ARISTOTLE
UNIVERSITY OF
THESSALONIKI

OPEN
COURSES

Special Topics on Genetics

Section 2: Structural Genomics

Triantafyllidis A.
School of Biology

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ
2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

License

Funding

- The offered educational material has been developed as part of the educational work of the Instructor.
- The project "Open Academic Courses at Aristotle University of Thessaloniki" has financially supported only the reorganization of the educational material.
- The project is implemented under the Operational Program "Education and Lifelong Learning" and is co-funded by the European Union (European Social Fund) and national resources.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ
2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Licensing of figures

We warmly thank the Pearson Education Inc for granting the right to use the following figures of this presentation:
Figures: 4,5

These figures come from the book Peter Russell, iGenetics: A Mendelian approach, 2006, Pearson Education Inc, publishing as Benjamin Cummings.

Section Contents

- Aims of study
- DNA – The carrier of genetic information
- Problems in sequencing and treatment methods
- Chromosomal, Genetic, physical maps and sequencing maps (and Exercises)
- New sequencing machines
- Assembling and storing data
- The future of analyses
- Finding genes in genomes sequences

Aims of study 1

- Strategies and challenges in the analysis of the genome
- The problems arising from the presence of the polymorphism in the genomes
- The study of three different genetic maps: high resolution linkage maps, large scale physical maps and sequencing maps

Aims of study 2

The main conclusions arising from integrated or semi-integrated programs of genomes sequencing of different species, such as :

- the number and type of genes,
- the percentage of repetitive sequences,
- the organization and the structure of genomes,
- their evolution, and
- the future directions of research

Aims of study 3

The techniques and methods of genomic analysis on a large scale, as :

- The automated sequencing machines (sequencers),
- Microarrays,
- Mass spectrometers.

DNA – Nucleic acids Carrier of genetic information (1/5)

Any molecule - Information Carrier should have the following properties:

1. Storage - Coding of Information
2. Doubling - Reproduction of information
3. Transcription-Transfer of information
4. Creation of Diversity - Evolution

DNA is such a molecule

DNA – Nucleic acids
Carrier of genetic information (2/5)

The C value paradox

C Value : The total amount of DNA in each haploid cell

Stable, **characteristic** for each organism

The most advanced organisms do not necessarily have higher C values, e.g. a protist, *Polychaos dubium* (former *Amoeba dubia)* with a 670 Gb genome has a >200 times larger genome than human.

http://www.genomesize.com/statistics.php

DNA – Nucleic acids
Carrier of genetic information (3/5)

Neither the number of protein genes is associated with complexity
For example the nematode, even though evolutionary inferior (simpler) from Drosophila, has a larger number of such genes.

Organism	# of Genes	Size of genome (Mb)
E. coli	4,288	4.64
Yeast	6,600	12.1
C. elegans	20,000	97
Drosophila	15,000	170
Arabidopsis	25,000	120
Mouse	~30,000	2,600
Human	~30,000	3,200

DNA – Nucleic acids
Carrier of genetic information (4/5)

What does actually count?

The transcription factors!

Yeast	300
Drosophila	1000
Human	3000

In more complex organisms there are:

- More complex regulatory regions
- More regulatory proteins

DNA – Nucleic acids Carrier of genetic information (5/5)

In the work of Gilad *et al*. (Nature 2006 440 242-245) the expression levels of 1056 genes common between human, chimpanzee, orangutan and macaque were analyzed. It was revealed that genes which encode transcription factors are highly active in humans. Some others have similar expression levels and therefore are necessary for all primates.

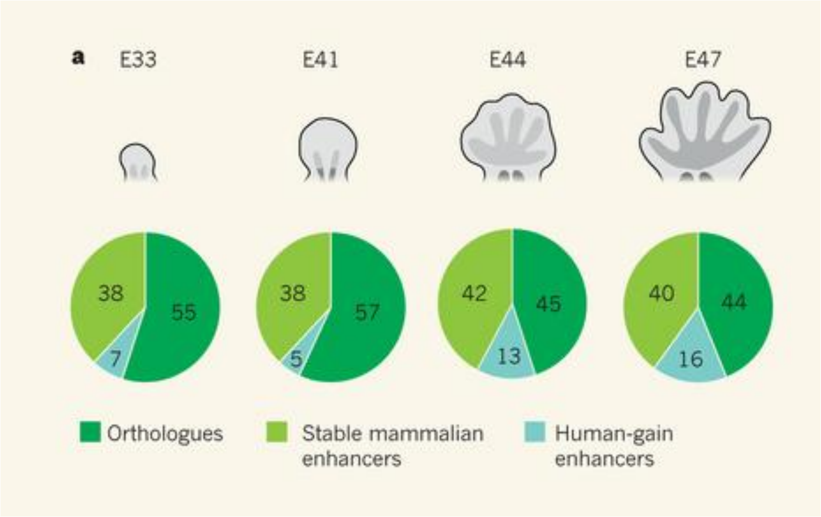


Figure 1: The evolution of genes' enhancers which are associated with the development of the limbs as studied for the species: mouse, macaque, human. Beyond the common genes, there are more new in human (Cotney *et al*. 2013 Cell, 154, 185-196)

Genome analysis

Researchers, in their efforts to study the genome of an organism, are interested in two main elements:

- **Finding the sequence** across all the chromosomes of a haploid genome.
- **Uncovering the diversity** or otherwise the existence of different alleles found among different individuals of the same species or in both series of homologous chromosomes (if we refer to a diploid organism).

Problems in sequencing (1/5)

Problems that may arise during sequencing are due to:

- Errors in the sequencing process
- The polymorphism within genomes
- Repeated sequences which are ample and identified with difficulty
- DNA segments that cannot be cloned and therefore cannot be sequenced

Problems in sequencing (2/5)

The size of a chromosome varies (e.g. in humans from 700kb to 500 Mb). However, the current technology does not allow the successful read of a sequence with a single continuous attempt for DNA fragments above 600 bp

How can the successful sequencing of a chromosome (500 Mb long) be possible with a rate of 600 bases each time?

How accurate is the result of sequencing when the error rate is about 1% per read?

☞ The HGP decided that the final deposited sequence of human genome should have an error rate not higher than 1 / 10.000 bp or 0,01%

Problems in sequencing (3/5)

Polymorphism

Higher organisms are diploid. Therefore genomes include both maternal and paternal chromosomes → different alleles of parents, create polymorphism

In human the polymorphism was calculated to be 1/500 bp

How is it possible to discern polymorphism from sequencing errors?

Problems in sequencing (4/5)

Repeated sequences

The genomes of higher organisms include numerous repeated sequences. These sequences may have a size of several tens of bases e.g. (GT)n or hundreds of kilobases.

When they are created (by doubling or independent assortment), they resemble each other. Over the years they differentiate because of mutations and degeneration. New sequences are very similar, while the oldest ones differ.

The problem is therefore big for new repeated sequences (which are very similar and are interspersed in the genome) of a size of > 600 bases. If a sequencing "read“ meets a sequence dispersed in 10,000 copies in different chromosomes, how can we identify which specific copy is actually analyzed and in which chromosome it is located?

Problems in sequencing (5/5)

DNA segments that cannot be cloned

The heterochromatic regions of the DNA pose the biggest problem since they are not easily cloned into cloning vectors.

Although heterochromatin contains very few genes it actually constitutes approximately 30% of the total DNA.

If these regions cannot be cloned, how will we know what they actually contain?

Troubleshooting methods (1/3)

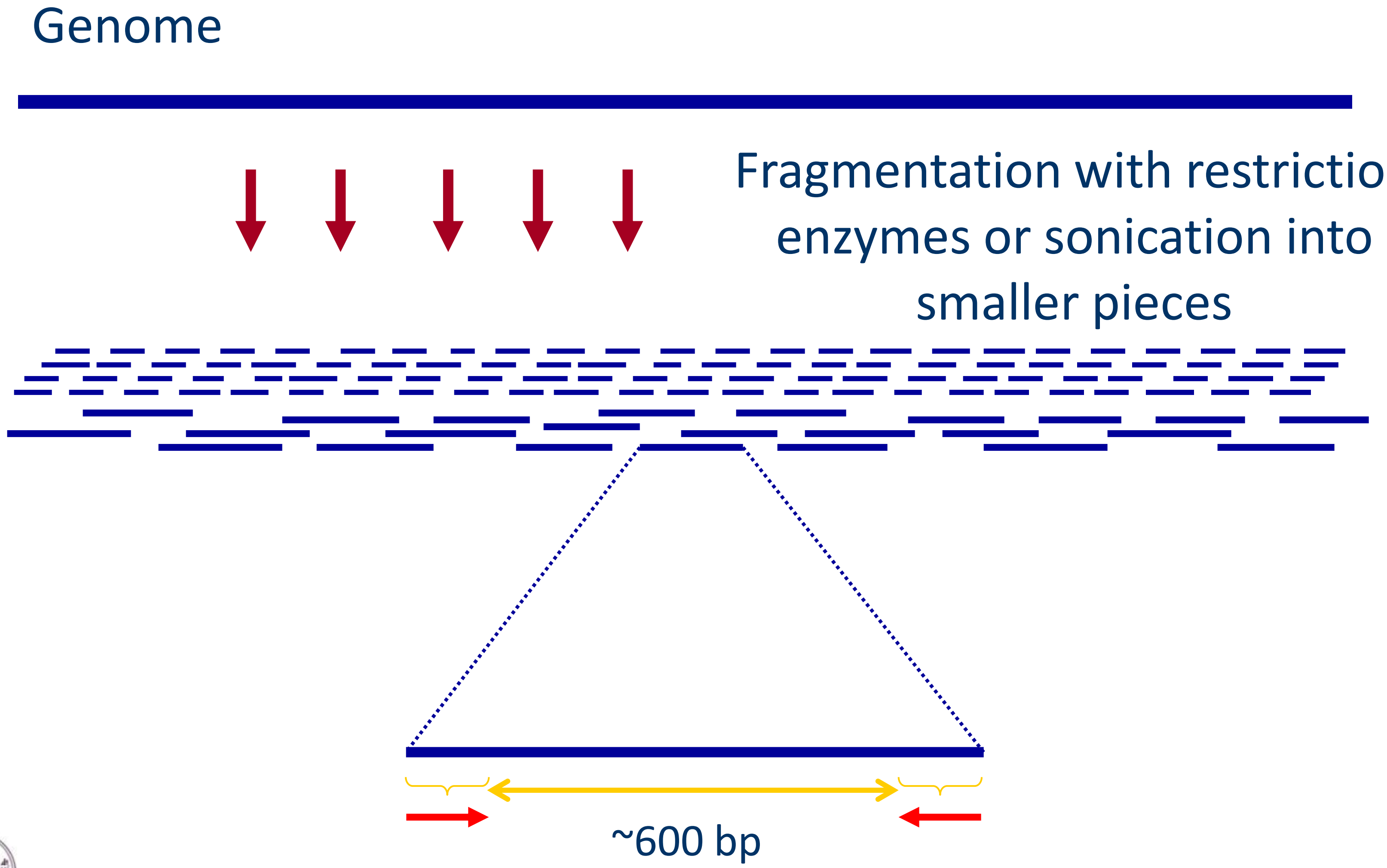
Scientists follow a “divide and conquer” approach

The chromosomes are broken into small and overlapping fragments and are cloned into libraries.

Sequencing can be achieved along almost all fragments. The overlapping of the fragments facilitates to reassemble the individual fragments in a chromosome.

The reassembly is not easy and certainly cannot be accomplished without the use of computers.

Troubleshooting methods (2/3)



Genome

Fragmentation with restriction enzymes or sonication into smaller pieces

~600 bp

Troubleshooting methods (3/3)

To achieve an error rate of 1/10,000, each DNA sequence fragment has to be sequenced multiple times from different clones. At least 10 different clones need to be sequenced→ tenfold cover of sequence genome.

Sequencing errors will occur in 1/10 reads, while the polymophisms in half (5/10) the reads.

However the drawback is that the sequence data which are processed by computers in order to assemble the genome are multiplied ten-fold, since each region is sequenced 10 times.

Two Basic Methods of Sequencing

Whole genome shotgun sequencing → Random digestion and sequencing of DNA fragments (Celera Company). It is faster but not very accurate.

Hierarchical shotgun sequencing → Sequencing of DNA fragments that the researchers had already placed on physical maps (HGP Program). It is extremely accurate but requires very costly and time-consuming preparation.

Techniques for genome analysis 1

Cloning: DNA size of 500 - 1000000 bp is cloned in various vectors. All cloned DNA fragments constitute a **library**.

In a library with good coverage, the DNA sequences should be encountered in at least 10 independent clones. So it will be possible to read the genome 10 independent times and avoid mistakes.

Techniques for genome analysis 2

Hybridization : It is used to identify the precise location of DNA fragments in a library. It is based on complementarity of a target DNA sequence with a DNA fragment from the library.

The clones that contain genes or are part of a chromosome are identified. Thus, the mapping and the continuation of sequencing is facilitated.

Techniques for genome analysis 3

PCR Amplification: Amplification of DNA fragments over one million times with sizes from 1 Kb to 20 Kb.

Primers specific for the particular sequence are required, otherwise multiple different fragments are amplified at the same time and problems in sequences are created.

DNA Sequencing: The sequencers read sequences of cloned DNA fragments of various sizes (from 50 bp to >600 bp).

Techniques for genome analysis 4

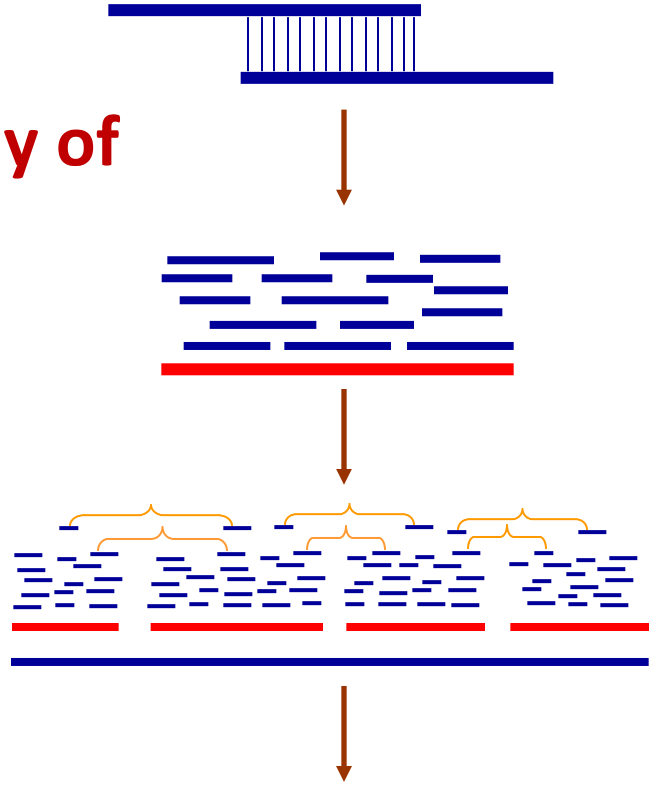
Computational tools

There are 4 different computer programs, those which:

1. Check if the sequence looks similar to already sequenced DNA fragments (available i.e. In databases)
2. Discover overlapping between DNA fragments and assemble them in a series
3. Estimate the error rate in the various sequences
4. Discover and identify genes on chromosomes

Computational tools

Computer programs for the assembly of the genome



PHRED, PHRAP, LUCY,
TGICL, CAP3, GAP4

..ACGATTACAATAGGTT..

http://seqanswers.com/wiki/Software/list,
http://www.jcvi.org/cms/research/software/

Chromosomal Maps (1/2)

They show the location of the genes, genetic markers, centromeres, telomeres and other sites along chromosomes.

They include the genetic maps (linkage maps), physical maps and sequence maps **on a large scale.**

The genome analysis process comprises three stages: Firstly the creation of a linkage map. Then the creation of a a physical map. Finally, the creation of the sequence map from the combination of the above.

Chromosomal Maps (2/2)

Genetic maps (linkage maps) – show the relative position of genes without knowing the actual distance. Based on the frequency of recombination between neighboring genes / genetic markers.

Physical maps – show the real distance between genes in bp.

Sequence maps – consist of the entire sequence of nucleotides in a gene or generally in a DNA segment

Click here for various types of maps in barley :

http://pgrc.ipk-gatersleben.de/kuenzel/barleymap.html

Genetic Maps of High Resolution (1/8)

- Created based on recombination frequency
- Distance of 1 centiMorgan (cM) corresponds to a chromosomal distance giving recombination rate of 1%
- In simple linkage maps, geneticists map the position of only a small number of genes that are quite close, through simple crossings, and some easily identifiable morphological or biochemical phenotypes are examined
- In high resolution maps, the known phenotypes are not enough in order to find genes that are closely spaced along chromosomes
- Also, there may be problems with phenotypes, whose inheritance is multigenic
- Controlled crossings of two and three points are not always possible because of ethical dilemmas, especially in humans
- It is therefore necessary to use genetic markers where tests are done on their molecular phenotype ie. their genotype

Genetic Maps of High Resolution (2/8)

Genetic Markers

- They do not necessarily have a functional role
- They are not transcribed (they are not always genes)
- They present high polymorphism (many alleles)
- There are millions of such DNA polymorphisms dispersed in all chromosomes
- There are two main genetic markers that are used for high resolution mapping
- Microsatellite DNA, Simple Sequence Repeats-**SSRs**
- **Single Nucleotide Polymorphisms-SNPs**

Genetic Maps of High Resolution (3/8)

Microsatellite DNA-SSRs

It is composed of multiple (4-50) repeats of 2-6 nucleotides (e.g. AGAGAGAG… ή GCAGCAGCAGCA…)

They are multi - allelic, since during DNA replication mistakes are made and one repeat is lost or added (e.g. AG or GCA respectively)

The alleles are determined by the number of repeats, after electrophoresis

Typically one locus is found every 2-30 Kb in the genome

Genetic Maps of High Resolution (4/8)

Single Nucleotide Polymorphisms -SNPs

- They are simple point mutations
- They are used a lot during recent years
- Frequency every 1/100 to 1/300 bases
- Genotyping costs 20-30 cents (with automatic machines based on GEL or PCR)

Genetic Maps of High Resolution (5/8)

Microsatellite DNA:

TGAACCGTTACTCG**TATATATATA**GCGTATGCT

↓

TGAACCGTTACTCG**TATATATATATATATA**GCGTATGCT

Polymorphism results from the different number of repetitions of dinucleotide TA

Single Nucleotide Polymorphisms

CGTTACT**T**GGTAACG

↓

CGTTACT**G**GGTAACG

Polymorphism: where there is simple substitution of one of the four nucleotides

Genetic Maps of High Resolution (6/8)

SNP Genotyping

Many different methods have been developed for the genotyping of Single Nucleotide Polymorphisms, some of which are listed below:

◈ Hybridization methods (microarrays, TaqMan, Molecular Beacons)

◈ Allele-specific PCR

◈ Analysis by restriction enzymes

☞ Each method has both **advantages** and **disadvantages**. So every time it is up to the researcher to choose the appropriate method to achieve its objectives.

http://www.nature.com/nrg/journal/v2/n12/full/nrg1201-930a.html

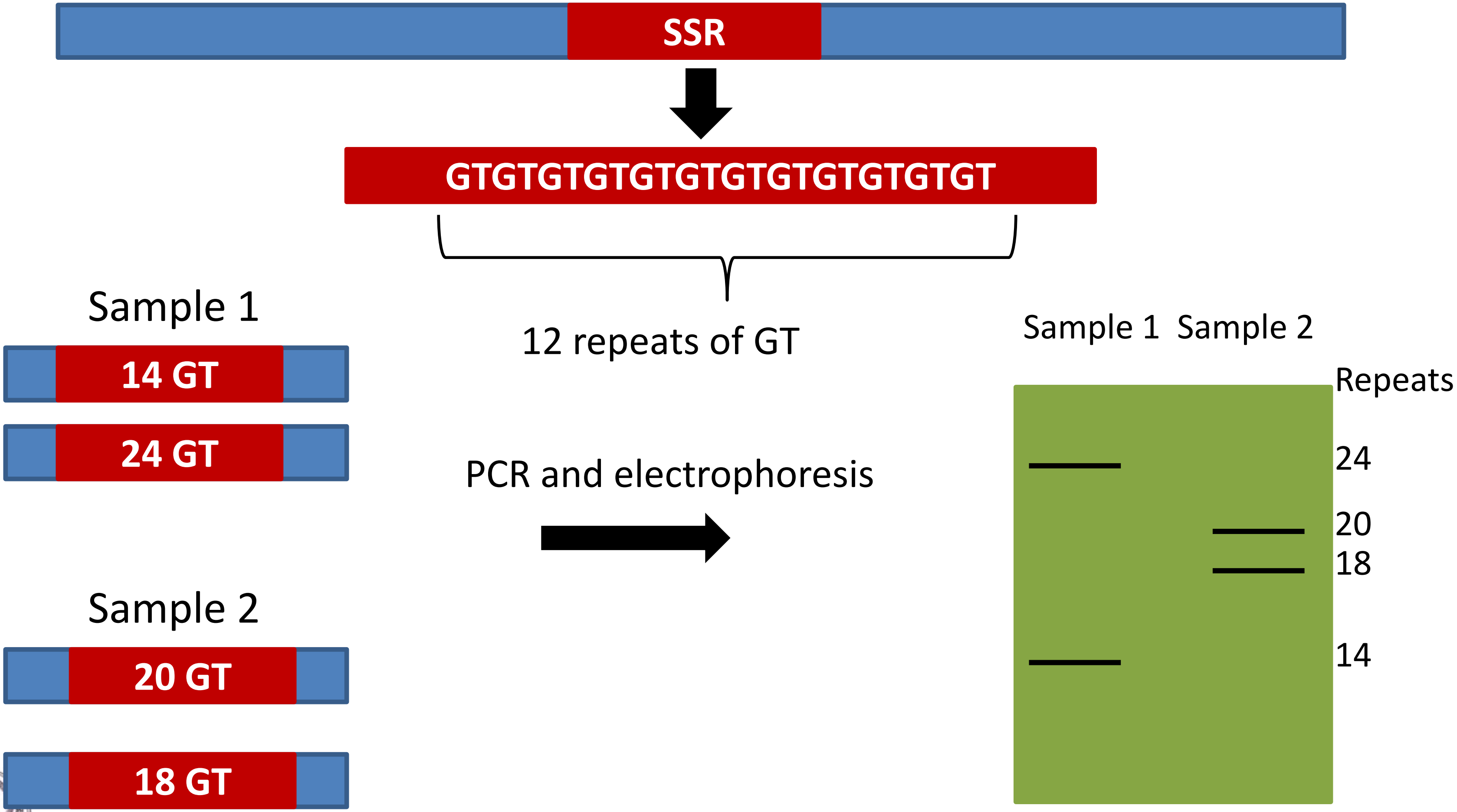
Genetic Maps of High Resolution (7/8)

Laboratory Methods for High-Throughput Genotyping
Howard J. Edenberg and Yunlong Liu
Genetics of Complex Human Diseases: A Laboratory Manual (eds. Al-Chalabi and Almasy). CSHL Press, Cold Spring Harbor, NY, USA, 2009.

http://cshprotocols.cshlp.org/content/2009/11/pdb.top62.long:
Figure 1: Different technologies (serial/parallel) for SNP genotyping for different types of projects
Figure 2: Examples of results of automatic SNP genotyping by using MassArray

Genetic Maps of High Resolution (8/8)

SSRs Genotyping



SSR

GTGTGTGTGTGTGTGTGTGTGTGT

12 repeats of GT

Sample 1

14 GT

24 GT

Sample 2

20 GT

18 GT

PCR and electrophoresis

Sample 1 Sample 2

Repeats

24

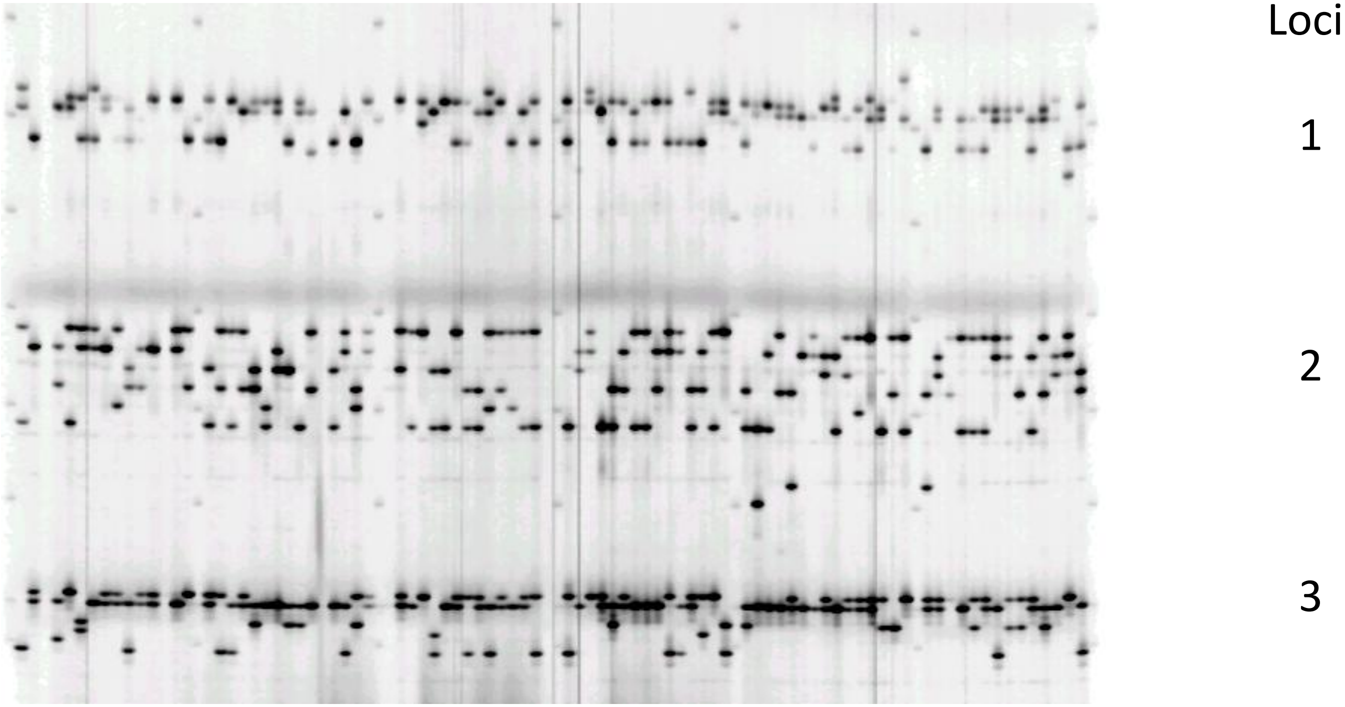
20

18

14

Genetic mapping with SSRs (1/3)

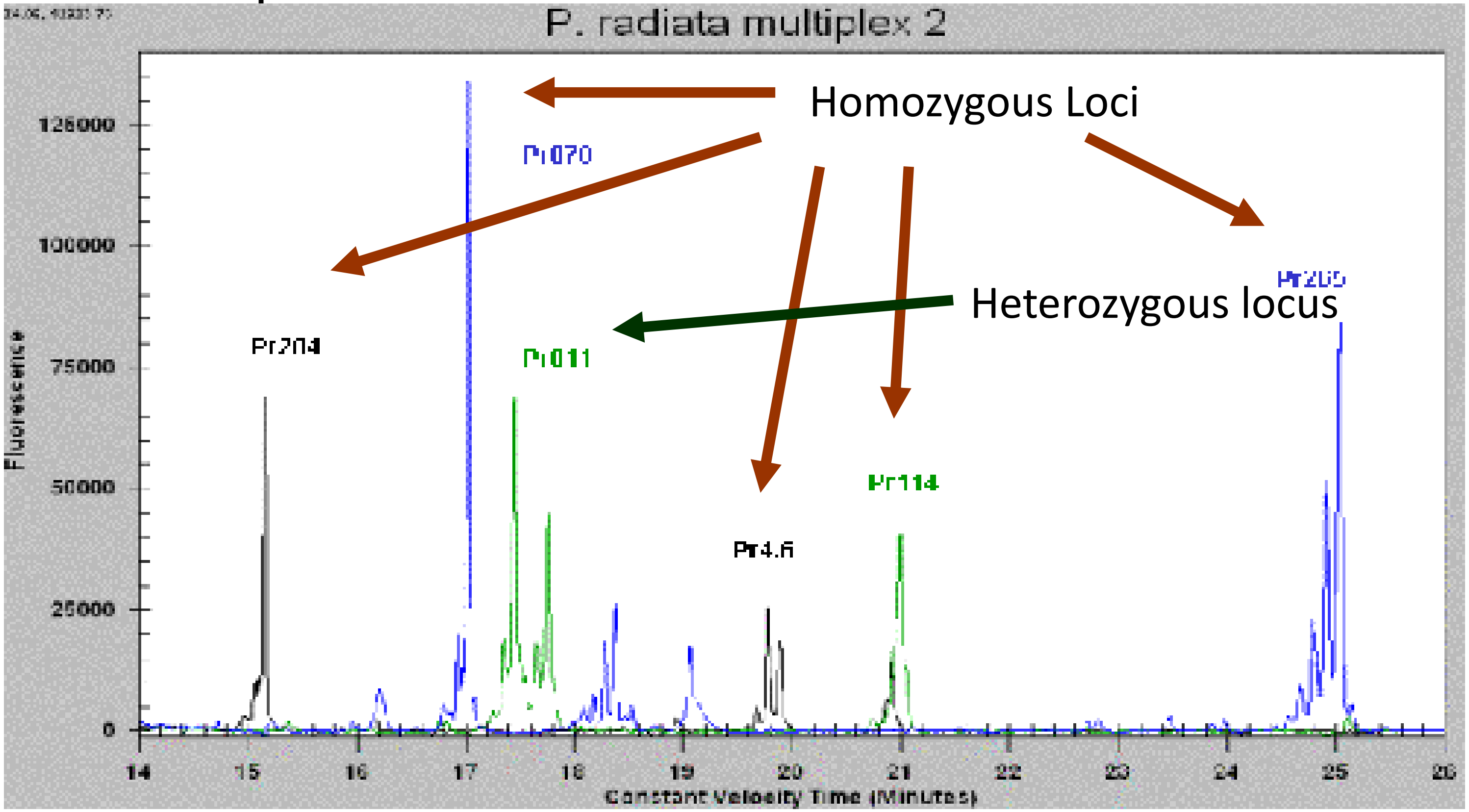
Figure 2: Electrophoresis results of three SSR loci in polyacrylamide gel electrophoresis. In each column, the PCR results after amplifying DNA from one individual at three loci have been electrophoresed. There are almost 100 individuals in this gel (i.e. 100 individuals X 3 loci = 300 genotypes)



Genetic mapping with SSRs (2/3)

Much faster process is the genotyping of hundreds to thousands of SSRs using automatic machines.

Εικόνα 3: The profile of an individual at six SSR loci



Homozygous Loci

Heterozygous locus

Genetic mapping with SSRs (3/3)

Primers labeled with different fluorescent dyes can be used in automatic machines. Even if the sizes of the PCR product for two different SSR loci are the same, the signal at a different wavelength helps to differentiate them.

Genetic Maps of High Resolution

- It is easy to analyze linkage today, as there are numerous data in special databases and websites for genetic markers
- In human and mouse there are over 150000 and 50000 SSRs and over 40,000,000 SNPs (230 million and 70 million redundant SNPs respectively).
- Researchers use about 500 SSRs in order to find the relative position of a gene, i.e. about 1 SSR / 6 Mb.
- Today with the use of SNPs, there are at least thousands such genetic markers, so there is a thousand fold increase in the ability to map a gene

http://hapmap.ncbi.nlm.nih.gov/

http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi

Exercise 1 – Solution

- Give the genotype of three individuals in two microsatellite (SSR) loci and in two SNP loci. These individuals must have different genotypes

Note the alleles you observe.

Alleles of SNP1 : A, T

Alleles of SNP2 : A, C

Alleles of SSR1 : $(GC)_3$, $(GC)_4$, $(GC)_5$, $(GC)_7$, $(GC)_{10}$

Alleles of SSR2 : $(AAT)_2$, $(AAT)_3$, $(AAT)_4$, $(AAT)_7$

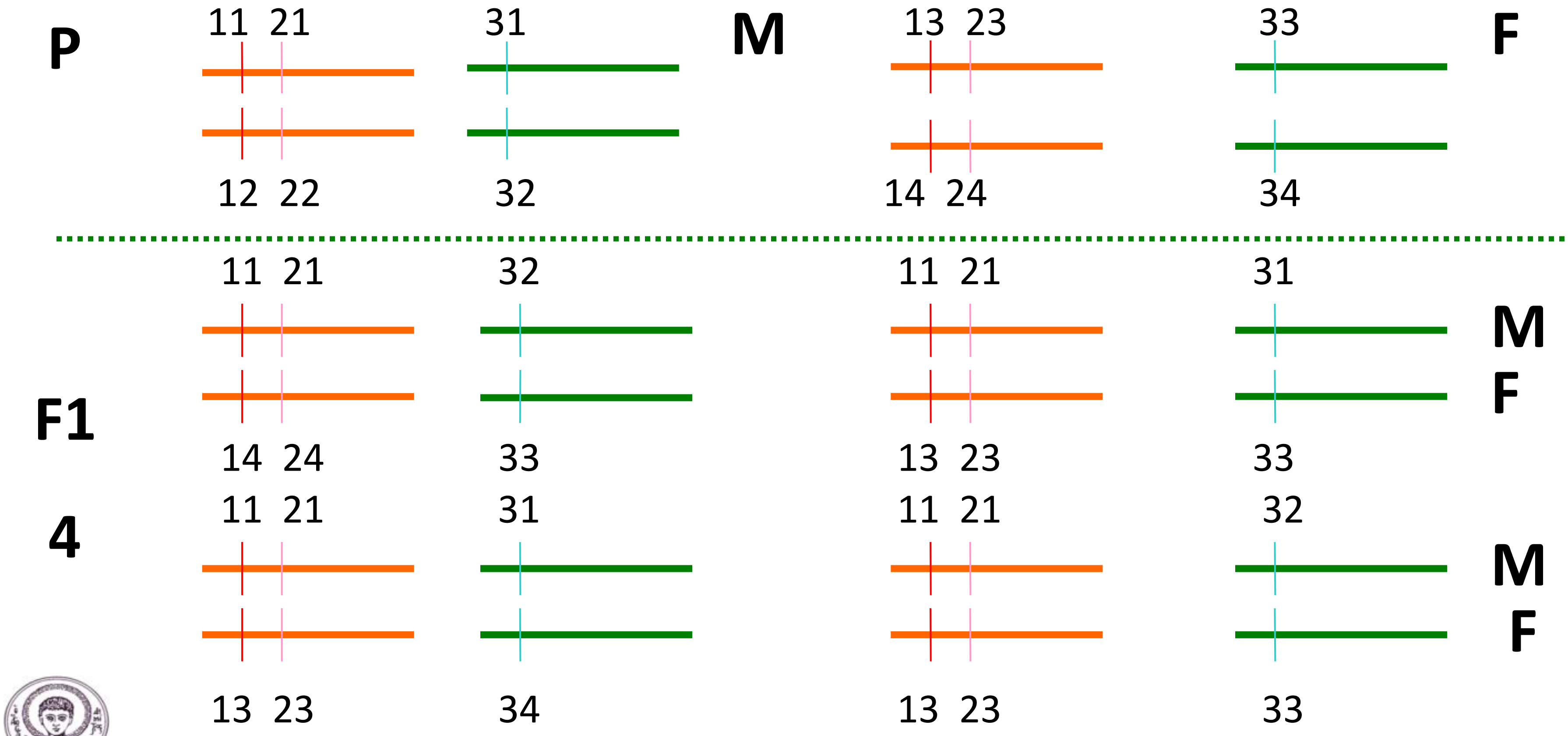
	SNP1 loci	**SNP2 loci**	**SSR1 loci**	**SSR2 loci**
Individual 1	AT	AA	$(GC)_7(GC)_4$	$(AAT)_2(AAT)_4$
Individual 2	AA	CC	$(GC)_{10}(GC)_3$	$(AAT)_3(AAT)_2$
Individual 3	TT	AC	$(GC)_5(GC)_3$	$(AAT)_4(AAT)_7$

Is there higher polymorphism in SNP or SSR loci;

Higher polymorphism is observed in SSR loci, as theoretically there is no upper bound in the number of repetitions (i.e. alleles) in one locus. On the other hand SNPs usually present only two (out of four possible) alleles

Genetic mapping with SSRs

The next step is to monitor the separation of a large number of genetic markers and co-inheritance in genealogical trees. Alleles of different loci grouped together in genealogical trees reveal the existence of linked loci (eg 1.1 and 2.1 below).



Genetic Maps of High Resolution

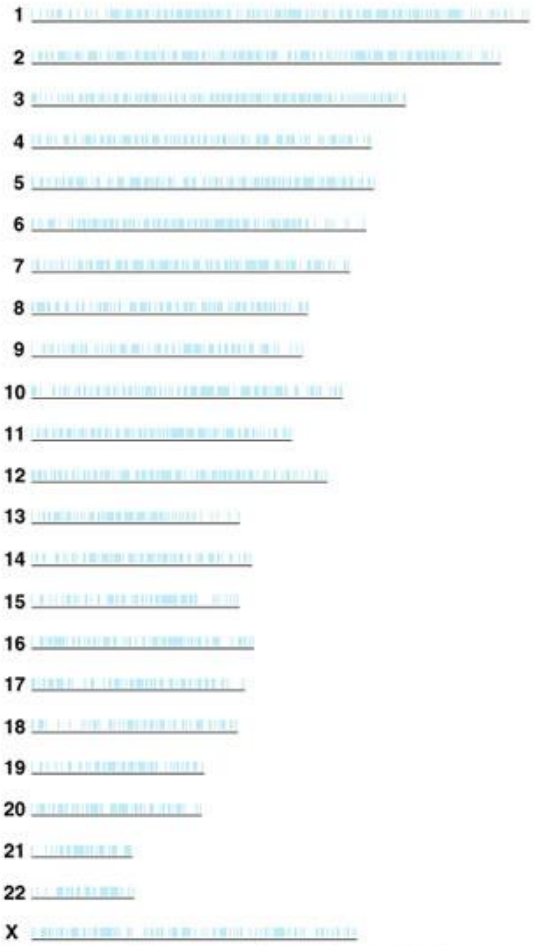
Thus, a high resolution linkage map is formed which includes the relative position of thousands, very nearby genetic markers.

http://www.funpecrp.com.br/gmr/year2007/vol3-6/gmr0340_full_text.html

Genetic Maps of High Resolution

For human, ~5000 markers were used in 500 individuals from 40 families from 3 generations each

500 X 5000 = 2,500,000 GENOTYPINGS!!!



Peter J. Russell, iGenetics: Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Figure 4: Genetic high density map of the human genome with 5,264 microsatellites (Dib *et al*. 1996 *Nature* 380: 152-154)

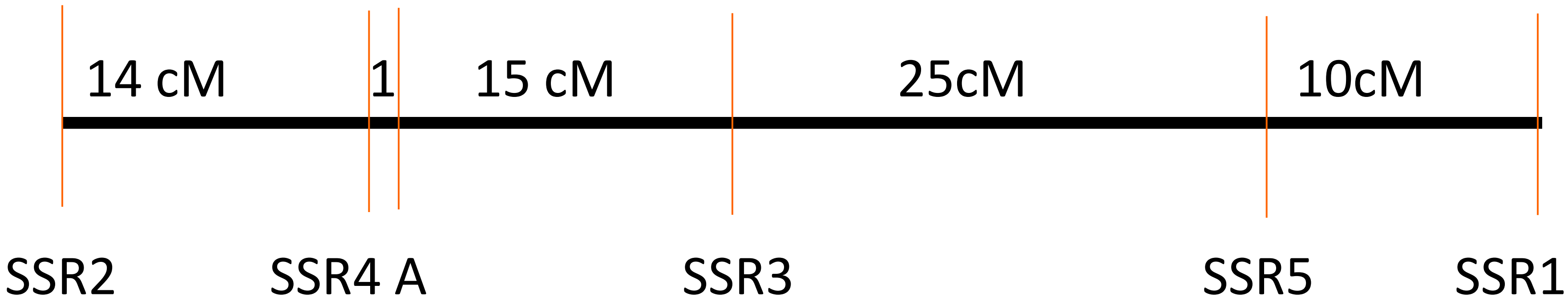
Exercise 2

The clones of a library are checked in order to find the exact position of gene A. Using SSR markers the following recombination frequencies were calculated:

Loci	% Recombination	Loci	% Recombination	Loci	% Recombination
A, SSR1	50	SSR1, SSR2	50	SSR2, SSR4	14
A, SSR2	15	SSR1, SSR3	35	SSR2, SSR5	50
A, SSR3	15	SSR1, SSR4	50	SSR3, SSR4	16
A, SSR4	1	SSR1, SSR5	10	SSR3, SSR5	25
A, SSR5	40	SSR2, SSR3	30	SSR4, SSR5	41

Draw a genetic map for the SSR markers and gene A.

Exercise 2 – Solution

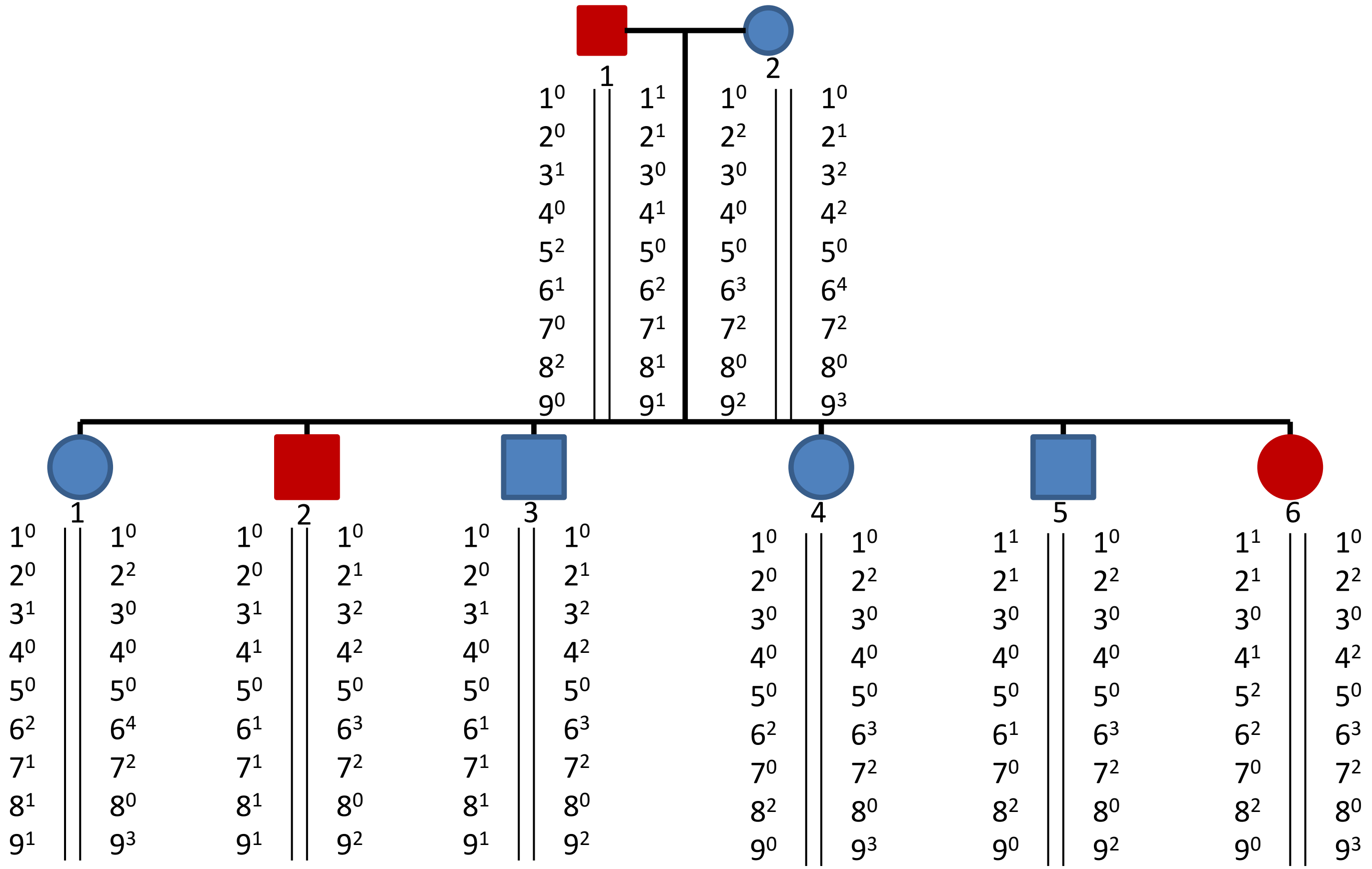


If the human genome is 3.3 Gb large and the genetic map of human includes 3300 cM what is the fragment size which is contained in this clone? *(65cm = 65 Mb)*

What is the distance in bp between gene A and the nearest SSR? *(1 cM = 1 Mb)*

Exercise 3

Best disease is associated with progressive blindness. It is an autosomal disease and is found on chromosome 11. Nine SSR markers were used in an attempt to find the exact position of the gene based on the study of a genealogical tree. Which of the nine markers is closest to the mutation that causes the disease?



1		2	
1^0	1^1	1^0	1^0
2^0	2^1	2^2	2^1
3^1	3^0	3^0	3^2
4^0	4^1	4^0	4^2
5^2	5^0	5^0	5^0
6^1	6^2	6^3	6^4
7^0	7^1	7^2	7^2
8^2	8^1	8^0	8^0
9^0	9^1	9^2	9^3

1		2		3		4		5		6	
1^0	1^0	1^0	1^0	1^0	1^0	1^0	1^0	1^1	1^0	1^1	1^0
2^0	2^2	2^0	2^1	2^0	2^1	2^0	2^2	2^1	2^2	2^1	2^2
3^1	3^0	3^1	3^2	3^1	3^2	3^0	3^0	3^0	3^0	3^0	3^0
4^0	4^0	4^1	4^2	4^0	4^2	4^0	4^0	4^0	4^0	4^1	4^2
5^0	5^0	5^0	5^0	5^0	5^0	5^0	5^0	5^0	5^0	5^2	5^0
6^2	6^4	6^1	6^3	6^1	6^3	6^2	6^3	6^1	6^3	6^2	6^3
7^1	7^2	7^1	7^2	7^1	7^2	7^0	7^2	7^0	7^2	7^0	7^2
8^1	8^0	8^1	8^0	8^1	8^0	8^2	8^0	8^2	8^0	8^2	8^0
9^1	9^3	9^1	9^2	9^1	9^2	9^0	9^3	9^0	9^2	9^0	9^3

Exercise 3 - Solution

The 4^0 allele characteristically occurs only in people with the disease.

Large-scale physical maps

A Physical Map consists of the combination of overlapped DNA fragments from library inserts placed with specific direction on each of the chromosomes. In other words clones that have emerged from chromosomes must be put in a specific order so as to form the physical map. The process requires the use of computers.

Using physical maps, the exact number of base pairs (bp, Kb, Mb) between a gene, a DNA locus or a DNA site and their neighboring elements (within a particular chromosome) are calculated

For each organism the correlation of physical and linkage maps can be calculated. In humans the correlation is 1 cM ~ 1 Mb and in mouse 1 cM ~ 2 Mb.

Meyers et al 2004 Nat. Rev. Genet 5. 578-588

Large-scale physical maps

Cloning vectors for libraries

The inserts must be as long as possible and stable. They must not break or recombine.

So far cloning vectors that have been used are:

- Yeast artificial chromosomes YACs: They receive large inserts (up to 1 Mb), but they are unstable and difficult to handle.
- Bacterial artificial chromosomes BACs. They receive inserts of 50-300 Kb and are more stable therefore were used more.

How many BACs (of average length 200 Kb) are needed to include the entire human genome;

3 X 10^9 bp the genome/ 200 Kb BAC = 15.000 BACs.

Large-scale physical maps

Vector	Size of partial digestion fragments of genomic DNA	Copy number	Number of clones required for 2x coverage of human genome
Cosmid	35-45 kb	50-100	~75000
Phasmid	35-45 kb	1 copy/cell	~75000
BAC	100-200 kb	1 copy/cell	15000-30000
PAC	100-200 kb	1 copy/cell	15000-30000
YAC	200-1000 kb	1 copy/cell	3000-15000

Cloning vectors of large fragments: These systems are essential for the study of the genome

Large-scale physical maps

Map	Number of BACs assembled	BACs coverage
Human	283,287	15x
Mouse	305,716	33x
D. melanogaster	10,253	14x
A. thaliana I	20,206	17x
A. thaliana II	9,389	7.2x
Rice I	21,087	6.9x
Rice II	65,287	20x
Soybean	78,001	9.6x
Rat	189,689	13.1x
Sorghum	22,233	4x
B. japonicum	4,608	77x

Cloning Vectors for libraries
Number of BACs that were needed to create physical maps in various species

Meyers et al 2004
Nat. Rev. Genet 5. 578-588

Large-scale physical maps

AIM→ finding the relative position of the inserts of the clones on the chromosome.

It can be achieved by a:

Technique from top to bottom → hybridization (FISH) with whole inserts onto the karyotype of the organism

Technique from bottom to top →the relative position of clones identified using 1) linkage maps or 2) restriction enzymes or 3) Sequence Tagged Sites–STS.

Large-scale physical maps

FISH (Fluorescence in situ hybridisation)

Based on the use of probes that are labeled with fluorescent dyes and used for hybridization on metaphase chromosomes

Easy mapping but of low resolution

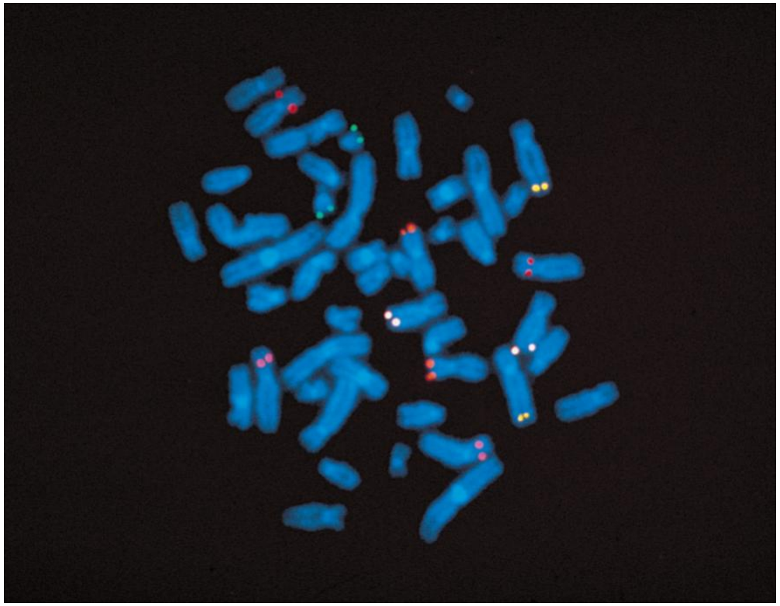


Figure 5: FISH Technique

https://www.youtube.com/watch?v=nm8Ai1CI9Is

Large-scale physical maps

To form the large-scale physical maps various approaches are used:

1) Linkage maps
2) Restriction enzymes
3) Sequence Tagged sites (STS)

☞ Researchers generally use a combination of these methods

Large-scale physical maps

1. Linkage maps

The existing information regarding the relative position of the genetic markers on the chromosomes is used.

Evenly distributed genetic markers spaced no more than 1 cM are required .

The genetic markers are used as hybridization probes.

Large-scale physical maps

Walking on chromosomes

- If 2 nearby markers M1 and M2, hybridize on YAC clones (with size of ~1 Mb, e.g. yM1 and yM2), probably these clones overlap
- The ends of one of the two clones are used, to find the relative position / overlapping of clones
- After using several consecutively nearby markers M1, M2, M3, M4, M5, M6, M7 which hybridize on corresponding clones… these clones are assembled in a continuous series
- If the genetic markers are not evenly distributed or a clone is too long, the clones do not form a continuous series, but there is a **gap**
- The overlapping clones which are assembled into a continuous genome fragment consists of a **contig**
- The compound of contigs creates the **scaffolds**

http://oregonstate.edu/dept/biochem/hhmi/hhmiclasses/bb451/figslett/FigBC.html

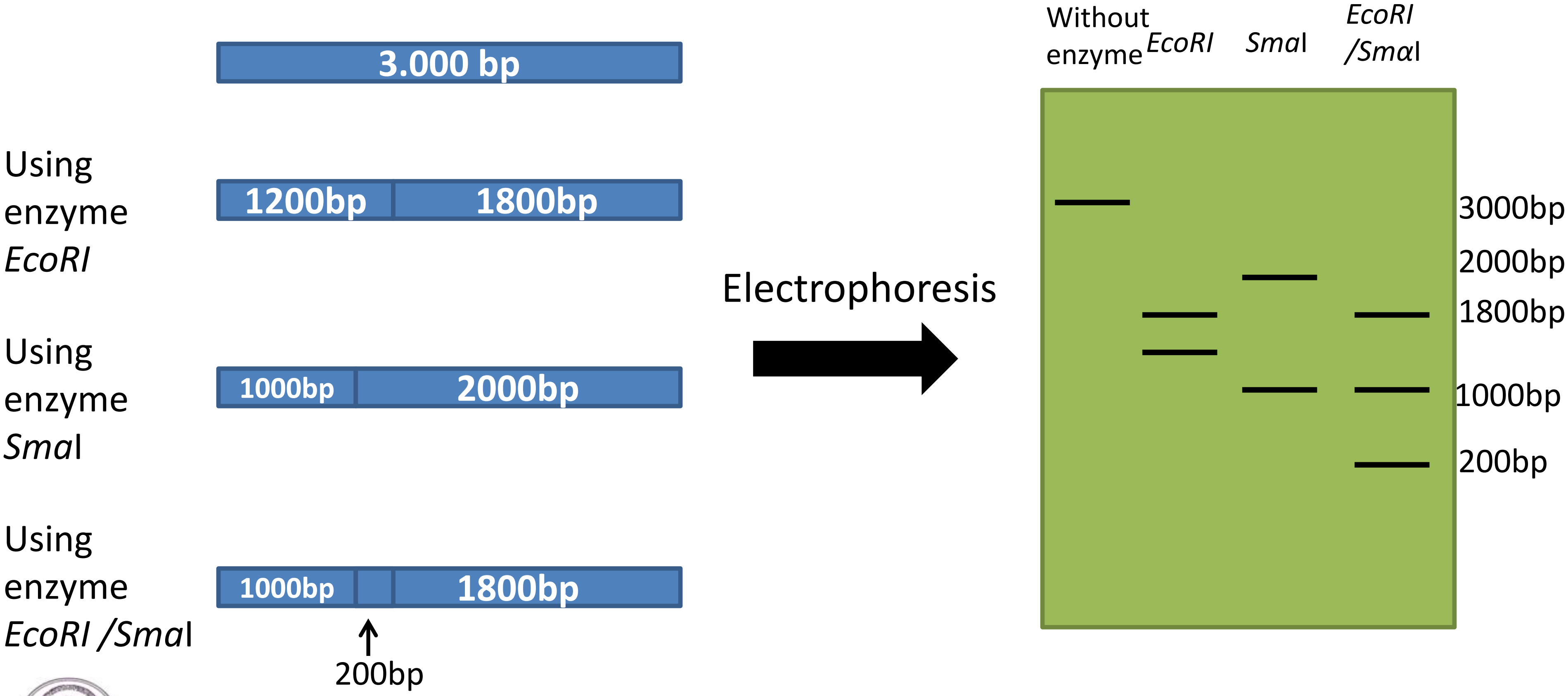
Comparison of the FISH method with linkage maps for constructing physical maps

- FISH allows mapping of any clone regardless of whether there is polymorphism, while linkage analysis requires polymorphism
- FISH needs only a small amount of DNA of the clone, while linkage analysis requires genotype information from many individuals
- FISH gives direct and comparable information about the relative position of clones on chromosomes even when they are modified through shifts
- FISH is applied on single clones, while linkage analysis always compares one clone with another
- **<u>Main Disadvantage</u>** → can not give the exact location of the clones

- It is usually used to find the location of a clone in a region of a chromosome 4-8 Mb long
- After that, the linkage analysis technique allows us to reach a deeper level of few cM

Large-scale physical maps

Mapping using restriction enzymes is based on the ability of restriction enzymes to cut DNA fragments at specific sites



3.000 bp

Using enzyme *EcoRI*

1200bp 1800bp

Using enzyme *Sma*I

1000bp 2000bp

Using enzyme *EcoRI /Sma*I

1000bp 1800bp

200bp

Electrophoresis

Without enzyme

EcoRI

*Sma*I

*EcoRI /Sma*I

3000bp

2000bp

1800bp

1000bp

200bp

Large-scale physical maps

2. Restriction enzymes

Chromosomes are packed in cosmid libraries .

The inserts in the cosmids are digested with specific restriction enzymes.

A characteristic **DNA fingerprint profile is created** consisting of all specific size fragments, resulting from the digestion.

If two cosmids overlap in a part of them, those restriction fragments that are produced from that overlapping region, will be common to the respective fingerprints.

Data from the digests are tens of thousands of fragments, and therefore the use of computers is required, for full analysis.

http://www.nature.com/nrg/journal/v5/n8/fig_tab/nrg1404_F1.html

Large-scale physical maps

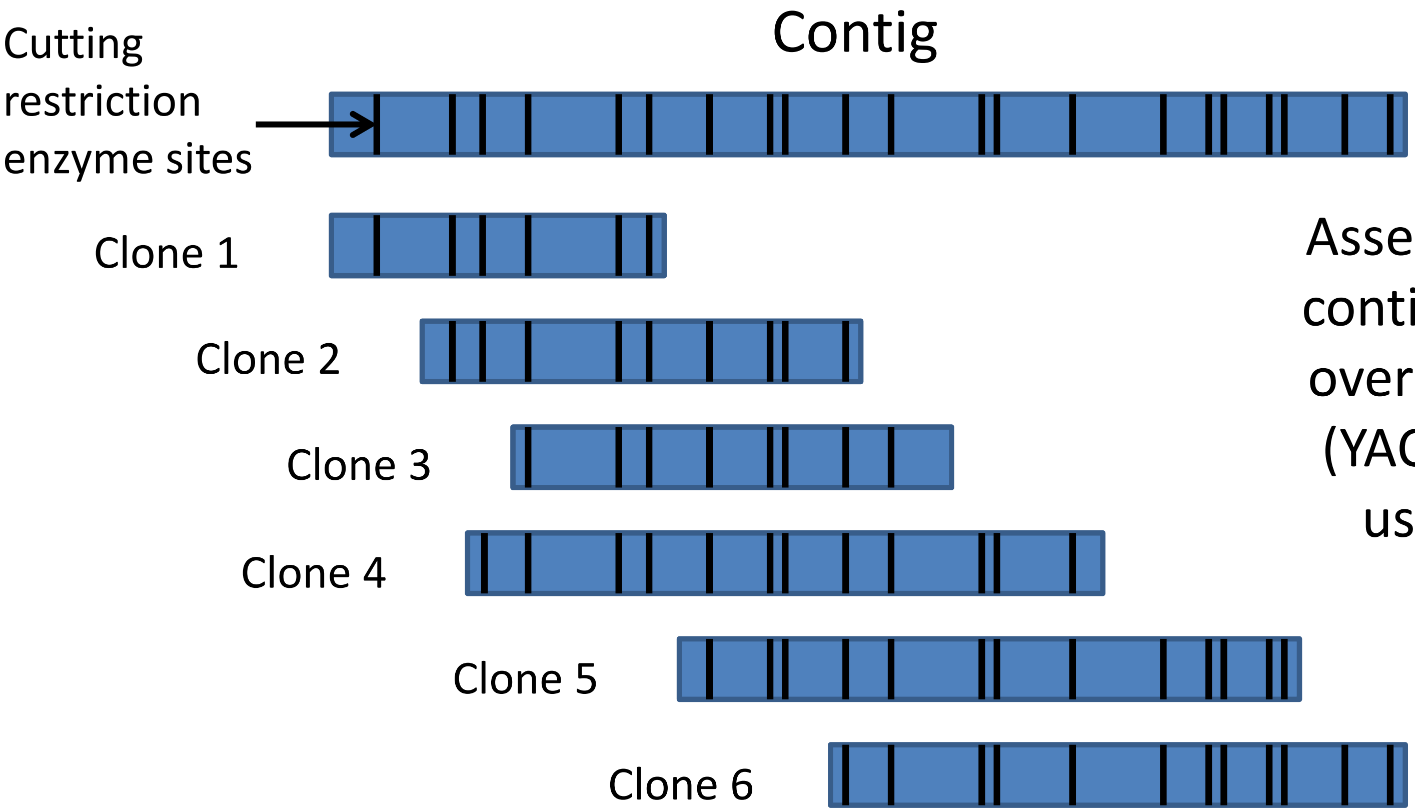
2. Restriction enzymes

Finferprint profile analysis is used for mapping and assembly of bacterial artificial chromosomes clones (BAC).

Mapping DNA fingerprints
http://www.yourgenome.org/downloads/animations.shtml

Large-scale physical maps

2. Restriction enzymes

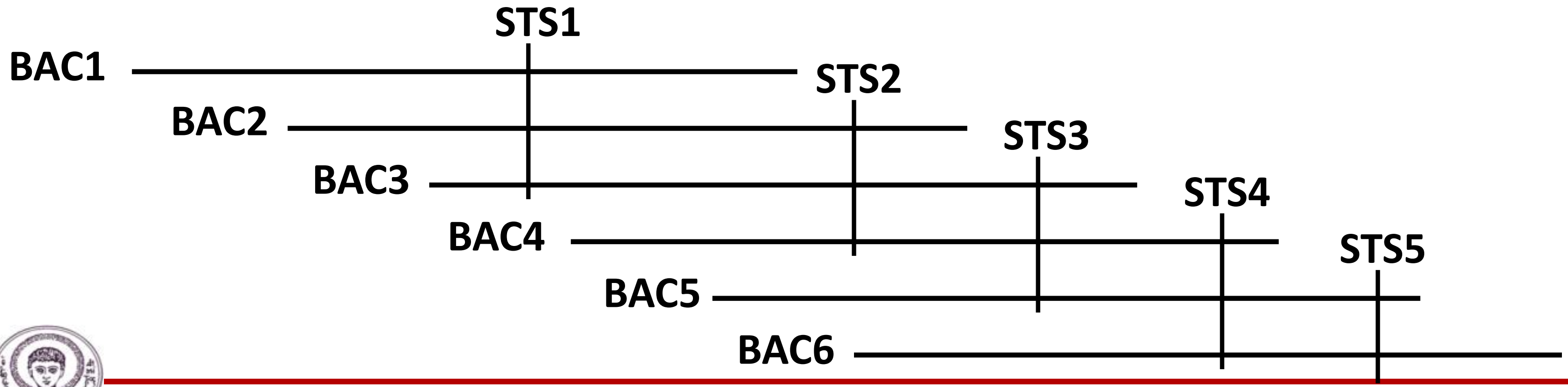


Assembly map of a contig consisting of overlapping clones (YACs or BACs are used for clone storage)

Large-scale physical maps

3. Sequence Tagged Sites (STS)

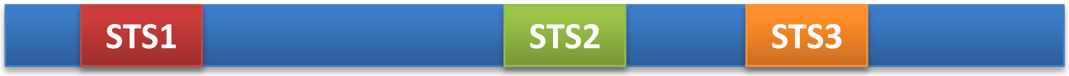
- A sequence that is unique and characteristic for a particular region/site of the chromosome is selected (sequence tagged site, STS). (i.e. SSRs / SNPs).
- Data from family trees are not needed
- Primers are designed to amplify STS by PCR
- All STS sites have a specific "address" on the chromosome.
- All BAC clones corresponding to the genome (~ 15000-20000) are checked whether they amplify this tagged site. If yes, then the clones overlap.
- As additional tagged sites are added, clones will be slowly assembled into a contig.



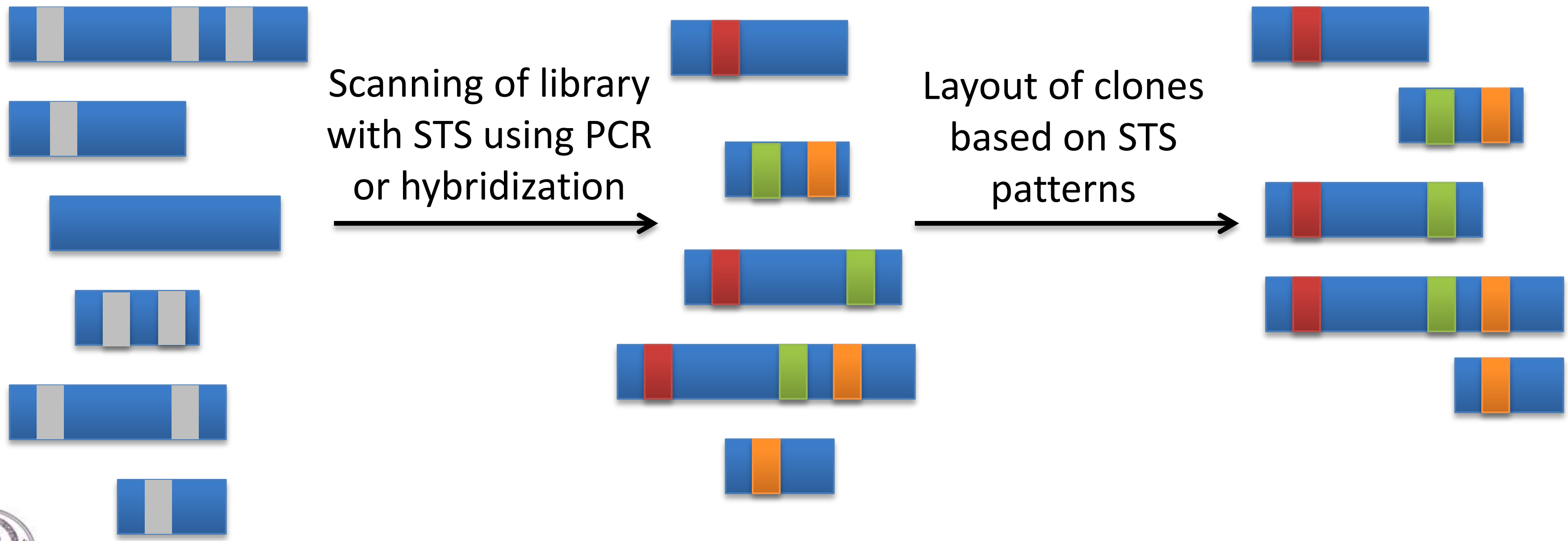
Large-scale physical maps

3. Sequence Tagged Sites (STS)

Mapping construction based on detecting STS on BAC clones



Library of BAC clones or other clones with large length

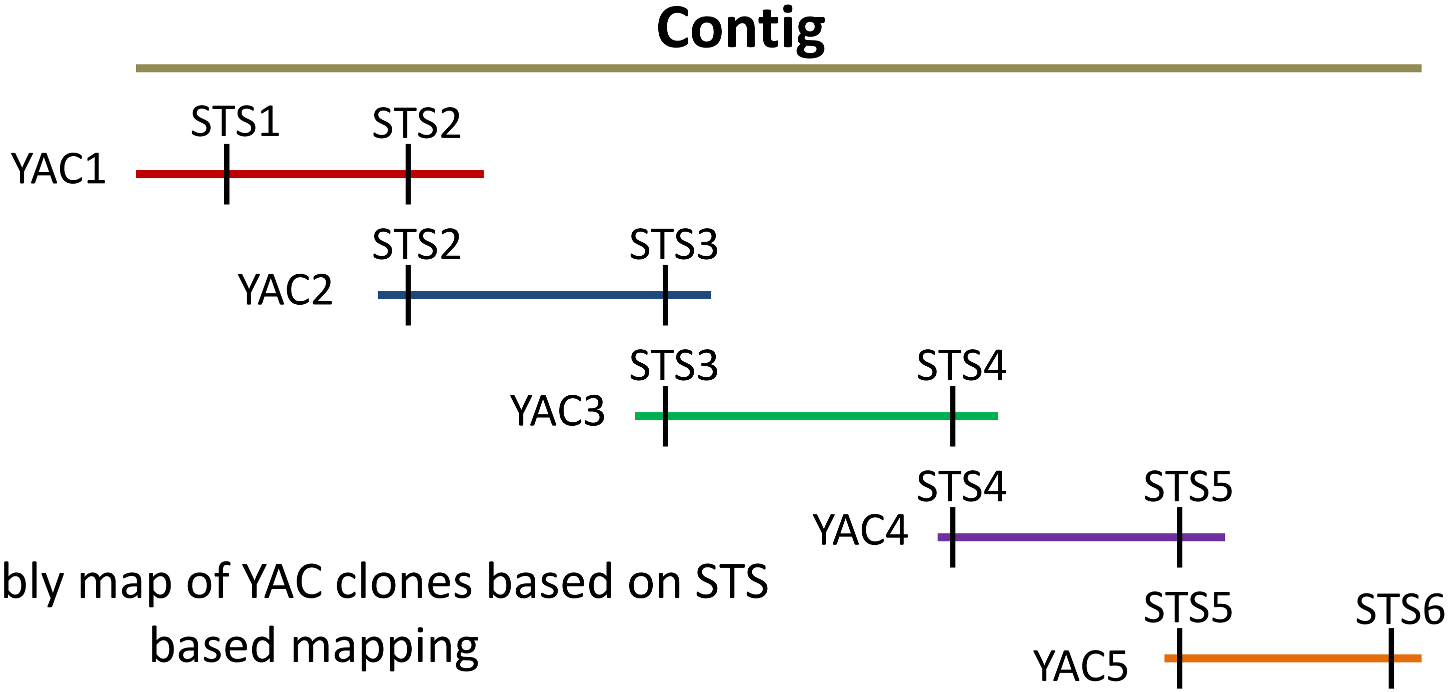


Scanning of library with STS using PCR or hybridization

Layout of clones based on STS patterns

Large-scale physical maps

3. Sequence Tagged sites (STS)

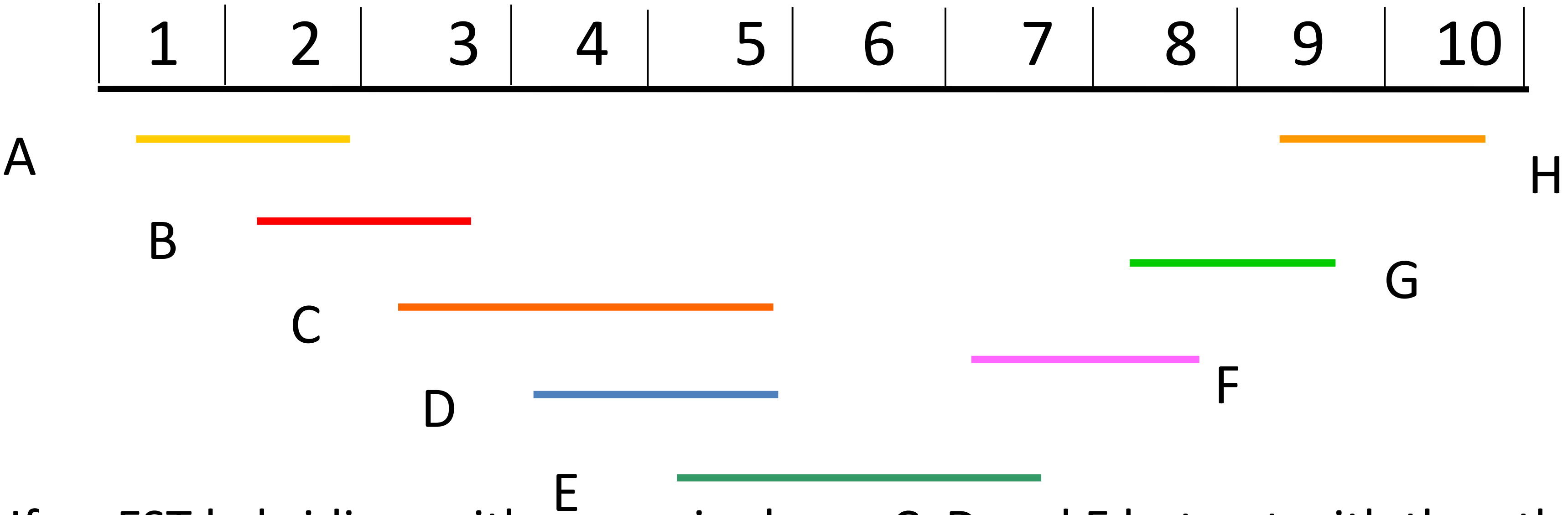


Assembly map of YAC clones based on STS based mapping

http://genome.cshlp.org/content/7/7/673/F4.expansion.html

Exercise 4

Here is a contig map of chromosome 3 of *Arabidopsis*



If an EST hybridizes with genomic clones C, D and E but not with the other clones, in which part of the chromosome 3 does this EST lie? If a clone of a gene hybridizes only with the genomic clones C and D in which part of chromosome 3 does it lie? If an STS hybridizes only with one genomic clone in which part of chromosome 3 does it lie?

Exercise 4 - Solution

If an EST hybridizes with genomic clones C, D and E but not with the other clones, in which part of chromosome 3 does this EST lie? The *EST is a DNA segement of a few hundred, which we try to locate on clones kilobases long. Consequently it is located in Section 5.*

If clone of a gene hybridizes only with the genomic clones C and D in which part of chromosome 3 does it lie? *In section 4.*

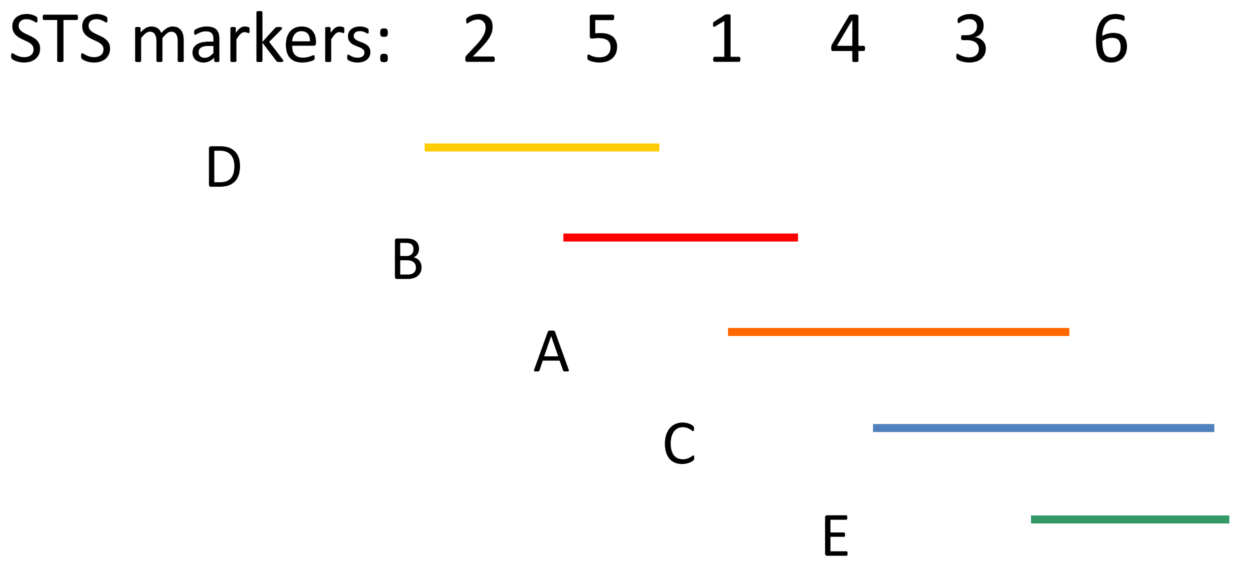
If an STS hybridizes only with one genomic clone in which part of chromosome 3 does it lie? *In 1,6 or 10.*

Exercise 5

Five genomic DNA clones (A-E) from YAC vectors were checked with hybridization for the presence of 6 STS. The results are shown in the following table. What is the order of the STS sites on the chromosome? Draw the contig map resulting from the combination of clones.

	STS1	STS2	STS3	STS4	STS5	STS6
A	+	-	+	+	-	-
B	+	-	-	-	+	-
C	-	-	+	+	-	+
D	-	+	-	-	+	-
E	-	-	+	-	-	+

Exercise 5 - Solution



STS markers: 2 5 1 4 3 6

D

B

A

C

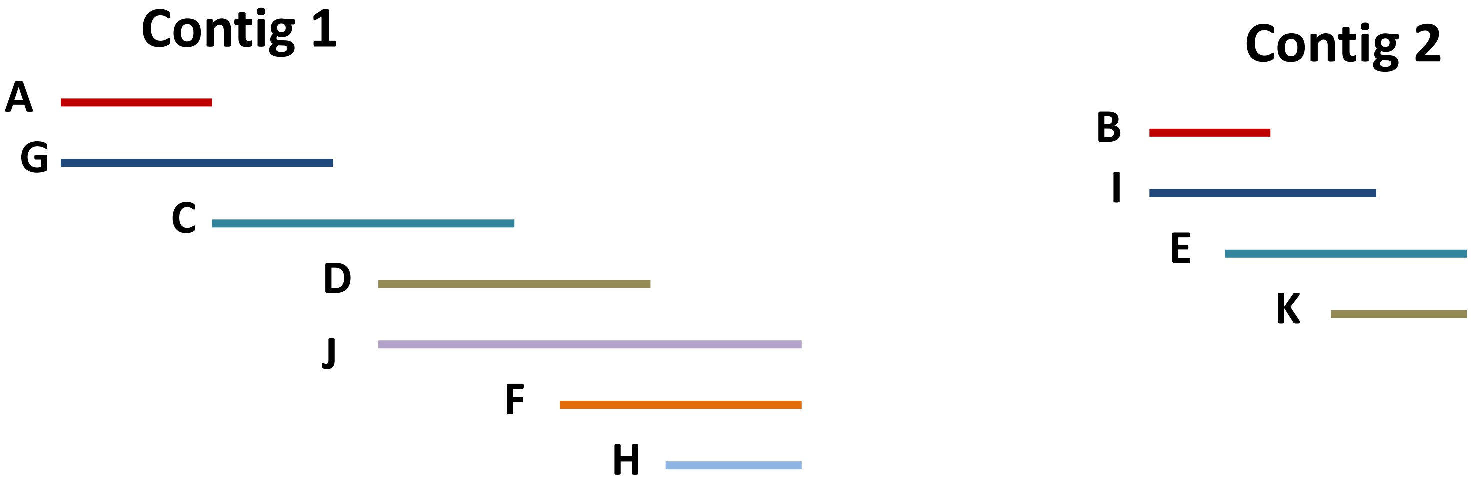
E

Exercise 6

11 genomic clones (A-K) from the chromosome 4 of *Drosophila melanogaster* were used to form a map of contigs with successive hybridizations. The results are shown in the following table. The (+) shows hybridization between the clones. How many contigs were assembled? What is the order of the clones in those contigs?

	A	B	C	D	E	F	G	H	I	J	K
K	-	-	-	-	+	-	-	-	-	-	+
J	-	-	+	+	-	+	-	+	-	+	
I	-	+	-	-	+	-		-	+		
H	-	-	-	-	-	+	-	+			
G	+	-	+	-	-	-	+				
F	-	-	-	+	-	+					
E	-	-	-	-	+						
D	-	-	+	+							
C	-	-	+								
B	-	+									
A	+										

Exercise 6 - Solution



Contig 1

A

G

C

D

J

F

H

Contig 2

B

I

E

K

Sequence maps (1/10)

The **Sequence Maps** show the sequence of nucleotides in a cloned DNA fragment

Two strategies are followed :

1. Hierarchical shotgun sequencing strategy
2. Whole-genome shotgun sequencing strategy-WGS

Shotgun : the cloned overlapping DNA fragments arise from random breakage of BAC inserts or the entire genome by ultrasound or partial digests with restriction enzymes (just like a shotgun hits the target in many places).

Sequence maps (2/10)

Hierarchical shotgun sequencing strategy

https://www.youtube.com/watch?v=-gVh3z6MwdU

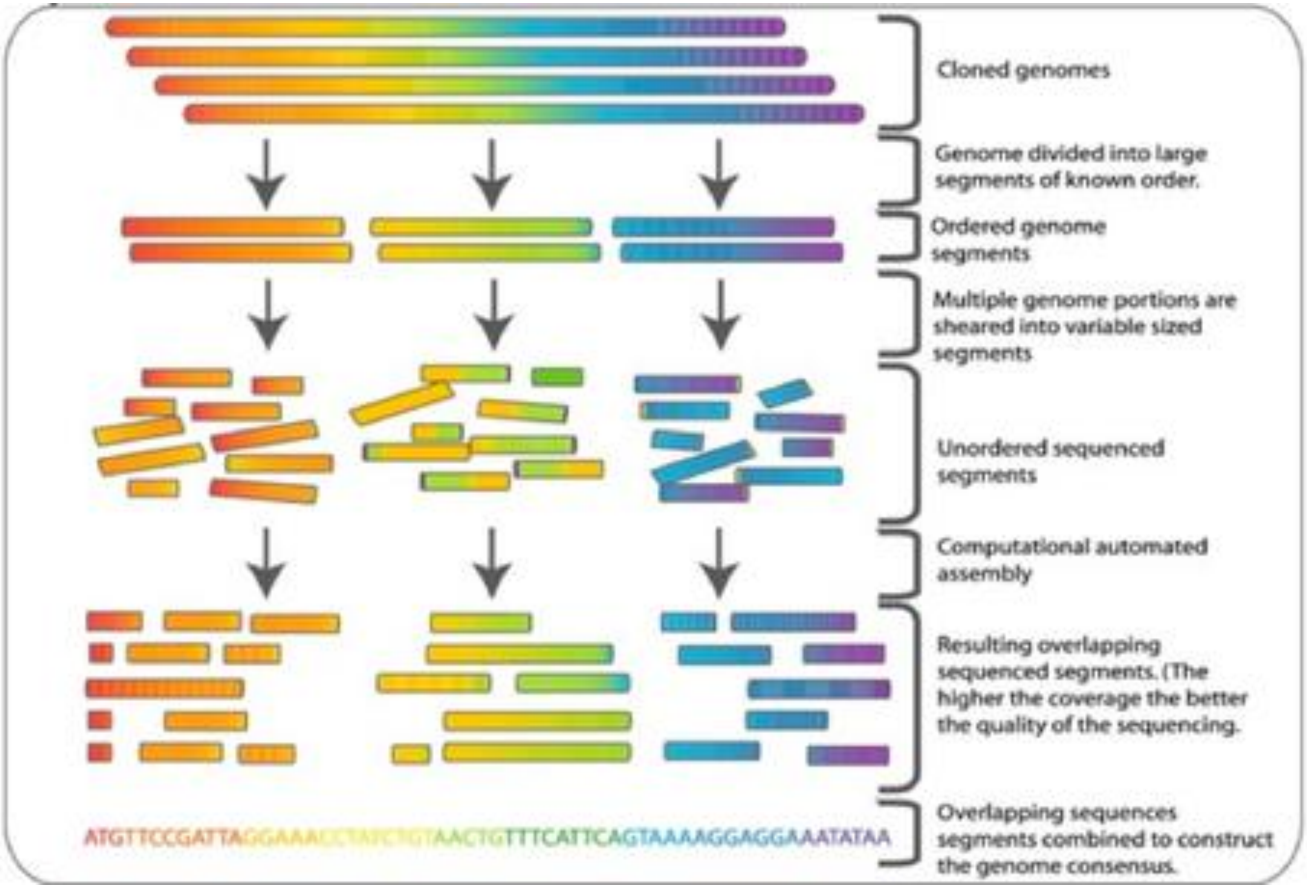
Used by the public sector within the HGP:

1. A genomic library is created in BACs
2. A map of overlapping BACs is created
3. As few aspossible BACs are selected in order to have a complete coverage of the genome
4. The BACs are broken into pieces with size of ~ 2 Kb which are cloned into plasmids, http://www.genome.gov/Edkit/flash/section3.html
5. Sequencing of plasmids follows until a 10 fold sequencing coverage of the BAC

http://www.yourgenome.org/landing_hgp.shtml

Sequence maps (3/10)

Figure 6: Hierarchical shotgun sequencing strategy


Cloned genomes
Genome divided into large segments of known order.
Ordered genome segments
Multiple genome portions are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequenced segments. (The higher the coverage the better the quality of the sequencing.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
Overlapping sequences segments combined to construct the genome consensus.

Sequence maps (4/10)

Hierarchical shotgun sequencing strategy

- Advantage: a small number of plasmids per BAC needs to be sequenced
- Disadvantage: It requires quite a lengthy and expensive preparatory work to create physical maps of BACs and plasmid libraries for about 15.000-20.000 BACs.

Sequence maps (5/10)

Whole-genome shotgun sequencing strategy

https://www.youtube.com/watch?v=vg7Y5EeZsjk&list=UUfe1_Tt2cyejbU3g0Hdo5Iw

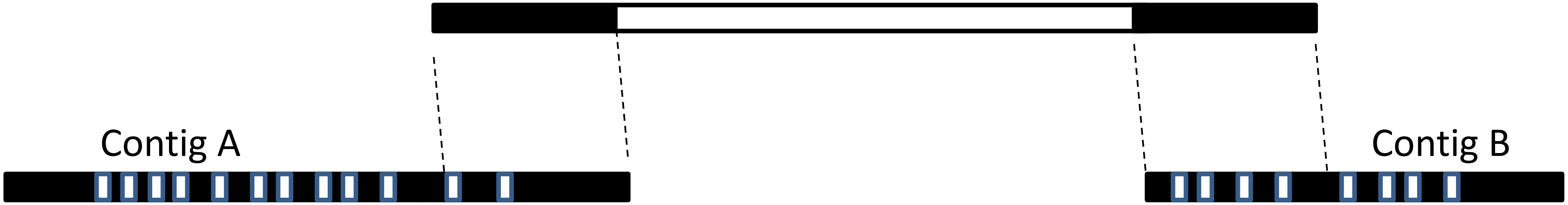
It was used by the private company Celera: The entire genome breaks three independent times so as to be cloned into

- A plasmid library with inserts ~ 2 Kb. It is used for 6 fold coverage of the genome
- A plasmid library with inserts ~ 10 Kb. It is used for 3 fold coverage of the genome
- A plasmid library with inserts ~ 200 Kb. It is used for single coverage of the genome

Inserts of different size are used in order to help the assembly of parts between which there are gaps. If two sequences are not assembled in one (they do not overlap), but are found in a larger insert then these sequences certainly are part of a larger contig (even though the in-between part is missing). We remind you that any library clone (no matter the size) can only be read from the beginning or the end (600 bases at most).

Sequence maps (6/10)

Whole-genome shotgun sequencing strategy



Explanation of genomic assembly read from libraries of different inserts: The bottom row shows the assembly of the inserts of the small libraries (2 and 10 kb) when two independent contigs have emerged (A and B). Although initially not joined in a contig, overlapping of each of the two ends of the insert clone with size of 200 kb (upper row) allows us to understand that they are part of a continuous chromosomal fragment, i.e. facilitates higher order assembly.

Sequence maps (7/10)

Whole-genome shotgun sequencing strategy

Advantages:

- No need to construct physical maps
- It requires the creation of a single BAC and two plasmid libraries
- Is based on a single automated and mature technique - sequencing

Disadvantages:

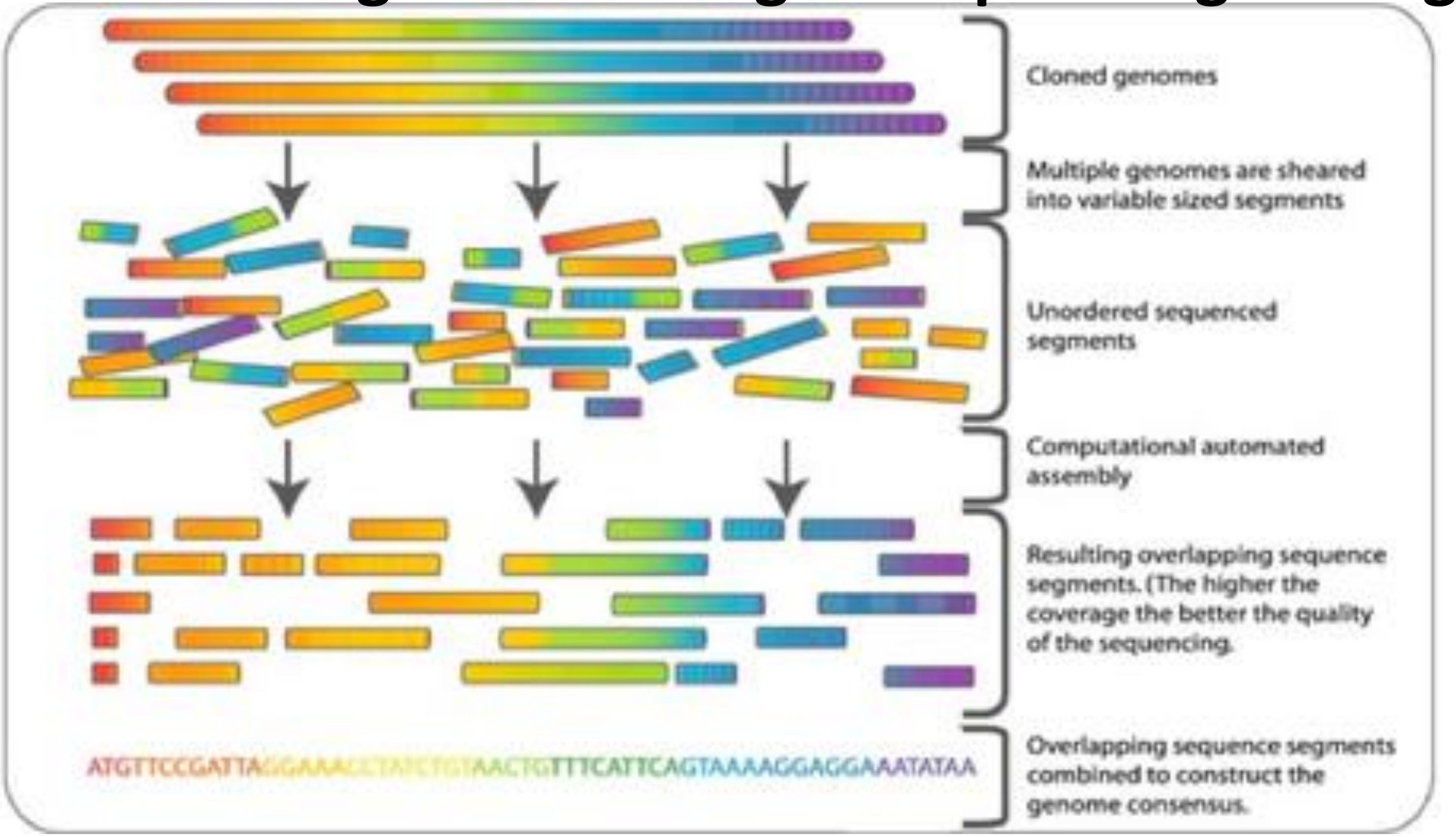
The problem created by the repetitive sequences in assembly.

Needs even more complicated bioinformatic analyses

The public sector, however, had accused Celera that it would not have succeeded so quickly to assemble / complete human genome, without integrating data from the creation of physical and genetic maps from the public effort.

Sequence maps (8/10)

Figure 7: Whole-genome shotgun sequencing strategy


Cloned genomes
Multiple genomes are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.
Overlapping sequence segments combined to construct the genome consensus.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA

See also: http://mmg-233-2014-genetics-genomics.wikia.com/wiki/Shotgun_Sequencing

Sequence maps (10/10)

Comparison of strategies

Disadvantages of both methods is that there are always heterochromatin compact areas that are impossible to be cloned and additionally some cloned areas can recombine and some of their parts are lost

With time, it has turned out that most research centers resort to the WGS approach, as it is faster and cheaper. Nowadays only when researchers are interested in a detailed study of the entire genome of a species both techniques will be used through a combination of them.

The problem of data storage

Data production rates 2004

… and rates are increasing

Most of the sequence data on the human genome was produced, basically, the last year before completion. At that time machines that were able to read 96 samples in 4 hours were created. Therefore, each machine could read 345.600 bp a day (600 bp / capillary X 96 capillary X 6 readings per day).

It was also possible to prepare and produce DNA fragments in factory rates for continuously supplying the automatic sequencing machines with DNA.

Finally, research units were created with 100-300 sequencing machines. Theoretically such units produce 103.680.000 bp per day (345.600 bp X 300 machines). Theoretically such a unit could provide a coverage of the human genome in 30 days (3 X 10^9 / 1,0368 X 10^8) .

Of course, there were always problems that slowed the pace, such as reduced production of DNA fragments, experimental errors, damage of machinery, low quality results.

And the rates are increasing

Until 2004 we relied on the same sequencing technique of Sanger!

From 2004 and beyond ... new machines were built with higher precision, lower costs and faster results

New Sequencing machines (1/6)

Roche-Pyrosequencing

- 2005
- Performs 1-2 million reactions
- Read length: 400-500 bases

http://454.com/products/technology.asp

https://www.youtube.com/watch?v=bFNjxKHP8Jc

New Sequencing machines (2/6)

Applied Biosystems SOLiD

- 2006
- Performs 100 millions reactions
- Read length: 35 bases

http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/solid-sequencing-chemistry.html
http://www.lifetechnologies.com/gr/en/home/life-science/sequencing/next-generation-sequencing/solid-next-generation-sequencing.html.html

https://www.youtube.com/watch?v=nlvyF8bFDwM

New Sequencing machines (3/6)

Illumina

- 2007
- Performs 80 million reactions
- Read length: 45 bases

http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf
http://www.illumina.com/systems/miseq.ilmn

https://www.youtube.com/watch?v=womKfikWlxM

New Sequencing machines (4/6)

1000 Genomes Project

- An integrated map of genetic variation from 1092 human genomes
The 1000 Genomes Project Consortium, 2012, Nature, 491(7422): 56–65

- The sequencing of 1000 human genomes from different populations will provide a more detailed picture of human genetic diversity. It was made possible thanks to three NGS (next generation sequencing) machines : 454, Illumina, SOLiD. The next slide shows the three stages of the program and the production of data

http://www.1000genomes.org/

New Sequencing machines (5/6)

Pilot 1

Low coverage

179 samples

@ 3.5 X

2.7 Tbp total

202 Gbp 454

1.8 Tbp Illumina

640 Gbp AB SOLiD

Pilot 2

Deep trios (CEU & YRI)

6 samples

@ 41X

1.1 Tbp total

87 Gbp 454

773 Gbp Illumina

270 Gbp AB SOLiD

Pilot 3

Exon capture

697 samples

@ 50X

2.2 Mbp of targets

8140 targets

New Sequencing machines (6/6)

In 2012 the genome of 2,500 people was already sequenced.
… production quantity 2 HG (6 X 10^{9}) per day
At the end of the 1000 genomes project within three years,
there were 3 X 10^{12}bp, i.e. 60 times more data than what had been deposited into public databases in the last 25 years.

Assessments…(old by now)…
2005 10^{1} HG
2015 10^{3} HG (3 Terabytes)
2025 10^{6} HG (3.000 Terabytes)
2035 10^{9} HG

Where will the data be stored?
How well will the data be analyzed?

Assembling and storing data (1/7)

- The 1000 Genomes project produced a total of 50 terrabytes (50.000.000.000.000) data.
- A strong computer (which has the correspondingly large hard drive) would need > 4.6 days **JUST** to download all data of the 1000 Genomes Project.

Assembling and storing data (2/7)

The Sequence Read Archive (SRA) stores raw sequence data from "next-generation" sequencing technologies including 454, IonTorrent, Illumina, SOLiD, Helicos and Complete Genomics. In addition to raw sequence data, SRA now stores alignment information in the form of read placements on a reference sequence.

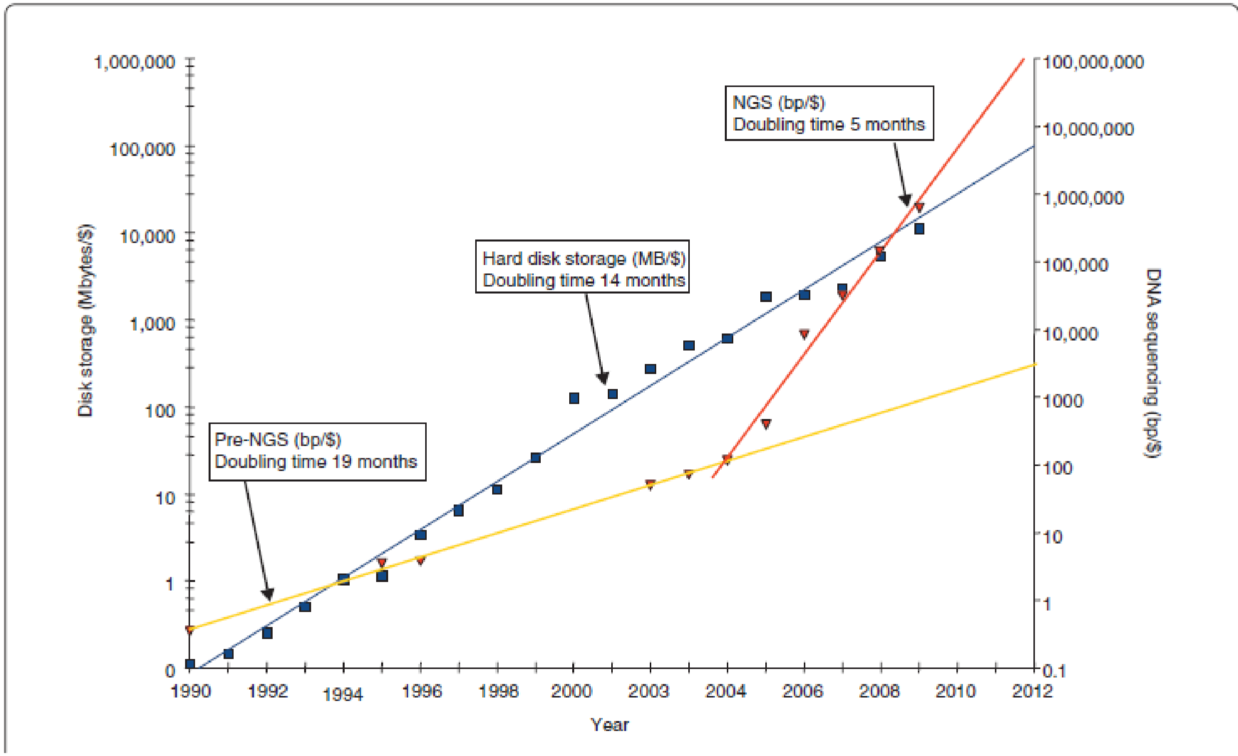
http://www.ncbi.nlm.nih.gov/Traces/home/

Assembling and storing data (3/7)

Sequencing vs data storage costs over time

Stein, **The case for cloud computing in genome informatics**
Genome Biology 2010, **11**:207

Figure 8: Since 2004 (when NGS machines appeared) and afterwards the production cost of DNA data decreases every six months (whereas it was 1.5 years before). After 2010, it actually costs less to produce DNA data than to save them.



NGS (bp/$)
Doubling time 5 months

Hard disk storage (MB/$)
Doubling time 14 months

Pre-NGS (bp/$)
Doubling time 19 months

Assembling and storing data (4/7)

How much data will be produced? 700 Mb, 600 Gb, 1 Tb, 1 Pb???

What will be the length of a single read? 50bp, 150 bp, 10Kbp, 100Kbp?

How much will it cost? 300M$ 100K$, 1000$, 1$???

How much time will be needed; 15 years, 1 week, 1 h, 1 min???

The Beijing Genome Institute produces 6 Tb data per day….

Assembling and storing data (5/7)

It is not surprising therefore that researchers are more concerned about where (in which computers / databases) will this growing data be deposited in the future and how it will be analyzed the, than how will this data be produced. And the answer is ... in the clouds !!! Large Internet companies such as Amazon, Google, Microsoft and others are already working along large genomic centers trying to find a solution. What they propose is ... **cloud computing**!

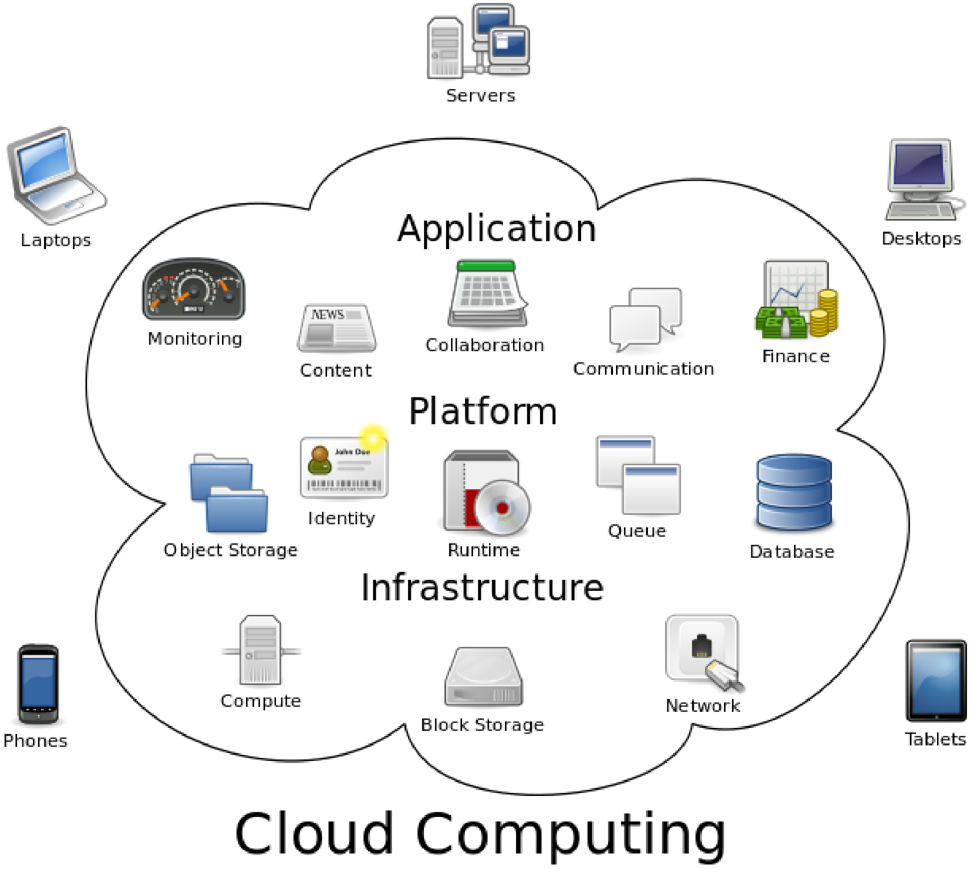


Figure 9: Cloud computing

Assembling and storing data (6/7)

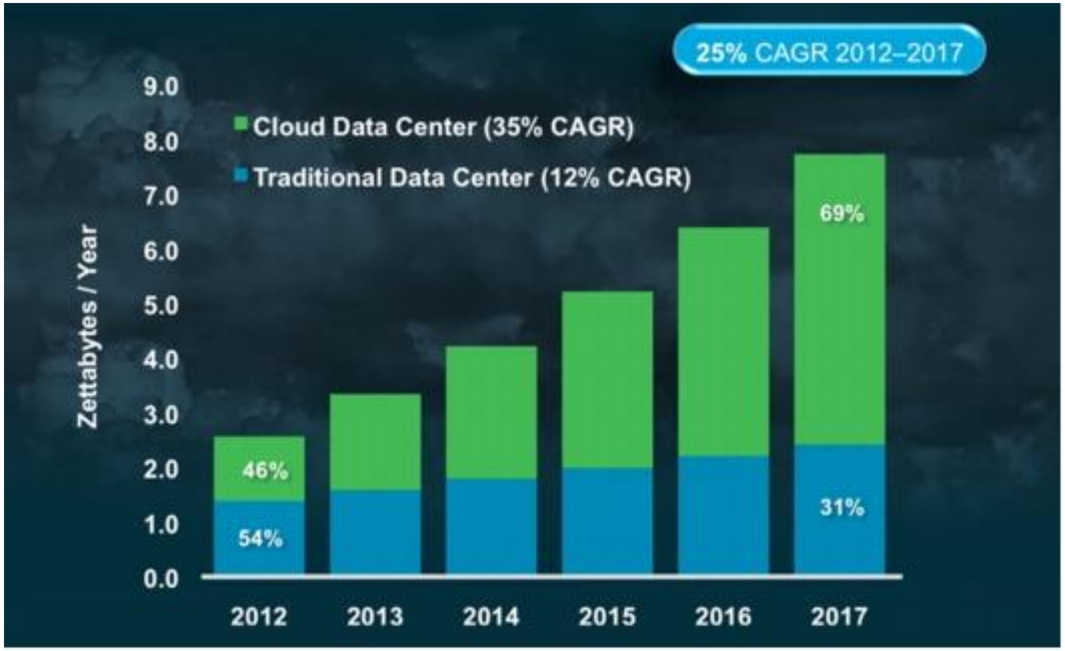
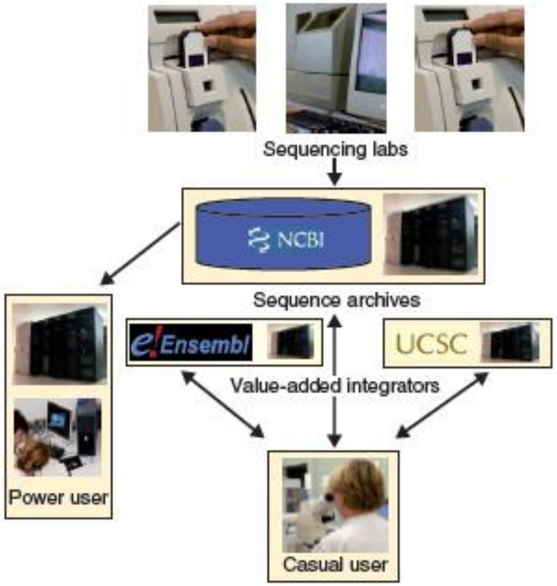


Figure 10: Until 2017> two thirds of all analyses on computers will be in the clouds

http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html

Assembling and storing data (7/7)



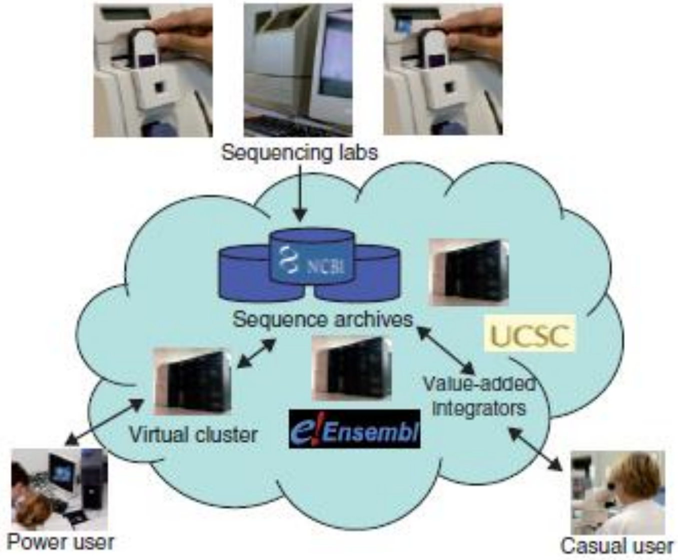


Figure 11: In the traditional data analysis, either a single researcher or a large laboratory had to download the data on personal computers.

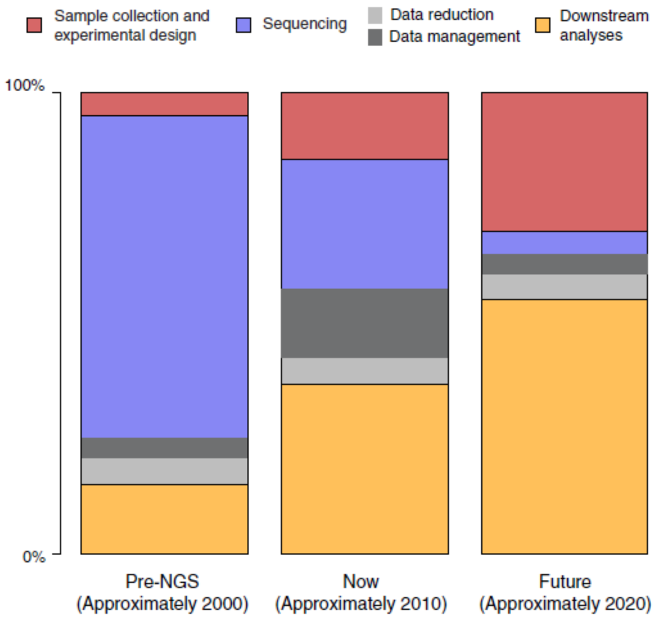
Figure 12: In the modern data analysis, all data will be stored online and analyses will be conducted in virtual machines.

Stein, Genome Biology 2010, **11**:207

Future of analyses (1/3)

Sboner et al. Genome Biology 2011, 12 :125

Nature 2013, 495, 293



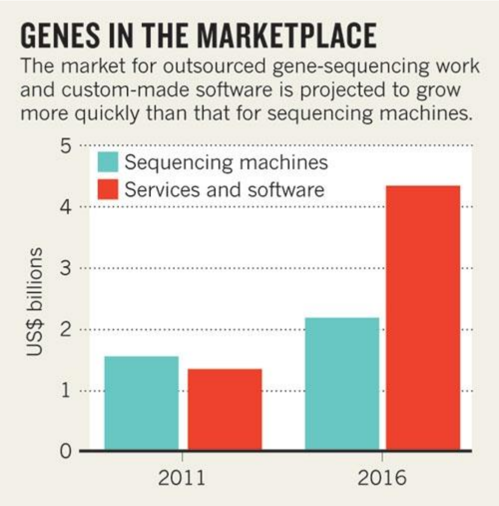


Figure 13: The analysis of data will gradually be more important and time consuming than their production.

Figure 14: Therefore genomic companies are involved in the provision of data analysis services and not just its production.

Future of analyses (2/3)

Primary Analysis	Secondary Analysis	Tertiary Analysis	Post Tertiary Analysis
Base Calling	Alignment	Variant Calling	In – Depth Annotation
Output	Output	Output	Output
Fastaq	SAM	VCF	**Biologically meaningful results**
XSq	BAM	Various tables	

We are still at the first three levels of analysis, we have not fully reached the fourth

Future of analyses (3/3)

The motto of the genomic company BGI, when launching the Easy Genomics service which offers processing of results from NGS sequencing, to the public (at a price), is the phrase of Albert Einstein : If you cant explain it simply, you don’t understand it well enough.

https://www.easygenomics.com/index

Illumina also offers a corresponding service, “BaseSpace”:

https://basespace.illumina.com/home/index

Biology: The big challenges of big data
http://www.nature.com/nature/journal/v498/n7453/full/498255a.html

Data protection

- Privacy protections: The genome hacker

Nature, 8/5/2013

http://www.nature.com/news/privacy-protections-the-genome-hacker-1.12940

- Personal Genome Project (PGP): is creating a freely available scientific resource that brings together genomic, environmental and human trait data. These data are donated by volunteers enrolled in a PGP study from our global network. Initiated by George Church at Harvard Medical School in 2005, the PGP has pioneered ethical, legal, and technical aspects related to the creation of public resources involving highly identifiable data like human genomes.

http://www.personalgenomes.org/organization/pgp

Finding the genes on genomic sequences (1/4)

It is necessary to use computers. They give confidence estimates for their forecasts :

- Searches based on "known" genes. Search homology. Comparison with cDNA sequences or ESTs (Expressed Sequence Tags) of the same species or with sequences of known genes of another species.
- “Blind" Searches of typical genes. Taking into account specific characteristics of genes, eg transcription initiation sites (ATG) and termination of translation (eg UAA), union positions exons / introns cutting (GT / AG) promoter regions (TATA boxes). Search for open reading frames-ORFs
- Comparison of whole genomes. Find conserved areas which will obviously correspond to functional areas.
- Existence of introns, junk DNA, high evolutionary changes hinder the analysis. Success rate is usually around 70-90%.

Find genes on the genomic sequences (2/4)

Reading frame

5' atgcccaagctgaatagcgtagaggggttttcatcatttgaggacgatgtataa 3‘

In most species an open reading frame starts with atg (met) and ends with (taa, tag, tga). The reading frames are three in the one chain and three in the complementary chain

Reading Frame 1

atg ccc aag ctg aat agc gta gag ggg ttt tca tca ttt gag gac gat gta taa

M P K L N S V E G F S S F E D D V *

Reading Frame 2

tgc cca agc tga ata gcg tag agg ggt ttt cat cat ttg agg acg atg tat

C P S * I A * R G F H H L R T M Y

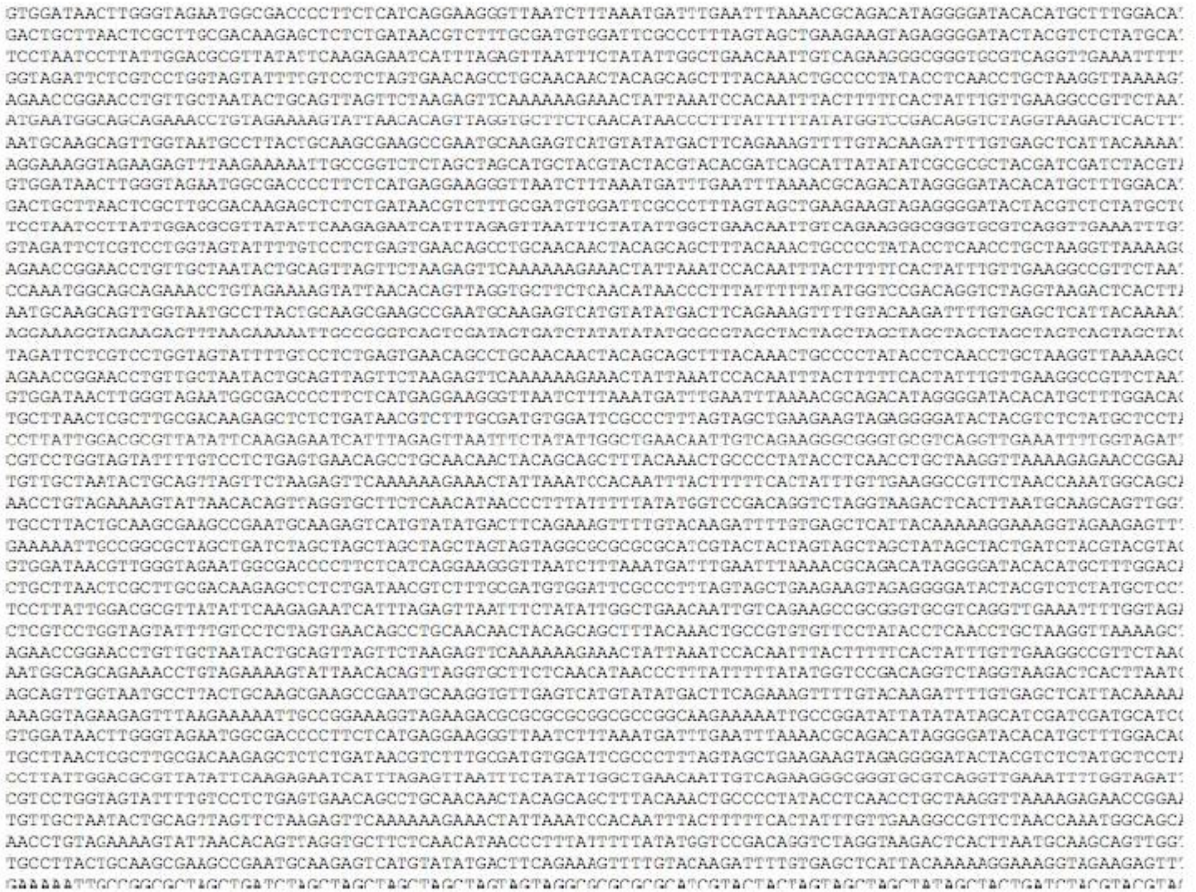
Reading Frame 3

gcc caa gct gaa tag cgt aga ggg gtt ttc atc att tga gga cga tgt ata

A Q A E * R R G V F I I * G R C I

Find genes on the genomic sequences (3/4)

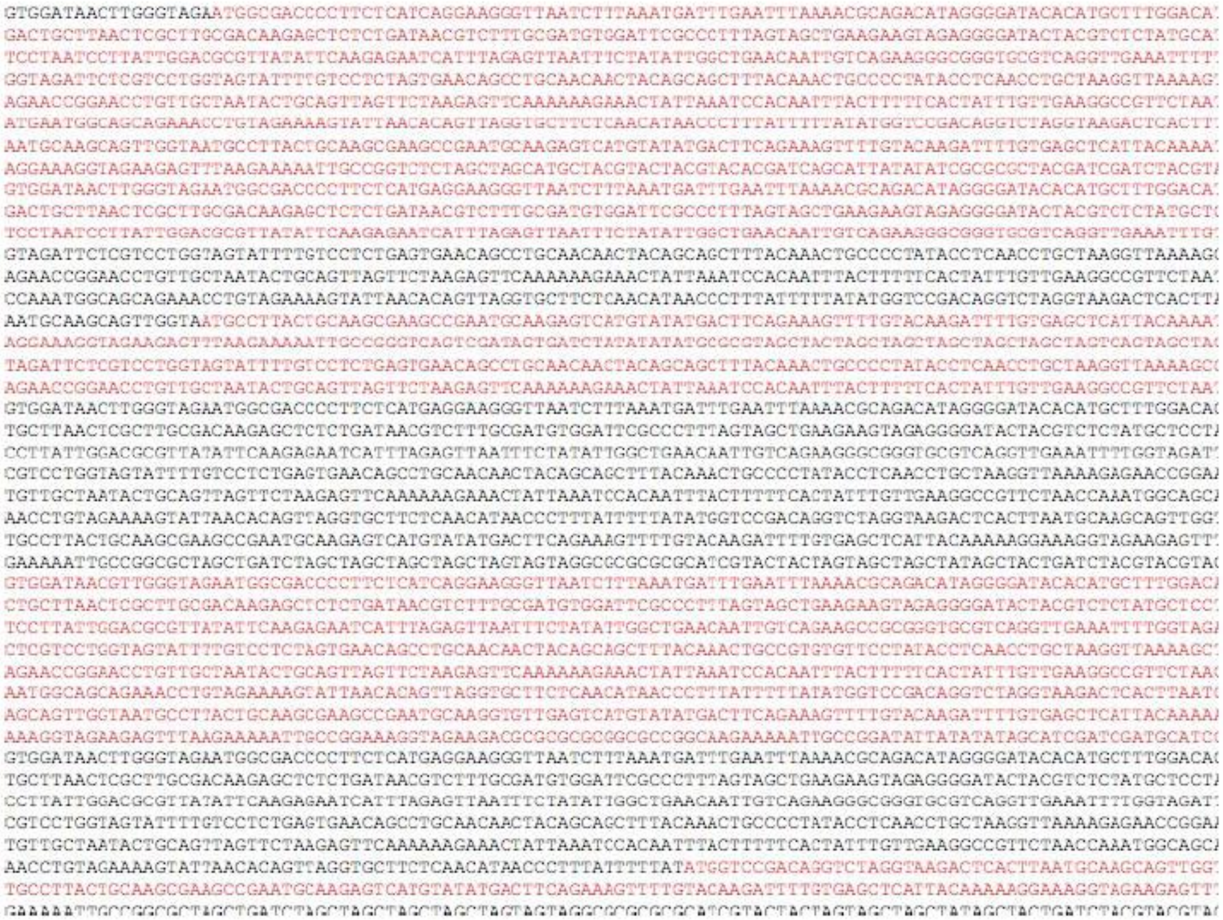
In **Figure 15** we have a genomic region in which we will find genes using computers.



```
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATCAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
GACTGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCA
TCCTAATCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTT
GGTAGATTCTCGTCCTGGTAGTATTTTGTCCTCTAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAG
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAA
ATGAATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTT
AATGCAAGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAA
AGGAAAGGTAGAAGAGTTTAAGAAAAATTGCCGGTCTCTAGCTAGCATGCTACGTACTACGTACACGATCAGCATTATATATCGCGCGCTACGATCGATCTACGT
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATGAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
GACTGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCT
TCCTAATCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTG
GTAGATTCTCGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAG
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAA
CCAAATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTT
AATGCAAGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAA
AGCAAAGCTAGAAGAGTTTAAGAAAAATTGCCCGGTCAGTCGATAGTGATCTATATATATGCCCGTAGCTACTAGCTAGCTAGCTAGCTAGCTACTCACTAGCTA
TAGATTCTCGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGC
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAA
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATGAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
TGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTCCT
CCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTTGGTAGAT
CGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGAGAACCGGA
TGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAACCAAATGGCAGC
AACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTAATGCAAGCAGTTGG
TGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAAAGGAAAGGTAGAAGAGTT
GAAAAATTGCCGGCGCTAGCTGATCTAGCTAGCTAGCTAGCTAGTAGTAGGCGCGCGCGCATCGTACTACTAGTAGCTAGCTATAGCTACTGATCTACGTACGTA
GTGGATAACGTTGGGTAGAATGGCGACCCCTTCTCATCAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGAC
CTGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTCC
TCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGCCGCGGGTGCGTCAGGTTGAAATTTTGGTAG
CTCGTCCTGGTAGTATTTTGTCCTCTAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCGTGTGTTCCTATACCTCAACCTGCTAAGGTTAAAAGC
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAA
AATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTAAT
AGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGGTGTTGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAA
AAAGGTAGAAGAGTTTAAGAAAAATTGCCGGAAAGGTAGAAGACGCGCGCGGCGCCGGCAAGAAAAATTGCCGGATATTATATATAGCATCGATCGATGCATC
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATGAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
TGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTCCT
CCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTTGGTAGAT
CGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGAGAACCGGA
TGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAACCAAATGGCAGC
AACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTAATGCAAGCAGTTGG
TGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAAAGGAAAGGTAGAAGAGTT
GAAAAATTGCCGGCGCTAGCTGATCTAGCTAGCTAGCTAGCTAGTAGTAGGCGCGCGCGCATCGTACTACTAGTAGCTAGCTATAGCTACTGATCTACGTACGTA
```

Find genes on the genomic sequences (4/4)

Using computers all open reading frames (ORFs) were found in this genomic region and displayed in red.



```
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATCAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
GACTGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCA
TCCTAATCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTT
GGTAGATTCTCGTCCTGGTAGTATTTTGTCCTCTAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAG
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAA
ATGAATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTT
AATGCAAGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAA
AGGAAAGGTAGAAGAGTTTAAGAAAAATTGCCGGTCTCTAGCTAGCATGCTACGTACTACGTACACGATCAGCATTATATATCGCGCGCTACGATCGATCTACGT
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATGAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
GACTGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCT
TCCTAATCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTG
GTAGATTCTCGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAG
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAA
CCAAATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTT
AATGCAAGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAA
AGGAAAGGTAGAAGACTTTAAGAAAAATTGCCCGGTCAGTCGATAGTGATCTATATATATGCGCGTACCTACTAGCTAGCTAGCTAGCTAGCTAGTCAGTAGCTA
TAGATTCTCGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGC
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAA
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATGAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
TGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTCCT
CCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTTGGTAGAT
CGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGAGAACCGGA
TGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAACCAAATGGCAGC
AACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTAATGCAAGCAGTTGG
TGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAAAGGAAAGGTAGAAGAGTT
GAAAAATTGCCGGCGCTAGCTGATCTAGCTAGCTAGCTAGCTAGTAGTAGGCGCGCGCGCATCGTACTACTAGTAGCTAGCTATAGCTACTGATCTACGTACGTA
GTGGATAACGTTGGGTAGAATGGCGACCCCTTCTCATCAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGAC
CTGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTCC
TCCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGCCGCGGGTGCGTCAGGTTGAAATTTTGGTAG
CTCGTCCTGGTAGTATTTTGTCCTCTAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCGTGTGTTCCTATACCTCAACCTGCTAAGGTTAAAAGC
AGAACCGGAACCTGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAA
AATGGCAGCAGAAACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTAAT
AGCAGTTGGTAATGCCTTACTGCAAGCGAAGCCGAATGCAAGGTGTTGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAA
AAAGGTAGAAGAGTTTAAGAAAAATTGCCGGAAAGGTAGAAGACGCGCGCGGCGCCGGCAAGAAAAAATTGCCGGATATTATATATAGCATCGATCGATGCATC
GTGGATAACTTGGGTAGAATGGCGACCCCTTCTCATGAGGAAGGGTTAATCTTTAAATGATTTGAATTTAAAACGCAGACATAGGGGATACACATGCTTTGGACA
TGCTTAACTCGCTTGCGACAAGAGCTCTCTGATAACGTCTTTGCGATGTGGATTCGCCCTTTAGTAGCTGAAGAAGTAGAGGGGATACTACGTCTCTATGCTCCT
CCTTATTGGACGCGTTATATTCAAGAGAATCATTTAGAGTTAATTTCTATATTGGCTGAACAATTGTCAGAAGGGCGGGTGCGTCAGGTTGAAATTTTGGTAGAT
CGTCCTGGTAGTATTTTGTCCTCTGAGTGAACAGCCTGCAACAACTACAGCAGCTTTACAAACTGCCCCTATACCTCAACCTGCTAAGGTTAAAAGAGAACCGGA
TGTTGCTAATACTGCAGTTAGTTCTAAGAGTTCAAAAAAGAAACTATTAAATCCACAATTTACTTTTTCACTATTTGTTGAAGGCCGTTCTAACCAAATGGCAGC
AACCTGTAGAAAAGTATTAACACAGTTAGGTGCTTCTCAACATAACCCTTTATTTTTATATGGTCCGACAGGTCTAGGTAAGACTCACTTAATGCAAGCAGTTGG
TGCCTTACTGCAAGCGAAGCCGAATGCAAGAGTCATGTATATGACTTCAGAAAGTTTTGTACAAGATTTTGTGAGCTCATTACAAAAAGGAAAGGTAGAAGAGTT
GAAAAATTGCCGGCGCTAGCTGATCTAGCTAGCTAGCTAGCTAGTAGTAGGCGCGCGCGCATCGTACTACTAGTAGCTAGCTATAGCTACTGATCTACGTACGTA
```

Note of use of third party works

Hierarchical shotgun sequencing strategy, http://en.wikipedia.org/wiki/File:Whole_genome_shotgun_sequencing_versusHierarchical_shotgun_sequencing.png, by Commins, J., Toft, C., Fares, M. A., CC-BY-SA-2.5 (http://creativecommons.org/licenses/by-sa/2.5/deed.en)
Whole-genome shotgun sequencing strategy, http://en.wikipedia.org/wiki/File:Whole_genome_shotgun_sequencing_versusHierarchical_shotgun_sequencing.png, by Commins, J., Toft, C., Fares, M. A., CC-BY-SA-2.5 (http://creativecommons.org/licenses/by-sa/2.5/deed.en)
Cloud computing, http://es.wikipedia.org/wiki/Archivo:Cloud_computing.svg, by Sam Johnston, CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/deed.en)

Reference note

Copyright Aristotle University of Thessaloniki, Triantafyllidis Alexandros. «Special Topics on Genetics. Structural Genomics». Edition: 1.0. Thessaloniki, 2015. Available from the web address: http://opencourses.auth.gr/eclass_courses.

Licensing note

[1] http://creativecommons.org/licenses/by-sa/4.0/

End of Section

Processing: Minoudi Styliani
Thessaloniki, Winter Semester 2014-2015

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ
2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Notes Preservation

Any reproduction or adaptation of the material should include:

- the Reference Note
- the Licence Note
- the Notes Preservation
- Note of use of third party works

accompanied with their hyperlinks.