



ΕΘΝΙΚΟ ΜΕΣΤΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Προσομοίωση Φυσιολογικών Συστημάτων

Εργαστηριακή Αναφορά 6

Θέμα: Βιοπληροφορική – Υπολογιστική Βιολογία

Στοιχεία Φοιτητή: Ονοματεπώνυμο: Αναστάσιος Παπαζαφειρόπουλος

Αριθμός Μητρώου: 03118079

Ακαδημαϊκό έτος: 2022-2023

3. Γονίδια και Μεταβολικές Αναθέσεις:

Ζητούμενο 3.1:

Γονίδιο INS: Το ανθρώπινο γονίδιο ινσουλίνης είναι ένα γονίδιο με παράγωγα πρωτεΐνης και αποτελείται από τρία εξόνια που κωδικοποιούν τις πολυπεπτιδικές αλυσίδες A, B, C. Πιο συγκεκριμένα, το εξόνιο 2 κωδικοποιεί την αλυσίδα B και μέρος της C, ενώ το εξόνιο 3 κωδικοποιεί την υπόλοιπη C και την A. Οι A και B αν ενωθούν μέσω της C συντελούν το μακρομόριο της προΐνσουλίνης. Η ινσουλίνη δημιουργείται από την ενζυμική αφαίρεση της C και την ένωση των A, B με δισουλφιδικούς δεσμούς. Στις δύο παρακάτω εικόνες παρουσιάζονται οι πληροφορίες του INS γονιδίου από τις δοθείσες διαδικτυακές πηγές <https://www.genenames.org/> και <https://omim.org/> αντίστοιχα:

Symbol report for INS ?

Report	HCOP homology predictions
HGNC data for INS	
Approved symbol ?	INS
Approved name ?	Insulin
Locus type ?	gene with protein product
HGNC ID ?	HGNC:6081
Symbol status ?	Approved
Previous symbols ?	IDDM2; IDDM1
Previous names ?	* Insulin-dependent diabetes mellitus 2 *
Chromosomal location ?	11p15.5
Gene groups ?	Neuropeptides
Gene resources for INS ?	
Ensembl	ENSG00000254647 gr Curated
	Ensembl region in detail gr
	Ensembl gene sequence gr
UC BC	uc001lvo.2 gr
NCBI Gene	3630 gr Curated
Alliance of Genome Resources	HGNC:6081 gr
Nucleotide resources for INS ?	
MANE Select	NM_000207.3 gr
	ENST00000381330.5 gr
RefSeq	NM_000207 gr Curated
	NCBI sequence viewer gr
INSDC	X70508 gr Curated
	ENA gr , GenBank gr , DDBJ gr
CCDS	CCDS7729 gr Curated
Protein resources for INS ?	
UniProt/Swiss-Prot	P01308 gr
	InterPro gr , PDBa gr , Reactome gr
AlphaFold	AF-P01308-F1 gr
Orthologs from selected species for INS ?	
Bos taurus	INS (VGNC:56268 gr) VGNC
Equus caballus	INS (VGNC:69316 gr) VGNC
Macaca mulatta	INS (VGNC:101288 gr) VGNC
Pan troglodytes	INS (VGNC:1070 gr) VGNC
Buc sorora	INS (VGNC:99778 gr) VGNC
Canis familiaris	INS (VGNC:97204 gr) VGNC
Felis catus	INS (VGNC:82037 gr) VGNC
Mus musculus	Ins2 (MGI:96573 gr) Curated
Rattus norvegicus	Ins2 (RGD:2916 gr)
Specialist resources for INS ?	
IUPHAR/BPS Guide to PHARMACOLOGY	6145 gr
Clinical resources for INS ?	
OMIM	176730 gr
MedlinePlus	Search via INS gr
ClinGen	Search via HGNC:6081 gr
ClinVar	Search via NCBI Gene ID 3630 gr
Orphanet	168345 gr
GenCC	HGNC:6081 gr
DECIPHER	Search via INS gr
Genetic Testing Registry	Search via NCBI Gene ID 3630 gr
dbVar	Search via NCBI Gene ID 3630 gr
Other resources for INS ?	
AmiGO	Search via P01308 gr
BioGPS	Search via NCBI Gene ID 3630 gr
Monarch	Search via HGNC:6081 gr
QuokkaGO	Search via P01308 gr
GeneCards	Search via HGNC:6081 gr
WikiGenes	Search via NCBI Gene ID 3630 gr
References for INS ?	
Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. Bennett ST et al. Nat Genet 1995 Mar;9(3):284-292 PMID: 7773291 Europe PMC gr , Pubmed gr gr	
Sequence of the human insulin gene. Bell GI et al. Nature 1980 Mar;284(5751):26-32 PMID: 6243748 Europe PMC gr , Pubmed gr gr	

Εικόνα 1: Πληροφορίες για το γονίδιο INS από την πηγή:

https://www.genenames.org/data/gene-symbol-report/#!/hgnc_id/HGNC:6081

*176730

Table of Contents

Title

Gene-Phenotype Relationships

Text

Description

Gene Structure

Mapping

Gene Function

Biochemical Features

Molecular Genetics

Animal Model

History

Allelic Variants

Table View

See Also

References

Contributors

Creation Date

Edit History

INSULIN; **INS**

Alternative titles: symbols

PROINSULIN

Other entities represented in this entry:

INS-IGF2 SPLICED READ-THROUGH TRANSCRIPTS, INCLUDED INSIGF, INCLUDED

HGNC Approved Gene Symbol: **INS**

Cytogenetic location: **11p15.5** Genomic coordinates (GRCh38): **11:2,159,779-2,161,209** (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	View Clinical Synopsis	Phenotype MIM number	Inheritance	Phenotype mapping key
11p15.5	Diabetes mellitus, insulin-dependent, 2		125852	AD	3
	Diabetes mellitus, permanent neonatal 4		618858	AD, AR	3
	Hyperproinsulinemia		616214	AD	3
	Maturity-onset diabetes of the young, type 10		613370	AD	3

PheneGene Graphics -

TEXT

Description

Insulin, synthesized by the beta cells of the islets of Langerhans, consists of 2 dissimilar polypeptide chains, A and B, which are linked by 2 disulfide bonds. However, unlike many other proteins, e.g., hemoglobin, made up of structurally distinct subunits, insulin is under the control of a single genetic locus; chains A and B are derived from a 1-chain precursor, proinsulin, which was discovered by [Steiner and Oyer \(1967\)](#). Proinsulin is converted to insulin by the enzymatic removal of a segment that connects the amino end of the A chain to the carboxyl end of the B chain. This segment is called the C (for 'connecting') peptide. +

Gene Structure

The human insulin gene contains 3 exons; exon 2 encodes the signal peptide, the B chain, and part of the C-peptide, while exon 3 encodes the remainder of the C-peptide and the A chain ([Steiner and Oyer, 1967](#)). +

The rat, mouse, and at least 3 fish species have 2 insulin genes ([Lomedico et al., 1979](#)). The single human insulin gene corresponds to rat gene II; each has 2 introns at corresponding positions. [Deltour et al. \(1993\)](#) showed that in the mouse embryo the 2 proinsulin genes are regulated independently, at least in part. The existence of a single insulin gene in man is supported by the findings in patients with mutations. The greatest variation among species is in the C-peptide. Receptor binding parts have been highly conserved. Some of these sites are involved with insulin-like activity, some with growth-factor activity, and some with both. +

INS-IGF2 Spliced Read-Through Transcripts

By EST database analysis and RT-PCR, [Monk et al. \(2006\)](#) identified 2 read-through transcripts, which they called the INSIGF long and short isoforms, that contain exons from both the **INS** gene and the downstream IGF2 gene (147470). The INSIGF short isoform contains **INS** exons 1 and 2 fused

External Links

Genome

DNA

Protein

Gene Info

Clinical Resources

Variation

ClinVar

gnomAD

CWAS Catalog

CWAS Central

HCMD

NHLBI EVS

Pharm gKB

Animal Models

Cellular Pathways

Εικόνα 2: Πληροφορίες για το γονίδιο INS από την πηγή: <https://omim.org/entry/176730?search=ins&highlight=ins>

Γονίδιο CLOCK:

Το όνομα του συγκεκριμένο γονιδίου σημαίνει: Circadian Locomotor Output Cycles Kaput και είναι υπεύθυνο για την ρύθμιση του κιρκάδιου μηχανισμού σε μοριακό επίπεδο. Η ρύθμιση αυτή πραγματοποιείται μέσω θετικών και αρνητικών επιδράσεων σε μεταγραφικό και μεταφραστικό επίπεδο. Πιο συγκεκριμένα, το γονίδιο CLOCK κωδικοποιεί μια πρωτεΐνη που είναι μέλος της οικογένειας των μεταγραφικών παραγόντων βασικής έλικας-αγκύλης-έλικας (bHLH)-

PAS. Η δομή bHLH διευκολύνει τη DNA-σύνδεση, αλλά και τον πρωτεϊνικό διμερισμό, δρώντας έτσι ως θετική συνιστώσα, η οποία κατευθύνει τη μεταγραφή άλλων γονιδίων του κιρκάδιου ρυθμού. Στις δύο παρακάτω εικόνες παρουσιάζονται οι πληροφορίες του CLOCK γονιδίου από τις δοθείσες διαδικτυακές πηγές <https://www.genenames.org/> και <https://omim.org/> αντίστοιχα:

Symbol report for CLOCK

Report HCOF homology predictions

HGNC data for CLOCK

Approved symbol: CLOCK
 Approved name: clock circadian regulator
 Locus type: gene with protein product
 HGNC ID: HGNC:2082
 Symbol status: Approved
 Previous names: "clock (mouse) homolog", "clock homolog (mouse)"
 Alias symbols: KIAA0334, KAT13D, bHLHa8
 Alias names: "Circadian locomotor output cycles protein kaput"
 Chromosomal location: 4q12
 Gene groups: Lysine acetyltransferases, Basic helix-loop-helix proteins, PAS domain containing

Gene resources for CLOCK

Ensembl: ENSG00000134852, Ensembl region in detail, Ensembl gene sequence
 UCSC: uc003hba.3
 NCBI Gene: 9575
 Alliance of Genome Resources: HGNC:2082

Nucleotide resources for CLOCK

MANE Select: NM_004898.4, ENST00000513440.6
 RefSeq: NM_004898, NCBI sequence viewer
 INSDC: AF011568, ENA, GenBank, DDBJ
 CCDB: CCDB33500

Protein resources for CLOCK

UniProt Swiss-Prot: O15516, InterPro, PDB, Reactome
 AlphaFold: AF-O15516-F1

Orthologs from selected species for CLOCK

Bos taurus: CLOCK (VGNC:27456)
 Equus caballus: CLOCK (VGNC:16627)
 Macaca mulatta: CLOCK (VGNC:71262)
 Pan troglodytes: CLOCK (VGNC:2027)
 Sus scrofa: CLOCK (VGNC:86774)
 Canis familiaris: CLOCK (VGNC:39353)
 Felis catus: CLOCK (VGNC:60970)
 Mus musculus: Clock (MGI:9698)
 Rattus norvegicus: Clock (RGD:620271)

Specialist resources for CLOCK

IUPHAR/EP3 Guide to PHARMACOLOGY: 2549

Clinical resources for CLOCK

OMIM: 601851
 DECIPHER: Search via CLOCK
 Genetic Testing Registry: Search via NCBI Gene ID 9575
 dbVar: Search via NCBI Gene ID 9575
 MedlinePlus: Search via CLOCK
 ClinGen: Search via HGNC:2082
 ClinVar: Search via NCBI Gene ID 9575

Other resources for CLOCK

AmiGO: Search via O15516
 BioGRID: Search via NCBI Gene ID 9575
 Monarch: Search via HGNC:2082
 Qeios: Search via O15516
 GeneCards: Search via HGNC:2082
 WikiGenes: Search via NCBI Gene ID 9575

References for CLOCK

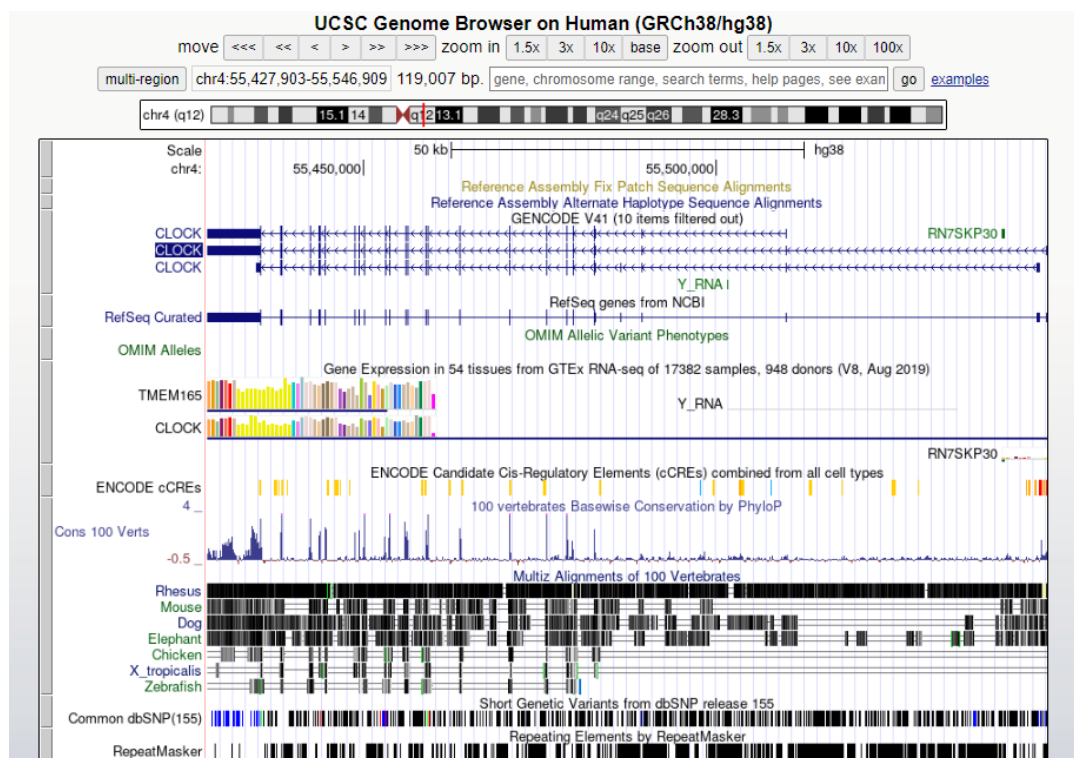
Molecular cloning and characterization of the human CLOCK gene: expression in the suprachiasmatic nuclei.
 Steeves TD et al. Genomics 1999 Apr;57(2):189-200
 PMID: 10198158 Europe PMC, Pubmed

Εικόνα 3: Πληροφορίες για το γονίδιο CLOCK από την πηγή: https://www.genenames.org/data/gene-symbol-report/#!/hgnc_id/HGNC:2082

Εικόνα 5: Επιβεβαίωση σχέσης Σακχαρώδους Διαβήτη Τύπου 2 και INS, πηγή:
<https://omim.org/entry/176730?search=diabetes%20%20ins&highlight=%28diabet e%7Cdiabetic%29%20%20ins>

Ζητούμενο 3.2:

Όπως φαίνεται και στην παρακάτω εικόνα, το γονίδιο CLOCK βρίσκεται στην περιοχή 4q12 του ανθρώπινου χρωμοσώματος 4 και πιο συγκεκριμένα ξεκινάει στο ζεύγος βάσης χρωμοσώματος 55.4427.903 και τελειώνει στο ζεύγος 55.546.909.



Εικόνα 6: Πληροφορίες για το γονίδιο CLOCK από την πηγή:
http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr4%3A55427903%2D55546909&hgid=1525670591_tO8nCCduiHvAlkAE74hz0cZ94ZkG

Ζητούμενο 3.3:

Στις παρακάτω εικόνες παρουσιάζονται κατά σειρά οι αλληλουχίες DNA του γονιδίου INS στον άνθρωπο, στον χιμπατζή και στον ποντικό:

```
gggcctcagctggggctgctgtcctaaggcaggggtgggaactaggcagccagcagggagg 60
ggacccctccctcactccctcctccacccccaccaccttgggccatccatggcggcat 120
cttgggccatccgggactggggacaggggtcctggggacaggggtgtggggacaggggtc 180
ctggggacaggggtctggggacaggggtcctggggacaggggtgtggggacaggggtgtg 240
gggacaggggtgtggggacaggggtcctggggacaggggtctggggacaggggtctgagg 300
acaggggtgtggggacaggggtgtggggacaggggtgtggggacaggggtgtggggacag 360
gggtctggggacaggggtccggggacaggggtgtggggacaggggtgtggggacaggggt 420
tgtggggacaggggtctggggacaggggtgtggggacaggggtcctggggacaggggtgt 480
ggggataggggtgtggggacaggggtgtggggacaggggtgtggggacaggggtctgggg 540
acagcagcgcgaagagccccgccctgcagcctccagctctcctggtctaagtggaaagt 600
ggccccaggtgagggctttgtctcctggagacatttgccccagctgtgagcagggacag 660
gtctggccacggggcccttggttaagactctaatgaccgctggtcctgaggaagaggtg 720
ctgacgaccaaggagatcttccacagaccagcaccagggaaatggtccggaaattgca 780
gcctcagccccagccatctgccgacccccccacccagggcctaattggggcagggcgga 840
gggggtgagaggtagggagatgggctctgagactataaagccagcggggggccagcagc 900
cctagccctccaggacagggctgcatcagaagaggccatcaagcaggtctgttccaaggg 960
cctttgcgtcaggtgggctcaggattccaggggtggctggacccagggccccagctctgca 1020
gcagggaggacgtggctgggctcgtgaagcatgtgggggtgagccagggggccccaaagg 1080
aggcacctggccttcagcctgcctcagcctgcctgtctccagatcactgtccttctg 1140
ccATGGCCCTGTGGATGCGCCTCTGCCCTGTGCGCTGTGCGCCTCTGGGGACCTG 1200
ACCCAGCCGACGCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAGCTCTCT 1260
ACCTAGTGTGCGGGGAACGAGGCTTCTTCTACACACCAAGACCCGCCGGGAGGCAGAGG 1320
ACCTCAGGGtgagccaactgccatttgcctggcctggcggccccagccacccctgct 1380
cctggcgtccacaccagcatgggcagaaggggcagggaggtgccaccacagcaggggt 1440
caggtgcaacttttttaaaaagaagtctcttggtcacgtctaaaagtgaccagctccct 1500
gtggccagtcagaaatctcagcctgaggacgggtgttggcttcggcagccccagatacat 1560
cagaggggtgggcacgctcctccctcactcgcctcaaacaaatgccccgcagccatt 1620
tctccaccctcatttgatgaccgcagattcaagtgtttgttaagtaaagtcctgggtga 1680
cctggggtcacagggtgccccacgctgcctgcctctgggcgaacccccatcacgcccgg 1740
aggagggcgtggctgcctgcctgagtgggccagaccctgtcgccagggcctcacggcagc 1800
tccatagtcaggagatggggaagatgctggggacagggccctggggagaagtactgggatc 1860
acctgttcaggctccactgtgacgctgccccggggcgggggaaggaggtgggacatgtg 1920
ggcgttggggcctgtaggctccacaccaggtgtgggtgacctccctctaacctgggtcca 1980
gccccgctggagatgggtgggagtgcgacctagggtggcgggcagggggcactgtgtc 2040
tccctgactgtgtcctcctgtgtccctctgcctcgccgctgttcgggaacctgctctg 2100
cggcacgtcctggcagTGGGGCAGGTGGAGCTGGGGCGGGGCCCTGGTGACGGCAGCCTG 2160
CAGCCCTTGGCCCTGGAGGGGTCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTAC 2220
AGCATCTGCTCCCTTACCAGCTGGAGAACTACTGCAACTagacgcagcccgaggcagc 2280
cccacacccgcgcctcctgcaccgagagagatggaataaagcccttgaaaccagccctgc 2340
tgtgcgtctgtgtgtcttggggccctgggccaagccccacttcccggcactgttgtga 2400
gccccctcccagctctctccacgctctctgggtgccacaggtgccaaagccggccagggc 2460
cagcatgcagtggctctcccaaaagcggccatgcctgtcggtgctgctgctgccccaccc 2520
tgtggtcaggggtccagtatgggagctgcgggggtctctgagggggcaggggtggtggg 2580
ccactgagaagtgacttcttgttcagtagctctggactcttggagtcccagagaccttg 2640
ttcaggaaggggaatgagaacattccagcaatttccccccacctagccctcccaggttc 2700
tattttagagttatttctgatggagtccctgtggaggaggaggtgggctgagggagg 2760
gggt 2820
```

Εικόνα 7: Η αλληλουχία του DNA του γονιδίου INS στον άνθρωπο.


```

aggtgctgttctgtgggagctgggagggccggaggggtgtacccaggggtcagcccag 60
atgacactatgggggtgatggtgtcgtgggacctggccaggagaggggagatgggctccc 120
agaagaggagtaggggtgagaggggtgctggggggcccgagagctgggcccagtgca 180
cagcttcccacacctgccacccccagagtcctgccgccacccccagatcacacggaaga 240
tgagggtccgagtgccctgctgaggacttgtgtcttgtccccgggtccccgggtcatgcc 300
tccttctgccacctcgggagctgagggccacagctgggggtgctgtcctacggcgggg 360
gggaactgggcagccagcaggggaggggacccctccctcactccactgtaccacccccac 420
caccttggcccatctatggcggtatcttggggcatcagggactggggacaggggtcctgg 480
ggacaggggtcctggggacaggggtcctggggacaggggtcctggggacaggggtcctgggg 540
acaggggtcctggggacaggggtcctggggacaggggtcctgggaacaggggtcctggggga 600
caggggtcctggggacaggggtcctggggacaggggtcctggggacaggggtcctggggaca 660
gggggtcctggggacaggggtcctggggacaggggtcctggggacaggggtcctggggacagg 720
gggtcctggggacaggggtcctggggacaggggtcctggggacaggggtcctggggacagcggt 780
gcaaaagagccccgcctgcagcctccagctcctcgtctaatgtggaagtgcccag 840
tgacggctttgtctcctcgtgagacatttgcctccagctgtgagcagggacaggtctggcc 900
accggggccccgtggttaagactctaatagcccgctggccctaaggaagaggtgctgacgac 960
caaggagatcttcccacagaccagcaccagggaatggtccggaaattgcagcctcagc 1020
ccccagccatctgccgacccccccacccagccctaattgggcagggcggcaggggttga 1080
caggtaggggagatgggtcctgagactataaagccagcggggggccagcagccctcagcc 1140
ctccaggacaggctgcacagaaagggccatcaagcaggtctgttccaaggcccttgcg 1200
tcagggtgggtcaggggtccagggtggctggacccagggccccagctctgcagcagggag 1260
gacgtggctgggtccttgaaagcatgtgggggtgagccccagggggccccagggcagccacc 1320
tgcccttcagcggcctcagccctgcctgtctccagatcactgtccttctgccATGGCC 1380
CTGTGGATGCGCCTCTGCCCTGCTGGTGTCTGTGGCCCTCTGGGGACCTGACCCAGCC 1440
TCGGCCTTTGTGAACCAACACCTGTGCGGCTCCACCTGGTGGAAAGCTCTCTACCTAGTG 1500
TGGGGGAACAGAGGCTTCTTACACACCCAAGACCAGCCGGGAGGCAGAGGACCTGCAG 1560
Ggtgagccaaccgcccgttgcctggccacccccagccacccccctgctcctggcgc 1620
tcccaccagcatgggcagaagggggcaggaggtgccacccagcaggggggtcaggtgca 1680
cttttttaaaaaaagaatgaagtctcttgggtcacatcctaaaaagtgaccagctcctgtg 1740
gcccagtcagaatctcagcctgagggacgggtgttggttcggcagccccgagatacatcag 1800
aggggtgggcacgctcctccctccactcggccctcaaacaaatgccccacagccatttct 1860
ccaccctcatttgatgaccgcagattcaagtgttttgttaagtaagtcctgggtgacct 1920
gggggtcacaggggtgccccacgctgcctgcctcctgggcgaacaccccatcacgccctgagg 1980
agggcgtggctgcctcctcgtgagtgggccagacccctgtcggcaggcctcacggcagctcc 2040
atagtcaggagatggggaagatgctggggacagggcctggggagaagtagtggggccacc 2100
tgttcaggctcccgtgtgacaccgccccggggcgggggaaggaggtaggacatgtgggc 2160
gttggggcctgtagggtccacacccagtggtgggtgacctccctctaacctgggtccagcc 2220
cggctggagatgggtgggagtgccagctagggtggtgggcagggcgggcactgtctctcc 2280
ctgactgtgtcctcctgtgtcctcctgcctcggcgtgttcgggaacctgtctgtgcgg 2340
cacgcctggcagTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAG 2400
CCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGTATCGTGGAAACAATGCTGTACCAGC 2460
ATCTGCTCCCTCTACCAGCTGGAGAATACTGCAACtagatggaataaagcccttgaacc 2520
agccctgctgtgccgtctgtgtgtcttgggggccctggggcaagccccacttcccggcac 2580
tgttgtgagccctcccagctctctccatgctctcgtgggtgccacaggtgccaatgccg 2640
ggcaggccccagcatgcagtggtctctcccaaaagcgccatgcctgtcggctgcctgtctac 2700
ccccaccctgtgggtcaggggtccagtatgggagctgcgggggtctctcaggggccagggg 2760
tggtgcagccactgagaatgaactcttggttcagtagctctgggaactcttgaggtcccca 2820
gagacctgttcaggaaaggggaatgagaacattccagcaatttccccccacctagccct 2880
cccaggttctattttagagttatctctgatggagtcctcgtggaggaggagggctgggc 2940
tgagggaggggggtcctgcagggcaggggggtgggaagggtggggagaggctgccgagagcc 3000

```

Εικόνα 8: Η αλληλουχία του DNA του γονιδίου INS στον χιμπατζή.

```

gaggggtgccgggaggggcacaaaggacacccccgggtccctgcagccccaggtctctgccg 60
actgctgcagaaaaacgtctctgaggagtcgggtggggccgtgctctcccgccaccctcgc 120
ccccagctgtgacagggacagctctgcagtcagggcgctcagggcctcgtaagacgctaa 180
tgaccgcgtggccccagcgagaggtgctgaccacggagagatgctcccgccccccgaag 240
cagggaaaatggctcggaaaactgcagcctcagcgccccccccccggccatctgccgacccc 300
cccaggccctaataatgggcccagggcgggcgggcgggggcagggaggtgggctcggggctata 360
aagccggcagcgccccggcagccccccagccctgcggaccagctgtttccccggcgctcagc 420
gagcaggtctgtgccaggggcctccggtccggggcggtgggacccgggaccccagctctgc 480
atgggtggtggtggggagggacgtgggctcctctcgtggggcatttgggggagcaagcggg 540
ggtcccgggggcagggcgccccgcccacgctggcctcagccccgctcctctcccaggtcttt 600
gcccccgccccgcATGGCCCTGTGGACGCGCCTCCTGCCCTGCTGGCCCTGCTGGCCC 660
TGTGGGGGCCCGAGCCCGCCCCCGCCTTCGTCAACCAGCACCTGTGCGGCTCCACCTGG 720
TGGAGGCCCTCTACCTGGTGTGCGGCGAGCGGGGCTTCTTCTACACGCCCAAGAGCGCC 780
GCGAGGTGGAGGACGCCCAGGgtgagcgccgcaggccccggggcagagggggcggggag 840
tgccaccgaaggaagggacctcttttggcctgccacgtcctgaaagcgccctgtggccc 900
ggccagagactctgggcttcaggacagtgcgcgctgcacgagcaggggccccgatggcacc 960
ctgaggccacgtggccctccctgcacctgccctccccaacaaacaccccagcccccg 1020
ccctacctcacaggacagcagcagcttccagggggacttttagtaaatgccaaaggccag 1080 Alu
gtgcggtgactcacgcctgtaatcccagcactctgggagggccaaggtgggcagactgctc 1140
gaggtcaggaggtttgaaaccagcctgagcaagagcgagaccccatctctactataaata 1200
aaaattaattggccaaactaatatataagaaaaaattagccgggcatggtggcacatgcc 1260
tgtagctcccagctactcgggagggctgaggcaggaggtatcgcttgagcccaggagatggag 1320
gttgcctgtgagctaggctgacgccacggcactcgctctagcctgggcaacaaagcgagac 1380
tctgtctcaaaaaaataaataaaaaaacaagcgccccaggctcctgggtgtgggtggtc 1440
tcagggtgaccttgggaagtggccccccacccgcccagtggggcgcggtagcaataggagg 1500
ctggcagtggggagtcgagggatgggcaaccccccggggacgggtgccctggtgggggcacc 1560
tgccacgttcccgcgcggcactgcccggggcggggacaggtggcaggggtgtgtgggct 1620
cggggcacgcctgtccctgccagcgccacccggggtagggggcgggcggtctctgtgtc 1680
cctgacgccttggcctgtctctctctctccctccctgtctgctggcacgcgcgcggcg 1740
gcagCGGGGCAGGTGGGGCCGGACGGCGGCCTGGGCGCGGGCGGCCTGCAGGCCCTGGCG 1800
CTGGAGGGGGCCCCGCAGAGCGCGGCATCGTGGAGCAGTGCTGCACCAGCATCTGCTCG 1860
CTGTACCAGCTGGAGAACTACTGCAACTagccaccgccccgccccgccccgggacggaa 1920
taaacctcttgaatggcccccggtgtctgtcttccgtgtgtctccgagccccgccccg 1980
cgccgcacccccctccccagcagcccccaaccccgagcgtcctccatgctccccggg 2040
ccggctaggccccgggtaccccaaggagcggggtgcctgccactgcccccccccggggtc 2100
tggggggcggtgggcagctgagatgccggggggctcttggggagctgacttccctcgttca 2160
gaagccctggaccctcaggggtccccagagaattttcagggaaagagaatgagaacattcc 2220
agtggcttcttcccaccta 2280

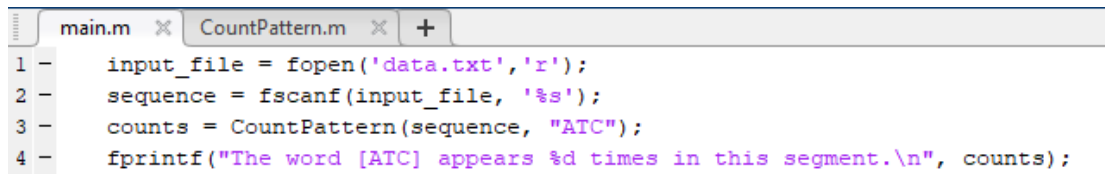
```

Εικόνα 9: Η αλληλουχία του DNA του γονιδίου INS στον ποντικό.

4. Εύρεση Θέσεων Αναπαραγωγής DNA:

Ζητούμενο 4.1:

Αρχικά σε ένα αρχείο data.txt βάλαμε το τμήμα γονιδιώματος που δίνεται στην εκφώνηση. Μετά εκτελέστηκε ο παρακάτω κώδικας των συναρτήσεων main.m και CountPattern.m στο περιβάλλον MATLAB:

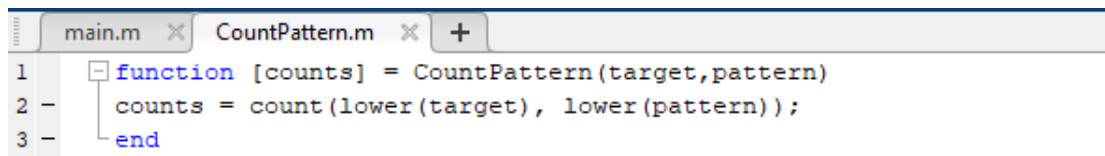


```

main.m  CountPattern.m  +
1 -   input_file = fopen('data.txt','r');
2 -   sequence = fscanf(input_file, '%s');
3 -   counts = CountPattern(sequence, "ATC");
4 -   fprintf("The word [ATC] appears %d times in this segment.\n", counts);

```

Εικόνα 10: Κώδικας Συνάρτησης main.m



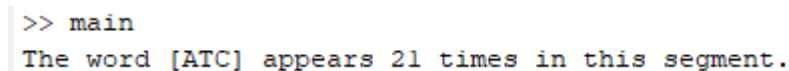
```

main.m  CountPattern.m  +
1 -   function [counts] = CountPattern(target,pattern)
2 -       counts = count(lower(target), lower(pattern));
3 -   end

```

Εικόνα 11: Κώδικας Συνάρτησης CountPattern.m

Και τυπώθηκε το αποτέλεσμα:



```

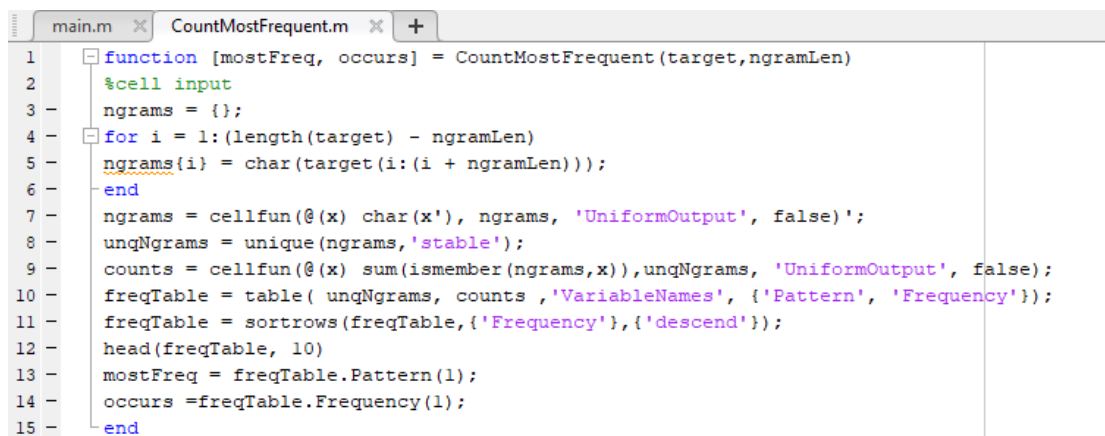
>> main
The word [ATC] appears 21 times in this segment.

```

Εικόνα 12: Αποτέλεσμα εκτέλεσης main.m

Ζητούμενο 4.2:

Τώρα, δημιουργήσαμε τη συνάρτηση CountMostFrequent.m, όπου δημιουργείται ένας πίνακας freqTable στον οποίο αποθηκεύονται η συχνότητα και το κάθε pattern της ακολουθίας. Έπειτα, στη main.m θέτουμε k=9 (όπως ζητείται από την εκφώνηση) και καλούμε τη συνάρτηση CountMostFrequent.m. Ο κώδικας των δύο συναρτήσεων καθώς και το αποτέλεσμα που τυπώνεται παρουσιάζονται παρακάτω:



```

main.m  CountMostFrequent.m  +
1 -   function [mostFreq, occurs] = CountMostFrequent(target,ngramLen)
2 -       %cell input
3 -       ngrams = {};
4 -       for i = 1:(length(target) - ngramLen)
5 -           ngrams{i} = char(target(i:(i + ngramLen)));
6 -       end
7 -       ngrams = cellfun(@(x) char(x'), ngrams, 'UniformOutput', false);
8 -       unqNgrams = unique(ngrams,'stable');
9 -       counts = cellfun(@(x) sum(ismember(ngrams,x)),unqNgrams, 'UniformOutput', false);
10 -      freqTable = table( unqNgrams, counts, 'VariableNames', {'Pattern', 'Frequency'});
11 -      freqTable = sortrows(freqTable,{'Frequency'},{'descend'});
12 -      head(freqTable, 10)
13 -      mostFreq = freqTable.Pattern(1);
14 -      occurs =freqTable.Frequency(1);
15 -   end

```

Εικόνα 12: Κώδικας Συνάρτησης CountMostFrequent.m

```
main.m x CountMostFrequent.m x +
1 - input_file = fopen('data.txt','r');
2 - sequence = fscanf(input_file, '%s');
3 - k = 9;
4 - [seq, freq] = CountMostFrequent(cellstr(sequence'), k);
5 - fprintf("The most frequent %d-digit sequence is [ %s ] and appears %d times in this segment.\n", k, char(seq), freq(1) );
6
```

Εικόνα 13: Κώδικας Συνάρτησης main.m

```
>> main

ans =

    10x2 table

    Pattern      Frequency
    _____
    'ctcttgatca'   [3]
    'tcttgatcat'   [3]
    'aagcatgata'   [2]
    'agcatgatca'   [2]
    'cttgatcata'   [2]
    'ttgatcatcg'   [2]
    'gctcttgatc'   [2]
    'atcaatgata'   [1]
    'tcaatgatca'   [1]
    'caatgatcaa'   [1]

The most frequent 9-digit sequence is [ ctcttgatca ] and appears 3 times in this segment.
```

Εικόνα 14: Αποτέλεσμα εκτέλεσης main.m

5. Ανάλυση Δεδομένων Πειράματος Γονιδιακής Έκφρασης:

Ζητούμενο 5.1:

Τα δείγματα που μετρήθηκαν στο πείραμα είναι 31. Οι διαφορετικές καταστάσεις του πειράματος είναι:

1. **to control group** (κατάσταση ελέγχου), που τοποθετήθηκαν 19 δείγματα (συνέχιση κανονικής διατροφής-χωρίς απώλεια βάρους)
2. **to intervention group** (κατάσταση παρέμβασης), που τοποθετήθηκαν 12 δείγματα (υποθερμική δίαιτα (AHA step I-βαθμιαία απώλεια βάρους).

Ο τύπος της πλατφόρμας μικροσυστοιχίας DNA με την οποία έγιναν οι μετρήσεις είναι:

GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array

Παρακάτω παρατίθεται και η σύνοψη του πειράματος:



Series GSE7117 [Query DataSets for GSE7117](#)

Status: Public on Jul 23, 2008

Title: Hepatic gene expression after a hypocaloric, low-fat diet in obese women and controls

Organism: [Homo sapiens](#)

Experiment type: Expression profiling by array

Summary: The prevalence of obesity has been increasing rapidly worldwide during the past two decades. This is alarming, since obesity has considerable effects on morbidity and mortality. The majority of gene expression studies about the effect of obesity and weight loss have been performed using the adipose tissue for mRNA extraction. However, also the liver plays a central role in maintaining energy balance. To our knowledge, no overall analysis of hepatic gene expression in response to changes in nutritional status has been made in humans. Therefore, it is important to investigate how a short-time hypocaloric diet affects overall hepatic gene expression and the metabolic profile in a group of overweight and obese women. The subjects (n=31) were middle-aged, overweight (BMI>25 kg/m²) women with gallstone disease scheduled for an elective gallbladder operation. The intervention subjects were placed on a hypocaloric AHA step I diet with a recommended daily energy intake of 5.0 MJ. The objective was to reduce 0.5 kg of body weight per week. The control subjects were instructed to continue their habitual diet and not to lose weight. Basic clinical measurements and laboratory analyses were performed twice at baseline and at two week intervals during the weight reduction period. Surgical liver biopsies were obtained at the end of the weight reduction period. RNA samples of 4 individuals from the intervention group and 4 individuals from the control group were selected for the microarray analysis. The results from the microarray analysis were fairly surprising. Only one gene was up-regulated and the rest 142 down-regulated in the diet intervention group compared to the control group when a minimum of 2-fold change was set as the limit. The global decrease in hepatic gene expression was unexpected but the results are interesting, with several genes not previously linked to weight reduction. The decrease in triglyceride and fasting plasma insulin concentrations observed in our study is in accordance with results from many previous weight-loss trials. Keywords: Overall hepatic gene expression associated with obesity and moderate weight loss in a group of overweight and obese middle-aged women.

Εικόνα 15: Σύνοψη του πειράματος με GEO accession id: GSE7117, πηγή: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi>

Ζητούμενο 5.2:

i. Οι τίτλοι των δειγμάτων είναι:

"Liver 1", "Liver 2", "Liver 3", "Liver 9", "Liver 12", "Liver 25",
"Liver 26", "MHLiv"

Τα αντίστοιχα Sample_geo_accession_id τους είναι:

"GSM162954", "GSM162956", "GSM162957", "GSM162958",
"GSM162959", "GSM162960", "GSM162961", "GSM162962"

- ii. Τα δείγματα προέρχονται από τον πανεπιστημιακό οδηγό της πόλης Oulu της Φινλανδίας.
- iii. Τα δείγματα: "Liver 1", "Liver 9", "Liver 25", "MHLiv" ανήκουν στην κατάσταση ελέγχου, ενώ τα δείγματα: "Liver 2", "Liver 3", "Liver 12", "Liver 26" στην κατάσταση παρέμβασης.
- iv. Τα ορίσματα της **read.table** είναι τα παρακάτω:
 - **file.path**: Το όνομα του αρχείου που θα χρησιμοποιήσει η συνάρτηση ("GSE7117_series_matrix.txt").
 - **skip**: Αγνοεί τις αρχικές 68 γραμμές στο αρχείο πριν ξεκινήσει να το διαβάσει.
 - **header**: Δηλώνει αν στο αρχείο υπάρχουν τα ονόματα των μεταβλητών όπως στην πρώτη γραμμή (True).
 - **sep**: Ο τρόπος που θα χωρίζονται τα στοιχεία του αρχείου σε κάθε γραμμή μεταξύ τους (sep="\t").
 - **row.names**: Ο αριθμός της στήλης που θα περιέχει τα ονόματα των σειρών (row.names=1).
- v. Με εκτέλεση της εντολής `dim(x)`, βλέπουμε ότι ο πίνακας `x` έχει 54675 γραμμές και 8 στήλες:

```
> dim(x)
[1] 54675      8
```

Εικόνα 16: Εκτέλεση εντολής `dim(x)`

- vi. Με την εκτέλεση της εντολής `colnames(x)` παίρνουμε:

```
> colnames(x)
[1] "GSM162954" "GSM162956" "GSM162957" "GSM162958" "GSM162959" "GSM162960" "GSM162961" "GSM162962"
```

Εικόνα 17: Εκτέλεση εντολής `colnames(x)`

Τα ονόματα των στηλών αντιστοιχούν στα 8 διαφορετικά βιολογικά δείγματα που πάρθηκαν προς ανάλυση με μικροσυστοιχίες DNA.

- vii. Με την εκτέλεση της εντολής `rownames(x) [1:15]`, προκύπτει:

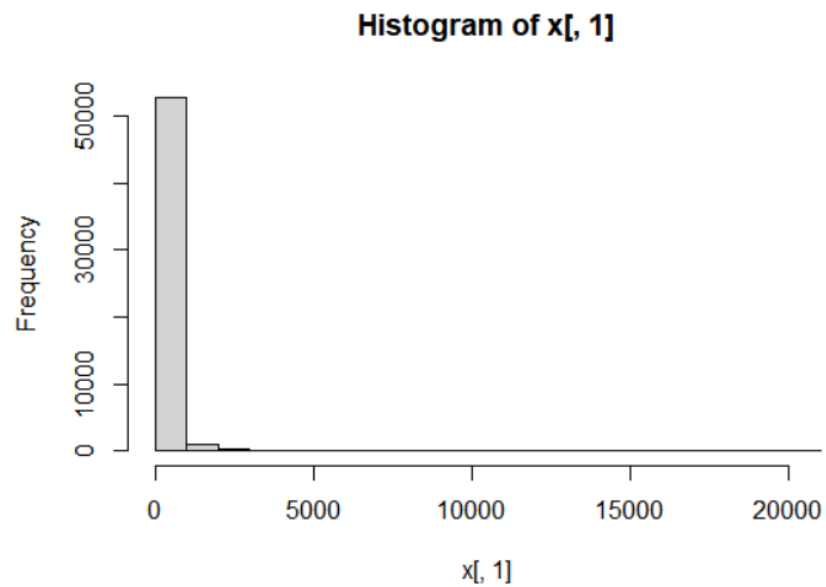
```
> rownames(x)[1:15]
[1] "1007_s_at" "1053_at" "117_at" "121_at" "1255_g_at" "1294_at" "1316_at" "1317_at"
[9] "1431_at" "1438_at" "1487_at" "1494_f_at" "1552256_a_at" "1552257_a_at" "1552258_a_at" "1552259_a_at"
```

Εικόνα 18: Εκτέλεση της εντολής `rownames(x) [1:15]`

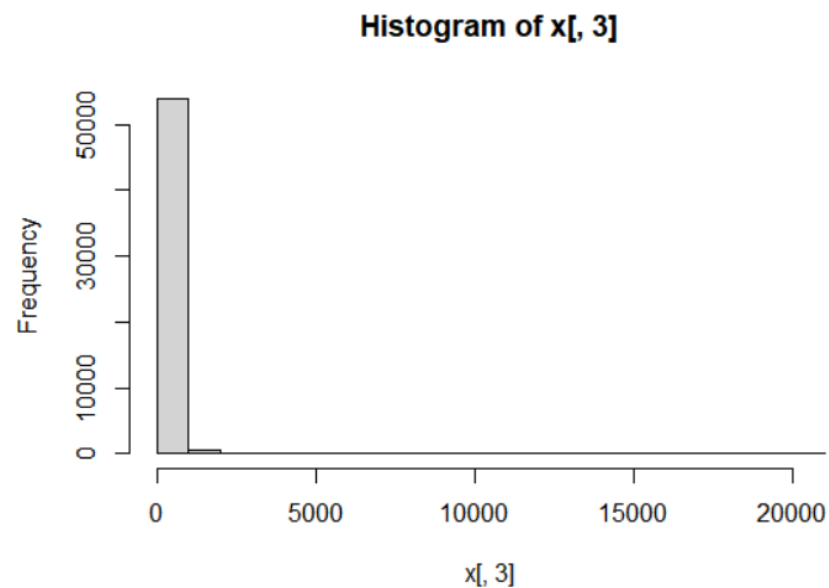
Οι γραμμές αντιστοιχούν σε διαφορετικά γονίδια.

- viii. Η τιμή του `x[3,5]` είναι: 42.91798 και περιγράφει το βαθμό και τη συγκέντρωση του 3^{ου} γονιδίου στο 5^ο δείγμα.

- ix. Η 200^η γραμμή μας δίνει τις συγκεντρώσεις του 200στού γονιδίου (1552540_s_at) και η 7^η στήλη τις τιμές συγκεντρώσεων των γονιδίων στο 7^ο δείγμα.
- x. Με την εκτέλεση των εντολών `hist(x[,1])` και `hist(x[,3])` παίρνουμε τα ακόλουθα αποτελέσματα:



Εικόνα 19: Ιστόγραμμα συχνοτήτων πρώτου δείγματος



Εικόνα 20: Ιστόγραμμα συχνοτήτων τρίτου δείγματος

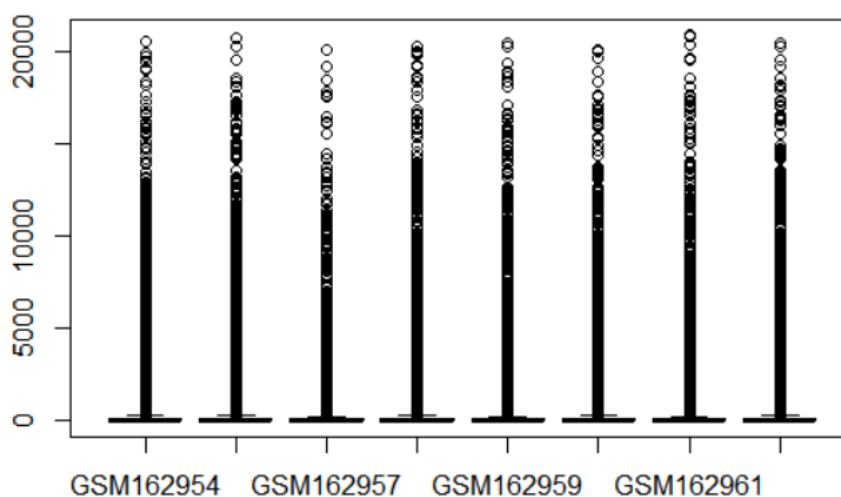
Ζητούμενο 5.3:

- i. Η εντολή `round` κάνει στρογγυλοποίηση στο διάνυσμα του πρώτου ορίσματος `apply(x, 2, summary)` κατά το δεύτερο όρισμα `digits=2`, δηλαδή στη συγκεκριμένη περίπτωση κατά δύο δεκαδικά ψηφία. Ειδικότερα, η `apply(x, 2, summary)` καλεί τη συνάρτηση `summary` και την εφαρμόζει στις στήλες του πίνακα `x` (υπολογίζοντας έτσι τα χαρακτηριστικά τους). Παρακάτω παρατίθεται και το αποτέλεσμα της εκτέλεσης:

```
> round(apply(x,2, summary),digits=2) # explain in your assignment the functionality/argument
code
      GSM162954 GSM162956 GSM162957 GSM162958 GSM162959 GSM162960 GSM162961 GSM162962
Min.          5.44      5.74      5.65      5.76      5.81      5.48      5.90      5.64
1st Qu.       16.63     16.33     17.55     16.24     17.54     16.18     17.14     16.49
Median        42.31     41.06     39.89     41.26     42.33     40.88     41.93     40.56
Mean         213.47     217.89     129.67     213.59     207.06     216.42     197.97     209.00
3rd Qu.      120.77     119.92     92.73     121.73     117.34     122.76     113.54     121.52
Max.        20596.82    20710.01    20130.98    20276.33    20518.67    20142.23    20892.18    20513.43
```

Εικόνα 21: Αποτέλεσμα εκτέλεσης εντολής: `round(apply(x,2, summary),digits=2)`

- ii. Με την εντολή `boxplot(x)` δημιουργούμε τα αντίστοιχα θηκογράμματα:



Εικόνα 22: Τα θηκογράμματα για κάθε δείγμα

Τα θηκογράμματα είναι κανονικοποιημένα, καθώς η διάμεσος των δειγμάτων βρίσκονται κοντά σε μια συγκεκριμένη τιμή.

Ζητούμενο 5.4:

- i. Η σωστή κλήση της συνάρτησης είναι: `c(0,1,1,0,1,0,1,0)`.
- ii. Παρατίθενται ο κώδικας που εκτελέστηκε και τα αποτελέσματα στις παρακάτω εικόνες:

```
p_values=matrix(0,nrow(x))
control = x[,c(1,4,6,8)]
intervent=x[,c(2,3,5,7)]
for(i in 1:nrow(x)){ p_values[i]<- t.test(control[i,],intervent[i,],alternative = c("two.sided"))
}
view(p_values[1:15])
```

Εικόνα 23: Κώδικας για t-test

```
> p_values[1:15]
[1] 0.1235861 0.2768388 0.6168126 0.1692107 0.174158
[7] 0.8137377 0.2075931 0.3956510 0.5966297 0.570401
[13] 0.7940872 0.7325208 0.9863738
```

Εικόνα 24: Τιμές p_value για τους 15 πρώτους ανιχνευτές

- iii. Παρατίθενται ο κώδικας και τα αποτελέσματα για τους ανιχνευτές με $p_value < 0.001$ και ο ανιχνευτής με το μικρότερο p_value .

```
> for(i in 1:nrow(x)){if (p_values[i]<0.001) print(row.names
[1] "1552912_a_at"
[1] "1552935_at"
[1] "1553222_at"
[1] "1555645_at"
[1] "1556049_at"
[1] "1562173_a_at"
[1] "200931_s_at"
[1] "204019_s_at"
[1] "204114_at"
[1] "204681_s_at"
[1] "204861_s_at"
[1] "205171_at"
[1] "208798_x_at"
[1] "212416_at"
[1] "214999_s_at"
[1] "215463_at"
[1] "215870_s_at"
[1] "221225_at"
[1] "224578_at"
[1] "225174_at"
[1] "227306_at"
[1] "227662_at"
[1] "229024_at"
[1] "231133_at"
[1] "232604_at"
[1] "232930_at"
[1] "234751_s_at"
[1] "234830_at"
[1] "235457_at"
[1] "235676_at"
[1] "236252_at"
[1] "239368_at"
[1] "239548_at"
[1] "240177_at"
[1] "243337_at"
[1] "243654_at"
[1] "243794_at"
[1] "36129_at"
.
```

Εικόνα 25: Κώδικας και ανιχνευτές με p_value <0.001

```
> row.names(x[which.min(p_values),])
[1] "243337_at"
.
```

Εικόνα 26: Κώδικας και ανιχνευτής με το μικρότερο p_value

Ζητούμενο 5.5:

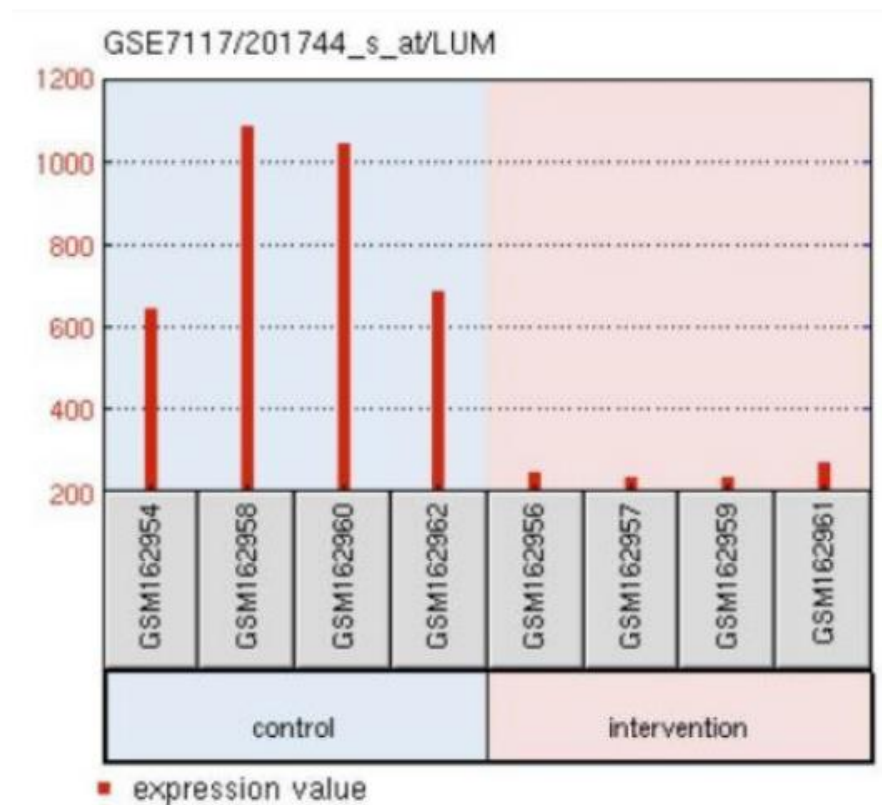
- i. Όλα αποτελέσματα για τα δύο πρώτα σημαντικά γονίδια απεικονίζονται στην ακόλουθη εικόνα:

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol
▶ 261744_s_at	0.4	0.00099731	-9.4	0.0462	-1.751	LUM
▶ 204019_s_at	0.544	0.0002013	-8.27	0.5363	-0.883	SH2YL1

Εικόνα 27: Όλα τα αποτελέσματα για τα δύο πρώτα σημαντικά γονίδια

Παρατηρούμε πως η τιμή p -value αυξάνεται σημαντικά από το πρώτο στο δεύτερο δείγμα. Η τιμή της $\log FC$ είναι κάτω από το μηδέν κάτι το οποίο δείχνει ότι τιμή για την κατάσταση παρέμβασης είναι μικρότερη από την τιμή για την κατάσταση ελέγχου.

Παρακάτω απεικονίζεται το γράφημα που αφορά τον πρώτο ανιχνευτή:



Εικόνα 28: Γράφημα για πρώτο ανιχνευτή

Παρατηρούμε πως η τιμή έκφρασης του γονιδίου είναι πολύ μεγαλύτερη στην κατάσταση ελέγχου.

- i. Το γονίδιο το οποίο εκφράζεται διαφορετικά σε στατιστικά σημαντικότερο βαθμό με το GEO2R είναι το γονίδιο Lumican. Συνοπτικές πληροφορίες του γονιδίου αυτού παρουσιάζονται στην παρακάτω εικόνα:

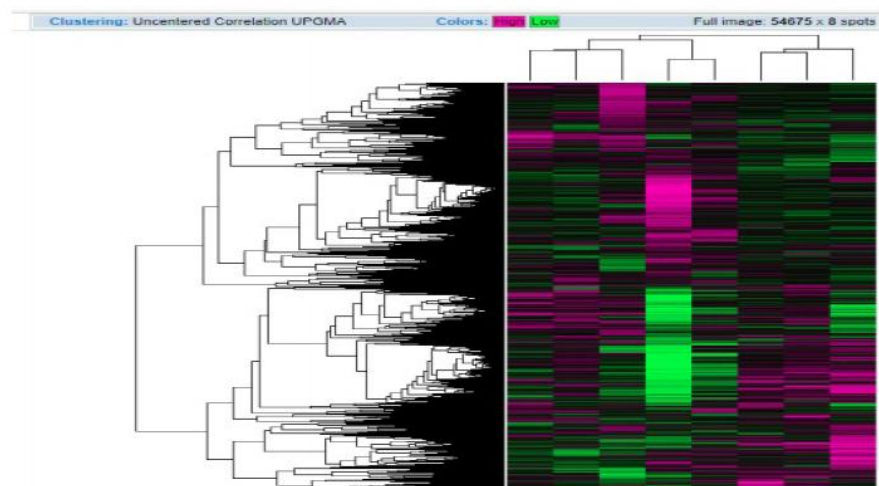
Official Symbol	LUM provided by HGNC
Official Full Name	lumican provided by HGNC
Primary source	HGNC:HGNC:6724
See related	Ensembl:ENSG00000139329 MIM:600616
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	LDC; SLRR20
Summary	This gene encodes a member of the small leucine-rich proteoglycan (SLRP) family that includes decorin, biglycan, fibromodulin, keratocan, epiphygan, and osteoglycin. In these bifunctional molecules, the protein moiety binds collagen fibrils and the highly charged hydrophilic glycosaminoglycans regulate interfibrillar spacings. Lumican is the major keratan sulfate proteoglycan of the cornea but is also distributed in interstitial collagenous matrices throughout the body. Lumican may regulate collagen fibril organization and circumferential growth, corneal transparency, and epithelial cell migration and tissue repair. [provided by RefSeq, Jul 2008]
Expression	Broad expression in gall bladder (RPKM 981.2), urinary bladder (RPKM 643.2) and 14 other tissues See more
Orthologs	mouse all

Εικόνα 29: Πληροφορίες του γονιδίου Lumican

Αξίζει να τονιστεί πως το γονίδιο αυτό ρυθμίζει τη διαφάνεια του κερατοειδούς, τη μετανάστευση των επιθηλιακών κυττάρων, την επισκευή των ιστών και την οργάνωση ινιδίων κολλαγόνου.

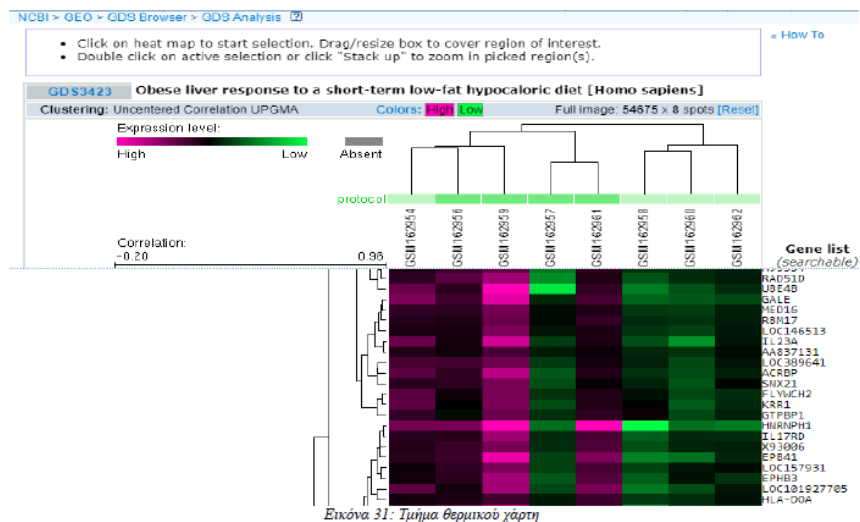
Ζητούμενο 5.6:

- Ο πλήρης θερμικός χάρτης που απεικονίζεται παρακάτω αποτελείται από 54.675 γραμμές και 8 στήλες:



Εικόνα 30: Πλήρης θερμικός χάρτης

- Παρουσιάζεται τμήμα της λεπτομέρειας του θερμικού χάρτη:

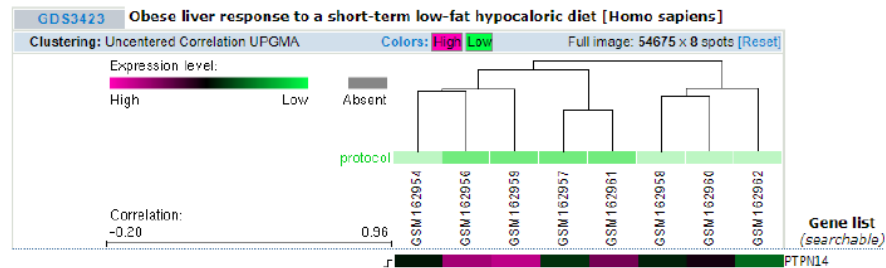


Εικόνα 31: Τμήμα θερμικού χάρτη

Εικόνα 31: Τμήμα της λεπτομέρειας του θερμικού χάρτη

Σε κάθε γραμμή είναι το κάθε γονίδιο και σε κάθε στήλη το κάθε δείγμα. Με διαφορετικά χρώματα παρουσιάζεται η ένταση έκφρασης των γονιδίων σε κάθε δείγμα. Ειδικότερα με μωβ χρώμα έχουμε ισχυρή ένταση έκφρασης ενώ με πράσινο χαμηλή. Είναι διακριτή επίσης η ομαδοποίηση του κάθε δείγματος.

- Παρουσιάζεται ο χρωματικός χάρτης του πρώτου γονιδίου:



Εικόνα 32: Θερμικός χάρτης του πρώτου γονιδίου

Με βάση τον χρωματικό κώδικα συμπεραίνουμε ότι η ένταση έκφρασης είναι μέτρια για το πρώτο γονίδιο.

- Σύμφωνα με την ομαδοποίηση που προέκυψε παρατηρούμε ότι υπάρχει διαφορά στην ένταση έκφρασης ανάμεσα στα γονίδια των δύο καταστάσεων (ελέγχου και παρέμβασης). Πιο συγκεκριμένα, στα γονίδια παρέμβασης υπάρχει πιο έντονη ένταση έκφρασης από τα γονίδια ελέγχου.