Special Topics on Genetics

Section 4: The human genome

Triantafyllidis A.

School of Biology

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

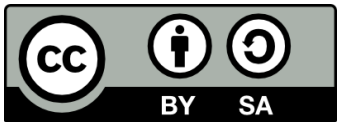
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

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.

Section Contents

- The human genome
- Applications and Aims
- Conclusions
- Organization of genes
- Enhancement of genetic diversity

The human genome (1/11)

- The first discussions about the mapping of the genome began in 1986
- The draft genome was announced in 2001
- 2 programs: IHGSC (International Human Genome Sequencing Consortium) & Celera
- Project Budget: 3 billion dollars

The human genome (2/11)

1988. Launch of the program. Countries that participates: USA, China, Japan, France, U. Kingdom, Germany

1988. James Watson was the initial director of the program

1992. Fr. Collins was appointed Director

1993. HGP is planned to be completed in **2005**

1994. The complete **genetic map** is published

1995. The complete **physical map** is published

1996. Deposit of DNA sequences in web officially starts

The human genome (3/11)

1998. C. Venter forms the Private Company Celera. His aim was the genome completion within 3 years with a cost of just 300 million dollars https://www.celera.com/

1998. The goal of the public sector for finishing the genome is now **2003**

1999. The complete sequencing of human chromosome 22 is announced

2000. German and Japanese scientists complete sequencing of the 21st human chromosome

The human genome (4/11)

2001. February. HGP and Celera announce the first draft of the human genome in Nature & Science respectively

http://www.nature.com/nature/journal/v409/n6822/fig_tab/409860a0_F1.html

2003 April. The completion of HGP is announced officially in the press, simultaneously with the 50th anniversary celebration of the discovery of the double helix

2013… New versions appear constantly (Ensembl No GRCh38, 12/2013) www.ensembl.org

The human genome (5/11)

- 21 October 2004 - New publication in Nature
- In 2001 10% of euchromatin and 30% of heterochromatin were missing
- There were gaps and incorrectly placed fragments
- Version of 2004 possessed only 341 gaps, 99% of euchromatin , 1 mistake in 100,000 bases
- Problems due to many mistakes in WGS
- 1% of euchromatin and regions near centromeres are still missing

The human genome (6/11)

- Whose is the genome?
- The state sector version was based on cloned haploid chromosome fragments
- Celera used fragments from 2 males and 3 females of different nationalities (from Africa, China, Latin America and Europe)
- For this reason the inter-individual variability was not correctly calculated

The human genome (7/11)

Genome Reference Consortium: provides the best possible reference assembly for human, e.g. by generating multiple representations for regions that are too complex.

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/index.shtml

Details and statistics about the assembly of the human genome are found in the following link:
http://www.ensembl.org/Homo_sapiens/Info/Annotation#assembly

http://www.youtube.com/watch?v=sPE9j_Hw9HU

The human genome (8/11)

2008: The 1000 genomes program is announced. It is an ambitious effort aiming to sequence the genome of at least 1000 people from all over the world, in order to create a more detailed picture of human genetic diversity.
http://www.1000genomes.org/sites/1000genomes.org/files/docs/1000genomes-newsrelease.pdf

October 2008: Complete Genomics announced that it can sequence the human genome with a cost of $5,000!!!
http://www.technologyreview.com/news/410939/five-thousand-bucks-for-your-genome/

November 2008: The complete genome of two men (a Chinese and a Nigerian) is publlished in Nature.
http://www.newscientist.com/article/mg20026813.400-chinese-and-nigerian-men-join-elite-genome-club.html

The human genome (9/11)

- Size: ~3 billion pairs of nucleotides (3096649726 bp)
- 30 times larger than that of *C. elegans* and 20 times than that of *D. melanogaster.*
- There are only 20-21,000 protein genes (4,000 are responsible for hereditary diseases)
- Two people differ only in 0.2% (although according to other estimations the difference is <3 %)
- The sequence homology between human and chimpanzees is 98.5%
- The size of autosomes ranges from 48 Mb - 279 Mb
- The size of X chromosome is 163 Mb, while that of Y only 51 Mb
- The book with the complete human genome sequence will never be printed

The human genome (10/11)

Applications

- Pre- and postnatal diagnosis and prognosis of hereditary diseases
- Predetermination and preselection of the individual phenotype (pre - implantation genetic diagnosis)
- Gene therapy
- Solving forensics problems
- Creation of a bank with genetic data
- Production of Vaccines and Diagnostics
- Transplants
- Production of drugs (Pharmacogenetics)

The human genome (11/11)

Objectives

The main objectives of the Human Genome Sequencing Project were the following :

- The construction of physical maps, genetic maps and sequence maps of human genome
- The sequencing of the genomes of a number of model organisms
- The development of advanced technologies for mapping and sequencing
- The development of computational methods for finding, analysis, display, storage of data from maps and the sequences
- The sequencing of ESTs from cDNA libraries and of whole cDNAs (expressed mRNA) in various human cell types
- The debate on ethical, social and legal challenges

The human genome – Conclusions (1/23)

There are less than 30,000 genes in the human genome.

They constitute only 5% of the total genome size.

• Small genes are difficult to be identified e.g. a large number of genes are transcribed into functional small RNA molecules (~ 25 bp). Their number may be too large.

• Genes that express small polypeptides are difficult to identify.

• Some genes are rarely expressed and may not follow the usual genetic code.

The function is known for only about half of the genes.

The human genome – Conclusions (2/23)

Known human genes (01/2013)	Autosomes	Chromosome X	Chromosome Y	Mitochondrial	**Total**
Known sequence	13370	651	48	35	14104
Known sequence and phenotype	124	4	0	2	130
Phenotypic description, molecular basis known	3371	271	4	28	3674
Mendelian Phenotype, molecular basis unknown	1627	133	5	0	1765
Other, mainly phenotypes with suspected mendelian basis	1765	125	2	0	1892
Total	20257	1184	59	65	**21565**

http://www.ncbi.nlm.nih.gov/Omim/mimstats.html

The human genome – Conclusions (3/23)

Genes encode for non translated RNAs or proteins

Non translated RNAs

There are ~ 22,170 (!!!) such genes, which are classified in the following categories:

- Transfer RNAs (tRNAs)
- Ribosomal RNAs (rRNAs)

There are numerous other small RNAs, expressed at low levels

As well as… circRNAs – circular RNA! (Nature 2013, 495)

The main categories and functions of mammal ncRNAs are shown at the link below:

http://www.nature.com/nature/journal/v482/n7385/full/482310a.html

The human genome – Conclusions (4/23)

Genes that code for proteins:

- It is established that 20,364 such genes exist (Ensembl GRCh38, 12/2013, http://www.ensembl.org/Homo_sapiens/Info/StatsTable).
- There are 14,415 pseudogenes !!
- Genes that code for proteins occupy a total region of only ~50 Mb
- The average size of human genes is 27 Kb (they are generally larger and with more and larger introns than invertebrates, such as Drosophila).
- There are ~10 exons/gene and
- Up to 30 to 50 introns in a gene →

There is variation from 0 (histones) to 234 (a muscle protein)

The human genome – Conclusions (5/23)

Genes smaller than 10kb:

- $tRNA^{tyr}$
- Histone H4
- Interferon-a
- Insulin
- B-globulin
- HLA class I

Genes smaller than 100kb:

- Albumin serum
- Collagen a1 II)
- HPRT
- Apolipoprotein B
- LDL receptor
- Phenylalanine hydroxylase

Genes larger than 100kb:

- Factor VIII
- CFTR
- NF1
- Utrophin
- Immunoglobulin heavy chain
- Immunoglobulin light chain kappa
- Dystrophin

The human genome – Conclusions (6/23)

The HUPO team (Human Proteome Organisation) aims to discover and define the structure and function of all human proteins.

The human proteome comprises **homologous** proteins (ie. proteins with common evolutionary ancestor) with 78.5% of mouse proteins, 61% of *Drosophila* proteins, 43% of *C. elegans* proteins, and 46% of yeast proteins.

Genes families exist which are found exclusively in vertebrates (and are usually associated with the immune and nervous systems).

The human genome – Conclusions (7/23)

The comparison with other genomes reveals that the distribution in other phylogenetic groups of possible homologous proteins to human proteins is as follows:

Vertebrates exclusively	22%
Vertebrates and other animals	24%
Animals and other eukaryotes	32%
Eukaryotes and prokaryotes	21%
No homology with animals	1%
Only prokaryotes	1%

The human genome – Conclusions (8/23)

Organization of genes

Genes are not randomly distributed in chromosomes. Thus we find:

- Gene families
- Regions rich in genes
- Regions poor in genes

The human genome – Conclusions (9/23)

Gene Families

- They can be found on a specific chromosome or not
- Known family genes: histones, haemoglobins, immunoglobulins, tubulins, heat shock proteins (1200 families)
- Gene families evolve by duplication of and evolutionary divergence from a common ancestral gene
- They show HOMOLOGY

The human genome – Conclusions (10/23)

Evolution of globulins

One of the major gene families are the alpha and beta globulins : http://www.mun.ca/biology/scarr/Globin_gene_families.html

- Gene duplications occurred in the evolution of human globulin gene families
- The initial separation resulted in two related lines: Myoglobin and haemoglobins
- A further duplication gave rise to the alpha- and beta-haemoglobins
- New duplications produced the family members of alpha- and beta-globulins

http://evolution-textbook.org/content/free/contents/ch27/ch27-f29.html

The human genome – Conclusions (11/23)

Gene Families

At some point an ancestral globulin gene copy duplicated and was translocated to another chromosome.

Creation of alpha and beta globulins

New duplications gave 3 copies of alpha globulins and 5 copies of beta globulins.

Other duplications resulted in PSEUDOGENES (due to among others: reading frame displacements, amino acid changes and premature end of translation mutations, or even loss of regulatory regions).

The human genome – Conclusions (12/23)

Regions rich in genes

✓ In human chromosome 6, 60 histocompatibility genes are found in a region of 700 Kb

✓ 70% of that DNA is transcribed

✓ The GC content of this region is 55%, while the average of the genome is 45%

http://www.nature.com/nrg/journal/v5/n12/full/nrg1489.html

The human genome – Conclusions (13/23)

Regions poor in genes

✓ The chromosomes with the smallest number of genes are 13 and Y

✓ There are ~ 80 gene "deserts", regions larger than 1 Mb without a gene

✓ New publications report that there are 545 such regions in mammals that are larger than 0.64 Mb

✓ They constitute 3% of chromosomes

✓ The largest region is 4.1 Mb

The human genome – Conclusions (14/23)

Regions poor in genes

It is difficult to find very small and very large genes

Large genes, over 500 Kb (e.g. the gene responsible for muscular dystrophy is ~ 2.3 Mb long): 124 large genes possess a total size of 112 Mb.

These genes surprisingly possess mRNAs with small size, usually only 1% of the total nuclear gene. Indicating presence of very large introns!

Large genes are slowly synthesized & usually expressed in nerve cells.

The human genome – Conclusions (15/23)

Enhancement of genetic diversity both at DNA and RNA levels is achieved through the combination of different basic strategies.

• *DNA level*

<u>T-lymphocytes Receptor genes</u>

They consist of three regions V, D, J. There are 45, 2 & 11 different copies respectively within a region of 700 Kb in chromosome 7.

D + J (through deletion) → DJ dimer

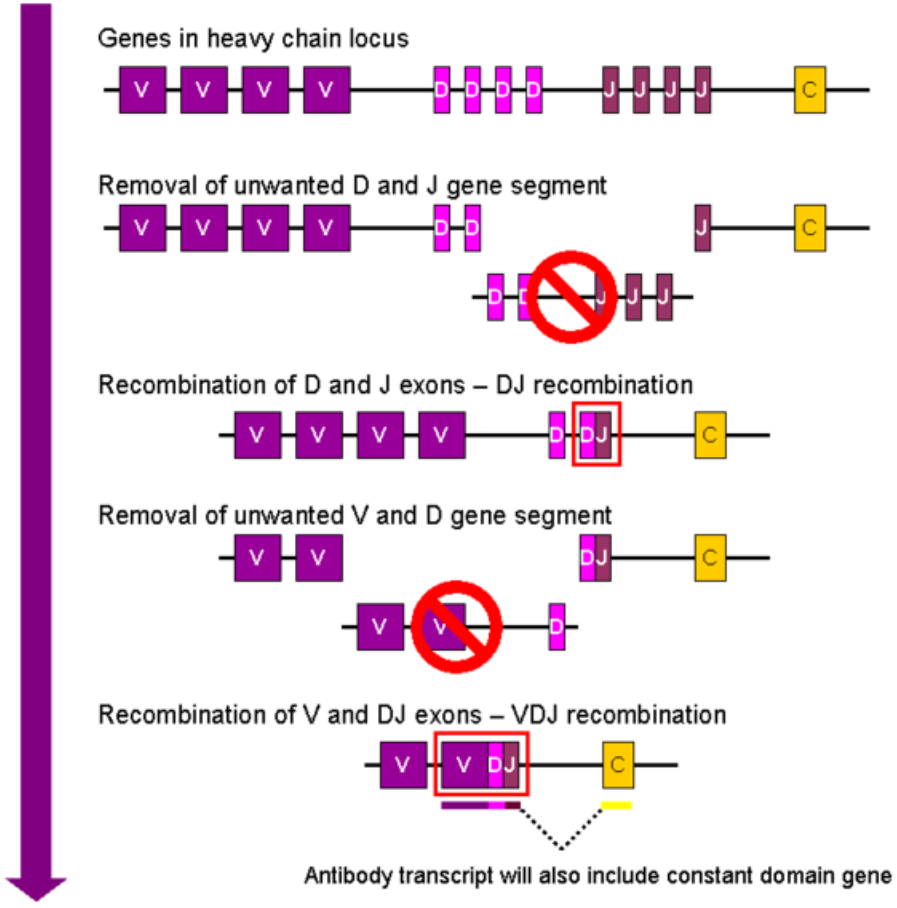
DJ + V (through subsequent deletion) → V-D-J

990 different V-D-J genes are produced

(45 X 2 X 11) from 58 copies of VDJ gene fragments

The human genome – Conclusions (16/23)

Figure 1: The diversity resulting from the recombination of V-D-J genes.



http://en.wikipedia.org/wiki/File:VDJ_recombination.png

The human genome – Conclusions (17/23)

- *RNA level*

It is possible to create greater genetic diversity through differential splicing as well as due to different transcription promoters.

The three neurexin genes have two transcription promoters (which produce alpha and beta mRNAs) as well as five positions in which differential spicing may occur.

More than 2000 mRNA types can be thus created.

The human genome – Conclusions (18/23)

ENCODE: Encyclopedia of DNA Elements

The goal of ENCODE is to build a comprehensive list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and gene activation.

http://www.nature.com/nature/journal/v447/n7146/pdf/nature05874.pdf
http://www.genome.gov/10005107
http://www.genome.gov/25521554

modENCODE: Model Organism Encyclopedia of DNA Elements

Goal: Studying the genome of *D. melanogaster* and *C. elegans*.

http://www.genome.gov/modencode/

The human genome – Conclusions (19/23)

6/9/2012

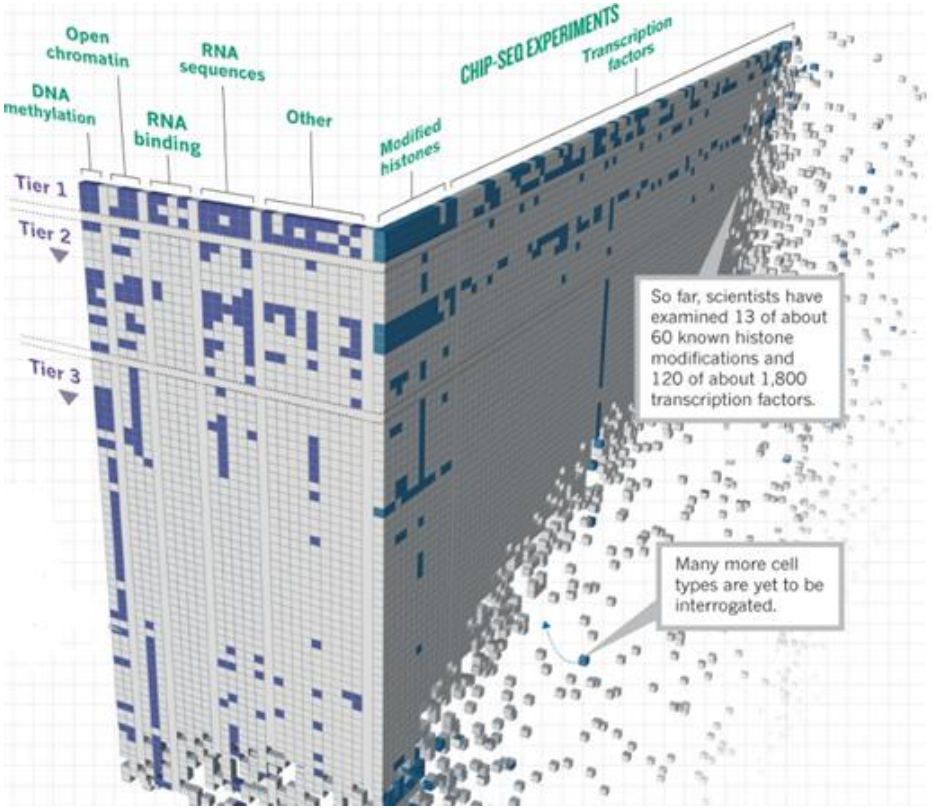


Figure 2: The results of a large survey of 32 research groups. They identified the DNA binding sites for about 120 transcription factors.

Roughly 80% of the genome has some sort of function with >70,000 promoter regions and 400,000 enhancer regions. It has been found that most of the genome is 'transcribed' into non-coding RNA molecules.

http://www.nature.com/news/encode-the-human-encyclopaedia-1.11312

The human genome – Conclusions (20/23)

Different mutation rates in the two human sexes

Male mutation rate is two times the female mutation rate

- In respectively placed repeated regions on chromosomes X and Y, it is found that the mutation rate is higher in males
- This higher mutation rate also affects autosomes in males, but this cannot be checked

- It probably occurs because of the higher rate of cell divisions in male germ cells
- Therefore, the majority of mutations in the human genome occurs in males
- These are not just harmful mutations but mutations that create diversity as well

The human genome – Conclusions (21/23)

Genome differences in different races

The human races have **very small differences**

Two same race individuals may differ much more than two individuals that belong to different races

•The human species is one and supposed differences are superficial

http://www.nature.com/nature/journal/v513/n7518/full/513306a.html

Gene mirror geography within Europe

Novembre *et al*. 2008, *Nature* **456**, 98-101

3000 samples were studied using half a million SNPs

The main conclusion is that the genetic distances between European populations reflect their geography and history.

The human genome – Conclusions (22/23)

The National Insitutes of Health of USA has a continuously updated list (http://www.genome.gov/26525384) indicating all SNP polymorphisms correlated with any human phenotype.

Genome wide association studies (GWAS) are association analyses based on the whole genome. They help us to understand that as regards the inheritance of characteristics, the monogenic explanation (1 gene = 1 phenotype) is probably the exception. The most likely model is the multigene, where many genes contribute to the final phenotype, which indeed is probably influenced by complex regulatory elements and epigenetic mechanisms.
http://www.nature.com/news/2008/081105/full/456018a.html

The human genome – Conclusions (23/23)

Mutational landscape and significance across 12 major cancer types

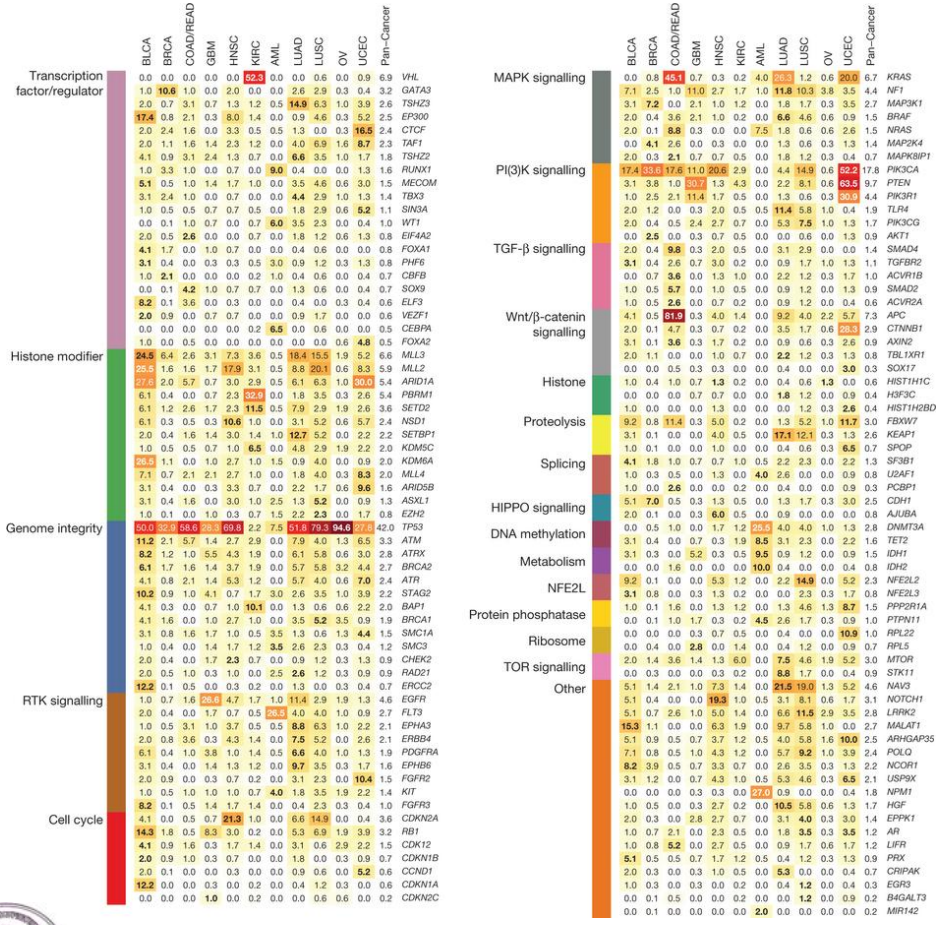


Figure 3: The 127 significantly mutated genes (SMGs) from 20 cellular processes in cancer identified in 12 cancer types (from 3281 tumors).
Nature 502, 333–339 (17 October 2013)
http://www.nature.com/nature/journal/v502/n7471/full/nature12634.html, CC-BY-NC-SA-3.0, http://creativecommons.org/licenses/by-nc-sa/3.0/

http://cancergenome.nih.gov/

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

accompanied with their hyperlinks.