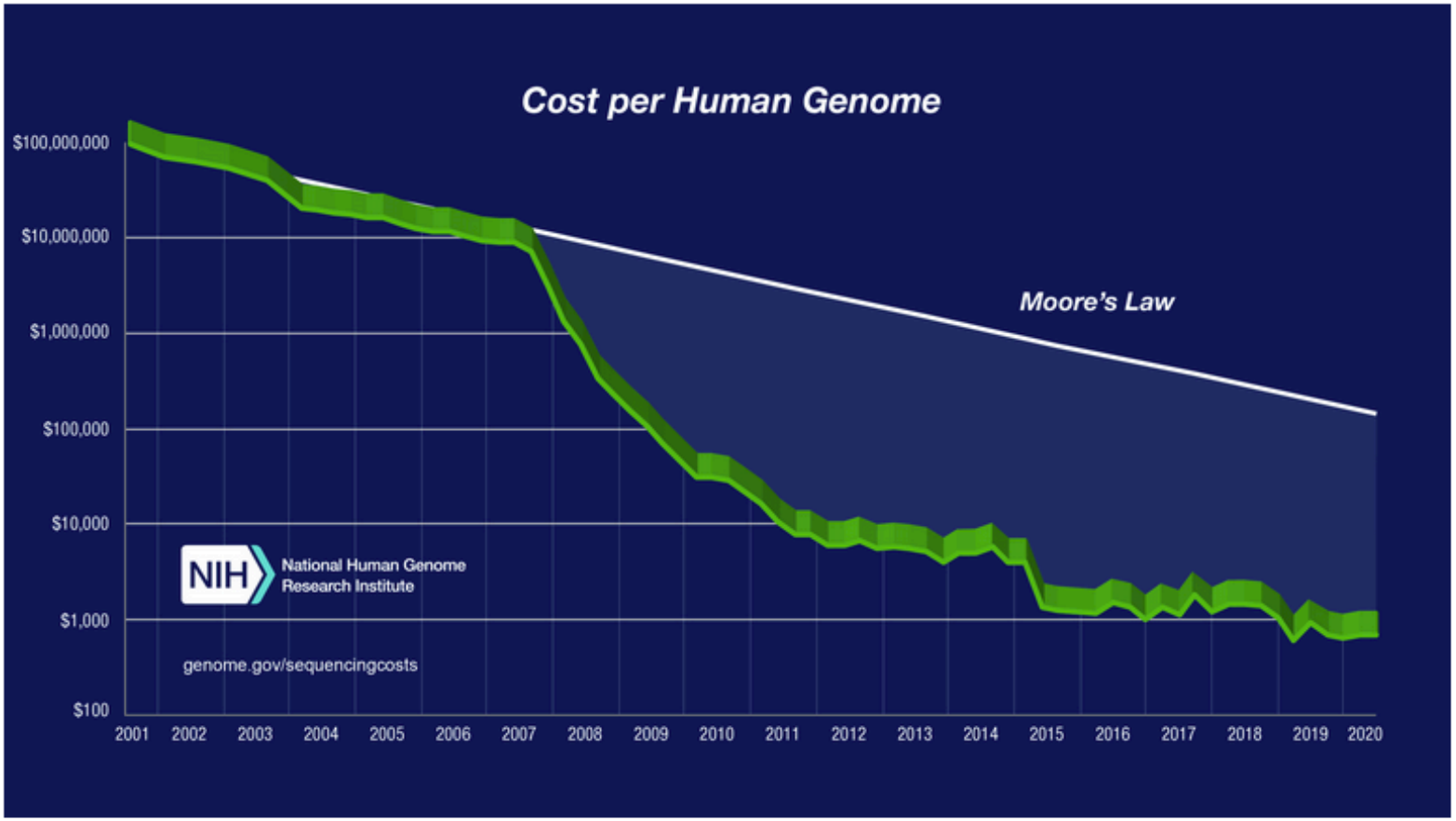
Βιοπληροφορική και Μοριακοί δείκτες ασθενειών

Γρηγόριος Αμούτζιας
Αν. Καθηγητής Βιοπληροφορικής με έμφαση στη Μικροβιολογία
Τμήμα Βιοχημείας και Βιοτεχνολογίας
Πανεπιστήμιο Θεσσαλίας

Next Generation Sequencing

Χαμηλό κόστος γενωμικών τεχνολογιών θα οδηγήσει σε καθημερινές εφαρμογές

- Κόστος αλληλούχισης
  - http://www.genome.gov/sequencingcosts/
- Ο νόμος του Moore προβλέπει διπλασιασμό της υπολογιστικής ισχύς κάθε δύο χρόνια.



Cost per genome data - 2020

Pacific Biosciences

Analytical Chemistry REVIEW

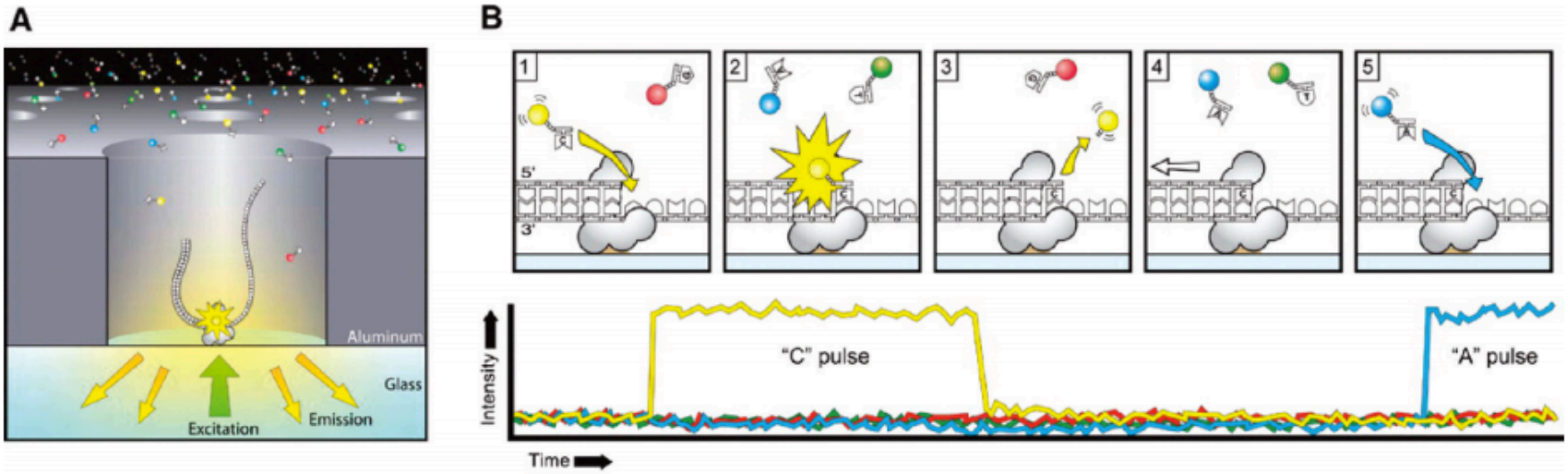
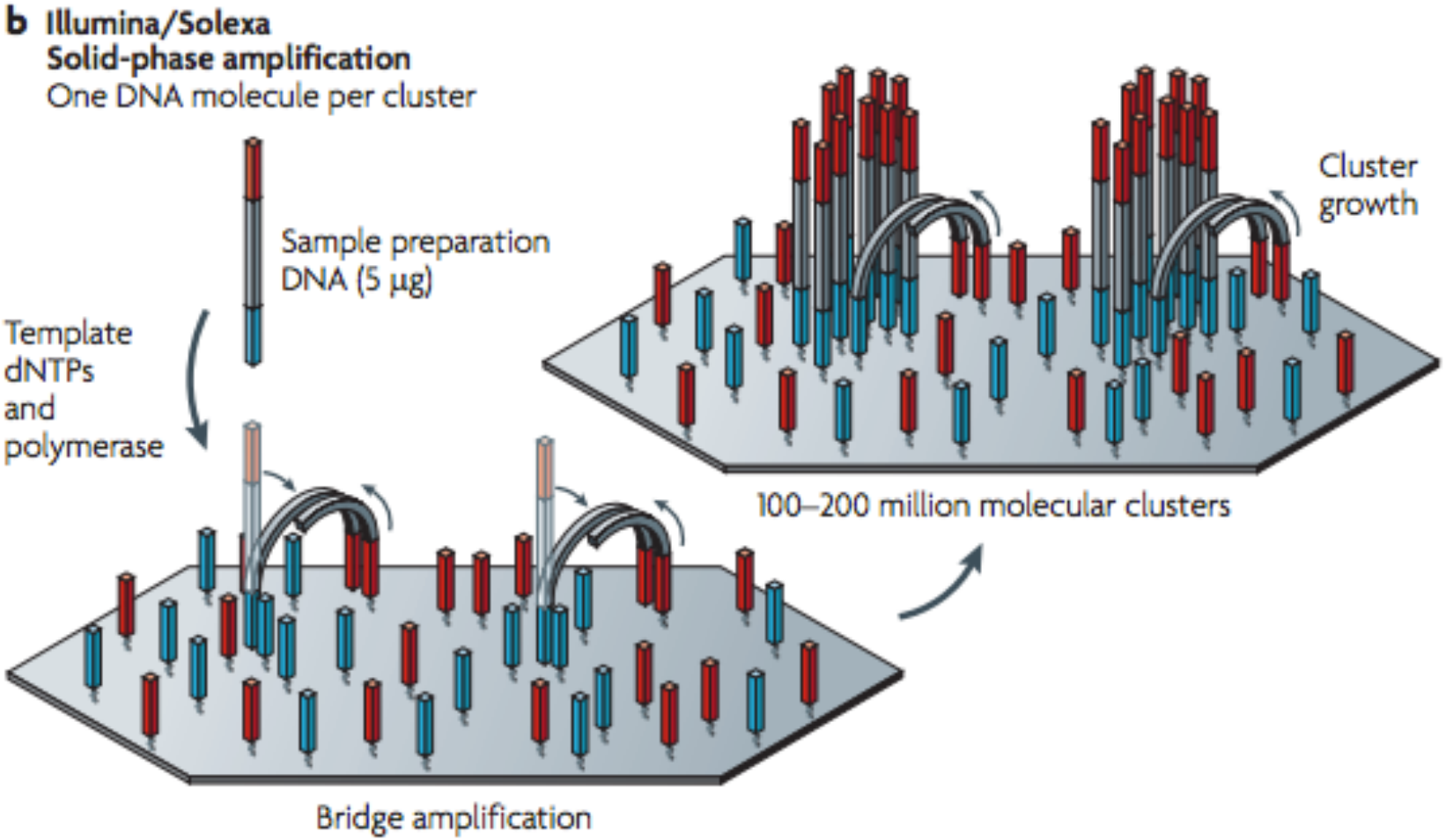


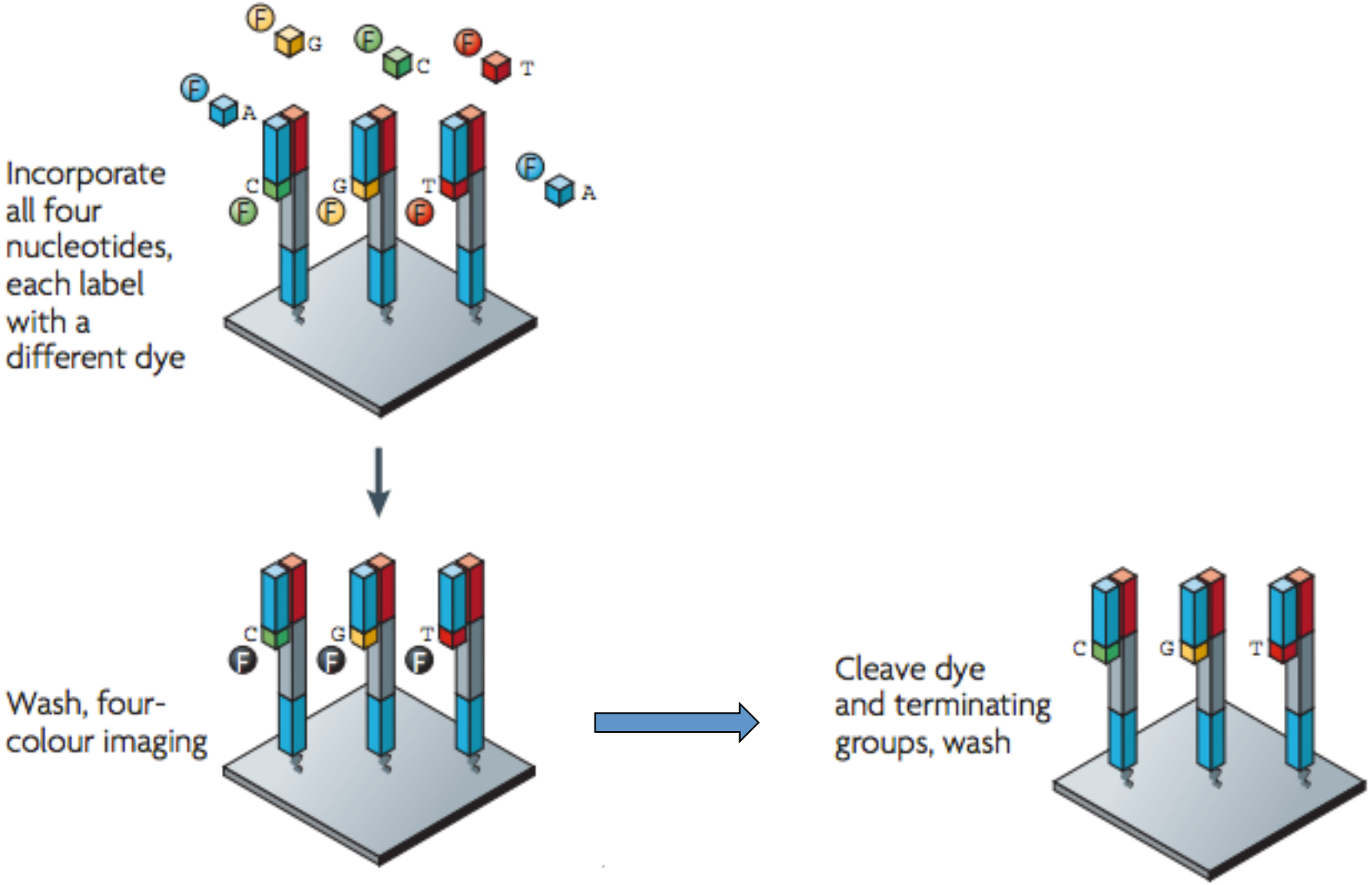
Figure 2. Schematic of PacBio's real-time single molecule sequencing. (A) The side view of a single ZMW nanostructure containing a single DNA polymerase (Φ29) bound to the bottom glass surface. The ZMW and the confocal imaging system allow fluorescence detection only at the bottom surface of each ZMW. (B) Representation of fluorescently labeled nucleotide substrate incorporation on to a sequencing template. The corresponding temporal fluorescence detection with respect to each of the five incorporation steps is shown below. Reprinted with permission from ref 39. Copyright 2009 American Association for the Advancement of Science.

http://www.ncbi.nlm.nih.gov/pubmed/21612267

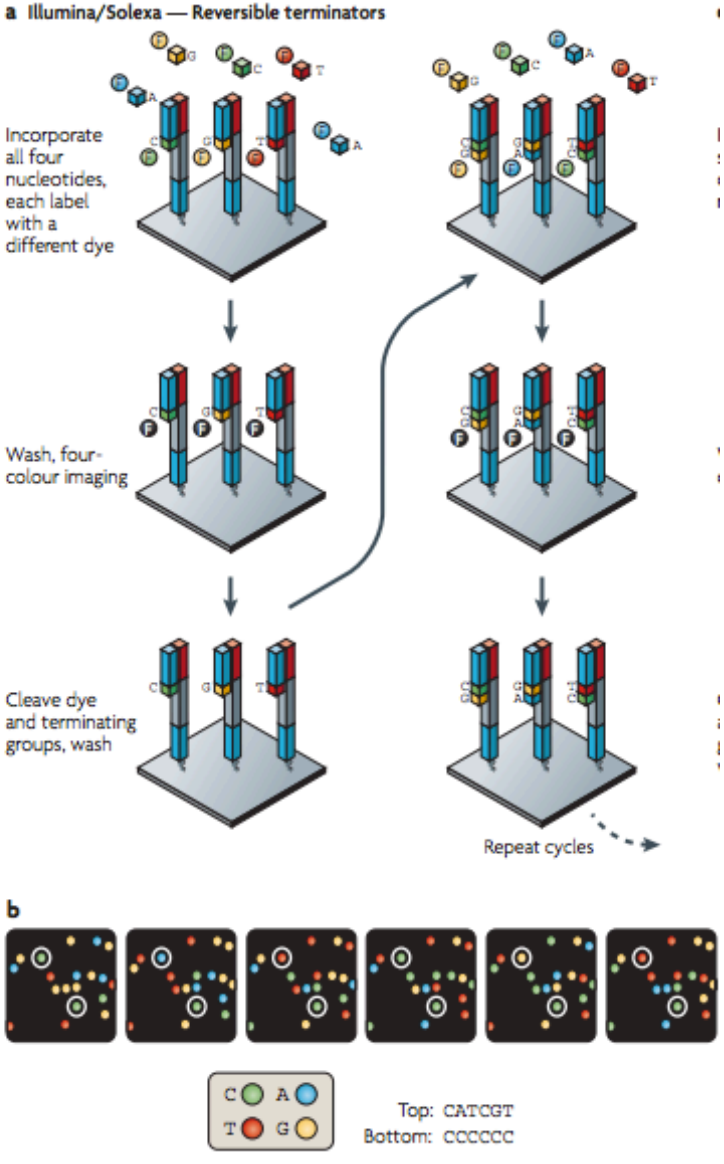
http://www.youtube.com/watch?v=NHCJ8PtYCFc
http://www.youtube.com/watch?v=GX6RSKh4J7E
SMRT techonology – real time single molecule sequencing



http://www.ncbi.nlm.nih.gov/pubmed/19997069



http://www.ncbi.nlm.nih.gov/pubmed/19997069



http://www.ncbi.nlm.nih.gov/pubmed/19997069

Ion Proton

http://www.lifetechnologies.com/global/en/home/about-us/news-gallery/press-releases/2012/life-techologies-itroduces-the-bechtop-io-proto.html

Press Releases

Life Technologies Introduces the Benchtop Ion Proton™ Sequencer; Designed to Decode a Human Genome in One Day for $1,000

SAN FRANCISCO, Jan. 10, 2012 /PRNewswire/ -- Life Technologies Corporation (NASDAQ: LIFE) today announced it is taking orders for the new benchtop Ion Proton™ Sequencer that is designed to sequence the entire human genome in a day for $1,000.

(Photo: http://photos.prnewswire.com/prnh/20120110/LA31914-a)

(Photo: http://photos.prnewswire.com/prnh/20120110/LA31914-b)

The Ion Proton™ Sequencer, priced at $149,000, is based on the next generation of semiconductor sequencing technology that has made its predecessor, the Ion Personal Genome Machine™ (PGM™), the fastest-selling sequencer in the world.

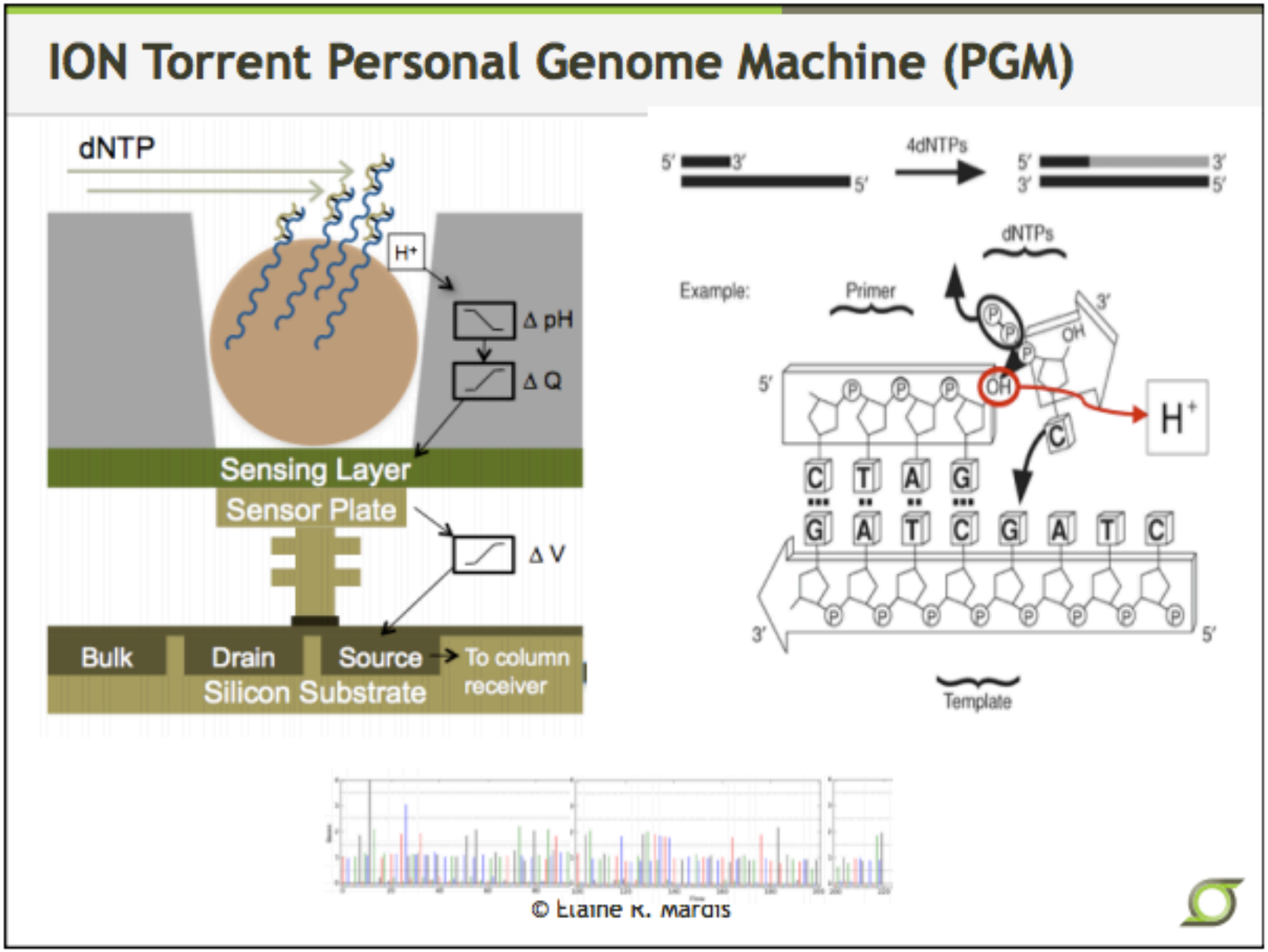
Up to now, it has taken weeks or months to sequence a human genome at a cost of $5,000 to $10,000 using optical-based sequencing technologies. The slow pace and the high instrument cost of $500,000 to $750,000 have limited human genome sequencing to relatively few research labs.

Ion Proton

Ion torrent chemistry

http://www.youtube.com/watch?v=yVf2295JqUg

Ουσιαστικά είναι ένα πολύ μικρό pH-meter
Δεν βασίζεται σε ανίχνευση φωτός!



Εικόνα Από Elaine Mardis

Oxford Nanopore

(Στο εγγύς μέλλον;)

Nanopore
http://www.youtube.com/watch?v=UWcCbIRPzvs

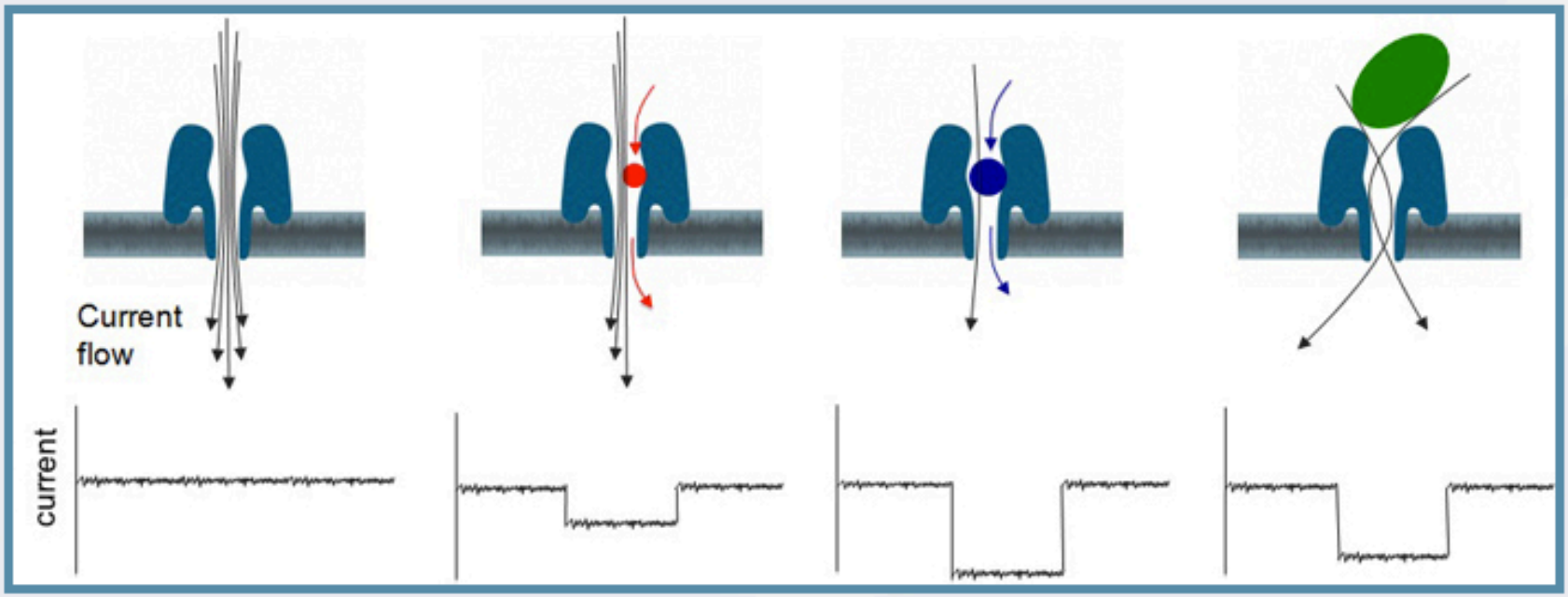
http://www.nanoporetech.com/technology/minion-a-miniaturised-sensing-instrument

Biological Nanopore

(Στο εγγύς μέλλον;)

Nanopore sensing

A nanopore may be used to identify a target analyte as follows.



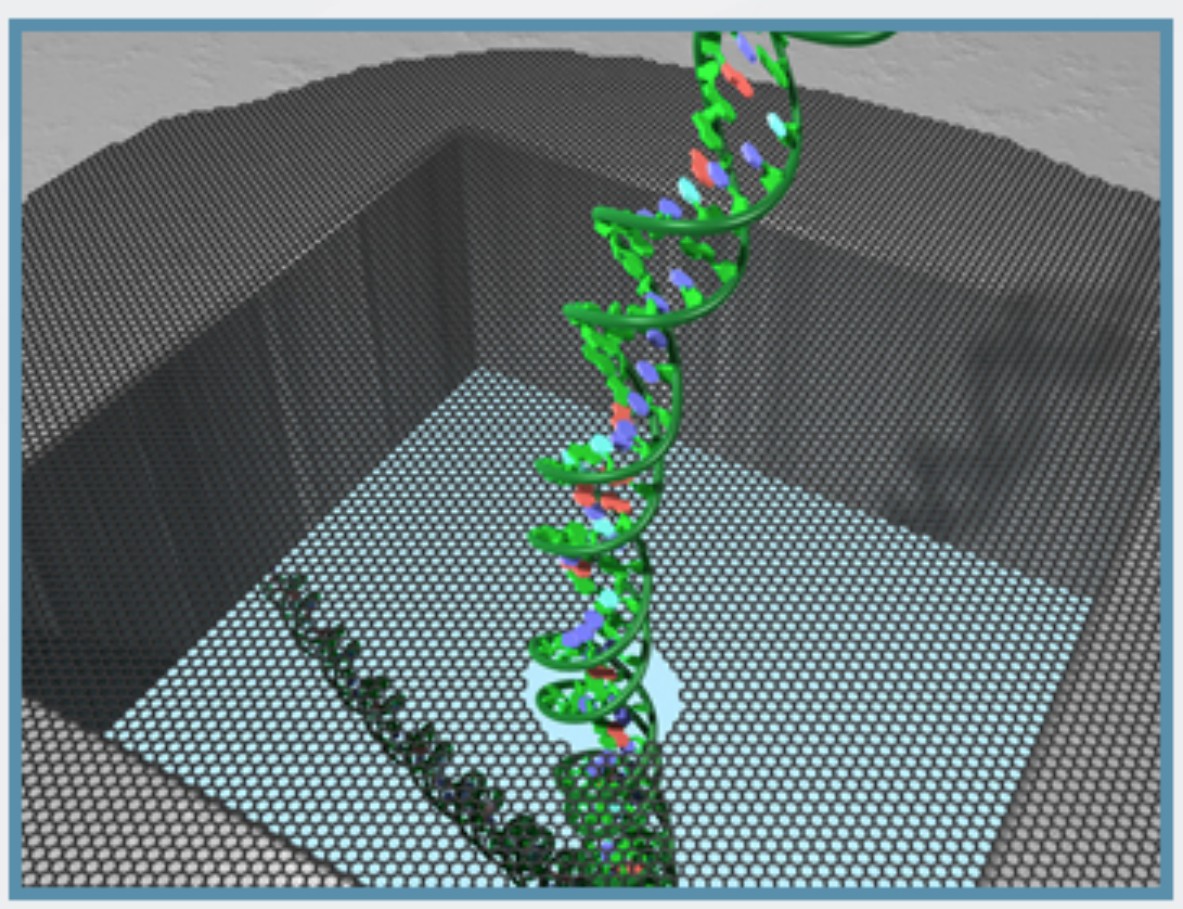
This diagram shows a protein nanopore set in an electrically resistant membrane bilayer. An ionic current is passed through the nanopore by setting a voltage across this membrane.

If an analyte passes through the pore or near its aperture, this event creates a characteristic disruption in current. By measuring that current, it is possible to identify the molecule in question. For example, this system can be used to distinguish between the four standard DNA bases G, A, T and C, and also modified bases. It can be used to identify target proteins, small molecules, or to gain rich molecular information, for example to distinguish the enantiomers of ibuprofen or molecular binding dynamics.

http://www.nanoporetech.com/technology/introduction-to-nanopore-sensing/introduction-to-nanopore-sensing

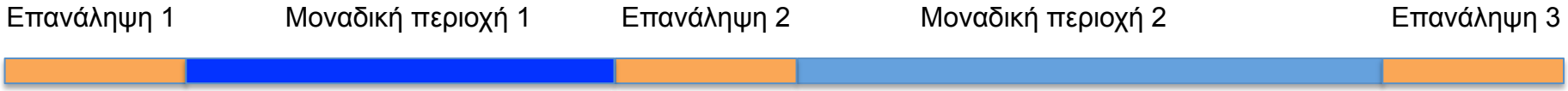
Solid state (Graphene) Nanopore

(Στο εγγύς μέλλον;)

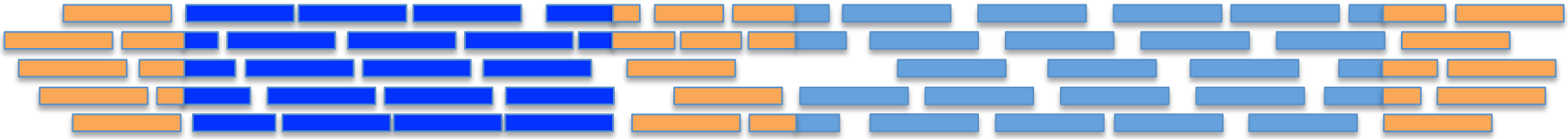


http://www.nanoporetech.com/technology/introduction-to-nanopore-sensing/solid-state-nanopores

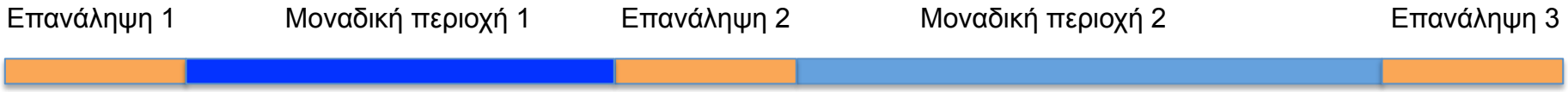
Reference assembly/alignment



Αλληλούχιση με Sequence Reads

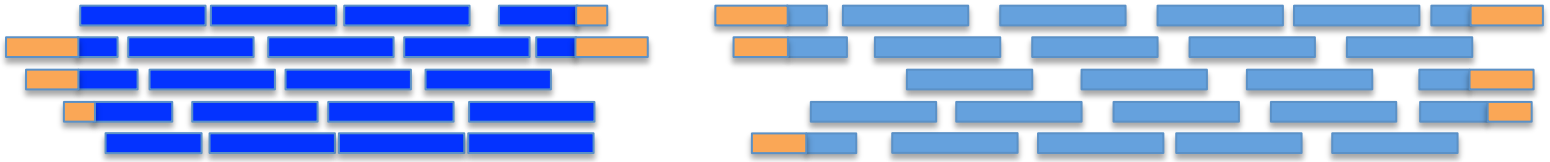


Συναρμολόγιση με βάση γονιδίωμα αναφοράς



Sequence Reads που μπορούν να στοιχιθούν
σε περισσότερες από μια θέσεις δεν στοιχίζονται

Μόνο στοίχιση των Sequence Reads που έχουν μια μοναδική θέση

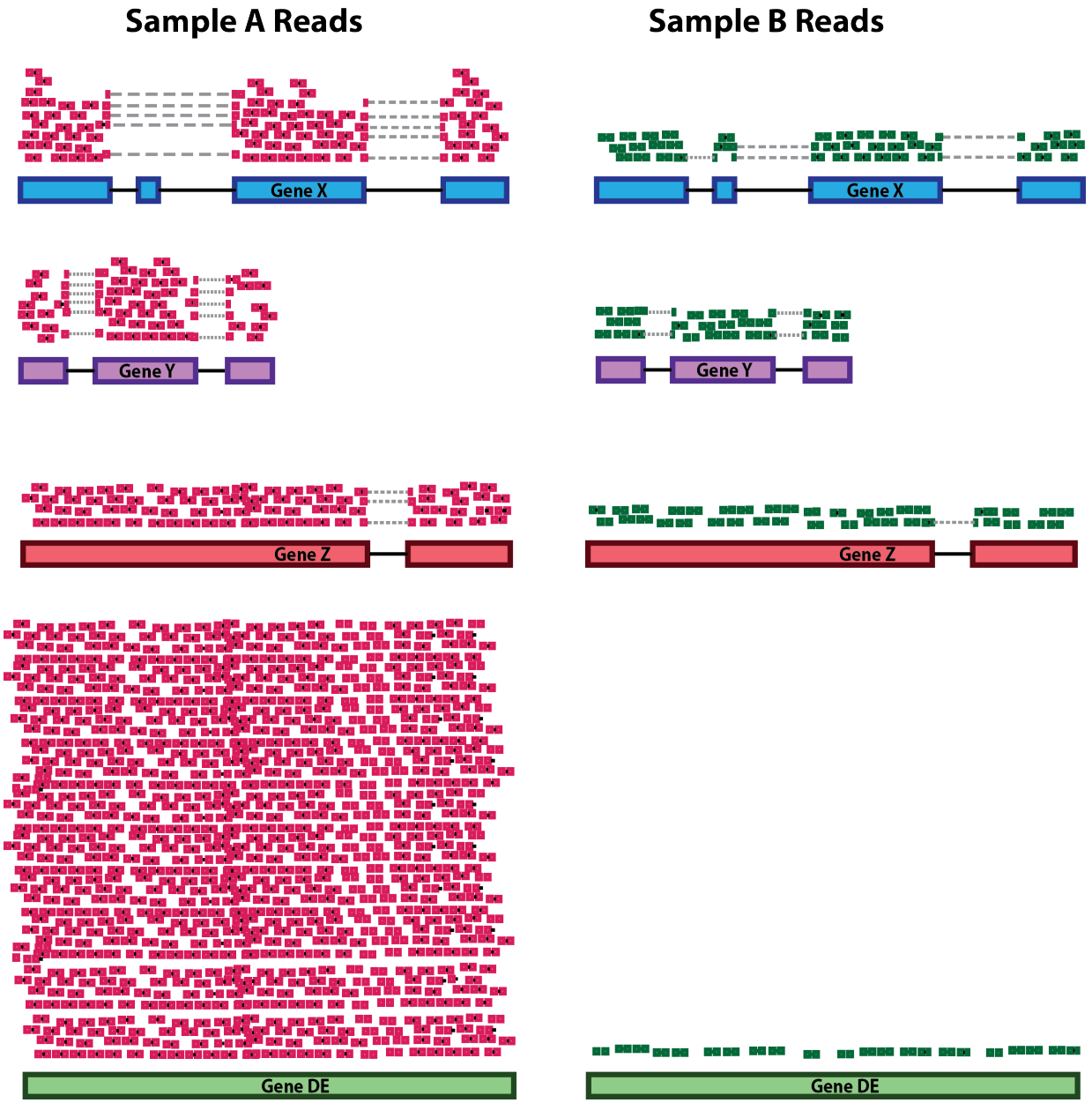


Reference assembly

Short read aligners

- Bowtie
- BWA
- STAR

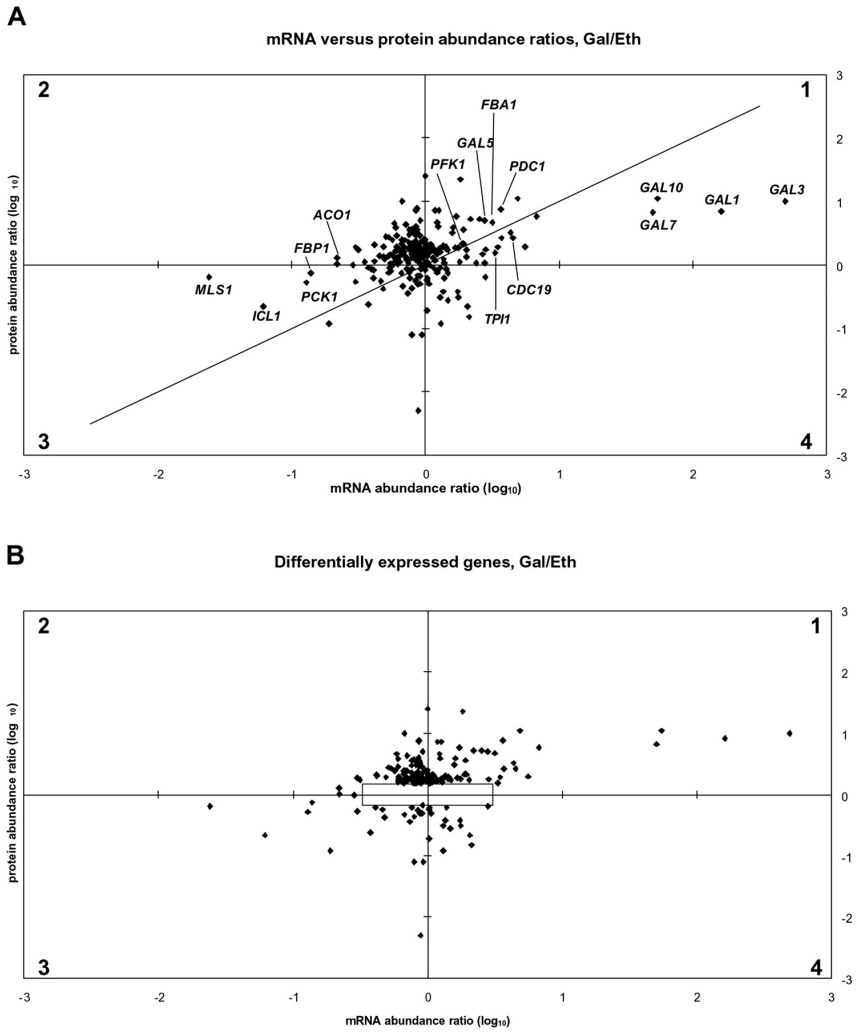
- RPKM – Reads per kilobase million
- FPKM – fragments per kilobase million
- TPM - Transcripts per million (TPM)



Διαφορική έκφραση γονιδίων

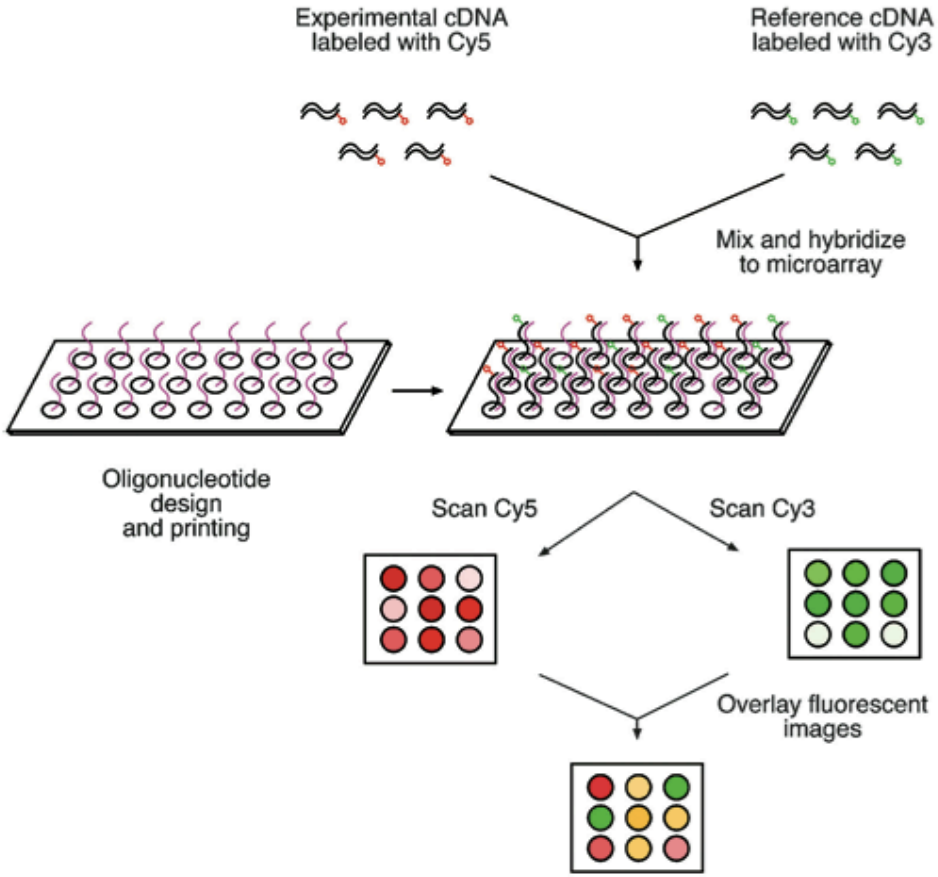
Microarrays & RNA-Sequencing

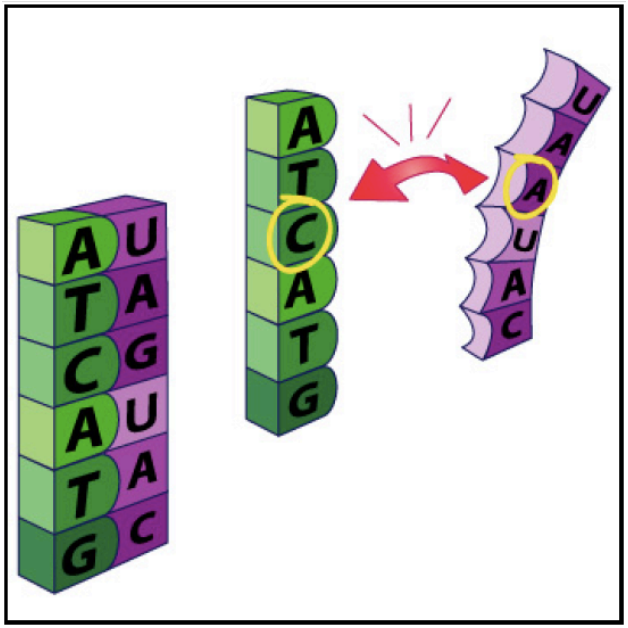
mRNA abundance ratios versus protein-abundance ratios.



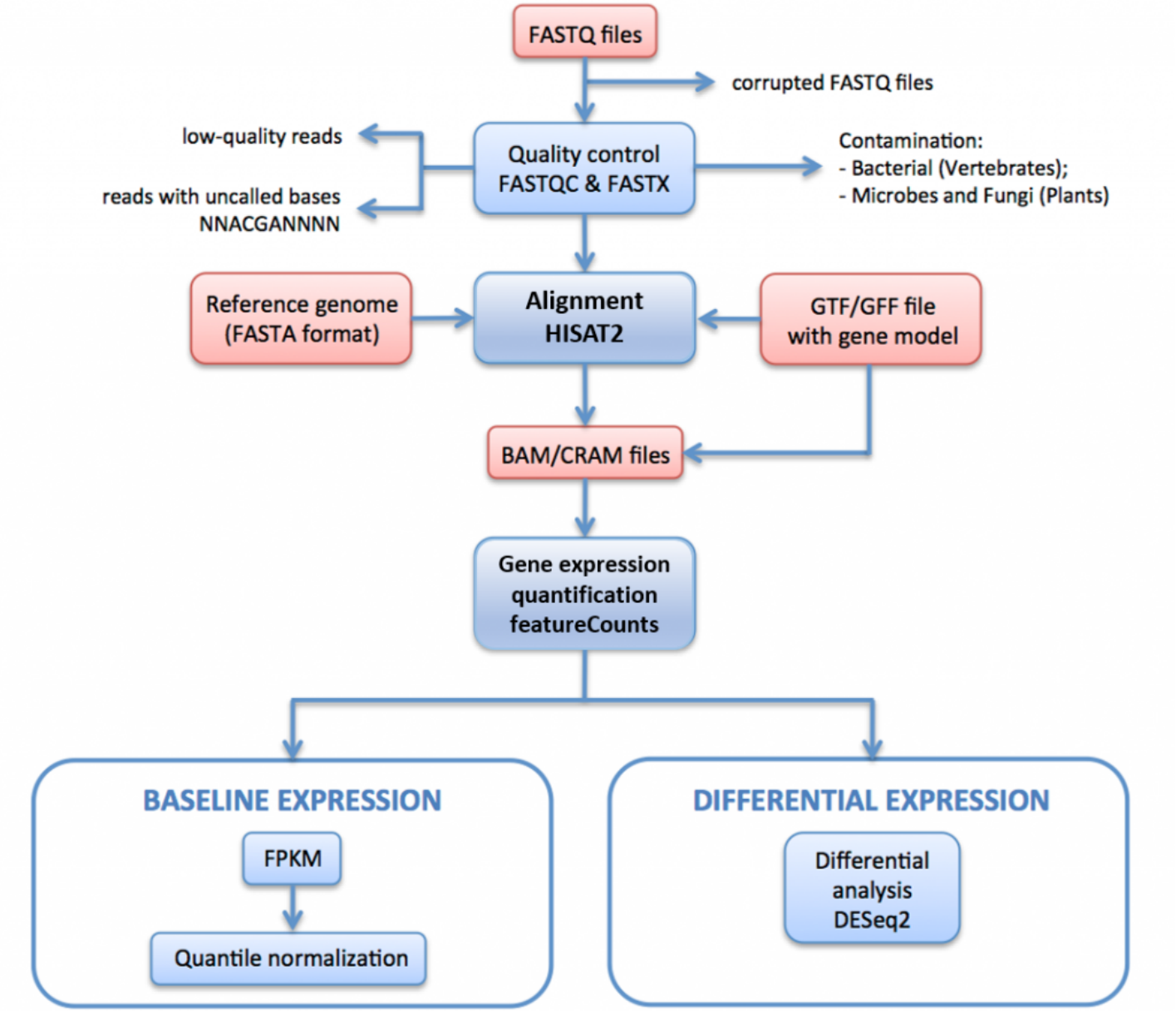
Griffin T J et al. Mol Cell Proteomics 2002;1:323-333

Μικροσυστοιχίες





RNA-SEQ


FASTQ files
corrupted FASTQ files
low-quality reads
reads with uncalled bases
NNACGANNNN
Quality control
FASTQC & FASTX
Contamination:
- Bacterial (Vertebrates);
- Microbes and Fungi (Plants)
Reference genome
(FASTA format)
Alignment
HISAT2
GTF/GFF file
with gene model
BAM/CRAM files
Gene expression
quantification
featureCounts
BASELINE EXPRESSION
FPKM
Quantile normalization
DIFFERENTIAL EXPRESSION
Differential
analysis
DESeq2

Log2

- Αν το γονίδιο εκφράζεται περισσότερο στην Α συνθήκη (κόκκινη χρωστική) από ότι στην control (πράσινη χρωστική), τότε ο λόγος συνθήκη_Α/control (κόκκινη/πράσινη) θα είναι $\lambda>1$, αλλιώς σε αντίθετη περίπτωση $0<\lambda<1$.
- Αν το γονίδιο εκφράζεται με διπλάσια ένταση στην συνθήκη Α, σε σχέση με την συνθήκη control, τότε ο λόγος θα είναι $\lambda=2$.
- Αν το γονίδιο εκφράζεται με τη μισή ένταση στην συνθήκη Α, σε σχέση με την συνθήκη control, τότε ο λόγος θα είναι $\lambda=0.5$.
- Μετατρέποντας τους λόγους σε $\log_2$, έχουμε:
  - $\lambda=2$ -> $\log_2\lambda=1$
  - $\lambda=0.5$ -> $\log_2\lambda=-1$
  - Με την κανονικοποίηση σε $\log_2$ τα δεδομένα γίνονται συμμετρικά.

Υπερ/υπο-έκφραση

- Πότε θεωρούμε ότι ένα γονίδιο υπερ/υπό-εκφράζεται σε μια συγκεκριμμένη συνθήκη.
  - $Log_2\lambda > 1$ ή $Log_2\lambda < -1$ (διπλάσια/υποδιπλάσια έκφραση σε σχέση με τη συνθήκη control).
  - Με στατιστικές μεθόδους (t-test, ANOVA).

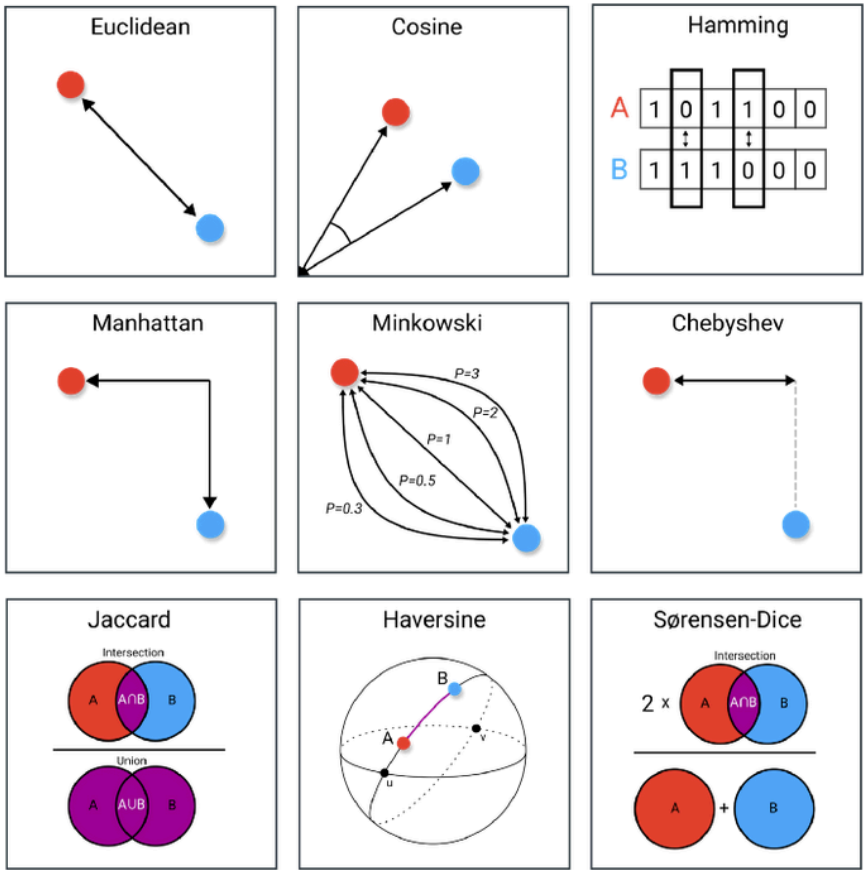
Ομαδοποίηση γονιδίων/συνθηκών με την ίδια συμπεριφορά.

- Χρειαζόμαστε αρκετά σημεία (διαφορετικές συνθήκες ή χρονικές στιγμές)
- Με μεθόδους αποστάσεων, όπου οι μετρήσεις ενός γονιδίου για διαφορετικές συνθήκες αποτελούν ένα διάνυσμα.
- Υπολογίζουμε αποστάσεις μεταξύ διαφορετικών διανυσμάτων (γονιδίων).
  - Ευκλείδια απόσταση
  - Συντελεστής συσχέτισης Pearson (Pearson correlation coefficient).
  - Δημιουργείται πίνακας αποστάσεων μεταξύ των γονιδίων.
  - Το αντίστοιχο μπορεί να γίνει και για να ομαδοποιήσουμε κοινές συνθήκες.

9 Distance Measures in Data Science

The advantages and pitfalls of common distance measures

Maarten Grootendorst Feb 1 · 10 min read ★



Distance Measures. Image by the author.

	Condition1	Condition2	Condition3	Condition4	Condition5
Gene1	1	−3	10	0	0
Gene2	−7	−2	−1	10	−8
Gene3	2	1	9	−9	5
Gene4	10	10	−4	0	−9
Gene5	−2	9	−7	0	−7
Gene6	−6	6	−5	−3	9
Gene7	2	1	8	−1	−2
Gene8	−3	−8	−1	−6	2
Gene9	−10	0	9	6	0
Gene10	−2	4	5	−7	−6
Gene11	−2	−2	0	−9	10
Gene12	−6	−10	−5	8	5
Gene13	2	−8	1	−1	2
Gene14	−7	−9	−7	1	1
Gene15	−6	4	−8	−1	−6
Gene16	−5	2	−5	8	−8
Gene17	8	−2	−7	0	2
Gene18	2	9	−9	9	3
Gene19	−3	−1	7	−1	6
Gene20	10	−4	3	−3	−1

	Condition1	Condition2
Gene1	1	−3
Gene2	−7	−2
Gene3	2	1
Gene4	10	10
Gene5	−2	9
Gene6	−6	6
Gene7	2	1
Gene8	−3	−8
Gene9	−10	0
Gene10	−2	4
Gene11	−2	−2
Gene12	−6	−10
Gene13	2	−8
Gene14	−7	−9
Gene15	−6	4
Gene16	−5	2
Gene17	8	−2
Gene18	2	9
Gene19	−3	−1
Gene20	10	−4

	Condition1	Condition2	Condition3	Condition4	Condition5
Gene1	1	−3	10	0	0
Gene2	−7	−2	−1	10	−8

	Condition1	Condition2	Condition3	Condition4	Condition5
Gene1	1	−3	10	0	0
Gene2	−7	−2	−1	10	−8

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8	Gene9	Gene10	Gene11	Gene12	Gene13	Gene14	Gene15	Gene16	Gene17	Gene18	Gene19	Gene20
Gene1																				
Gene2																				
Gene3																				
Gene4																				
Gene5																				
Gene6																				
Gene7																				
Gene8																				
Gene9																				
Gene10																				
Gene11																				
Gene12																				
Gene13																				
Gene14																				
Gene15																				
Gene16																				
Gene17																				
Gene18																				
Gene19																				
Gene20																				

	Condition1	Condition2
Gene1	1	−3
Gene2	−7	−2
Gene3	2	1
Gene4	10	10
Gene5	−2	9
Gene6	−6	6
Gene7	2	1
Gene8	−3	−8
Gene9	−10	0
Gene10	−2	4
Gene11	−2	−2
Gene12	−6	−10
Gene13	2	−8
Gene14	−7	−9
Gene15	−6	4
Gene16	−5	2
Gene17	8	−2
Gene18	2	9
Gene19	−3	−1
Gene20	10	−4

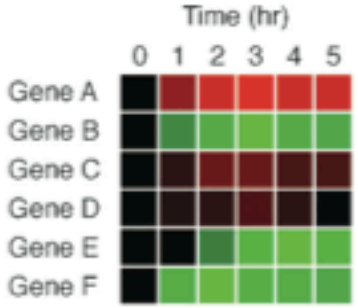
	Condition1	Condition2	Condition3	Condition4	Condition5
Condition1					
Condition2					
Condition3					
Condition4					
Condition5					

Ομαδοποίηση

	0 hr	1 hr	2 hr	3 hr	4 hr	5 hr
Gene A	1	4	6	8	6	6
Gene B	1	0.6	0.3	0.1	0.3	0.4
Gene C	1	2	4	4	3	3
Gene D	1	1.5	2	3	2	1
Gene E	1	1	0.5	0.2	0.1	0.2
Gene F	1	0.3	0.1	0.2	0.3	0.4

convert to false colors

$\log_2$ conversion



	Gene B	Gene C	Gene D	Gene E	Gene F
Gene A	-0.82	0.96	0.65	-0.68	-0.79
Gene B		-0.85	-0.86	0.66	0.67
Gene C			0.70	-0.65	-0.87
Gene D				-0.41	-0.72
Gene E					0.26

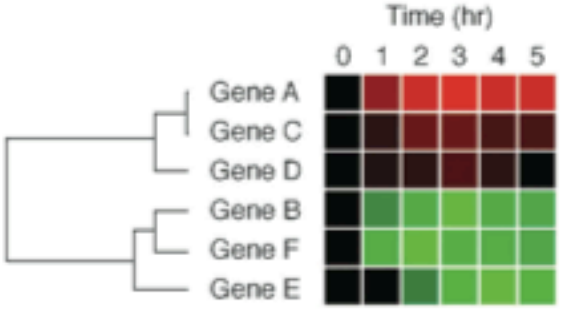
calculating Pearson correlation coefficients between genes

	0 hr	1 hr	2 hr	3 hr	4 hr	5 hr
Gene A	0	2	2.6	3	2.6	2.6
Gene B	0	-0.7	-1.7	-3.3	-1.7	-1.3
Gene C	0	1	2	2	1.6	1.6
Gene D	0	0.6	1	1.6	1	0
Gene E	0	0	-1	-2.3	-3.3	-2.3
Gene F	0	-1.7	-3.3	-2.3	-1.7	-1.3

conversion of coefficients to positive distance values

	Gene B	Gene C	Gene D	Gene E	Gene F
Gene A	1.82	0.04	0.35	1.68	1.79
Gene B		1.85	1.86	0.34	0.33
Gene C			0.30	1.65	1.87
Gene D				1.41	1.72
Gene E					0.74

hierarchical clustering



Οντολογίες

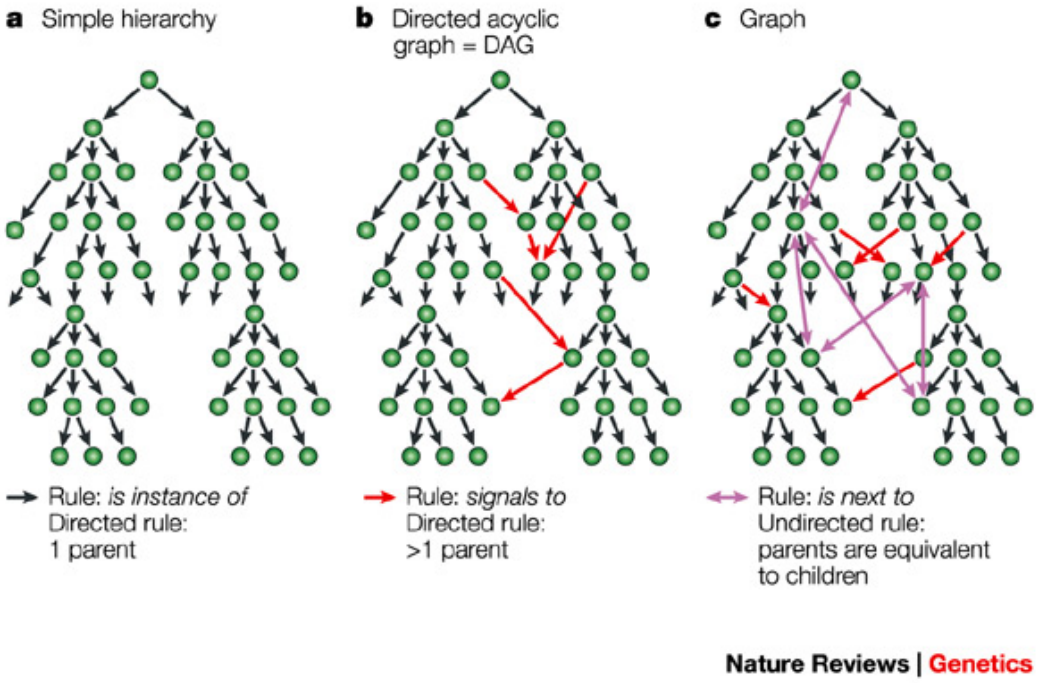
- www.geneontology.org

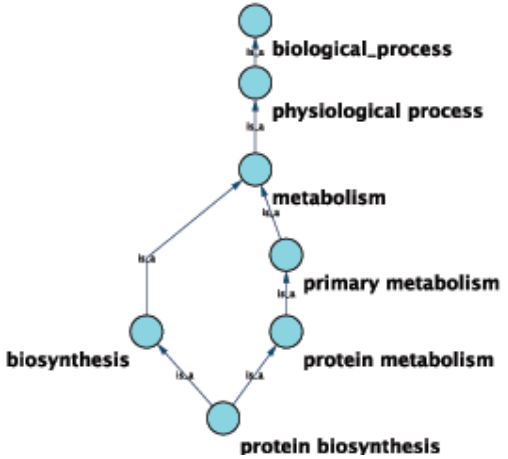
- Ελεγχόμενο λεξιλόγιο για την περιγραφή των ιδιοτήτων των γονιδίων και των πρωτεϊνών.

- Περιγράφουν:
  - Μοριακές λειτουργίες του βιομορίου (1 ή περισσότερες).
  - Βιολογικές διαδικασίες στις οποίες εμπλέκεται το βιομόριο (1 ή περισσότερες).
  - Κυτταρικό διαμέρισμα στο οποίο συναντάται το βιομόριο (1 ή περισσότερα).

Οντολογίες: Η δομή τους

- Δείχνει τις σχέσεις μεταξύ των διαφορετικών όρων.
- Ένας όρος μπορεί να αποτελεί πιο εξειδικευμένη περιγραφή ενός άλλου όρου.
- Είναι κατευθυνόμενα ακυκλικά γραφήματα (DAG).
- Παρόμοια με ιεραρχίες.
- Η διαφορά είναι ότι ένας κόμβος-απόγονος μπορεί να έχει περισσότερους από έναν προγόνους.





Gene ontology

Table 1 | **Evidence codes used by GO**

Evidence code	Evidence code description	Source of evidence	Manually checked	Current number of annotations*
IDA	Inferred from direct assay	Experimental	Yes	71,050
IEP	Inferred from expression pattern	Experimental	Yes	4,598
IGI	Inferred from genetic interaction	Experimental	Yes	8,311
IMP	Inferred from mutant phenotype	Experimental	Yes	61,549
IPI	Inferred from physical interaction	Experimental	Yes	17,043
ISS	Inferred from sequence or structural similarity	Computational	Yes	196,643
RCA	Inferred from reviewed computational analysis	Computational	Yes	103,792
IGC	Inferred from genomic context	Computational	Yes	4
IEA	Inferred from electronic annotation	Computational	No	15,687,382
IC	Inferred by curator	Indirectly derived from experimental or computational evidence made by a curator	Yes	5,167
TAS	Traceable author statement	Indirectly derived from experimental or computational evidence made by the author of the published article	Yes	44,564
NAS	Non-traceable author statement	No 'source of evidence' statement given	Yes	25,656
ND	No biological data available	No information available	Yes	132,192
NR	Not recorded	Unknown	Yes	1,185

*October 2007 release

Οντολογίες: στατιστική ανάλυση

- Παράδειγμα:
  - 1 γονιδίωμα με 10.000 γονίδια.
  - 1.000 γονίδια εμπλέκονται στον κυτταρικό κύκλο (GO_term: cell-cycle). (10% του γονιδιώματος).
  - Αν επιλέξουμε τυχαία έναν αριθμό Χ γονιδίων, θα περιμέναμε (από τύχη) περίπου το 10% (με κάποιες διακυμάνσεις) να έχουν τον όρο “κυτταρικός κύκλος”.
  - Η τυχαία διακύμανση εξαρτάται από τον αριθμό των γονιδίων.
  - Έστω ότι με τα microarrays σε ένα πείραμα βρήκαμε ότι Χ αριθμός γονιδίων υπερεκφράζονται.
  - Σε αυτό τον Χ αριθμό, βρήκαμε ότι 20% των γονιδίων ανήκουν στον κυτταρικό κύκλο.
  - Αυτή η απόκλιση (20% παρατηρούμενο - 10% αναμενόμενο) είναι στα όρια των τυχαίων διακυμάνσεων, ή είναι στατιστικά σημαντική?
    - Στατιστικά σημαντική, σημαίνει ότι τα υπερεκφρασμένα γονίδια είναι εμπλουτισμένα για την κατηγορία “κυτταρικός κύκλος ”. Δηλαδή, ο κυτταρικός κύκλος εμπλέκεται στην διαδικασία που μελετάμε.

Οντολογίες: στατιστική ανάλυση

- Η στατιστική ανάλυση γίνεται με το υπεργεωμετρικό τεστ.
- Παίρνουμε ένα p-value.
- Αν p-value < 0.05, τότε είναι στατιστικά σημαντικό.

- Αν στις οντολογίες μας είχαμε 100 όρους, θα επαναλαμβάναμε τα παραπάνω τεστς για τον κάθε όρο.
- Όμως, όσο περισσότερα τεστ κάνουμε για το πείραμά μας, τόσο αυξάνει ή πιθανότητα να βρούμε κάτι στατιστικά σημαντικό (p-value < 0.05) καθαρά από λάθος.
- Άρα, πρέπει να λάβουμε υπόψην μας πόσα τεστ διενεργούμε και να διορθώσουμε τα p-values (multiple testing correction).
  - False discovery rate (Benjamini-Hochberger)
  - Bonferroni correction

Βάσεις Δεδομένων

Ετήσιος κατάλογος ΒΔ

- Κάθε Ιανουάριο στο Nucleic Acids Research (Special database issue)

OXFORD ACADEMIC | Journals

You are here: NAR Journal Home » Database Summary Paper Categories

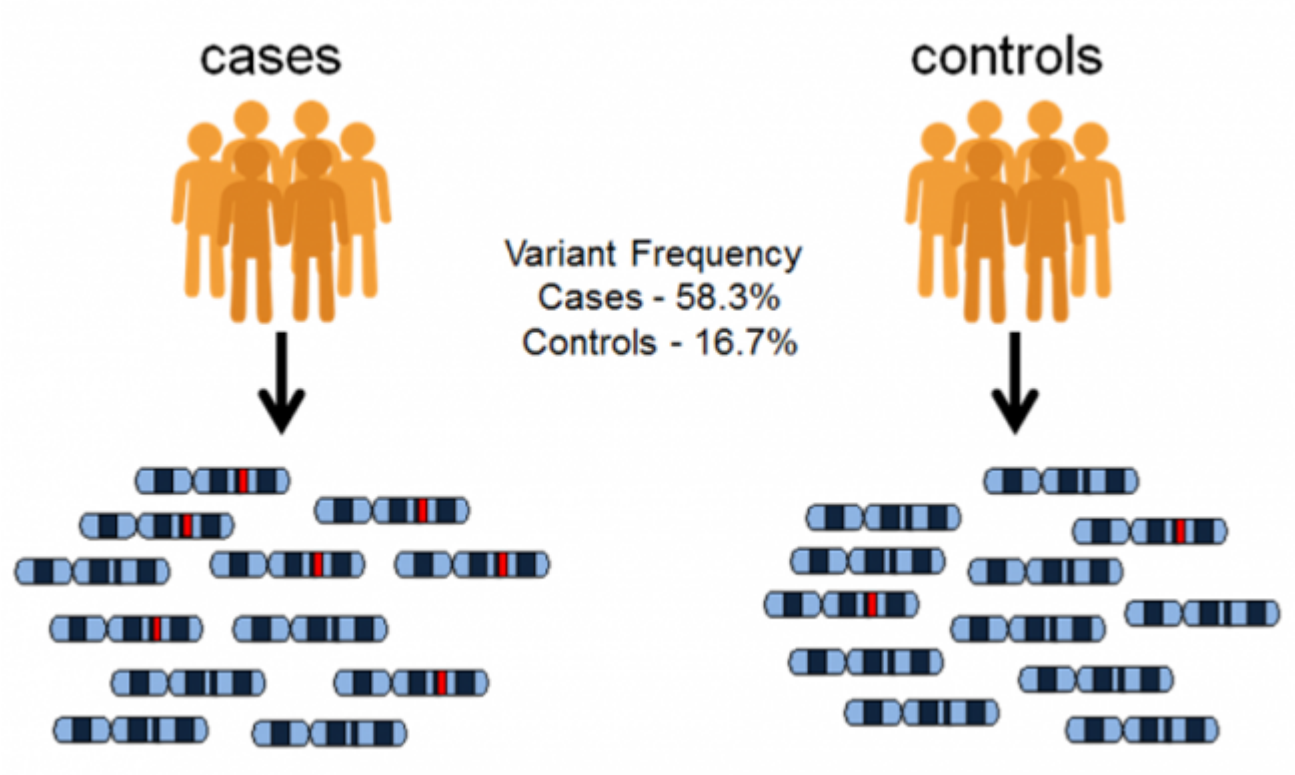
NAR Database Summary Paper Category List

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
CancerResource
General human genetics databases
General polymorphism databases
Cancer gene databases
Gene-, system- or disease-specific databases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases
Cell biology

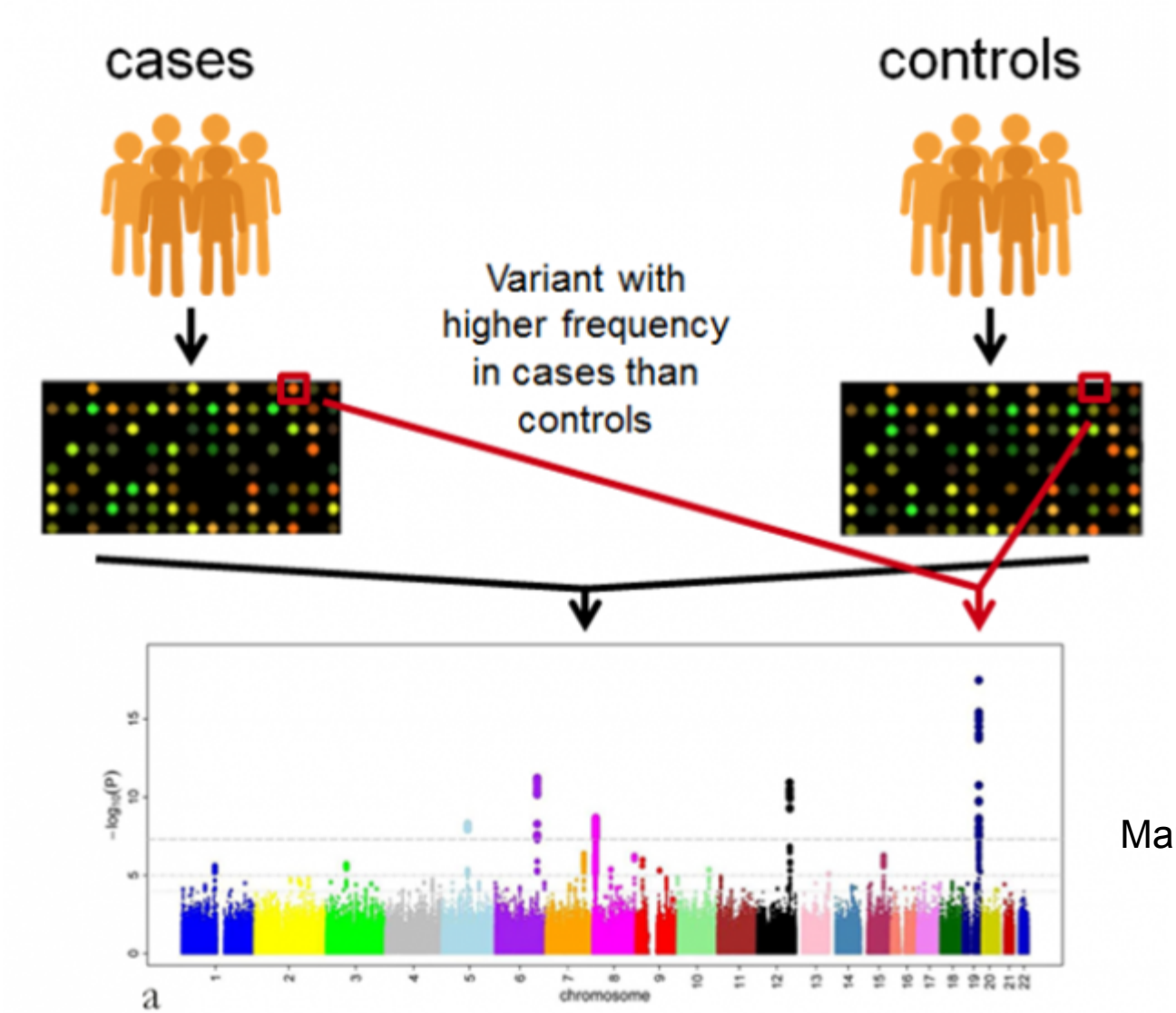
- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

GWAS



GWAS
SNParrays − WGS - WES

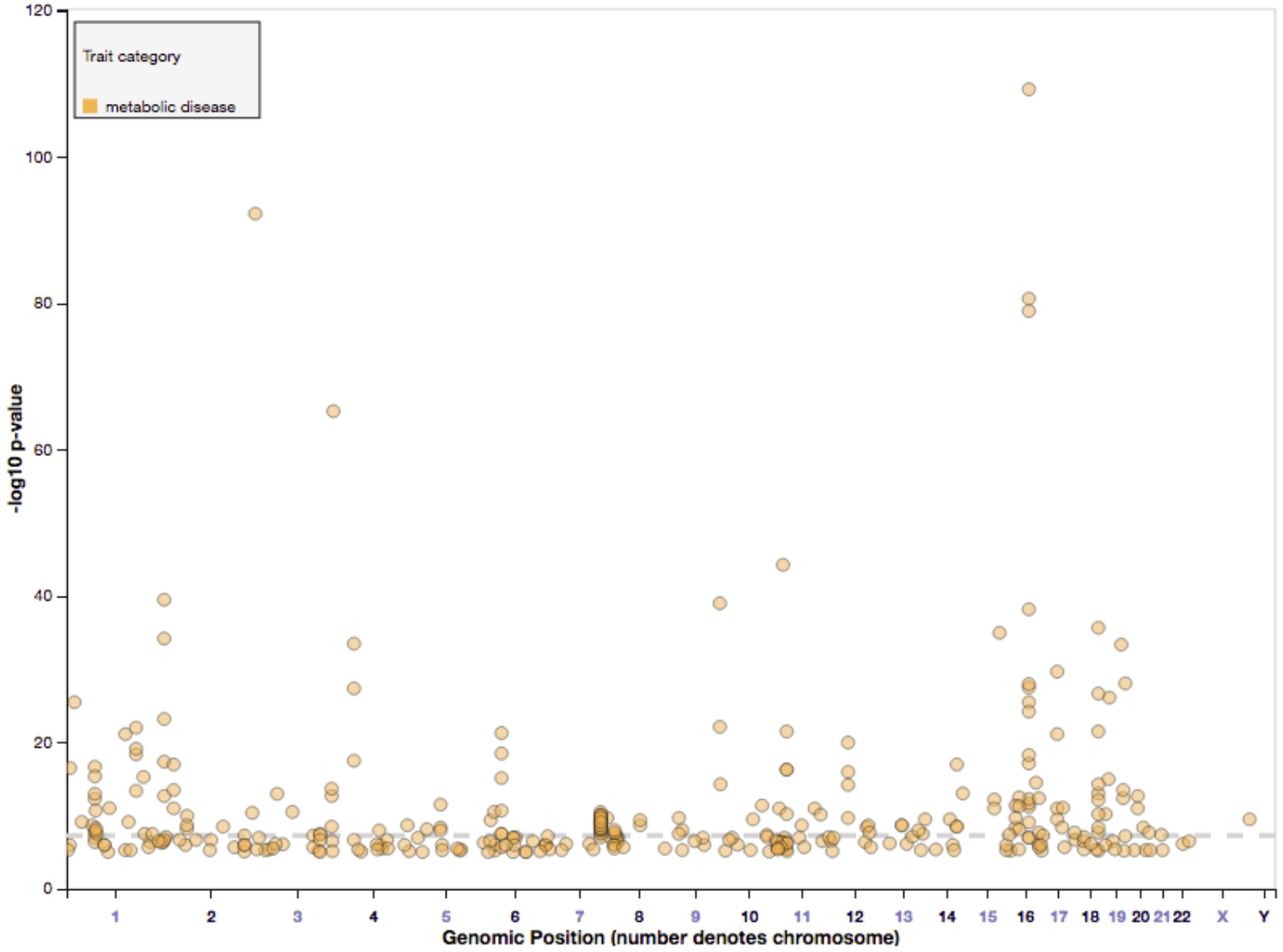
Cases vs Controls − pvalue corrected for multiple testing



Manhattan plots

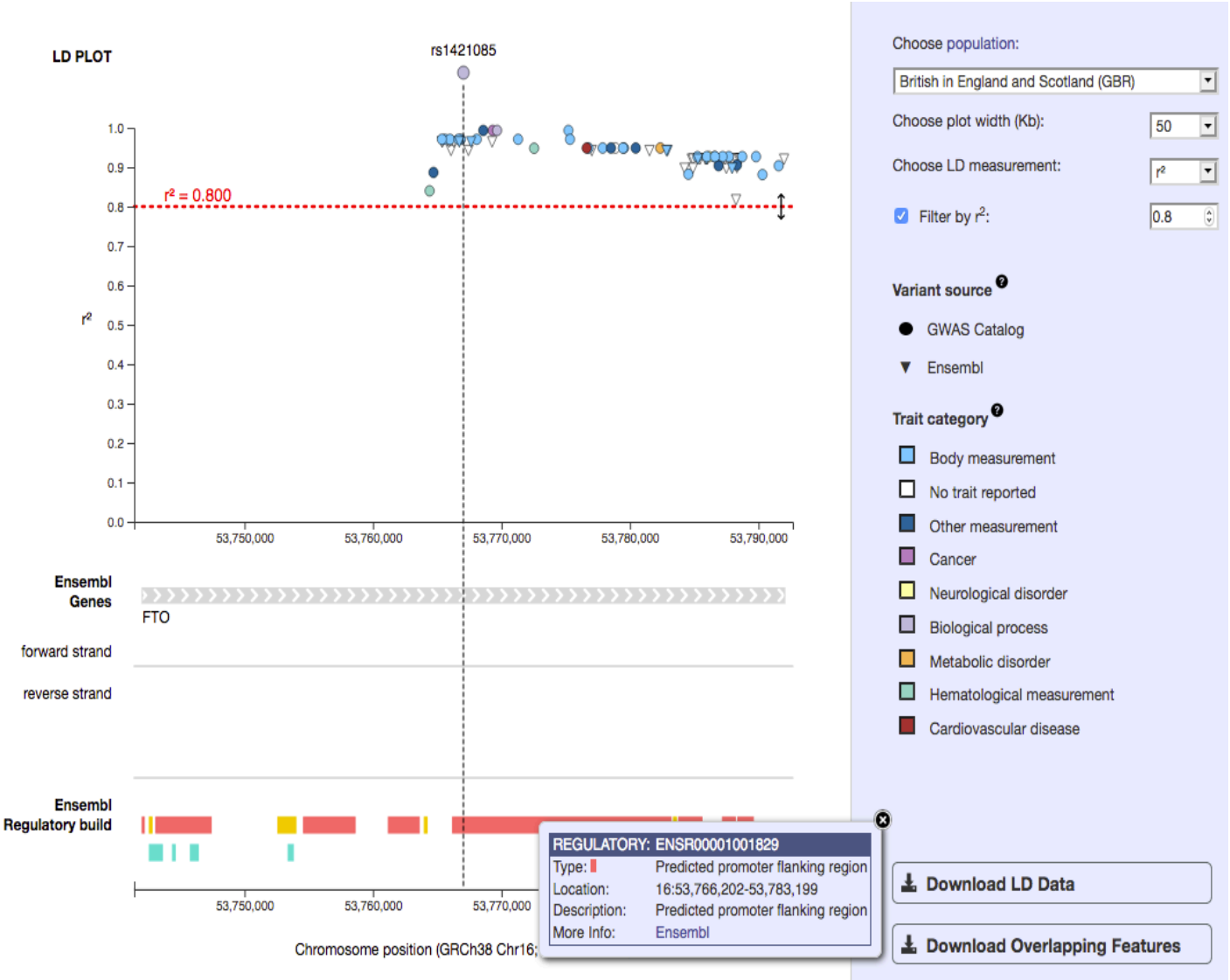
GWAS catalogue - EMBL-EBI and NHGRI.

Παχυσαρκία: SNP rs1421085

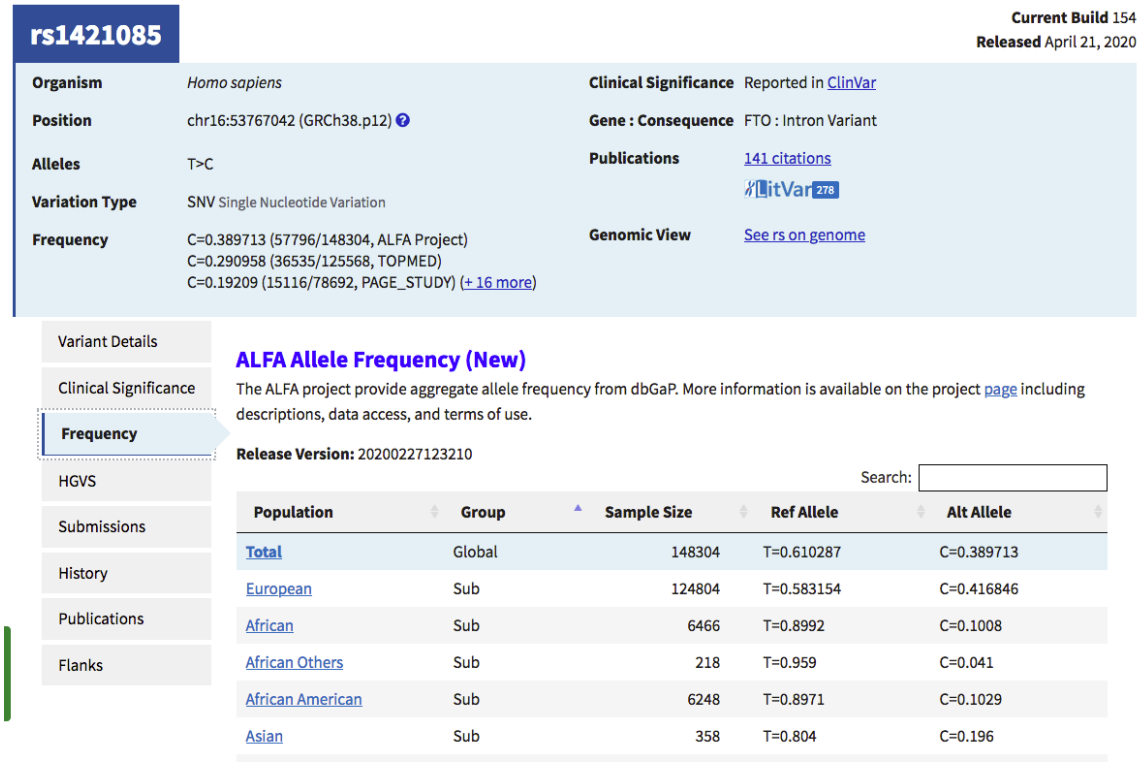


GWAS catalogue

Παχυσαρκία: SNP rs1421085



dbSNP - NCBI



rs1421085

Current Build 154
Released April 21, 2020

Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr16:53767042 (GRCh38.p12)	**Gene : Consequence**	FTO : Intron Variant
Alleles	T>C	**Publications**	141 citations, LitVar 278
Variation Type	SNV Single Nucleotide Variation		
Frequency	C=0.389713 (57796/148304, ALFA Project) C=0.290958 (36535/125568, TOPMED) C=0.19209 (15116/78692, PAGE_STUDY) (+ 16 more)	**Genomic View**	See rs on genome

Variant Details
Clinical Significance
Frequency
HGVS
Submissions
History
Publications
Flanks

ALFA Allele Frequency (New)

The ALFA project provide aggregate allele frequency from dbGaP. More information is available on the project page including descriptions, data access, and terms of use.

Release Version: 20200227123210

Search:

Population	Group	Sample Size	Ref Allele	Alt Allele
Total	Global	148304	T=0.610287	C=0.389713
European	Sub	124804	T=0.583154	C=0.416846
African	Sub	6466	T=0.8992	C=0.1008
African Others	Sub	218	T=0.959	C=0.041
African American	Sub	6248	T=0.8971	C=0.1029
Asian	Sub	358	T=0.804	C=0.196

- Στην dbSNP, ο χρήστης μπορεί να αναζητήσει πληροφορίες για SNPs, χρησιμοποιώντας το reference SNP number, το όνομα του γονιδίου, γονιδιωματικές συντεταγμένες και άλλα στοιχεία.
- Για ένα συγκεκριμένο SNP, ο χρήστης μπορεί να δει σε τι συχνότητες εμφανίζεται στους διάφορους πληθυσμούς, σε ποιά γονιδιακή περιοχή εντοπίζεται, άλλα SNPs που υπάρχουν στην εγγύς γειτονιά, τι επιπτώσεις έχει (αν π.χ. εντοπίζεται σε ιντρόνιο, αν αλλάζει κάποιο αμινοξύ κ.τ.λ.), ποιές δημοσιευμένες εργασίες το αναφέρουν, όπως και εάν έχει κάποια κλινική σημασία
- Σε αυτή την περίπτωση, δίνεται ο αντίστοιχος σύνδεσμος στην βάση δεδομένων ClinVar (επίσης του NCBI), που περιέχει κλινικά σημαντικά SNPs και τις αντίστοιχες πληροφορίες τους.

KEGG Search Help

» Japanese

KEGG Home
Release notes
Current statistics

KEGG Database
KEGG overview
Searching KEGG
KEGG mapping
Color codes

KEGG Objects
Pathway maps
Brite hierarchies
KEGG DB links

KEGG Software
KEGG API
KGML

KEGG FTP
Subscription
Background info

GenomeNet

DBGET/LinkDB

Feedback
Copyright request

Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
See Release notes (April 1, 2021) for new and updated features.

New article KEGG: integrating viruses and cellular organisms

Main entry point to the KEGG web service

KEGG2 KEGG Table of Contents [Update notes | Release history]

Data-oriented entry points

KEGG PATHWAY	KEGG pathway maps	Pathway
KEGG BRITE	BRITE hierarchies and tables	Brite
KEGG MODULE	KEGG modules	Brite table
KEGG ORTHOLOGY	KO functional orthologs [Annotation]	Module
KEGG GENOME	Genomes [Pathogen | Virus | Plant]	Network
KEGG GENES	Genes and proteins [SeqData]	KO (Function)
KEGG COMPOUND	Small molecules	Organism
KEGG GLYCAN	Glycans	Virus
KEGG REACTION	Biochemical reactions [RModule]	Compound
KEGG ENZYME	Enzyme nomenclature	Disease (ICD)
KEGG NETWORK	Disease-related network variations	Drug (ATC)
KEGG DISEASE	Human diseases	Drug (Target)
KEGG DRUG	Drugs [New drug approvals]	Antiinfectives
KEGG MEDICUS	Health information resource [Drug labels search]	

Organism-specific entry points

KEGG Organisms Enter org code(s) Go hsa hsa eco

Analysis tools

KEGG Mapper	KEGG PATHWAY/BRITE/MODULE mapping tools
BlastKOALA	BLAST-based KO annotation and KEGG mapping
GhostKOALA	GHOSTX-based KO annotation and KEGG mapping
KofamKOALA	HMM profile-based KO annotation and KEGG mapping
BLAST/FASTA	Sequence similarity search
SIMCOMP	Chemical structure similarity search

KEGG Pathway Maps

Option

One-click mode

Row border shading

Search

Go

ID search

Go

Join

Metabolism

Genetic Information Processing

Environmental Information Processing

Cellular Processes

Organismal Systems

Human Diseases

- Cancer: overview
  - 05200 Pathways in cancer
  - 05202 Transcriptional misregulation in cancer
  - 05206 MicroRNAs in cancer
  - 05205 Proteoglycans in cancer
  - 05204 Chemical carcinogenesis
  - 05203 Viral carcinogenesis
  - 05230 Central carbon metabolism in cancer
  - 05231 Choline metabolism in cancer
  - 05235 PD-L1 expression and PD-1 checkpoint pathway in cancer
- Cancer: specific types
  - 05210 Colorectal cancer
  - 05212 Pancreatic cancer
  - 05225 Hepatocellular carcinoma
  - 05226 Gastric cancer
  - 05214 Glioma
  - 05216 Thyroid cancer
  - 05221 Acute myeloid leukemia
  - 05220 Chronic myeloid leukemia
  - 05217 Basal cell carcinoma
  - 05218 Melanoma
  - 05211 Renal cell carcinoma
  - 05219 Bladder cancer
  - 05215 Prostate cancer
  - 05213 Endometrial cancer
  - 05224 Breast cancer
  - 05222 Small cell lung cancer
  - 05223 Non-small cell lung cancer
- Infectious disease: viral
  - 05166 Human T-cell leukemia virus 1 infection
  - 05170 Human immunodeficiency virus 1 infection
  - 05161 Hepatitis B
  - 05160 Hepatitis C
  - 05171 Coronavirus disease - COVID-19
  - 05164 Influenza A
  - 05162 Measles
  - 05168 Herpes simplex virus 1 infection
  - 05163 Human cytomegalovirus infection
  - 05167 Kaposi sarcoma-associated herpesvirus infection
  - 05169 Epstein-Barr virus infection

KEGG **Breast cancer - Homo sapiens (human)**

[Pathway menu | Pathway entry | Download KGML | Show description | Image (png) file | Help]

Change pathway type

Option

Scale: 100%

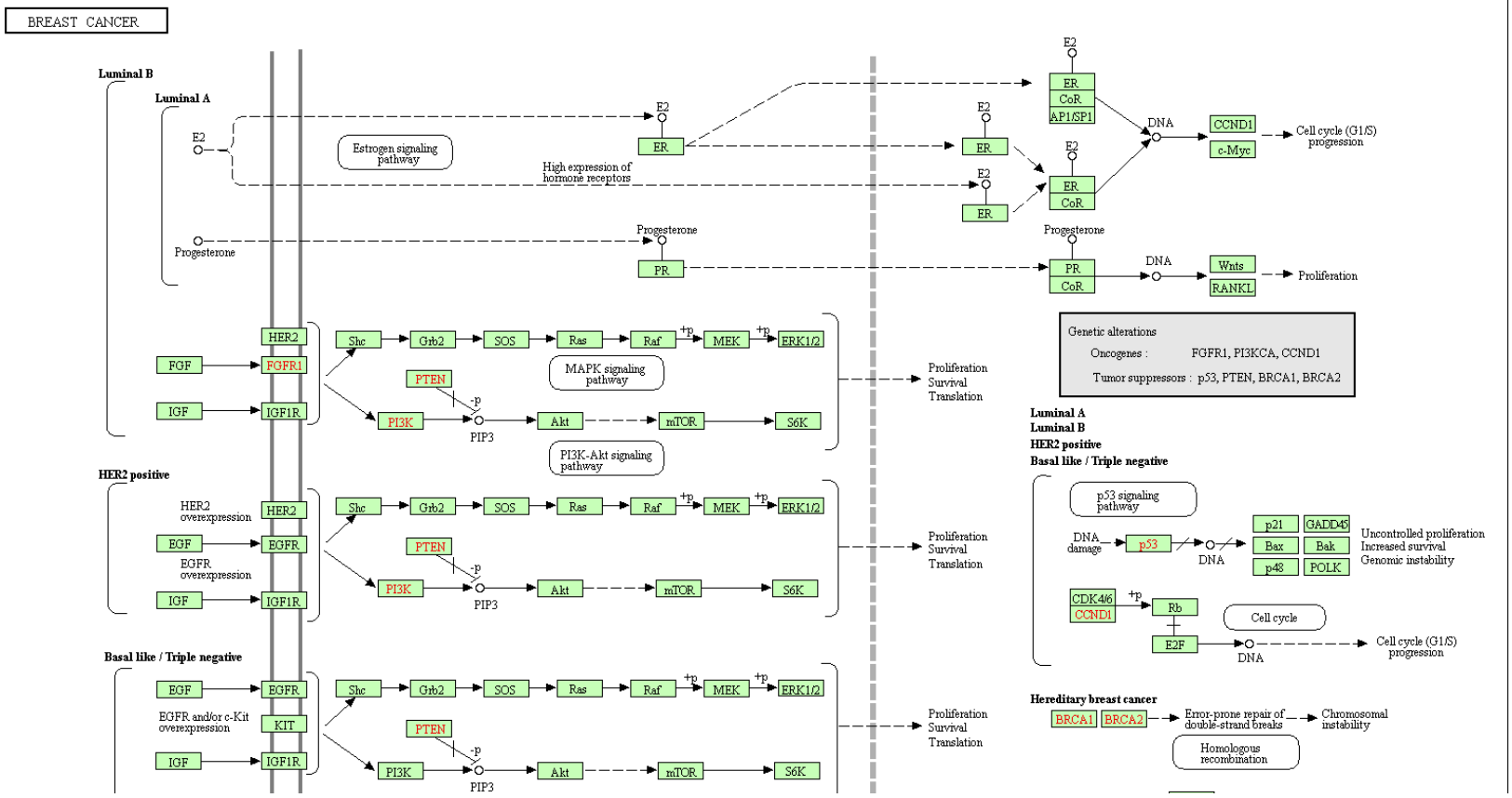
Search

Go

User data +

Network

- nt06210 ERK signaling
  - N00041 EGFR-overexpression
  - N00022 ERBB2-overexpressio
  - N00020 Amplified FGFR to RA
- nt06214 PI3K signaling
  - N00042 EGFR-overexpression
  - N00034 ERBB2-overexpressio
  - N00038 Amplified FGFR to PI3
  - N00049 Mutation-activated PI
  - N00050 Amplified PI3K to PI3
  - N00052 Mutation-inactivated
- nt06215 WNT signaling
  - N00059 FZD7-overexpression
  - N00060 LRP6-overexpression
- nt06216 NOTCH signaling
  - N00087 NOTCH-overexpressio
- nt06227 Nuclear receptor signal
  - N00287 ESR1-positive to nucl
- nt06230 Cell cycle G1/S
  - N00275 Amplified CCND1 to c
- nt06240 Transcription
  - N00115 Mutation-inactivated



KEGG

Homo sapiens (human): 2099

Help

Entry	2099 CDS T01001
Gene name	ESR1, ER, ESR, ESRA, ESTRR, Era, NR3A1
Definition	(RefSeq) estrogen receptor 1
KO	K08550 estrogen receptor alpha
Organism	hsa Homo sapiens (human)
Pathway	hsa01522 Endocrine resistance hsa04915 Estrogen signaling pathway hsa04917 Prolactin signaling pathway hsa04919 Thyroid hormone signaling pathway hsa04961 Endocrine and other factor-regulated calcium reabsorption hsa05200 Pathways in cancer hsa05205 Proteoglycans in cancer hsa05224 Breast cancer
Network	nt06227 Nuclear receptor signaling nt06270 Breast cancer nt06323 KISS1-GnRH-LH/FSH-E2 signaling
Element	N00286 Nuclear-initiated estrogen signaling pathway N00287 ESR1-positive to nuclear-initiated estrogen signaling pathway N00887 Mutation-inactivated ESR1 to nuclear-initiated estrogen signaling pathway
Disease	H00031 Breast cancer H02061 Estrogen resistance syndrome
Drug target	Acolbifene hydrochloride: D02758 Afimoxifene: D06551 Alfatradiol: D07121 Arzoxifene hydrochloride: D02993 Bazedoxifene: D03062<JP> Brilanestrant: D11264 Chlorotrianisene: D00269 Clometherone: D03551 Clomifene (DG00474): D00962<JP/US> D07726 Danazol: D00289<JP/US> Delmadinone (DG02935): D03675 D07783 Desogestrel: D02367 Dienestrol: D00898 Diethylstilbestrol (DG00466): D00577 D00946 D07826 Droloxifene (DG01265): D03911 D03912 Elacestrant (DG03078): D11671 D11672 Enclomiphene (DG01992): D08910 D10876 Epitiostanol: D01265 Equilin: D04041 Estetrol (DG03027): D11513 D11514 Estradiol (DG00462): D00105<JP/US> D01413<JP/US> D01617 D01953 D04061<US> D04063<US> D04064 D04065 D07918<US>

All links

Ontology (3)
- KEGG BRITE (3)

Pathway (8)
- KEGG PATHWAY (8)

Network (3)
- KEGG NETWORK (3)

Disease (7)
- KEGG DISEASE (2)
- OMIM (5)

Drug (81)
- KEGG DRUG (81)

Genome (1)
- KEGG GENOME (1)

Gene (40)
- KEGG ORTHOLOGY (1)
- RefGene (2)
- NCBI-PROTEINID (1)
- NCBI-Gene (1)
- HGNC (1)
- Ensembl (1)
- RIKEN BRC-DNA (2)
- OC (1)
- PHAROS (1)
- PRONIT (28)
- VEGA (1)

Protein sequence (27)
- UniProt (2)
- SWISS-PROT (1)
- RefSeq(pep) (24)

DNA sequence (224)
- RefSeq(nuc) (26)
- GenBank (99)
- EMBL (99)

Protein domain (4)
- Pfam (4)

All databases (398)

Download RDF

OMIM

ICD+

114480

BREAST CANCER

Alternative titles; symbols

BREAST CANCER, FAMILIAL

Other entities represented in this entry:

BREAST CANCER, FAMILIAL MALE, INCLUDED

Phenotype-Gene Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key	Gene/Locus	Gene/Locus MIM number
1p34.1	{Breast cancer, invasive ductal}	114480	AD, SMu	3	RAD54L	603615
2q33.1	{Breast cancer, protection against}	114480	AD, SMu	3	CASP8	601763
2q35	{Breast cancer, susceptibility to}	114480	AD, SMu	3	BARD1	601593
3q26.32	Breast cancer, somatic	114480		3	PIK3CA	171834
5q34	{Breast cancer, susceptibility to}	114480	AD, SMu	3	HMMR	600936
6p25.2	{?Breast cancer susceptibility}	114480	AD, SMu	1	NQO2	160998
6q25.1-q25.2	Breast cancer, somatic	114480		3	ESR1	133430
8q11.23	Breast cancer, somatic	114480		3	RB1CC1	606837
11p15.4	Breast cancer, somatic	114480		3	SLC22A1L	602631
11q22.3	{Breast cancer, susceptibility to}	114480	AD, SMu	3	ATM	607585
12p12.1	Breast cancer, somatic	114480		3	KRAS	190070
13q13.1	{Breast cancer, male, susceptibility to}	114480	AD, SMu	3	BRCA2	600185
14q32.33	{Breast cancer, susceptibility to}	114480	AD, SMu	3	XRCC3	600675
14q32.33	Breast cancer, somatic	114480		3	AKT1	164730
15q15.1	{Breast cancer, susceptibility to}	114480	AD, SMu	3	RAD51	179617
16p12.2	{Breast cancer, susceptibility to}	114480	AD, SMu	3	PALB2	610355
16q22.1	{Breast cancer, lobular}	114480	AD, SMu	3	CDH1	192090
17p13.1	Breast cancer, somatic	114480		3	TP53	191170
17q21.33	{Breast cancer, susceptibility to}	114480	AD, SMu	3	PHB	176705
17q23.2	Breast cancer, somatic	114480		3	PPM1D	605100
17q23.2	{Breast cancer, early-onset, susceptibility to}	114480	AD, SMu	3	BRIP1	605882
22q12.1	{Breast cancer, susceptibility to}	114480	AD, SMu	3	CHEK2	604373

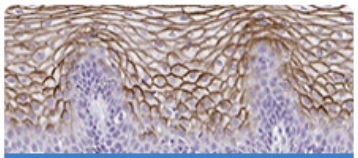
THE HUMAN PROTEIN ATLAS

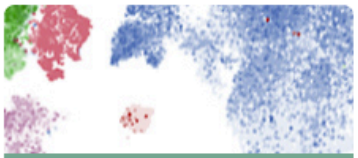
MENU HELP NEWS

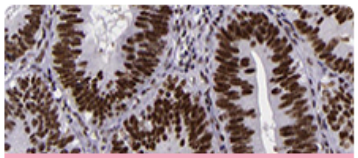
SEARCH

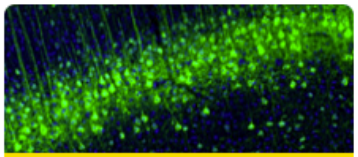
Search Fields »

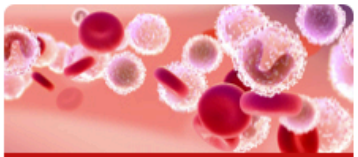
e.g. ACE2, GFAP, EGFR

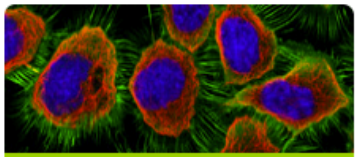

TISSUE ATLAS


SINGLE CELL TYPE ATLAS


PATHOLOGY ATLAS


BRAIN ATLAS


BLOOD ATLAS


CELL ATLAS

News

Movie of the month: The kidneys

The kidneys are responsible for filtering a mindblowing 180 liters of blood daily, eliminating toxins, regulating blood pressure and creating urine.... Read more

read the latest article - published Thu, 15 Apr 2021

Recent news

Thu, 15 Apr 2021
Movie of the month: The kidneys

Mon, 15 Mar 2021
Movie of the month: the nervous gut

Wed, 24 Feb 2021
Towards a Cell Cycle Atlas

all news articles

PRESS ROOM

contact@proteinatlas.org

INTRODUCTION
PUBLICATIONS
LICENCE & CITATION
DOWNLOADABLE DATA

Version: **20.1**
Atlas updated: 2021-02-24
release history

Proteome analysis based on
26941 antibodies targeting
17165 unique proteins

TISSUE ATLAS

PRIMARY DATA

TISSUES

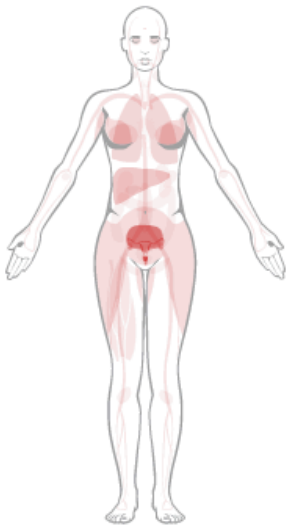
GENE/PROTEIN

ANTIBODIES AND VALIDATION

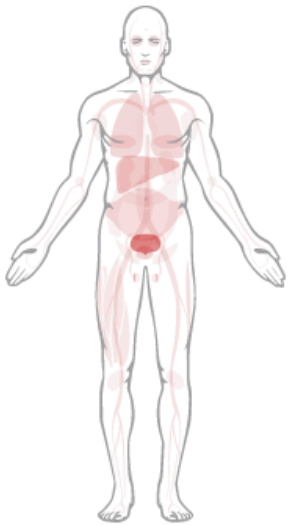
Dictionary

Tissue proteome

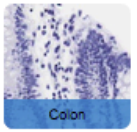
RNA AND PROTEIN EXPRESSION SUMMARY[i]

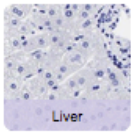


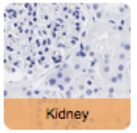
Expression | Detection | All organs

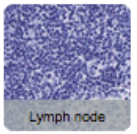


	RNA expression (NX)[i]	Protein expression (score)[i]
Brain		
Eye		
Endocrine tissues		
Lung		
Proximal digestive tract		
Gastrointestinal tract		
Liver & gallbladder		
Pancreas		
Kidney & urinary bladder		
Male tissues		
Female tissues		
Muscle tissues		
Adipose & soft tissue		
Skin		
Bone marrow & lymphoid tissues		



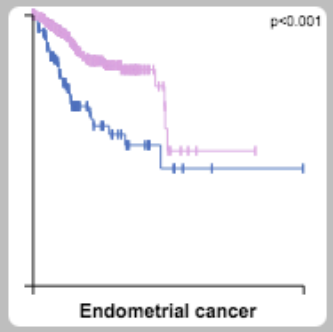






PROGNOSTIC SUMMARY[i]

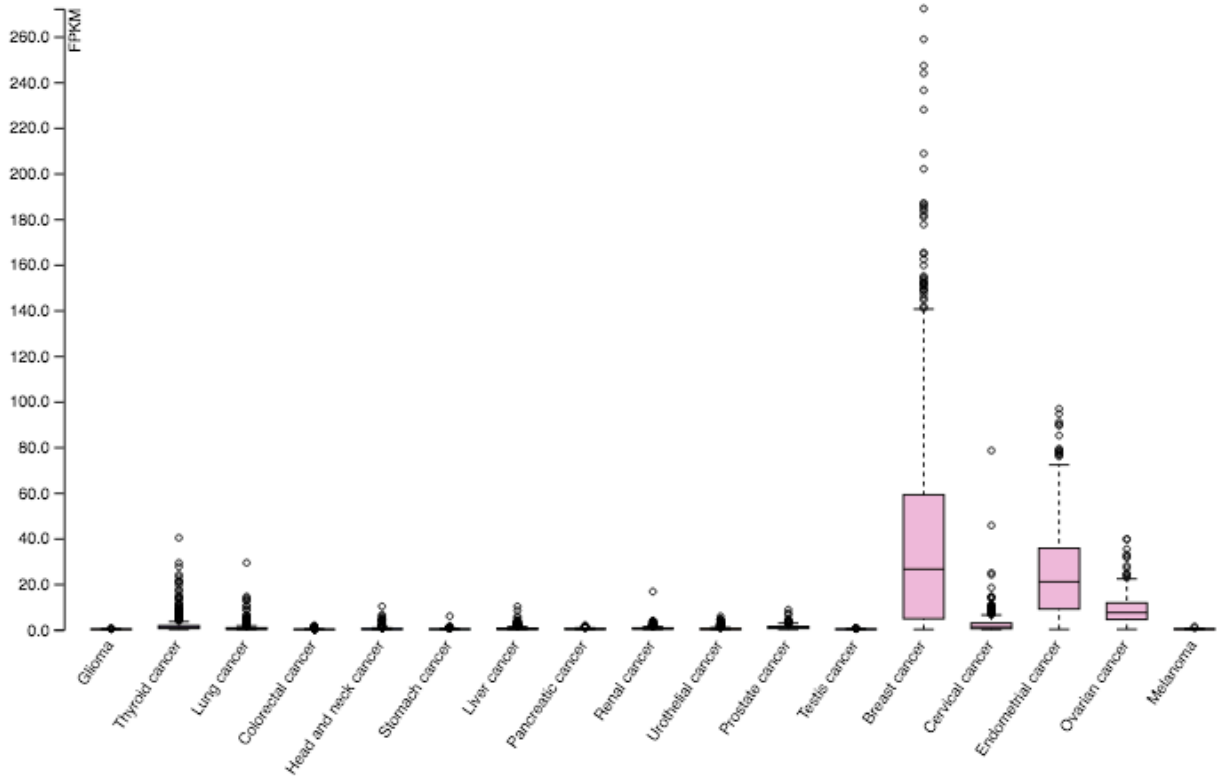
Prognostic marker in endometrial cancer (favorable)



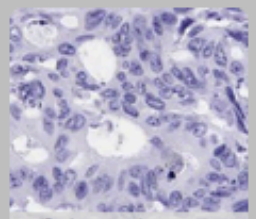
RNA EXPRESSION OVERVIEW[i]

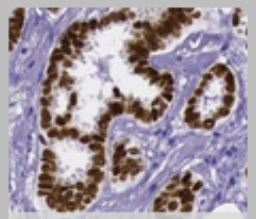
TCGA dataset[i]

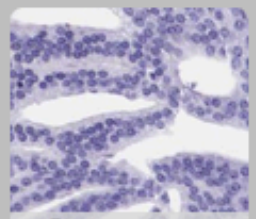
RNA cancer category: Group enriched (breast cancer, endometrial cancer, ovarian cancer)

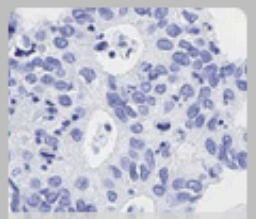


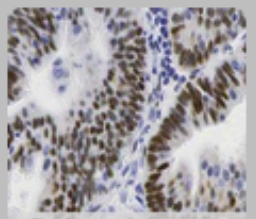
PROTEIN EXPRESSION[i]


Colorectal cancer


Breast cancer


Prostate cancer


Lung cancer

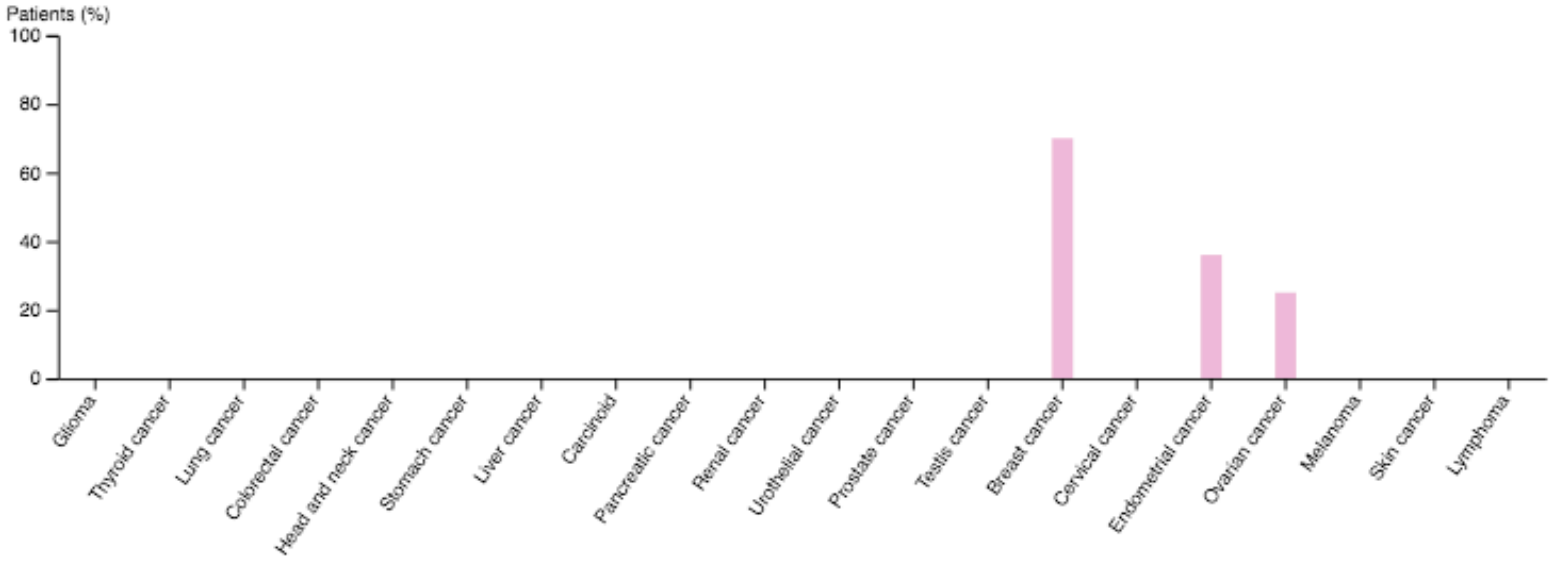

Endometrial cancer

PROTEIN EXPRESSION SUMMARY[i]

Organ | Expression | Alphabetical

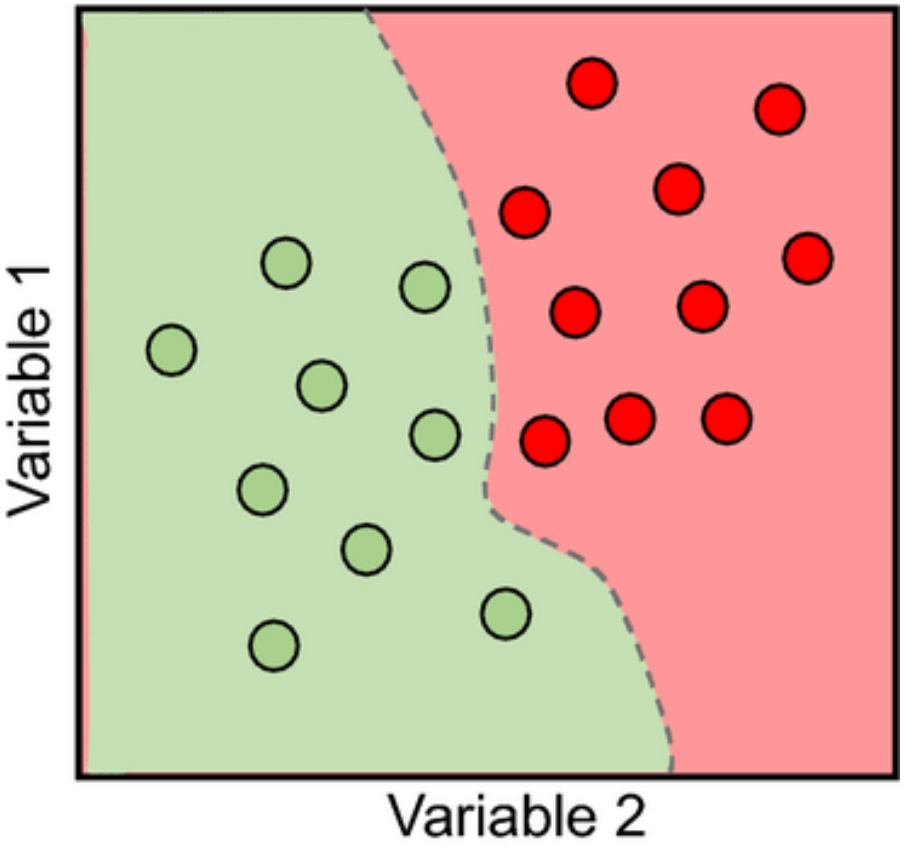
HPA000449 HPA000450 **CAB000037** CAB055099 CAB072858

Brest, ovarian and endometrial cancers displayed strong nuclear positivity. Remaining cancers were negative.



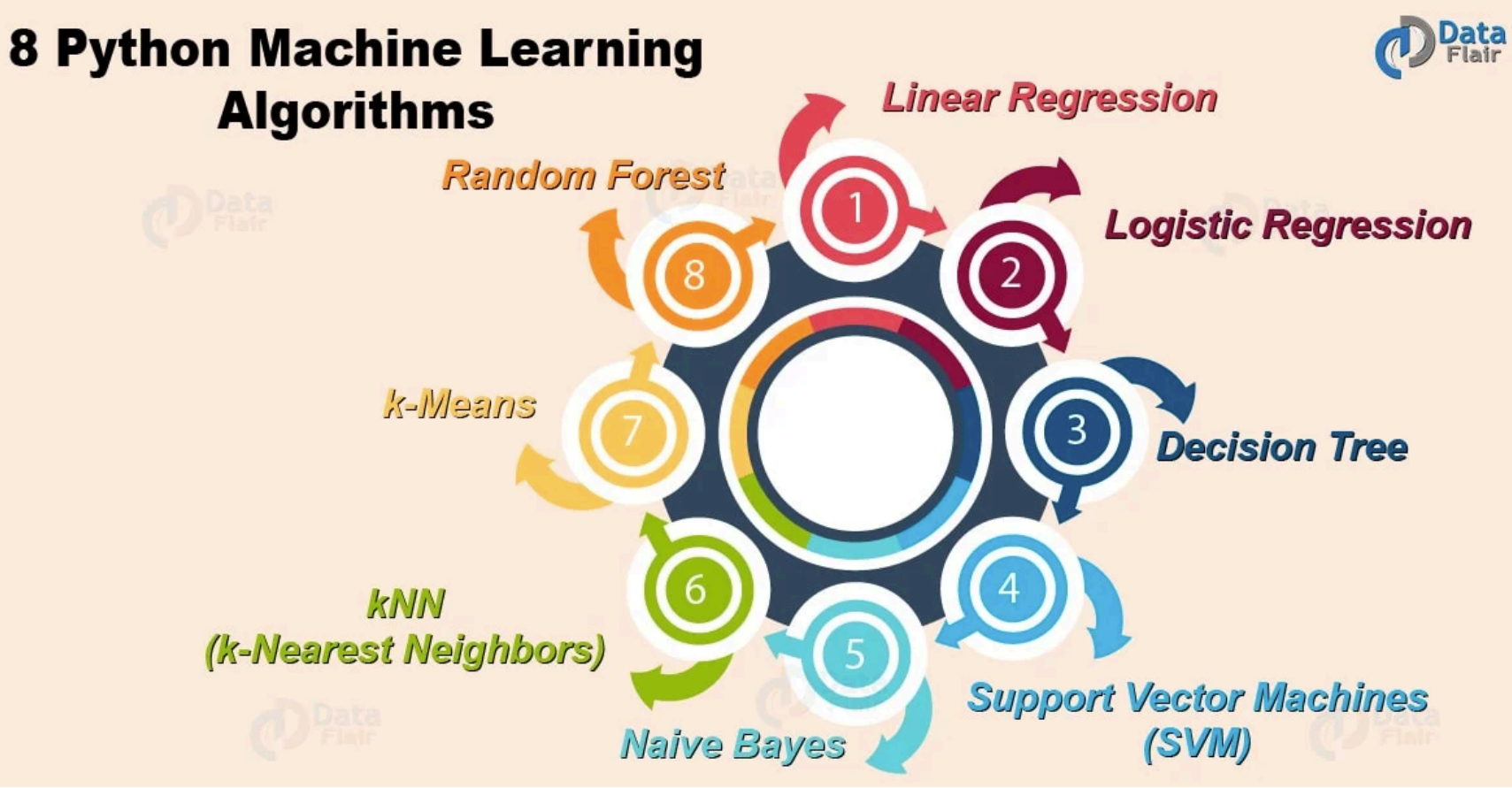
Μηχανική μάθηση – Machine Learning

- Supervised learning

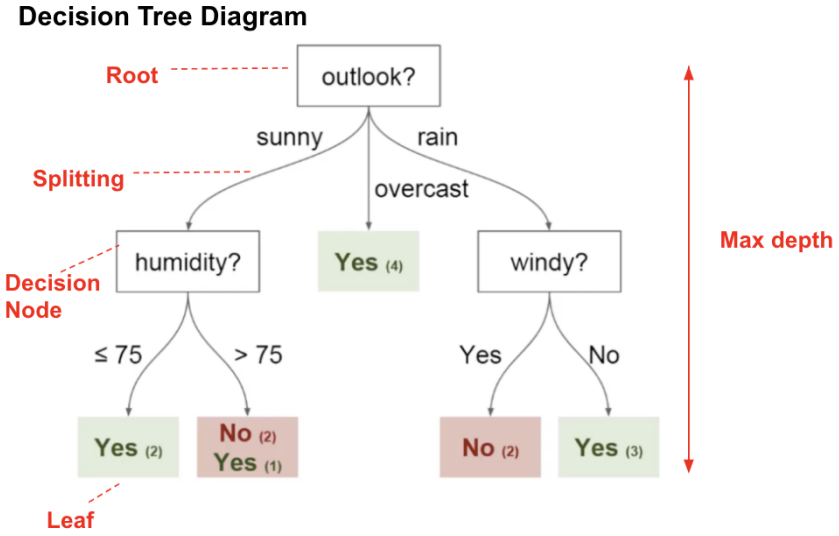


Μηχανική μάθηση – Machine Learning

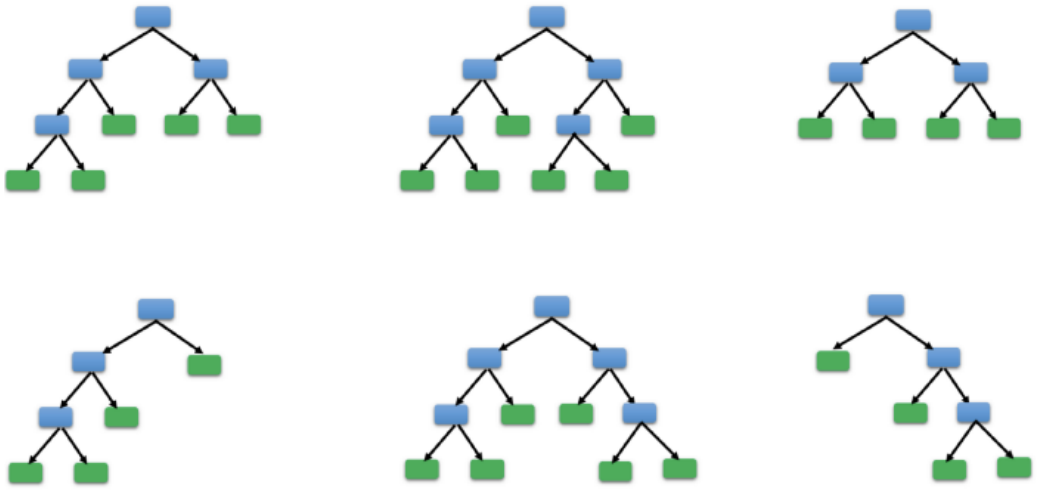
- Supervised learning



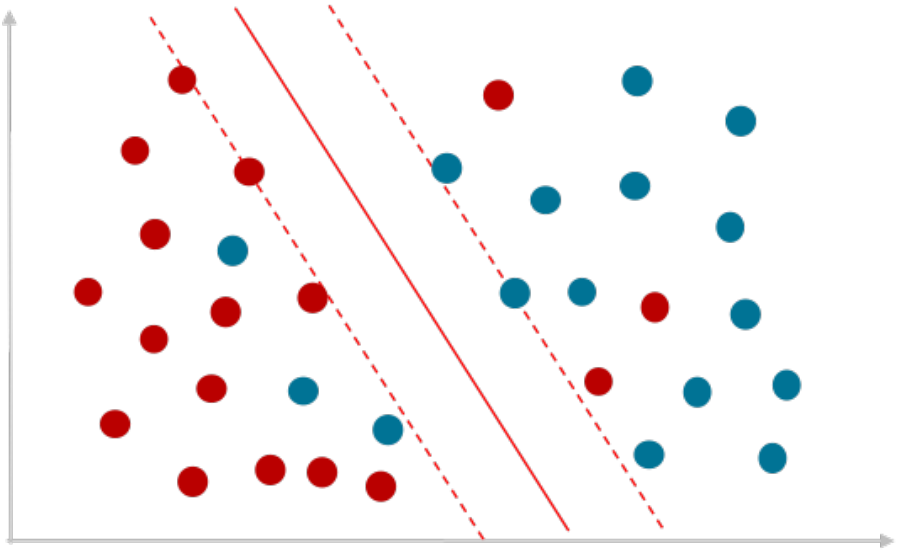
Decision Trees / Random Forests



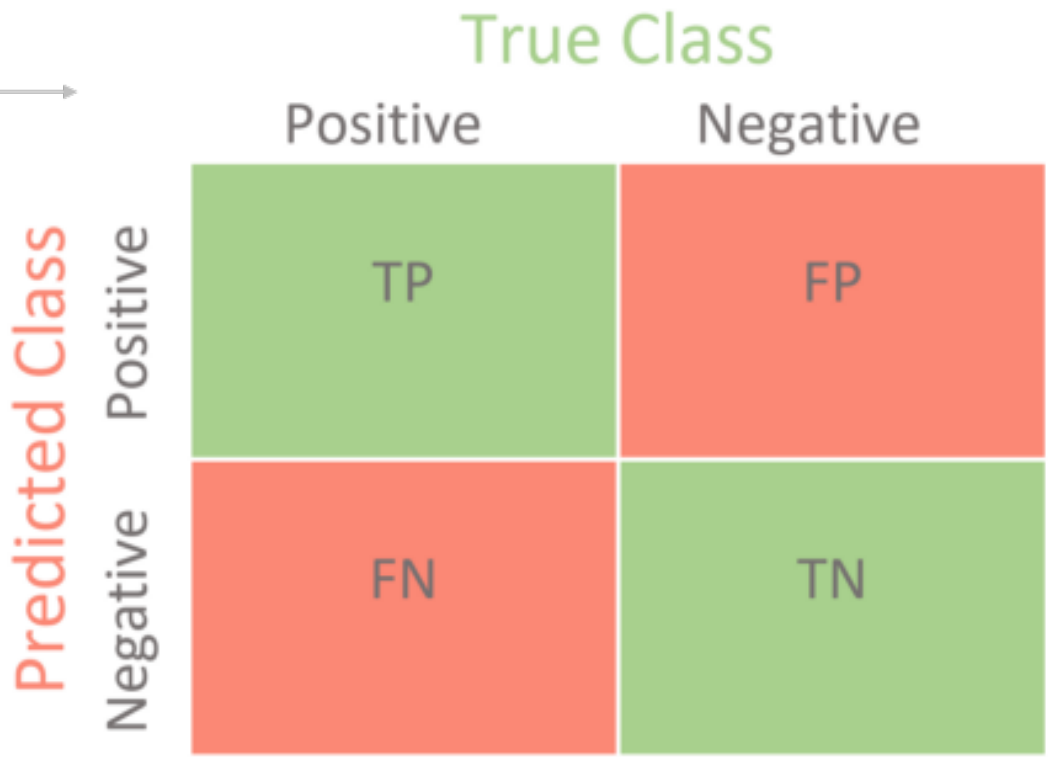
- Πολλοί μέτρια εκπαιδευμένοι κατηγοριοποιητές
- Η απόφαση βασίζεται στο σύνολο των επιμέρους αποφάσεων



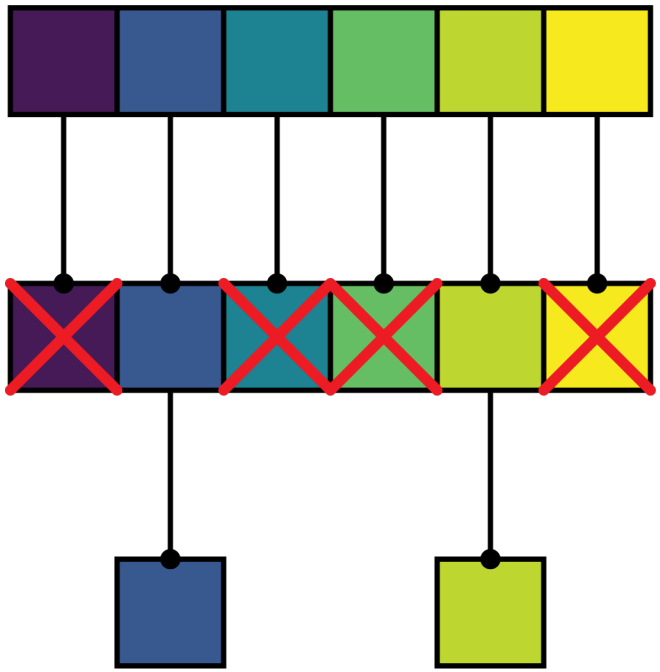
Απόδοση



- Confusion matrix
- Accuracy = (TP + TN)/(TP + TN + FP + FN)
- Precision
- Recalll
- Sensitivity
- Specificity
- F1-score

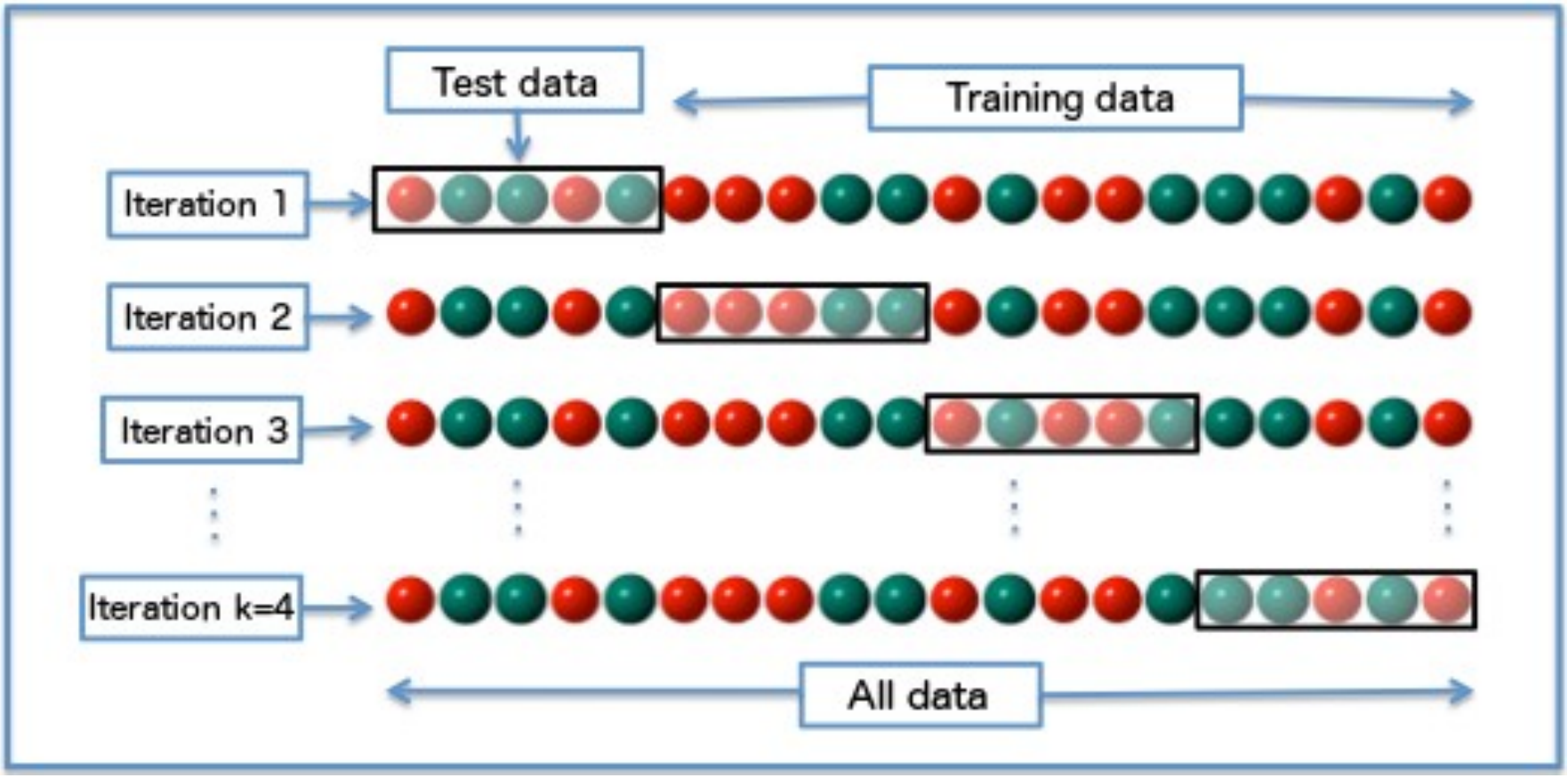


Μηχανική μάθηση – Feature selection

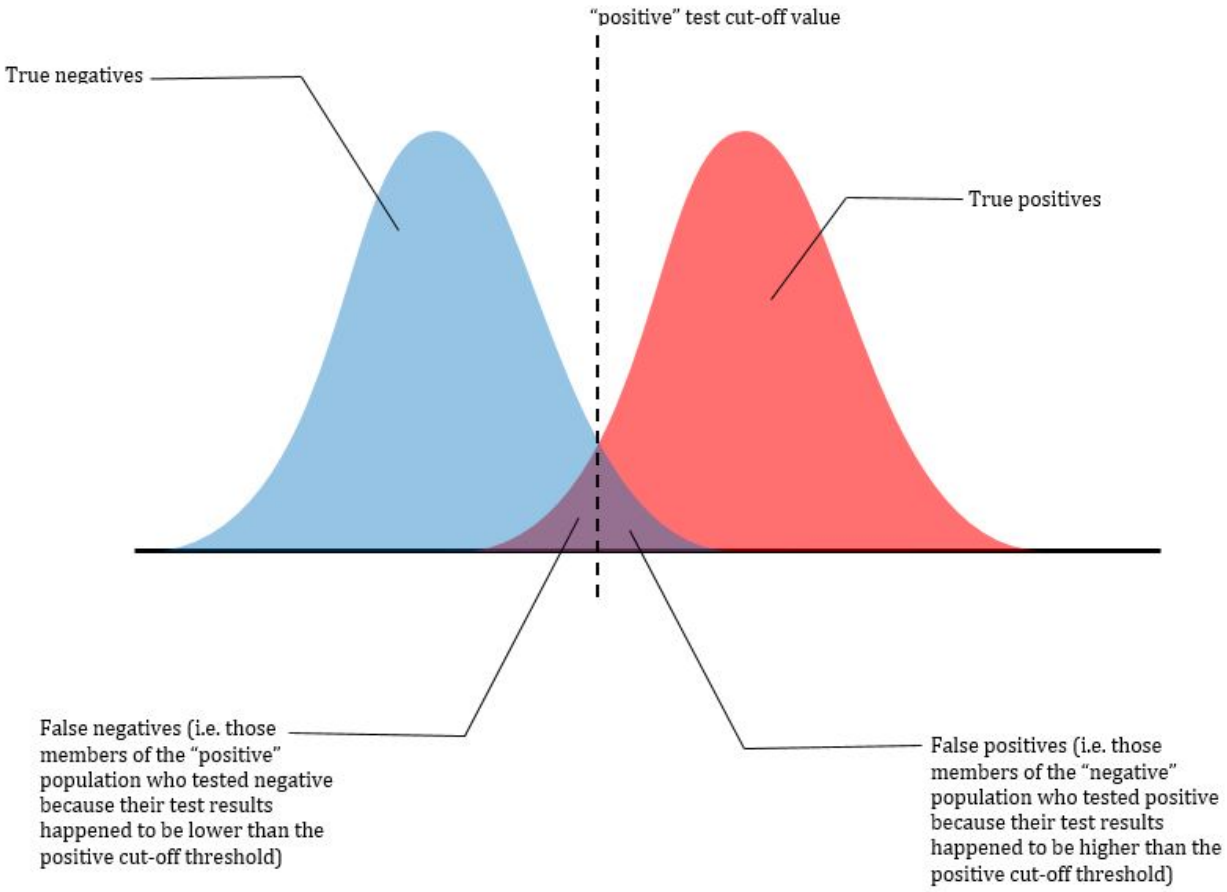


- Wrapper methods – δοκιμάζει πολλά διαφορετικά υποσύνολα και κρατάει αυτό που έχει την καλύτερη απόδοση – Αργές
- Filter methods – ελέγχουν με στατιστικές μεθόδους το κάθε ένα χαρακτηριστικό ξεχωριστά πόση πληροφορία περιέχει - Γρήγορες

Αντεπικύρωση – Cross-validation



Μηχανική μάθηση – Machine Learning



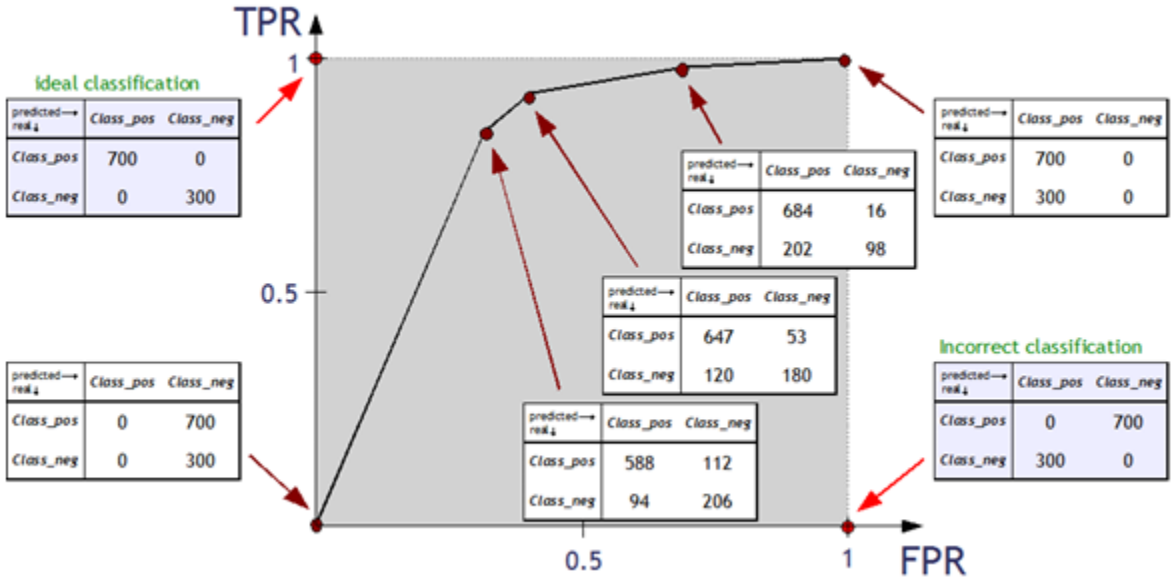
Μηχανική μάθηση – ROC curve

From Confusion Matrix to ROC Curve

Actual Class \ Predicted Class	Yes	No
Yes	True Pos (Hit)	False Neg (Type I Error)
No	False Pos (Type II Error)	True Neg (Correct Rejection)

$$TPR = \frac{TP}{TP + FN}$$

→ Y Axis on ROC Curve



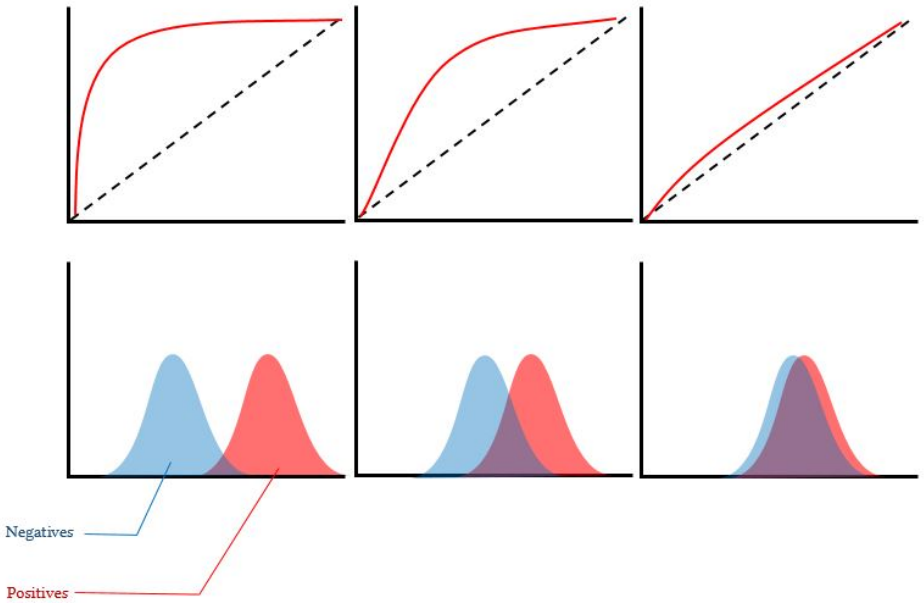
From Confusion Matrix to ROC Curve

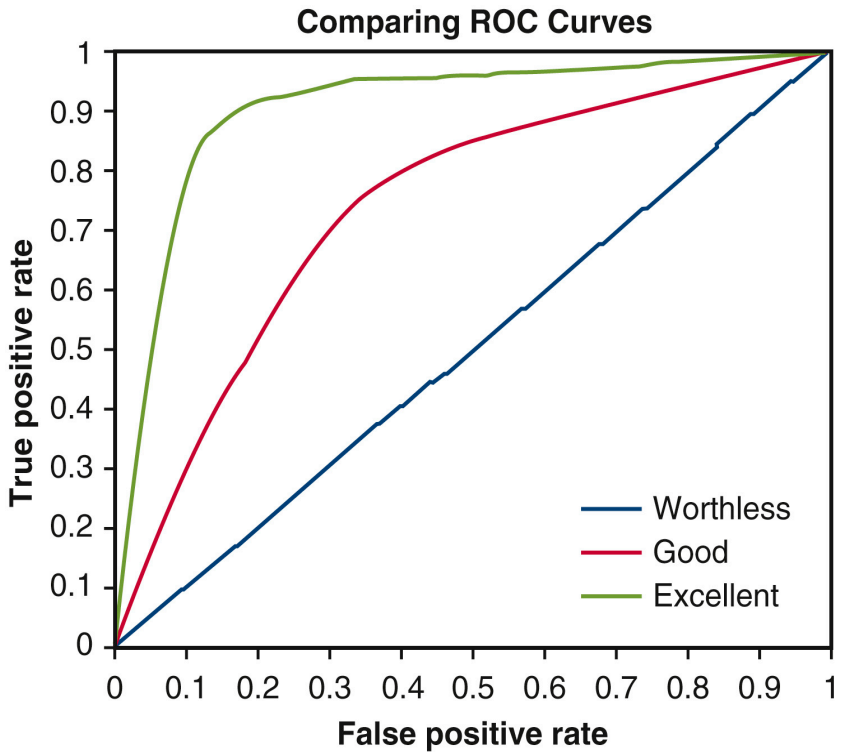
Actual Class \ Predicted Class	Yes	No
Yes	True Pos (Hit)	False Neg (Type I Error)
No	False Pos (Type II Error)	True Neg (Correct Rejection)

$$FPR = \frac{FP}{FP + TN}$$

→ X Axis on ROC Curve

Μηχανική μάθηση – ROC curve





In vitro διαγνωστικά τεστ που βασίζονται σε μικροσυστοιχίες

FDA: In Vitro Diagnostic Multivariate Index Assays (IVDMIAs)

- FDA's **In Vitro Diagnostic Product Database**
- http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfivd/index.cfm

- http://www.ivdtechnology.com/article/exploring-fda-approved-ivdmias

- Some IVDMIAs are laboratory-developed tests (LDTs). LDTs are tests that are developed by a single clinical laboratory for use only in that laboratory.

- http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm079148.htm

- IVDMIAs raise significant issues of safety and effectiveness. These types of tests are developed based on observed correlations between multivariate data and clinical outcome, such that the clinical validity of the claims is not transparent to patients, laboratorians, and clinicians who order these tests. Additionally, IVDMIAs frequently have a high risk intended use. FDA is concerned that patients are relying upon IVDMIAs with high risk intended uses to make critical healthcare decisions when FDA has not ensured that the IVDMIA has been clinically validated and the healthcare practitioners are unable to clinically validate the test themselves. Therefore, there is a need for FDA to regulate these devices to ensure that the IVDMIA is safe and effective for its intended use.

Mammaprint - Tissue of origin

- http://www.ivdtechnology.com/article/exploring-fda-approved-ivdmias
- **MammaPrint.**
  The first IVDMIA, the MammaPrint system, made by Agendia Inc., is a qualitative IVD test service performed in a single lab outside the United States using a 70-gene expression profile of fresh frozen breast cancer tissue samples to assess a breast cancer patient¬s risk for distant metastasis. FDA approved MammaPrint in February 2007 under de novo classification procedures.

- **Tissue of Origin Test**
  In July 2008, the Tissue of Origin Test, made by Pathwork Diagnostics, was cleared. This microarray RNA profiling test is to be used on clinical, formalin-fixed, paraffin-embedded (FFPE) biopsy tissue to aid in the classification of the origin of the tumor tissue. In June 2010 a second clearance introduced a different specimen and specimen-preparation method, and the algorithm for analysis of the expression data to create a diagnostics report and interpretation. The test uses microarray technology by Affymetrix Inc. and advanced analytics to measure the gene-expression patterns of challenging tumors, including metastatic, poorly differentiated, and undifferentiated cancer. It is intended to measure the degree of similarity between the RNA expression patterns in a patient's tumor tissue with the RNA expression patterns in a database of fifteen known tumor types.

Mammaprint


70 significant breast cancer prognosis genes
Tumor samples
78
70
60
50
40
30
20
10
Metastases: white = +
low risk
Threshold
high risk

Καρκίνοι αγνώστου προελεύσεως

- Σε κάποιες περιπτώσεις εμφάνισης/επανεμφάνισης καρκίνου είναι άγνωστη η πρωταρχική πηγή (ιστός), ακόμα και μετά από μια σειρά διαγνωστικών τεστ/βιοψία.
- Αυτό δεν επιτρέπει να χρησιμοποιηθεί ένα κατάλληλο θεραπευτικό σχήμα.
- Οι μικροσυστοιχίες επιτρέπουν να δημιουργηθεί το προφίλ γονιδιακής έκφρασης του συγκεκριμένου καρκίνου και να συγκριθεί με το προφίλ καρκίνων γνωστής προέλευσης.

Καρκίνοι αγνώστου προελεύσεως

- Δημιουργείται μια βάση από δεδομένα μεταγραφωμικής (από άλλες βάσεις δεδομένων και βιβλιογραφία).
- Τα δεδομένα είναι από γνωστούς καρκίνους, κανονικούς ιστούς, και από άλλες ασθένειες.
- Τα δεδομένα φιλτράρονται, κανονικοποιούνται.
- Στη συνέχεια γίνεται σύγκριση.

Καρκίνοι αγνώστου προελεύσεως

- http://genomemedicine.com/content/3/9/63/abstract
- **Classification of unknown primary tumors with a data-driven method based on a large microarray reference database**
- **Kalle A Ojala**, **Sami K Kilpinen** and **Olli P Kallioniemi**

IVDMIA - FDA

- http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2007/ucm108836.htm
- The MammaPrint is the first cleared in vitro diagnostic multivariate index assay (IVDMIA) device.

- http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2008/ucm116931.htm
- **FDA Clears Test that Helps Identify Type of Cancer in Tumor Sample**
- The Pathwork Tissue of Origin test compares the genetic material of a patient's tumor with genetic information on malignant tumor types stored in a database.It uses a microarray technology to analyze thousands of pieces of genetic material at one time. The test considers 15 common malignant tumor types, including bladder, breast, and colorectal tumors.

Preoperative prediction of ovarian cancer in patients presenting with pelvic masses

The FDA has cleared three IVDMIAs that employ measurements of serum biomarkers for the evaluation of women presenting with a pelvic mass:

- MIA or Ova1®,
- a second-generation multivariate index assay (MIA2G, Overa) and
- the Risk of Ovarian Malignancy Assay (ROMA).

Compared with measurement of CA-125 levels, an approach used by many clinicians to assess the malignant potential of adnexal masses despite the fact that the sole approved indication for the use of CA-125 is for the detection of ovarian cancer recurrence.

MIA2G Test

The MIA2G multivariate index assay combines the results of the biomarker concentrations from the Cobas assays for:

- apolipoprotein A-1 (APO-A1),
- cancer antigen 125 (CA 125-II),
- human epididymis protein 4 (HE4),
- follicle-stimulating hormone (FSH) and
- transferrin (TRF).

The MIA2G risk score was calculated using OvaCalc software version 4.0.0, which uses the five biomarker values and a proprietary algorithm to return a dimensionless numerical score from 0.0 to 10.0.

ROM:Risk Of Malignancy
SI: Symptoms Index

Gynecologic Oncology 2016 141DOI: (10.1016/j.ygyno.2016.04.459)

A value < 5 is considered low risk, and a value ≥ 5 is considered elevated risk, regardless of menopausal status.

ROMA Calculations

- ROMA was calculated from the same Cobas 6000 biomarker values used to calculate the MIA2G result. The calculation used to determine ROMA values was found in (Moore et al., 2009). The biomarkers used to calculate ROMA were the same CA 125 and HE4 values used in the MIA2G algorithm.

	CA125		ROMA		MIA2G	
	Pre	Post	Pre	Post	Pre	Post
	(*N* = 506)	(*N* = 487)	(*N* = 506)	(*N* = 487)	(*N* = 506)	(*N* = 487)
Sensitivity, %	50.7	79.7	78.1	79.7	90.4	91.3
n/N	37/73	137/172	57/73	137/172	66/73	157/172
95% CI	39.5–61.8	73.0–85.0	67.3–86.0	73.0–85.0	81.5–95.3	86.1–94.6
Specificity, %	95.8	81.3	76.2	82.5	70	59.4
n/N	415/433	256/315	330/433	260/315	303/433	187/315
95% CI	93.5–97.4	76.6–85.2	72.0–80.0	78.0–86.3	65.5–74.1	53.9–64.6
Positive likelihood ratio	12.193	4.253	3.282	4.562	3.011	2.246
95% CI	7.353–20.217	3.338–5.418	2.666–4.041	3.547–5.868	2.561–3.541	1.950–2.587
Negative likelihood ratio	0.515	0.25	0.288	0.247	0.137	0.147
95% CI	0.407–0.650	0.185–0.338	0.186–0.445	0.183–0.333	0.068–0.278	0.090–0.240
Pre-test odds of ovarian malignancy	0.17–1	0.55–1	0.17–1	0.55–1	0.17–1	0.55–1
Post-test odds of ovarian malignancy with high risk score	2.06–1	2.32–1	0.55–1	2.49–1	0.51–1	1.23–1
Post test odds of no ovarian malignancy with low risk score	11.53–1	7.31–1	20.63–1	7.43–1	43.29–1	12.47–1
Positive test rate (as overall %)	10.9	40.2	31.6	39.4	38.7	58.5
False positive rate (as overall %)	3.6	12.1	20.4	11.3	25.7	26.3
False negative rate (as overall %)	7.1	7.2	3.2	7.2	1.4	3.1

CA125 high-risk cutoff: premenopausal subjects > 200 U/ml, postmenopausal subjects > 35 U/ml; ROMA high-risk cutoff: premenopausal subjects ≥ 11.4, postmenopausal subjects ≥ 29.9

Open Access Article

Target Analysis of Volatile Organic Compounds in Exhaled Breath for Lung Cancer Discrimination from Other Pulmonary Diseases and Healthy Persons

by Michalis Koureas [1], Paraskevi Kirgou [2], Grigoris Amoutzias [3], Christos Hadjichristodoulou [1], Konstantinos Gourgoulianis [2] and Andreas Tsakalof [1,4,*]

[1] Department of Hygiene and Epidemiology, University Hospital of Larissa, Faculty of Medicine, University of Thessaly, 22 Papakyriazi Street, 41222 Larissa, Greece

[2] Respiratory Medicine Department, University Hospital of Larissa, Faculty of Medicine, University of Thessaly, 41500 Larissa, Greece

[3] Bioinformatics Laboratory, Department of Biochemistry and Biotechnology, University of Thessaly, 41500 Larissa, Greece

[4] Department of Biochemistry, Faculty of Medicine, University of Thessaly, 41500 Larissa, Greece

* Author to whom correspondence should be addressed.

Metabolites **2020**, *10*(8), 317; https://doi.org/10.3390/metabo10080317

Received: 29 May 2020 / Revised: 29 July 2020 / Accepted: 31 July 2020 / Published: 3 August 2020

(This article belongs to the Special Issue Volatile Metabolites' New Frontier for Metabolomics)

- The aim of the present study was to investigate the ability of breath analysis to distinguish lung cancer (LC) patients from patients with other respiratory diseases and healthy people.

- The population sample consisted of 51 patients with confirmed LC, 38 patients with pathological computed tomography (CT) findings not diagnosed with LC, and 53 healthy controls.

- The concentrations of 19 volatile organic compounds (VOCs) were quantified in the exhaled breath of study participants by solid phase microextraction (SPME) of the VOCs and subsequent gas chromatography-mass spectrometry (GC-MS) analysis.

- Kruskal–Wallis and Mann–Whitney tests were used to identify significant differences between subgroups.

- Machine learning methods were used to determine the discriminant power of the method. Several compounds were found to differ significantly between LC patients and healthy controls. Strong associations were identified for 2-propanol, 1-propanol, toluene, ethylbenzene, and styrene (p-values < 0.001–0.006)

- The random forest machine learning algorithm achieved a correct classification of patients of 88.5% (area under the curve—AUC 0.94).

- However, none of the methods used achieved adequate discrimination between LC patients and patients with abnormal computed tomography (CT) findings.