# Special Topics on Genetics

## Section 3: Structural Genomics of Organisms

**Triantafyllidis A.**
School of Biology

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

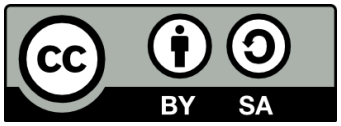
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# License

- The offered educational material is subject to Creative Commons licensing.
- For educational material, such as images, that is subject to other form of licensing, the license is explicitly referred to within the presentation.



# 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.

# Licensing of figures

We warmly thank the Pearson Education Inc for granting the right to use the following figures of this presentation:
Figures: 10, 17, 18

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

# Section Contents

- Genomes of prokaryotic organisms
- Metagenomic programs
- Genomes of eukaryotic organisms
- Overall Genomic Analysis of Biodiversity

# Model Organisms

The six first model organisms, in which genomic analyses were conducted:

**Figure 1:** *E. coli*

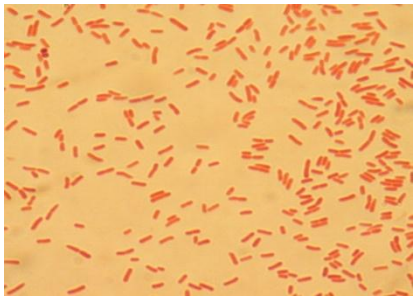


**Figure 2:** *S. cerevisiae*

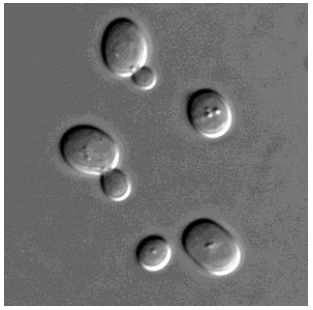


**Figure 3:** *C. elegans*

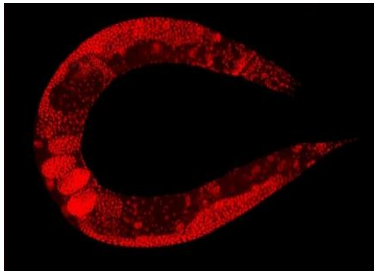


**Figure 4:** *D. melanogaster*

**Figure 5:** *A. thaliana*

**Figure 6:** *M. musculus*

# Genomes of prokaryotic organisms (1/5)

- **1995**: Sequencing of *Haemophilus influenzae* completed
- **2014**: ~3,200 complete genomes of bacteria (3,000 Eubacteria + 175 Archaea). An additional 25,000 programs are in progress (ongoing or finishing/in assembly) http://www.ncbi.nlm.nih.gov/genome/browse/

Recent years have witnessed a significant increase in sequenced genomes
http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_growth.html

# Genomes of prokaryotic organisms (2/5)

Most bacteria sequenced are related to human diseases: meningitis, cholera, tuberculosis, leprosy, pneumonia, ulcers and typhus

Project relevance of bacterial projects

| Sector | Percentage (2008) |
|---|---|
| Biomedical | 60% |
| Biotechnology | 14% |
| Environment | 6% |
| Agricultural | 5% |
| Phylogeny | 3% |
| Evolution | 2% |
| Other | 10% |

| Sector of bacterial genomic programs | Percentage (2013) |
|---|---|
| Medical | 46.9% |
| Human Pathogen | 6.4% |
| Human Microbiome Project | 6.4% |
| Environmental | 5.7% |
| Agricultural | 4.3% |
| Biotechnological | 3.4% |
| Tree of Life | 3.3% |
| Other | 23.7% |

# Genomes of prokaryotic organisms (3/5)

| | Genome size (Mb) | Number of genes |
|---|---|---|
| **Archaea** | | |
| *Methanosarcina acetivorans C2A* | 5.75 (Max) | 4540 |
| *Archaeglobus fulgidis* | 2.17 | 2493 |
| *Methanoccus jannaschii* | 1.66 | 1738 |
| *Thermoplasma acidophilum* | 1.56 | 1509 |
| **Eubacteria** | | |
| *Escherichia coli* | 4.64 | 4397 |
| *Bacillus subtilis* | 4.21 | 4212 |
| *Haemophilus influenzae* | 1.83 | 1791 |
| *Aquifex aeolicus* | 1.55 | 1552 |
| *Rickettsia prowazekii* | 1.11 | 834 |
| *Mycoplasma pneumoniae* | 0.82 | 710 |
| *Mycoplasma genitalium* | 0.58 | 503 |

# Genomes of prokaryotic organisms (4/5)

- The size of these bacterial genomes usually varies from 580,000 bp for *Mycoplasma genitalium*, up to ~ 4.5 Mb for *Mycobacterium tuberculosis* and *Escherichia coli*
- **Exceptions**: *Sorangium cellulosum* So0157-2 with very large size of 14.78 Mb

And the following bacteria with very small sizes

- *Nasuia deltocephalinicola* 112,000 bp
- *Candidatus Carsonella ruddii ruddii* 160,000 bp,
- *Buchnera aphidicola* 450,000 bp
- http://www.cbs.dtu.dk/services/GenomeAtlas-2.0/show-atlas.php?type=genomeatlas&KLSO=ASC&KLSK=ORGANISMSORT&kingdom=Bacteria&tableType=Protein%20Length&segmentid=Cruddii_PV_Main
- http://wishart.biology.ualberta.ca/BacMap/graphs_cgview.html

# Genomes of prokaryotic organisms (5/5)

## *The tree of Life…after 1990*

Woese et al. (1990) used phylogenetic analysis of ribosomal RNA to show that living organisms are grouped into three rather than two domains. The new domain was the Archaea.

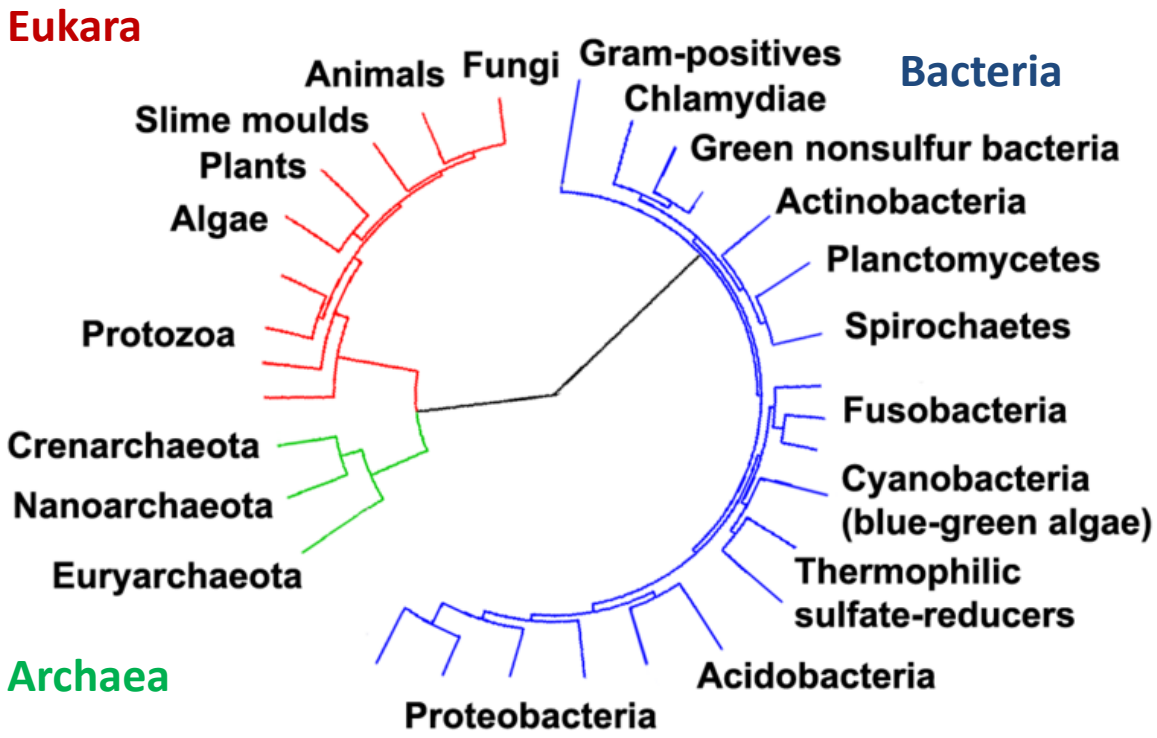



**Figure 7: The tree of life**
by Tim Vickers, http://commons.wikimedia.org/wiki/File:Collapsed_tree_labels_simplified.png

# *Archaea*

- Archaea are prokaryotes, such as eubacteria
- They are found in extreme environments (high temperature, salinity, pressure, pH, ocean depths). ... **though we now know that they are also plentiful in the oceans and in the human body!**
- They resemble bacteria structurally
- Replication, transcription, translation functions similar to eukaryotes
- Sequenced Genomes are still limited (~175)

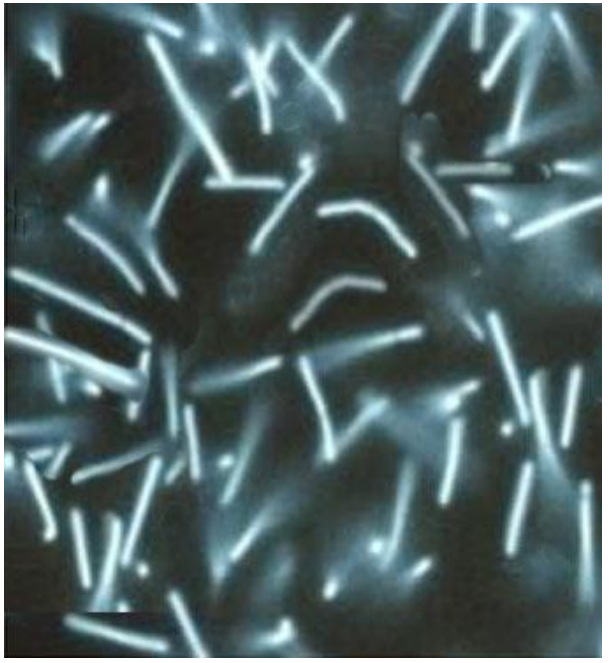


**Figure 9:** *Methanopyrus kandleri*

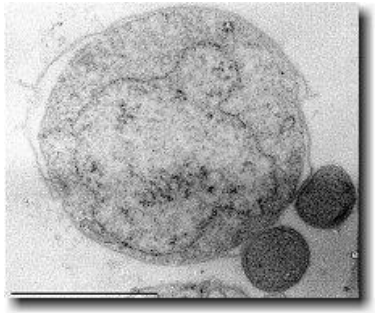


**Figure 8:** *Nanoarchaeum equitans*

490,885 bp

# *Haemophilus influenzae*

- 1995: The first free organism that had its genome sequenced (WGS)
- Contains 1743 genes, which are located very close and / or overlapping
- 1/3 of genes have an unknown function

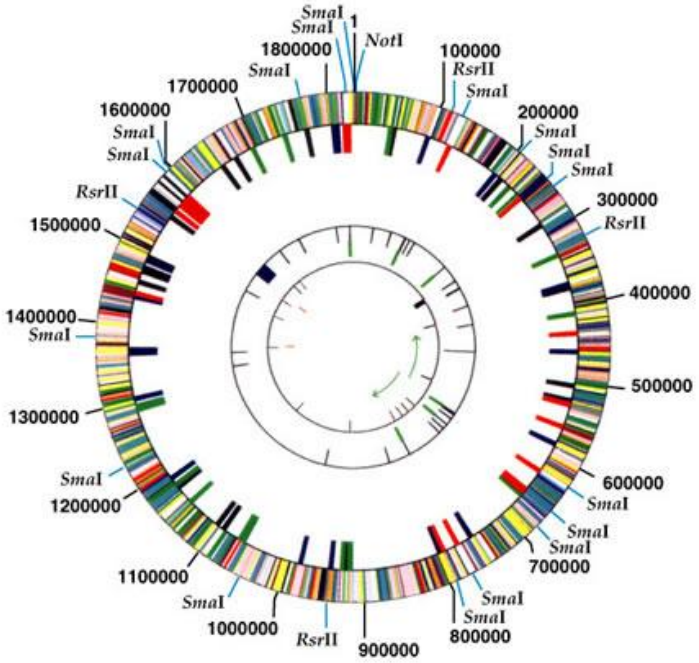





**Figure 10:** The genome of *H. influenzae*

# *M. genitalium*

- Its genome shows the minimum number of genes that an organism needs for autonomous replication
- It includes 503 genes either closely located and / or are overlapping
- Most of these are related to basic functions of a cell (replication, transcription, translation)
- 15% of the genome does not seem to code for proteins

http://en.wikipedia.org/wiki/File:Mycoplasma_genitalium.gif

# *Vibrio cholerae*

- The first prokaryote discovered to possess two chromosomes (3 & 1 Mb).
- The second chromosome seems to originate from a plasmid, trapped in an ancient species.
- Numerous plasmid genes have been found to be integrated into various bacterial genomes.

http://www.pnas.org/content/106/36/15442/F3.large.jpg

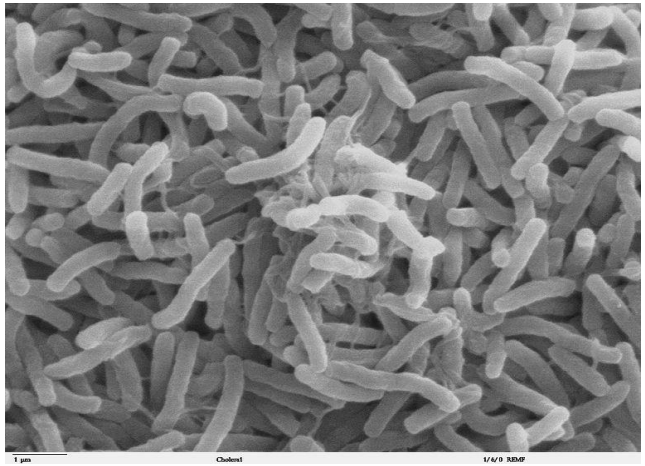


**Figure 11:** *Vibrio cholerae*

# *Escherichia coli* (1/2)

- The genome of type strain MG1655 was completed in 1997 (using Whole Genome Sequencing)
- Size ~ 4.5Mb.
- ~ 90% codes for genes
- ~ 20% of those genes originated through horizontal gene transfer
- Just 10% corresponds to regulatory areas, intergenic areas, repeated domains, gene residues (remnants of horizontal transfer) or unknown areas.
- 4397 potential genes have been identified, 1/3 of which produce known proteins.

http://www.pnas.org/content/103/34/12879/F1.large.jpg

# *Escherichia coli* (2/2)

- 1/3 of *E. coli* genes codes for enzymes → this allows for adjustment to a large scale of metabolic conditions:
- Composes all necessary proteins and nucleic acids
- Shows metabolic versatility: Development is possible under aerobic and anaerobic conditions using different energy generation pathways
- Metabolic pathways are activated depending on organismal needs – these pathways are not continuously active
- There are numerous enzymes even for particular metabolic reactions
- Development is possible using different sources of nitrogen and carbon
- Uses a wide range of carriers for substrate arrest and transfer

# *The tree of Life…postgenome*

**From The Tree of Life towards The Web of Life**

• Horizontal gene transfer complicates relationships between species, as well as the understanding of the phylogenetic tree of life, where all species were simply believed to be derived from a common ancestor

• The genes in a genome can represent different evolutionary histories, since even a rare gene transfer can cause different molecular genealogical relations

• The realistic scenario does not actually correspond to the Tree of life, but to the much more complex Web of life.

http://www.texscience.org/reports/sboe-tree-life-2009feb7.htm

# The Tree of Life – 4th domain? (1/2)

The discovery of large viruses (with extra large genomes) has prompted researchers to assume the existence of a fourth so far unknown domain

Mimivirus… Raoult, D. *et al. Science 306, 1344–1350 (2004).*

**Figure 12:** Mimivirus

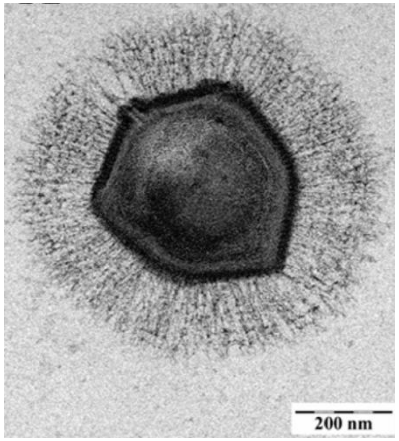
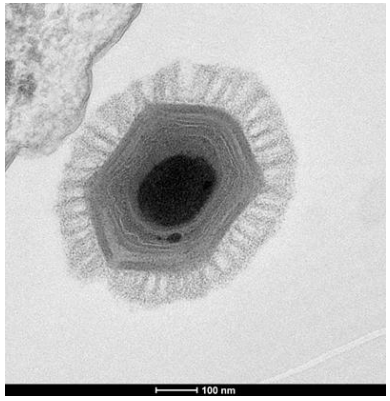



**Figure 13:** *Megavirus chilensis*




The largest known viral genomes are indicated in the following link, with the genome of ***Pandoravirus salinus*** (2,473,870 bp) being the largest http://www.giantvirus.org/top.html

# The Tree of Life – 4th domain? (2/2)

Metagenomic studies have been carried out in order to study the large viruses and to clarify whether they belong to a fourth Domain

Wu, D. *et al. PLoS ONE 6, e18011 (2011).*
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0018011

- **Figure 14** shows the phylogenetic tree constructed based on the sequence of the beta subunit of RNA polymerase II
- The nucleocytoplasmic large DNA viruses, NCLDVs are placed together in a possible fourth Domain

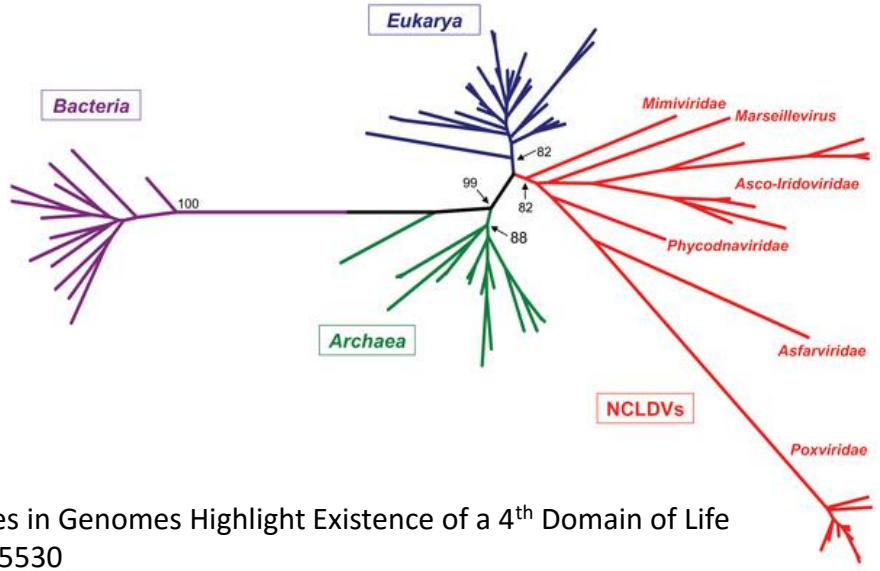



**Figure 14:** Phylogenetic and Phyletic Studies of Informational Genes in Genomes Highlight Existence of a 4th Domain of Life Including Giant Viruses. 2010. Boyer M., *et al*. PLoS ONE, 5(12): e15530
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0015530

# Metagenomes (1/9)

Metagenomics (Environmental Genomics, Ecogenomics, Community Genomics)

- The study of genetic material resulting directly from environmental samples
- Avoids problems related to the culture of microorganisms

| Sector | Percentage |
|---|---|
| Environmental | 58% |
| Host-associated | 33% |
| Engineered | 9% |

Distribution of metagenomics projects in GOLD (September 2011)

Instructions on how to submit results of a metagenomic project : http://www.ncbi.nlm.nih.gov/genbank/metagenome

# Metagenomes (2/9)

2002, Breitbart *et al*. showed that there are > 5,000 different viruses in 200 liters of seawater.

2004: Sample analysis of the Sargasso Sea identified DNA from> 2,000 different species with 148 new bacteria.

http://www.jcvi.org/cms/research/projects/gos/overview/

The new sequencing technologies and data analysis will allow even more impressive discoveries.

http://www.iscb.org/

# Metagenomes (3/9)

## The Sargasso Sea Experiment

The power of environmental metagenomics

J. Craig Venter, *et al*. Science 2 April 2004:Vol. 304. pp. 66 - 74

- Researchers sequenced microbial populations of water samples from the Sargasso Sea near Bermuda
- Production of > 1 billion bases of sequencing data
- High diversity and abundance levels of organisms were revealed
- > 1800 species genomes sequenced, 148 of which were previously unknown
- > 1.2 millions new genes identified

http://www.sciencemag.org/content/304/5667/66
http://www.jcvi.org/cms/research/projects/gos/overview/
http://blogs.jcvi.org/2010/10/second-leg-of-greek-sampling/

# Metagenomes (4/9)

- **Tara Oceans Expedition**: The first attempt of a global study on marine plankton
- Goal: Understanding the plankton ecosystem, by exploring the countless species and studying the interactions between them and their environment
- Over 12 multidisciplinary scientific specialties were involved in the program (oceanographers, ecologists, biologists, geneticists, physicists)
- Duration: 2009-2012
- The route followed: http://www.nature.com/news/systems-ecology-biology-on-the-high-seas-1.13665

- ***Tara Oceans Polar Circle* expedition**: Started in May 2013
- This program focuses on the plankton communities of the Arctic Ocean

http://oceans.taraexpeditions.org/ http://www.youtube.com/watch?v=eOMs8squk64

# Metagenomes (5/9)

- Timetable of microbial communities studies using high-performance sequencing.
- **Increase of data for microbial communities with NGS sequencing.**

Each cycle corresponds to biomonitoring projects deposited in NCBI (until May 2012), and shows the amount of produced data (based on the diameter of the circle) at the time of publication (x-axis). Red - human, black - other animals, green - environmental samples.

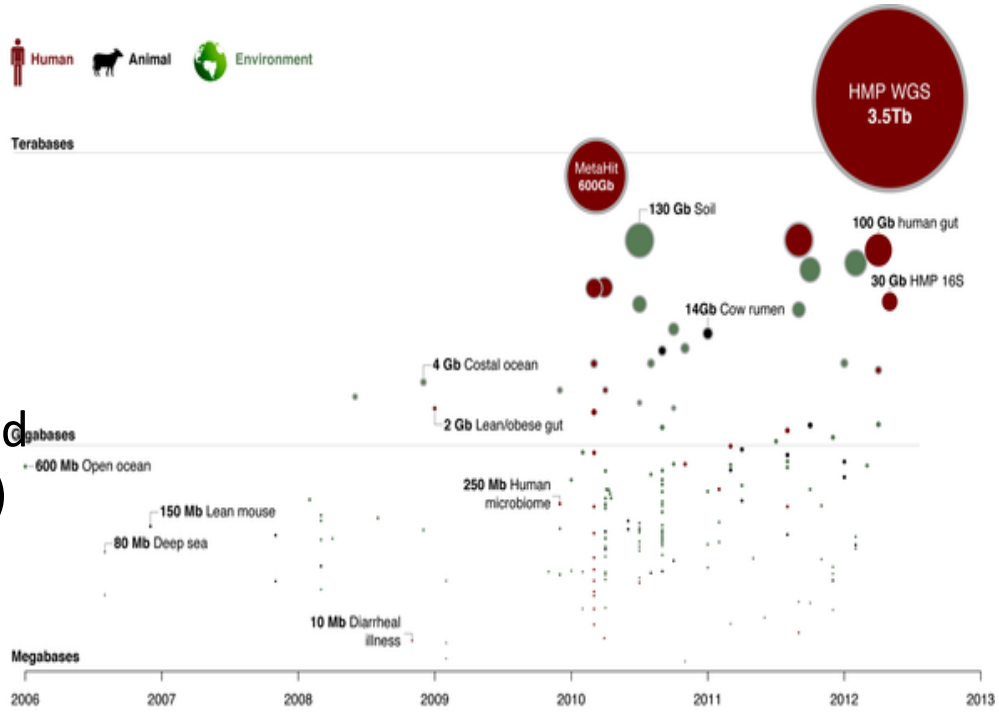



**Figure 15:** The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. 2012. Gevers D, *et al*. PLoS Biol 10(8): e1001377. doi:10.1371/journal.pbio.1001377
http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001377

# Metagenomes (6/9)

**Diet rapidly and reproducibly alters the human gut microbiome**

Consuming only animal or plant foods changes the composition of microbial community in the human gut rapidly.
Microbes colonize the intestine through food very quickly.
The microbiome can cope with the diversity and the changes in diet rapidly.

Nature 505, 559–563 (23 January 2014)
http://www.nature.com/nature/journal/vaop/ncurrent/full/nature12820.html

# Metagenomes (7/9)

**mtDNA haplogroup and single nucleotide polymorphisms structure human microbiome communities**
http://www.biomedcentral.com/1471-2164/15/257/abstract

Distribution of basic families of microbes in the intestine and the type of microbial community in the uterus in different European (H, J, K, T, U, V, W, X, I), Asian (B, F), African American (L2, L3) and Latin American (A, C) haplogroups
These data provide initial evidence for the association between host ancestral genome with the structure of its microbiome

# Metagenomes (8/9)

**Chimpanzees and humans harbour compositionally similar gut enterotypes**
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3520023/pdf/nihms425633.pdf
**Genomic variation landscape of the human gut microbiome**
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3536929/pdf/nihms-417625.pdf
**Alterations of the human gut microbiome in liver cirrhosis**
http://www.ncbi.nlm.nih.gov/pubmed/25079328
**Biogeography and individuality shape function in the human skin metagenome**
http://www.nature.com/nature/journal/v514/n7520/full/nature13786.html

# Metagenomes (9/9)

**Fifty thousand years of Arctic vegetation and megafaunal diet**

The first large-scale ancient DNA metabarcoding study of circumpolar plant diversity.

Number of species before, during and after the great Ice period
Analysis of the diet of eight large animals
http://www.nature.com/nature/journal/v506/n7486/full/nature12921.html

# New Large Bacterial Programs

http://jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity/

December 3, 2013

Pacific Biosciences, the Wellcome Trust Sanger Institute and Public Health England Collaborate to Finish Genomes of 3,000 Bacterial Strains. Rapid Finishing of Reference Genomes Possible Through Long-Read, High-Quality SMRT(R) Sequencing

# Genomes of prokaryotic organisms

## Generalizations for Bacterial Genomes

1. There are very few non-coding regions
2. Coding regions:
   - 25% of genes are unique
   - 50% of genes have unidentified function
   - There is frequent horizontal gene transfer
3. Huge intraspecific diversity (→ 22% differences within species)
4. The genes of bacteria are mainly organized in operons
5. Abundance: 1 gene / 1 Kb

# Genome size / Number of genes

- The diagram shows the correlation between genome size and number of genes
- The large genome size does not always correspond to an increased number of genes

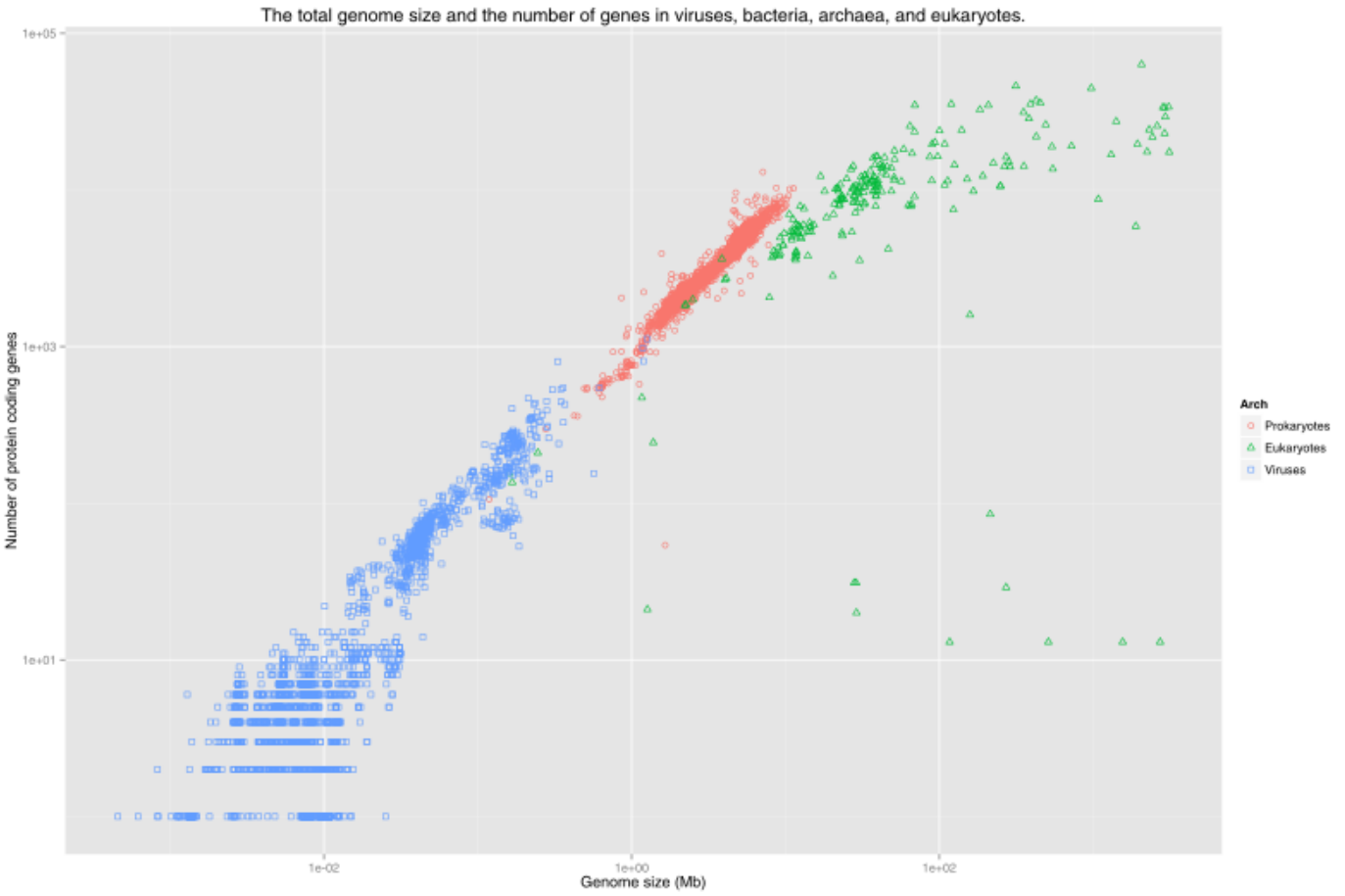



**Figure 16:** Genome size / Number of genes

# Eukaryotic genomes (1/3)

- A rapid increase in the number of published completed and draft genomes has been observed in recent years
- The list of published genomes is continuously updated :
  - Complete genomes: 334
  - Assembled / Contigs: 1985

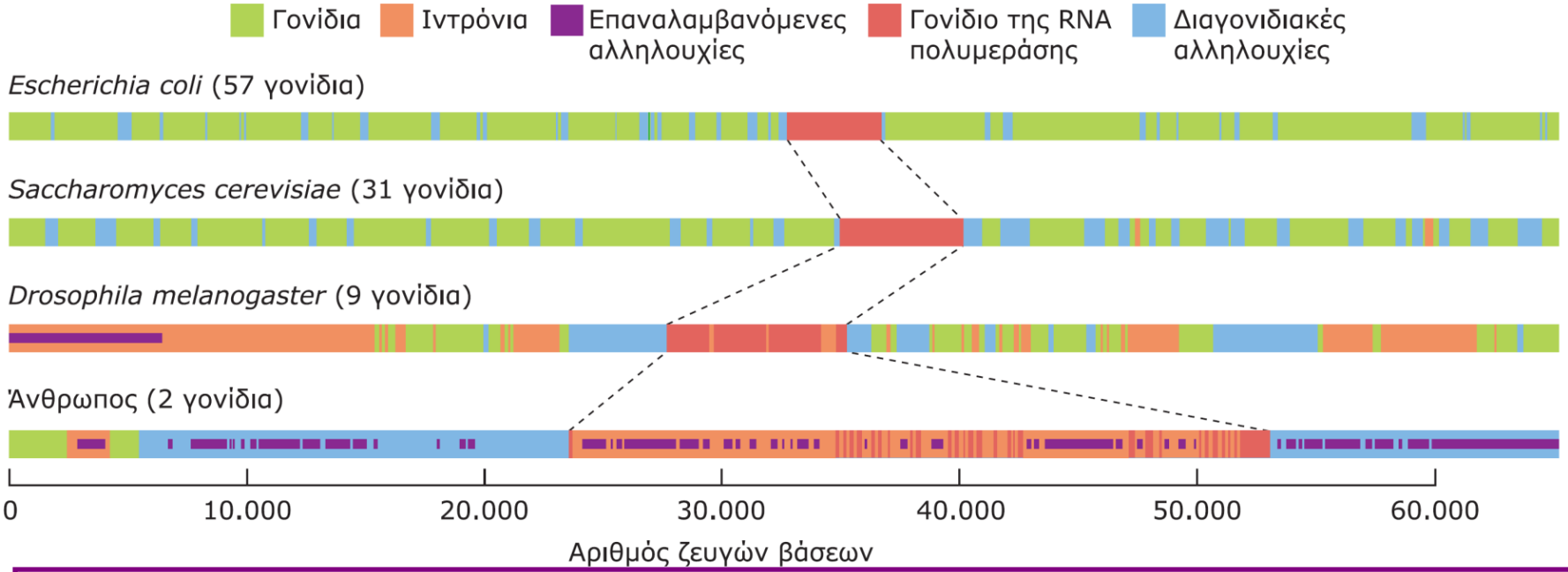
http://www.ncbi.nlm.nih.gov/genome/browse/

# Eukaryotic genomes(2/3)

| | Genome size (Mb) | Number of genes |
|---|---|---|
| *Saccharomyces cerevisiae* | 12.1 | ~6,000 |
| *Plasmodium falciparum* | 30 | ~6,500 |
| *Caenorhabditis elegans* | 97 | ~20,000 |
| *Arabidopsis thaliana* | 120 | ~25,000 |
| *D. Melanogaster* | 170 | ~16,000 |
| *Oryza sativa* | 415 | ~20,000 |
| *Zea mays* | 2500 | ~20,000 |
| *Homo sapiens* | 3200 | ~30,000 |
| *Hordeum vulgare* | 5300 | ~20,000 |
| *Picea abies* | 20,000 | 28,354 |

# Eukaryotic genomes (3/3)

**Figure 17:** Comparison of the chromosomal regions around the rRNA polymerase gene between *E.coli*, yeast, *Drosophila* and human species, which reveals the gene density variation in various genomes.




As we go up in complexity, genes are more sparsely distributed and there are more intronic and repeated regions.

# The yeast genome (1/7)

- The first eukaryotic organism sequenced (1996)
- Genome size of ~ 12 Mb in 16 chromosomes with a characteristic structure

http://medakagb.lab.nig.ac.jp/Saccharomyces_cerevisiae/index.html

# The yeast genome (2/7)

- Includes 5885 possible protein genes and 455 genes coding for various types of RNA. Many of these genes have homologues in prokaryote genes. Additionally, 46% of yeast genes are associated with homologous in human. The yeast genome is more compact than the genomes of nematodes, *Drosophila* and human. It has few genes (5%) with introns, which are small in size. It has few repeated sequences.
- Abundance 1 gene / 1.5-2 Kb
- A key difference from the bacterial genome is that in yeast (and generally in eukaryotes) many genes (15%) are present in the form of multiple copies. This is an indication of the origin of the yeast genome by duplication of the entire genome of an ancestral organism (Whole Genome Duplication -WGD)!

# The yeast genome (3/7)

- Whole Genome Duplication (WGD): There are indications that the yeast genome originated through he duplication of the entire genome of a ancestral organism
- Raw and abundant Material for Evolution!

- www.yeastgenome.org

Depiction of duplicated sections of yeast chromosomes due to WGD
http://www.nature.com/nature/journal/v428/n6983/fig_tab/nature02424_F3.html

# The yeast genome (4/7)

**An exercise on how to calculate the number of genes present in a genome bioinformatically**

The calculation of the hypothetical number of genes depends on the minimum allowable size of the protein:

- 260,000 ORFs in total with a size 2-99 codons
- 114,000 ORFs corresponding to proteins with size of >15 amino acids
- 7,505 ORFs corresponding to proteins with size of 100 amino acids

- Depiction of the total number of ORFs of indicated length encoded in the *S. cerevisiae* genome (Basrai et al 1997, Genome Res 7, 768-771)

http://genome.cshlp.org/content/7/8/768.long

# The yeast genome (5/7)

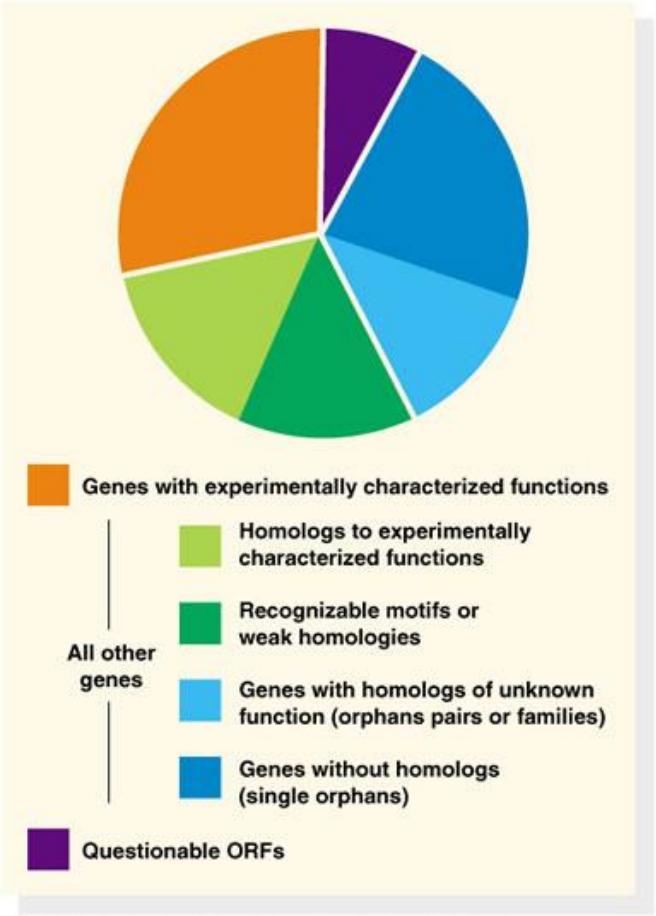





**Figure 18:** Distribution of what we know regarding potential ORFs in the yeast genome. (From B. Dujon. 1996. *Trends Genet* 12: 263-270.)

# The yeast genome (6/7)

To facilitate the study of genes, researchers group them together in some broader categories, depending on their function. Such groups are :

- Metabolism / Energy
- Cell growth, division and DNA synthesis
- Transcription
- Protein synthesis
- Protein destination
- Cellular transport and transport mechanisms
- Cell communication
- Cellular defense, cell death and aging
- Cellular organization
- Transposable elements, viral and plasmid proteins
- Unknown role

# The yeast genome (7/7)

## Saccharomyces Genome Resequencing Project

ABI sequencing machines were used
37 *S. cerevisiae* and 27 *S. paradoxus* strains were sequenced at small coverage (1x and 3x),
1.42 million reads in total and more than 1 Gb of sequence were produced

**WHY** was this program conducted? To:

- Compare sequences
- Find SNPs & INDELs
- Study the variation of copy numbers (CNVs)
- Study functional Diversity – Pseudogenes
- Study the population structure of *S. paradoxus* & *S. cerevisiae*
- Find data supporting selection phenomena
- Compare laboratory strains

# The nematode genome (1/3)

- The nematode has many advantages for studying growth in organisms:
- Easy and fast culture
- Ability to create homozygous lines by self fertilization
- Easy observation due to its transparent body
- Existence of 6 specific cell lines that grown into various tissues

- Since 1960, it has been a model organism for biology and genetics

http://www.nature.com/news/neuroscience-as-the-worm-turns-1.12461

# The nematode genome (2/3)

- The first multicellular organism sequenced (sequencing completed in 1998 with some errors, fully finished in 2002)
- Its size is relatively small ~ 97 Mb, organized in 6 **telocentric** chromosomes
- Includes 20,470 protein genes, more sparsely arranged than those of yeast
- … but also includes 16,000 RNA genes!
- Abundance: 1 gene / 5 Kb
- 43% of the genes are homologous to humans
- ~ 30% of the genome corresponds to non-coding regions
- *C. briggsae genome* recently completed
- Additionally three species are sequenced

http://www.wormbase.org/ http://www.sanger.ac.uk/Projects/C_elegans/

# The nematode genome (3/3)

- The number of genes that perform basic functions is about the same as yeast
- **Unexpected** as the nematode has ~ eightfold genome size
- Many genes are involved in intracellular communication

http://www.affymetrix.com/catalog/131405/AFFY/C.-elegans-Genome-Array#1_1

# The genome of *D. melanogaster* (1/5)

- It is one of the most important model organisms for Biology
- It was the second multicellular organism in which the deciphering of its genome was completed
- Partially sequenced in 2000 (by Celera company and state universities)
- First eukaryotic genome where whole genome shotgun technology approach was used

http://flybase.bio.indiana.edu/

# The genome of *D. melanogaster* (2/5)

- Genome size: ~ 170 Mb, organized into 5 different chromosomes
- ~ 1/3 constitutes of centromeric heterochromatin, multiple repetitions (AATAACATAG) & few genes
- Telomeres mainly include transposable elements

- In the euchromatin region (~ 120 Mb) only 15,000 genes are found
- Average abundance just one gene / 9 Kb
- Mostly unique genes
- Only 4,000 are essential for *Drosophila* survival!

http://www.ncbi.nlm.nih.gov/pubmed/10731132

# The genome of *D. melanogaster* (3/5)

- ~ 60% of the genes are homologous to mammalian genes.
- 177 genes concern also human diseases (cancer and cardiomuscular, neurological, endocrine, metabolic and hematological diseases).
- Transgenic, knock-out, and mutated insects with added human genes, such as those associated with Parkinson's disease can be created.
- Affymetrix Chips exist to analyse gene expression. But for > 20% of genes we do not know their function.
- In November 2007, the analysis of 12 other species of the genus was completed!
  http://www.nature.com/nature/journal/v450/n7167/fig_tab/nature06340_F1.html
- September 2009: sequencing of 50 *D. melanogaster* individuals.

http://www.bbc.co.uk/news/science-environment-12686745

# The genome of *D. melanogaster* (4/5)

## The transcriptome of *Drosophila*

- Researchers in the modENCODE program used RNA-seq, microarrays and cDNA sequencing to analyze gene expression in 30 different developmental stages!!!
- They Identified > 110.000 new elements in the genome, ie. Genes, coding and noncoding transcripts, exons, introns and protein isoforms that until now were unknown
- 1500 genes with unknown function stil exist

http://www.youtube.com/watch?v=jmR0uK8KbO8&list=PL029759EA7A7D4607&index=5

Nature 471, 473–479 (24 March 2011)

# The genome of *D. melanogaster* (5/5)

## *Drosophila* Genetic Reference Panel (DGRP)

- 192 homozygous series have been created with fully sequenced genome
- More than 4,5 M SNPs are recorded
- The DGRP is a living library of common polymorphisms that affect the complex characteristics and a resource of gene association mapping of quantitative characteristics
- Has allowed for the analyses of genotype and phenotype interactions

Nature 482, 173–178 (9 Feb 2012)
http://www.nature.com/nature/journal/v482/n7384/full/nature10811.html

http://dgrp.gnets.ncsu.edu/

# The genome of *A. thaliana* (1/3)

- Has a small genome ~ 120 Mb in 5 chromosomes
- Its reproduction is fast
- It produces a large number of seeds
- It is self- / hetero- sexually fertilized
- It has a small size
- Its cultivation is easy

- Small presence of repeated sequences

# The genome of *A. thaliana* (2/3)

- Completed (92% of it, including all euchromatin regions) in 2000 by a multinational group
- The following link shows the groups that contributed to the sequencing of chromosome 5:

http://openi.nlm.nih.gov/detailedresult.php?img=138921_gb-2001-2-4-comment2004-1&req=4

- Includes 25,498 coding genes
- Includes a small number of introns and small exons
- Abundance: 1 gene / 4.6 Kb
- Transposable elements: 10%, much less than corn with a value of 50-80%

- http://www.nature.com/nature/journal/v408/n6814/full/408796a0.html

# The genome of *A. thaliana* (3/3)

- 70% of genes are duplicated or members of gene families
- The genome of the species originated from the union-duplication of two genomes 110 million years ago

http://www.nature.com/nrg/journal/v2/n7/fig_tab/nrg0701_493a_F3.html

- There are only ~ 15,000 completely different genes
- Probably, in the progenitor species of *Arabidopsis* the genome duplicated in order to generate a tetraploid species

2008- 1001 genome project! - www.1001genomes.org
Sept 2014- 1100 individuals have been sequenced, results expected

# The mouse genome (1/6)

- 2001: Celera completed 5X coverage of the genome and sold it to private pharmaceutical companies
- 2002: 96% of the genome was published by an international group of 27 laboratories, supported by government
- It was the second finished mammalian genome after human
- Announced much earlier than expected, and with a small budget (**$ 130 million**)
- A combination of hierarchical and WG strategy was used
- 2005: 99.9% of the genome has been completed

http://www.nature.com/nature/journal/v420/n6915/full/nature01262.html

# The mouse genome (2/6)

- Size: ~ 2.6 Gb approximately 14% smaller than the human genome → lower amount of repetitive DNA & higher loss rate of nonfunctional DNA fragments.
- Twice (and above) mutation rate in mice compared to humans.

The reasons for this difference are not known. It is due to either the shorter reproductive generation or the smaller body size in combination with a higher metabolic rate.

Consequence: More mutations are inherited through the germ cell in the same time period in mouse compared to humans.

# The mouse genome (3/6)

- ~ 90% of the mouse genome has corresponding chromosomal regions in humans (**synteny**) – enabling the faster assembly of the mouse genome.
- At the nucleotide level, 40% of the mouse genome is conserved in comparison to humans.
- Synteny has applications in discovering genes (e.g. alcoholism genes).

http://www.mun.ca/biology/scarr/MGA2-11-33smc.html

Areas functionally important - have been preserved for 75 million years.
They include protein genes, regulatory regions and structural chromosomal regions.

# The mouse genome (4/6)

## The genome of 17 mouse strains

**Mouse genomic variation and its effect on phenotypes and gene regulation**

Keane T.M. *et al*. 2011 Nature, 477, 289–294

- The genomics analysis of 17 different wild and laboratory mouse strains was completed.
- The four basic genomes of wild mice from which these strains were generated were analyzed.
- Over 8 million new SNPs and 718 loci associated with various regions of quantitative interest, were identified.
- This work contributes to a better understanding of the function of the mouse genome and the association of phenotype with the molecular genetics.
- These sequences provide a starting point for a new era in the functional analysis of a key model organism.

http://www.nature.com/nature/journal/v477/n7364/fig_tab/nature10413_F1.html

# The mouse genome (5/6)

- Includes <30,000 genes, a relatively small number.
- The largest percentage, deducting the basics, is associated with reproduction, smell and disease resistance.
- ~80% of the genes are also found in humans.
- Just 1% of genes (118 genes) has no resemblance to those of other organisms.

Until 2014 genetically modified laboratory knock-out mice strains will be created, each one missing one of the >~20,000 protein genes.

# The mouse genome (6/6)

- **Genome-wide Generation and Systematic Phenotyping of Knockout Mice Reveals New Roles for Many Genes, Cell 154 2013 452-464**
- Sanger Institute Mouse Genetics Project

  - Analysis of 489 genes with knock-out mutants
  - More than 40% of these were necessary for life
  - Many non-studied genes had unexpected phenotypes

http://www.cell.com/abstract/S0092-8674(13)00761-7

# The rat genome (1/3)

- April 2004: a draft of 90% of the rat genome was announced by an international group of 20 laboratories
- A combination of hierarchical and WG strategy was used → (7X coverage)
- Rat: model for studies of drug effects and their toxicity
- Comparative genomics were done between the three available mammals genomes, based on the evolutionary divergences between rodents - humans (75 million years ago) and rat-mouse (12-24 million years ago)

http://www.nature.com/nature/focus/ratgenome/

# The rat genome (2/3)

- Size ~2,75 Gb, smaller than human, but bigger than mouse.
- Missing mainly telomeric, centromeric regions.

- ~ 40% of the rat genome is fully conserved compared to human and mouse.
- These areas mainly code for proteins but not only.
- They also include regulatory regions,

chromosomal structural regions, and
coding regions for RNA molecules.

**Figure 19:** *Rattus auratus*

# The rat genome (3/3)

- Includes ~ 30,000 genes.
- 90% common genes between rat and mouse. The remaining 10% concerns protein families highly abundant in one of the two species particularly, and in rat mainly concern genes producing pheromones and contributing to detoxification and immunity.
- Of particular interest, in rodents, are P450 genes that help in processing of drugs and protection from toxic side effects. This gene family is more abundant in rodents than in humans. Therefore, it is difficult to use rodents in drug toxicity tests in humans in some cases, since these genes may be absent from the human genome. But, by sequencing the genome we can now fully analyze these genes and find if for some drugs genes are activated that have homologues in humans.
- Also, the number of these genes in rat is more similar to humans (as compared to mouse) and this perhaps explains why rats were always better models of toxicity studies than mice.
- It should also be noted that all the 'disease' genes in humans have orthologous genes in rat.
- The mutation rate is 3 times higher than in humans.

# The genome of *Tetraodon nigroviridis*

- Published in October 2004 based mainly on use of WGS (as in *Takifugu rubripes*)
- More than 10% of euchromatin missing
- Size: 340 Mb (one of the smallest vertebrate genomes)
- Includes 20,000-25,000 genes (36 Mb, 11% of the genome)

- Includes 21 chromosomes in phenomenon of whole genome duplication (WGD) is strongly evident
- Divergence from mammals 450 million years ago
- Based on comparative genomics studies it is now hypothesized that the ancestor of vertebrates had 12 chromosomes
- In man only one chromosome has not undergone major transformations

http://www.nature.com/nature/journal/v431/n7011/full/nature03025.html

**Figure 20:** *Tetraodon nigroviridis*

# The chicken genome (1/2)

- Published in December 2004
- Combination of HGS & WGS (7X coverage, in 1 year), still missing 2% of the genome
- 2 sex chromosomes : ZW females ZZ males, autosomes = 10 macrochromosomes & 28 microchromosomes (without repetitive sequences)
- Size: 1 Gb (much less repetitive regions)
- Includes 20,000-23,000 genes (many quantitative)

**Importance**

- Finding regulatory elements
- Divergence of chicken: 300 million years ago
- A parallel analyses of three farm types
- Discovery of 2.8 million SNPs
- Polymorphism 5 SNPs/Kb, 7-fold compared to human

**Figure 21:** *Gallus gallus*

# The chicken genome (2/2)

- There is a widespread gene family for keratin protein specific for flakes, legs and wings, but not for the keratin type of mammalian hair
- Missing genes responsible for milk production, creating teeth and pheromones
- There is an extensive family of genes for smell (!)
- There are few genes for taste

http://www.nature.com/nature/focus/chickengenome/

# The bee genome (1/2)

- Published in November 2006, the genome of *Apis mellifera,* was the third insect genome sequenced.
- WGS (7.5X coverage), still missing more than 4%
- Includes 16 chromosomes (no sex chromosomes present)
- Size: 236 Mb
- Perhaps includes just ~ 10,000 genes

**Figure 22:** *Apis mellifera*

**Importance**

- Social bee behavior. How can the same genome resulting into queens and worker bees? →
- Existence of 65 microRNAs with different functionality depending on the role of each bee. Monitoring gene activation with microarrays.

http://www.nature.com/nature/focus/honeybee/

# The bee genome (2/2)

- There is an extensive gene family for pheromones
- It includes reduced genes for taste (fewer risks due to feeding mode)
- There are fewer immune genes!

- In 2011 a metagenomic analysis of the beehive and its microbiome was conducted

http://www.nature.com/nature/journal/v443/n7114/fig_tab/nature05260_F5.html

# The genome of cod

The genome sequence of Atlantic cod reveals a unique immune system

Star *et al*. 2011, Nature, 477(7363): 207–210

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3537168/pdf/emss-50854.pdf

- 22,154 genes were identified, using 454 sequencing
- A complex system for thermal adaptation is evident
- It has lost genes for MHCII and CD4
- Although lacking MHCII complex, it has developed adaptations in its immune system

# The genome of African cichlid fish

**The genomic substrate for adaptive radiation in African cichlid fish**

A number of genome adaptations were found in these species associated with rapid rates of evolution and divergence in both coding and regulatory regions

Nature (2014) 513, 375-381

# Overall Genomics Analysis of Biodiversity (1/2)

BGI – Beijing Genome Institute (http://www.genomics.cn/en/index_
Pioneer in the analysis of biodiversity. In 2012 it had published 80 animal genomes and had finished a total of 540 genomes of animals and plants.
Study of species like the naked mole rat demonstrates that there are important adaptations to organisms with human interest

BGI has already initiated ambitious programs like
Million Plant & Animal Genomes Project
http://www.genomics.cn/en/navigation/show_navigation?nid=5657
One million micro-ecosystem Genomes Project
http://www.genomics.cn/en/navigation/show_navigation?nid=5659

# Overall Genomics Analysis of Biodiversity (2/2)

## GENOME 10K Project

- *Analysis of 10,000 vertebrate genomes*
- *The purpose is the study of biodiversity*
- *Budget 100 M $*
- *Participation of 68 scientists from five continents*
- *http://genome10k.soe.ucsc.edu*

The number of vertebrate species to be analyzed under the program per taxonomic level

| | Orders | | | Families | | | Genera | | | Species | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Groups | With G10K samples | Total | % of total | With G10K samples | Total | % of total | With G10K samples | Total | % of total | With G10K samples | Total | % of total |
| Mammals | 27 | 27 | 100 | 145 | 150 | 97 | 763 | 1230 | 62 | 1826 | 5416 | 34 |
| Birds | 32 | 34 | 94 | 182 | 199 | 91 | 1587 | 2172 | 73 | 5074 | 9723 | 52 |
| Amphibians | 3 | 3 | 100 | 50 | 56 | 89 | 301 | 510 | 59 | 1760 | 6570 | 27 |
| Reptiles | 4 | 4 | 100 | 63 | 65 | 97 | 751 | 1087 | 69 | 3297 | 9002 | 37 |
| Fishes | 62 | 62 | 100 | 424 | 532 | 80 | 1777 | 4956 | 36 | 4246 | 31 564 | 13 |
| Totals | 128 | 130 | 98 | 864 | 1002 | 86 | 5179 | 9955 | 52 | 16 203 | 62 275 | 26 |

Journal of Heredity 2009:100(6):659–674

# Note of use of third party works (1/2)

***Escerichia coli***, http://commons.wikimedia.org/wiki/File:E_choli_Gram.JPG, by Bobjgalindo, CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/deed.en).
***S. cerevisiae*** under DIC microscopy, http://commons.wikimedia.org/wiki/File:S_cerevisiae_under_DIC_microscopy.jpg, by Masur.
***C. elegans***, http://commons.wikimedia.org/wiki/File:C_elegans_stained.jpg, by Public Library of Science journal, CC-BY-2.5 (http://creativecommons.org/licenses/by/2.5/deed.en).
***D. melanogaster***, http://commons.wikimedia.org/wiki/File:Drosophila_melanogaster-side_(aka).jpg, by André Karwath, CC-BY-SA-2.5 (http://creativecommons.org/licenses/by-sa/2.5/deed.en).
***A. thaliana***, http://commons.wikimedia.org/wiki/File:Arabidopsis_thaliana_inflorescencias.jpg, CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/deed.en).
***M. musculus***, http://commons.wikimedia.org/wiki/File:Kletterk%C3%BCnstler_Hausmaus.JPG, by 4028mdk09, CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/deed.en).
***Nanoarchaeum equitans***, http://commons.wikimedia.org/wiki/File:Urzwerg.jpg, by Karl Stetter.
***Methanopyrus kandleri***, http://en.wikipedia.org/wiki/File:Arkea.jpg, by ms: User:PM Poon, CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/deed.en).
***Vibrio cholerae*** , http://en.wikipedia.org/wiki/File:Cholera_bacteria_SEM.jpg, by Zeimusu.
**Mimivirus**, http://commons.wikimedia.org/wiki/File:Electron_microscopic_image_of_a_mimivirus_-journal.ppat.1000087.g007_crop.png, by Ghigo E, Kartenbeck J, Lien P, Pelkmans L, Capo C, Mege JL, Raoult D, CC-BY-SA-2.5 (http://creativecommons.org/licenses/by-sa/2.5/deed.en).

# Note of use of third party works (2/2)

***Megavirus chilensis***, http://commons.wikimedia.org/wiki/File:Megavirus.jpg, by Chantal Abergel, CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/deed.en).

**Genome size / Number of genes**, http://upload.wikimedia.org/wikipedia/commons/0/0d/Genome_size_vs_number_of_genes.svg, by Estevezj, CC-BY-SA-3.0 (http://creativecommons.org/licenses/bysa/3.0/deed.en).

***Rattus rattus***, http://commons.wikimedia.org/wiki/File:Rattus_norvegicus_-_Fairlands_Valley_Park,_Stevenage,_England-8.jpg, by Anemone Projectors, CC-BY-SA-2.0 (http://creativecommons.org/licenses/by-sa/2.0/deed.en).

***Tetraodon nigroviridis***, http://commons.wikimedia.org/wiki/File:Gr%C3%BCner_Kugelfisch.png, by Viktoria Schmuck, CC-BY-SA-3.0 (http://creativecommons.org/licenses/bysa/3.0/deed.en).

***Gallus gallus*** , http://commons.wikimedia.org/wiki/File:Gallus_gallus_male_big.jpg, by Guam, CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/deed.en).

***Apis mellifera***, http://commons.wikimedia.org/wiki/File:Apis_mellifera_carnica_worker_hive_entrance_2.jpg, by Bartz R., CC-BY-SA-2.5 (http://creativecommons.org/licenses/by-sa/2.5/deed.en).

# Reference note



# Licensing note

# 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.