGAGAGAAGGACTTTACTCTTTGGAATTATCTTTTGTGTTGATGTTATCCACCTTTTGT
ACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCTAGTCTTATAAAGTCAA**AATGTTAATTGGCATAAATTATAGACTTTTTTTAGCAGAGAACTTT**
GAGGAACCTAAATGCCAACCACTCTAAAAATGCAGTTTTTCAGAAGAATGAATATTTTCATGGATAGTTCTAAATACTAATGAACCTTAAATAGCTTACTATTG
ATCTGTCAAAGTGGGTTTTATATAATTTCTTTTACAAATCACCTGACACATTTAATATAGGTTAAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTAT
TACAAAGGCTGCTAAGTGT**GTTAAGAGCATACTCATTCTGTTCTCCAAATATTTTATAAGGTGCTTTAAGAATA**AGGTATGTTTTTAAAGTTAAGTTCCTAC
TATTTATAGGAAGTGAACATCACCTAAAAATACCAATGATTACAAACTTCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGAAAAACATATTTTCTGCATT
AAGTTAGGCAGTATTGCTTAGTTTTCAAAGTGGTAGGCTTTGGAGTCAGATTATTTTGAATCAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGC
AGGTCCCTTAACATCTCTGTGTGTGACTTGACCTTTAAATTTGGAGACTGTCATAGGGGTAAATCCCTTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTT
AACTGCTATTATTATGGATTACCATATTTTACATTTCATCAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTGAGTATAATTTTGTGACTCTCAA
TCTGGATATGCAATGAGTGGGCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTACATGGCCTTACCAGATATACAGGAAACACGTACATGTTTCTATT
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

TGCCGCGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAATGCACCTTGCAAACGTAACAGGAACCCGA
CTAGGATCATCGGGAAAAGGAGGAGGAGGAGGAAGGCAGGCTCCGGGGAAGCTGGTGGCAGCGGGTCTGGGTCTGGCGGACCTGACGCGAAGGA
GGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCTCCAGCGTTGCCAACTGGACCTAAAGAGAGGGCCGCGACTGTGCCCC
ACCTGCGGGATGGGCTGGTGTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGGAGGGAGGGAGGCTGGGAGTCAGAATCGGGAAAGGGA
GGTGCGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCCTAGTGGGTGGAGAAAGCCGCTAGAGCAA
ATTTGGGGCCGGACCAGGCAGCACTCGGCTTTTAACCTGGGCAGTGAAG**GCGGGGGAAAGAGCAAAAGGAAGGGGTGGTGTGCGG**AGTAGGGGTGGG
TGGGGGGAATTGGAAGCAAATGACATCA**CAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAGGTCTTTGGCATTAGGAGCTTGA**
GCCCAGACGGCCCTAGCAGGGACCCAGCGCCGAGAGACCATGCAGAGGTGCGCTCTGGAAAAGGCCAGCGTTGTCTCCAAACTTTTTTTCAGGTGAGA
AGGTGGCCAACCGAGCTTCGGAAAGACACGTGCCACGAAAGAGGAGGGCGTGTGTATGGGTGGGTTGGGGTAAAGGAATAAGCAGTTTTTAAAAA
GATGCGCTATCATTATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTAGAAGAGATGGAAGAATGAAGTGAAGCTGATTGAAT
AGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACTATGCACTTGTTTTATTTCTATTTTCTAAGAACTAAAAATACTT
GTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCTTC**ATGCTTGGA**CACTTGATTGTCTTCTTGGCACATACAGGTG**CCATG**CCCTGCATATAGTAAGTG
CTCAGAAAACATTTCTTGACTGAATTCAGCCAACAAAAATTTGGGGTAGGTAGAAAATATATGCTTAAAGTATTTATTGTTATGAGACTGGATATATCTAGTA
TTTGTACAGGTAAATGATTCTTCAAAAATTGAAAGCAAATTTGTTGAAATATTTATTTTGAAGAAAGTTACTTCACAAGCTATAAATTTTAAAGCCATAGG
AATAGATACCGAAGTTATATCCAAGTACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACTTTAATGAGATTTTTTAAAT
AGCATCTAAGTTCGGAATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCTGTTATTATACTGG
TGTGAATGAATGAATAGGTACTGCTCTCTTGGGACATTACTTGACACATAATTACCCAATGAATAAGCATACTGAGGTATCAAAAAAGTCAAATATGTTATA
AATAGCTCATATATGTGTGTAGGGGGGAAGGAAT**TAGCTTTACATCTCTCTTATGTTTAGTTCTCTGCATGTGCAGTTAATCCTGGA**ACTCCGGTGCTAAGG
AGAGACTGTTGGCCCTTGAAGGAGAGCTCCTCCCTGTGGATGAGAGAGAAGGACTTTACTCTTTGGAATTATCTTTTGTGTTGATGTTATCCACCTTTTGT
ACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCTAGTCCTTATAAAGTCAA**AATGTTAATTGGCATAAATTATAGACTTTTTTTAGCAGAGAACTTT**
GAGGAACCTAAATGCCAACAGTCTAAAAATGCAGTTTTCAGAAGAATGAATATTTATGGATAGTTCTAAATACTAATGAACCTTAAATAGCTTACTATTG
ATCTGTCAAAGTGGGTTTTATATAATTTCTTTTACAAATCACCTGACACATTTAATATAGGTTAAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTAT
TACAAAGGCTGCTAAGTGT**GTTAAGAGCATACTCATTCTGTTCTCCAAATATTTATAAGGTGCTTTAAGAATA**AGGTATGTTTTTAAAGTTAAGTTCCTAC
TATTTATAGGAAGTACAATCACCTAAAATACCAATGATTACAAACTTCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGAAAAACATATTTTCTGCATT
AAGTTAGGCAGTATTGCTTAGTTTTCAAAGTGGTAGGCTTTGGAGTCAGATTATTTGATTGAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGC
AGGTCCCTTAACATCTCTGTGTGTGACTTGACCTTTAAATTTGGAGACTGTCATAGGGGTAAATCCCTTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTT
AACTGCTATTATTATGGATTACCATATTTTACATTCATCACAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTGAGTATAATTTGTGACTCTCAA
TCTGGATATGCAATGAGTGGGCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTACATGGCCTTACCAGATATACAGGAAACACGTACATGTTTCTATT
GTATGTTGTTAAATGCCTTAGAATTTAACTTTCTGAATAGGATCCCTCAGTTTGAGAGTCATAAAAGAGTAAAATTATTATGGTAT

~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*
<i>Saccharomyces cerevisiae</i> (unicellular budding yeast)	12 million	6,000
<i>Trichomonas vaginalis</i>	160 million	60,000
<i>Plasmodium falciparum</i> (unicellular malaria parasite)	23 million	5,000
<i>Caenorhabditis elegans</i> (nematode)	95.5 million	18,000
<i>Drosophila melanogaster</i> (fruit fly)	170 million	14,000
<i>Arabidopsis thaliana</i> (mustard; thale cress)	125 million	25,000
<i>Oryza sativa</i> (rice)	470 million	51,000
<i>Gallus gallus</i> (chicken)	1 billion	20,000-23,000
<i>Canis familiaris</i> (domestic dog)	2.4 billion	19,000
<i>Mus musculus</i> (laboratory mouse)	2.5 billion	30,000
<i>Homo sapiens</i> (human)	2.9 billion	20,000-25,000

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

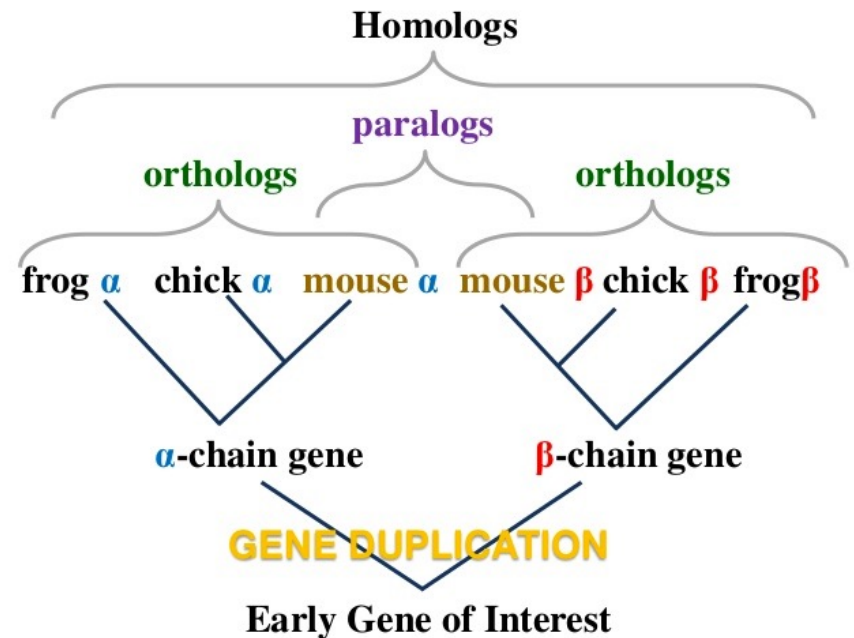
A gene related to a second gene by descent from a common ancestral DNA sequence. The term, homolog, may apply to the relationship between genes separated by the event of **speciation** (see ortholog) or to the relationship between genes separated by the event of genetic **duplication** (see paralog).

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.).

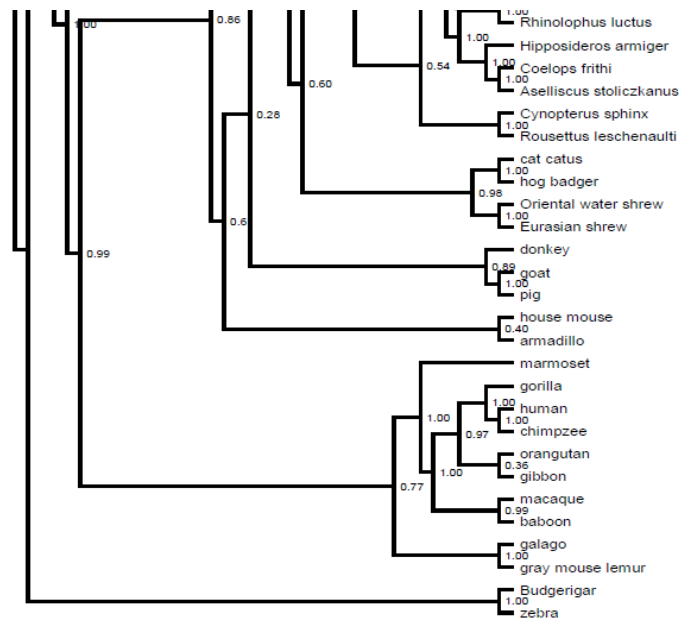
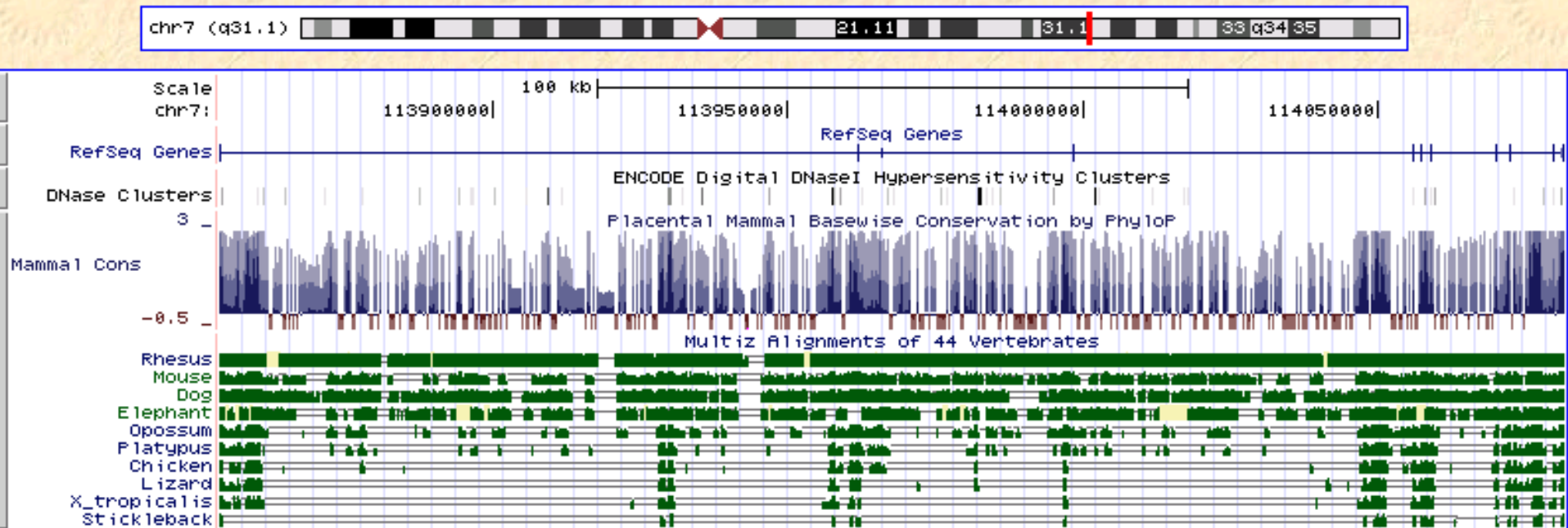
Paralog:

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.



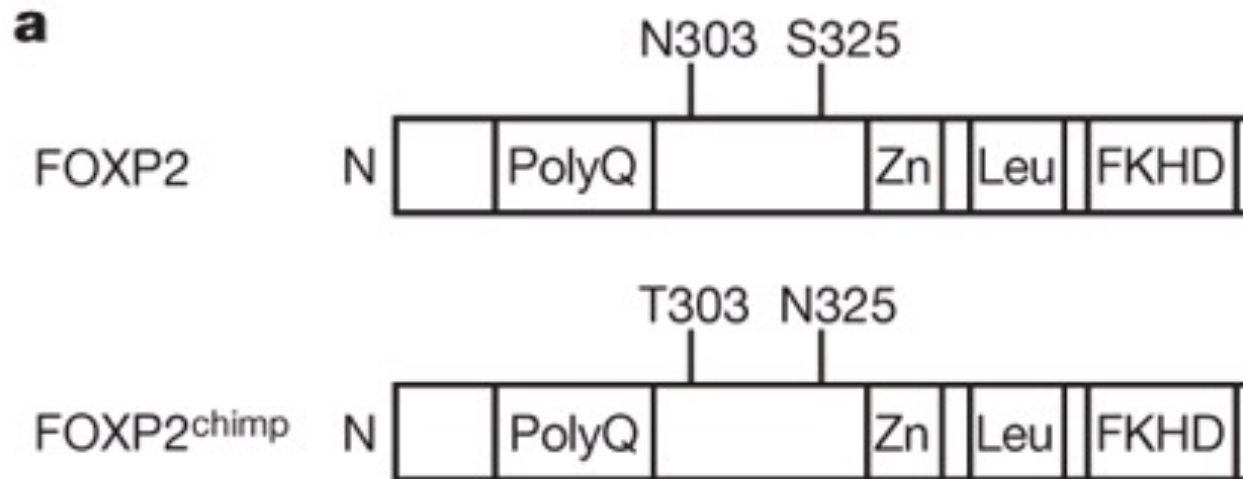
Comparative Genomics: a test case → FOXP2

Gene alignments: FOXP2



Comparative Genomics: a test case → FOXP2

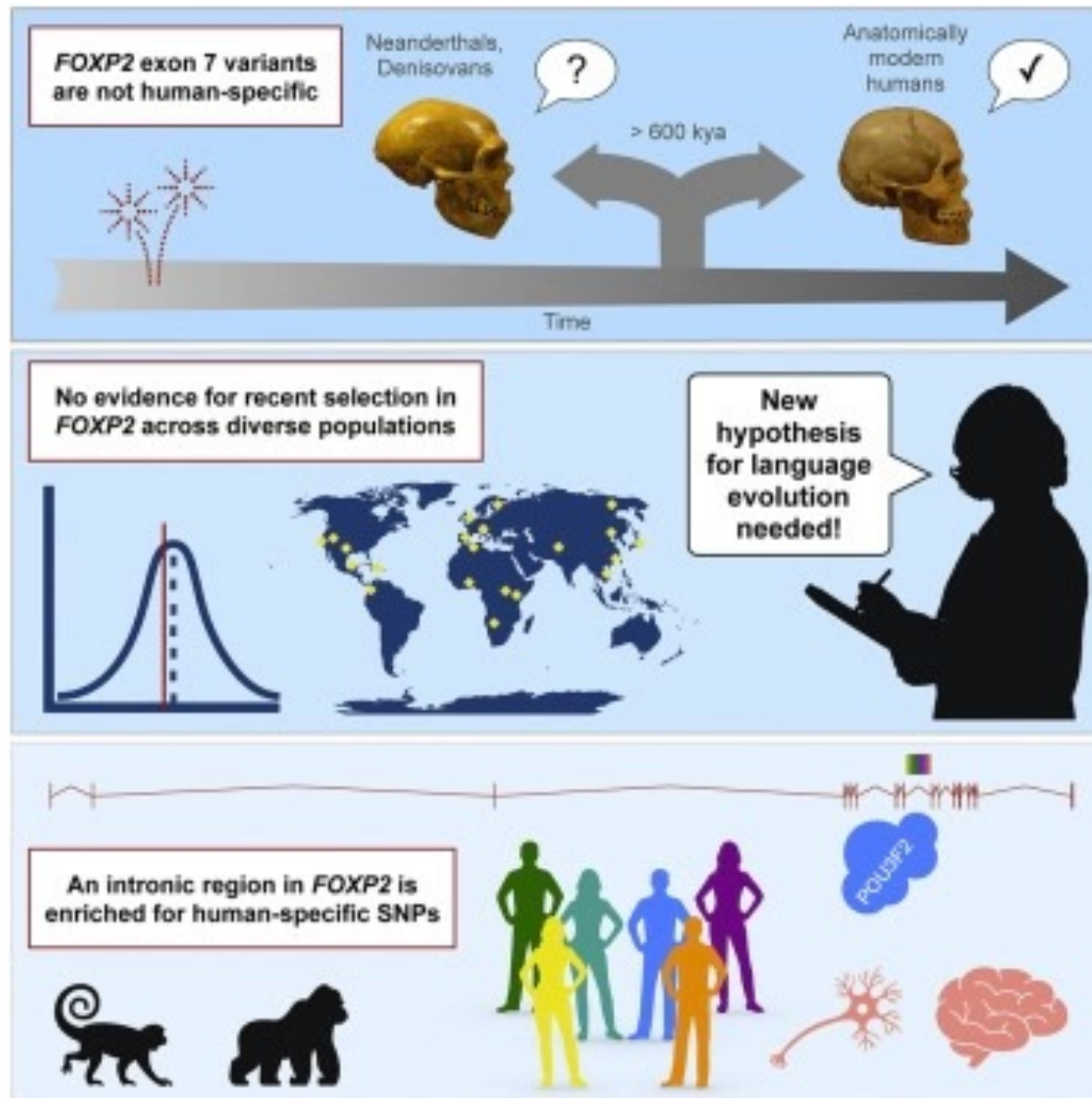
Bonkowsky and Chien, Dev. Dyn., 2005



- 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

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

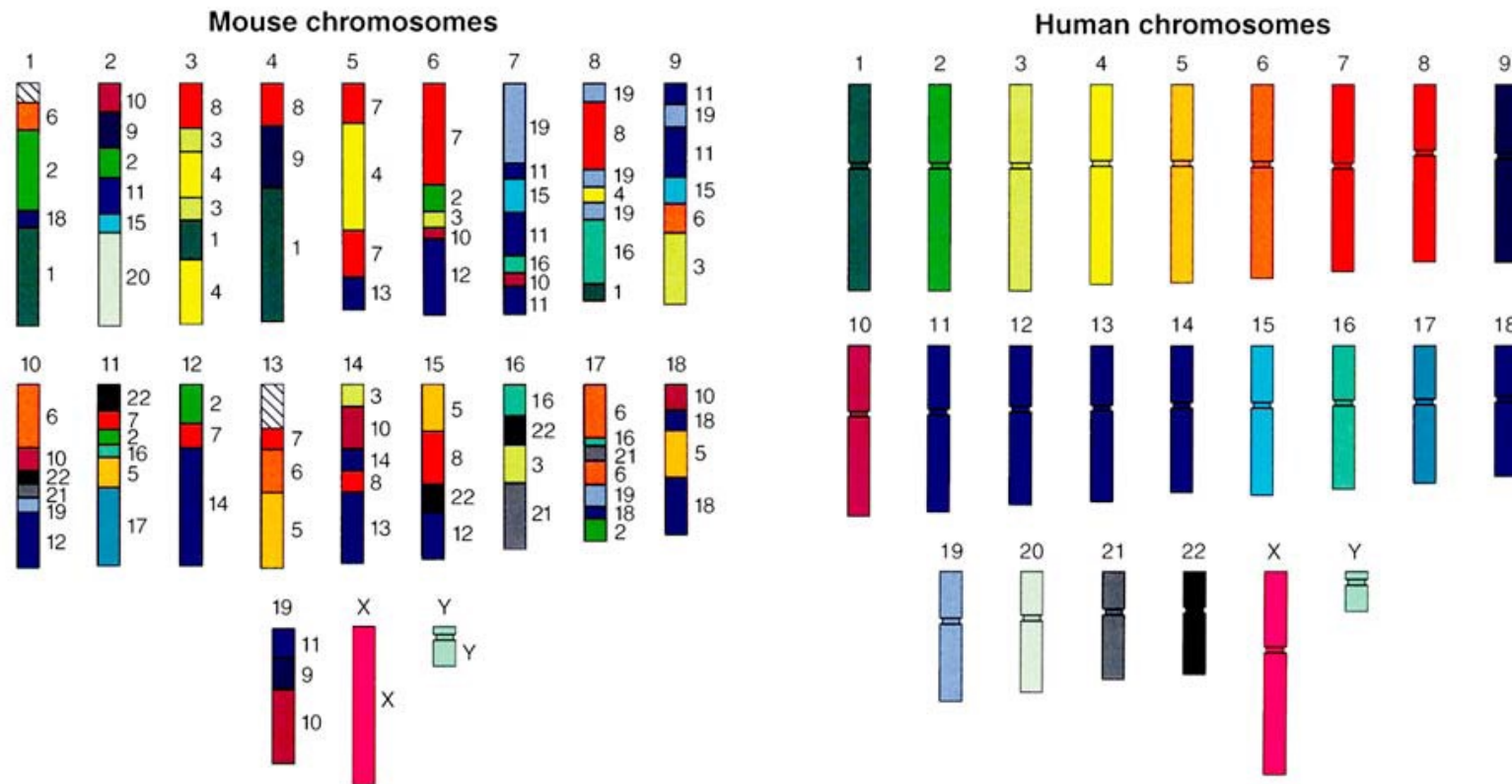
How unique are humans?

- 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

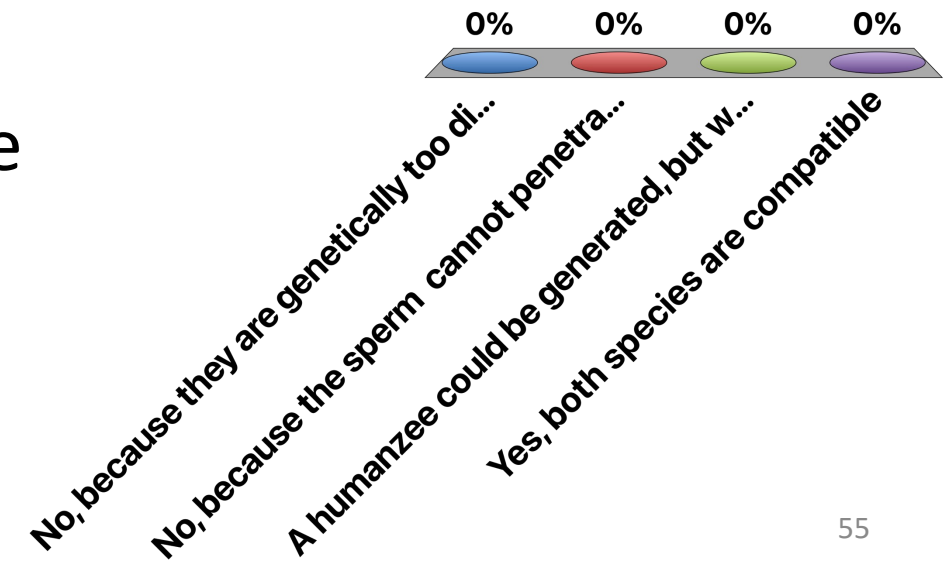
Co-localization of genes on chromosomes of different species = “Syntenic regions”



Courtesy Lisa Stubbs
Oak Ridge National Laboratory

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

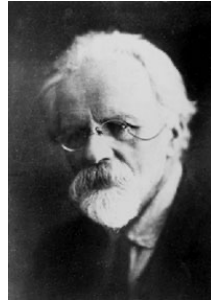
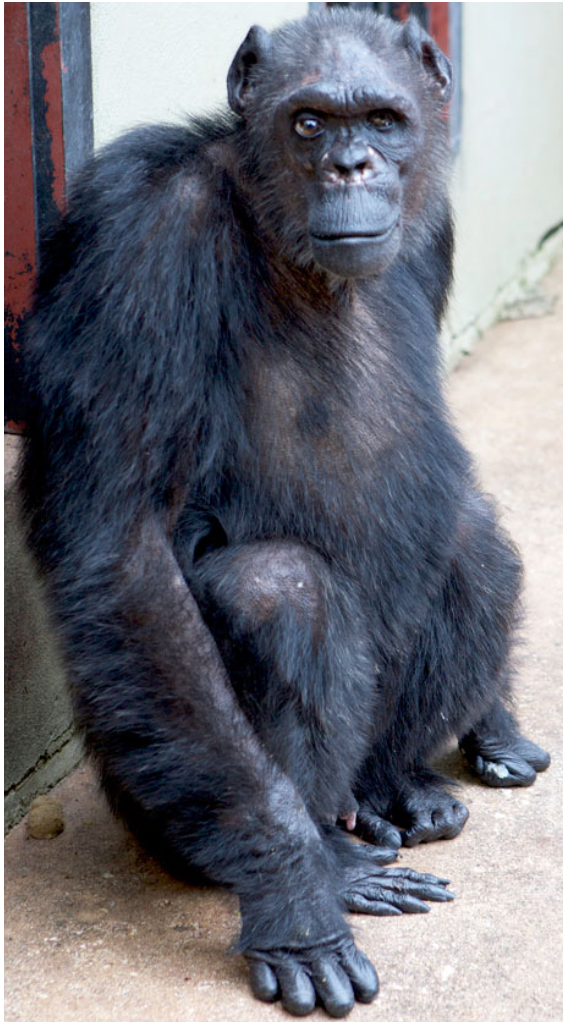
- A. No, because they are genetically too distinct
- B. No, because the sperm cannot penetrate the egg membrane
- C. A humanzee could be generated, but would be sterile
- D. Yes, both species are compatible



The humanzee

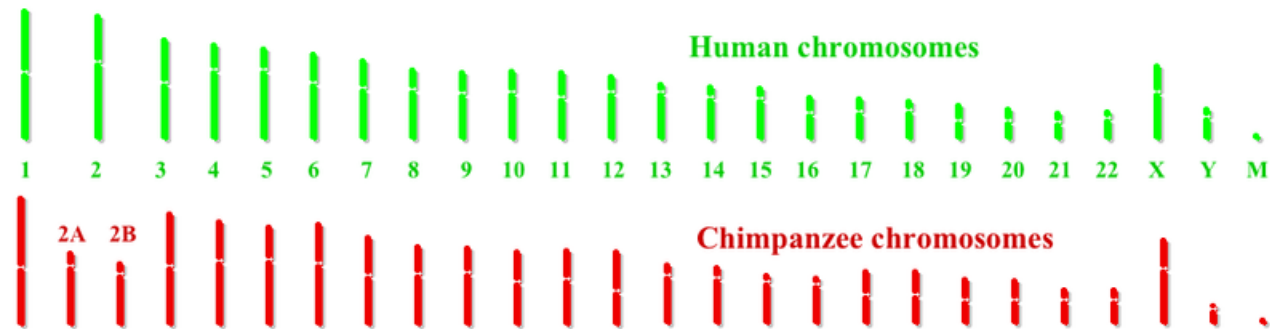
Q10

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 autosomal chromosomes) (Patterson et al., Nature, 2006)

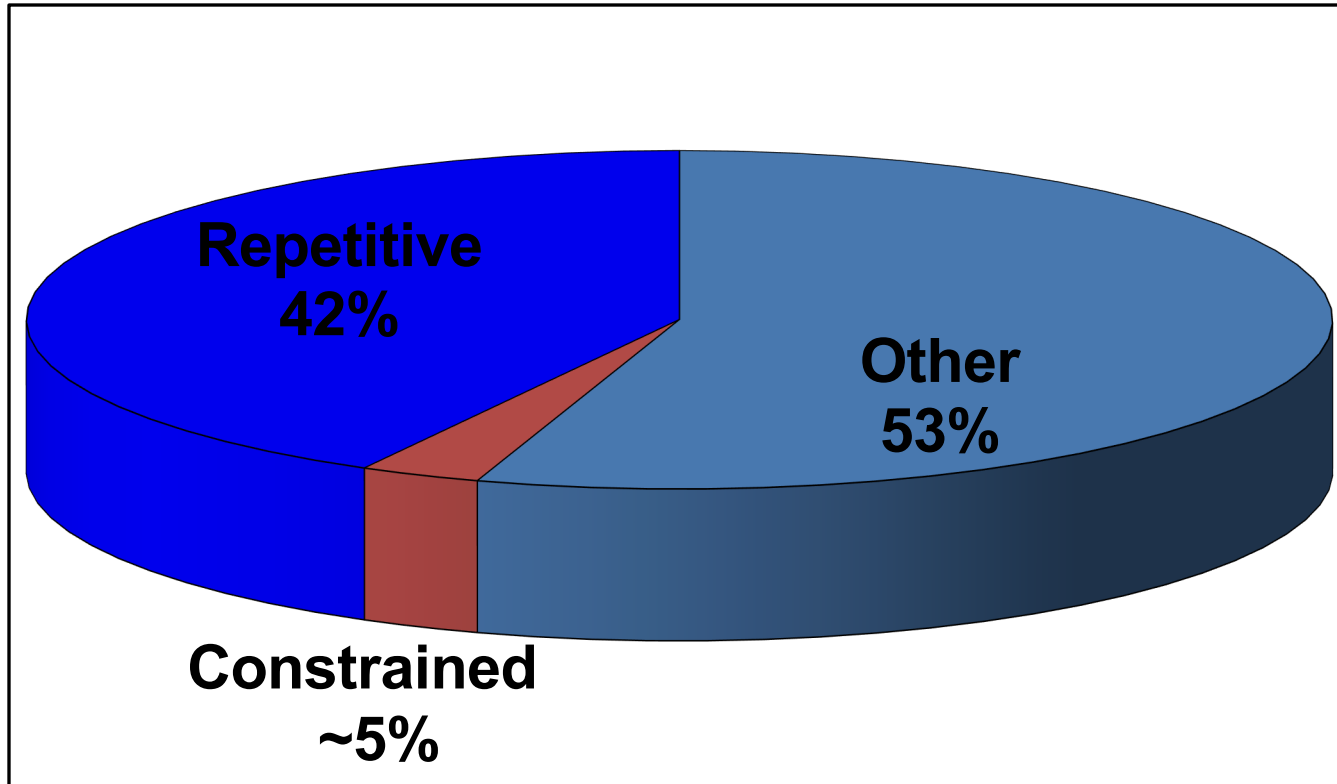


Humans have 23, apes have 24 chromosome pairs

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

What about the rest of the genome?

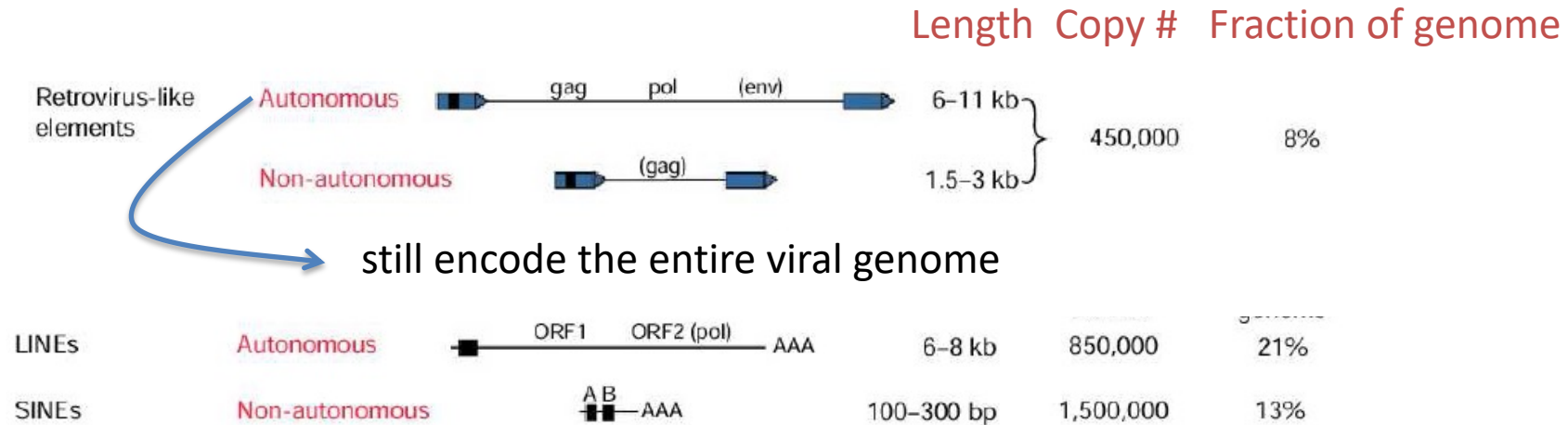
Q11



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

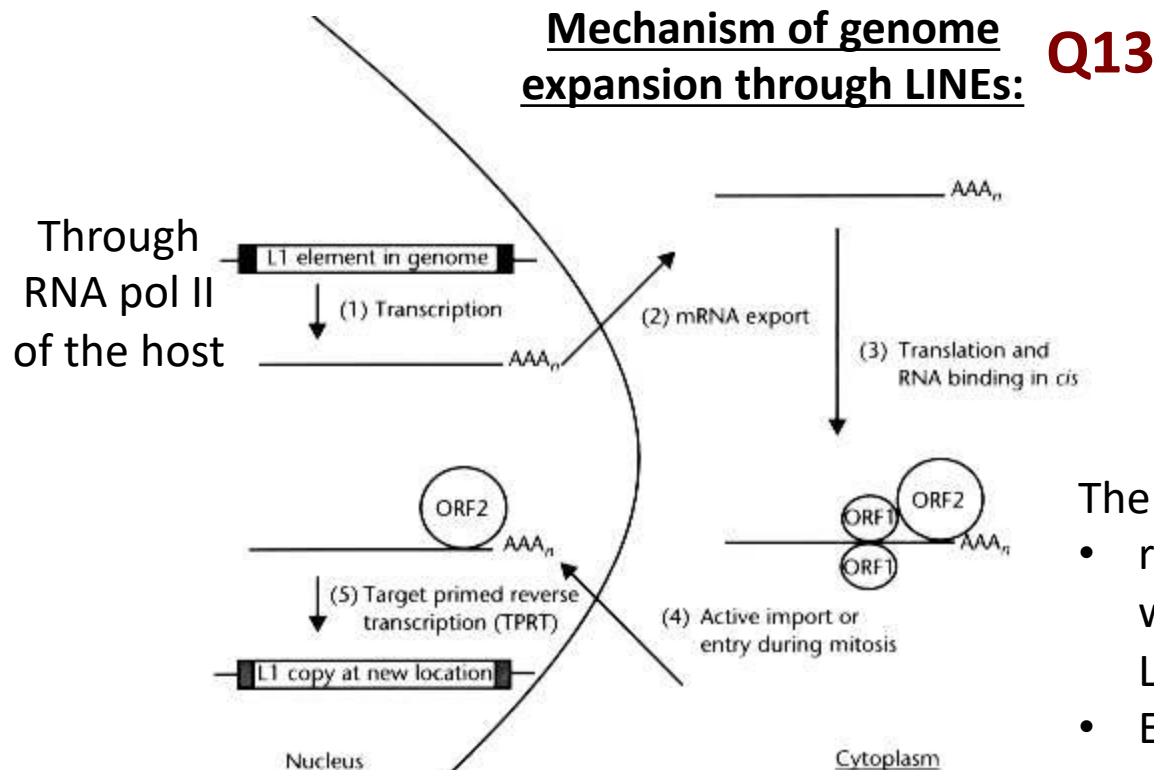
Env is the viral envelope protein

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.



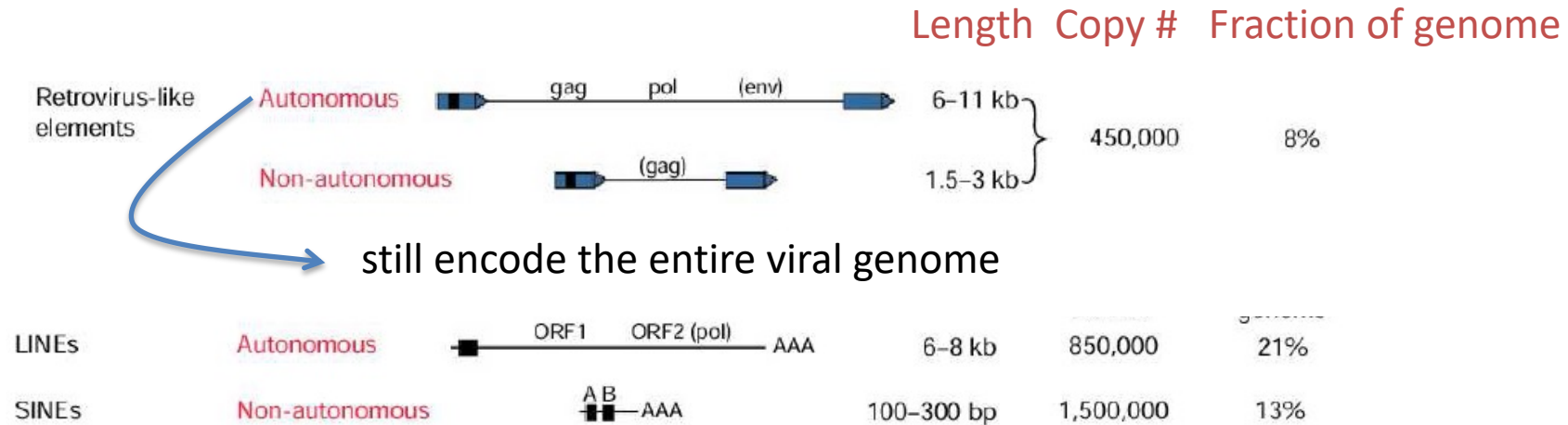
The ORFs code for:

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

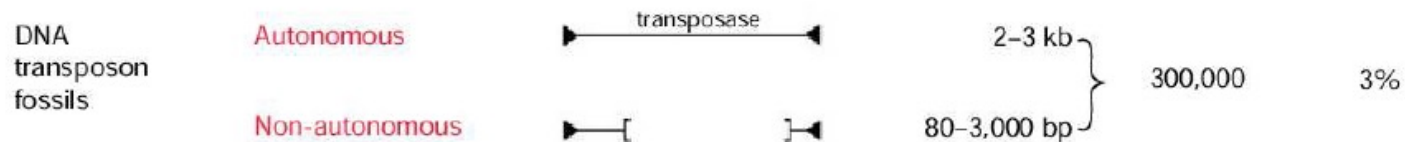
What about the rest of genome

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.)



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

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	

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.

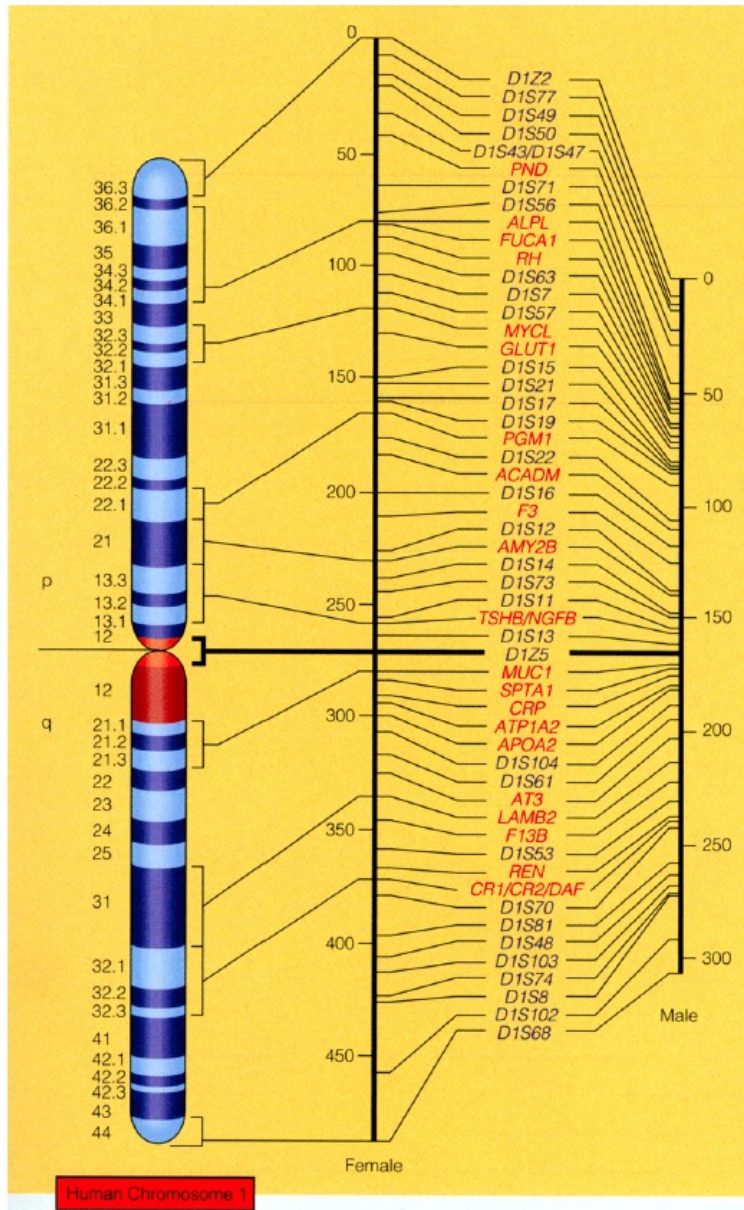
In mice, transposable elements are much more active: 10% of spontaneous mutations causing a noticeable effect are due to transposable elements.



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

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
<i>Microsatellite</i>	30–300 bp	2–10	200,000	10^{-3}
<i>Multilocus Minisatellite</i>	1–20 kb	2–10	30,000	10^{-3}

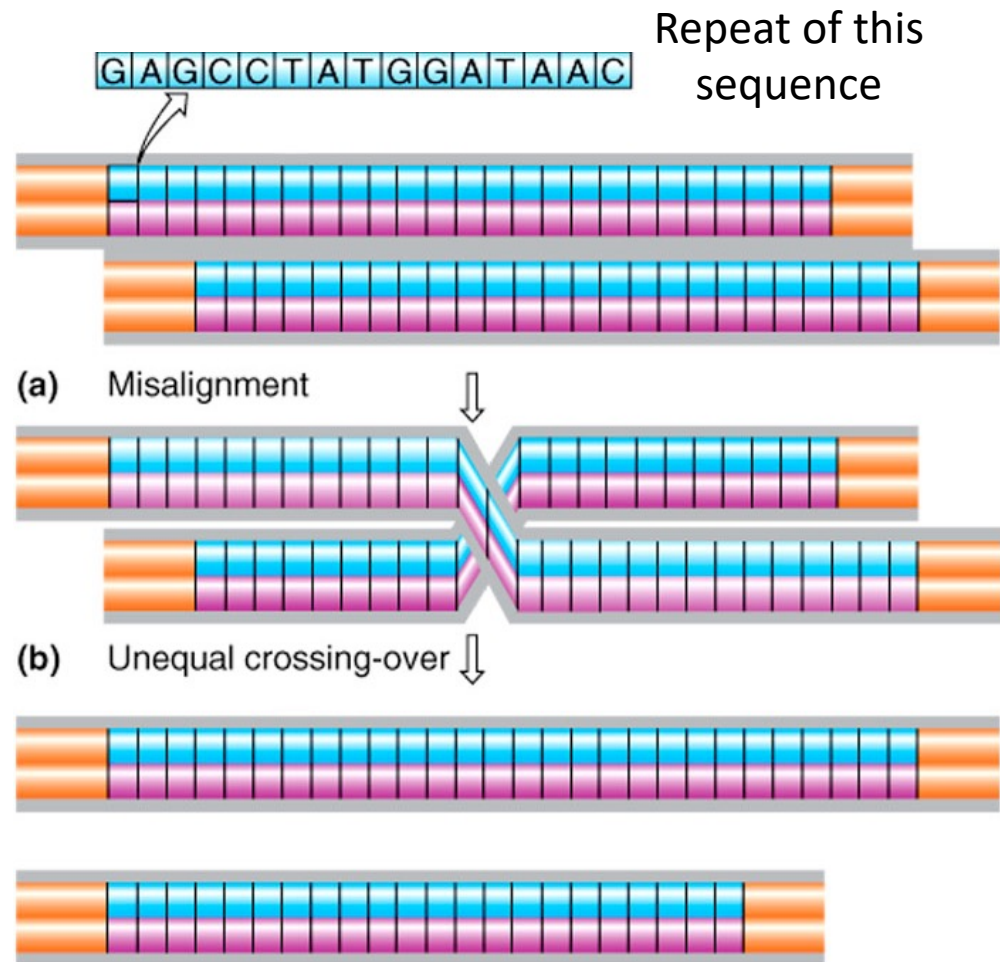
2–5 bp
repeat unit

10–100 bp
repeat unit

What about the rest of genome

Repeat elements

minisatellite



Recombinant products

The two alleles have now a differentially sized minisatellite

What about the rest of genome

Repeat elements: disease

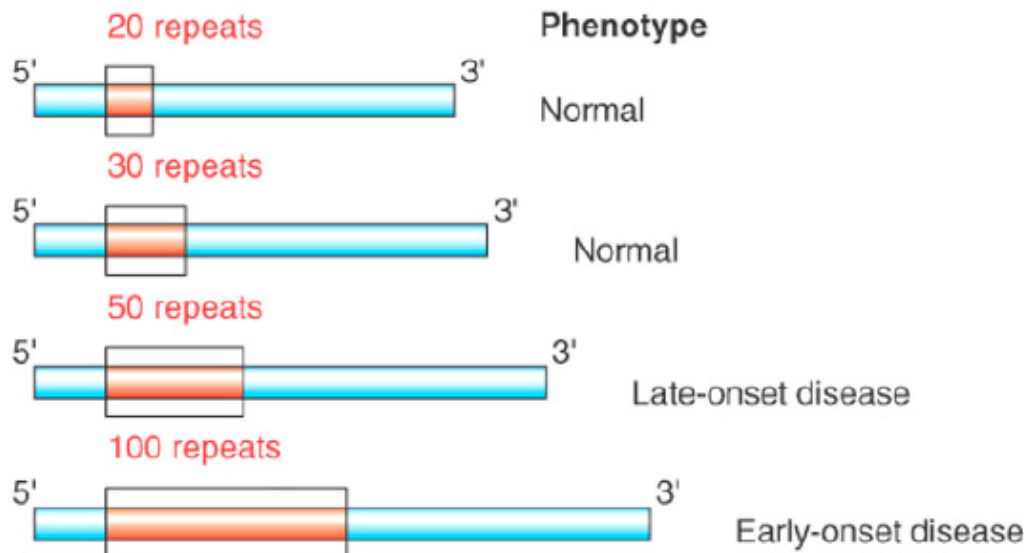
Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

(a) Basic structure of the *HD* gene



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

(b) Some alleles at the *HD* locus



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

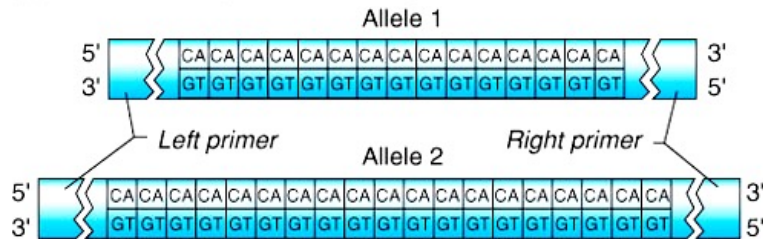
What about the rest of genome

Q15 Repeat elements: PCR-based DNA profiling

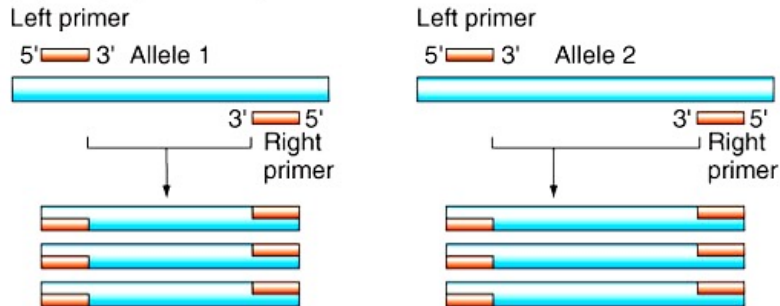
Increased sensitivity and allowed automation

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

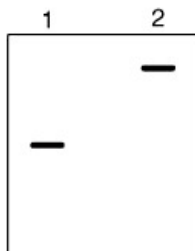
(a) Determine sequences flanking microsatellites.



(b) Amplify alleles by PCR.



(c) Analyze PCR products by gel electrophoresis and staining.

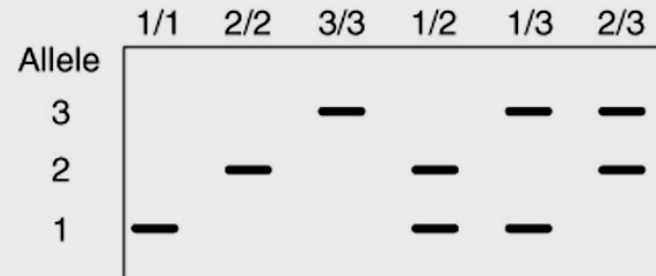


Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

(d) Example of population with three alleles.



Six diploid genotypes are present in this population.



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”?