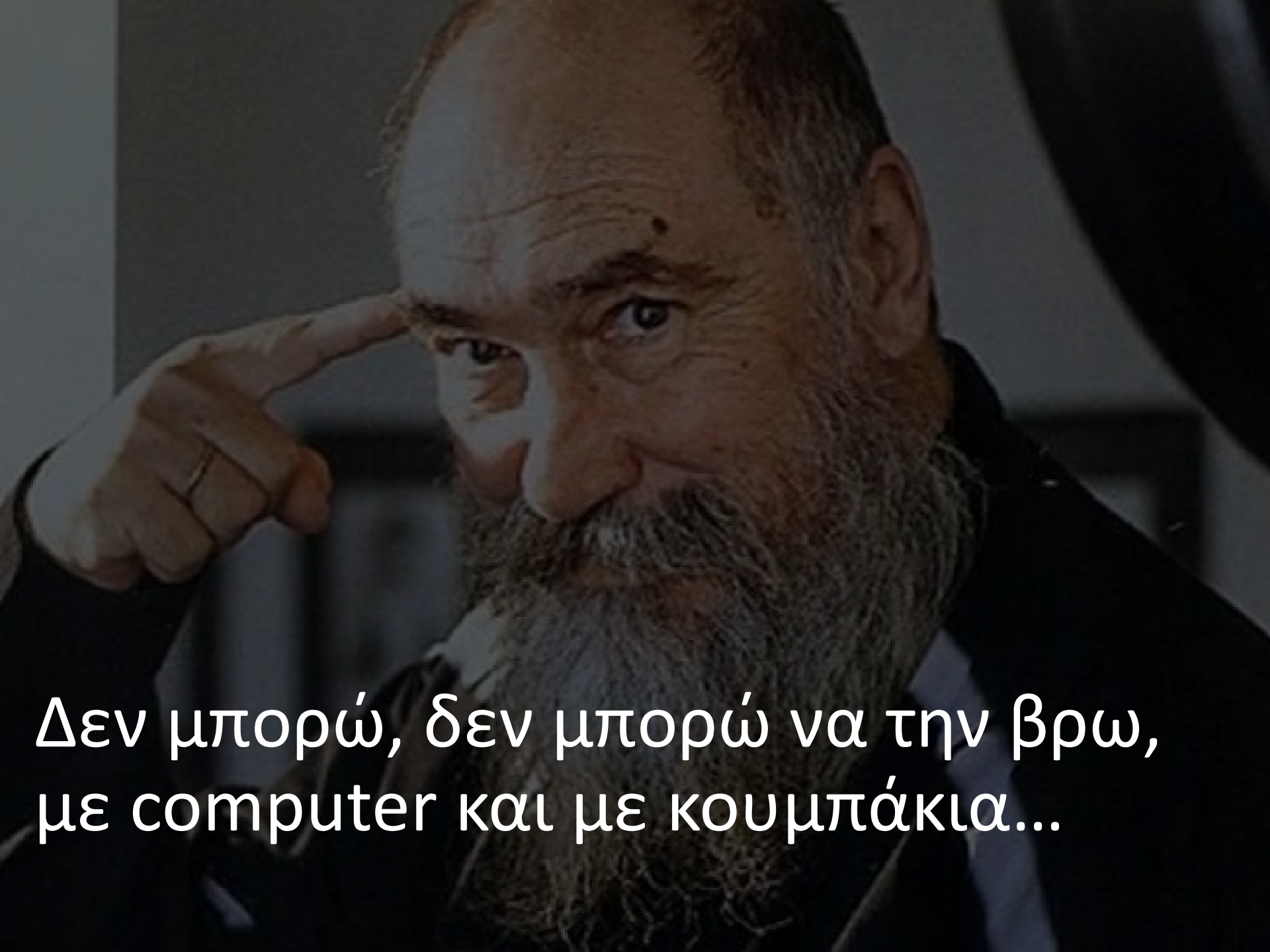


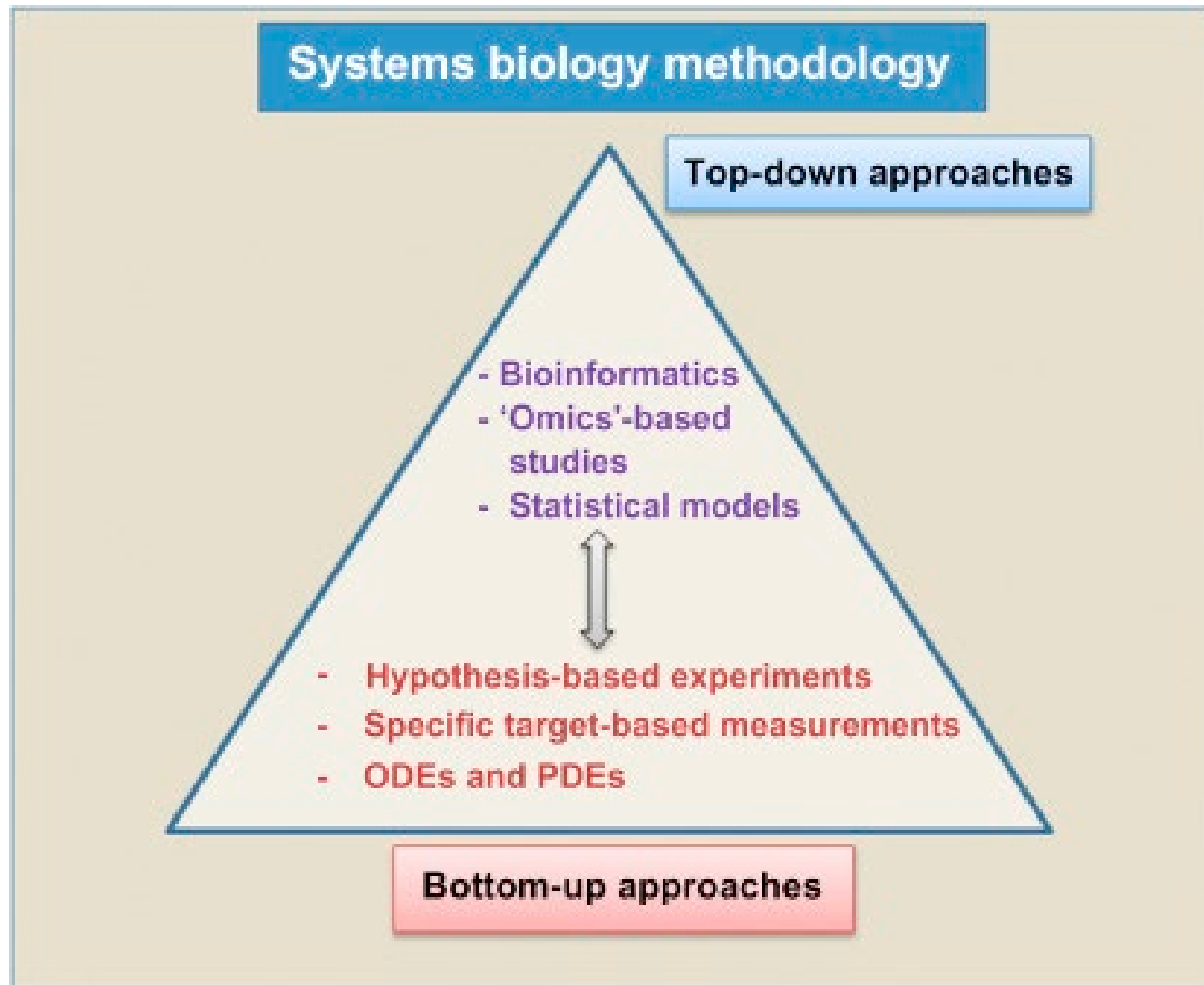
Βιολογία Συστημάτων

Βιοπληροφορική





Δεν μπορώ, δεν μπορώ να την βρω,
με computer και με κουμπάκια...



Η **Top-down** είναι «από τα δεδομένα προς τη θεωρία».

Η **Bottom-up** είναι «από τη θεωρία προς τα δεδομένα».

Η **Συστημική Βιολογία** χρησιμοποιεί και τις δύο για να συνδέσει την παρατήρηση με τη μηχανιστική κατανόηση.

-Omics: Ορισμός

- Επίθημα που καταδεικνύει την διεξοδική μελέτη βιολογικών συστημάτων και παραγόντων που μπορεί να εμπλέκονται στην παθογένεια της νόσου καθώς και στην συλλογή και ψηφιοποίηση βιολογικών δεδομένων ΜΕΓΑΛΗΣ ΚΛΙΜΑΚΑΣ
- Genomics, Transcriptomics, Metabolomics, Metagenomics κ.λπ.

BIG data

Big data is a term for data sets that are so large or complex that traditional data processing application software and hardware are inadequate to deal with them.



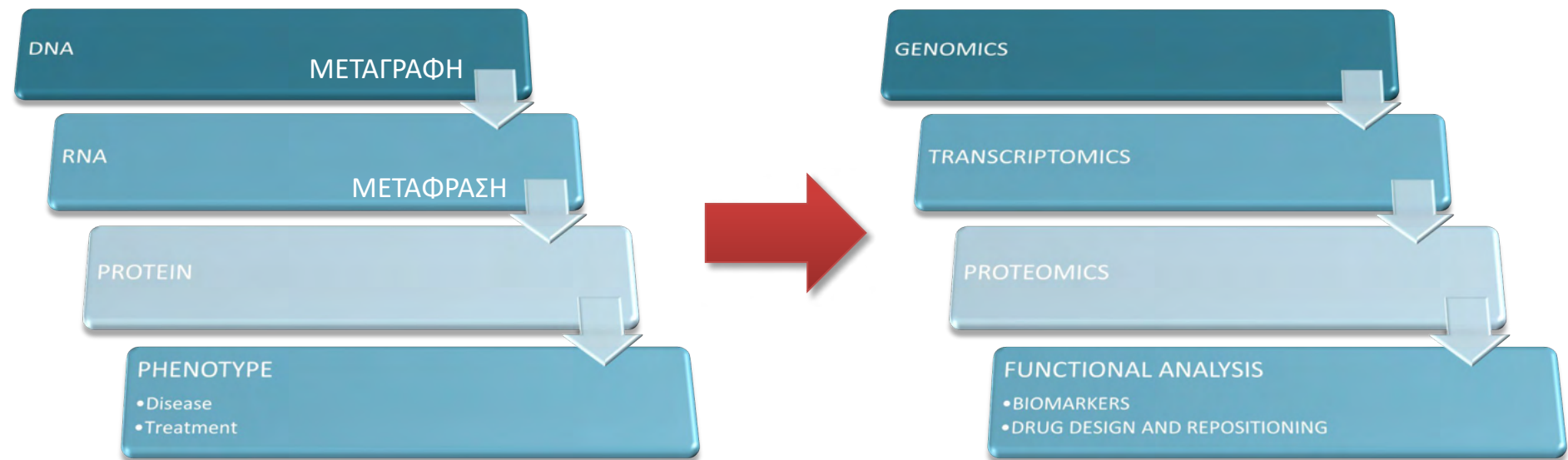
*everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it*

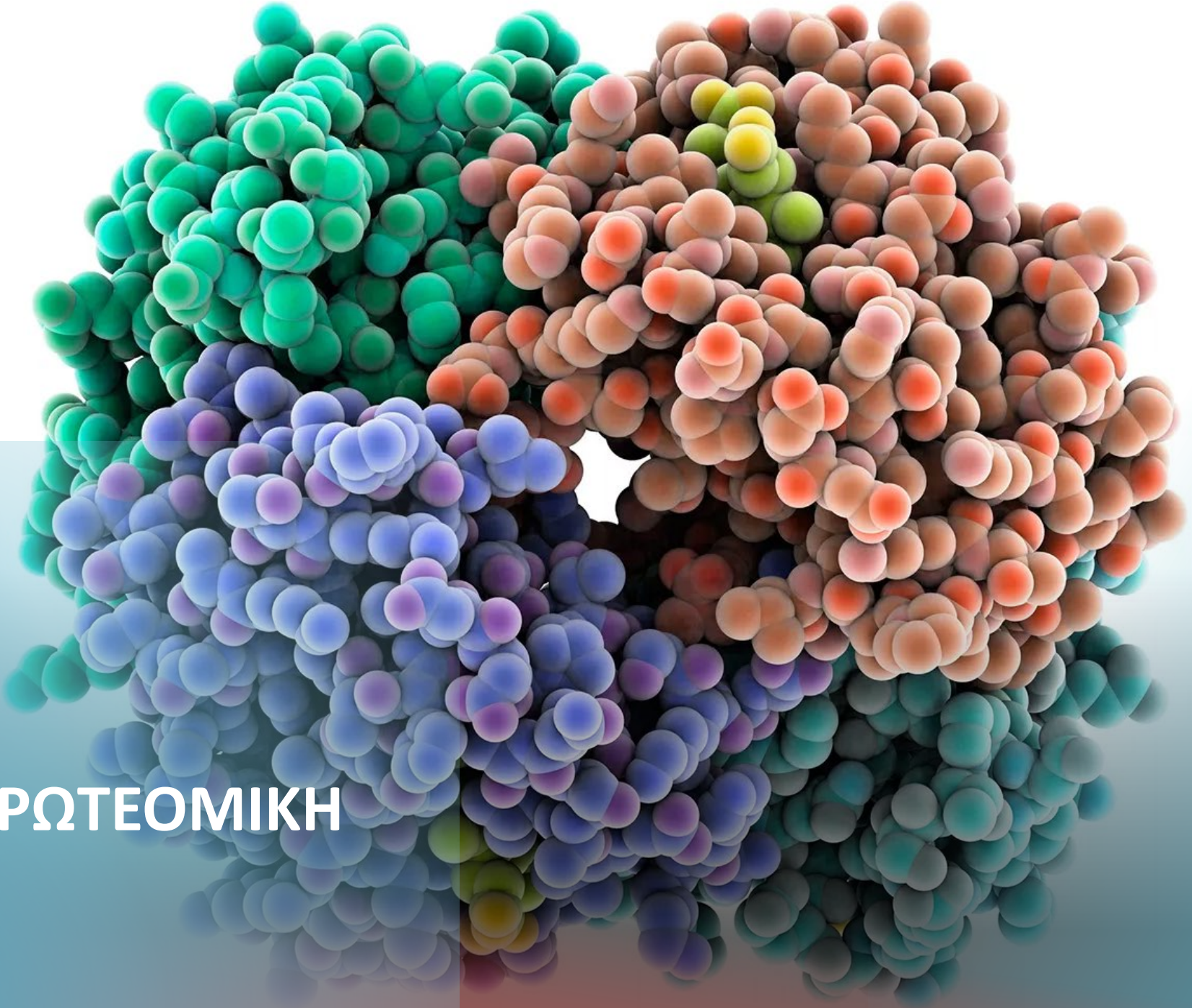


Τύπος Omics	Ορισμός	Αναμενόμενο αποτέλεσμα	Πιθανές εφαρμογές
Γονιδιωματική	Πολυεπιστημονική προσέγγιση της μελέτης, του ποσοτικού προσδιορισμού και του χαρακτηρισμού όλων των γονιδίων και των μεταλλάξεων τους σε ένα δείγμα	Αλληλούχηση DNA, ταυτοποίηση SNP	Προδιάθεση, Διάγνωση, Πρόγνωση, Σταδιοποίηση, Ταξινόμηση υπο-φαινοτύπων ασθενειών, Ανταπόκριση στη θεραπεία
Επιγονιδιωματική	Το επιστημονικό πεδίο που μελετά τις αλληλεπιδράσεις του περιβάλλοντος με το γονιδίωμα και τον τρόπο με τον οποίο αυτές μπορούν να ρυθμίζουν τη γονιδιακή έκφραση	Διαφορικά μοτίβα τροποποιήσεων DNA και ιστονών	
Μεταγραφομική	Το πεδίο της επιστήμης που μετρά την έκφραση του RNA και ανιχνεύει ποσοτικές διαφορές και τους τρόπους αλληλεπιδράσεων των γονιδίων	Διαφορική έκφραση RNA (mRNA, rRNA, miRNA, lncRNA, tRNA, snRNA)	
Πρωτεομική	Δίνει την ευκαιρία ανίχνευσης, ταυτοποίησης και χαρακτηρισμού ολόκληρης της πρωτεϊνικής έκφρασης ενός δεδομένου κυττάρου ή ιστού σε ευρεία κλίμακα	Ανίχνευση, ταυτοποίηση και χαρακτηρισμός των επιπέδων έκφρασης του πρωτεόματος	
Μεταβολομική	Η μελέτη των μεταβολικών διεργασιών και των αλλαγών στην παραγωγή μεταβολιτών σε έναν οργανισμό	Ανίχνευση και ταυτοποίηση της σύνθεσης μεταβολιτών	
Μικροβίωμα (μετα-γονιδιωματική, μετα-μεταβολομική, κλπ)	Η συνολική γενετική σύνθεση της μικροχλωρίδας σε ένα συγκεκριμένο όργανο	Ανίχνευση, ταυτοποίηση, χαρακτηρισμός και ποσοτικοποίηση της μικροβιακής σύνθεσης και του μεταβολικού της προφίλ	

Μονάδες ζωής σε ψηφιακή πληροφορία

How life can be translated into data



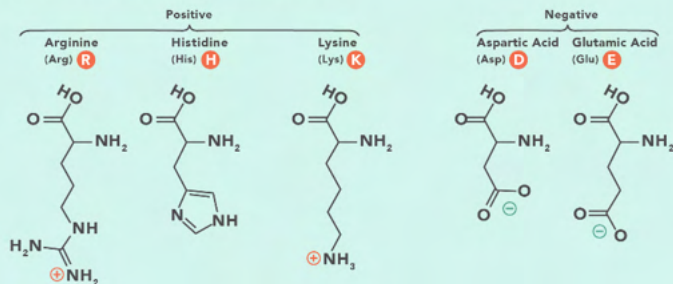


ΠΡΩΤΕΟΜΙΚΗ

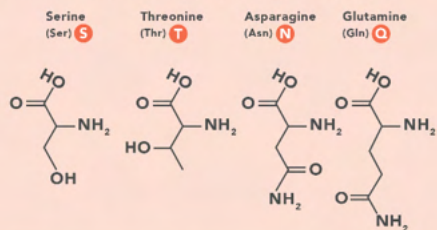
20 Αμινοξέα

64 Κωδικόνια

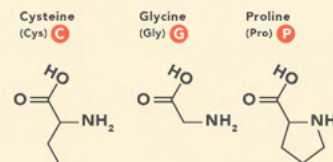
A. Amino Acids with Electrically Charged Side Chains



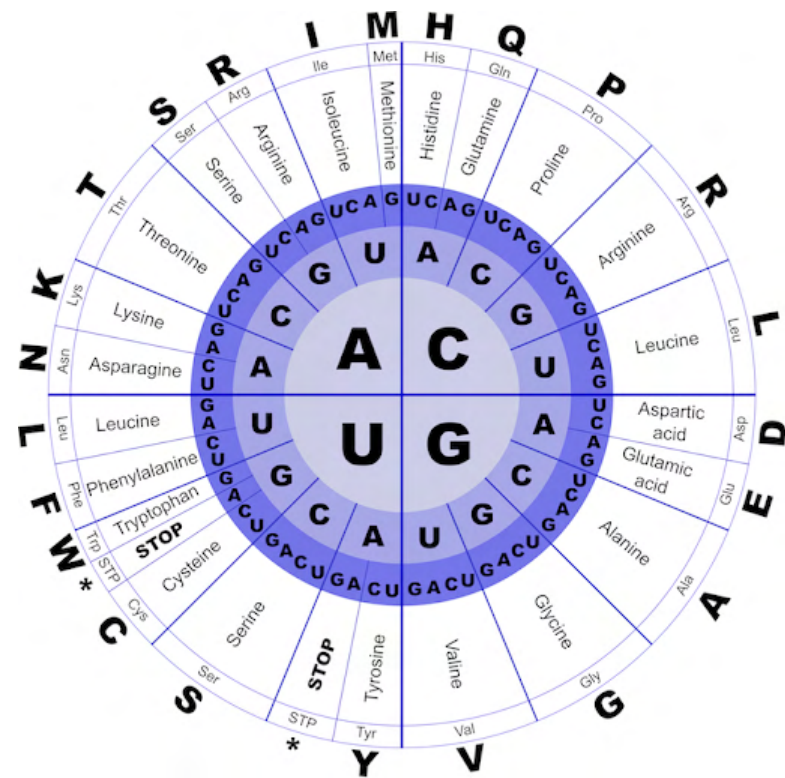
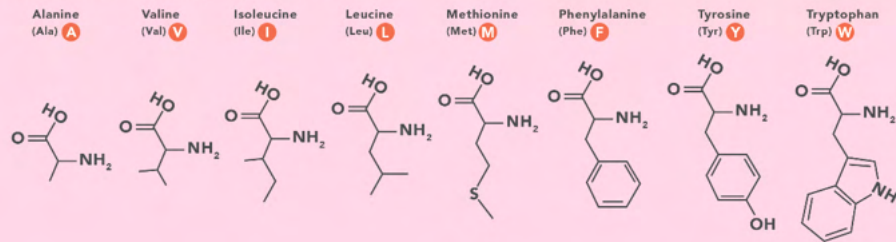
B. Amino Acids with Polar Uncharged Side Chains



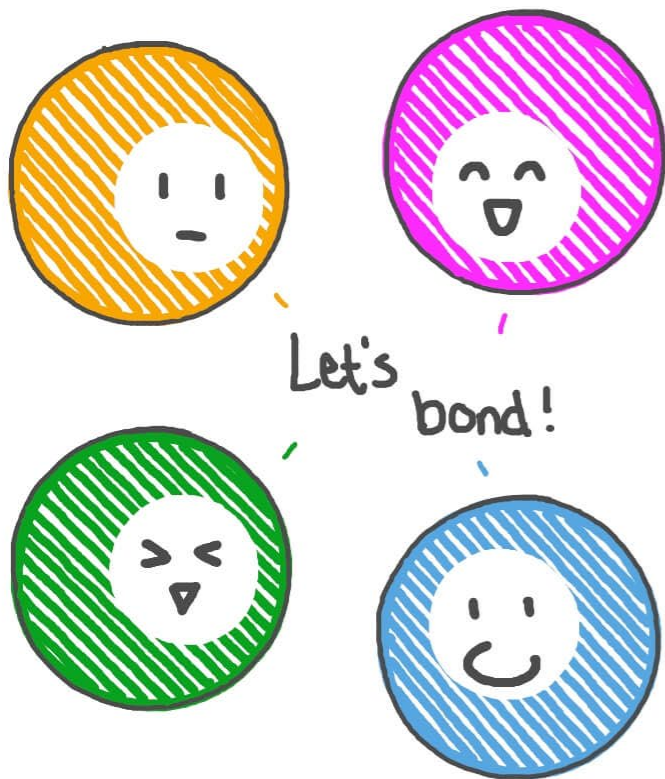
C. Special Cases



D. Amino Acids with Hydrophobic Side Chains



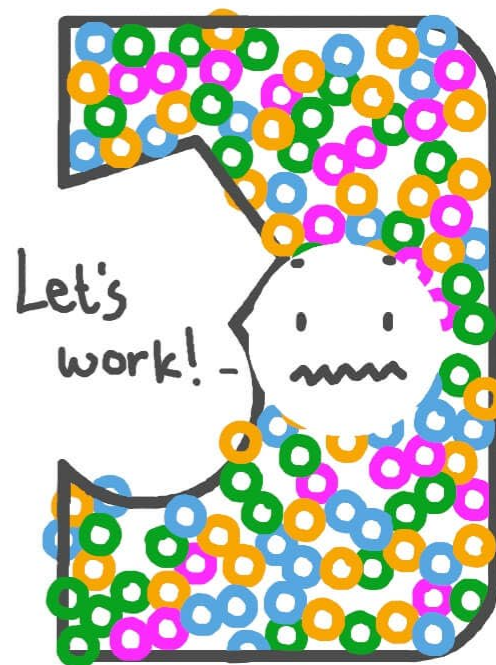
PROTEINS



amino acids



polypeptide



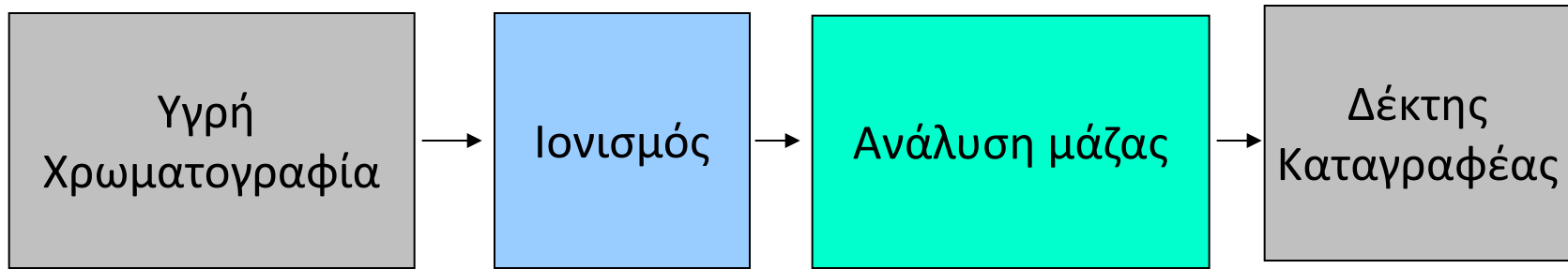
enzyme (protein)



LC-MS/MS είναι σαν ένα μηχάνημα που πρώτα **χωρίζει** τα μόρια (LC), μετά τα **ζυγίζει** και τα «**θρυμματίζει**» (MS/MS), και στο τέλος μας λέει **ποια περιέχονται μέσα** στο δείγμα και σε τι ποσότητα.



LC-MS Ανάλυση πρωτεϊνών



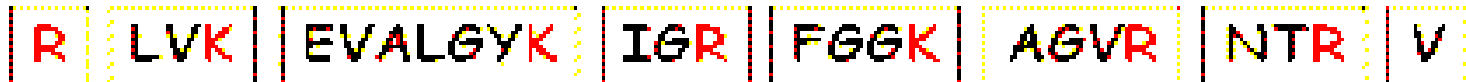
1. **Υγρή Χρωματογραφία (Liquid Chromatography):** Πρόκειται για την τεχνική που χρησιμοποιείται για τον διαχωρισμό των πρωτεϊνών ή των πεπτιδίων, βασιζόμενη σε φυσικοχημικές ιδιότητες όπως το μέγεθος, το φορτίο ή την υδροφοβικότητα. Τα συστατικά του δείγματος διαχωρίζονται μέσα σε μια στήλη που περιέχει υλικό διαχωρισμού.
2. **Ιονισμός (Ionization):** Τα πεπτίδια ή οι πρωτεΐνες που εξέρχονται από τη χρωματογραφία μετατρέπονται σε ιόντα. Τεχνικές όπως η ηλεκτροψεκασμός (Electrospray Ionization, ESI) ή η MALDI (Matrix-Assisted Laser Desorption/Ionization) χρησιμοποιούνται συνήθως για τη δημιουργία φορτισμένων μορίων.
3. **Ανάλυση Μάζας (Mass Analysis):** Η φάση αυτή περιλαμβάνει τη μέτρηση της μάζας των ιόντων. Οι μάζες ανιχνεύονται με φασματογράφους μάζας (Mass Spectrometers), οι οποίοι διαχωρίζουν τα ιόντα με βάση τον λόγο μάζας/φορτίου (m/z).
4. **Δέκτης Καταγραφέας (Detector/Recorder):** Τα δεδομένα που προκύπτουν καταγράφονται από τον δέκτη, ο οποίος αποτυπώνει τις τιμές των m/z . Αυτά τα δεδομένα χρησιμοποιούνται για την αναγνώριση και την ποσοτικοποίηση των πρωτεϊνών ή των πεπτιδίων.

Διάσπαση των πρωτεϊνών σε πεπτίδια

Peptide sequence



Σημεία Διάσπασης (Cleavage Sites):



Arginine: R

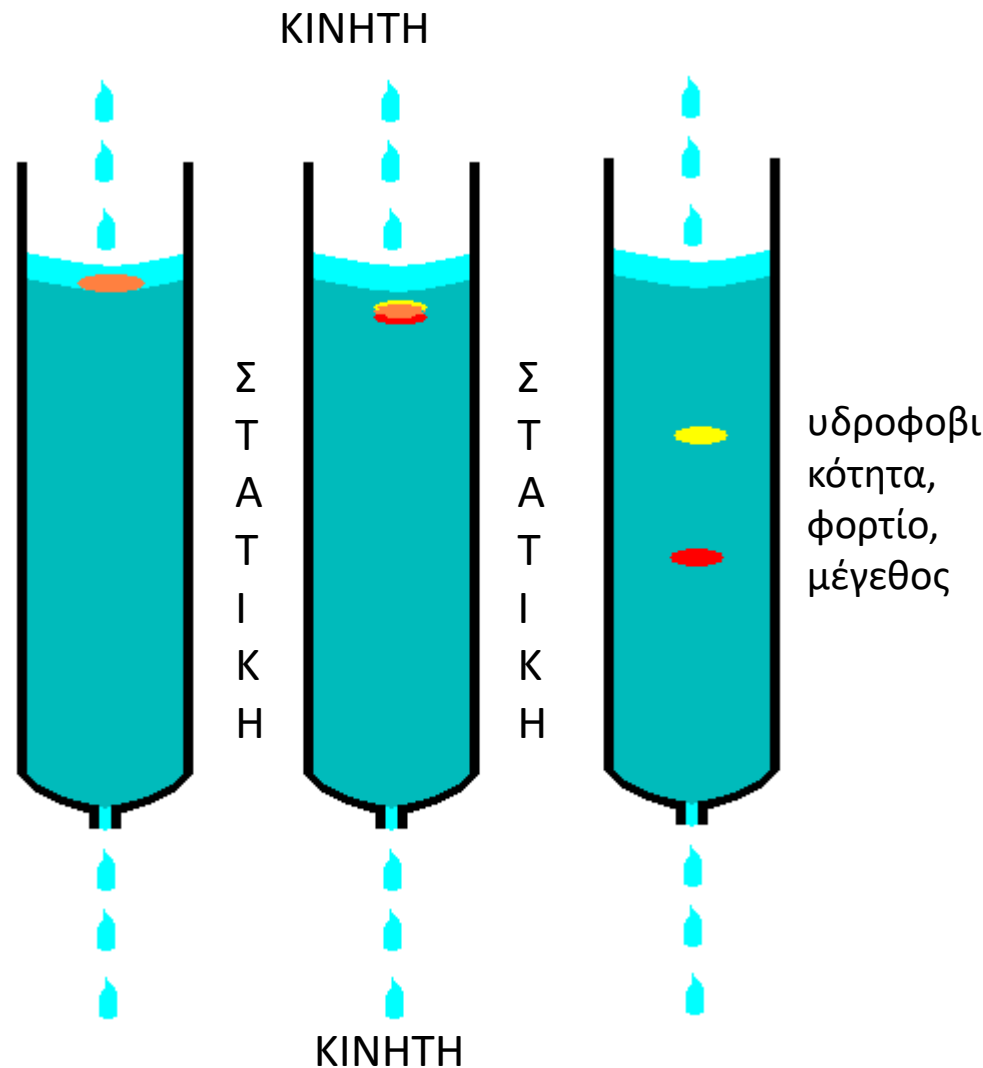
Lysine: K

Παραγόμενα Πεπτίδια

Η τρυψίνη είναι ένα πρωτεολυτικό ένζυμο που κόβει συγκεκριμένα μετά από αμινοξέα **αργινίνη (R)** και **λυσίνη (K)**, αρκεί να μην ακολουθείται από προλίνη (P).

χρωματογραφία

Η χρωματογραφία υψηλής πίεσης ανήκει στις χρωματογραφικές τεχνικές, άρα ο διαχωρισμός είναι αποτέλεσμα της συνδυαστικής δράσης μιας στατικής και μιας κινητής φάσης. Στην HPLC, το δείγμα εισάγεται στη κορυφή της στήλης και με τη βοήθεια της κινητής φάσης, τα συστατικά του μετακινούνται με τη μορφή ζωνών και τελικά εκκλύονται το ένα μετά το άλλο. Οι αναλυόμενες ουσίες κατανέμονται μεταξύ της στατικής και της κινητής φάσης, με αποτέλεσμα να μετακινούνται με διαφορετικές ταχύτητες κατά μήκος της στήλης



Κινητή φάση (Mobile phase)

- Είναι **υγρό που ρέει** μέσα από τη στήλη.
- Συνήθως **μίγμα νερού + οργανικού διαλύτη** (π.χ. ακτονιτρίλιο, μεθανόλη).
- Η σύστασή του **αλλάζει** κατά τη διάρκεια της ανάλυσης (gradient).

Παράδειγμα Reversed-Phase LC:

- Mobile phase A: νερό + 0.1% formic acid
- Mobile phase B: ακτονιτρίλιο + 0.1% formic acid
- Αυξάνουμε σταδιακά το B → αποκολλιούνται τα πιο υδρόφοβα μόρια.

Στατική φάση (Stationary phase)

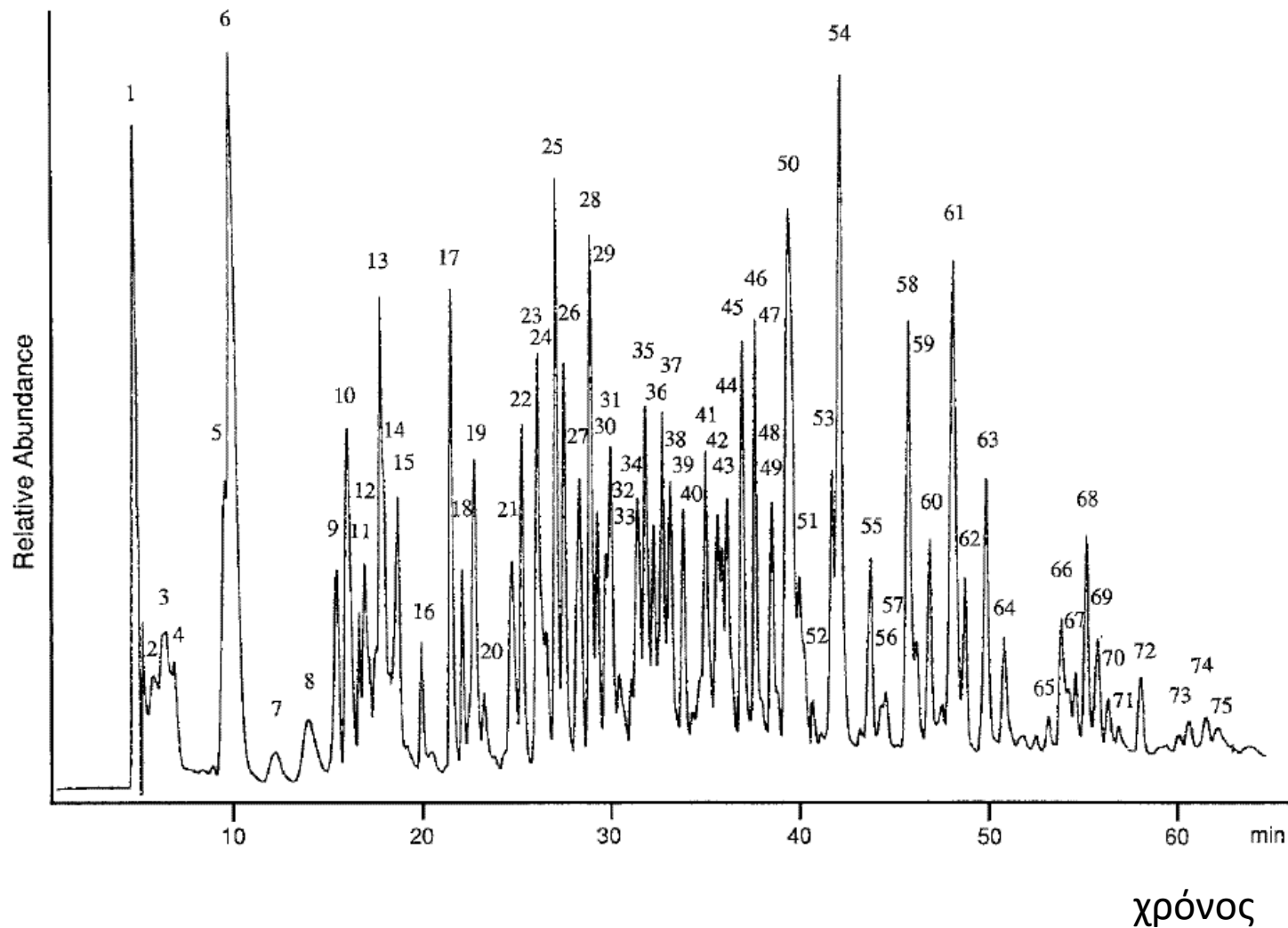
- Είναι **στερεό υλικό που μένει ακίνητο**.
- Αποτελείται από **μικροσκοπικά σφαιρίδια** (silica beads).
- Πάνω τους υπάρχουν χημικές ομάδες που **κρατούν προσωρινά** τα μόρια.

Παράδειγμα:

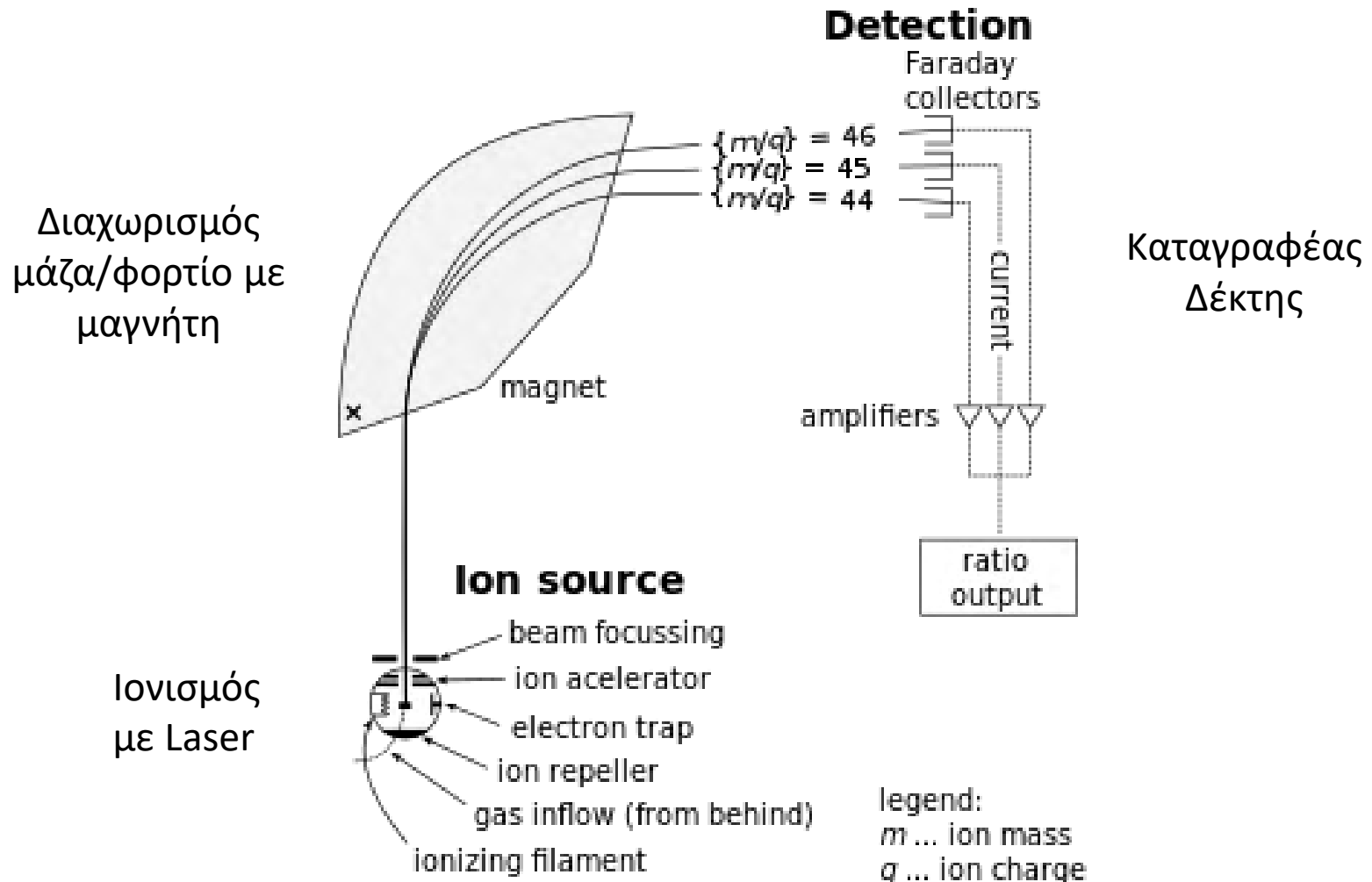
- C18 (υδρόφοβες αλυσίδες) για διαχωρισμό πεπτιδίων στην LC-MS/MS.



Αποτέλεσμα της HPLC η έξοδος σε διαφορετικό χρόνο των διαφορετικών πεπτιδίων

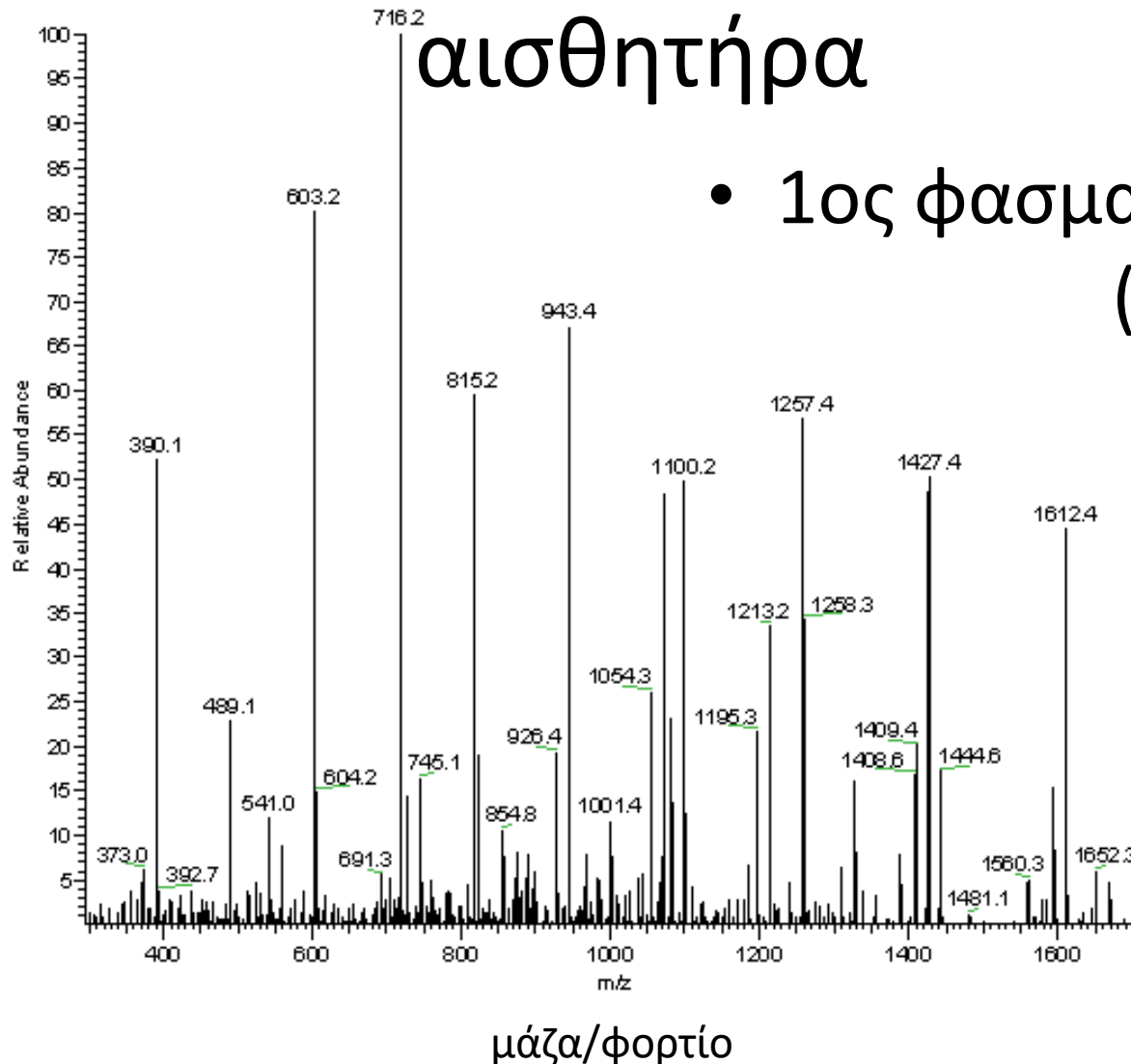


Ιονισμός και μαγνητικός διαχωρισμός των πεπτιδίων MS

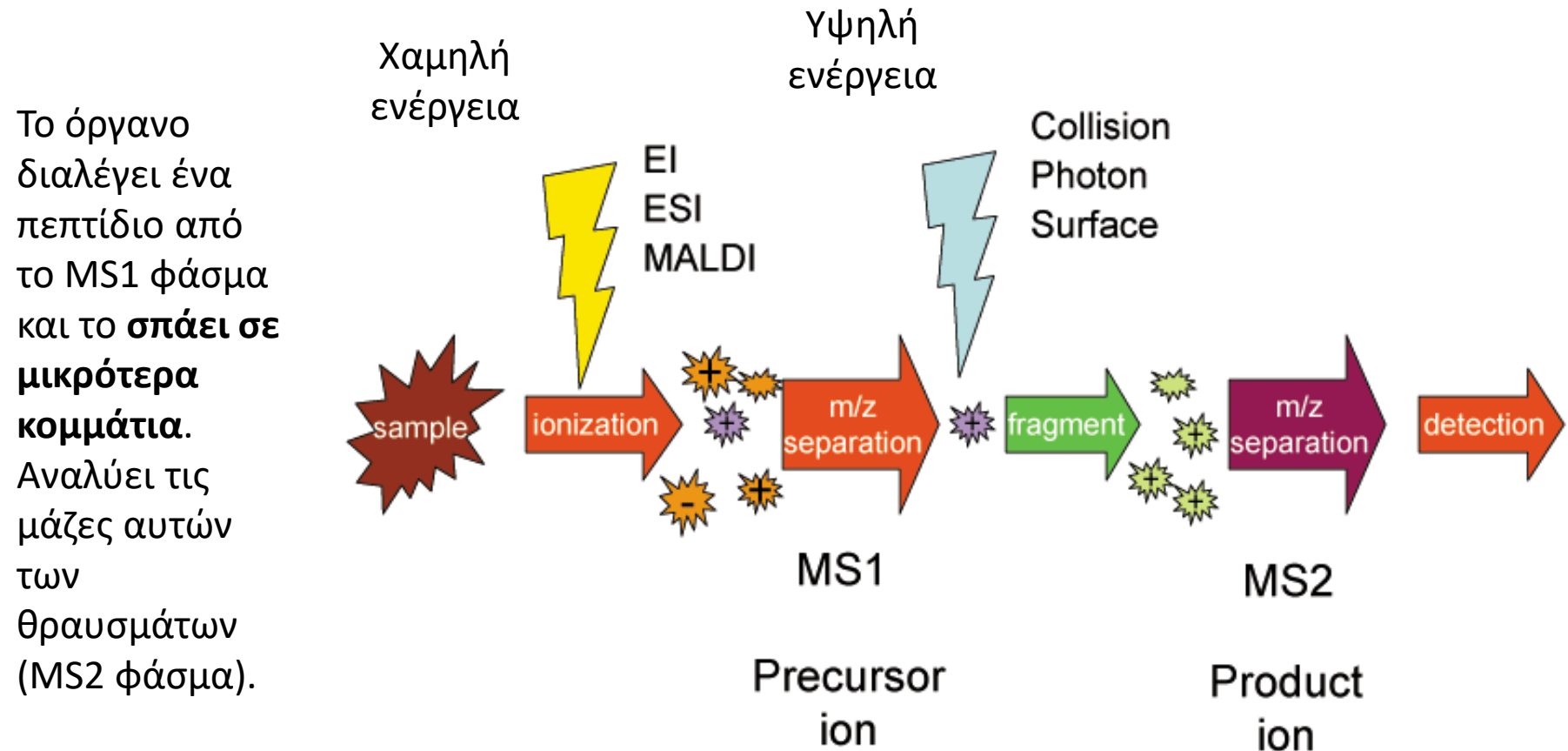


Τα πεπτίδια αναλύονται πάνω στον αισθητήρα

- 1ος φασματοσκόπος
(MS)



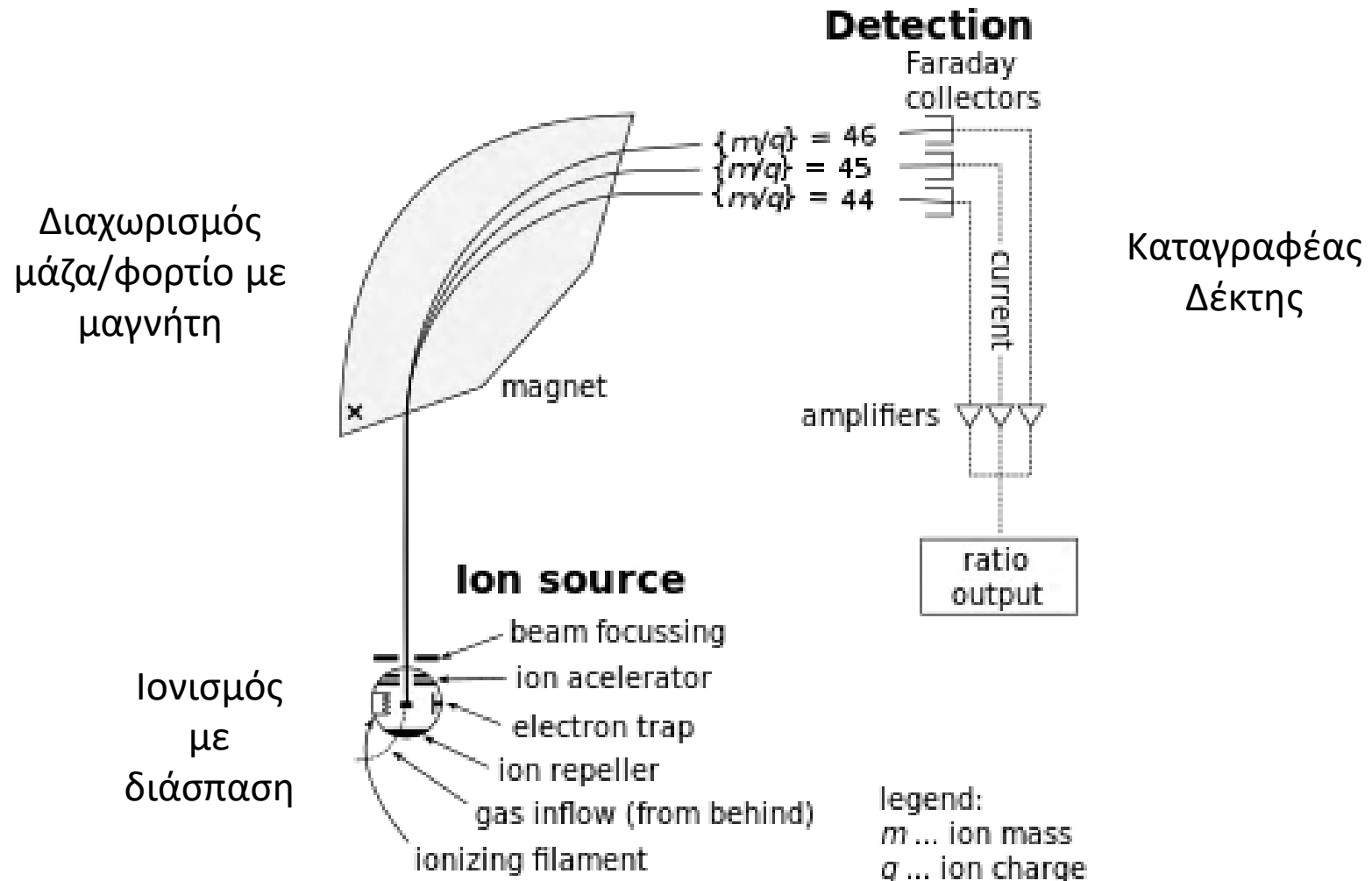
Δεύτερος ιονισμός με καταστροφή και μαγνητικός διαχωρισμός των πεπτιδίων MS/MS



MS1 (Πρόδρομα Ιόντα): Επιτρέπει την ανίχνευση των μορίων που αποτελούν στόχο.

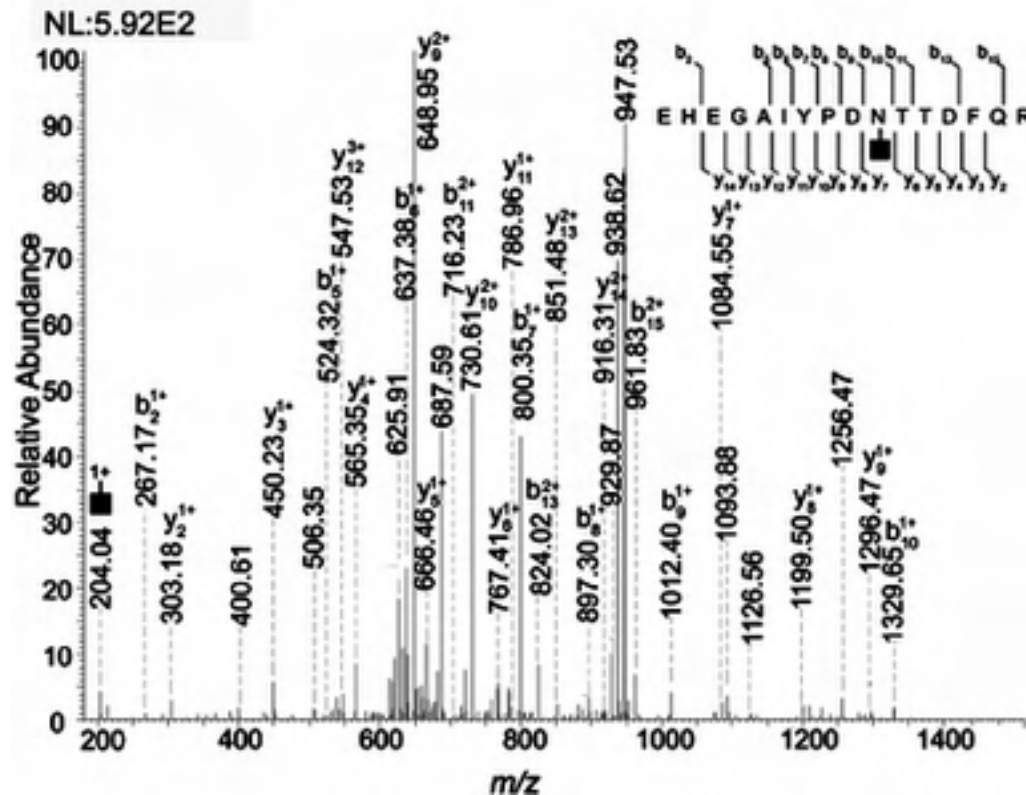
MS2 (Θραύσματα Ιόντων): Δίνει πληροφορίες για την αλληλουχία των πεπτιδίων, καθώς τα θραύσματα αντιστοιχούν σε συγκεκριμένα αμινοξέα.

Δεύτερος ιονισμός με καταστροφή και μαγνητικός διαχωρισμός των πεπτιδίων



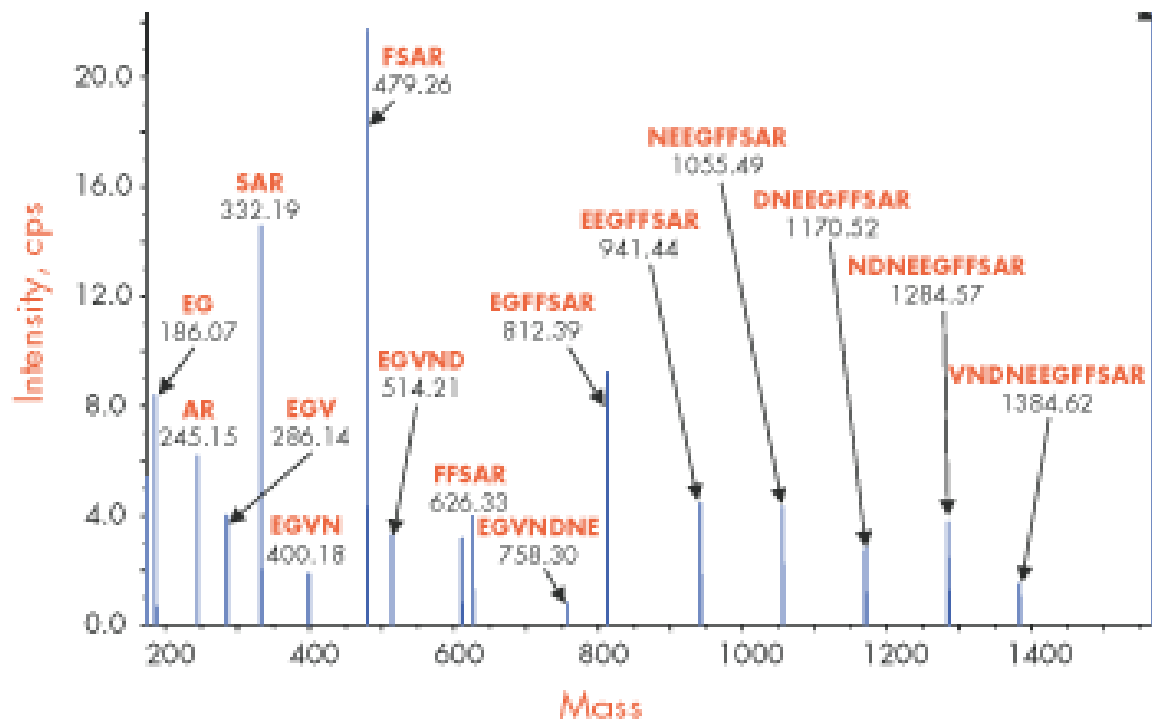
Ανάλυση των κατακερματισμένων πεπτιδίων δίνει μια υπογραφή

- 2ο MS



Μετά το δεύτερο MS μπορώ να αντιστοιχίσω
την κορυφή με συγκεκριμένη αλληλουχία
αμινοξέων

- 1ο MS με σημείωση των πεπτιδικών
αλληλουχιών από το 2^ο MS



Βιοπληροφορική στη πρωτεωμική

- Αναζητώντας το πεπτίδιο σε μια βάση δεδομένων μπορώ να το αντιστοιχίσω με την πρωτεΐνη από την οποία προήλθε
- Επίσης το ύψος της κορυφής μου δείχνει τη σχετική ποσότητα ενός πεπτιδίου και άρα μίας πρωτεΐνης (διαφορές στην έκφραση μεταξύ καταστάσεων)
- Το λογισμικό (π.χ. MaxQuant, Proteome Discoverer, Skyline κ.λπ.) συνδυάζει όλα τα φάσματα, τα συγκρίνει με βάσεις δεδομένων, και **αναγνωρίζει και ποσοτικοποιεί** τις πρωτεΐνες του δείγματος.



MaxQuant

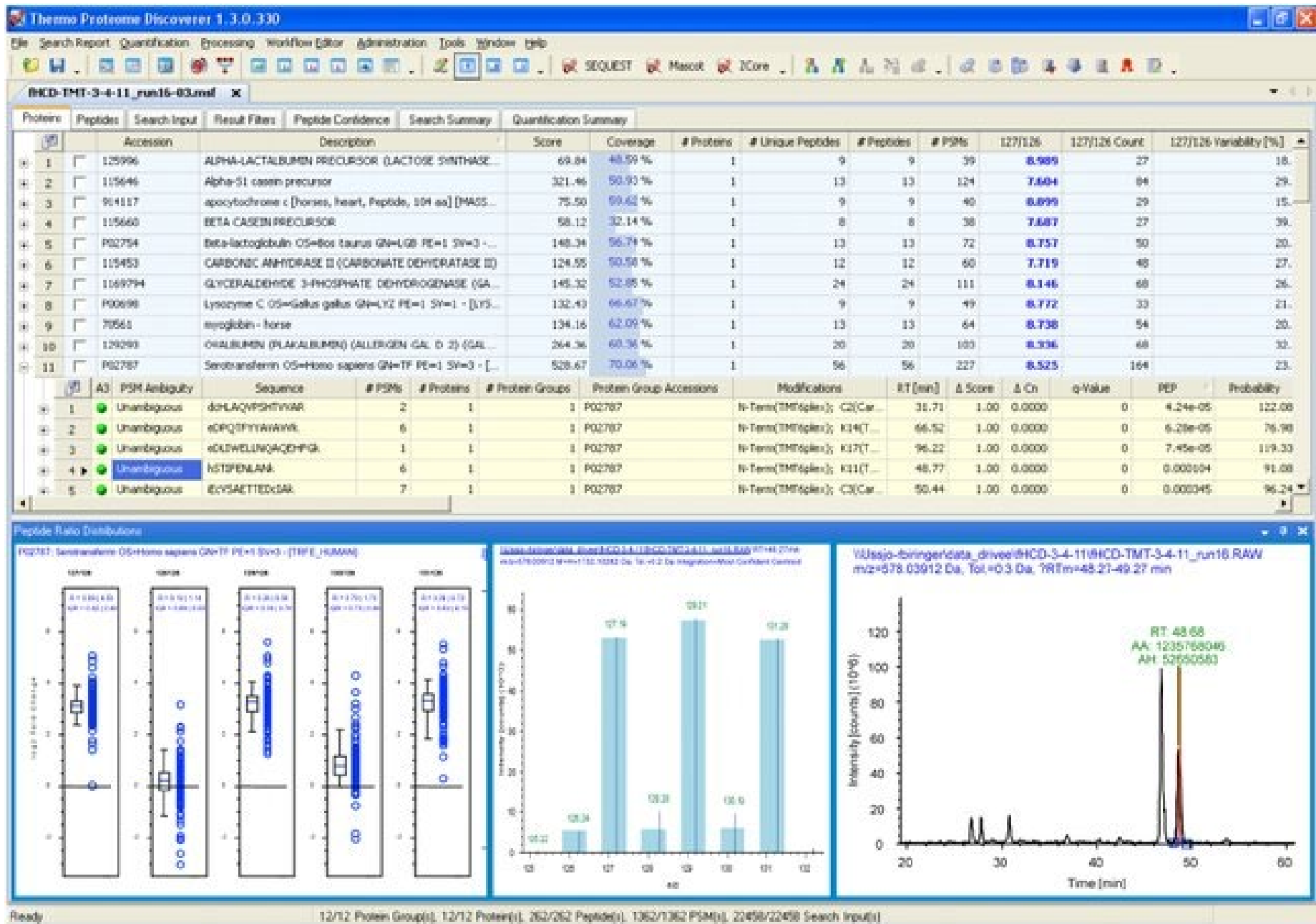
Skyline

Skyline Targeted Mass Spec Environment

PROTEOME DISCOVERER 1.4
Mass Informatics Platform for Protein Scientists

© Copyright 2008 - 2012 Thermo Fisher Scientific Inc.
All rights reserved. This program is protected by copyright
law and international treaties as described in Help About.

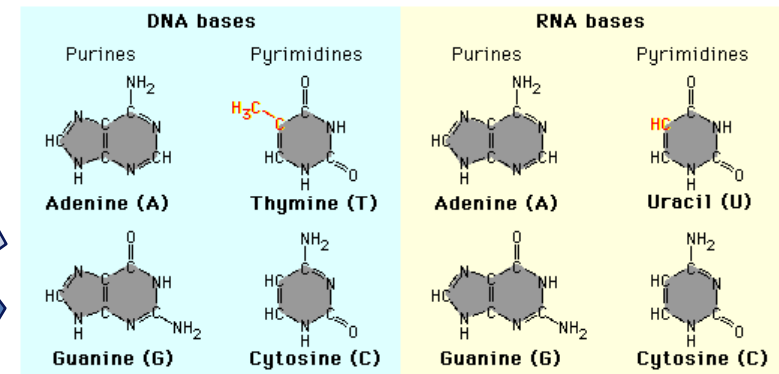
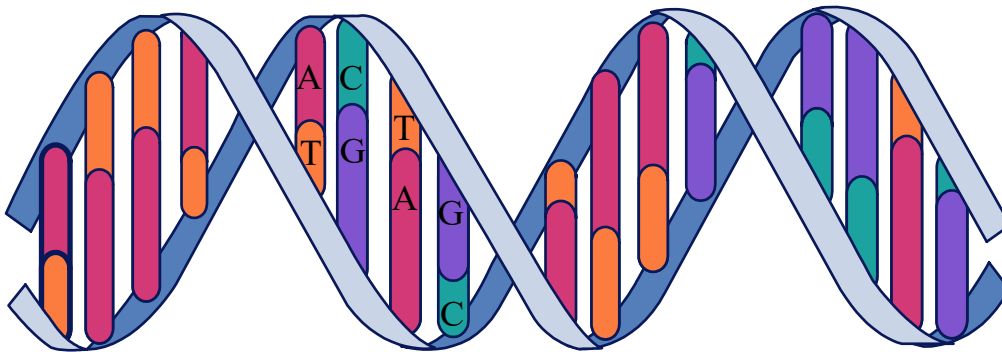
Thermo
SCIENTIFIC





Γονιδιωματική

Το DNA ως αποθήκη πληροφορίας



 Αδενίνη (A)

Προσδένει πάντα με

 Θυμίνη (T)

*Ουρακίλη (U)

 Κυτοσίνη (C)

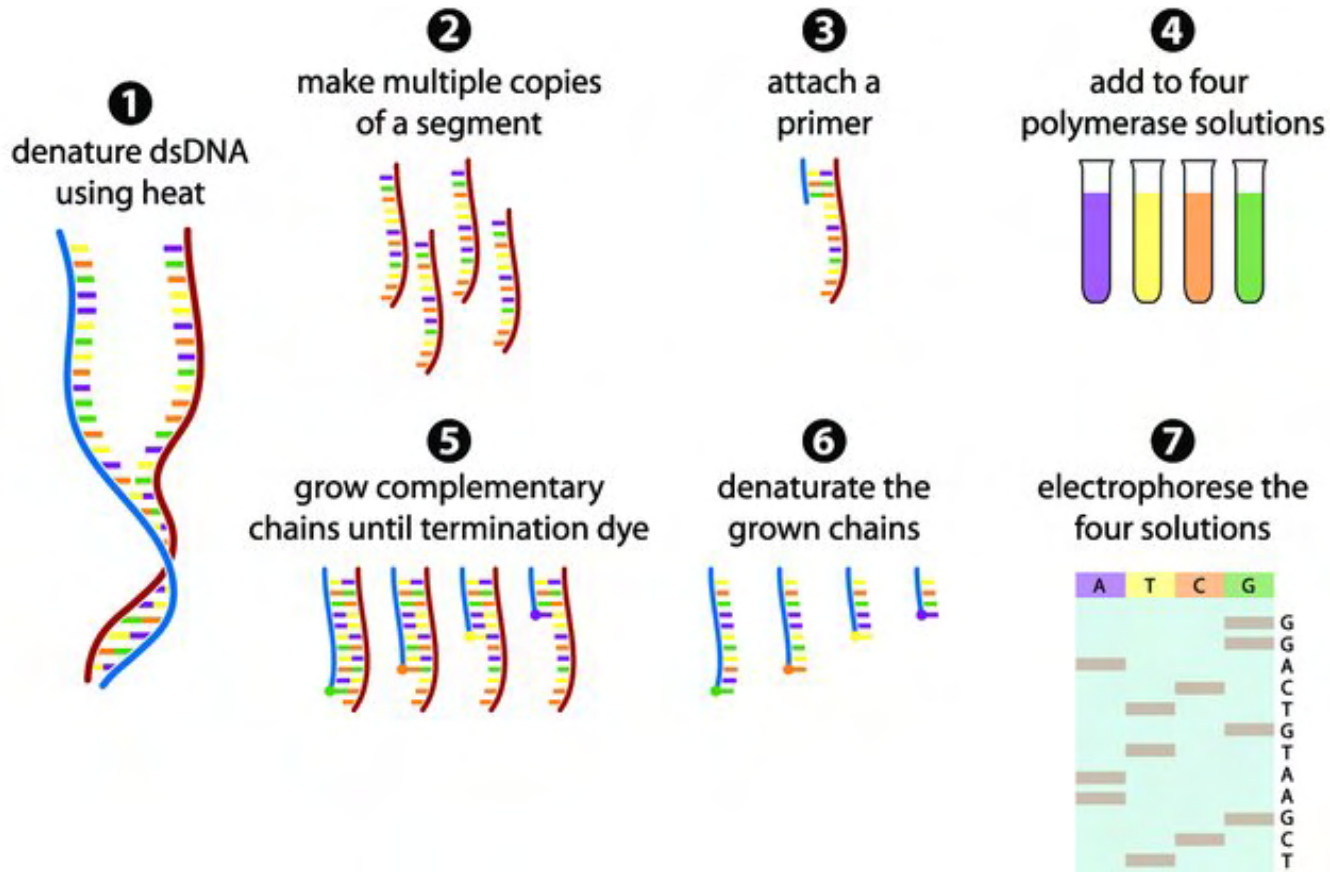
Προσδένει πάντα με

 Γουανίνη (G)

NGS

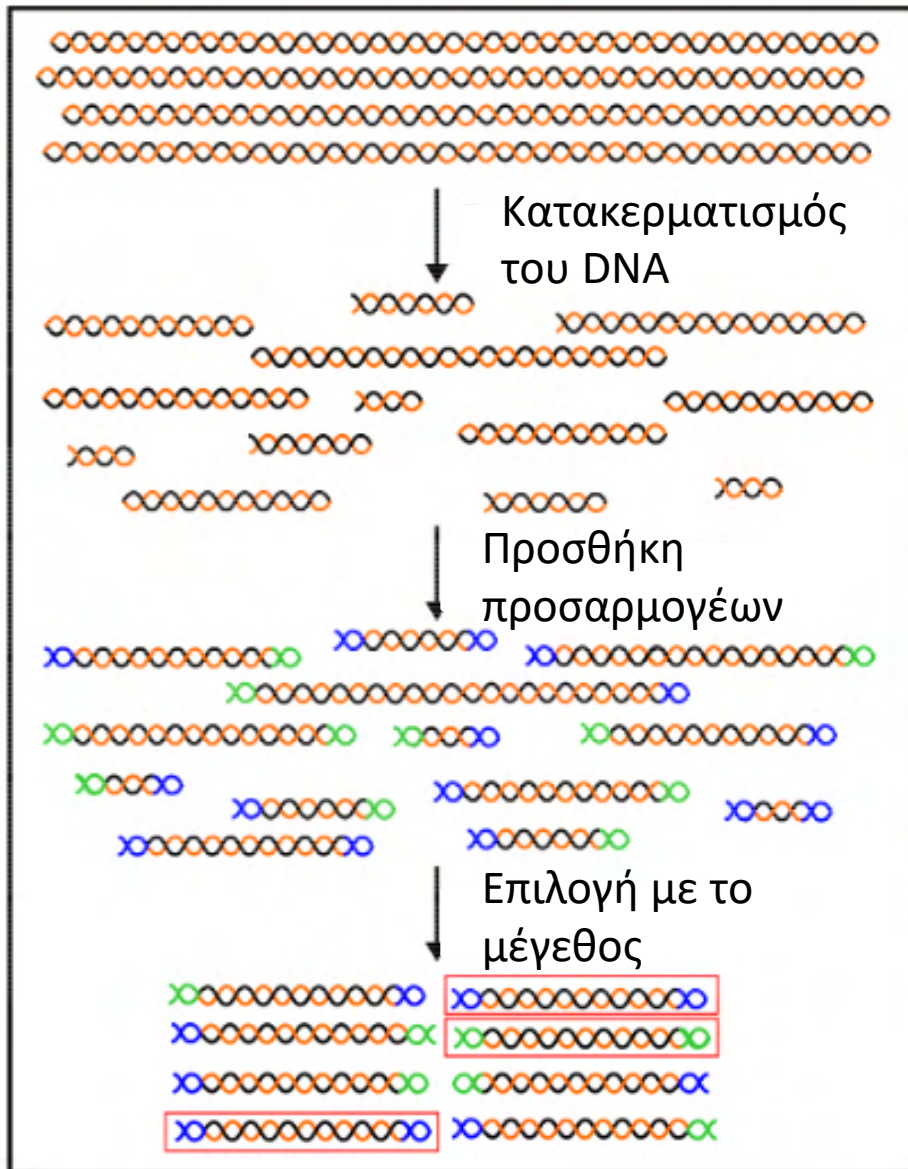
1. Φανταστείτε το DNA σας ως ένα μακρύ εγχειρίδιο οδηγιών. Κατακερματισμός σημαίνει ότι αυτό το εγχειρίδιο οδηγιών κόβεται σε μικρότερα κομμάτια. Αυτό μπορεί να συμβεί φυσικά ή να γίνει επίτηδες σε ένα εργαστήριο.
2. Σκεφτείτε τους προσαρμογείς σαν μικρές ετικέτες που κολλάτε στα εγχειρίδια οδηγιών κοπής. Αυτές οι ετικέτες τους βοηθούν να μείνουν μαζί με τη σωστή σειρά αργότερα.
3. Το PCR σημαίνει Αλυσιδωτή Αντίδραση Πολυμεράσης. Είναι ένας τρόπος δημιουργίας αντιγράφων των θραυσμάτων DNA. Υπάρχουν διαφορετικοί τρόποι (πρωτόκολλα) για να κάνετε PCR, και ο καθένας μπορεί να σας δώσει μια δέσμη (συστοιχία) μικροσκοπικών αποικιών, οι οποίες είναι ομάδες πανομοιότυπων θραυσμάτων DNA.
4. Φανταστείτε ένα ειδικό μηχάνημα που μπορεί να ξαναφτιάξει τα εγχειρίδια οδηγιών χρησιμοποιώντας τα μικρά κομμάτια. Αυτό το μηχάνημα χρησιμοποιεί ειδικά δομικά στοιχεία (νουκλεοτίδια) που λάμπουν (με φθορισμό) για να επανατοποθετήσει τα κομμάτια μαζί.
5. Τα ανακατασκευασμένα εγχειρίδια οδηγιών με τα λαμπερά δομικά στοιχεία είναι σαν ένα βιβλίο με εικόνες. Ένα μηχάνημα σαρώνει αυτό το εικονογραφημένο βιβλίο (διαβάζει) για να δει ποια λαμπερά κομμάτια υπάρχουν (απεικονίζοντας τη συστοιχία). Αυτό βοηθά τους επιστήμονες να κατανοήσουν την αρχική αλληλουχία DNA.

Sanger sequencing

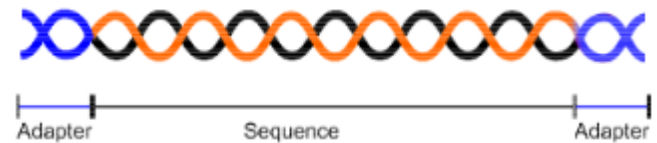


Gauthier, Michel G. *Simulation of polymer translocation through small channels: A molecular dynamics study and a new Monte Carlo approach*. Diss. University of Ottawa (Canada), 2008.

Αρχή λειτουργίας NGS

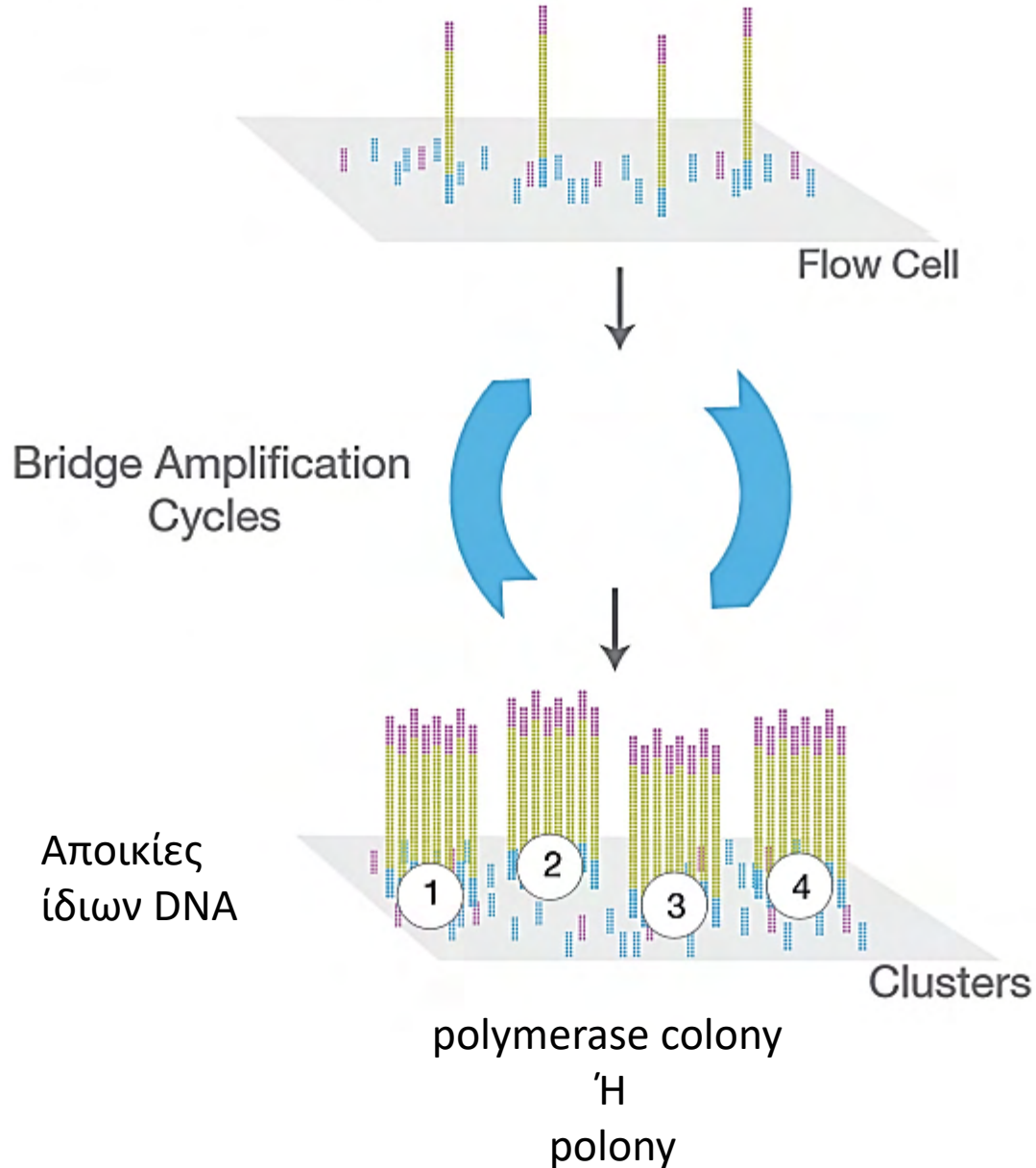


Κατακερματισμός με υπερήχηση



Δημιουργία της Βιβλιοθήκης με μέγεθος βάσει του ορίου διαβάσματος του μηχανήματος

B. Cluster Amplification

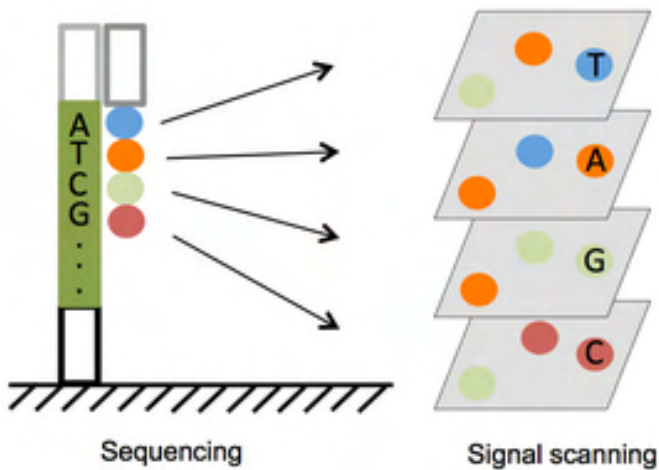


Το flow cell (επιφάνεια γυαλιού) καλύπτεται με ειδικά προσαρμοσμένα ολιγονουκλεοτίδια (adapters). Αυτά τα ολιγονουκλεοτίδια είναι συμπληρωματικά στα adapters που έχουν προστεθεί στα άκρα των μορίων DNA του δείγματος

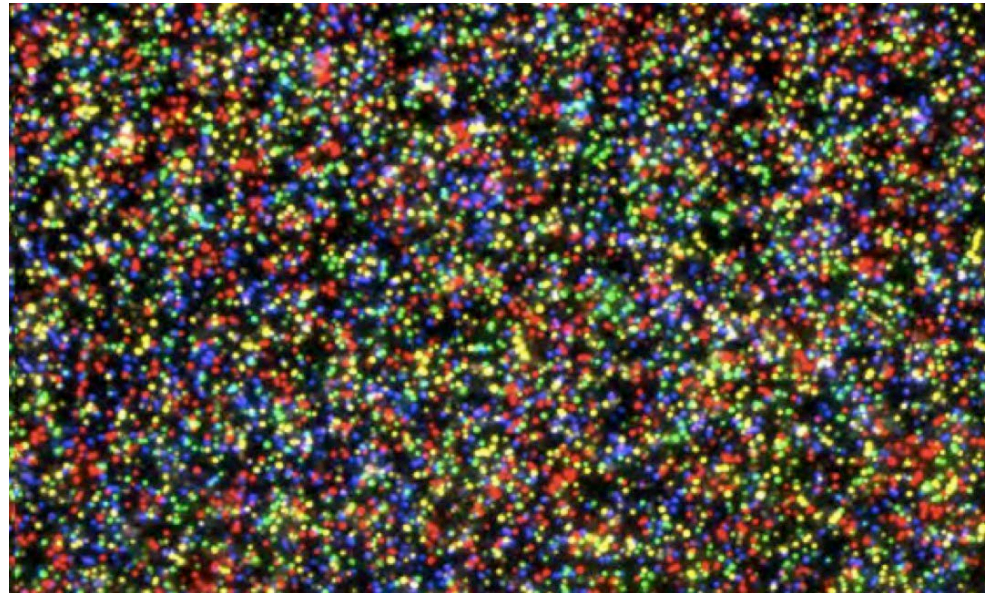
Bridge Amplification: Το DNA διπλώνει για να σχηματίσει μια γέφυρα (bridge) και υβριδοποιείται με ένα κοντινό ολιγονουκλεοτίδιο. Μετά από κύκλους αντιγραφής (cycles of replication) μέσω DNA πολυμεράσης, σχηματίζονται αντίγραφα του μορίου DNA.

Δημιουργία Clusters : Με τη διαδικασία επαναλαμβανόμενης αντιγραφής, κάθε αρχικό μόριο DNA αντιγράφεται πολλές φορές στο ίδιο σημείο, σχηματίζοντας ένα cluster που αποτελείται από εκατοντάδες ή χιλιάδες αντίγραφα της ίδιας αλληλουχίας. Είναι αρκετά πυκνά ώστε να παράγουν ανιχνεύσιμο σήμα κατά τη διαδικασία αλληλούχησης.

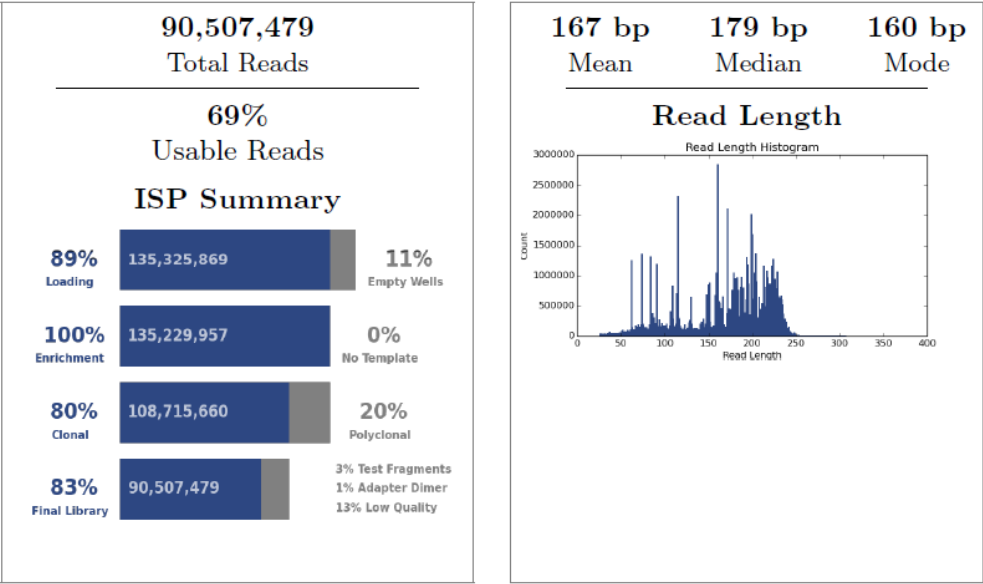
Η κάθε αποικία φωτίζει ανάλογα με το νουκλεοτίδιο που προστίθεται



Κατά τη διάρκεια της αλληλούχησης, **φθορίζοντα νουκλεοτίδια** (A, T, C, G) ενσωματώνονται στις αναπτυσσόμενες αλυσίδες DNA στις αποικίες.



Μέρος του chip
όταν λειτουργεί



Barcode Name	Sample	Bases	$\geq Q20$	Reads	Mean Read Length	Read Length Histogram
No barcode	none	335,828,653	285,070,290	2,097,701	160 bp	
IonXpress_001	Sample 1	19,019,728	17,357,836	121,298	156 bp	
IonXpress_002	Sample 2	72,656,371	64,421,723	384,755	188 bp	
IonXpress_003	Sample 3	44,755,483	39,680,941	241,586	185 bp	
IonXpress_004	Sample 4	99,213,528	88,927,887	533,459	185 bp	
IonXpress_005	Sample 5	41,004,703	36,996,373	228,438	179 bp	
IonXpress_006	Sample 6	206,427,006	185,192,801	1,132,005	182 bp	
IonXpress_007	Sample 7	292,593,417	262,802,403	1,609,115	181 bp	

FASTQ FILE

@JXIW1:03984:06407

AGTCGGCAAATTAGATCCGTAACCTTCGGGATAAGGATTGGCTCTGAAGACTGGGATGACTCGGGCTATGCCACGGCTTCGCGGCCGGG
GTTTGCCTCAATGCGTTCCTCCGTCGGGCCGGCCGGGGCCTTCGCGGGCCTCCGGGTGCGTTCGGCGGGCGCTGCAGTCATCACCAA
CAGTCAATTCAGAACTGGCACGGCTGAGGGAATCCGACTGTCTAAT

+

CCBBB@BBB=C@BCCCC?BBC@BC@BBB>BEE?B?AA>@?B?@ @BCAF;;;CD?CCBBBCCD>CCCCC@EBB=AB=???@=@=@
BB:BF>BB?BBC?CCCC?A?BB=@>;;*718=@?CCC;C?C@BB???8?<@ @=AB<BBAAA?CC?CCC=BB?BBC?;6//0>>@9>@>ACC
CCC?C?C>??=?@D@BCBD?BAABBF=B@AA>ACC@AA999@C

@JXIW1:03985:06415

TGAGATACCACCACTCTTACTGTTGCCTTACTTACATGATCAGGTGGAACAAGTGCTAAGGTTATTGCAACCTCCTAGCATAAGGTTTCTT
GTTTCAGCGTTTCAGTCATGTCGTCATTCTGGCCATAAGGCAGAAGACCGAGCCTGTGGTCATCGCGTCCGTCGCGCGGCTCTGGGCGTG
CGATCCGAATGGGCGGTTTCATCGCCGCACTAGTAGGGTCCCGCCCTCGGGCCCGTACGGCACACACCGGCCACAGCTTTGCTCAACTT
TCACACCGTACGCCGATGACAGTAAGACATCTGGATACTCCAAGTCATGGACAGTGCCAGGTGCGGAGTTTGACTGGGGCGGTACATCT
CCAAAACAATAACGGAGGTGTCCAAAGGTCAGCTCAGTGTGGACAGAAA

+

:CCCCBAA=:6;ACBCACDCCC@BB?B@DCC?BBCBBBCBCC@BB?B@CC?BBBCDD@B@BGCC;;B?B@CE@CCDEDAB?B=BB;B
EADG@CBCCCC@CCFGCDCCCCDCCCC@C?A@>A6;;A>A?BBBB?DCB=@ @AA>ABBB?BBBBCBC;;=@BABBBCC@C@AAEE<
AA;;BBCDADD@BBB=BB>BB=AAAAA>???ABDDDC=@@>9::5:DAB9@?:?AB??<?ADA??>??B>AACDCDC<AABBB?BB
B>BCFBB?>>CCCD@CCBDBBCCCC>AACAAABB>AABAAA?B@CCC@AA@ABBCCBB@BB?@ @??:<777*889<<<<2?@=BCCC
CBBB@FDC8;;6:@<@D<@ @=@BCD@CC=B?BB>??BBBBBBBB>A@AAAA

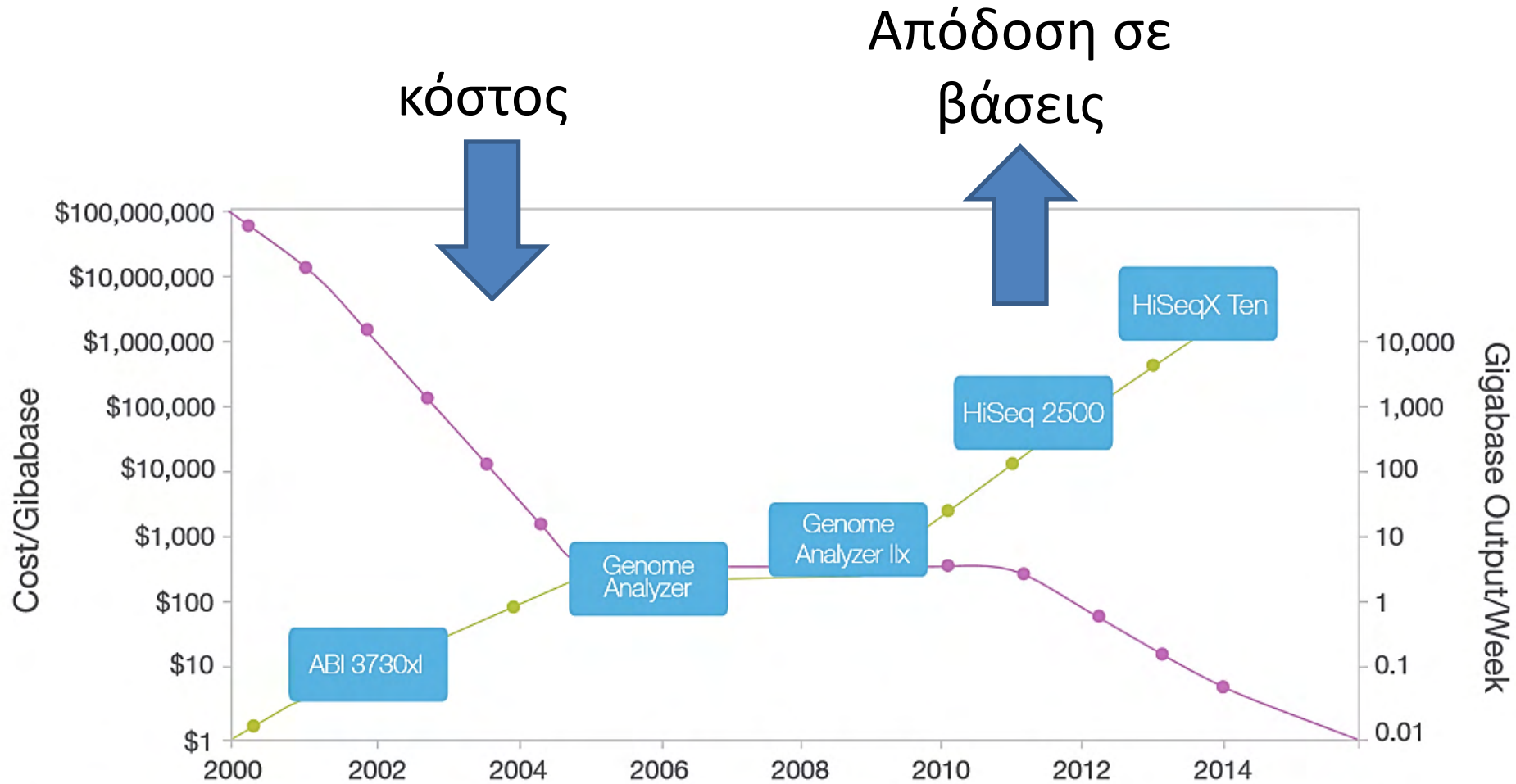


Figure 1: Sequencing Cost and Data Output Since 2000—The dramatic rise of data output and concurrent falling cost of sequencing since 2000. The Y-axes on both sides of the graph are logarithmic.

Και τώρα τι;

Χιλιάδες γραμμές αλληλουχιών DNA RNA και πρωτεϊνών.

Τι σημαίνουν οι διαφορές μεταξύ ατόμων;

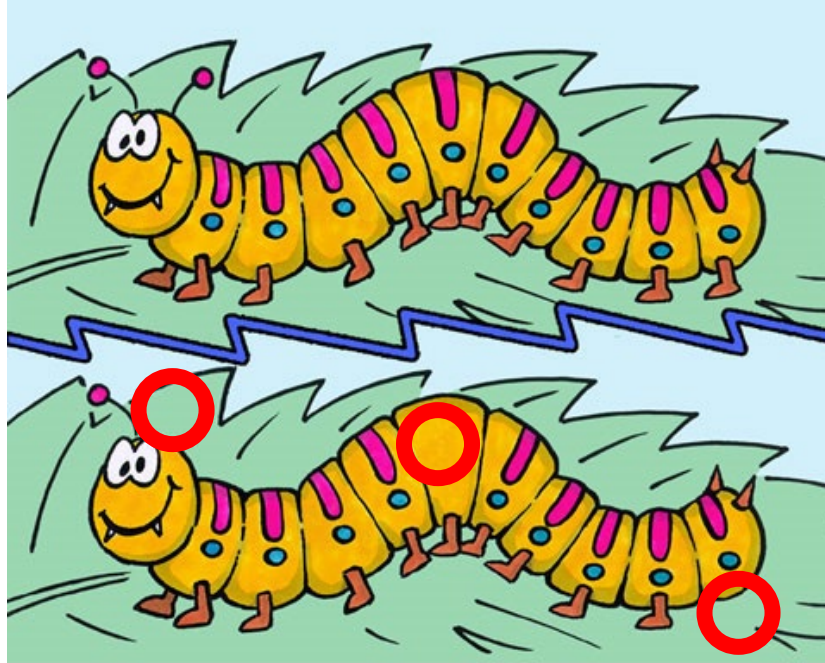
ΓΟΝΟΤΥΠΟΣ



ΦΑΙΝΟΤΥΠΟΣ

Πως τις εντοπίζουμε;

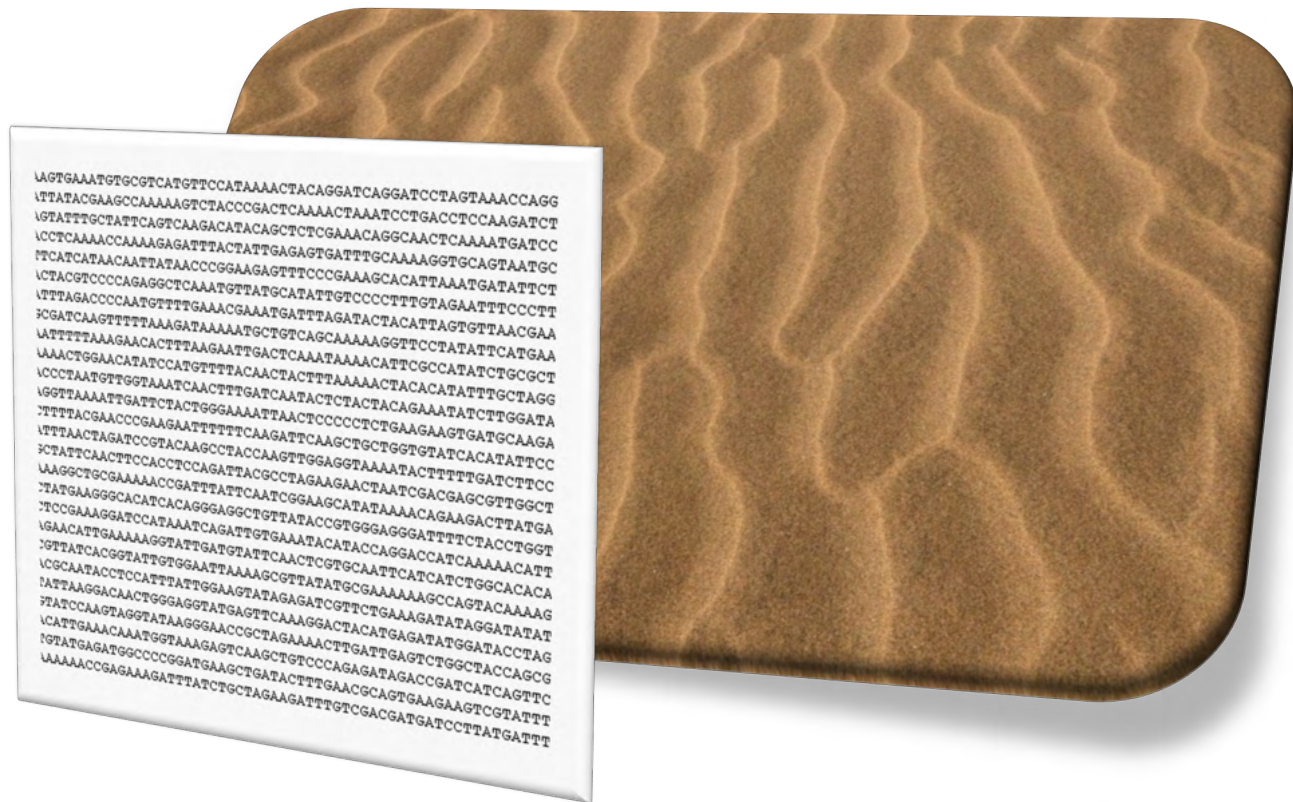
Ας παίξουμε ένα παιχνίδι...



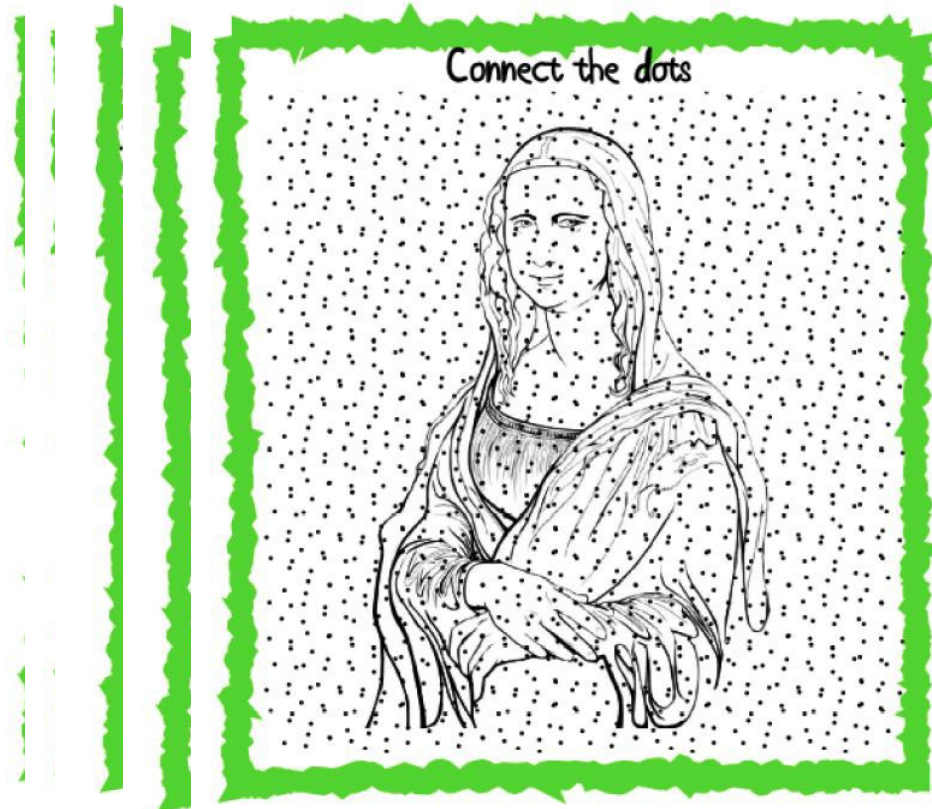
<http://www.scraggie.co.uk>

Εύκολο δεν ήταν;

Στην δικιά μας περίπτωση όμως...



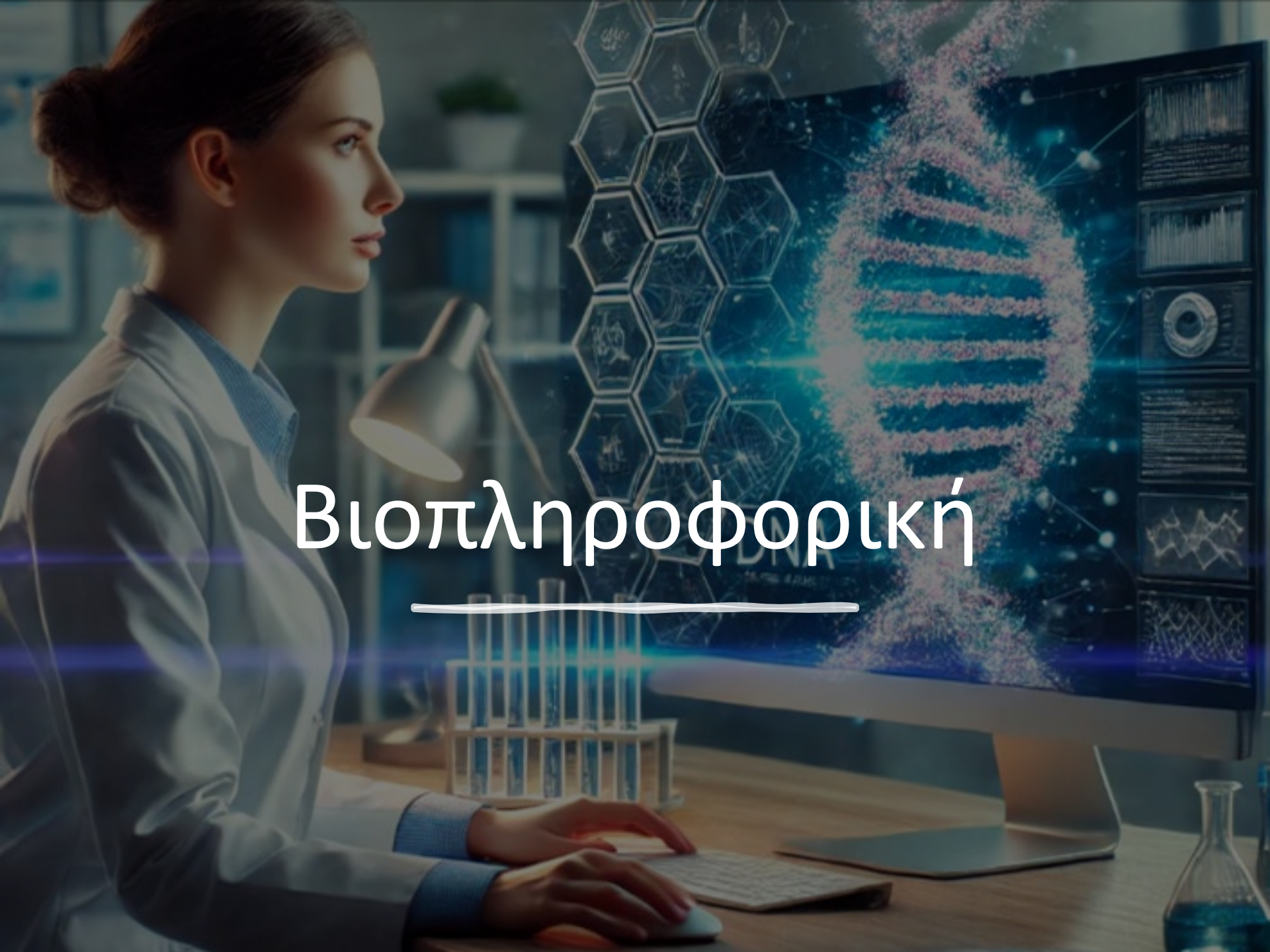
«Χάσμα Γενεών»



Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Τι καλούμαστε να κάνουμε

- Whole genome sequencing (για Genome wide association studies, για είδη που θέλουμε ολόκληρο το γονιδίωμα)
- Whole exome sequencing (ο φτωχός συγγενής του WGS)
- Amplicon sequencing (πχ microbiome)
- ChIP-seq για επιγενετικές αλλαγές
- RNA-seq (για διαφορική έκφραση γονιδίων – Differential Gene Expression Analysis - DGEA)
- Και άλλα πολλά



Βιοπληροφορική

Οι πιο συχνές χρήσεις των εξαγόμενων αλληλουχιών από Genomic sequencing

- Στοίχιση αλληλουχιών
(alignment)

AATTCAG	GAATTCA
GA-TC-G	GCAT-C-
AATTC-A	GAATTC-
GA-TCGA	GCAT-CG

- Συναρμολόγηση
αλληλουχιών
(Assembly)

- Ανίχνευση
πολυμορφισμών
(Variant calling)

Στοίχιση αλληλουχιών (Alignment)

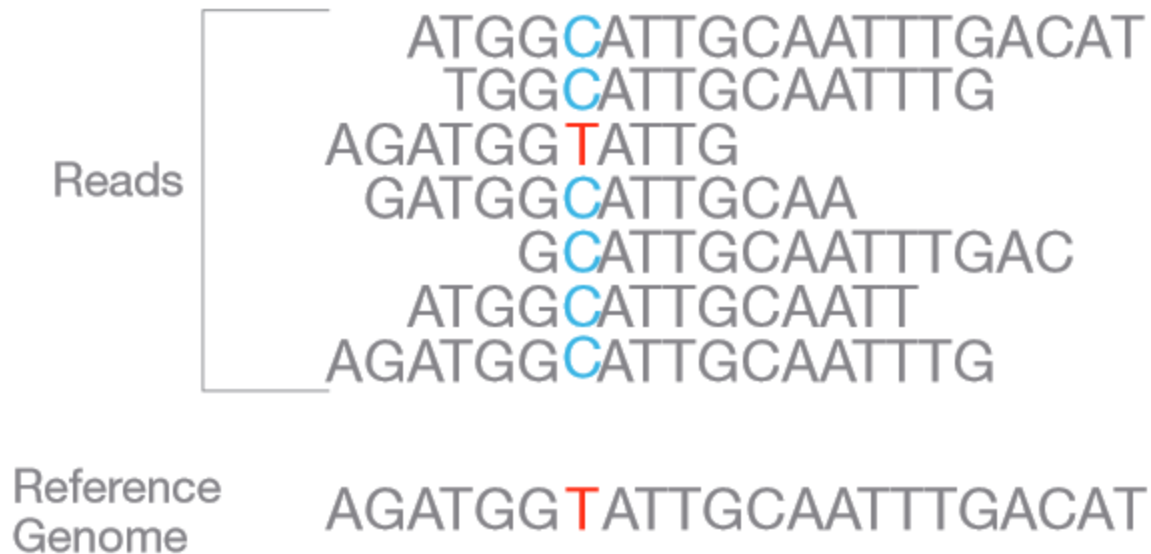
- Αναζητάμε τις διαφορές αλλά και τις ομοιότητες
- Ανάγουμε σχέσεις σύμφωνα με τις ομοιότητες
 - Μπορούμε να ανιχνεύσουμε σχέσεις κληρονομικές, αναπτυξιακές, λειτουργικές

Πώς;

Παίζουμε το παιχνίδι αλλά «κλέβουμε» με την χρήση υπολογιστών.

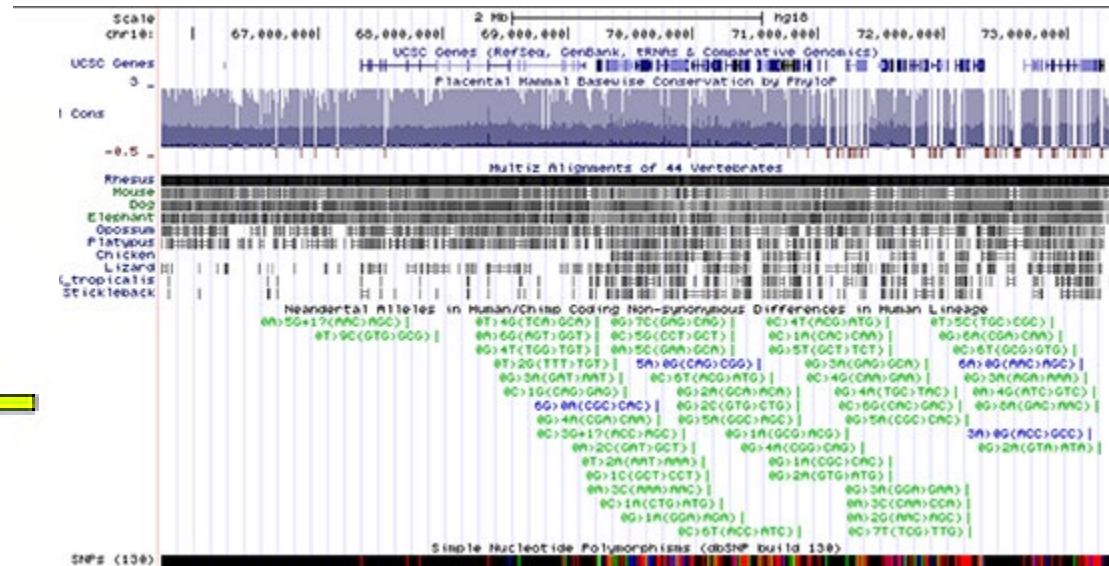
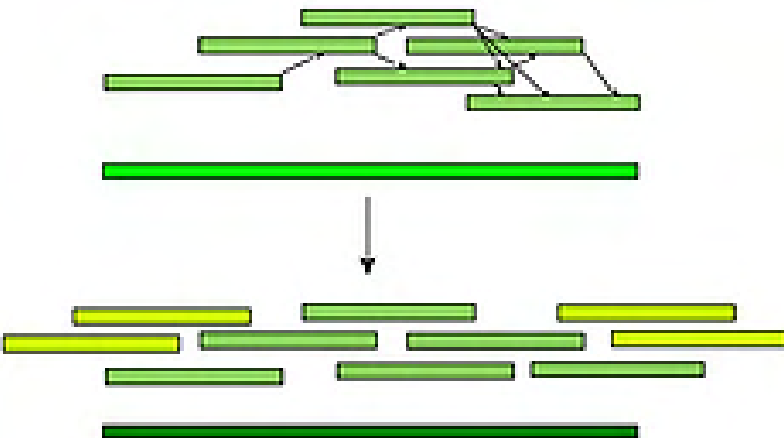
Στοίχιση των αλληλουχιών

D. Alignment & Data Analysis



Βιοπληροφορική στη γονιδιωματική

Στοίχιση αλληλουχιών και τοποθέτηση των ίδιων περιοχών τη μία πάνω στην άλλη ώστε να βγει η συνολική



Reference sequences

<https://www.ncbi.nlm.nih.gov/nucleotide/>

NCBI Resources ☒ How To ☒ [ndovro@gmail.com](#) [My NCBI](#) [Sign Out](#)

Nucleotide [Advanced](#) [Help](#)

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.coronavirus.gov>.
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

https://www.ensembl.org/Homo_sapiens/Info/Index

e!Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog [Login/Register](#)

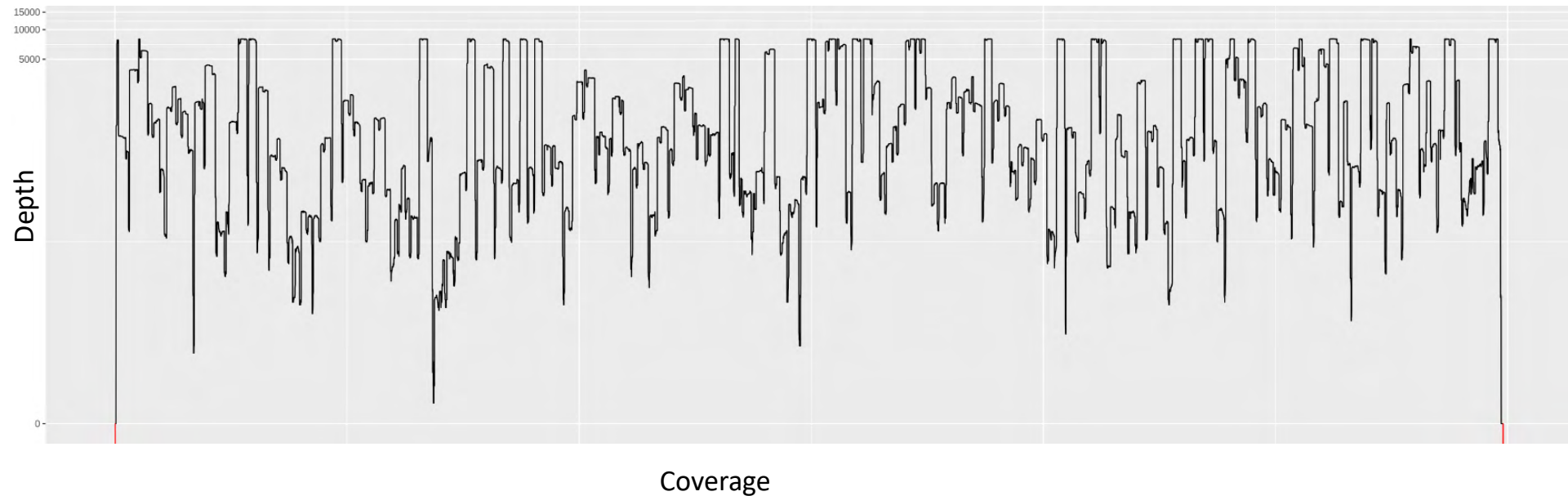
 **Human** (GRCh38.p13) ▼

Search Human (Homo sapiens)

Search all categories

e.g. [BRCA2](#) or [17:63992802-64038237](#) or [rs699](#) or [osteoarthritis](#)

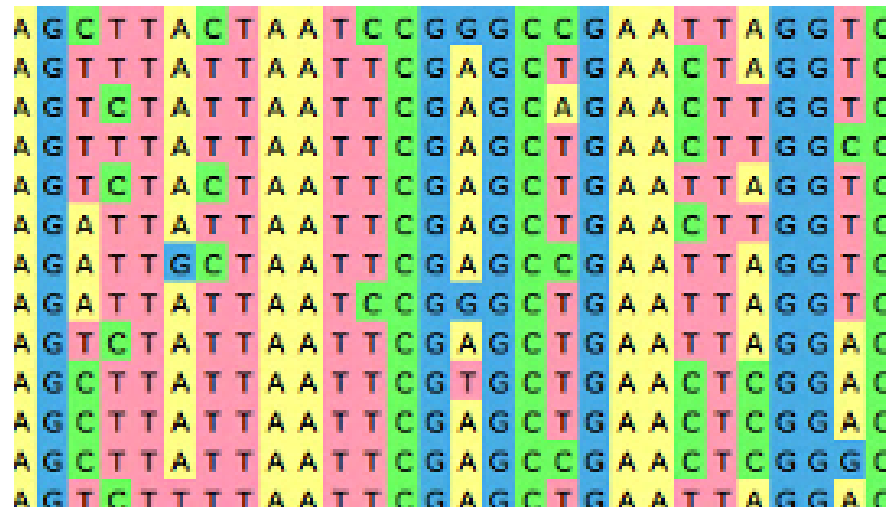
Depth and coverage



Depth= πόσες βάσεις έχουμε διαβάσει στα reads μας σε μία συγκεκριμένη θέση
Coverage= πόσες βάσεις του reference sequence έχουμε «πιάσει»

Multiple Sequence alignment (MSA)

- Όταν θέλουμε να ελέγξουμε πολλαπλές αλληλουχίες ίδιου μήκους για ομολογία μεταξύ τους
- Συνήθως κάνει ανά-δύο συγκρίσεις (pairwise).
- Συχνότεροι Αλγόριθμοι : Clustal (W or OMEGA), MUSCLE, MAFFT



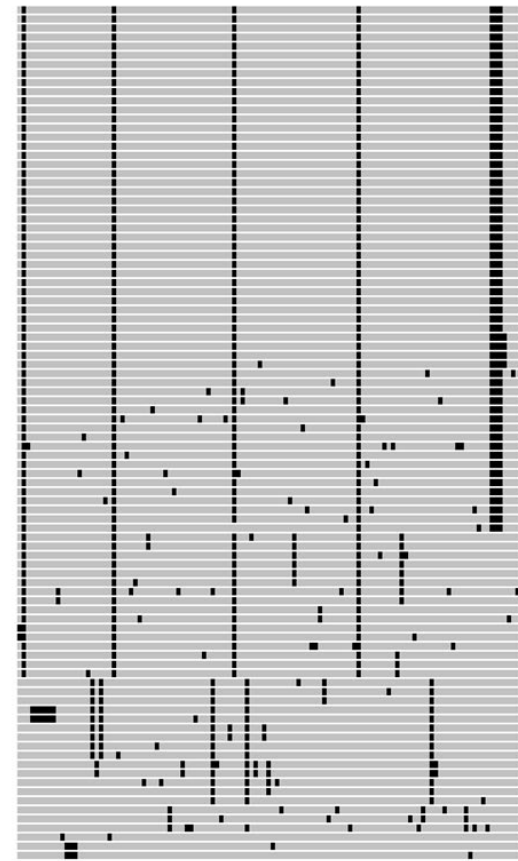
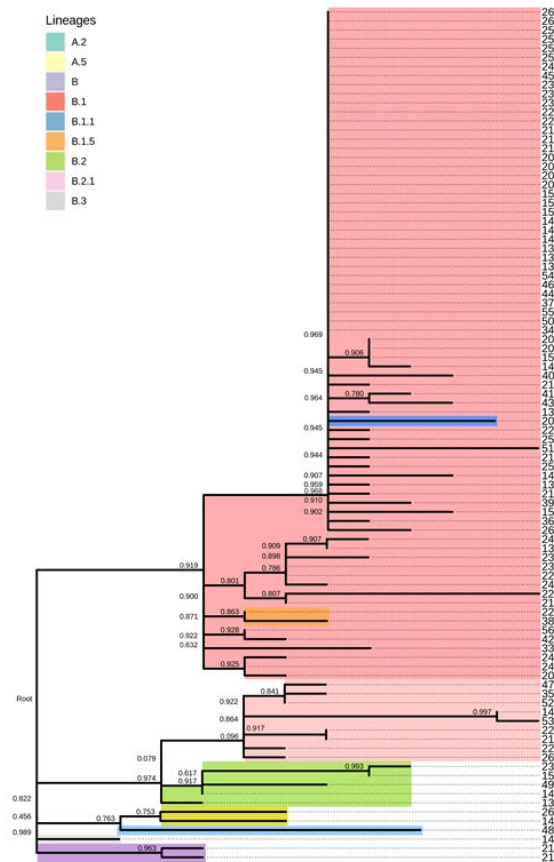
```
A G C T T A C T A A T C C G G G C C G A A T T A G G T C
A G T T T A T T A A T T C G A G C T G A A C T A G G T C
A G T C T A T T A A T T C G A G C A G A A C T T G G T C
A G T T T A T T A A T T C G A G C T G A A C T T G G C C
A G T C T A C T A A T T C G A G C T G A A T T A G G T C
A G A T T A T T A A T T C G A G C T G A A C T T G G T C
A G A T T G C T A A T T C G A G C C G A A T T A G G T C
A G A T T A T T A A T C C G G G C T G A A T T A G G T C
A G T C T A T T A A T T C G A G C T G A A T T A G G A C
A G C T T A T T A A T T C G T G C T G A A C T C G G A C
A G C T T A T T A A T T C G A G C T G A A C T C G G A C
A G C T T A T T A A T T C G A G C C G A A C T C G G G C
A G T C T T T T A A T T C G A G C T G A A T T A G G A C
```

MSA στην πράξη

Figure 1. Phylogram with polymorphism profile of SARS-Cov2 variants from Athens, Greece. Tree calculations were based on maximum-likelihood method. The ratio of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. Colours represent SARS-Cov2 lineages.

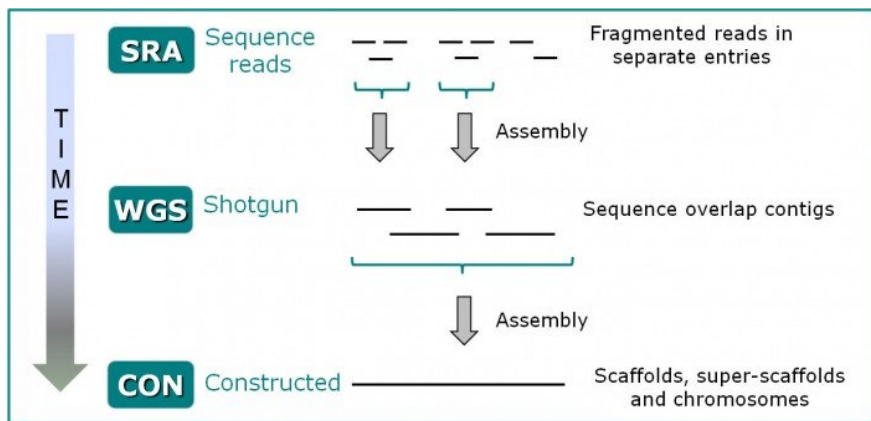
Spanakis, N.*, Kassela, K.*, Dovrolis, N.*, Bampali, M., Gatzidou, E., Kafasi, A., Froukala, E., Stavropoulou, A., Lilakos, K., Veletza, S., Tsiodras, S., Tsakris, A. and Karakasilioti, I. 2020 "A main event and multiple introductions of SARS-CoV-2 initiated the COVID-19 epidemic in Greece." *Journal of medical virology* 93.5 (2021): 2899-2907.

*Equal Contribution



Assembly – Συναρμολόγηση

- De novo
- Genome guided



<https://www.ebi.ac.uk/training/online/course/nucleotide-sequence-data-resources-ebi/what-ena/how-sequence-assembled>

Modern assemblers use refined graph algorithms to deduce a genome sequence *de novo*

Name	Algorithm	Reference
Celera WGA Assembler	Best Overlap Graph	Myers, E.W., <i>et al.</i> (2000) <i>Science</i> Huson, D.H., <i>et al.</i> (2001) <i>Bioinformatics</i>
EULER	De Bruijn Graph	Pevzner, P.A., <i>et al.</i> (2001) <i>PNAS</i>
Velvet	De Bruijn Graph	Zerbino, D.R. & Birney, E. (2008) <i>Genome Res.</i>
ALLPATHS	De Bruijn Graph (Unipath)	Butler, J., <i>et al.</i> (2008) <i>Genome Res.</i>
ABYSS	Distributed De Bruijn Graph	Simpson, J.T., <i>et al.</i> (2009) <i>Genome Res.</i>
SOAPdenovo (MEGAHIT)	(Succinct) De Bruijn Graph	Li, R., <i>et al.</i> (2009) <i>Genome Res.</i>
SPAdes	Paired De Bruijn Graph	Bankevich, A., <i>et al.</i> (2012) <i>J. of Comp. Biol.</i>
MaSuRCA	De Bruijn Graph for "Super-reads", then an Overlap Graph	Zimin, A.V., <i>et al.</i> (2013) <i>Bioinformatics</i>

Additionally, there are specialized transcriptome assemblers like: Trinity and Trans-ABYSS.

<https://www.slideshare.net/JosHctorGlvez/basics-of-genome-assembly>

Assembly performed using:
Shepard, Samuel S., et al. "Viral deep sequencing needs an adaptive approach: IRMA, the iterative refinement meta-assembler." *BMC genomics* 17.1 (2016): 708.

Assembly performed using:
Shepard, Samuel S., et al. "Viral deep sequencing needs an adaptive approach: IRMA, the iterative refinement meta-assembler." *BMC genomics* 17.1 (2016): 708.

Variant calling

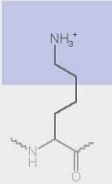
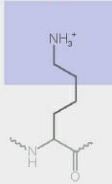
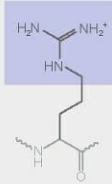
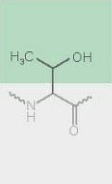
- Ανίχνευση σημειακών και μη πολυμορφισμών (SNPs – SNVs)
- Ανίχνευση INDELs (insertions – deletions)
- Ανίχνευση CNV (copy number variation)
- Απαιτεί και πάλι 2 αλληλουχίες
 - 1 δικιά μας που μπορεί να προκύψει από assembly ή consensus από reads
 - 1 reference
- Αποτελείται από 2 βασικά βήματα και ένα τρίτο συμπληρωματικό
 - Base calling (που είναι ουσιαστικά ίδιο με το alignment + depth)
 - Δημιουργία VCF Αρχείων (variant calling format)
 - VCF annotation (την επίδραση των αλλαγών μας στα πεπτίδια και στις πρωτεΐνες τους) από βάσεις δεδομένων.

VCF file

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT ../sort_R10.bam
R10 241 . C T 104 . DP=178;VDB=3.06419e-20;SGB=-0.693147;MQ0F=0;AC=2;AN=2;DP4=0,0,133,0;MQ=60
GT:PL 1/1:134,255,0
R10 3037 . C T 110 . DP=43;VDB=0.907023;SGB=-0.693145;MQ0F=0;AC=2;AN=2;DP4=0,0,40,0;MQ=59 GT:PL
1/1:140,120,0
R10 14408 . C T 198 . DP=175;VDB=1;SGB=-0.693147;MQ0F=0;AC=2;AN=2;DP4=0,0,169,0;MQ=59 GT:PL 1/1:228
,255,0
R10 14786 . C T 225 . DP=259;VDB=1.21219e-07;SGB=-0.693147;MQ0F=0;AC=2;AN=2;DP4=0,0,252,0;MQ=59
GT:PL 1/1:255,255,0
R10 18126 . T C 228 . DP=393;VDB=0.00600597;SGB=-0.693147;RPB=1;MQB=1;BQB=1;MQ0F=0;AC=2;AN=2;DP4=1,0,
384,0;MQ=59 GT:PL 1/1:255,255,0
R10 23403 . A G 138 . DP=306;VDB=5.22786e-06;SGB=-0.693147;RPB=1;MQB=1;BQB=1;MQ0F=0;AC=2;AN=2;DP4=1,0,
,193,0;MQ=60 GT:PL 1/1:165,255,0
R10 28881 . G A 225 . DP=219;VDB=1.60741e-16;SGB=-0.693147;MQ0F=0;AC=2;AN=2;DP4=0,0,195,0;MQ=59
GT:PL 1/1:255,255,0
R10 28882 . G A 142 . DP=208;VDB=2.58008e-17;SGB=-0.693147;MQ0F=0;AC=2;AN=2;DP4=0,0,186,0;MQ=59
GT:PL 1/1:172,255,0
R10 28883 . G C 207 . DP=210;VDB=3.9167e-17;SGB=-0.693147;RPB=0.173007;MQB=0.995834;BQB=0.0131085;MQ0
F=0;AC=2;AN=2;DP4=3,0,195,0;MQ=59 GT:PL 1/1:234,255,0
(END)
```

Αποτέλεσμα πολυμορφισμών

- Synonymous (Silent)
- Non-Synonymous
 - Nonsense
 - A random stop codon is introduced, and the protein is incomplete
 - Missense
 - In a conservative missense mutation, the amino acid replaced is similar in function and shape to the amino acid being replaced.
 - In a non-conservative missense mutation, a completely different kind of amino acid is added to the chain.

	Point mutations				
	No mutation	Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr
					
	basic			polar	

<https://biologydictionary.net/missense-mutation/>

Annotated VCF

rs_ids	gene	chrom	start	end	ref	alt	aa_change	impact	impact_se	aaf_1kg_a	aaf_exac	aaf_esp_a	clinvar_dis	clinvar_ge	polyphen	polyphen	sift_pred	sift_score	cadd_raw	gts.Samp34			
rs7770385	DDX11L1	chr1	13416	13417	C	CGAGA		non_codin	LOW	-1	0.067	-1	None	None		None		None	None	C/C			
None	DDX11L1	chr1	13655	13658	CAG	C		non_codin	LOW	-1	-1	-1	None	None		None		None	None	CAG/C			
None	RP5-857K2	chr1	664485	664488	TAG	T		upstream_	LOW	-1	-1	-1	None	None		None		None	None	TAG/TAG			
rs1891476	RP11-206L	chr1	721756	721757	T	A		non_codin	LOW	0.028155	0.051	-1	None	None		None		None	None	T/T			
rs7404521	LINC00111	chr1	762329	762330	G	T		non_codin	LOW	-1	0.014	-1	None	None		None		None	None	G/G			
rs7404521	LINC00111	chr1	762344	762345	A	G		non_codin	LOW	-1	0.017	-1	None	None		None		None	None	A/A			
rs4128579	SAMD11	chr1	865627	865628	G	A	G/S	missense_	MED	0.002796	0.003051	0.003465	None	None	benign	0.014	deleteriou	0.01	None	G/G			
rs5683401	SAMD11	chr1	874777	874826	GCCTCCCC	G		splice_don	HIGH	-1	0.047	-1	None	None		None		None	None	GCCTCCCCAGCCACGGTGAGGAC			
rs4372192	SAMD11	chr1	876498	876499	A	G		intron_var	LOW	0.913139	0.92	0.91929	None	None		None		None	None	G/G			
rs6672356	SAMD11	chr1	877830	877831	T	C	W/R	missense_	MED	1	0.999	-1	None	None	benign	0	tolerated	1	None	C/C			
rs1425582	SAMD11	chr1	878313	878314	G	C	G	synonymo	LOW	0.043131	0.04	0.048918	None	None		None		None	None	G/G			
rs1483278	SAMD11	chr1	878330	878331	C	T	P/L	missense_	MED	0.005192	0.005314	0.007161	None	None	benign	0.013	tolerated	0.25	None	C/C			
rs3701049	SAMD11	chr1	878663	878664	G	A	E	synonymo	LOW	0.000399	0.000511	7.69E-05	None	None		None		None	None	G/G			
rs7523549	SAMD11	chr1	879316	879317	C	T	Y	synonymo	LOW	0.091254	0.048	0.082384	None	None		None		None	None	C/C			
rs1446562	NOC2L	chr1	880482	880483	A	G	D	synonymo	LOW	0.002196	0.0015	0.001692	None	None		None		None	None	A/A			
rs2272757	NOC2L	chr1	881626	881627	G	A	L	synonymo	LOW	0.441893	0.564	0.474777	None	None		None		None	None	A/A			
rs3547188	NOC2L	chr1	881917	881918	G	A	S/L	missense_	MED	0.029353	0.046	0.045825	None	None	benign	0.183	deleteriou	0.01	None	G/G			
rs7601133	NOC2L	chr1	883591	883592	C	G	L	synonymo	LOW	-1	8.32E-06	-1	None	None		None		None	None	C/C			
rs3828047	NOC2L	chr1	887800	887801	A	G	T	synonymo	LOW	0.922324	0.935	0.927418	None	None		None		None	None	G/G			
rs3748596	NOC2L	chr1	888638	888639	T	C	E	synonymo	LOW	0.922724	0.936	0.927264	None	None		None		None	None	C/C			
rs3748597	NOC2L	chr1	888658	888659	T	C	I/V	missense_	MED	0.922724	0.935	0.92733	None	None	benign	0	tolerated	1	None	C/C			
rs3828049	NOC2L	chr1	889237	889238	G	A	A/V	missense_	MED	0.05651	0.057	0.047216	None	None	possibly_d	0.88	tolerated	0.06	None	G/G			
rs1386722	NOC2L	chr1	891590	891591	C	T	R/H	missense_	MED	0.000599	0.001269	0.001307	None	None	benign	0	tolerated	0.79	None	C/C			
None	NOC2L	chr1	892397	892398	C	T	S/N	missense_	MED	-1	-1	-1	None	None	possibly_d	0.511	deleteriou	0.01	None	C/C			
rs1330301	NOC2L	chr1	894572	894573	G	A		intron_var	LOW	0.634385	0.809	0.690167	None	None		None		None	None	A/A			
rs1130343	KLHL17	chr1	896063	896064	C	G		5_prime_L	LOW	0.092252	0.008508	-1	None	None		None		None	None	C/C			
rs4970441	KLHL17	chr1	897324	897325	G	C	A	synonymo	LOW	0.859225	0.907	0.85889	None	None		None		None	None	C/C			
rs1330322	KLHL17	chr1	897563	897564	T	C		intron_var	LOW	0.89996	0.899	-1	None	None		None		None	None	C/C			

Το «οπλοστάσιο μας»

Δημιουργία
μεθοδολογικών
ροών
ανάλυσης
(pipelines)
από
σύνθεση
πηγών και
υλοποιήσεω
ν

Ανάλυση γονιδιακών πολυμορφισμών

Εργαλεία και αλγόριθμοι όπως το PLINK, τα bedtools κ.α. που ανιχνεύουν στατιστικά σημαντικούς πολυμορφισμούς ανάμεσα σε ομάδες δειγμάτων

Annotation πολυμορφισμών (πχ Ensembl VEP) από διάφορες βάσεις δεδομένων

Ανάλυση γονιδιακής έκφρασης

Βάσεις ανοιχτών δεδομένων (repositories) όπως το GEO, το TCGA κ.α. που προσφέρουν δεδομένα γονιδιακής έκφρασης

Αλγόριθμοι και υλοποιήσεις στατιστικής ανάλυσης διαφορικής έκφρασης αναλόγως τον τρόπο ανίχνευσης της έκφρασης. πχ GEO2R, Limma για δεδομένα μικροσυστοιχειών, DESEQ2, EdgeR για RNASEQ.

Ανάλυση μικροβιώματος

QIIME, PICRUST, Calypso, MicrobiomeAnalyst μπορούν να συνδυαστούν και να χρησιμοποιηθούν παράλληλα

Vegan και phyloseq (πακέτα της R) καθώς και αλγόριθμοι πολύ-παραγοντικής στατιστικής ανάλυσης όπως LEfSe, MIXMC, Canonical Correspondence Analysis κ.α συνδέουν το μικροβίωμα με συγκεκριμένες υποθέσεις

Το «οπλοστάσιο μας»

Δημιουργία
μεθοδολογικ
ών ροών
ανάλυσης
(pipelines)
από
σύνθεση
πηγών και
υλοποιήσεω
ν

Ανάλυση μοριακών Οδών

Βάσεις δεδομένων όπως KEGG, Reactome, GO, Panther και άλλες που σύμφωνα με την είσοδο των γονιδίων μας επιστρέφουν μοριακά μονοπάτια στα οποία συμμετέχουν

PathwayConnector που προσφέρει συμπληρωματικά μονοπάτια για την επίτευξη ενός πλήρους συνδεδεμένου δικτύου μοριακών οδών

Ανάλυση δικτύων συσχετίσεων γονιδίων και πρωτεϊνών

Πλατφόρμες όπως το STRING, με είσοδο γονίδια που μας ενδιαφέρουν, επιστρέφουν ποιες πρωτεΐνες αυτά μεταφράζουν και εμπλουτισμένα δίκτυα στα οποία αυτές συσχετίζονται

Υλοποιήσεις του αλγορίθμου genemania προσφέρουν εμπλουτισμένα δίκτυα γονιδίων και συσχτίσεων τους μέσω βιβλιογραφικών πληροφοριών

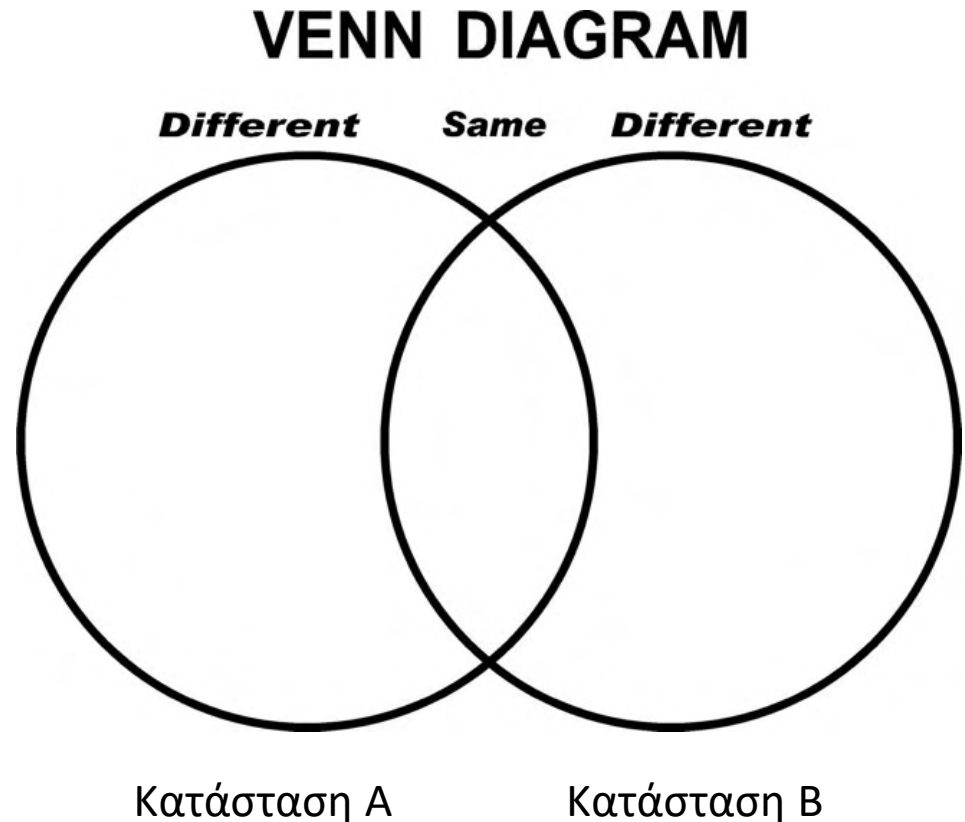
Ανάλυση βιολογικών δικτύων (γενικότερα)

Cytoscape ως μία εύκολη εφαρμογή απεικόνισης και ανάλυσης δικτύων γενικότερα

Igraph μέσω της R για απεικόνιση και ανάλυση δικτύων (απαιτεί προγραμματιστικές γνώσεις)

Βιοπληροφορική στην έκφραση

- Ανάλυση στη διαφορά των επιπέδων
 - Πρωτεϊνών
 - mRNAs
 - Μεταβολιτών
 - κτλ



RNA-seq

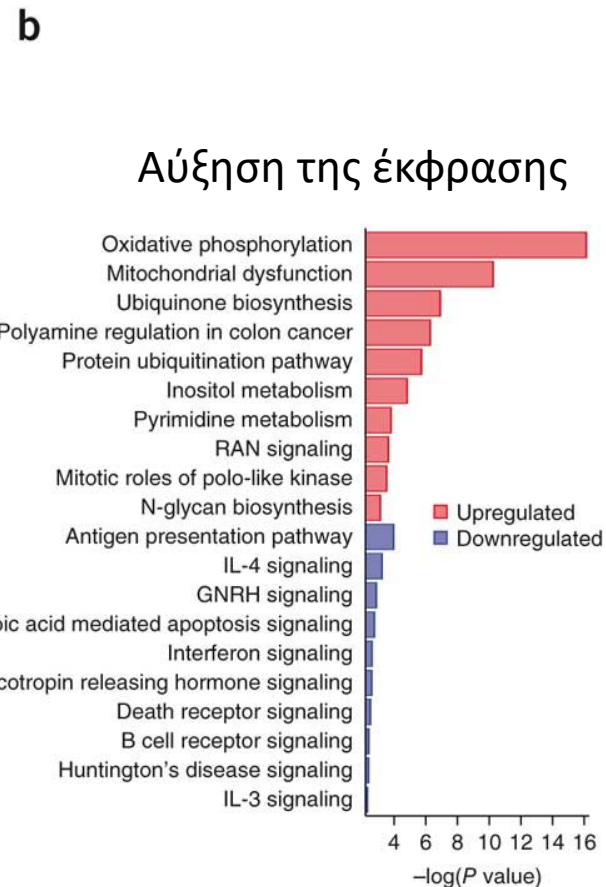
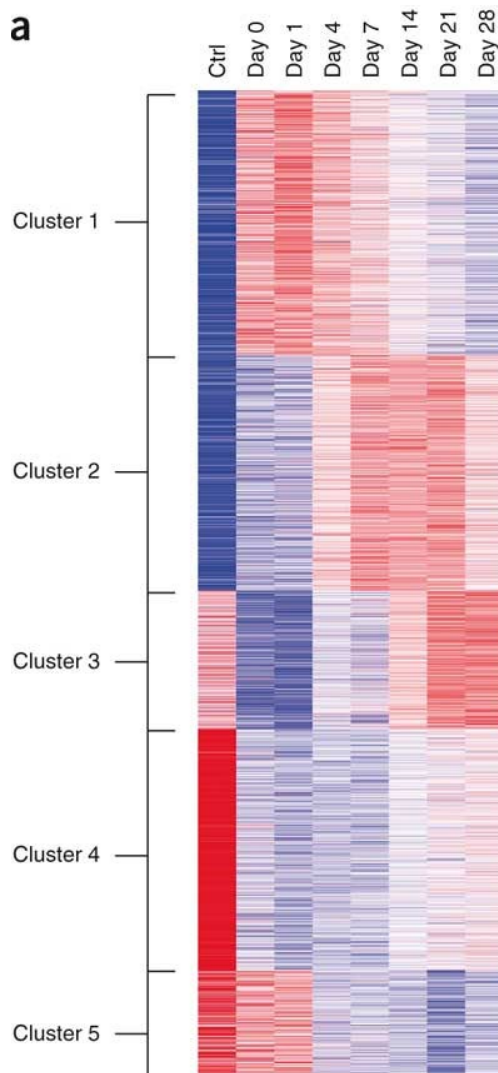
- Το χρησιμοποιούμε για διαφορική έκφραση
- Ήρθε να συμπληρώσει και ίσως να υπο/αντικαταστήσει ή συμπληρώσει μεθόδους όπως τα microarrays και η real time quantitate PCR (qPCR) όταν θέλουμε να ελέγξουμε πάρα πολλά γονίδια μαζί
- Συνήθως έχει 2 στάδια
 - Alignment πάνω σε transcriptome και read count
 - Στατιστική Ανάλυση και pathway analysis
- Μπορεί να χρησιμοποιηθεί πάνω σε mRNAs και miRNAs με διαφορετικά pipeline για το καθένα

ensembl	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	
ENSG00000103569	395.1546867	6.485820338	0.380377799	17.05099603	3.44E-65	8.14E-61	
ENSG00000162747	470.1872928	5.742798983	0.349090355	16.45075234	8.28E-61	9.81E-57	
ENSG00000143546	1054.332041	6.089059037	0.378057651	16.10616535	2.31E-58	1.82E-54	
ENSG00000255398	479.4583455	6.296735991	0.394427284	15.96425058	2.27E-57	1.34E-53	
ENSG00000169429	833.0229258	5.770030574	0.364670649	15.82258017	2.17E-56	1.03E-52	
ENSG00000149968	4293.281475	6.7841815	0.443133008	15.30958285	6.60E-53	2.61E-49	
ENSG00000163464	165.8081686	6.410991973	0.419111048	15.29664274	8.05E-53	2.72E-49	
ENSG00000157551	231.8355143	5.810430644	0.387379035	14.99934202	7.42E-51	2.20E-47	
ENSG00000203747	881.8225046	3.281300895	0.223258066	14.69734533	6.70E-49	1.66E-45	
ENSG00000163220	2131.626531	5.206438449	0.354313868	14.69442468	7.00E-49	1.66E-45	
ENSG00000184956	37.80892497	24.04597247	1.640622407	14.65661591	1.22E-48	2.63E-45	
ENSG00000196611	5553.711467	6.028803877	0.415531897	14.50864284	1.07E-47	2.11E-44	
ENSG00000140274	546.8379873	5.284472721	0.375468196	14.07435509	5.46E-45	9.95E-42	
ENSG00000163421	88.66321553	6.134933722	0.447268433	13.71644692	8.09E-43	1.37E-39	
ENSG00000123610	213.3111931	5.555693248	0.406886031	13.65417543	1.91E-42	3.01E-39	
ENSG00000099985	176.7986793	5.186713839	0.380530476	13.63021931	2.65E-42	3.92E-39	
ENSG00000134339	169.2528619	4.649528588	0.342255509	13.58496349	4.92E-42	6.85E-39	
ENSG00000163735	1052.78467	5.020379809	0.373148432	13.45410934	2.91E-41	3.83E-38	
ENSG00000163221	87.29142814	6.594850941	0.497386505	13.25900657	4.00E-40	4.99E-37	
ENSG00000125538	1408.987545	4.813114647	0.365926837	13.15321577	1.63E-39	1.93E-36	
ENSG00000085265	337.6442983	3.383346632	0.260936304	12.96617827	1.90E-38	2.15E-35	
ENSG00000171051	457.416417	4.356148964	0.336899522	12.93011321	3.04E-38	3.28E-35	
ENSG00000150337	193.6648928	3.531393813	0.278439772	12.68279234	7.37E-37	7.59E-34	
ENSG00000258227	46.1933647	6.254938937	0.493871566	12.66511249	9.23E-37	9.11E-34	
ENSG00000182782	291.9296006	4.346919717	0.344388716	12.62213166	1.59E-36	1.51E-33	
ENSG00000108342	125.1942536	7.438191246	0.59323442	12.53836764	4.60E-36	4.20E-33	
ENSG00000133048	1623.019741	4.405669986	0.352164714	12.51025391	6.56E-36	5.76E-33	
ENSG00000124731	107.4973978	4.782902761	0.3842842	12.44626442	1.47E-35	1.24E-32	
ENSG00000166670	183.498394	5.414090695	0.435193209	12.44065989	1.57E-35	1.28E-32	
ENSG00000136689	824.9608588	4.310801638	0.352243048	12.23814539	1.94E-34	1.54E-31	
ENSG00000096088	162.375491	6.56334203	0.536438593	12.23502954	2.02E-34	1.54E-31	
ENSG00000186407	168.5371786	4.556525679	0.378786668	12.02926625	2.49E-33	1.85E-30	
ENSG00000183019	42.09179229	6.155277543	0.51605114	11.92765031	8.49E-33	6.10E-30	
ENSG00000119535	1010.95183	3.411126199	0.294783539	11.57163054	5.74E-31	4.00E-28	
ENSG00000128383	146.2192839	3.691837809	0.319565744	11.55267072	7.16E-31	4.84E-28	
ENSG00000142748	24.81730327	3.71795333	0.326014283	11.40426518	3.98E-30	2.62E-27	
ENSG00000262156	28.94885427	7.290261965	0.640192108	11.38761611	4.82E-30	3.09E-27	

Βιοπληροφορική στην έκφραση

Ανάλυση
διαφορικής
έκφρασης

Γονιδίων και
ομάδων
γονιδίων



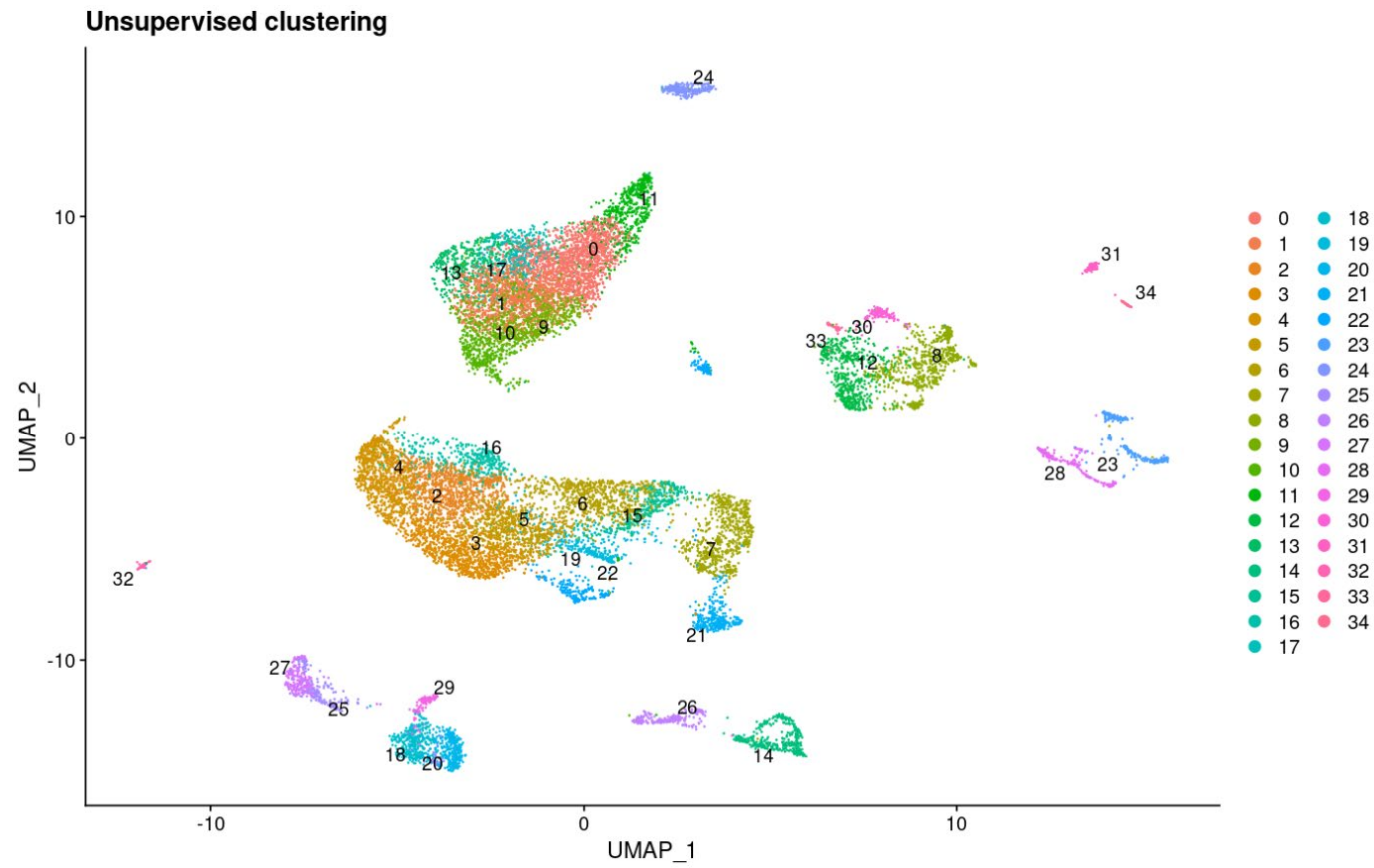
Single-Cell RNAseq

- Ίδια φιλοσοφία με το προηγούμενο (bulk) αλλά
 - Χρησιμοποιούμε διαφορετικό τρόπο για την βιβλιοθήκη
 - Διαφορετικό τύπο sequencing μηχανήματος
 - Παίρνουμε διαφορετικούς τύπους αρχείων για την ανάλυση μας
- Σκοπός είναι να μελετήσουμε τις διαφορές μεταξύ διαφορετικών κυττάρων

Single-Cell RNAseq

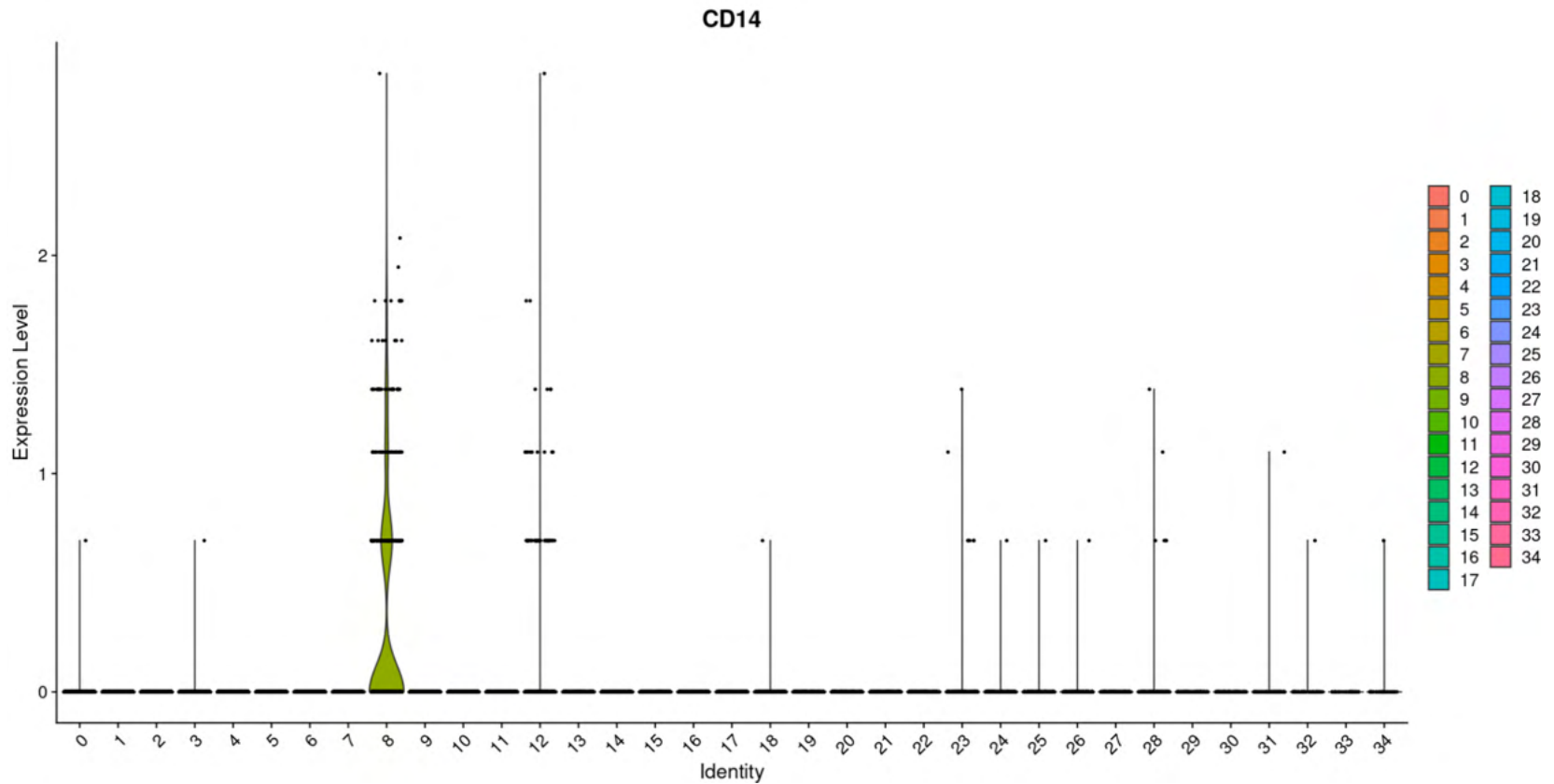
- Βήματα
 - Εισαγωγή δεδομένων (reads ανά κύτταρο, κύτταρα ανά sample)
 - Αφαίρεση νεκρών κυττάρων (περιεκτικότητα σε μιτοχονδριακά reads)
 - Αφαίρεση Doublets (κυττάρων που συνενώθηκαν (τεχνητά κατά το χτίσιμο της βιβλιοθήκης)
 - Κανονικοποίηση (μεταξύ κυττάρων και μεταξύ samples)
 - Clustering κυττάρων
 - Differential expression μεταξύ clusters
 - Χαρακτηρισμός των clusters

Single-Cell RNAseq



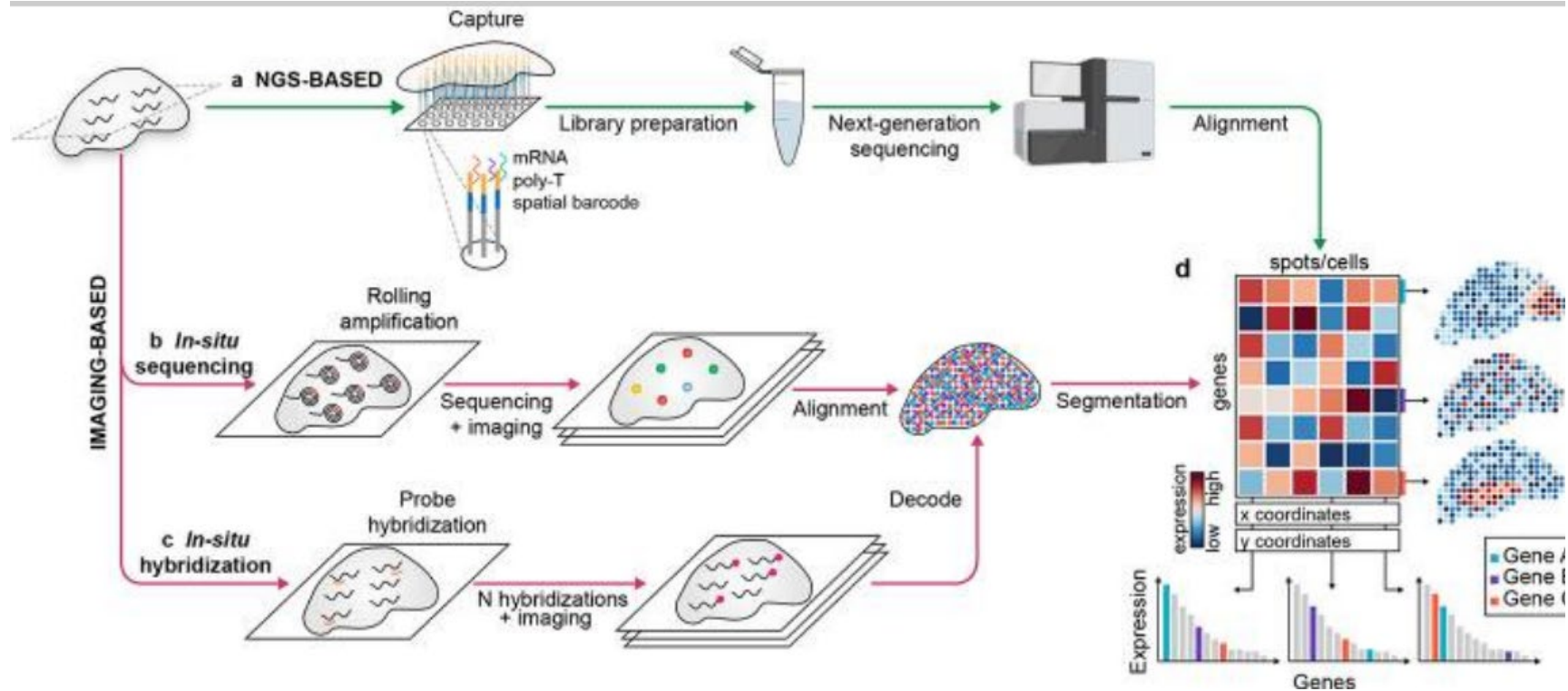
Single-Cell RNAseq

CD14 = marker molecule for
monocytes and macrophages



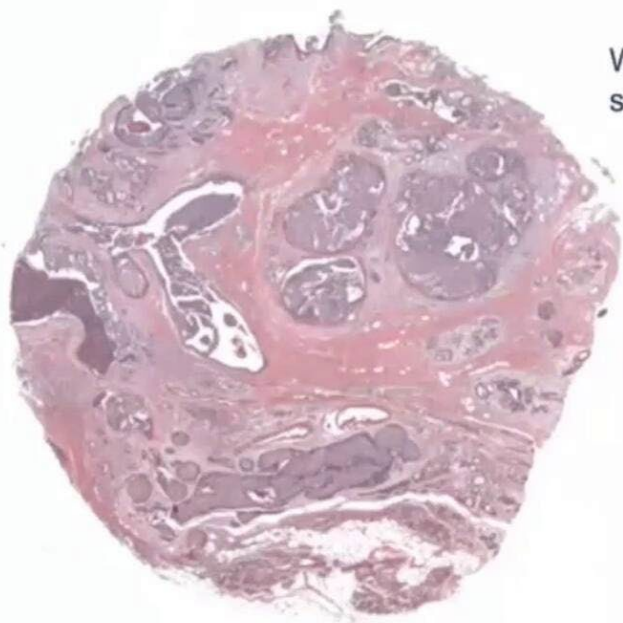
Spatial Transcriptomics

PMID: [34381231](#)



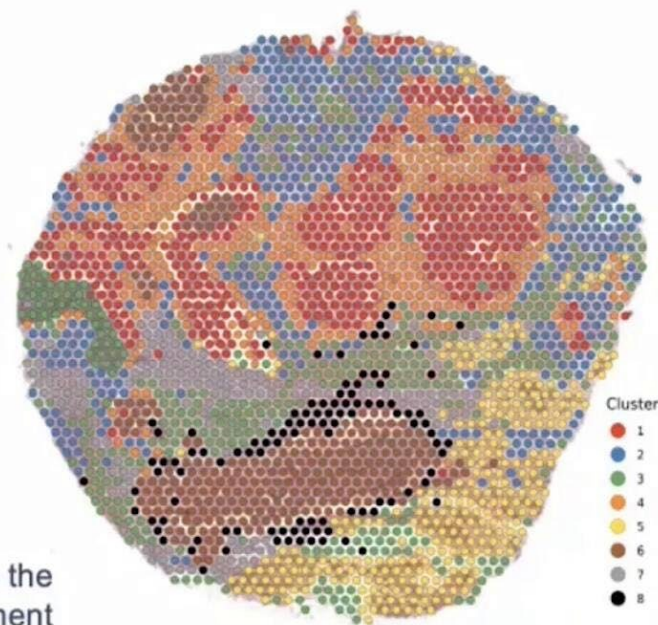
a, NGS-based spatial transcriptomic methods barcode transcripts according to their location in a lattice of spots. **b**, *In situ* sequencing approaches directly read out the transcript sequence within the tissue. **c**, *In situ* hybridization methods detect target sequences by hybridization of complementary fluorescent probes. **d**, The product of spatial transcriptomics is the gene expression matrix - where the rows and columns correspond to genes and locations.

Significance of spatial biology



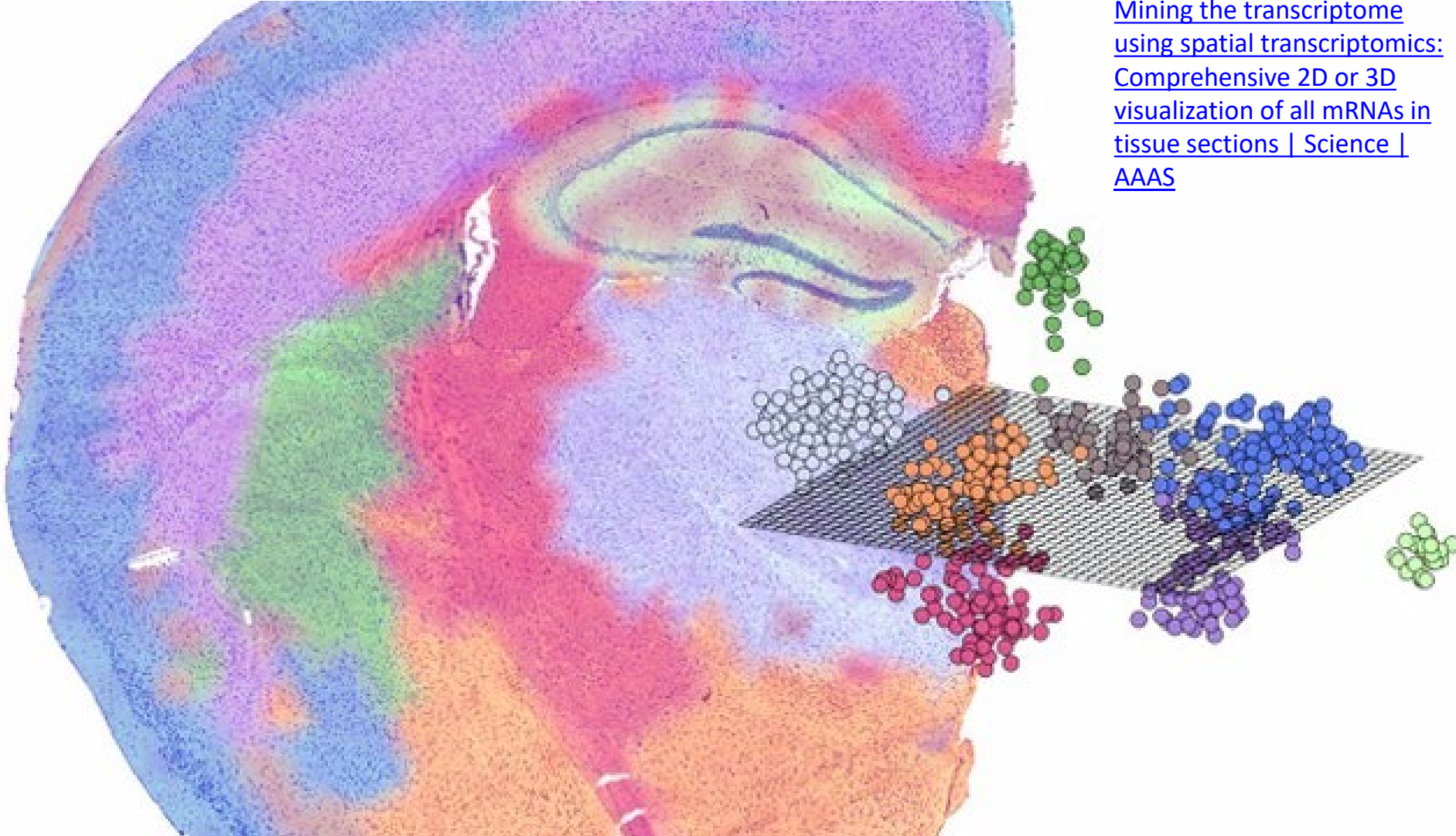
Human breast ductal carcinoma in situ

What's lying beneath the surface?



Visium reveals the complex environment

[Mining the transcriptome
using spatial transcriptomics:
Comprehensive 2D or 3D
visualization of all mRNAs in
tissue sections | Science |
AAAS](#)



INTEGRATION -ΣΥΝΕΝΩΣΗ

