Γονιδιωματική

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

Microbial Genomics

1992: the sequencing of an entire chromosome, chromosome III, of the unicellular eukaryotic microbe Saccharomyces cerevisiae, by Steve Oliver and colleagues.

Comparative Study > Nature. 1992 May 7;357(6373):38-46. doi: 10.1038/357038a0.

The complete DNA sequence of yeast chromosome III

S G Oliver [1], Q J van der Aart, M L Agostoni-Carbone, M Aigle, L Alberghina, D Alexandraki, G Antoine, R Anwar, J P Ballesta, P Benit, et al.

1995: the publication of the first complete bacterial genome of the opportunistic human pathogen Haemophilus influenza, by Craig Venter and colleagues

> Science. 1995 Jul 28;269(5223):496-512. doi: 10.1126/science.7542800.

Whole-genome random sequencing and assembly of Haemophilus influenzae Rd

R D Fleischmann [1], M D Adams, O White, R A Clayton, E F Kirkness, A R Kerlavage, C J Bult, J F Tomb, B A Dougherty, J M Merrick, et al.

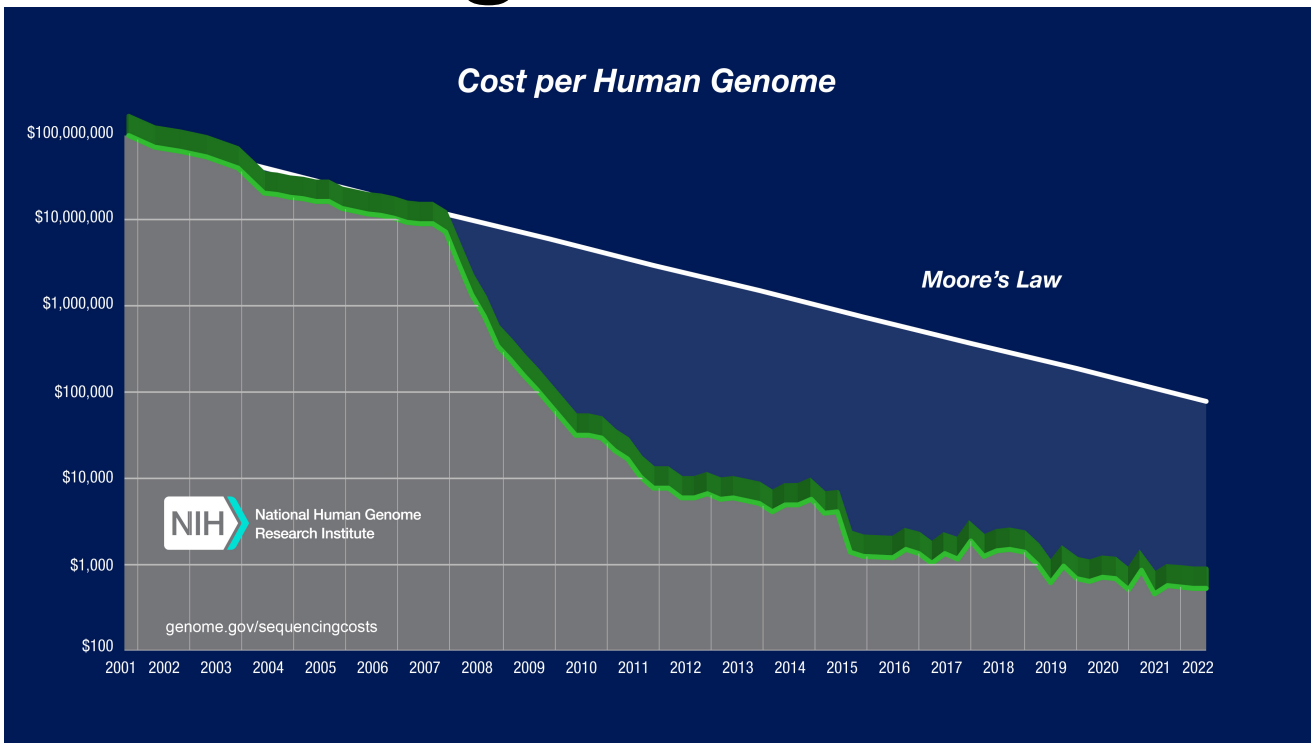
Οι τεχνολογίες

The main sequencing technologies

Short read technologies
- Illlumina
- BGI/MGI

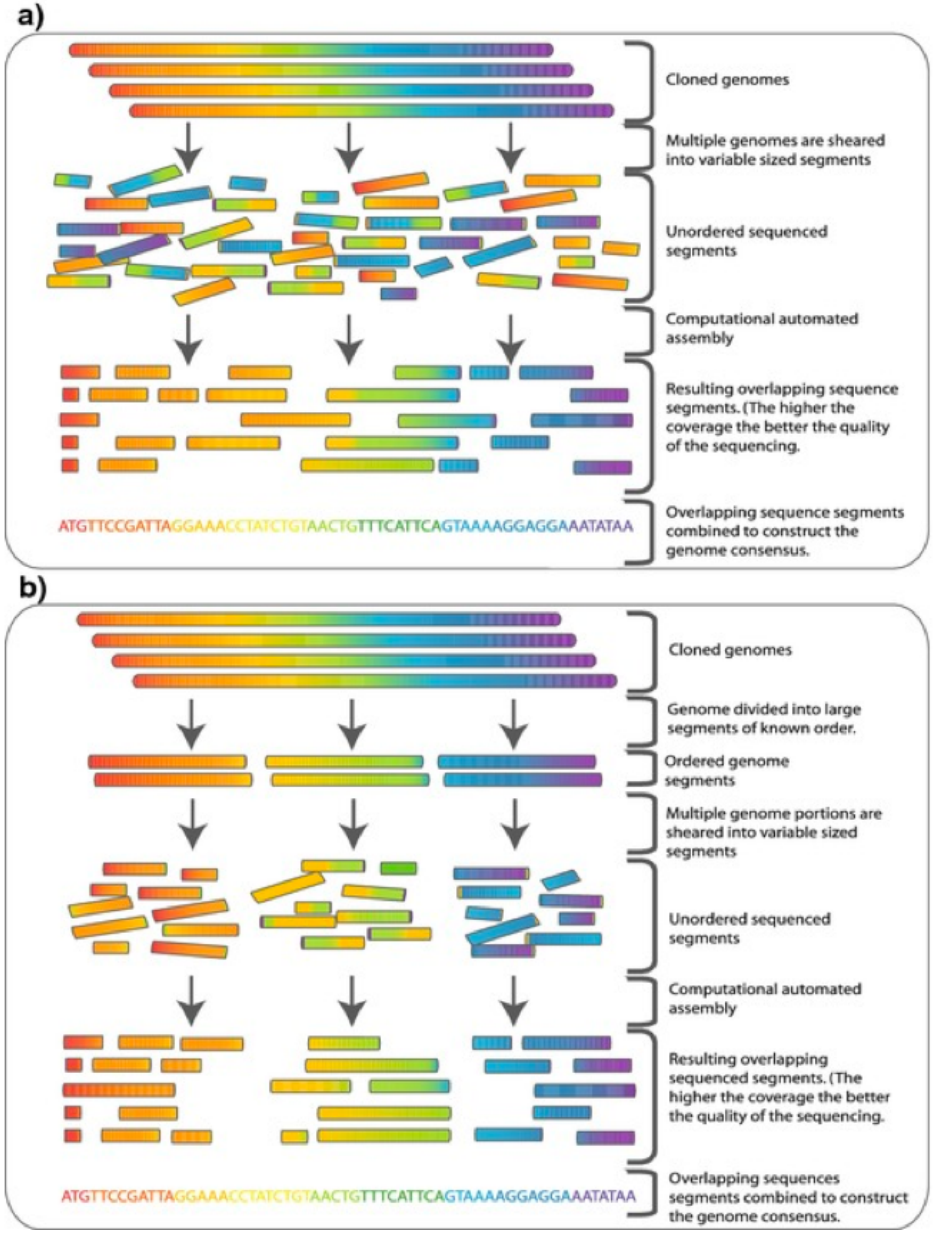
Long read technologies
- Pacific Biosciences
- Oxford Nanopore

These two technologies can sequence non-amplified input DNA, and tend to require simpler protocols (minimum time of 10 min for library construction) than those needed for short-read sequencing (several days)

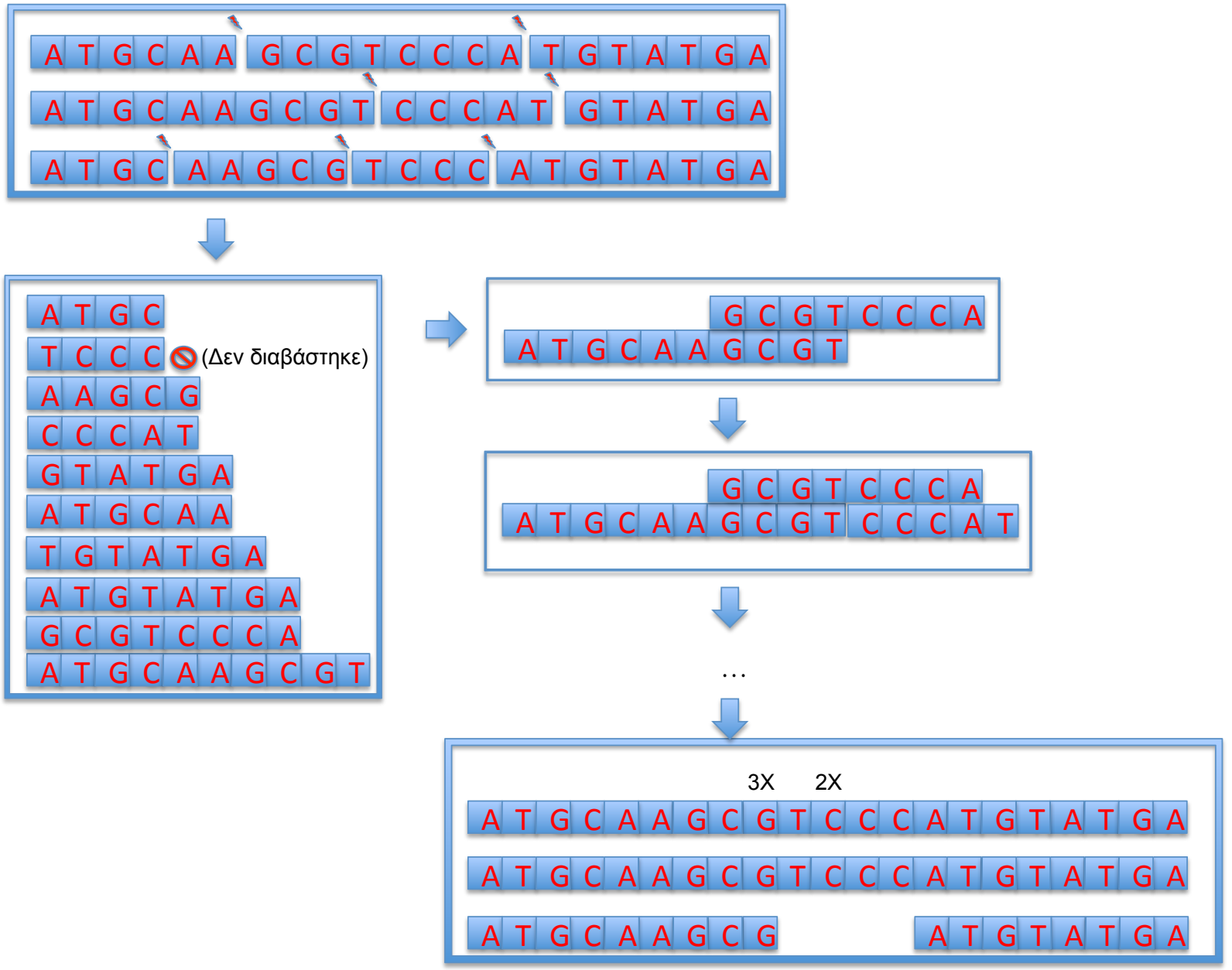


https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

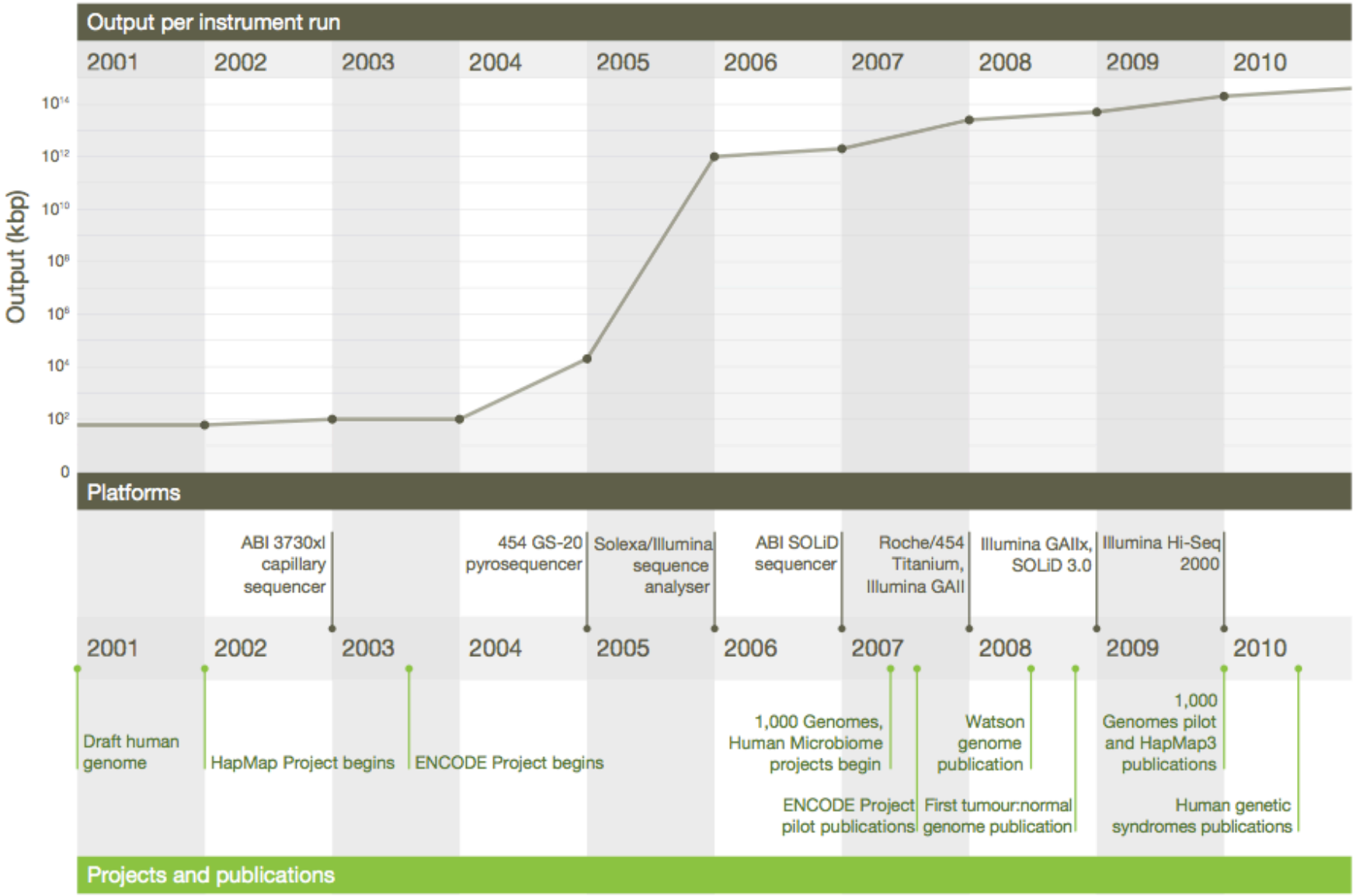
Shotgun sequencing



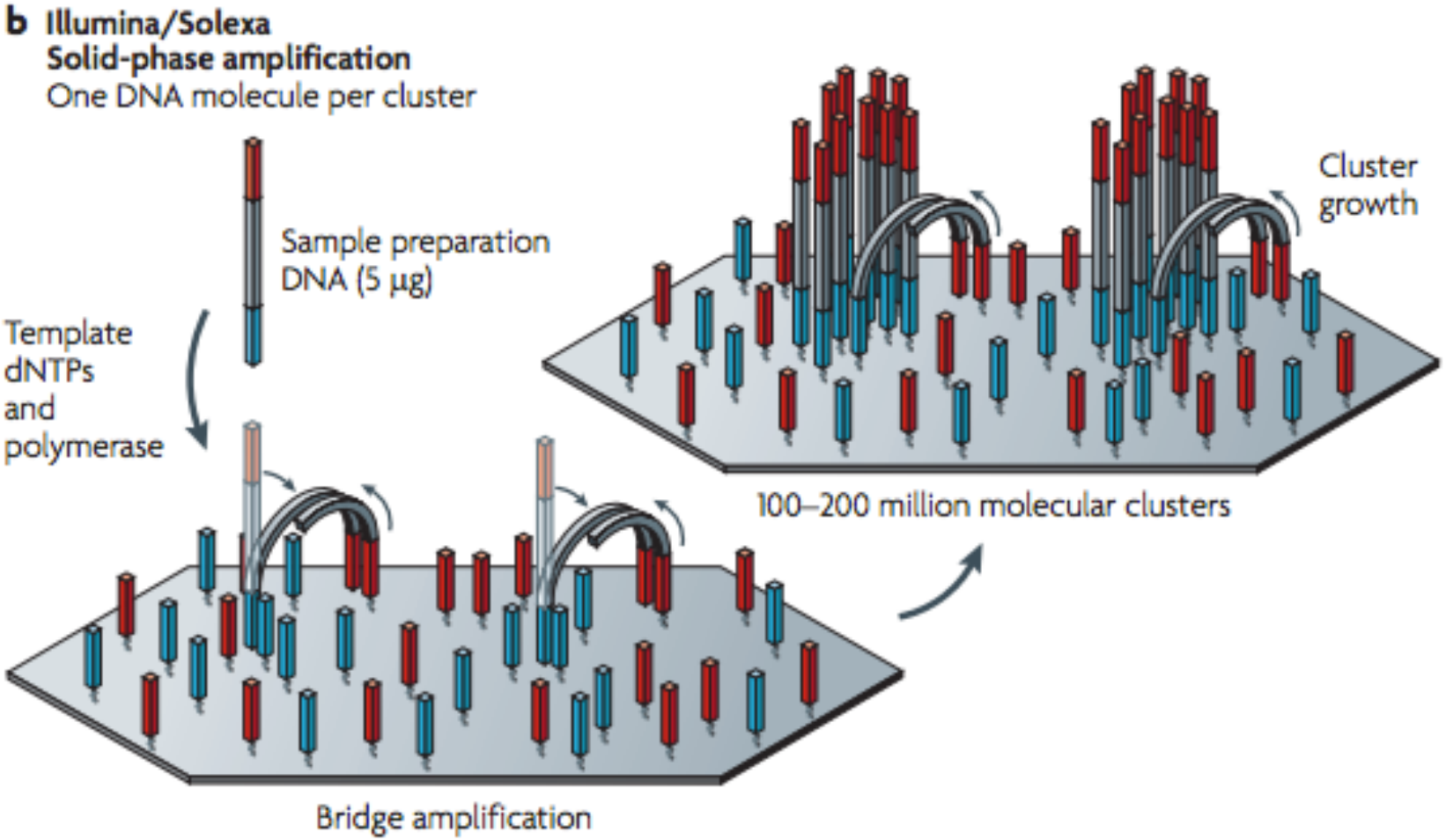
Shotgun sequencing - Βάθος αλληλούχισης (coverage)



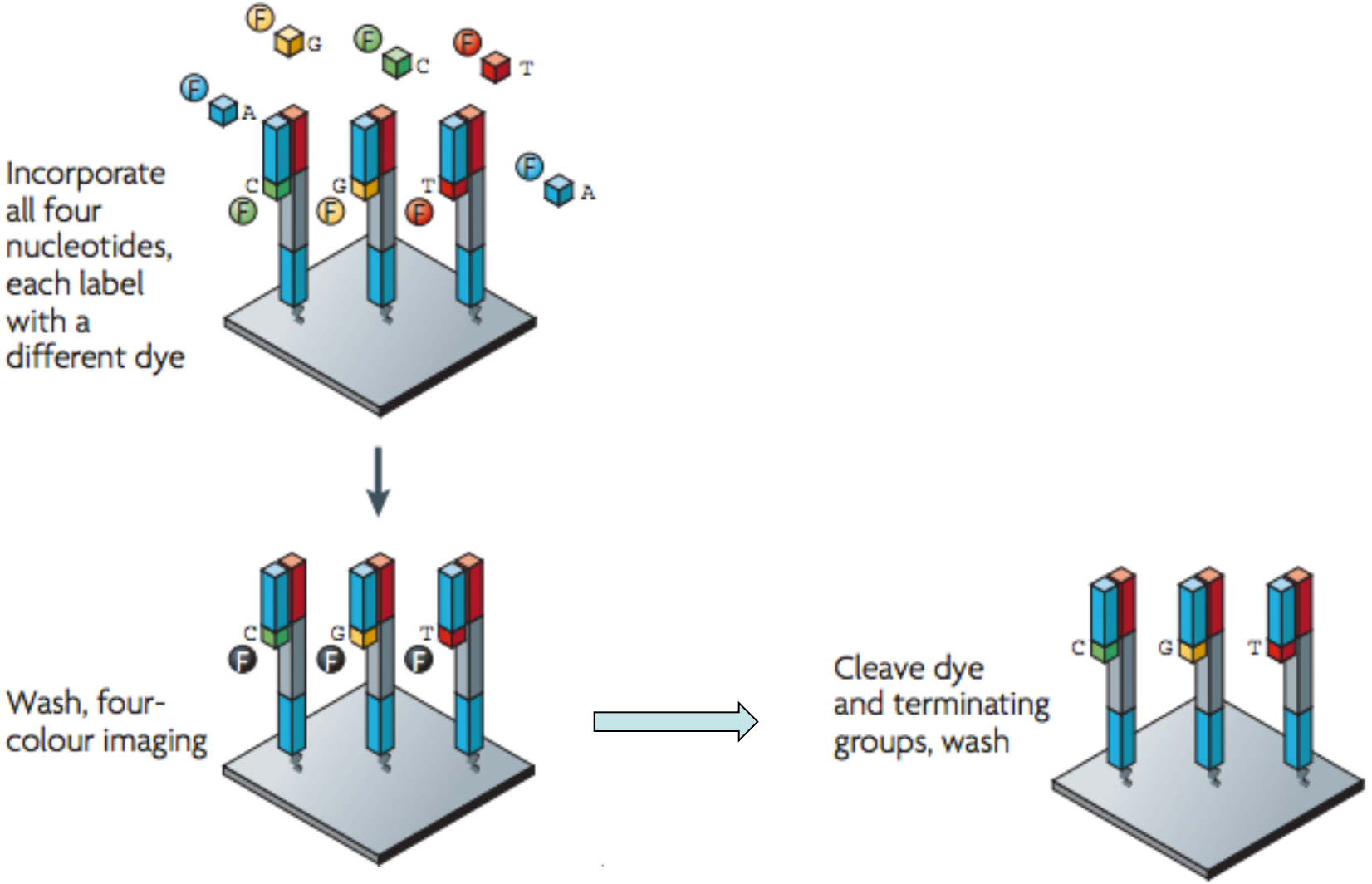
- http://www.nature.com/nature/journal/v470/n7333/pdf/nature09796.pdf
- A decade's perspective on DNA sequencing technology
- Elaine R. Mardis



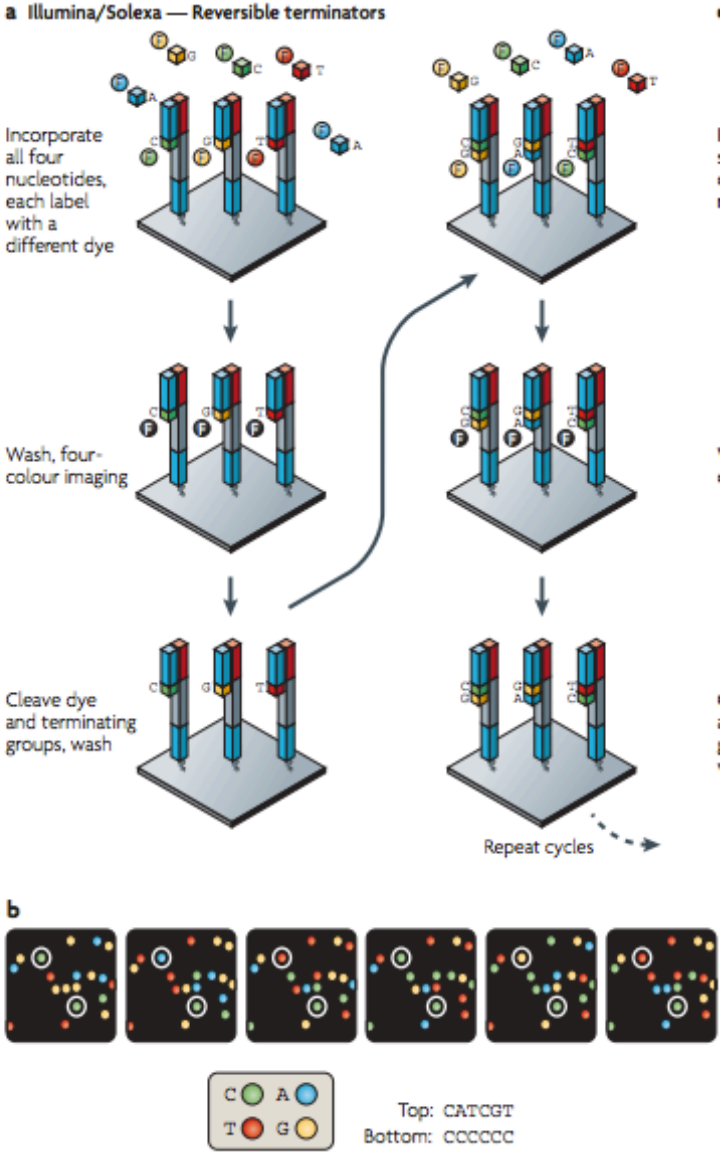
Illumina
http://www.youtube.com/watch?v=77r5p8IBwJk&feature=related



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.

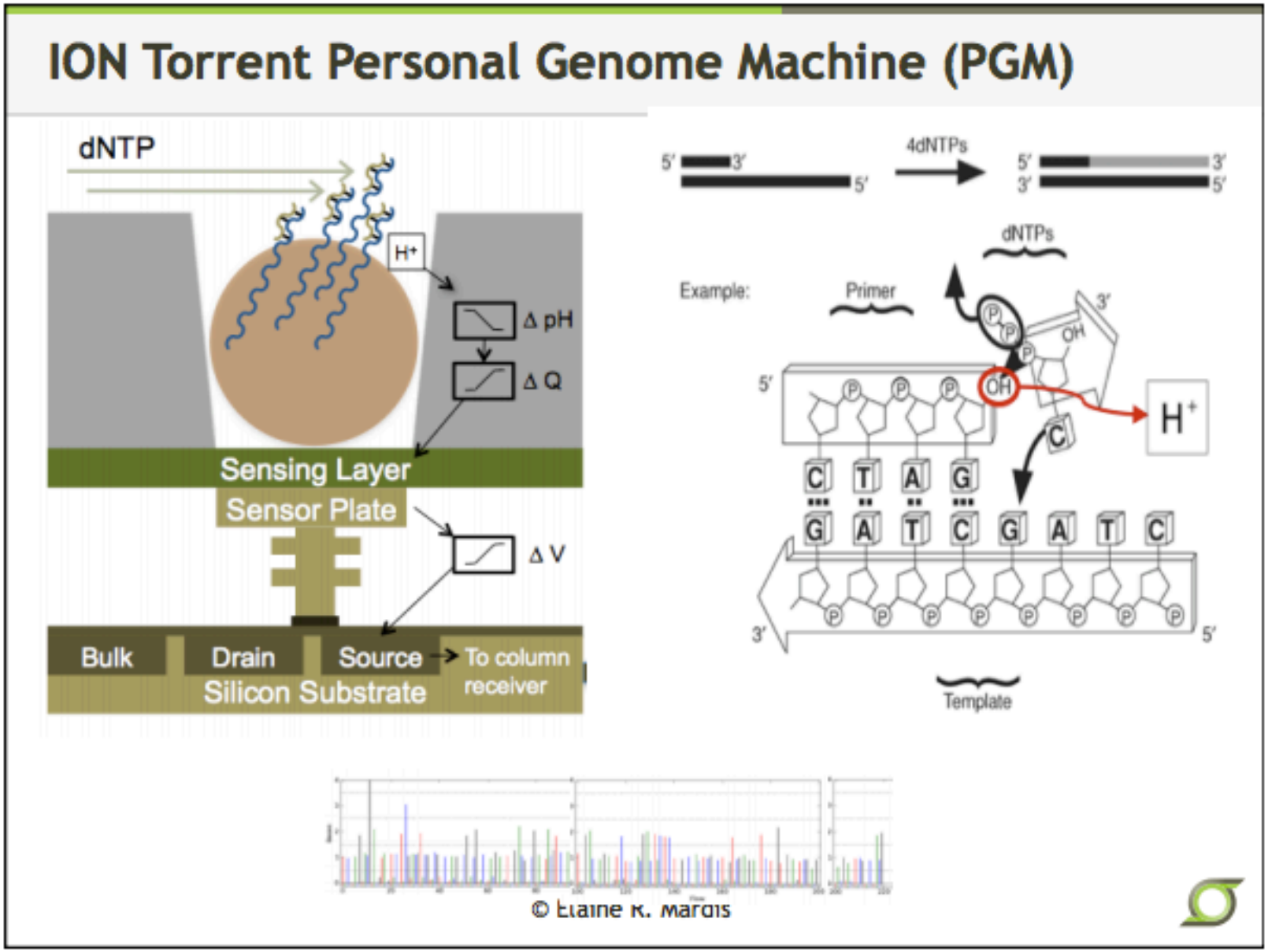
Ion Proton

Ion torrent chemistry

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

http://www.youtube.com/iontorrent

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



Εικόνα Από Elaine Mardis

Pacific Biosciences

Analytical Chemistry REVIEW

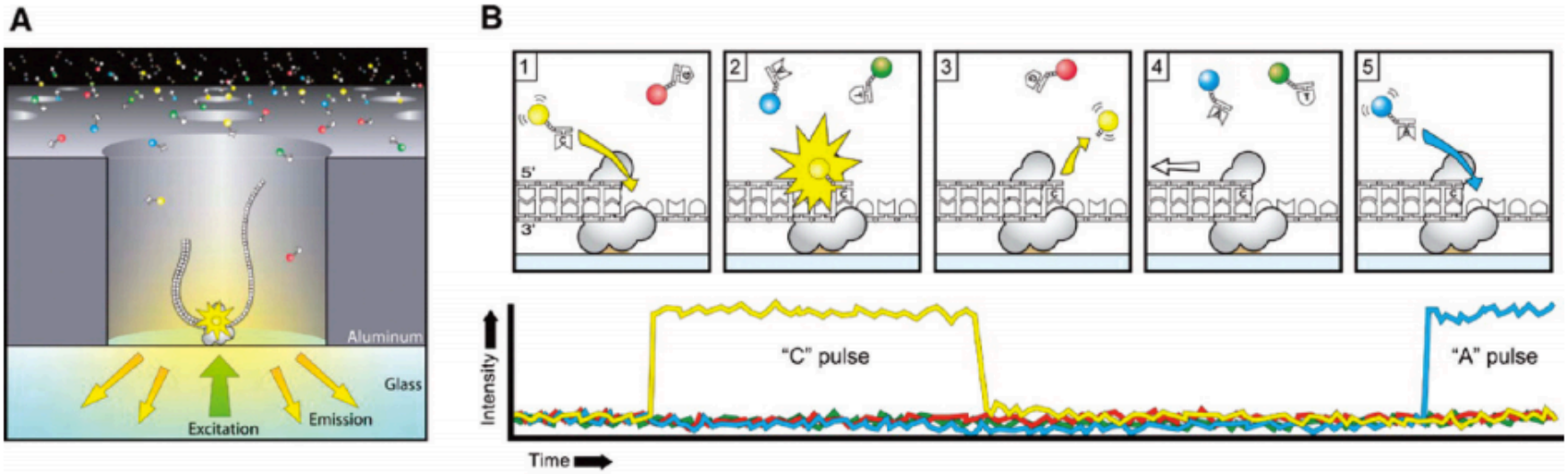


Figure 2. Schematic of PacBio's real-time single molecule sequencing. (A) The side view of a single ZMW nanostructure containing a single DNA polymerase (Φ29) bound to the bottom glass surface. The ZMW and the confocal imaging system allow fluorescence detection only at the bottom surface of each ZMW. (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

Pacific Biosciences

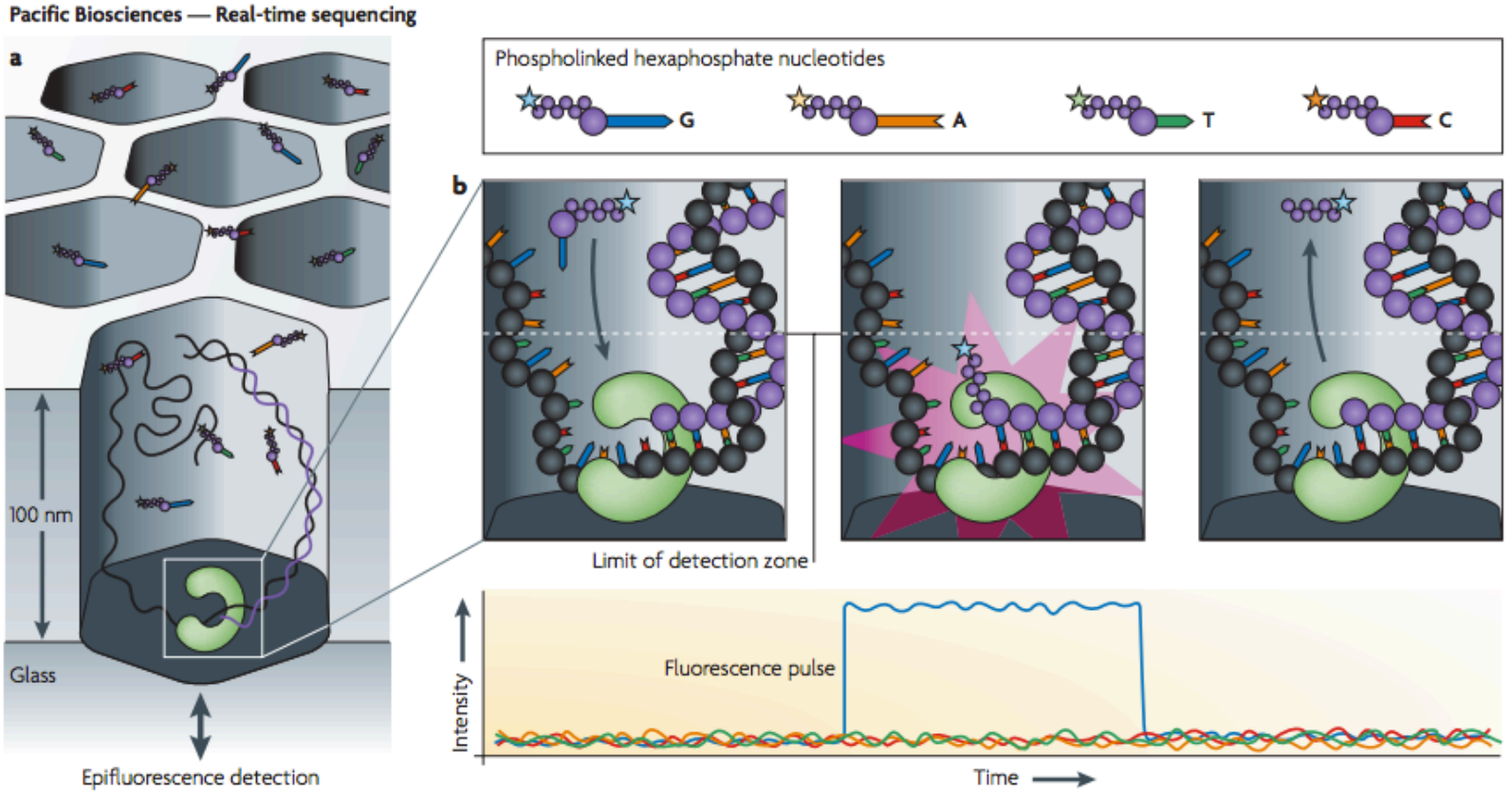
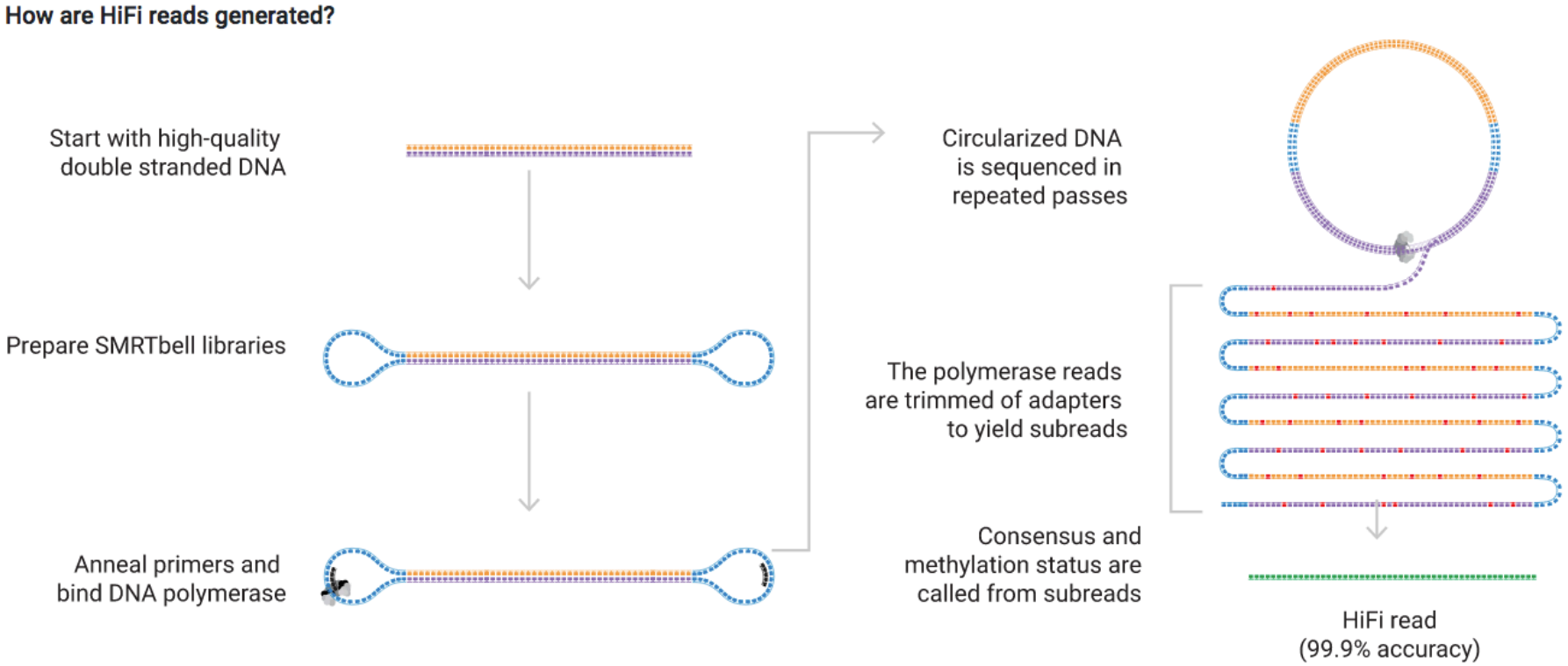


Figure 4 | **Real-time sequencing.** Pacific Biosciences' four-colour real-time sequencing method is shown. **a** | The zero-mode waveguide (ZMW) design reduces the observation volume, therefore reducing the number of stray fluorescently labelled molecules that enter the detection layer for a given period. These ZMW detectors address the dilemma that DNA polymerases perform optimally when fluorescently labelled nucleotides are present in the micromolar concentration range, whereas most single-molecule detection methods perform optimally when fluorescent species are in the pico- to nanomolar concentration range[42]. **b** | The residence time of phospholinked nucleotides in the active site is governed by the rate of catalysis and is usually on the millisecond scale. This corresponds to a recorded fluorescence pulse, because only the bound, dye-labelled nucleotide occupies the ZMW detection zone on this timescale. The released, dye-labelled pentaphosphate by-product quickly diffuses away, dropping the fluorescence signal to background levels. Translocation of the template marks the interphase period before binding and incorporation of the next incoming phospholinked nucleotide.

Pacific Biosciences HiFi reads

- Initially, PacBio reads had a high error rate (15%) using Single Molecular, Real-Time (SMRT) sequencing.
- Later, they developed high-fidelity reads (HiFi) with an error rate lower than 1%.
- HiFi reads are typically 15–20 kbp long.
- Sequel II or Revio instruments

https://www.nature.com/articles/s41467-024-44804-3
https://www.youtube.com/watch?v=_ID8JyAbwEo



https://www.pacb.com/technology/hifi-sequencing/

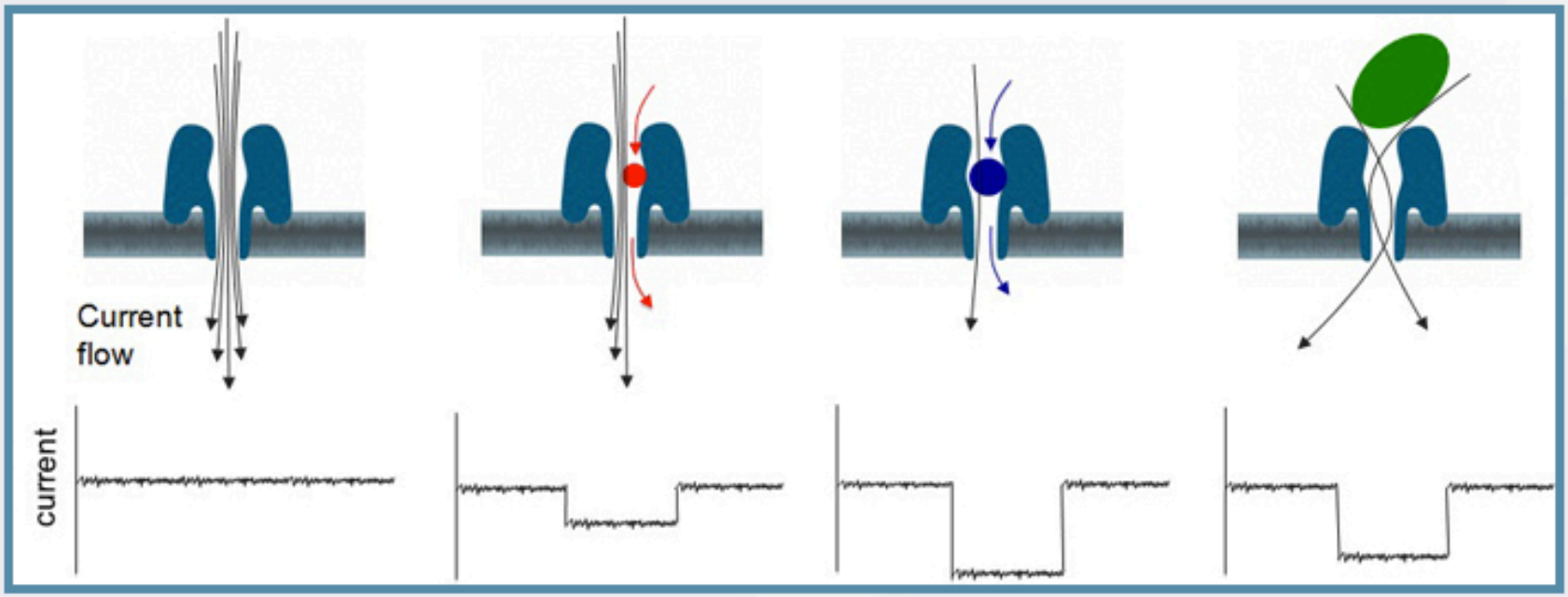
Oxford Nanopore

Longer reads, up to 4Mbp!
New R10 chemistry provides high accuracy at more than 99%.
https://www.youtube.com/watch?v=RcP85JHLmnI

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

Biological Nanopore

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

Oxford Nanopore

- Capable of displaying the generated sequence in real-time, enabling adaptive sampling of the reads being sequenced.
- Very portable, with a proven track record even in extreme environments such as the International Space Station.

Nanopore DNA Sequencing and Genome Assembly on the International Space Station

Sarah L. Castro-Wallace, Charles Y. Chiu, Kristen K. John, Sarah E. Stahl, Kathleen H. Rubins, Alexa B. R. McIntyre, Jason P. Dworkin, Mark L. Lupisella, David J. Smith, Douglas J. Botkin, Timothy A. Stephenson, Sissel Juul, Daniel J. Turner, Fernando Izquierdo, Scot Federman, Doug Stryke, Sneha Somasekar, Noah Alexander, Guixia Yu, Christopher E. Mason & Aaron S. Burton

Scientific Reports **7**, Article number: 18022 (2017) | Cite this article

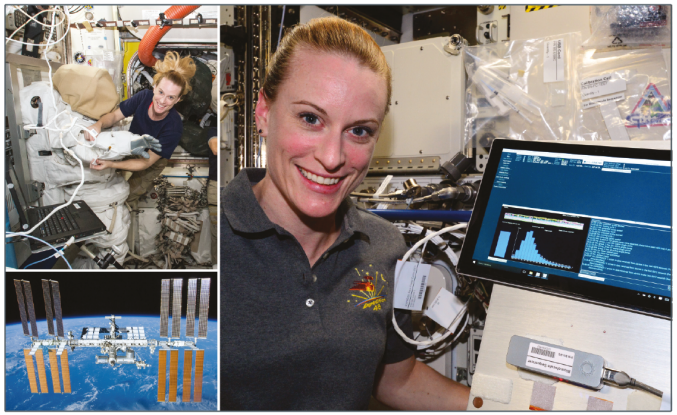


Fig. 1 Astronaut Kate Rubins on the ISS

https://nanoporetech.com/resource-centre/dna-sequencing-microgravity-international-space-station-iss-using-minion

http://genomebiology.com/content/pdf/gb-2011-12-8-125.pdf

Sboner *et al. Genome Biology* 2011, **12**:125
http://genomebiology.com/2011/12/8/125

Genome Biology

OPINION

The real cost of sequencing: higher than you think!

Andrea Sboner[1,2], Xinmeng Jasmine Mu[1], Dov Greenbaum[1,2,3,4,5], Raymond K Auerbach[1] and Mark B Gerstein*[1,2,6]

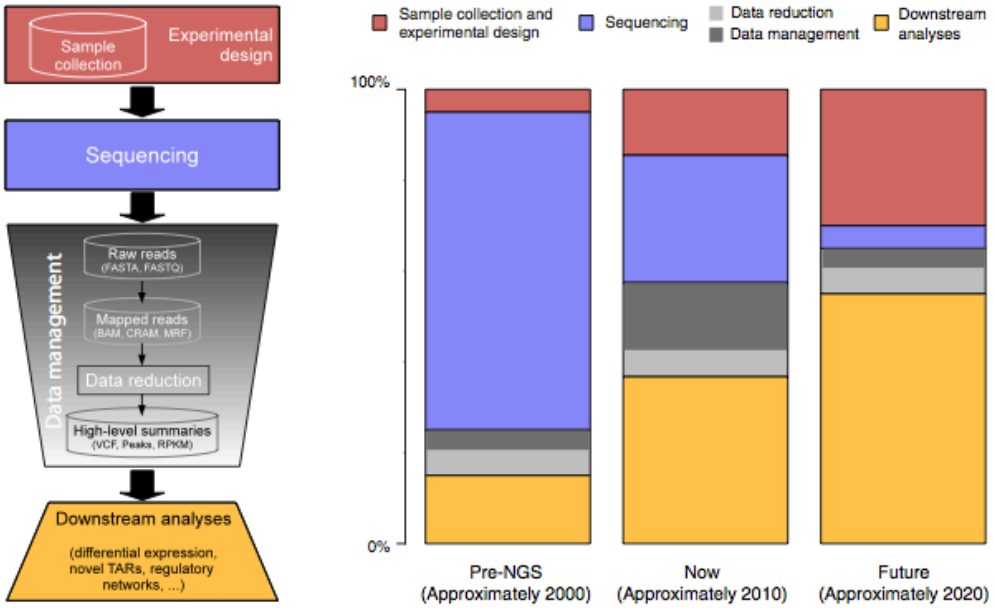
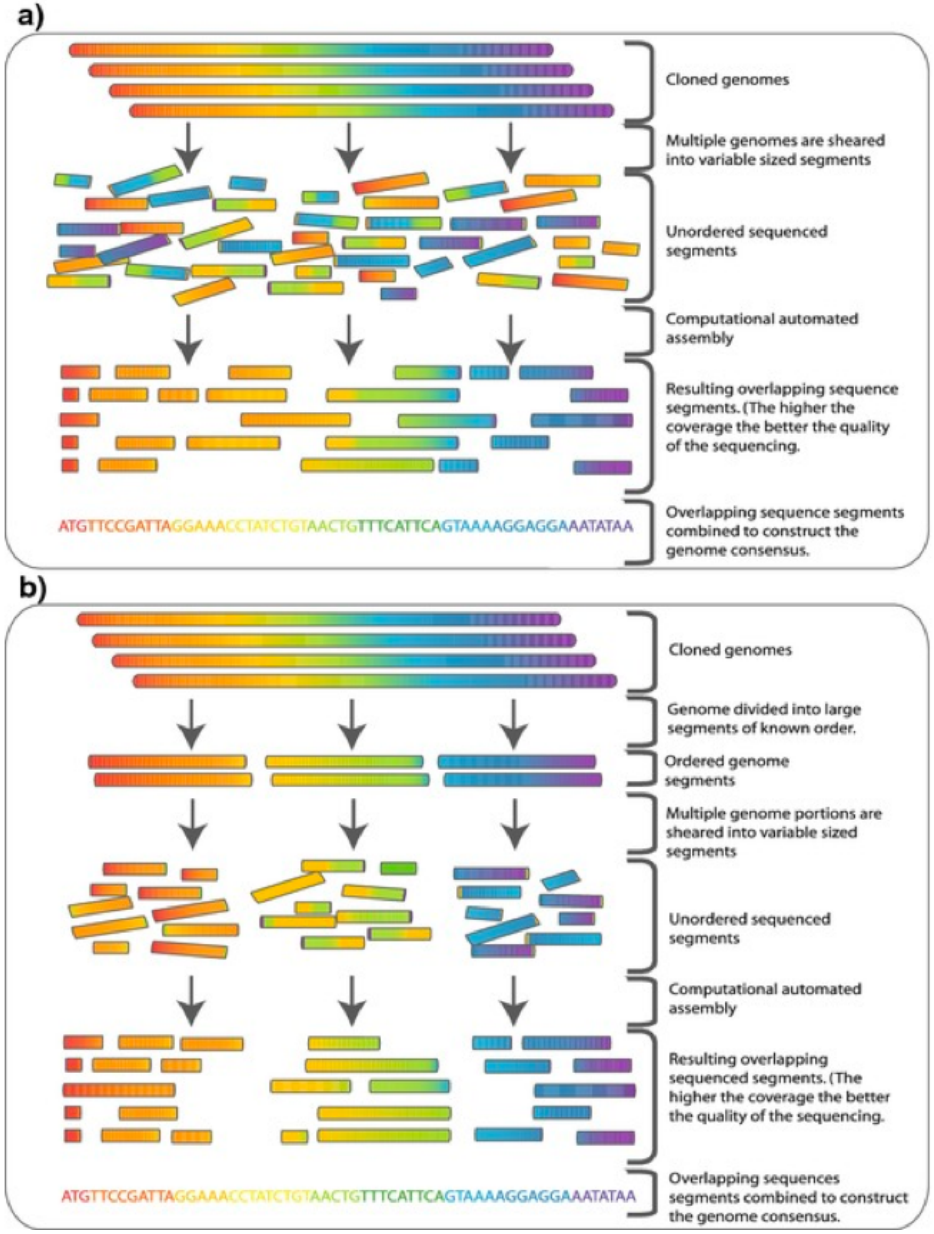


Figure 1. Contribution of different factors to the overall cost of a sequencing project across time. Left, the four-step process: (i) experimental design and sample collection, (ii) sequencing, (iii) data reduction and management, and (iv) downstream analysis. Right, the changes over time of relative impact of these four components of a sequencing experiment. BAM, Binary Sequence Alignment/Map; BED, Browser Extensible Data; CRAM, compression algorithm; MRF, Mapped Read Format; NGS, next-generation sequencing; TAR, transcriptionally active region; VCF, Variant Call Format.

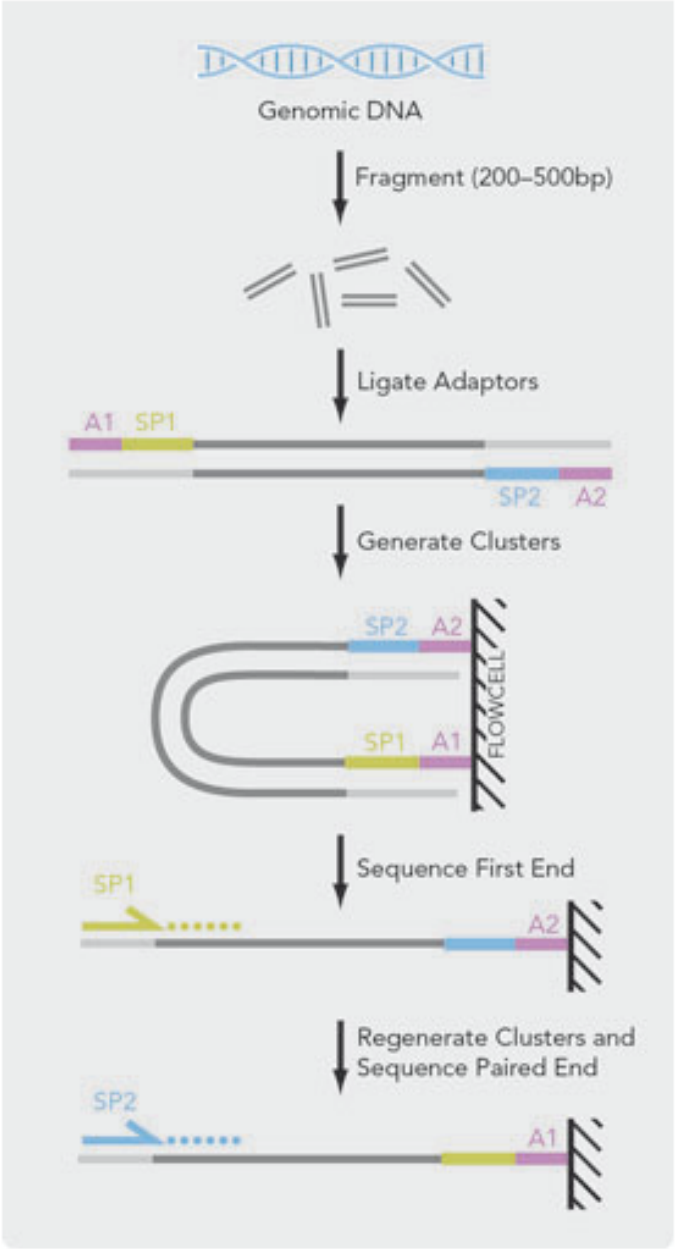
Συναρμολόγιση Γονιδιωμάτων Με Βιοπληροφορική

Shotgun sequencing

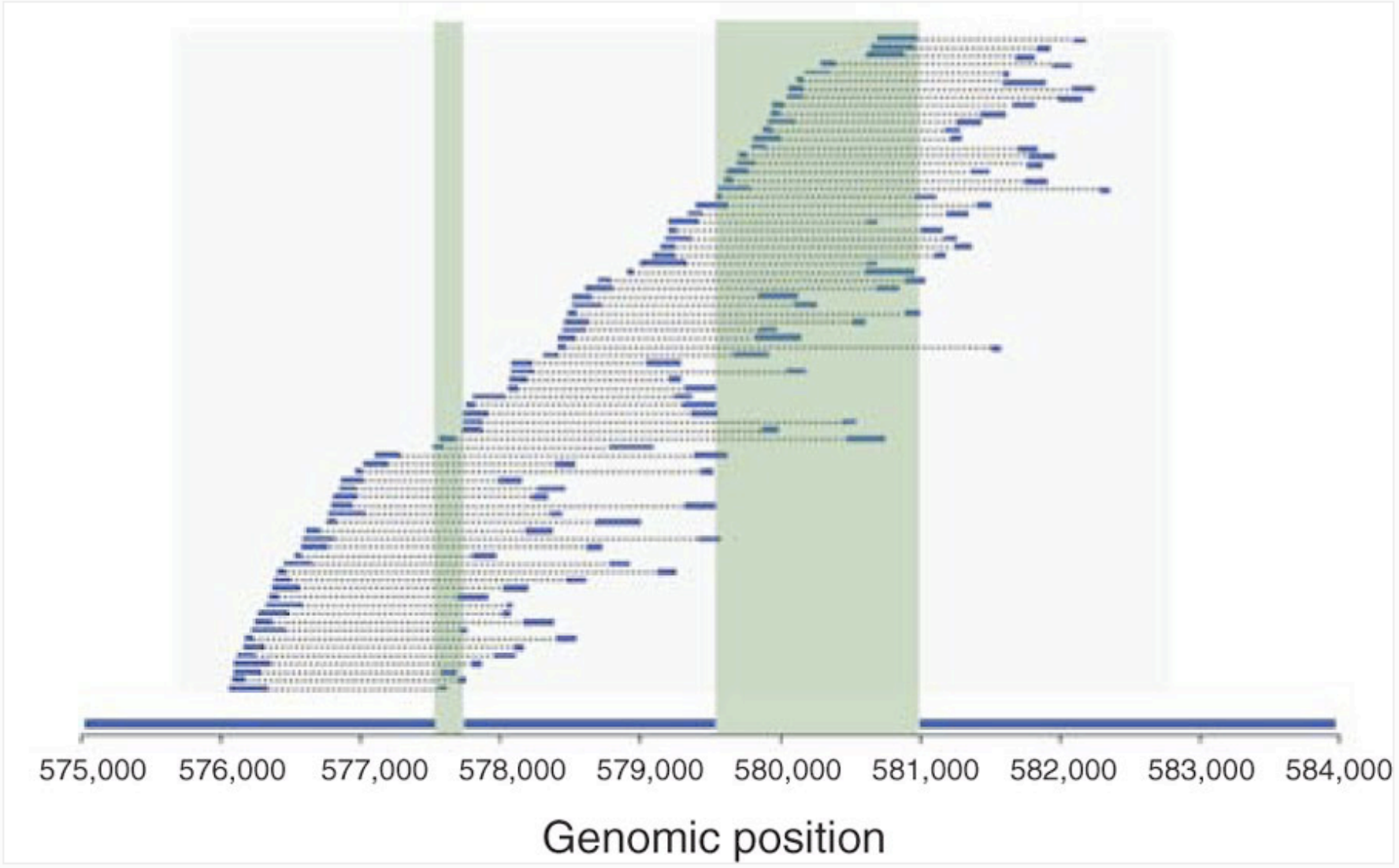


Sequencing

- Single end reads
- Paired end reads



Sequencing - paired end reads



A region of the *de novo* assembly of *E. coli* K-12, with the *de novo*–assembled contigs covering the region shown in blue along the bottom axis. The paired-end reads generated with this protocol are capable of bridging the 0.2-kb and 1.5-kb gaps between the contigs, highlighted in green.

Reads

- 454
- Illumina
- SOLID

Table 2. Comparison of mapping.

Method	Ratio of mapped reads	Accuracy per base
FLX	89.0	99.9
GA	63.7	96.7
SOLiD	47.3	99.8

Filtered data set of GA was shown.
doi:10.1371/journal.pone.0019534.t002

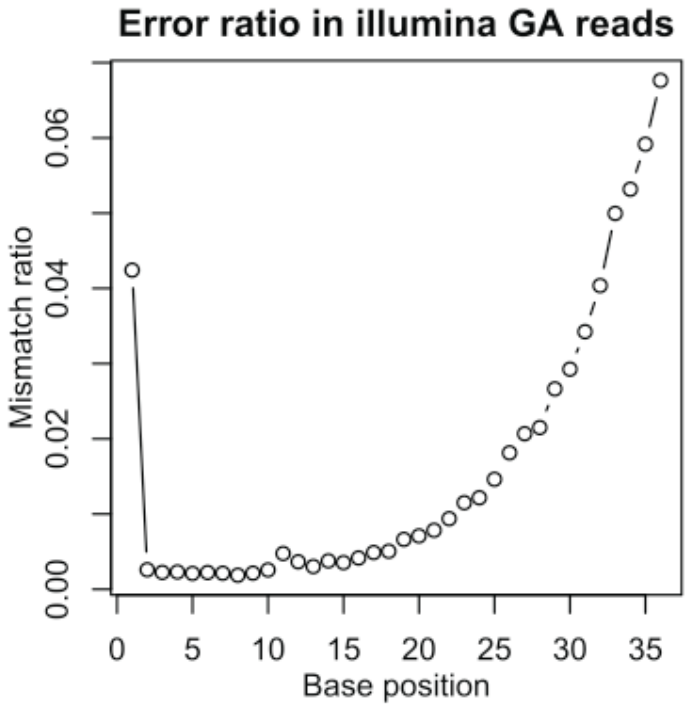


Figure 1. Error ratio in GA reads depending on the base position of the read. Ratio of mismatch between mapped reads and reference sequence to the total number of mapped reads was plotted against base position in the reads. The mismatch ratio increases along with the base position indicating decrease of accuracy of base calls.
doi:10.1371/journal.pone.0019534.g001

SOLID: ~50% reads δεν στοιχίζονται στο γονιδίωμα, από το οποίο έγινε το Sequencing!

Πρόβλημα στις χημικές αντιδράσεις μάλλον.

Εδώ, το πρόβλημα εντοπίζεται στην συσσώρευση λαθών κατά την ενσωμάτωση φθοριζόντων dNTPs.

Sequence read – Fastq format

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Τα σύμβολα στην τελευταία γραμμή αντιστοιχούν σε τιμές Q, για την κάθε μια βάση που αλληλουχίθηκε.

Το Q-score είναι μια ακέραια τιμή που προκύπτει από την πιθανότητα να έχει γίνει λάθος στην αλληλούχιση μιας συγκεκριμένης βάσης.

Αν p = πιθανότητα να έχει γίνει λάθος στην αλληλούχιση της συγκεκριμένης βάσης, τότε:

$Q=-10\log_{10}(p)$

Q=30 -> p=0.001 (πολύ καλής ποιότητας αλληλούχιση)

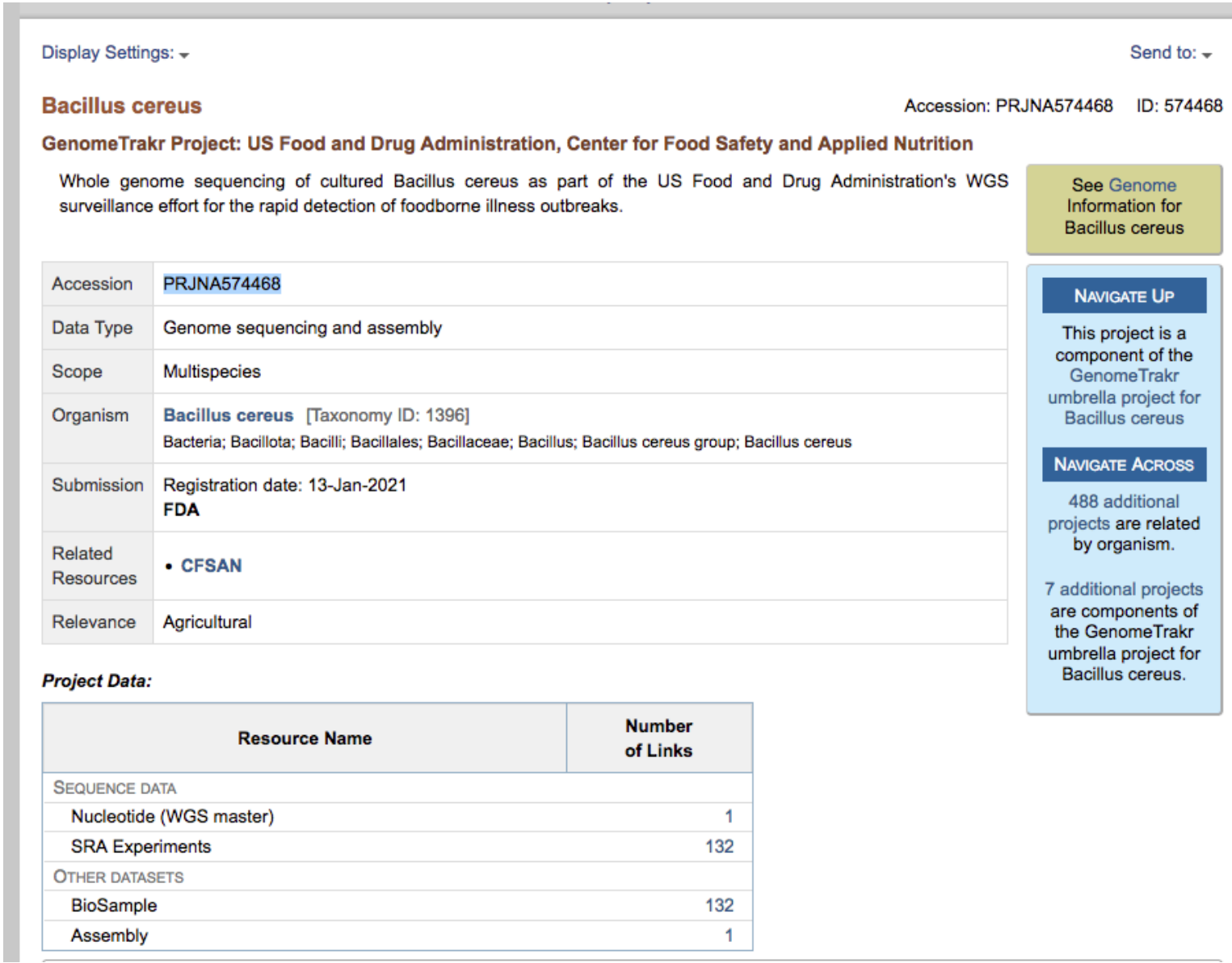
Q=13 -> p=0.05

Sequence Read Archive SRA - Bioproject

ΒΔ όπου κατατίθενται τα δεδομένα αλληλούχισης γονιδιωμάτων (raw sequence reads), είτε Whole genome sequencing (WGS) είτε RNA-Seq.

Τα δεδομένα είναι οργανωμένα ανά

Bioproject & Biosample



Display Settings: ▾ Send to: ▾

Bacillus cereus Accession: PRJNA574468 ID: 574468

GenomeTrakr Project: US Food and Drug Administration, Center for Food Safety and Applied Nutrition

Whole genome sequencing of cultured Bacillus cereus as part of the US Food and Drug Administration's WGS surveillance effort for the rapid detection of foodborne illness outbreaks.

See Genome Information for Bacillus cereus

Accession	PRJNA574468
Data Type	Genome sequencing and assembly
Scope	Multispecies
Organism	**Bacillus cereus** [Taxonomy ID: 1396] Bacteria; Bacillota; Bacilli; Bacillales; Bacillaceae; Bacillus; Bacillus cereus group; Bacillus cereus
Submission	Registration date: 13-Jan-2021 **FDA**
Related Resources	• **CFSAN**
Relevance	Agricultural

NAVIGATE UP

This project is a component of the GenomeTrakr umbrella project for Bacillus cereus

NAVIGATE ACROSS

488 additional projects are related by organism.

7 additional projects are components of the GenomeTrakr umbrella project for Bacillus cereus.

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (WGS master)	1
SRA Experiments	132
OTHER DATASETS	
BioSample	132
Assembly	1

Sequence Read Archive SRA - Biosample

Pathogen.env.1.0

Identifiers	BioSample: SAMN39831649; SRA: SRS20383910; CFSAN: CFSAN133898
Organism	Bacillus cereus cellular organisms; Bacteria; Terrabacteria group; Bacillota; Bacilli; Bacillales; Bacillaceae; Bacillus; Bacillus cereus group
Package	Pathogen: environmental/food/other; version 1.0

Attributes

collected by	CFSAN/DFPST
collection date	2023
geographic location	USA:IL
isolation source	powdered infant formula
latitude and longitude	missing
strain	DFPST-SP1
isolate name alias	CFSAN133898
source type	food
project name	GenomeTrakr
attribute_package	environmental/food/other

Links

BioProject	PRJNA574468 Bacillus cereus Retrieve all samples from this project
Submission	FDA; 2024-02-05

Accession: SAMN39831649 ID: 39831649
SRA

Sequence Read Archive SRA – SRA experiment

Περιέχει πληροφορίες για την αλληλούχιση του συγκεκριμένου δείγματος

Links from BioSample

SRX23535113: Whole genome Illumina MiSeq sequence of Bacillus cereus
1 ILLUMINA (Illumina MiSeq) run: 963,136 spots, 397.7M bases, 212.5Mb downloads

External Id: EXT00466158

Design: Whole genome library prepared from a cultured bacterial isolate.

Submitted by: FDA

Study: GenomeTrakr Project: US Food and Drug Administration, Center for Food Safety and Applied Nutrition
PRJNA574468 • SRP235182 • All experiments • All runs
show Abstract

Sample: Pathogen.env.1.0
SAMN39831649 • SRS20383910 • All experiments • All runs
Organism: Bacillus cereus

Library:
Name: Illumina DNA Prep library SEQ000139830
Instrument: Illumina MiSeq
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

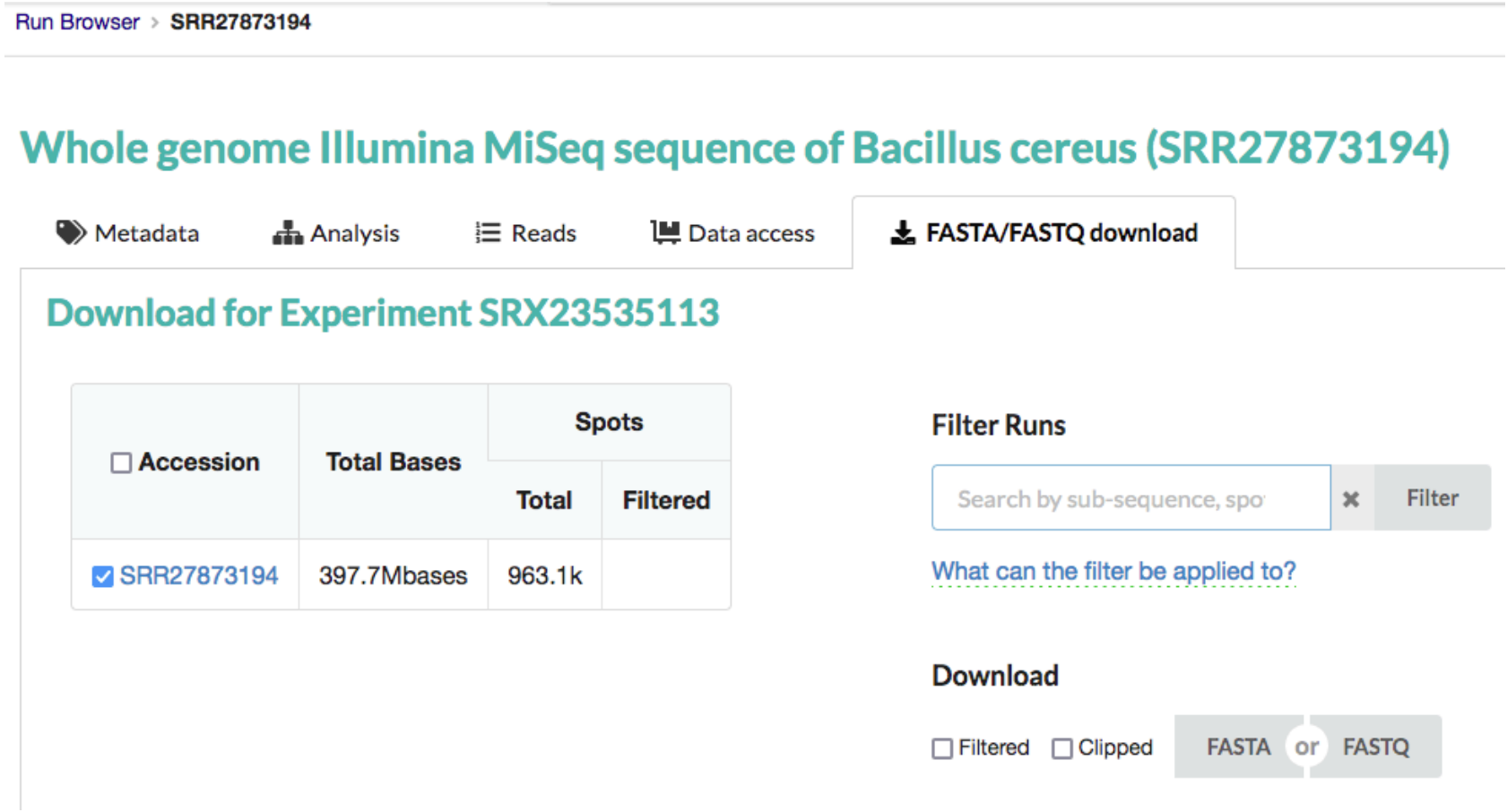
Spot descriptor:

Runs: 1 run, 963,136 spots, 397.7M bases, 212.5Mb

Run	# of Spots	# of Bases	Size	Published
SRR27873194	963,136	397.7M	212.5Mb	2024-02-05

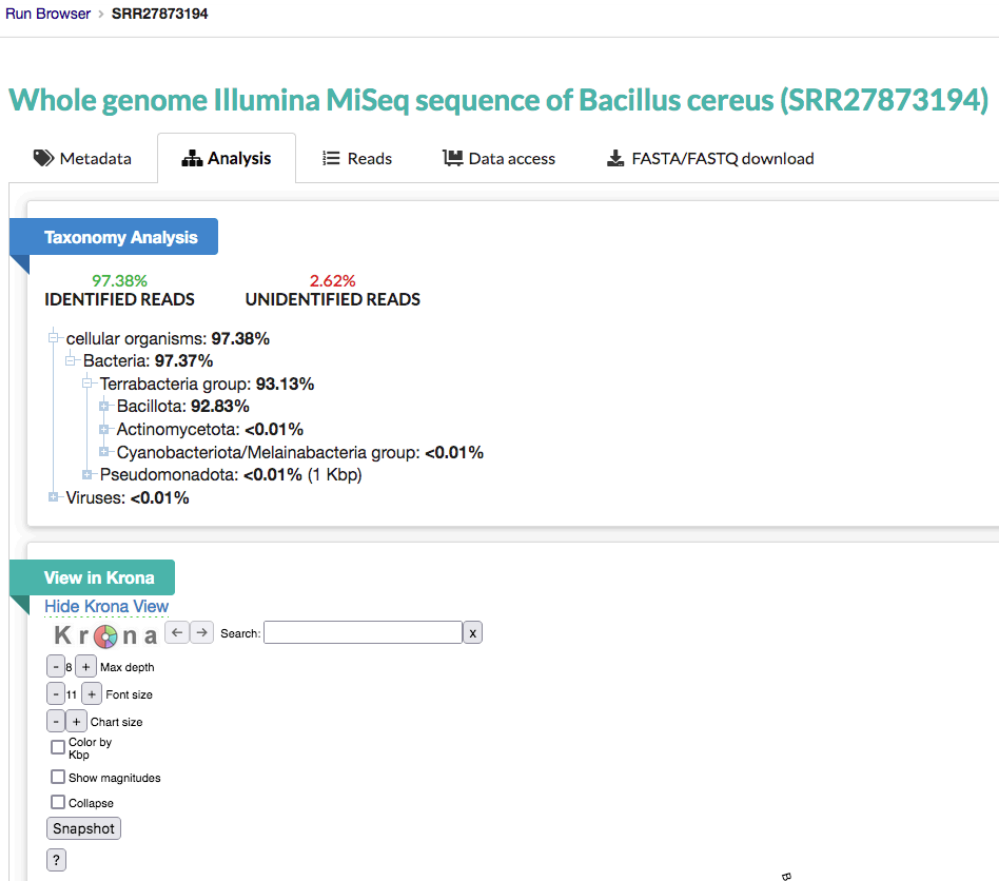
Sequence Read Archive SRA – SRA run SRR

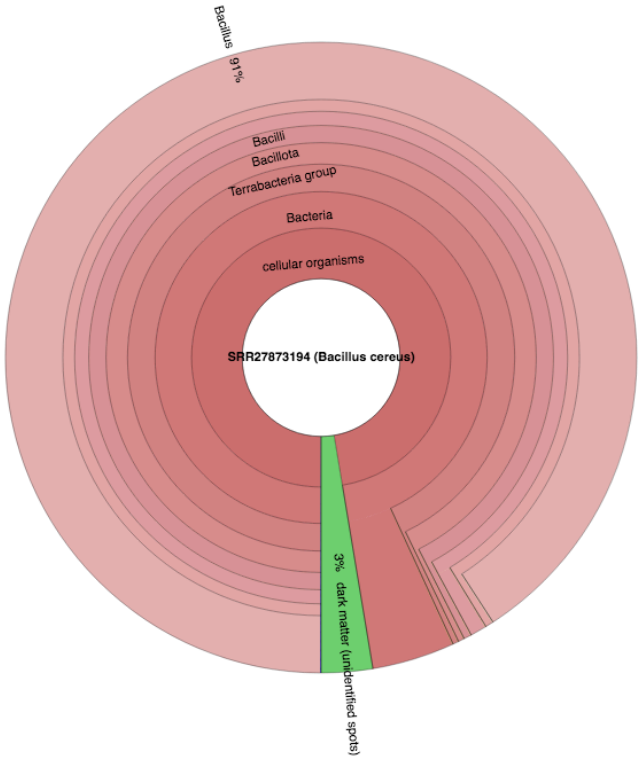
Από εδώ μπορούμε να αποκτήσουμε τα fasta/fastq δεδομένα



Sequence Read Archive SRA – SRR

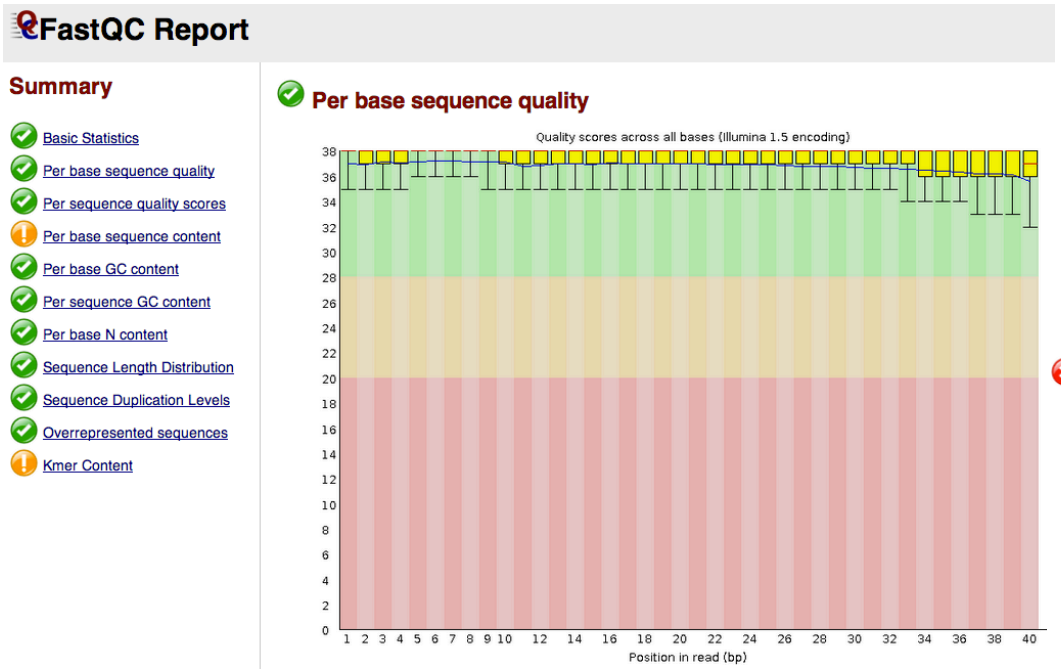
Από το Analysis tab, μέσω του Krona view μπορούμε να δούμε σε ποιές ταξινομικές ομάδες ανήκουν οι ακολουθίες (αν π.χ. Έχουμε επιμόλυνση από άλλο είδος).



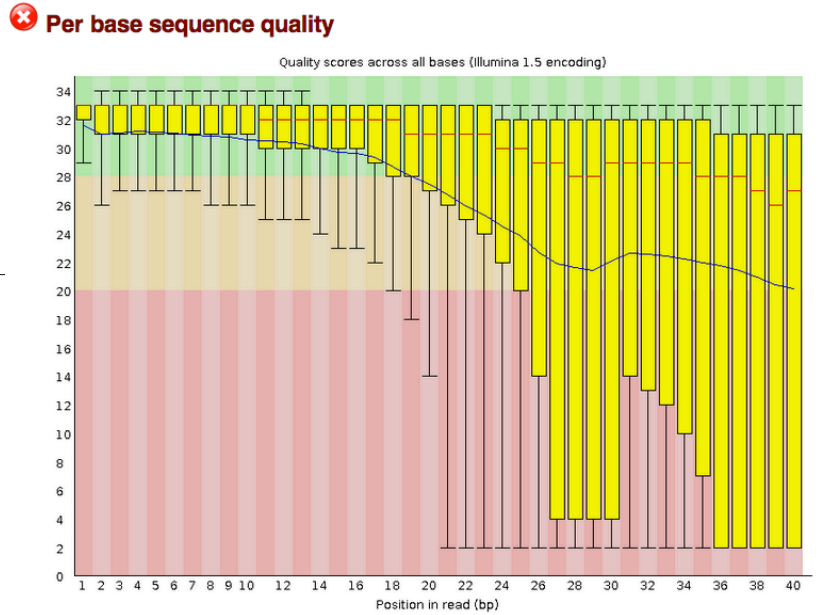


Sequence reads – Έλεγχος ποιότητας δεδομένων (quality control)

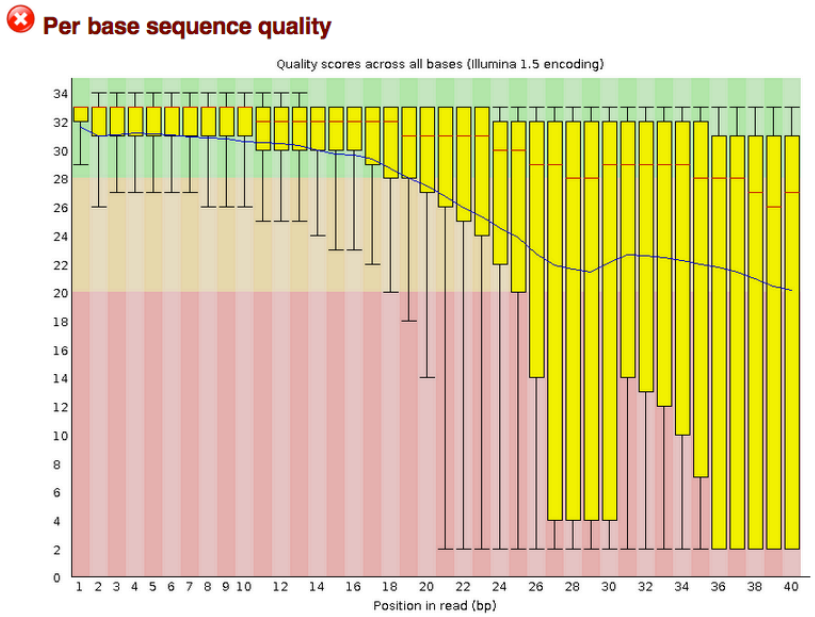
Πολύ υψηλής ποιότητας δεδομένα.



Χαμηλής ποιότητας δεδομένα.



Sequence reads – Φιλτράρισμα/trimming



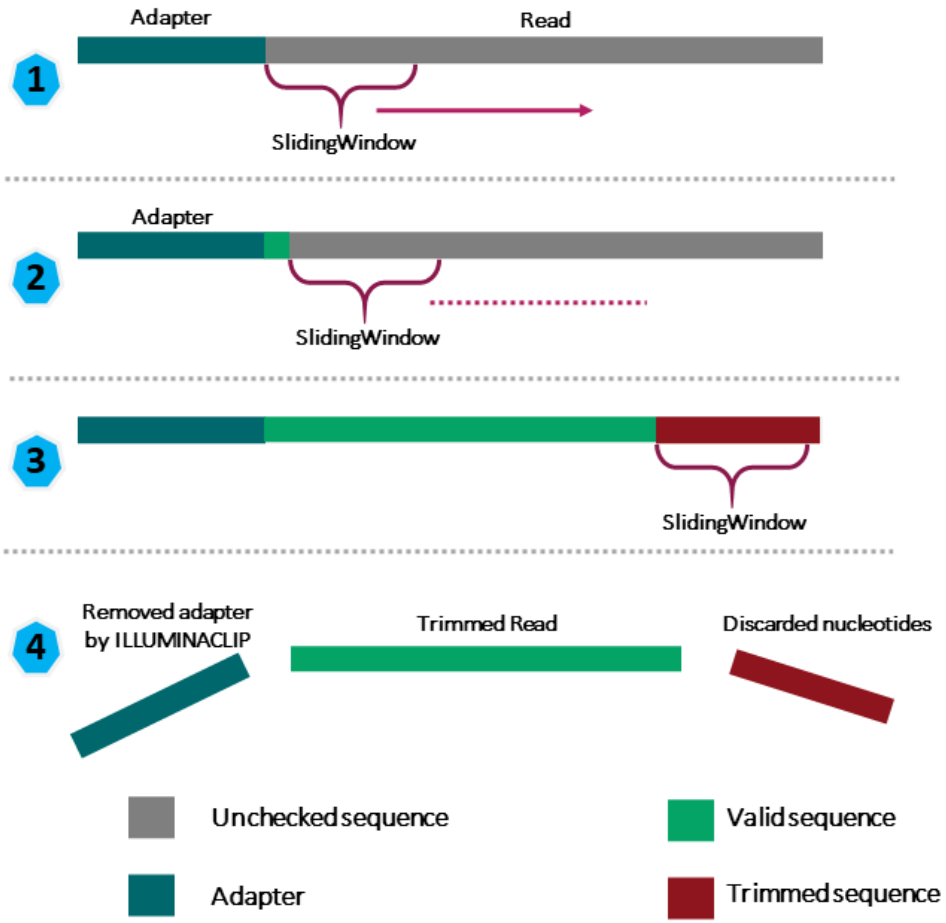
Είτε θα αποφασίσουμε να κόψουμε όλα τα sequence reads σε μια συγκεκριμένη θέση, μετά την οποία η ποιότητα αλληλούχισης πέφτει σημαντικά στα περισσότερα

Είτε θα κόψουμε τα προβληματικά κομμάτια για το κάθε sequence read χωριστά. Μετά θα απορριφθούν όλες τα κομμένα sequence reads που έχουν πολύ μικρό μήκος.

Trimmomatic

Είτε θα κόψουμε τα προβληματικά κομμάτια για το κάθε sequence read χωριστά με συρρόμενα παράθυρα. Μετά θα απορριφθούν όλες τα κομμένα sequence reads που έχουν πολύ μικρό μήκος.

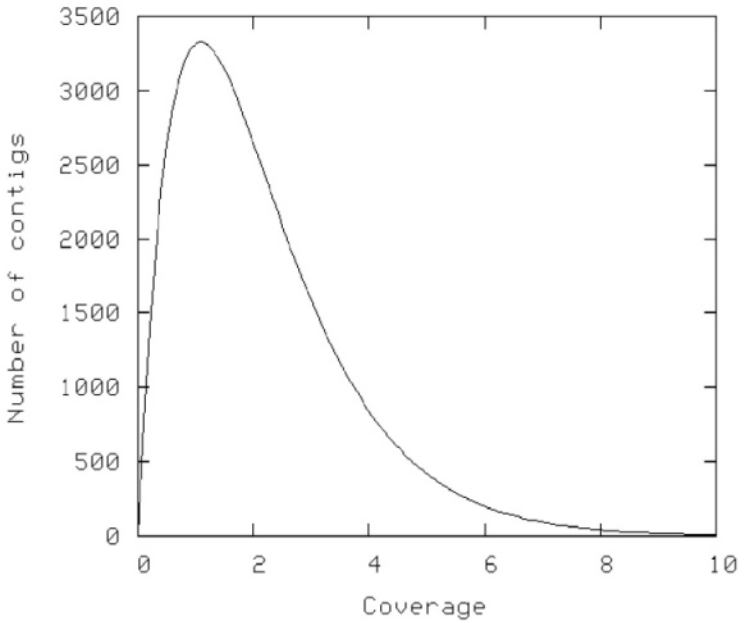
Επίσης, αφαιρούμε τους adapters.



Lander - Waterman

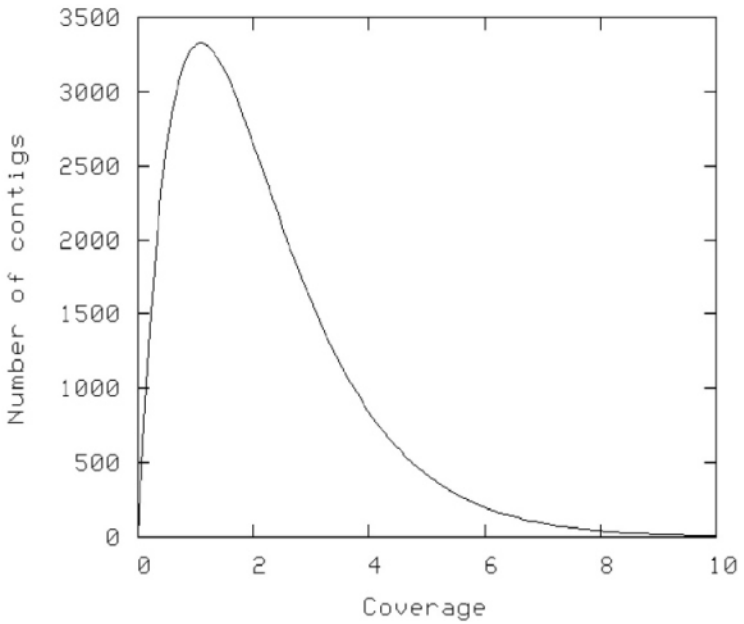
- Πόσο sequencing coverage απαιτείται για να μπορεί να συναρμολογηθεί ένα γονιδίωμα?
- Τουλάχιστον 8-10Χ

- Το παράδειγμα δείχνει πόσα contigs θα δημιουργηθούν θεωρητικά, ανάλογα με την κάλυψη (coverage) του χρωμοσώματος.
- Όσο μεγαλύτερη η κάλυψη, σε τόσο λιγότερα κομμάτια θα είναι σπασμένο το ανακατασκευασμένο χρωμόσωμα
- Στην πράξη, ο αριθμός των contigs είναι μεγαλύτερος από το αναμενόμενο.



Lander - Waterman

- Στην πράξη, ο αριθμός των contigs είναι μεγαλύτερος από το αναμενόμενο, γιατί:
- Πάντα υπάρχει μια πιθανότητα για μια περιοχή να μην αλληλουχιθεί
- Κάποια κομμάτια σπασμένου DNA είναι τοξικά σε φορείς κλωνοποίησης (π.χ. στην E.coli).
- Επαναλήψεις



Προβλήματα συναρμολόγισης από επαναλήψεις - contigs

The ability of an assembly program to produce a single contig is also limited by regions of the genome that occur in multiple near-identical copies throughout the genome (**repeats**). The reads originating from different copies of a repeat appear identical to the assembler and cause assembly errors. A simple example is shown in Figure 5, where the assembler incorrectly collapses the two copies of repeat A leading to the creation of two contigs instead of one (Figure 6).

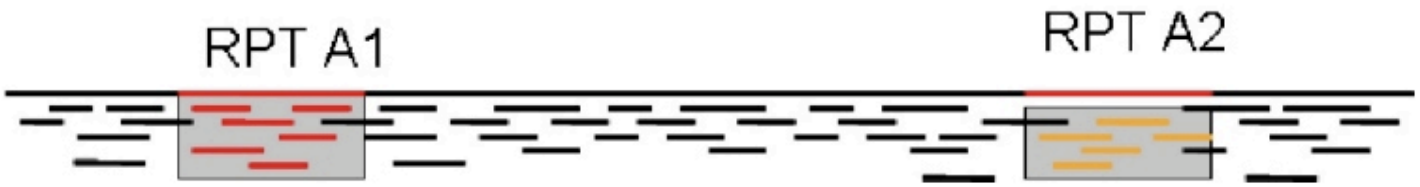


Figure 5. Two copies of a repeat along a genome. The reads colored in red and those colored in yellow appear identical to the assembly program.

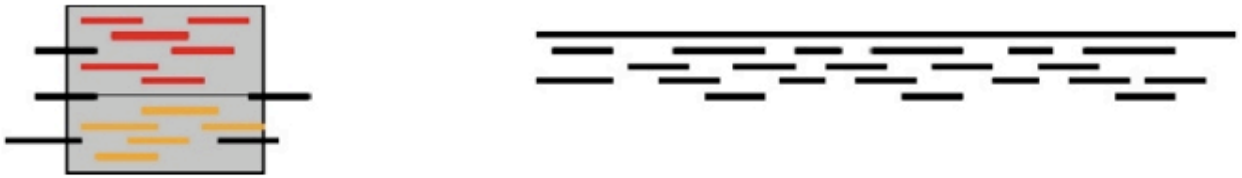
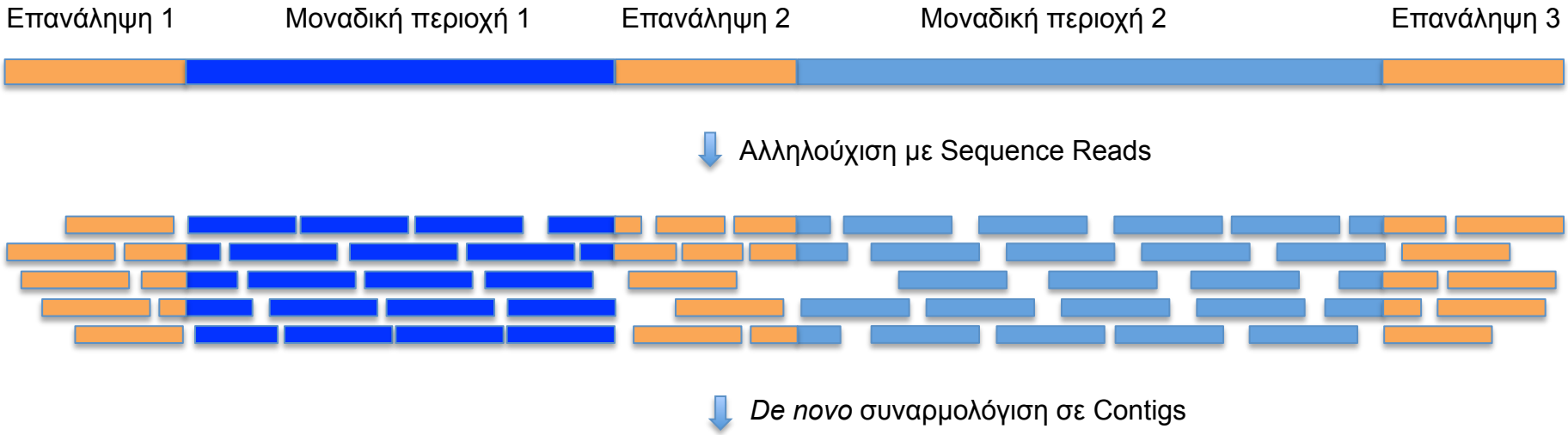


Figure 6. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs

Προβλήματα συναρμολόγισης



Προβλήματα συναρμολόγισης από επαναλήψεις - scaffolds

Scaffolding

The contigs produced by an assembly program can be ordered and oriented along a chromosome using additional information contained in the shotgun data. In most sequencing projects, the sizes of the fragments generated through the shotgun process are carefully controlled, thus providing a link between the sequence reads generated from the ends of a same fragment (called **paired ends** or **mate pairs**). In a typical shotgun project, multiple **libraries** -- collections of fragments of similar sizes -- are usually generated, providing the assembler with additional constraints: within the assembly the paired end reads must be placed at a distance consistent with the size of the library from which they originate and must be oriented towards each other. Within an assembly each read is assigned an orientation corresponding to the DNA strand from which the read was generated. The constraints provided by mate pairs lead to constraints on the relative order and orientation of the contigs (Figure 7). The process through which the read pairing information is used to order and orient the contigs along a chromosome is called **scaffolding**.

Figure 7. A scaffold of 3 contigs (the thick arrows) held together by mate pairs. Thin lines connect the paired ends.

Αφού έχουν γίνει τα scaffolds, όποια κενά υπάρχουν καλύπτονται με στοχευμένη αλληλούχιση - gap closure

Διαφορετικά προγράμματα

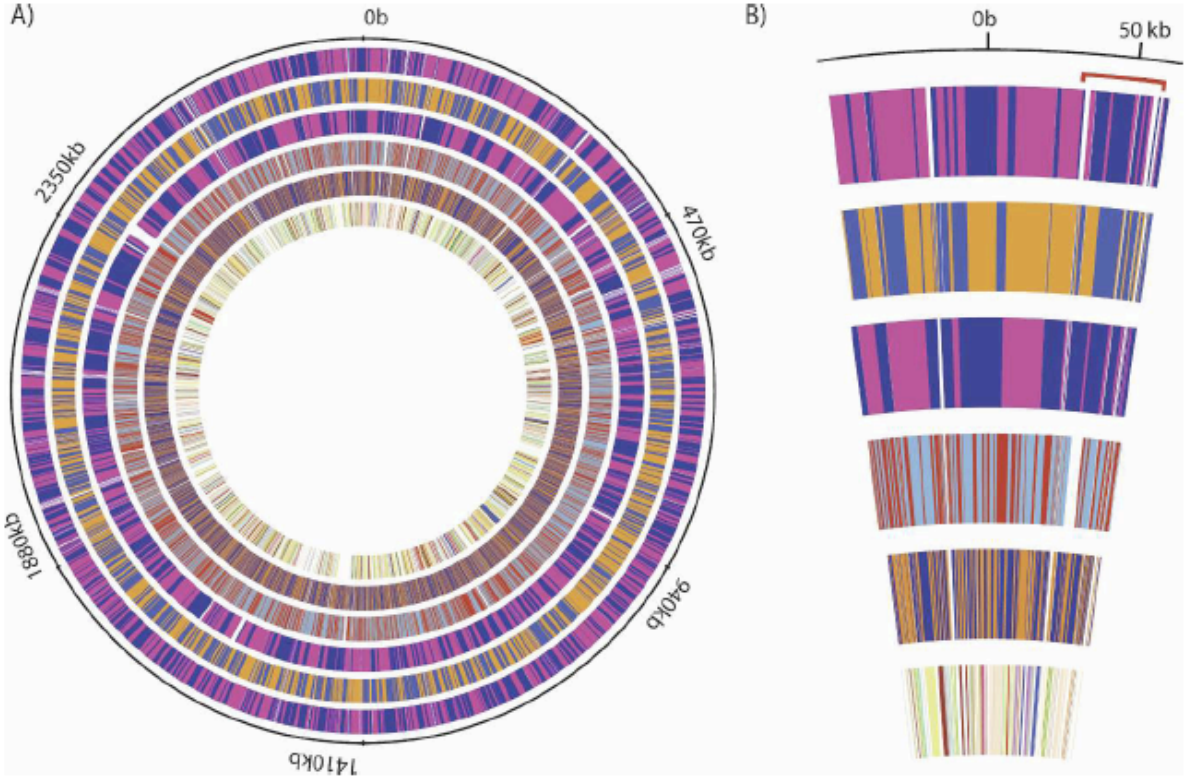


Figure 1. Mapping of the contigs on the reference *Staphylococcus aureus* MW2 genome. (*A*) From external to internal, the circles correspond to the contigs produced by (1) Edena strict, (2) Velvet, (3) Edena nonstrict, (4) SSAKE, and (5) SHARCGS. The contigs are colored by alternating two different colors, which allows distinguishing contig boundaries. The last inner circle shows the coding sequences. The gaps in the Edena nonstrict assembly correspond to large misassembled contigs that did not properly map the reference genome. (*B*) The magnification of the region around the origin of replication provides a better view to compare the contigs length and layout between the different assembly methods. It can be seen that the contigs assembled by Edena and Velvet are long enough to reveal entire genes. More importantly, significant overlaps exist between the contigs assembled by the two programs, which also means that even larger contigs could be assembled by merging both approaches. The position of the SSC*mec* cassette of type IV.1 (Chongtrakool et al. 2006) is indicated by the red line.

Κενά μετά την συναρμολόγιση

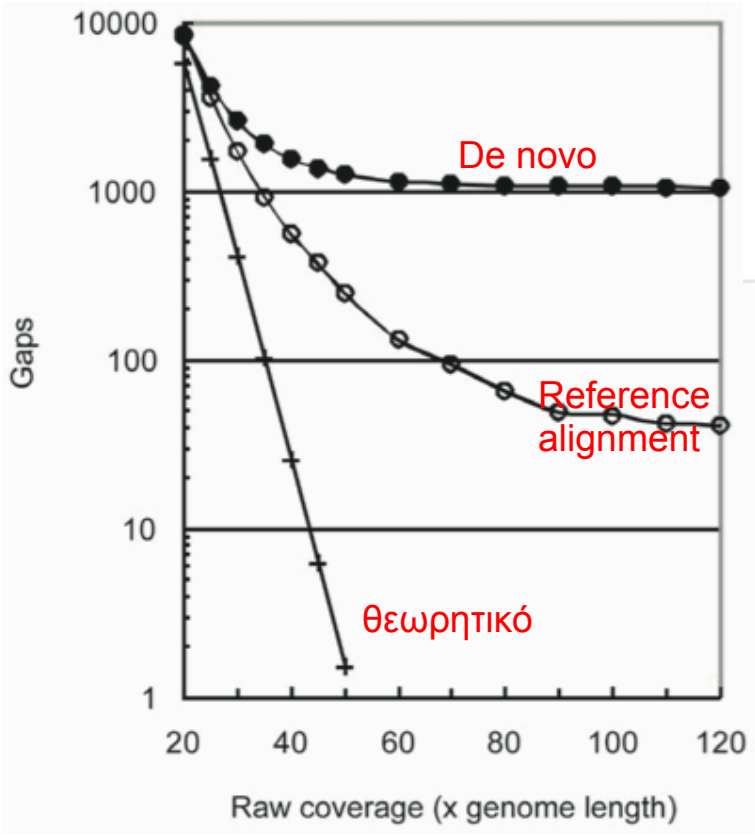


Figure 1. Assessing the cause of gaps in an assembly of 36nt reads. The predicted number of sequence gaps based on the Lander-Waterman model (+) is presented along with the actual number of sequence gaps in sets of 36nt Illumina reads (○). This was determined by aligning the reads in each set to the reference sequence. The total number of gaps present in Velvet assemblies of the various read sets is also included (●). The numerous additional gaps observed in the assemblies are due to unresolvable repeats (○ vs. ●). Additional details can be found in the Supplementary Methods (File S1).
doi:10.1371/journal.pone.0011518.g001

OPEN ACCESS Freely available online PLoS one

Read Length and Repeat Resolution: Exploring Prokaryote Genomes Using Next-Generation Sequencing Technologies

Matt J. Cahill[1], Claudio U. Köser[1], Nicholas E. Ross[2], John A. C. Archer[3]*

1 Department of Genetics, University of Cambridge, Cambridge, United Kingdom, 2 Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, United Kingdom, 3 Division of Chemical and Biological Engineering, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

- Οι επαναλήψεις μπορεί να εμποδίσουν την πλήρη συναρμολόγιση του γονιδιώματος

Κενά μετά την συναρμολόγιση

- Το επιλεγμένο μήκος του sequence read καθορίζει αν θα μπορέσει να συναρμολογηθεί μια επανάληψη

OPEN ACCESS Freely available online PLoS one

Read Length and Repeat Resolution: Exploring Prokaryote Genomes Using Next-Generation Sequencing Technologies

Matt J. Cahill[1], Claudio U. Köser[1], Nicholas E. Ross[2], John A. C. Archer[3]*

1 Department of Genetics, University of Cambridge, Cambridge, United Kingdom, 2 Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, United Kingdom, 3 Division of Chemical and Biological Engineering, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

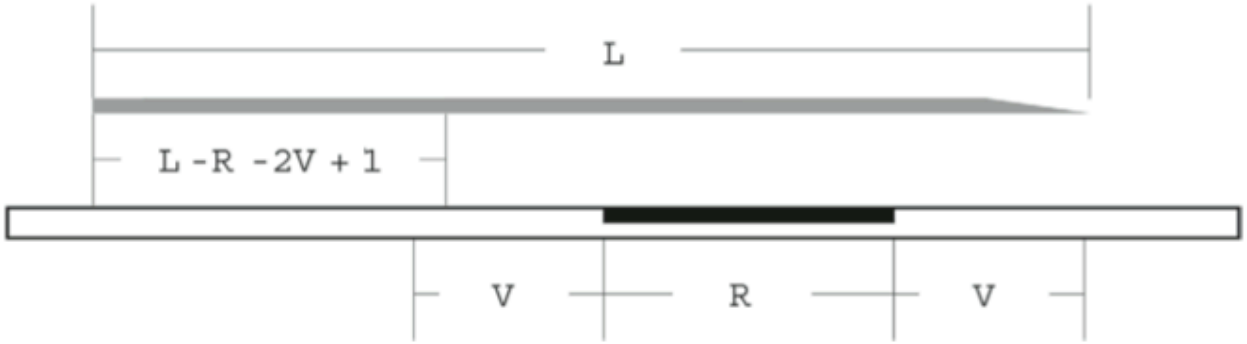


Figure 2. A model of repeat assembly. To unambiguously assemble a repeat (black rectangle), a read must encompass the entirety of the repeat and extend, in both directions, into unique sequence. If the repeat has a length of R nt, and the adjacent unique sequence must be at least V nt, then resolution of the repeat requires that a read starts in a $L-(R+2V-1)$ window next to the repeated sequence. The likelihood of this failing to occur in an assembly of a given number of reads of a particular length, can be estimated using an approach analogous to that used to compute sequence gaps [13,14].
doi:10.1371/journal.pone.0011518.g002

Κενά μετά την συναρμολόγιση

OPEN ACCESS Freely available online

PLoS one

Read Length and Repeat Resolution: Exploring Prokaryote Genomes Using Next-Generation Sequencing Technologies

Matt J. Cahill[1], Claudio U. Köser[1], Nicholas E. Ross[2], John A. C. Archer[3]*

1 Department of Genetics, University of Cambridge, Cambridge, United Kingdom, **2** Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, United Kingdom, **3** Division of Chemical and Biological Engineering, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Μεγαλύτερο μήκος sequence read = λιγότερα κενά

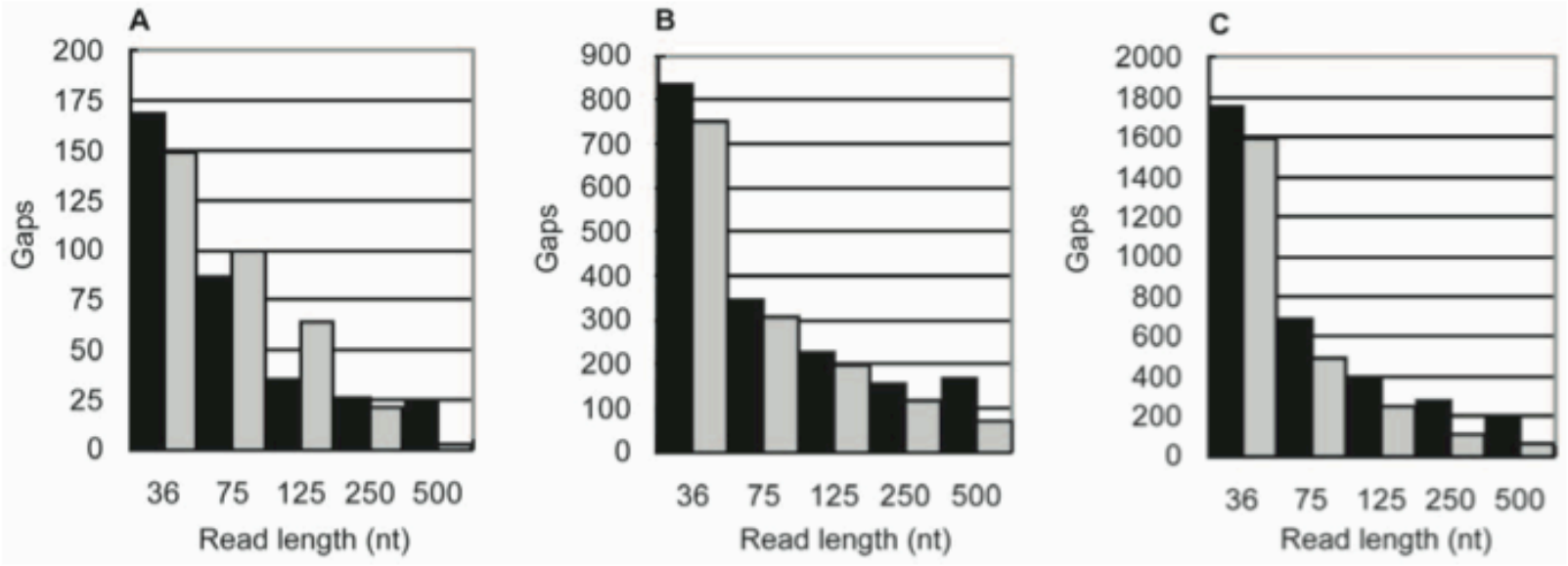


Figure 3. Assessing the accuracy of the algorithm. The number of repeat-induced gaps predicted by the algorithm (grey bars) compared to the number of gaps observed (black bars) in actual assemblies of 36, 75, 125, 250, and 500nt simulated reads from **A**) *M. genitalium*, **B**) *E. coli* and **C**) *S. coelicolor*. The observed gaps are those between unique, non-redundant contigs larger than the read length. The coverage depth of each read set was the threshold at which random gaps are no longer predicted by the Lander-Waterman model. This occurs at effective coverage depths of 9–17×.
doi:10.1371/journal.pone.0011518.g003

Κενά μετά την συναρμολόγιση

- Οι επαναλήψεις μπορεί να εμποδίσουν την πλήρη συναρμολόγιση του γονιδιώματος

OPEN ACCESS Freely available online PLoS one

Read Length and Repeat Resolution: Exploring Prokaryote Genomes Using Next-Generation Sequencing Technologies

Matt J. Cahill[1], Claudio U. Köser[1], Nicholas E. Ross[2], John A. C. Archer[3]*

1 Department of Genetics, University of Cambridge, Cambridge, United Kingdom, 2 Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, United Kingdom, 3 Division of Chemical and Biological Engineering, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

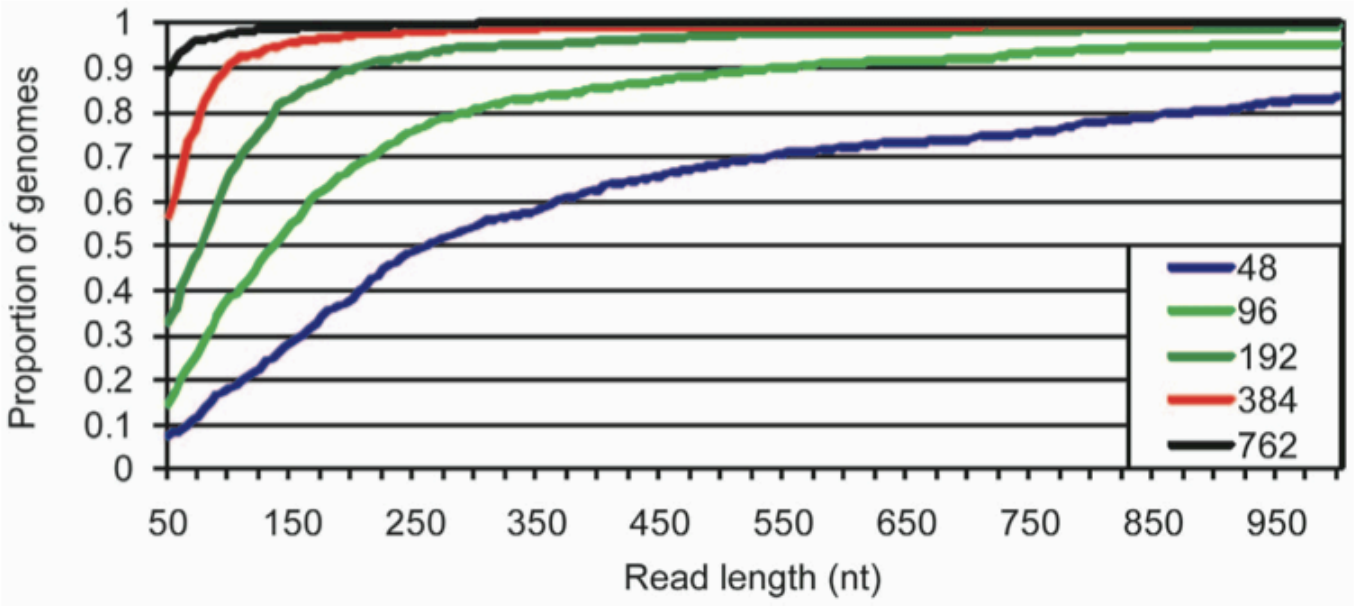


Figure 4. Assessing the performance of a range of read lengths. The fraction of the 818 genomes that meet gap benchmarks as a function of read length was calculated. The benchmarks were 762, 384, 192, 96, and 48 repeat-induced gaps. For example, assuming reads of 150nt, ~50% of the genomes can be assembled with fewer than 96 gaps.
doi:10.1371/journal.pone.0011518.g004

Κενά μετά την συναρμολόγιση

Κάλυψη αλληλούχισης 100Χ για 6 οργανισμούς

OPEN ACCESS Freely available online

PLoS one

Read Length and Repeat Resolution: Exploring Prokaryote Genomes Using Next-Generation Sequencing Technologies

Matt J. Cahill[1], Claudio U. Köser[1], Nicholas E. Ross[2], John A. C. Archer[3]*

1 Department of Genetics, University of Cambridge, Cambridge, United Kingdom, 2 Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, United Kingdom, 3 Division of Chemical and Biological Engineering, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Limits of Repeat Resolution

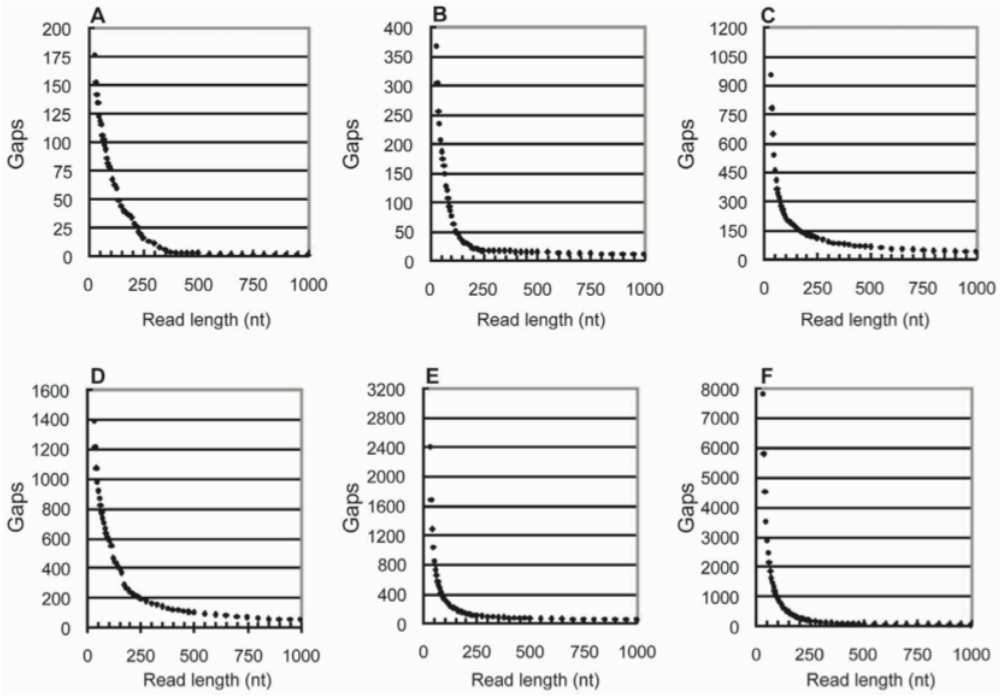


Figure 5. Read length and repeat resolution in 6 genomes. The algorithm was used to predict the occurrence of repeat-induced gaps in assemblies of six bacterial genomes from a range of read lengths. A raw coverage of 100× was used for all genome/read length pairings. Assembly results were predicted for read lengths at increments between 30–1,000nt. Between 30 and 100nt the increment was 5nt; 100–250nt, 10nt; 250–500nt, 25nt; and 500–1,000nt, 50nt. **A**) *M. genitalium* (580 kb), **B**) *H. Influenza* (1.8 Mb), **C**) *E. coli* (4.6 Mb), **D**) *N. meningitidis* (2.3 Mb), **E**) *S. coelicolor* (8.7 Mb) and **F**) *S. cellulosum* (13.0 Mb).

doi:10.1371/journal.pone.0011518.g005

Κενά μετά την συναρμολόγιση

Τα κενά δεν εξαρτώνται μόνο από το βάθος κάλυψης αλληλούχισης και το μήκος των sequence reads, αλλά και από τον ίδιο οργανισμό

36nt reads

125nt reads

500nt reads

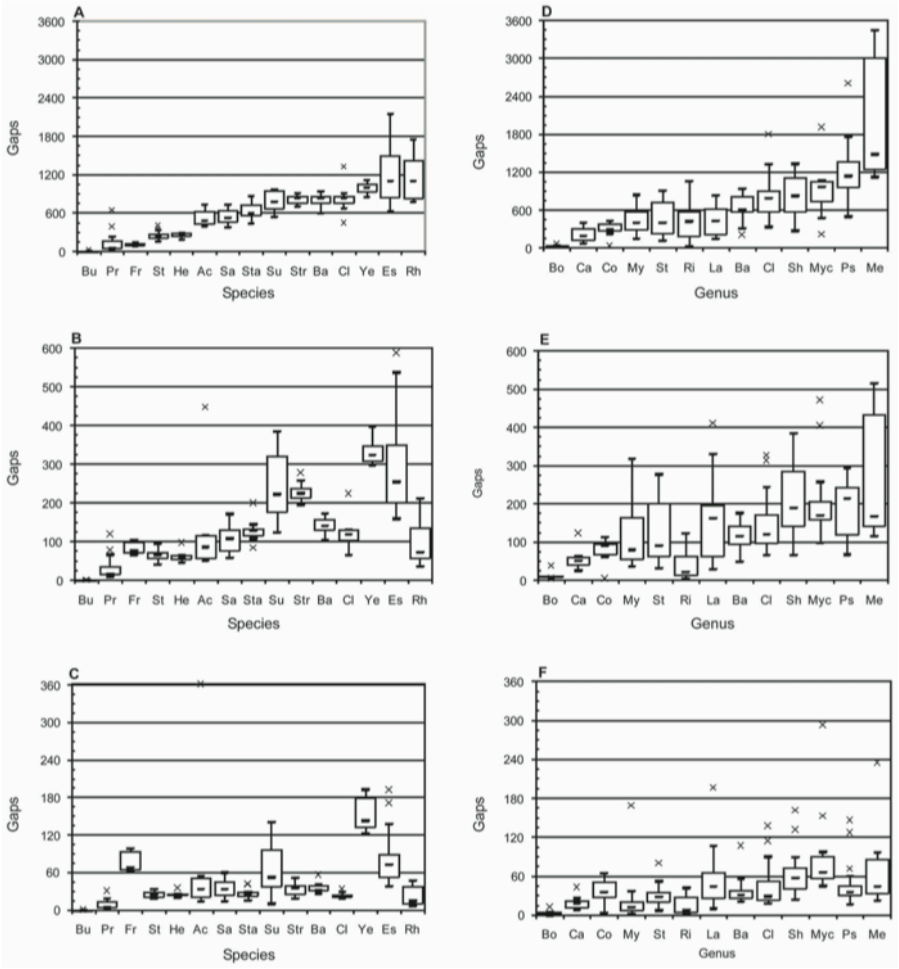


Figure 6. Variation in assembly results within taxa. The median number of repeat-induced gaps for all members of a group is represented by (−). The lower and upper bounds of the hollow rectangle correspond to the first and third quartile, and the range is indicated by the whiskers. Any outliers are plotted as (×). In **A**)–**C**), the species are are *Buchnera aphidicola*, *Prochlorococcus marinus*, *Francisella tularensis*, *Streptococcus pyogenes*, *Helicobacter pylori*, *Acinetobacter baumannii*, *Salmonella enterica*, *Staphylococcus aureus*, *Sulfolobus islandicus*, *Streptococcus pneumoniae*, *Bacillus cereus*, *Clostridium botulinum*, *Yersinia pestis*, *Escherichia coli*, *Rhodopseudomonas palustris*. In **D**)–**F**), the genera are *Borrelia*, *Campylobacter*, *Corynebacterium*, *Mycoplasma*, *Streptococcus*, *Rickettsia*, *Lactobacillus*, *Bacillus*, *Clostridium*, *Shewanella*, *Mycobacterium*, *Pseudomonas*, *Methylobacterium*. For *Methylobacterium*, outliers at 36nt (6,307) and 125nt (1,219) have been omitted. Gap predictions are for reads of **A**)/**D**) 36nt, **B**)/**E**) 125nt, and **C**)/**F**) 500nt.
doi:10.1371/journal.pone.0011518.g006

Τα περισσότερα βακτηριακά γονίδια μπορούν να συναρμολογηθούν

Kingsford *et al. BMC Bioinformatics* 2010, **11**:21
http://www.biomedcentral.com/1471-2105/11/21

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Assembly complexity of prokaryotic genomes using short reads

Carl Kingsford*, Michael C Schatz, Mihai Pop

- Μικρού μήκους reads μπορούν να συναρμολογήσουν τα περισσότερα γονίδια, αλλά σπάνε το γονιδίωμα σε πολλά μικρά κομμάτια (contigs)

Table 2 Median N50 and reconstructible genes.

k	N50 (%)	Genes (%)
25	1.14	96.29
35	2.41	98.12
50	3.90	98.94
100	8.12	99.51
250	13.52	99.84
500	18.03	100
1000	46.57	100

Median N50 as a percentage of the chromosome size and median number of genes that are reconstructible for various read lengths *k*.

Τα περισσότερα βακτηριακά γονίδια μπορούν να συναρμολογηθούν

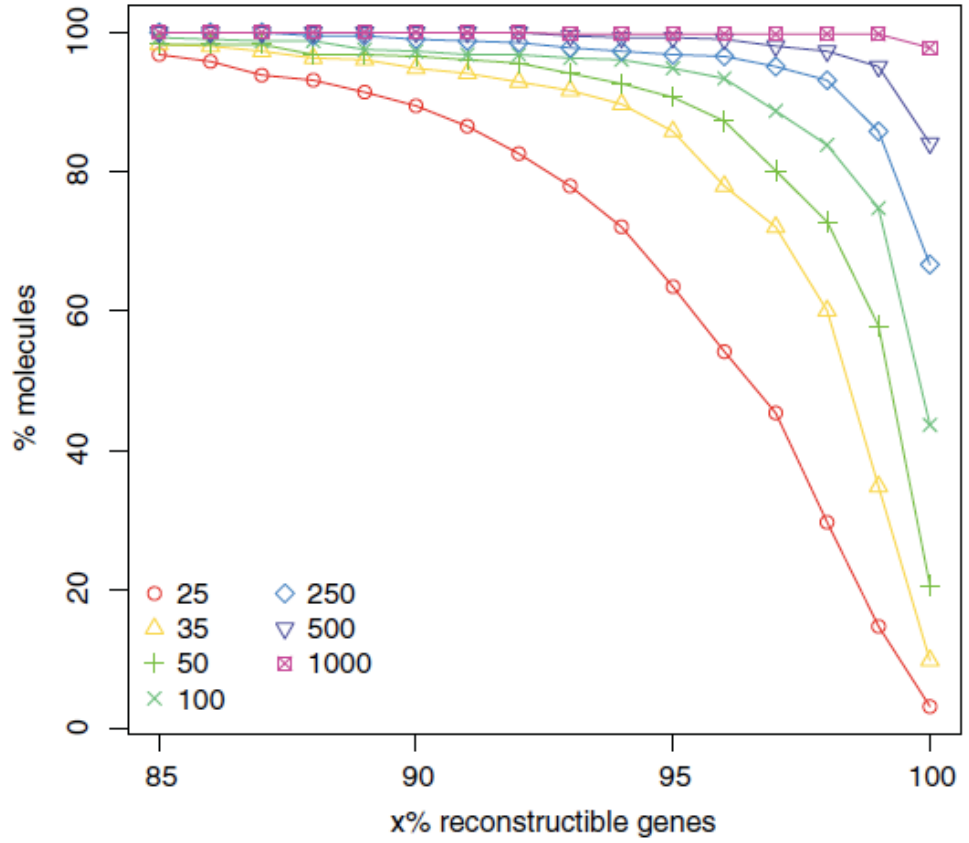


Figure 4 Fraction of reconstructible genes. Cumulative histogram plotting a percentage of genes (x-axis) against the percentage of chromosomes for which at least that many genes can be reconstructed. When $k = 1000$, nearly all chromosomes have all their genes completely reconstructed. When $k = 25$, the number of reconstructible genes falls off quicker, but many genes can still be reconstructed: 90% or more of the genes can be reconstructed in 89% of the chromosomes when $k = 25$.

Μικρού μήκους reads μπορούν να συναρμολογήσουν τα περισσότερα γονίδια, αλλά σπάνε το γονιδίωμα σε πολλά μικρά κομμάτια (contigs)

Τα περισσότερα βακτηριακά γονίδια μπορούν να συναρμολογηθούν

Γονιδιωματικά στοιχεία που προκαλούν προβλήματα στην συναρμολόγιση:

Μεταθετά στοιχεία

transposons

Intergenic repeats

Insertion sequences

prophages

Γονίδια που συνήθως δεν μπορούν να συναρμολογηθούν:

Transposases

Phages

Integrases

Γονίδια που σχετίζονται με την αποφυγή του ανοσοποιητικού συστήματος (έχουν επαναλήψεις)

De novo Sequence assemby

- http://www.cbcb.umd.edu/research/assembly_primer.shtml

- De novo assembly
  - Greedy extention
  - OLC
  - De Bruijn graph
  - Hybrid

Greedy assemblers

Greedy assemblers - The first assembly programs followed a simple but effective strategy in which the assembler greedily joins together the reads that are most similar to each other. An example is shown in Figure 8, where the assembler joins, in order, reads 1 and 2 (overlap = 200 bp), then reads 3 and 4 (overlap = 150 bp), then reads 2 and 3 (overlap = 50 bp) thereby creating a single contig from the four reads provided in the input. One disadvantage of the simple greedy approach is that because local information is considered at each step, the assembler can be easily confused by complex repeats, leading to mis-assemblies.

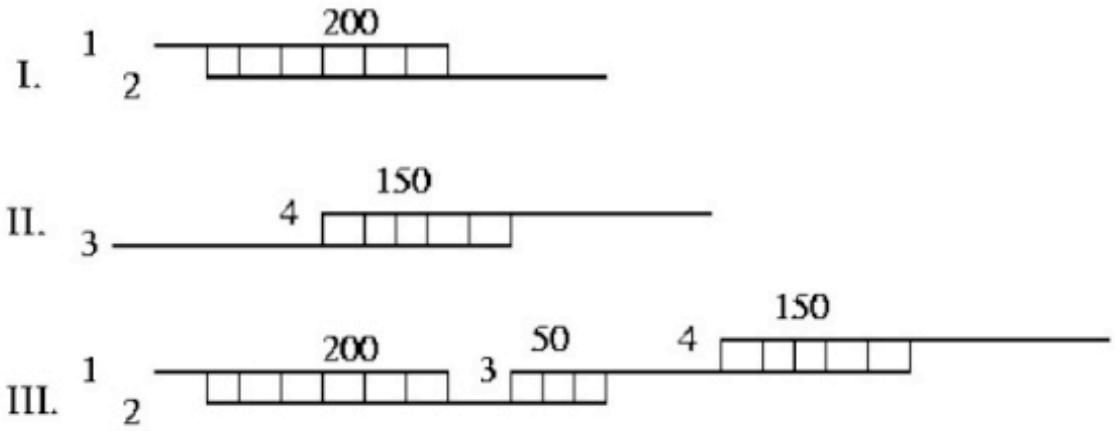


Figure 8. Greedy assembly of four reads.

Overlap - layout - consensus (OLC)

Overlap-layout-consensus - The relationships between the reads provided to an assembler can be represented as a graph, where the nodes represent each of the reads and an edge connects two nodes if the corresponding reads overlap. The assembly problem thus becomes the problem of identifying a path through the graph that contains all the nodes - a **Hamiltonian path** (Figure 9). This formulation allows researchers to use techniques developed in the field of **graph theory** in order to solve the assembly problem. An assembler following this paradigm starts with an **overlap** stage during which all overlaps between the reads are computed and the graph structure is computed. In a **layout** stage, the graph is simplified by removing redundant information. Graph algorithms are then used to determine a layout (relative placement) of the reads along the genome. In a final **consensus** stage, the assembler builds an alignment of all the reads covering the genome and infers, as a consensus of the aligned reads, the original sequence of the genome being assembled.

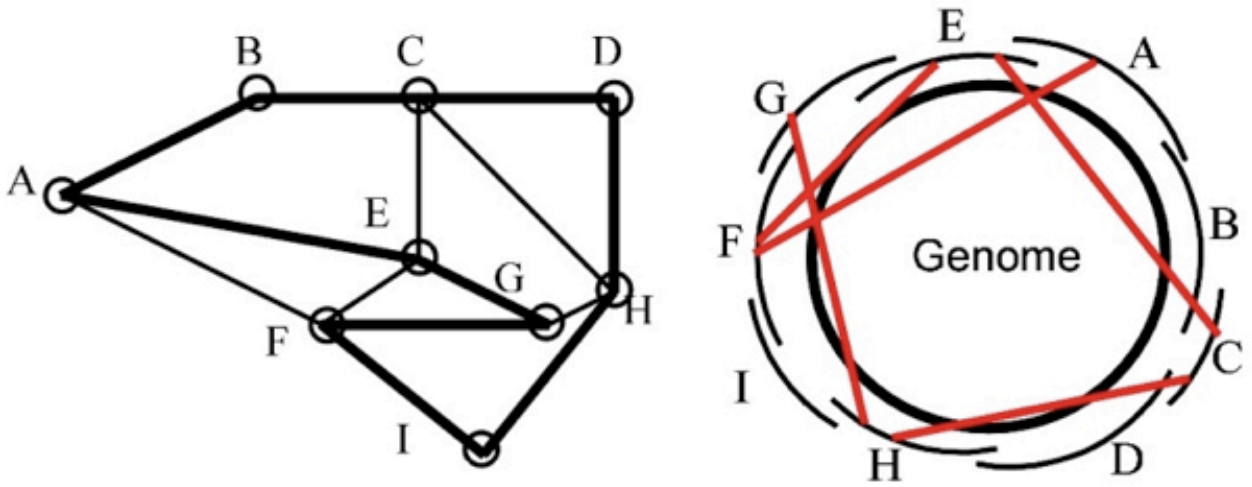
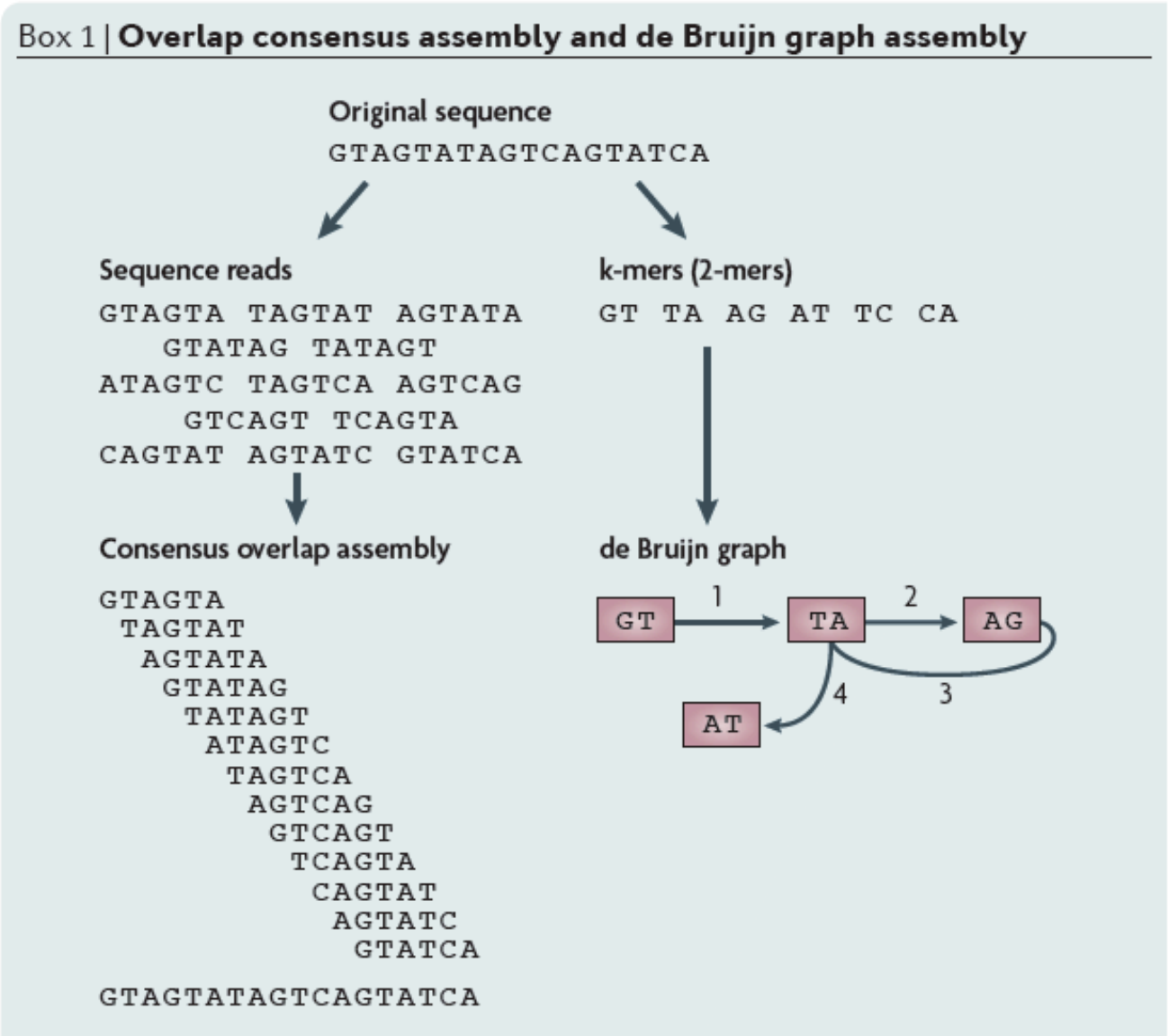


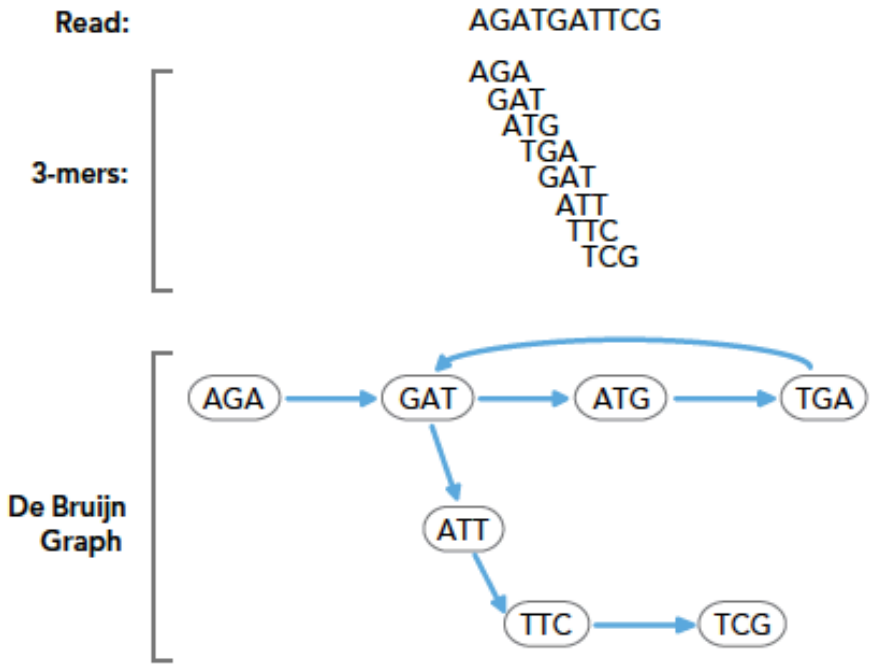
Figure 9. Overlap graph for a bacterial genome. The thick edges in the picture on the left (a Hamiltonian cycle) correspond to the correct layout of the reads along the genome (figure on the right). The remaining edges represent false overlaps induced by repeats (exemplified by the red lines in the figure on the right)

Γραφήματα De Bruijn



De bruijn graph

Figure 3: De Bruijn Graph for Read with K=3



The length of overlaps is k-1=2. Gray arrows indicate where all the k-mers derived from the one read are placed in the graph. Blue arrows indicate the order of the k-mers and their overlaps.

Comparative assembly

Align-layout-consensus - As more and more genomes become available in public databases, it is increasingly the case that a completed genome exists that is closely related to the genome being assembled. The assembly problem thus becomes easier as the relative placement of reads can be inferred from their alignment to the related genome (or **reference**), in a process called **comparative assembly**. Thus, the overlap stage of assembly (often one of the most computationally intensive assembly tasks) is replaced by an alignment step. The layout stage is also greatly simplified due to the additional constraints provided by the alignment to the reference.

BAC-by-BAC sequencing

BAC-by-BAC (hierarchical) sequencing - In order to avoid some of the complexity involved in assembling large genomes, scientists developed a hierarchical approach. First, the genome is broken up into a collection of large fragments (between 40 and 200 kbp) called **Bacterial Artificial Chromosomes** or **BAC**s. The BACs location along the genome is then mapped using specialized laboratory experiments. A **minimal tiling path** of BACs is chosen such that each base in the genome is covered by at least one BAC, and the overlap between BACs is minimized. Each BAC is then sequenced through the standard shotgun method, the resulting assemblies being combined into an assembly for each chromosome using the information provided by the tiling paths (Figure 10).

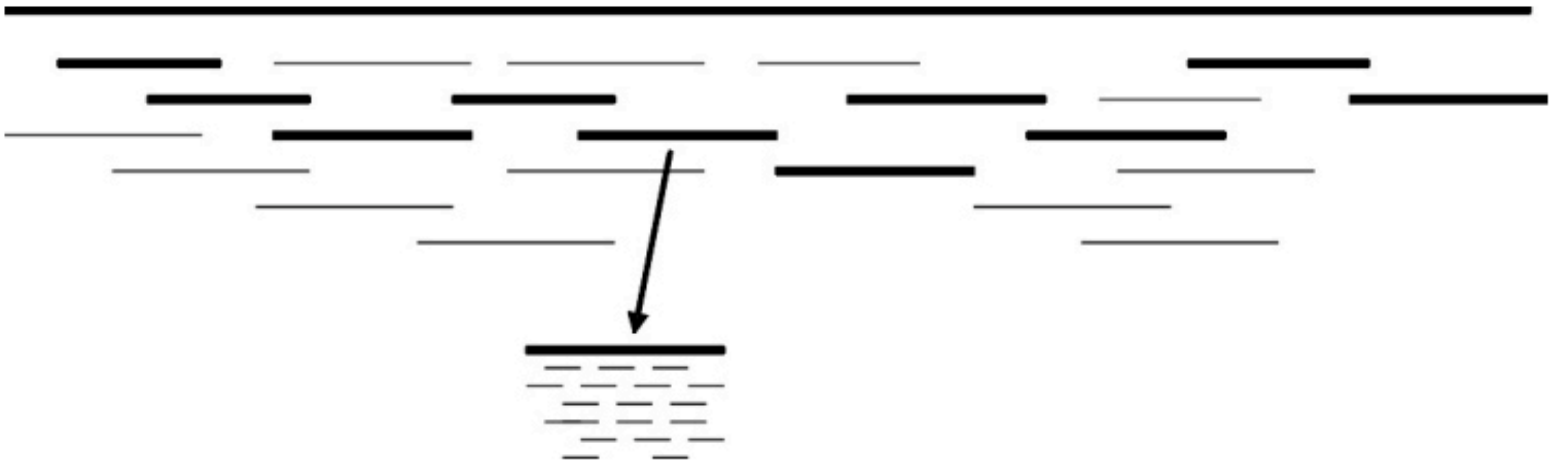
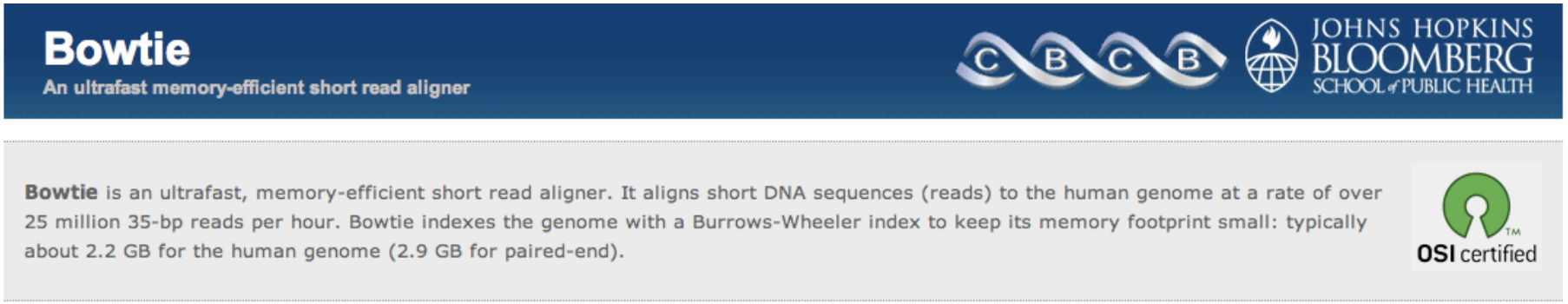


Figure 10. BAC-by-BAC approach. The long lines represent individual BACs. The minimal tiling path is represented by thick lines. Each BAC in the tiling path is then sequenced through the shotgun method.

Short read alignment

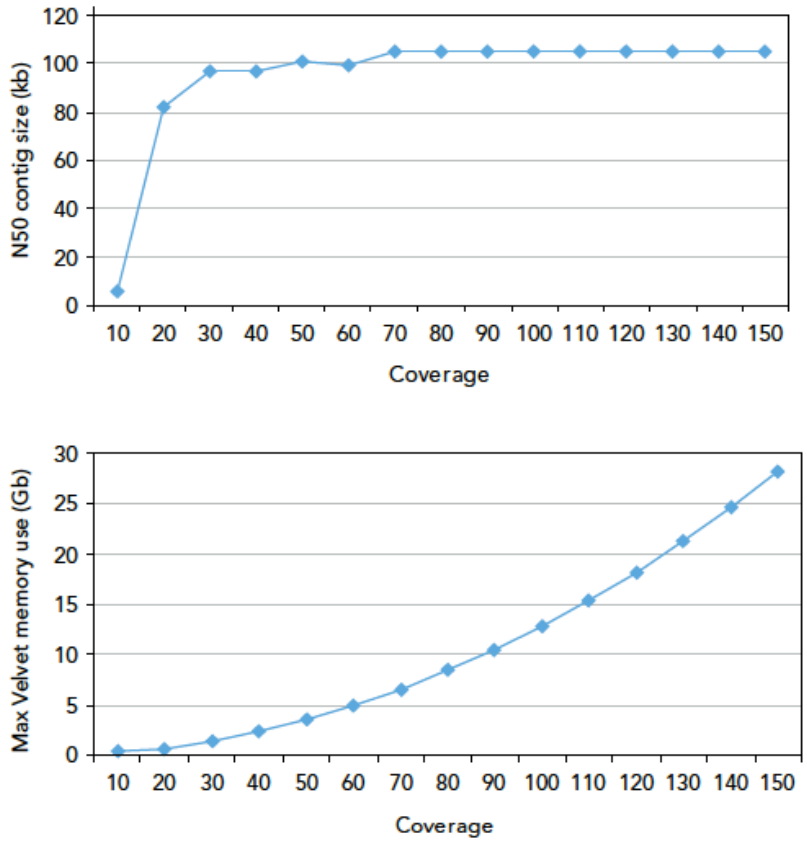


Τιμή N50

- Η τιμή αυτή αντιστοιχεί σε εκείνο το μήκος contigs, ώστε το 50% του γονιδιώματος (μετά από de novo assembly) να εντοπίζεται σε contigs αυτού το μήκους ή μεγαλύτερου.
- Μεγάλη τιμή του N50 σημαίνει ότι το μεγαλύτερο μέρος του γονιδιώματος βρίσκεται σε λίγα και μεγάλα contigs.
- Δηλαδή, τόσο καλύτερη η συναρμολόγιση.
- Μικρή τιμή σημαίνει ότι το γονιδίωμα δεν έχει συναρμολογηθεί καλά.

Κάλυψη του γονιδιώματος και κορεσμός

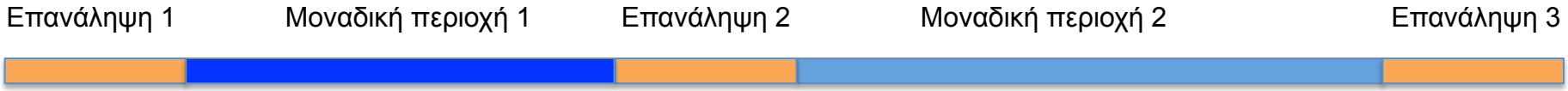
Figure 4: Effect of Coverage



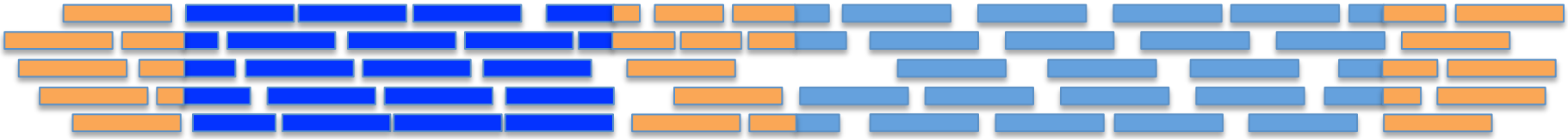
Effect of coverage on N50 contig size and memory requirements in an E. coli de novo assembly.

- Δεν έχει νόημα να αλληλουχίσουμε ένα γονιδίωμα με υπερβολικά μεγάλη κάλυψη (coverage), για μια συγκεκριμένη τεχνολογία και μήκος sequence reads, γιατί από ένα σημείο και μετά έχει επέλθει κορεσμός.

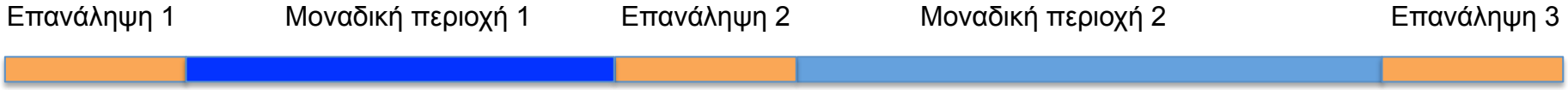
Reference assembly/alignment



Αλληλούχιση με Sequence Reads

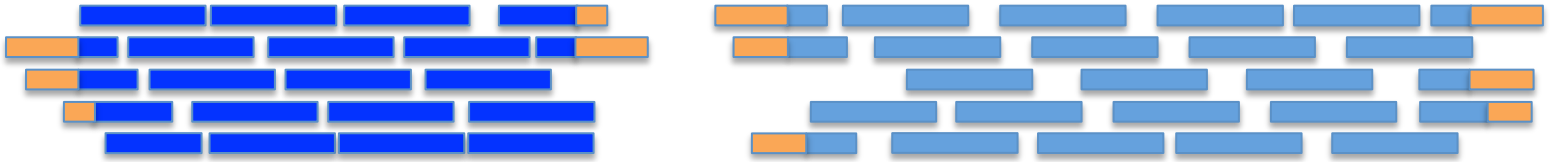


Συναρμολόγιση με βάση γονιδίωμα αναφοράς



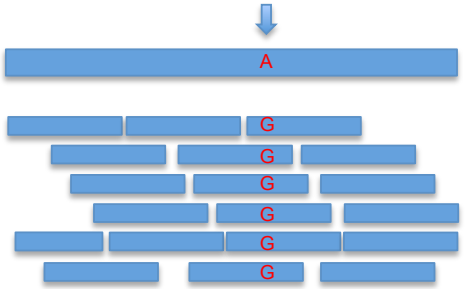
Sequence Reads που μπορούν να στοιχιθούν
σε περισσότερες από μια θέσεις δεν στοιχίζονται

Μόνο στοίχιση των Sequence Reads που έχουν μια μοναδική θέση



Reference assembly/alignment/SNPs

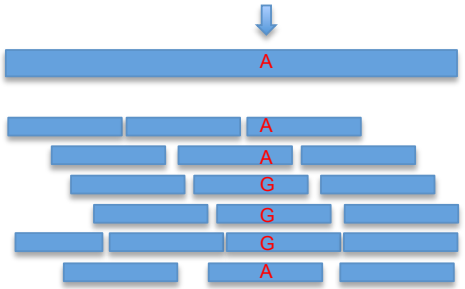
Πολυμορφισμός σε ομοζυγωτία



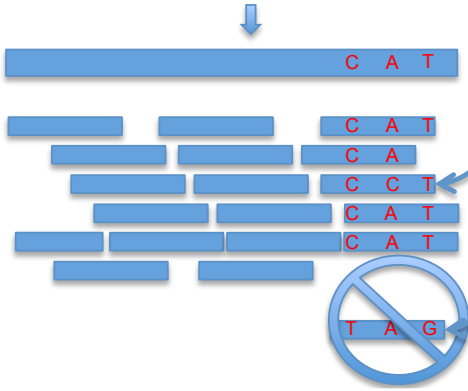
Γονιδίωμα Αναφοράς

Sequence Reads του ατόμου

Πολυμορφισμός σε ετεροζυγωτία



Λάθος σε κάποιο Sequence Read

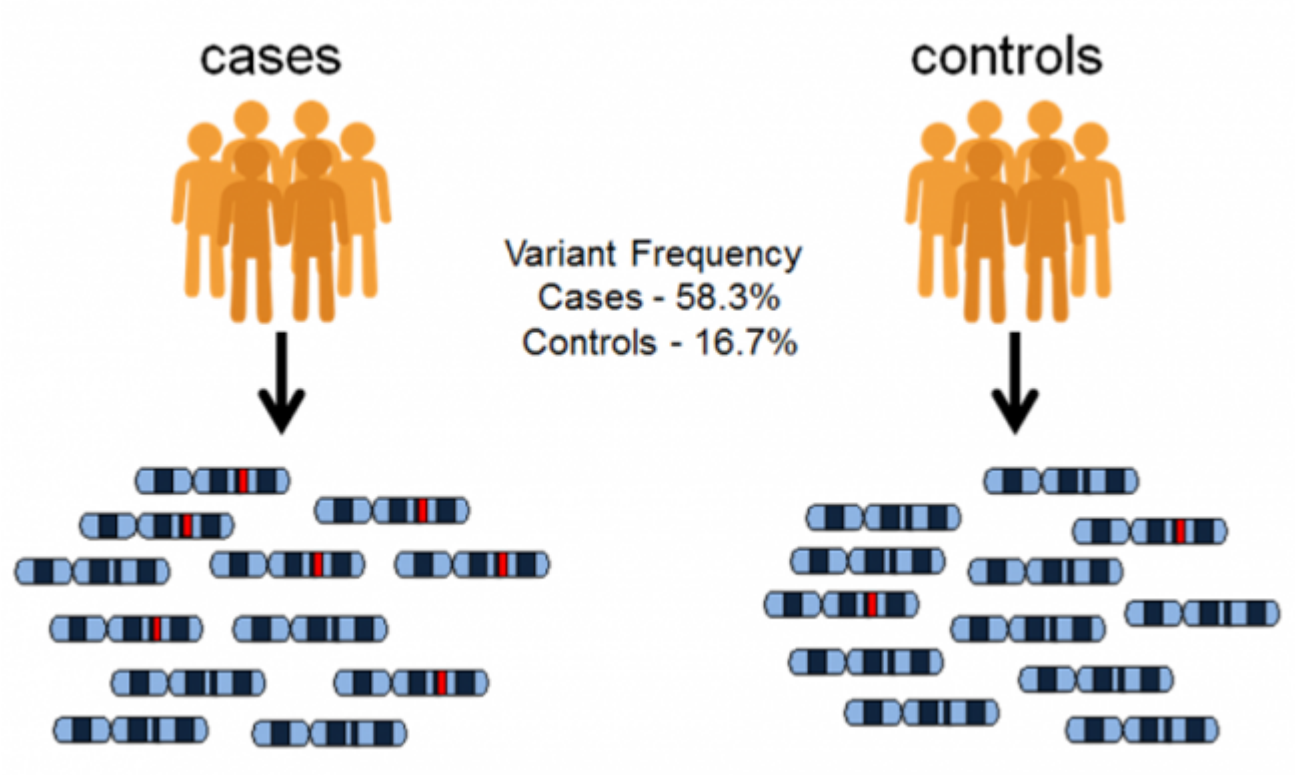


Επιτρεπόμενο όριο: Μία αναντιστοιχία

Το sequence read έχει μόνο ένα λάθος, αλλά θα στοιχιθεί

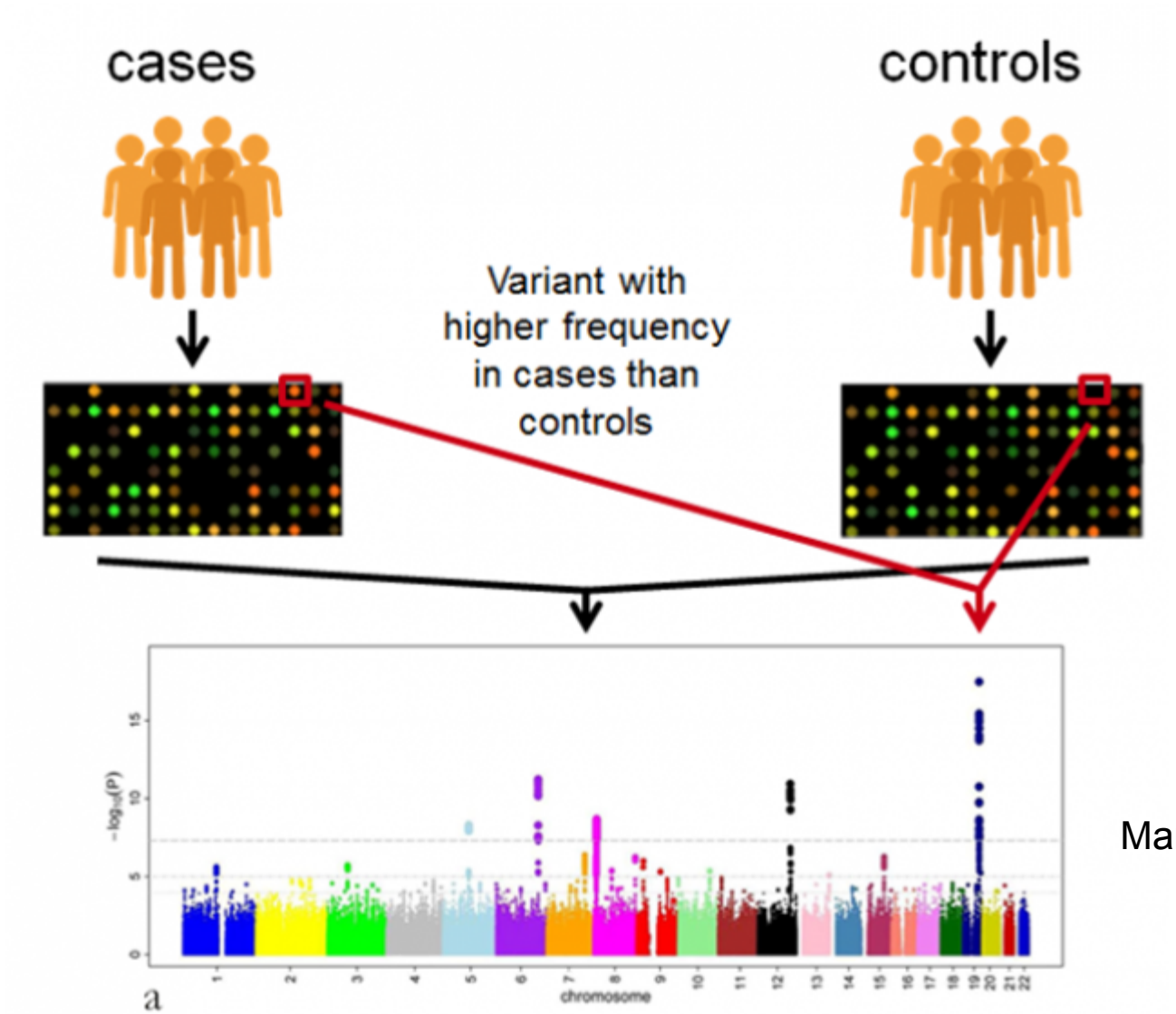
Το sequence read έχει περισσότερα λάθη
από το επιτρεπόμενο όριο και δεν θα στοιχιθεί

GWAS



GWAS
SNParrays − WGS - WES

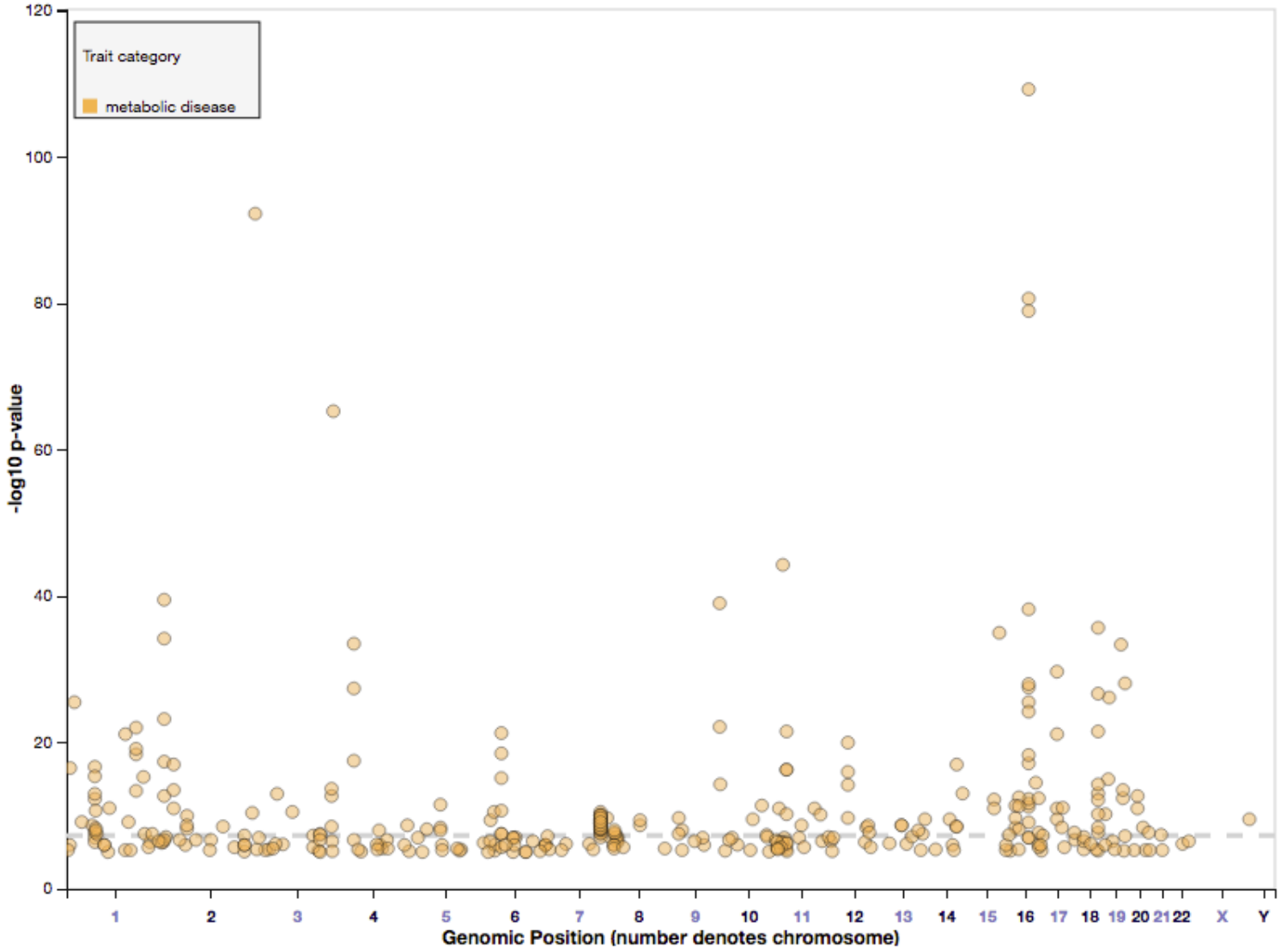
Cases vs Controls − pvalue corrected for multiple testing



Manhattan plots

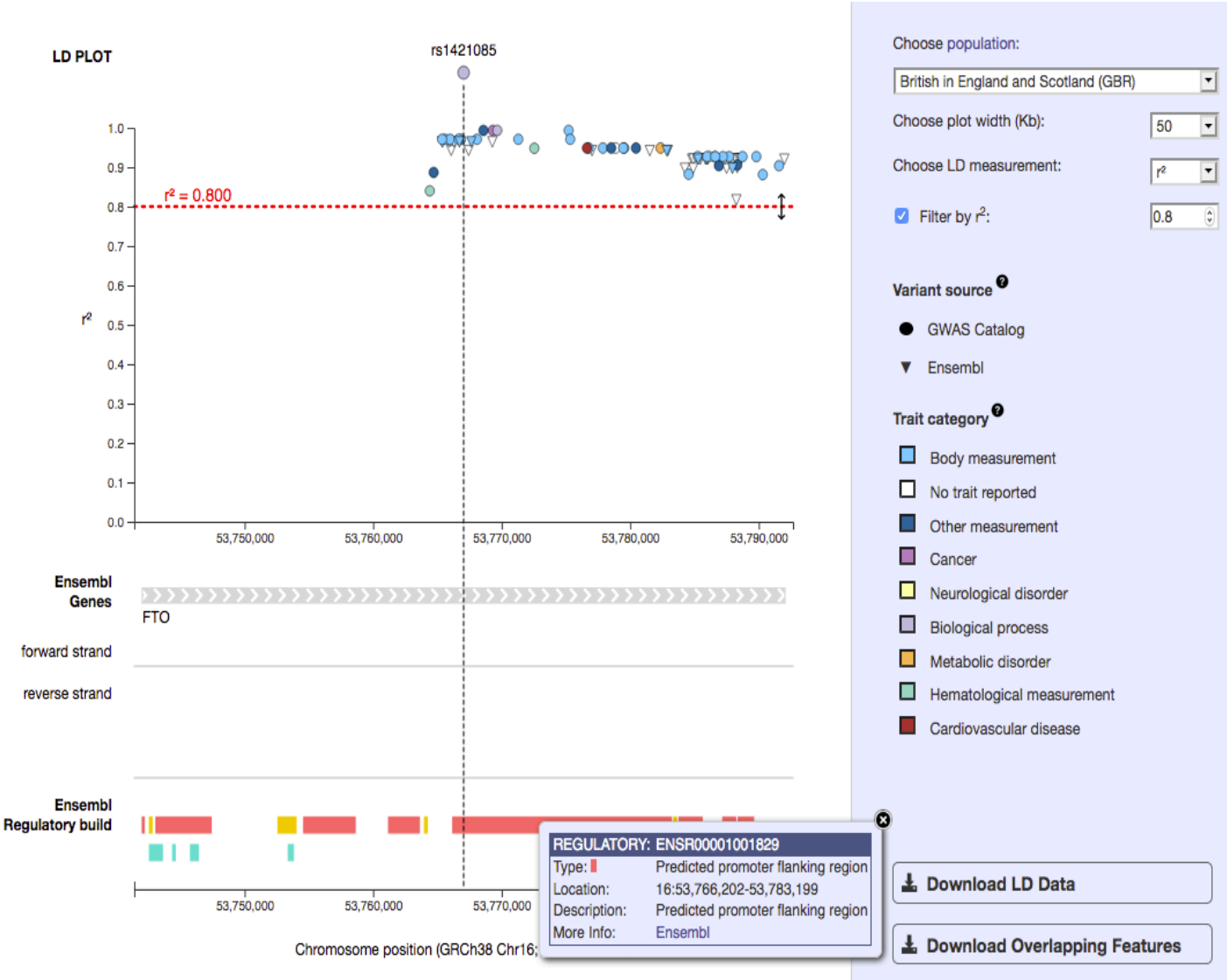
GWAS catalogue - EMBL-EBI and NHGRI.

Παχυσαρκία: SNP rs1421085

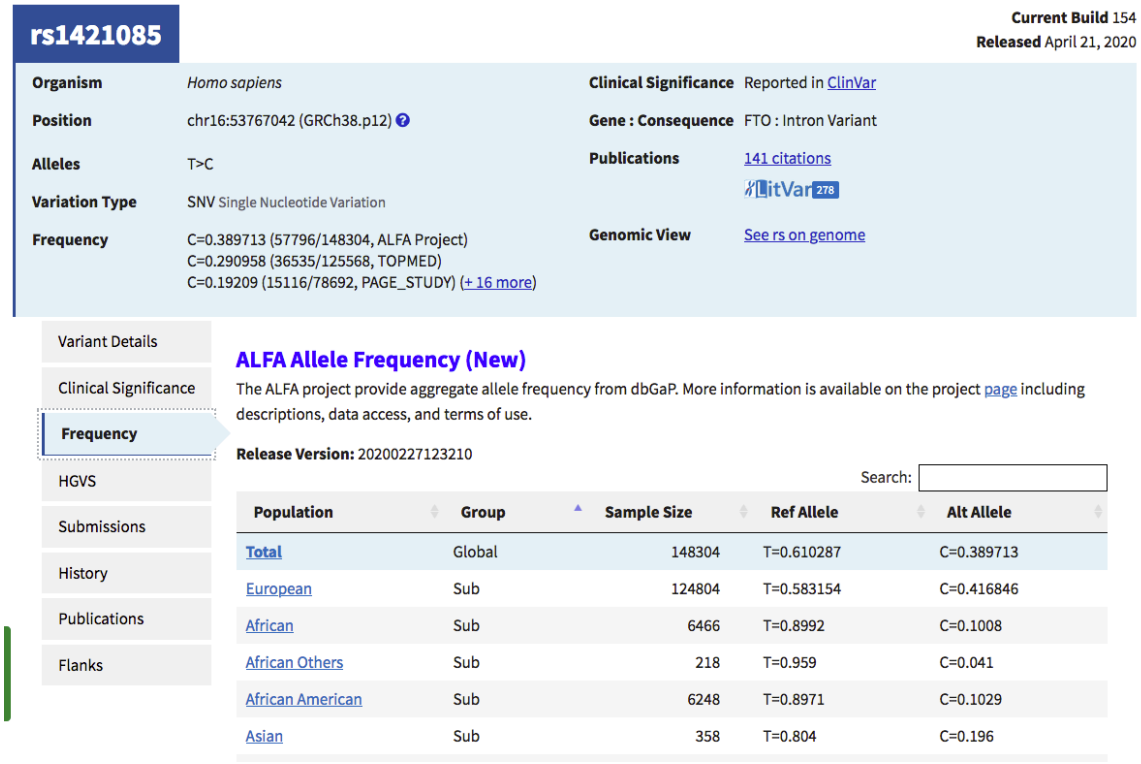


GWAS catalogue

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.

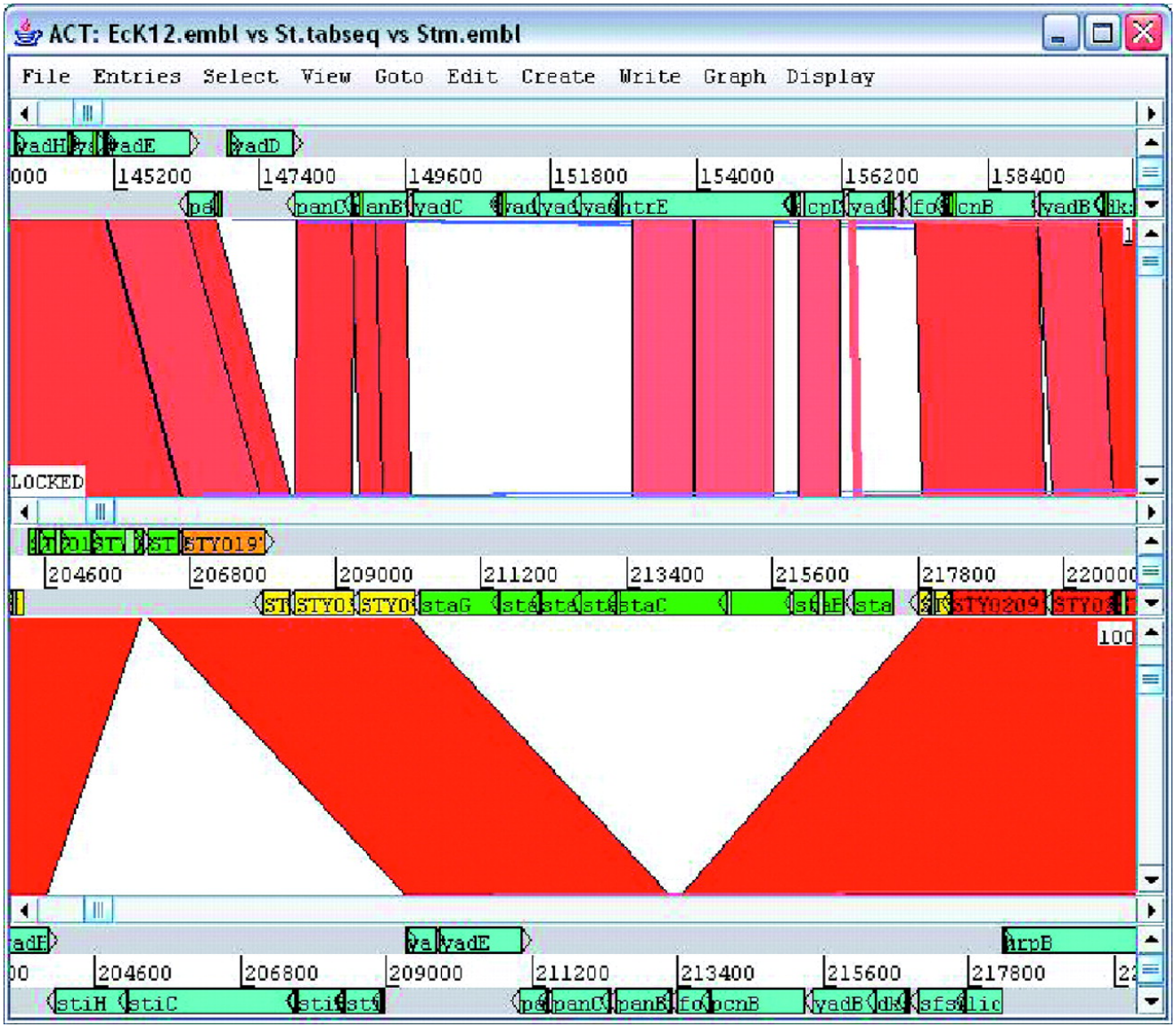
Εφαρμογές

‘Ελεγχος εξελικτικών υποθέσεων -

Προέλευση -

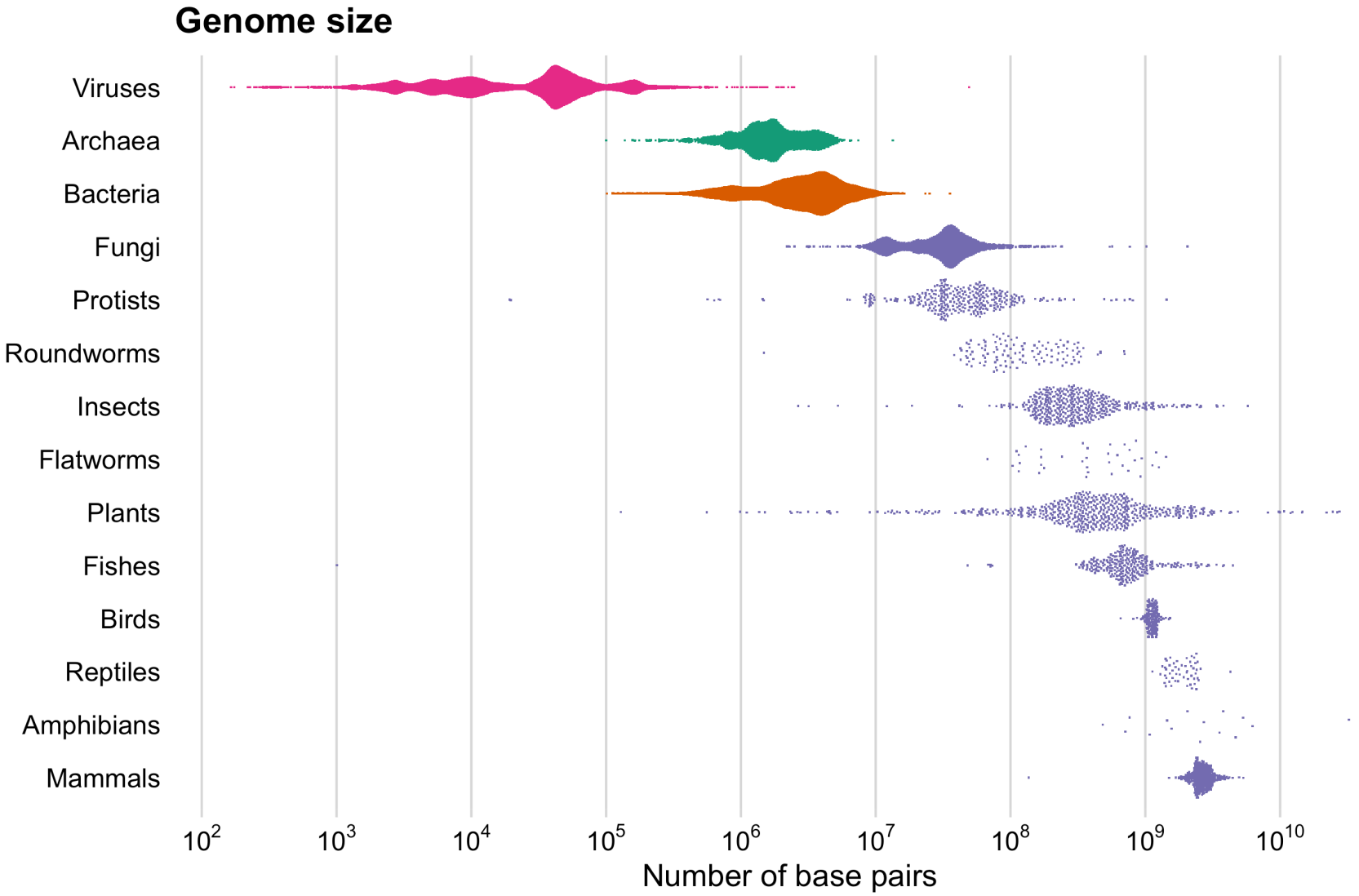
Επιδημιολογία

Σύγκριση γονιδιωμάτων - ACT



BLASTN comparison of part of three sequences: Escherichia coli K12, Salmonella Typhi CT18 and Salmonella Typhimurium LT2 (from top to bottom).

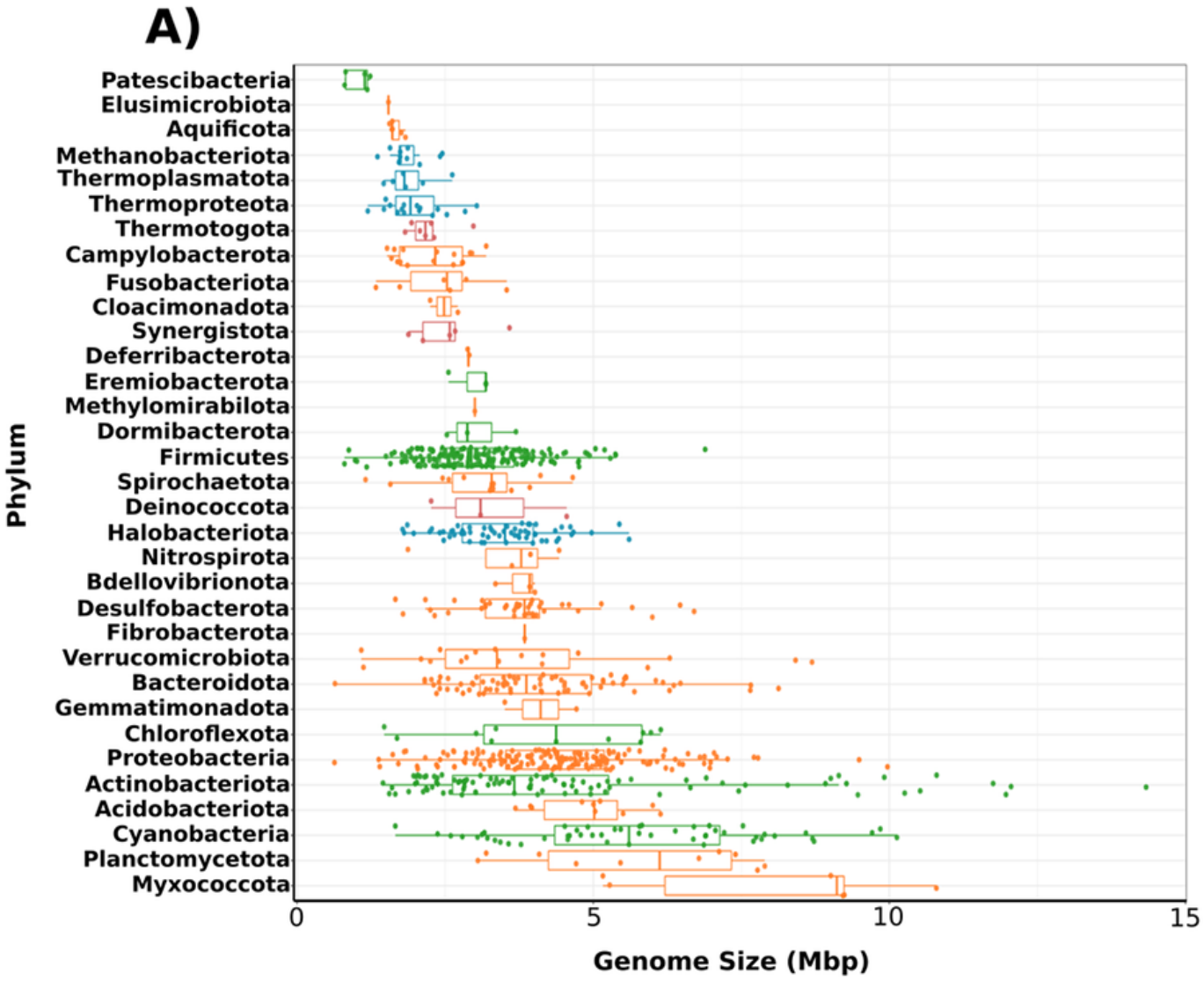
Genome size



Data source: National Center for Biotechnology Information

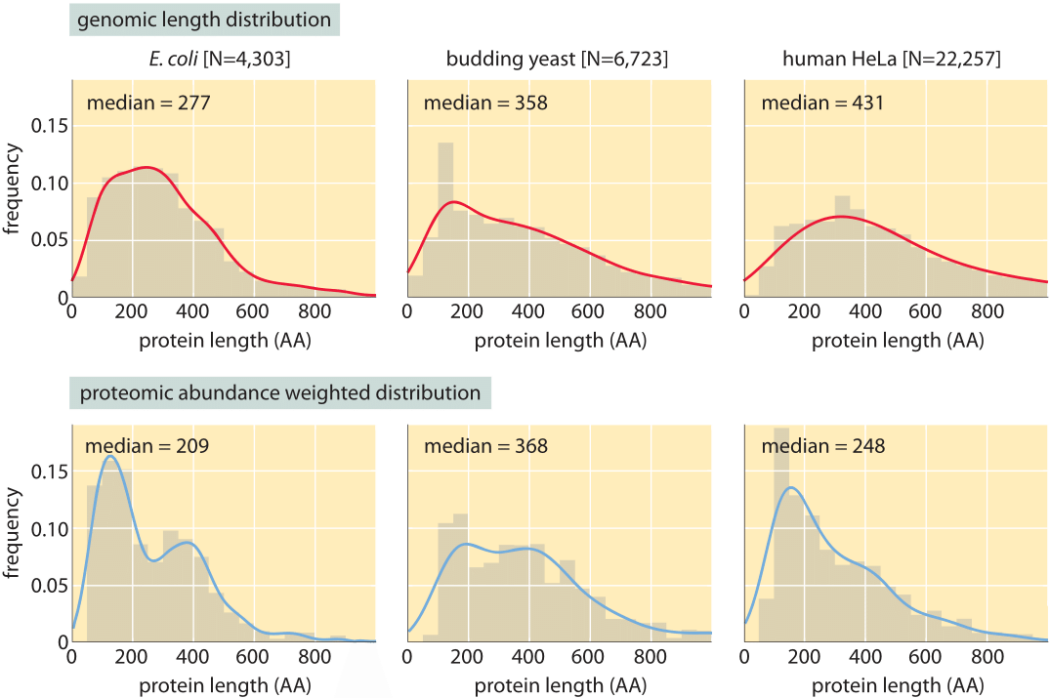
http://tom-e-white.com/datavision/05-genome-size.html

Genome size



Martinez-Gutierrez and Aylward, 2022. DOI: 10.1371/journal.pgen.1010220

How big is the average protein



organism	median protein length (amino acids)
H. sapiens	375
D. melanogaster	373
C. elegans	344
S. cerevisiae	379
A. thaliana	356
5 eukaryotes (above)	361
67 bacteria	267
15 archaea	247

http://book.bionumbers.org/how-big-is-the-average-protein/

Μέγεθος μικροβιακού γονιδιώματος

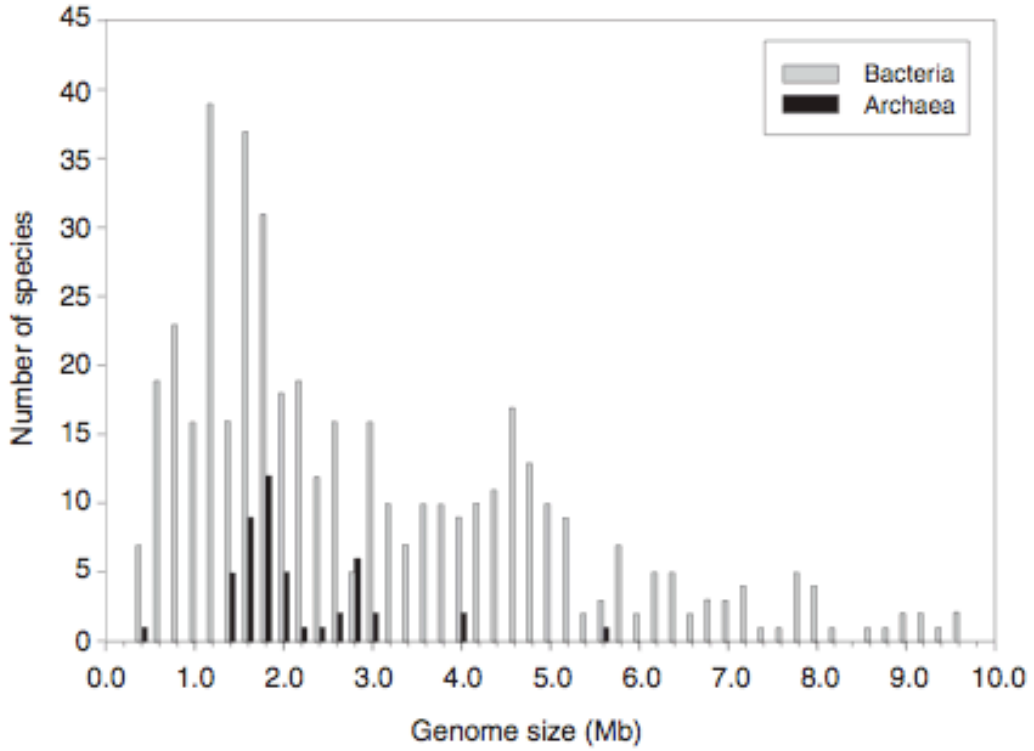


FIGURE 10.12 The distribution of genome size variation among prokaryotes based on complete genome sequencing and pulse-field gel electrophoresis (PFGE) estimates. This includes 18 complete sequences and 29 PFGE measurements for Archaea (black bars), and 125 complete sequences and 323 PFGE estimates for Bacteria (gray bars). Based on this combined dataset, the mean genome size for the Archaea is 2.22 ± 0.13 Mb, and for the Bacteria is 3.10 ± 0.09 Mb. For sequenced genomes alone, the means are 2.19 ± 0.27 Mb for Archaea and 3.40 ± 0.17 Mb for Bacteria. Values from multiple strains per species were averaged, and complete sequencing data were used preferentially where measurements had been made by both methods. Complete genome sequence data were taken from the Center for Biological Sequence Analysis (CBS) Genome Atlas Database (www.cbs.dtu.dk/services/GenomeAtlas) in the spring of 2004, and the PFGE estimates were taken from the dataset compiled by Islas *et al.* (2004), now available as part of the *Prokaryote Genome Size Database* (www.genomesize.com/prokaryotes).

Μέγεθος γονιδιώματος και τρόπος διαβίωσης

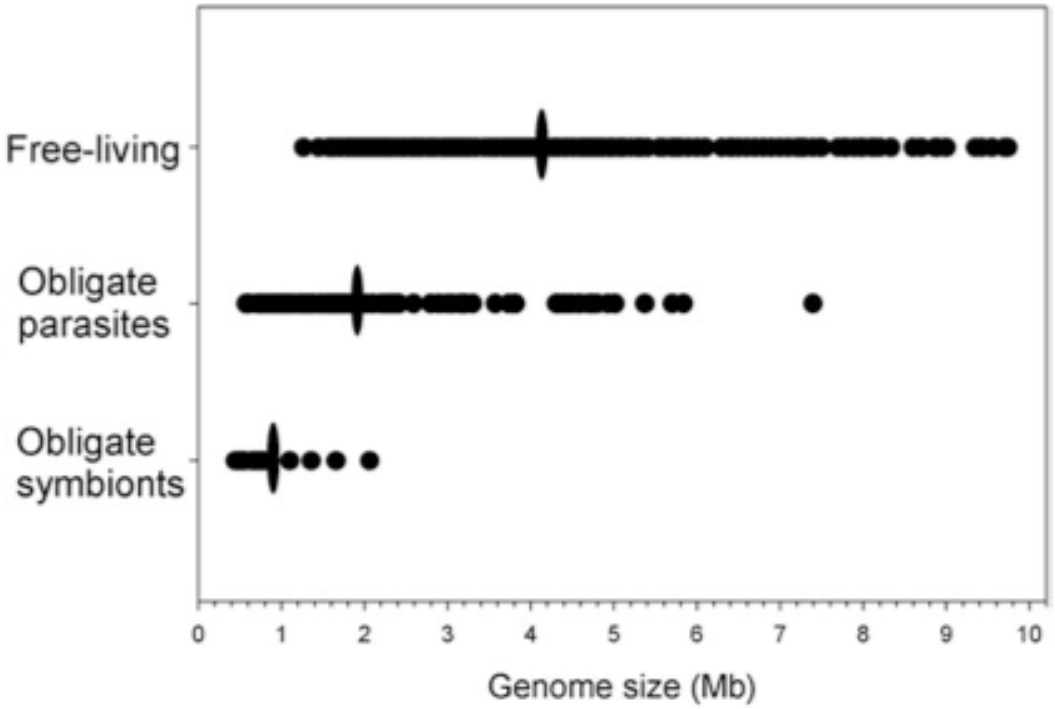


FIGURE 10.13 The distribution of genome sizes according to lifestyle in the Bacteria. Each point represents the genome size (measured by pulse-field gel electrophoresis) of one species or strain of bacteria categorized as either free-living (n = 398), obligately parasitic (n = 227), or obligately symbiotic (n = 20) as in Islas *et al.* (2004). The means for each category are indicated with vertical ellipses. Data were provided by S. Islas and A. Lazcano, Universidad Nacional Autónoma de México.

Στους προκαρυώτες, ο αριθμός γονιδίων συσχετίζεται με το μέγεθος του γονιδιώματος

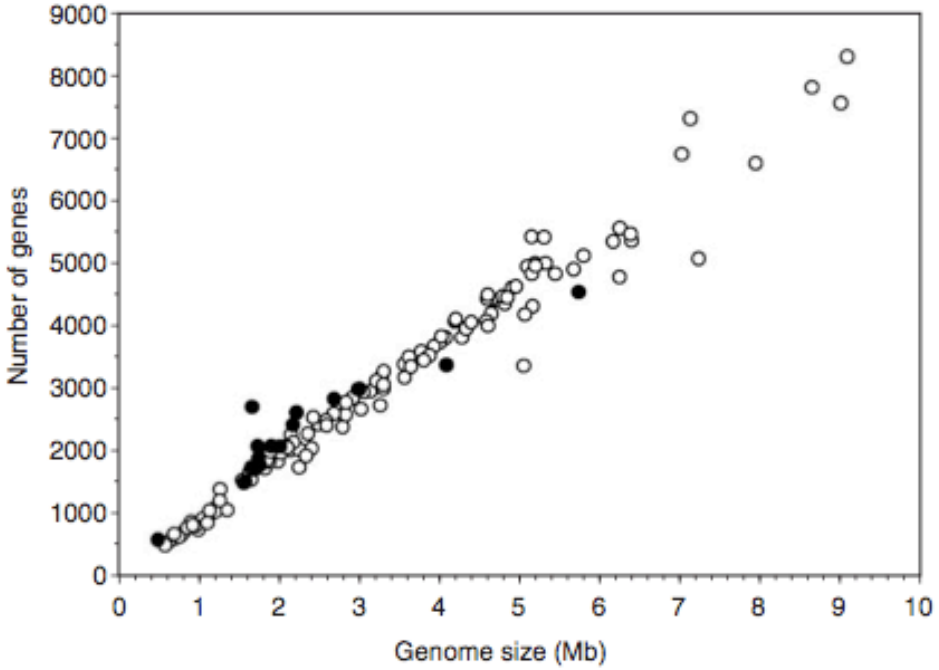


FIGURE 10.11 The relationship between gene (i.e., open reading frame) number and genome size in prokaryotes, as revealed by data from 140 completely sequenced genomes. Unlike in eukaryotes, gene number is strongly positively correlated with genome size in both Archaea (●) and Bacteria (○). The regression statistics were as follows, Archaea: $r^2 = 0.88$, $P < 0.0001$, $n = 18$; Bacteria: $r^2 = 0.97$, $P < 0.0001$, $n = 122$; all prokaryotes: $r^2 = 0.97$, $P < 0.0001$, $n = 140$. The regressions were very slightly stronger following log-transformation, but not substantially different. It has been reported that the archaeon *Aeropyrum pernix* and the bacterium *Mycobacterium leprae* represent exceptions to this trend, with the former having more than the expected number of genes and the latter exhibiting fewer than expected (Doolittle, 2002; Tanaka *et al.*, 2003). However, that these two species are distinct outliers is not so readily apparent with the large dataset used here, in which the relationship generally becomes slightly looser at the higher end of the distribution. Moreover, if the large number of pseudogenes in the *M. leprae* genome are included, this species falls on the line as well (see Mira *et al.*, 2001). Data were taken from the Center for Biological Sequence Analysis (CBS) Genome Atlas Database (www.cbs.dtu.dk/services/GenomeAtlas) in the spring of 2004.

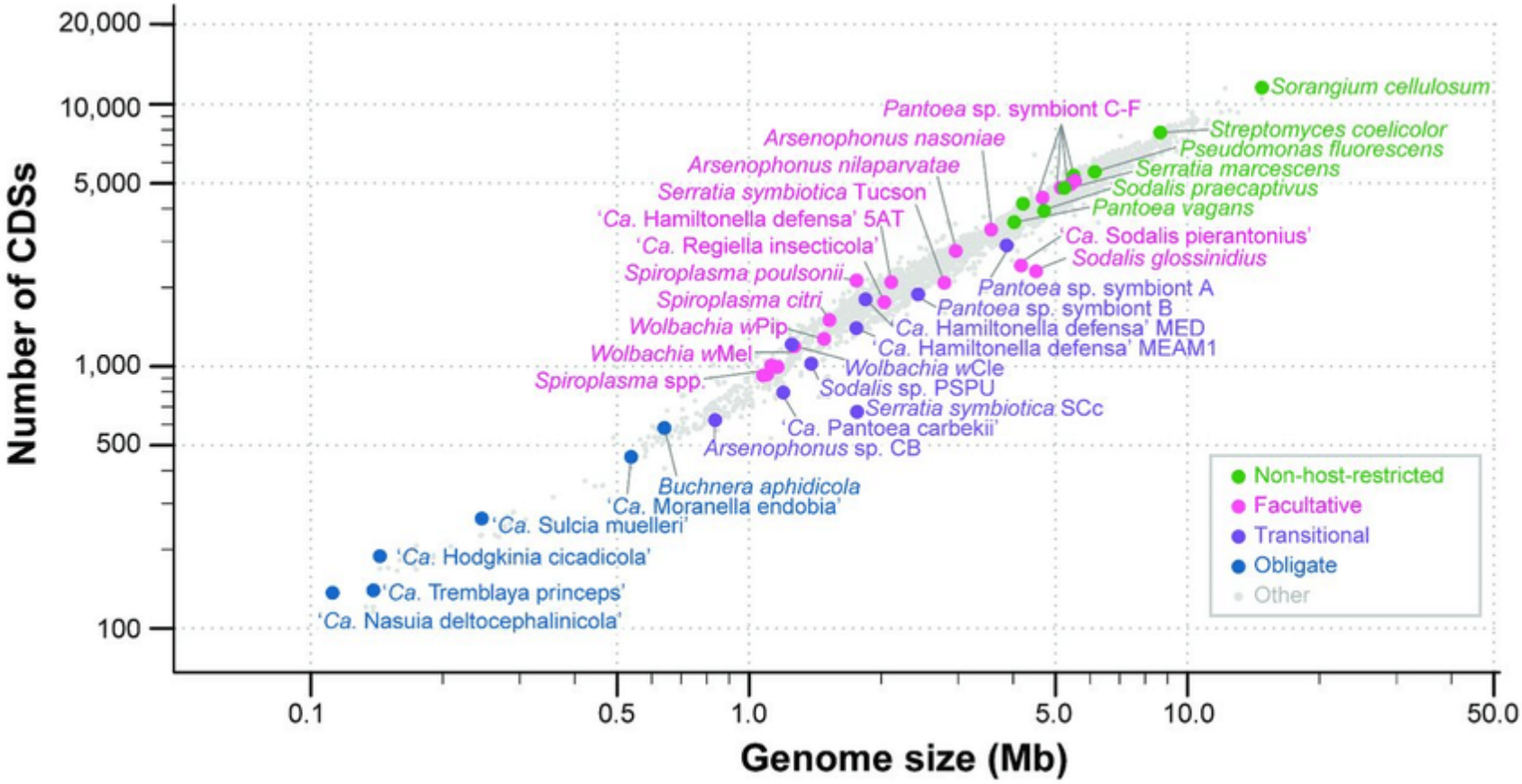
Μικρές διαγονιδιακές περιοχές (intergenic regions).

Ίσως το πολύ υψηλό effective population size στους προκαρυώτες επιτρέπει να διατηρούν τόσο συμπυκνωμένο γονιδίωμα.

Πολυπλοκότητα των οργανισμών και παράδοξο της τιμής Ν.

Genome size and coding potential

C value paradox doesn't hold for Bacteria



Lo et al., 2016. DOI:10.1093/femsre/fuw028

Association between genome size and the number of CDSs in bacteria. Based on ~ 5000 complete genomes available from GenBank as of March 2016. Representative lineages are color-coded by the lifestyle

Προφάγοι στο γονιδίωμα

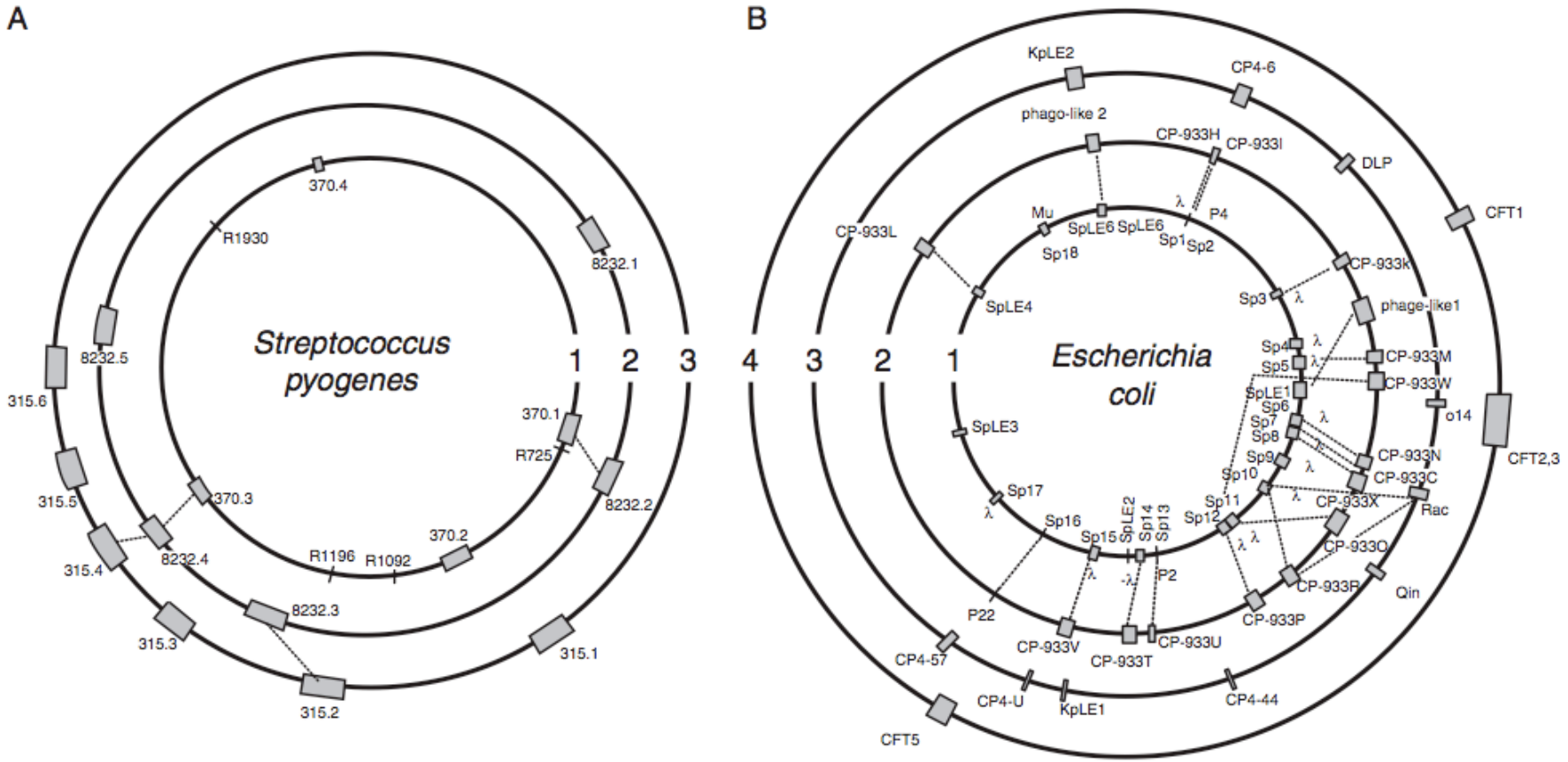


FIGURE 10.9 Prophage content and locations (gray boxes) in several strains of two species of bacteria. (A) *Streptococcus pyogenes*, also known as "group A *Streptococcus*" (GAS), which causes a wide range of infections. The numbered rings represent the genomes of three different serotypes: (1) M1, (2) M18, (3) M3. (B) *Escherichia coli*, a normally benign gut bacterium that includes some enterohemorrhagic and uropathogenic strains. The numbered rings represent the genomes of four different strains: (1) O157:H7 VT2-Sakai, (2) O157:H7 EDL933, (3) K12-MG1655, (4) CFT073. Prophages account for about 12% and 16% of the *S. pyogenes* and pathogenic *E. coli* genomes, respectively (Canchaya *et al.*, 2003). Note that the circumferences of these schematic circular drawings are not to scale and therefore do not reflect the real relative lengths of the chromosomes depicted. Adapted from Canchaya *et al.* (2003), reproduced by permission (© American Society for Microbiology).

Πόσο σταθερή είναι η αρχιτεκτονική ενός γονιδιώματος.

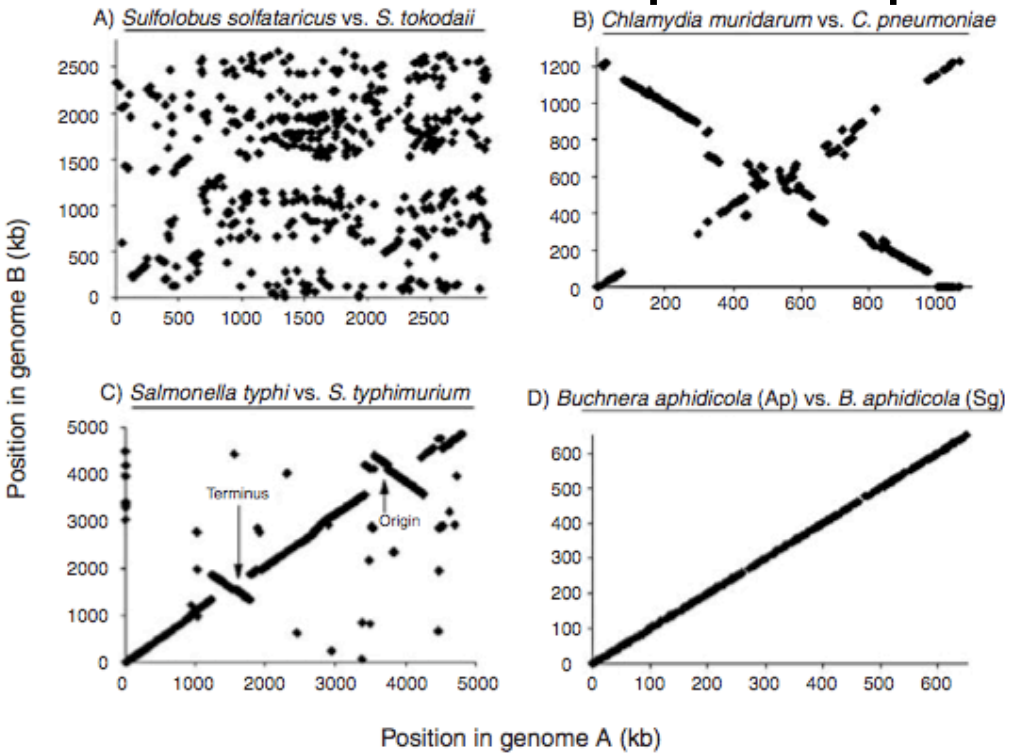


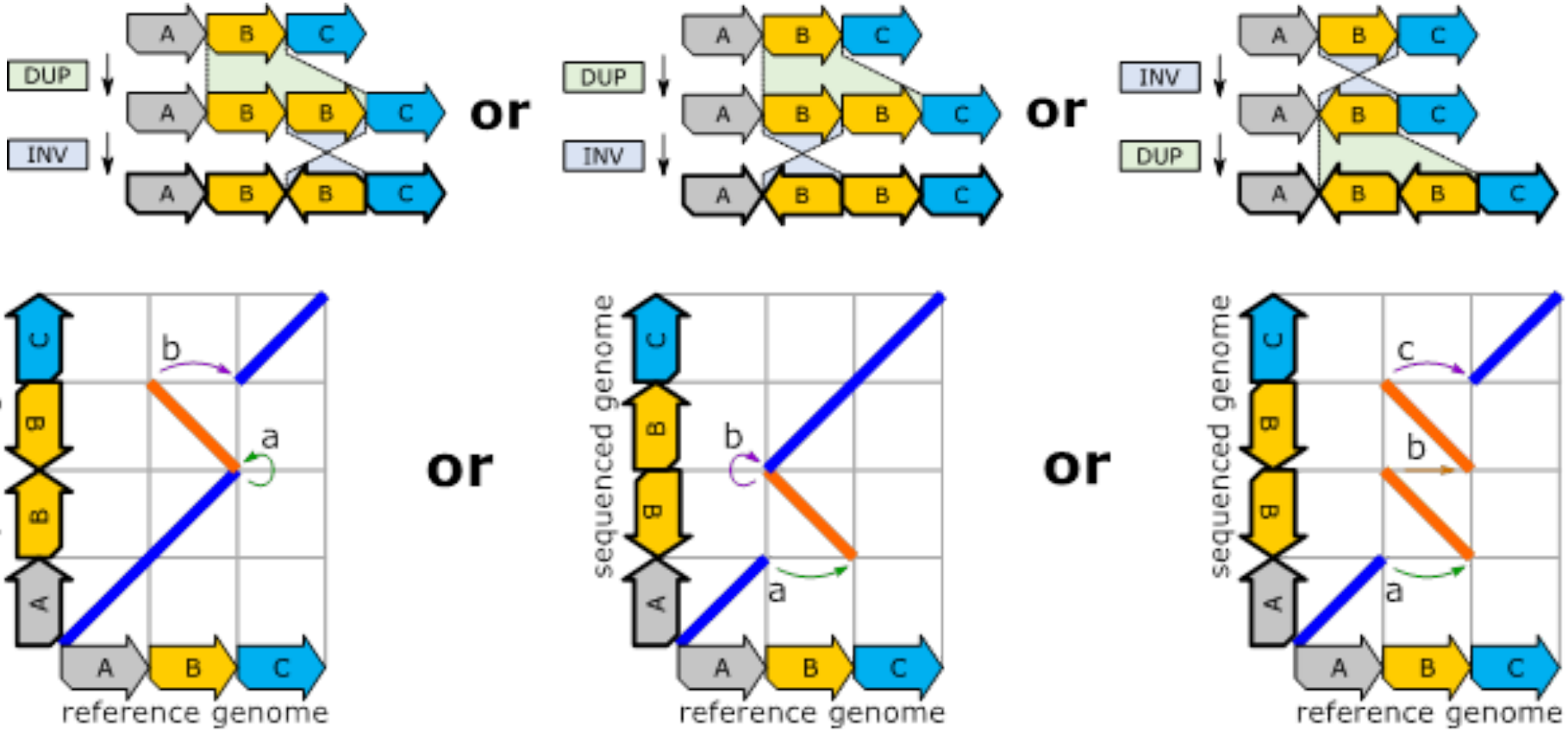
FIGURE 10.6 Gene position plots showing examples of both plasticity and stability in gene order between closely related species of prokaryotes. In these plots, the location of a given gene, measured as its distance from a given starting point in kilobases (kb), is plotted on one axis each for the two species being compared. Unless otherwise indicated, the origin of the axes represents the origin of replication in the chromosomes. (A) The archaeons *Sulfolobus solfataricus* and *S. tokodaii*, whose genomes share very little common gene order and are clearly extremely dynamic. (B) The bacteria *Chlamydia muridarum* and *C. pneumoniae*, which exhibit a clear "X-alignment," indicating a single, large, symmetrical inversion around the origin of replication (see also Eisen *et al.*, 2000; Hughes, 2000). (C) The bacteria *Salmonella typhi* and *S. typhimurium*, which show evidence of two smaller symmetrical inversions, one around the origin of replication and one around the replication terminus. (D) Two strains (or possibly species) of the endosymbiotic bacterium *Buchnera aphidicola* living in distantly related aphid hosts (Ap = *Acyrthosiphon pisum*; Sg = *Schizaphis graminum*). In this case, there has been remarkable stasis in gene order for 50–70 million years, despite considerable sequence divergence (see Tamas *et al.*, 2002). Based on a figure presented by Mira *et al.* (2002), reproduced by permission (© Elsevier Inc.).

Dotplot για ορθόλογα γονίδια μεταξύ δύο προκαρυωτών του ίδιου είδους.

Κάθε κουκίδα στο Dotplot είναι η θέση του ορθόλογου γονιδίου σε δύο διαφορετικά γονιδιώματα.

Κάποιοι οργανισμοί έχουν σταθερή γονιδιωματική αρχιτεκτονική και κάποιοι άλλοι όχι.

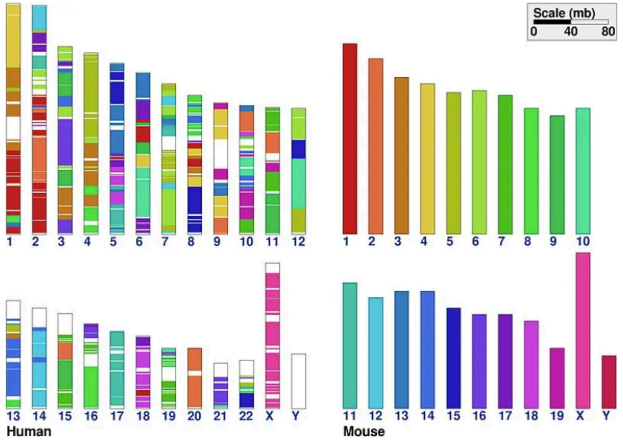
Dotplot: Πόσο σταθερή είναι η αρχιτεκτονική ενός γονιδιώματος;



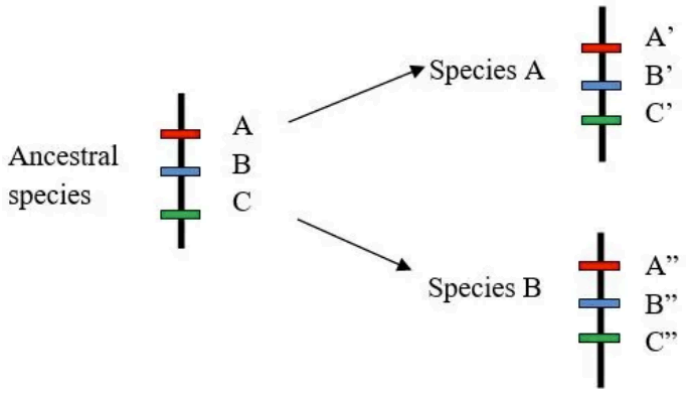
http://itbe.hanyang.ac.kr/ak/MSV/

Γονιδιωματικό Dotplot: συνταινικότητα και συγγραμικότητα;

Συνταινικότητα (synteny): Όταν μεταξύ δύο γονιδιωμάτων τα ορθόλογα παραμένουν στη ίδια γειτονιά (π.χ. ίδιο χρωμόσωμα). Δεν είναι απαραίτητο να παραμένουν με την ίδια σειρά.



Συγγραμικότητα (collinearity): Όταν μεταξύ δύο γονιδιωμάτων τα ορθόλογα παραμένουν στη ίδια γειτονιά με την ίδια σειρά.



https://www.differencebetween.com/difference-between-synteny-and-collinearity/

Οι τεχνολογίες αλληλούχισης νέας γενιάς στην κλινική μικροβιολογία

http://www.ncbi.nlm.nih.gov/pubmed/22868263

Transforming clinical microbiology with bacterial genome sequencing.
Didelot X1, Bowden R, Wilson DJ, Peto TE, Crook DW.

Nat Rev Genet. 2012 Sep;13(9):601-12. doi: 10.1038/nrg3226. Epub 2012 Aug 7.

Σκοπός της κλινικής μικροβιολογίας είναι

- η γρήγορη ταυτοποίηση παθογόνων μικροοργανισμών σε κλινικά δείγματα, για την αποτελεσματικότερη αντιμετώπιση του ασθενούς (διαγνωστική μικροβιολογία)
- η παρακολούθηση της εξάπλωσης μια επιδημίας (μικροβιολογία δημόσιας υγειινής)

Τρείς βασικές χρησιμότητες της αλληλούχισης μικροβιακών γονιδιωμάτων:

- Ταυτοποίηση του είδους.
- Έλεγχος ιδιοτήτων, όπως ανθεκτικότητα σε αντιβιοτικά ή ένταση παθογένεσης
- Παρακολούθηση της εμφάνισης και εξάπλωσης μιας επιδημίας.

Σύντομα, η τεχνολογία θα είναι γρήγορη, ακριβής, και φτηνή για κλινική ρουτίνα σε όλους τους παθογόνους μικροοργανισμούς, όπως:

- Ιούς
- Βακτήρια
- Μήκυτες
- Παράσιτα

Οι τεχνολογίες νέας γενιάς αλληλούχισης στην κλινική μικροβιολογία

Η διαδικασία της ταυτοποίησης βακτηρίων με κλασικές μεθόδους μπορεί να αποτελείται από πολλές, χρονοβόρες και εξειδικευμένες αναλύσεις.

Το βακτήριο πρέπει πρώτα να καλλιεργηθεί από το δείγμα.

Να ταυτοποιηθεί το είδος.

Να καθοριστεί η παθογένειά του.

Να ελεγχθεί για ανθεκτικότητα/ευαισθησία σε αντιβιοτικά.

Η διαδικασία αυτή μπορεί να διαρκέσει από μέρες έως και εβδομάδες ή σε κάποιες περιπτώσεις (αργά αναπτυσσόμενα βακτήρια - Mycobacterium tuberculosis) μέχρι και μήνες.

Οι περισσότερες βακτηριακές μολύνσεις οφείλονται σε ~20 είδη.

Large scale bacterial genome projects

An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes

Blanca M. Perez-Sepulveda, Darren Heavens, Caisey V. Pulford, Alexander V. Predeus, Ross Low, Hermione Webster, Gregory F. Dykes, Christian Schudoma, Will Rowe, James Lipscombe, Chris Watkins, Benjamin Kumwenda, Neil Shearer, Karl Costigan, Kate S. Baker, Nicholas A. Feasey, Jay C. D. Hinton, Neil Hall & The 10KSG consortium

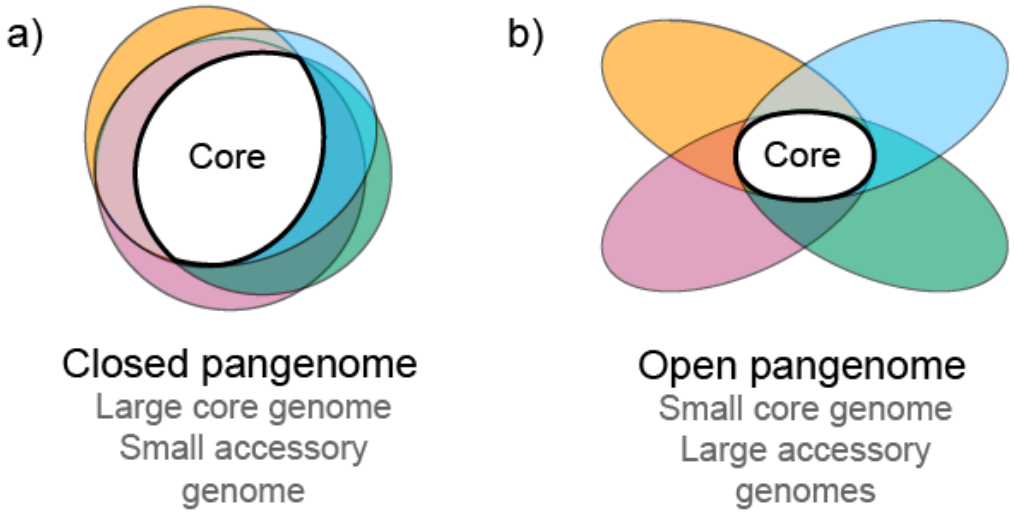
Genome Biology **22**, Article number: 349 (2021) | Cite this article

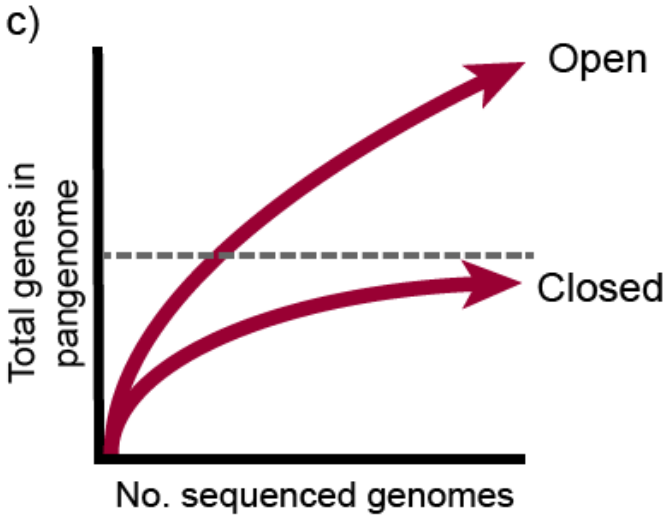
- Salmonella enterica
- A diverse collection of 10,419 isolates from low- and middle-income countries.
- The genomes were sequenced with Illumina, with a total reagent cost (DNA extraction and genome sequence generation—excluding staff time) of **less than USD$10 per genome**.

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02536-3

Πανγονιδίωμα
Ανοιχτό και κλειστό

a) Closed pangenomes are characterized by large core genomes and small accessory genomes. b) Open pangenomes tend to have small core genomes and large accessory genomes. c) The size of open pangenomes tends to increase with every added genome, meanwhile closed pangenome's size tends to be asymptotic despite adding more genomes. Due to this characteristic, complete pangenome size for closed pangenomes can be predicted.





diversity

MDPI

Article

Comparative Analysis of the Core Proteomes among the *Pseudomonas* Major Evolutionary Groups Reveals Species-Specific Adaptations for *Pseudomonas aeruginosa* and *Pseudomonas chlororaphis*

Marios Nikolaidis [1], Dimitris Mossialos [2], Stephen G. Oliver [3] and Grigorios D. Amoutzias [1,*]

1 Bioinformatics Laboratory, Department of Biochemistry and Biotechnology, University of Thessaly, 41500 Larissa, Greece; marionik23@gmail.com

2 Microbial Biotechnology-Molecular Bacteriology-Virology Laboratory, Department of Biochemistry and Biotechnology, University of Thessaly, 41500 Larissa, Greece; mosial@bio.uth.gr

3 Cambridge Systems Biology Centre & Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK; sgo24@cam.ac.uk

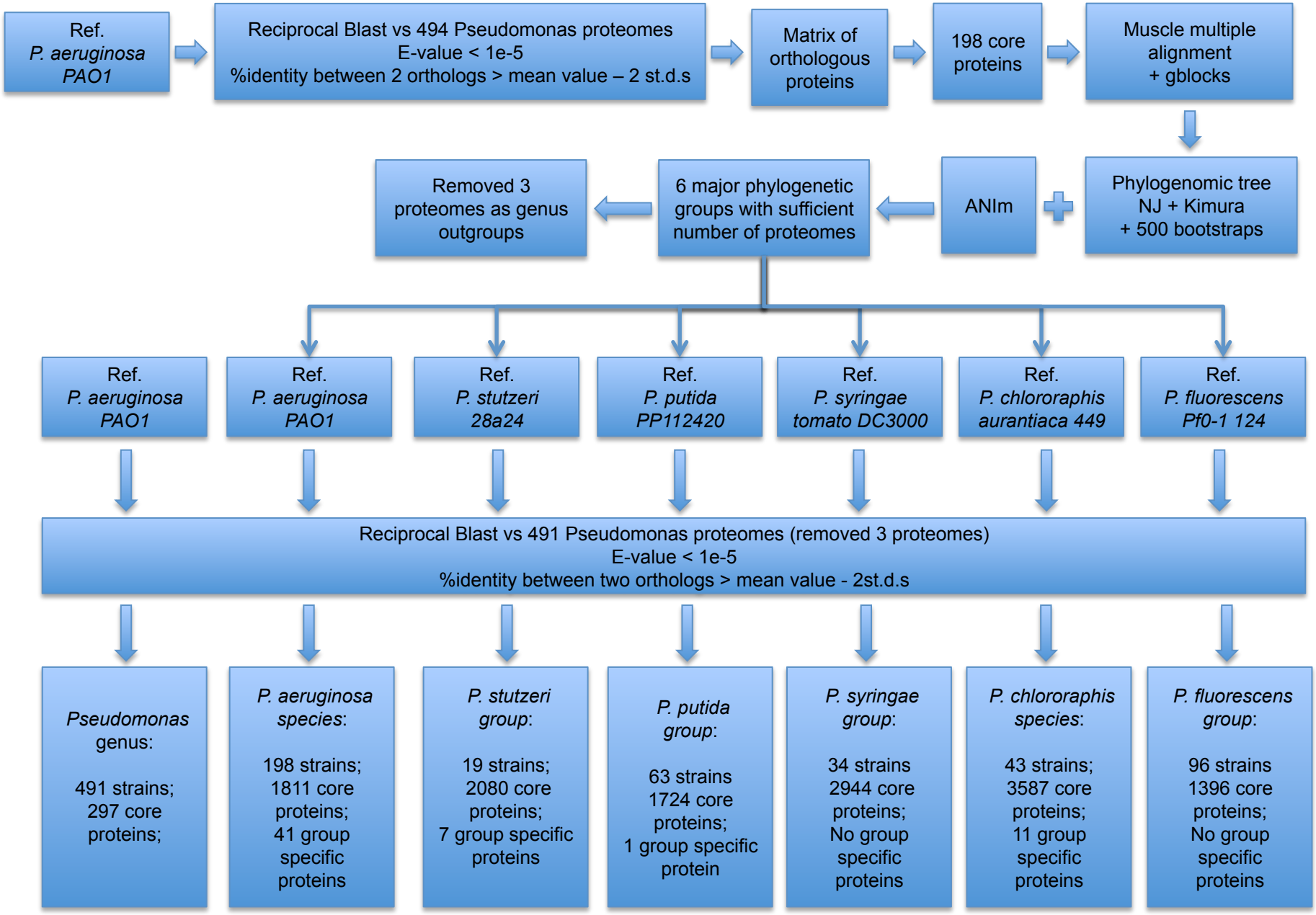
* Correspondence: amoutzias@bio.uth.gr; Tel.: +30-2410-565-289; Fax: +30-2410-565-290

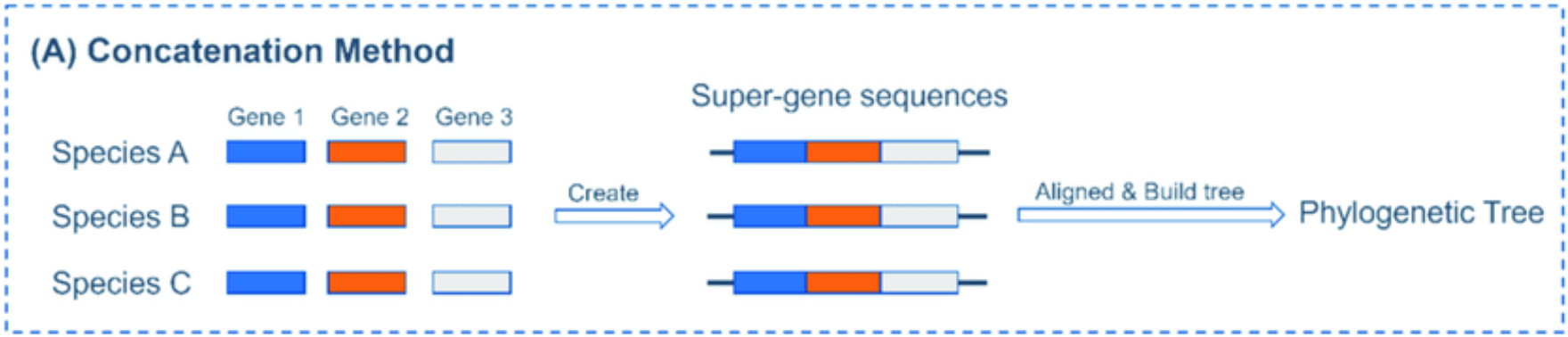
Received: 22 June 2020; Accepted: 22 July 2020; Published: 24 July 2020

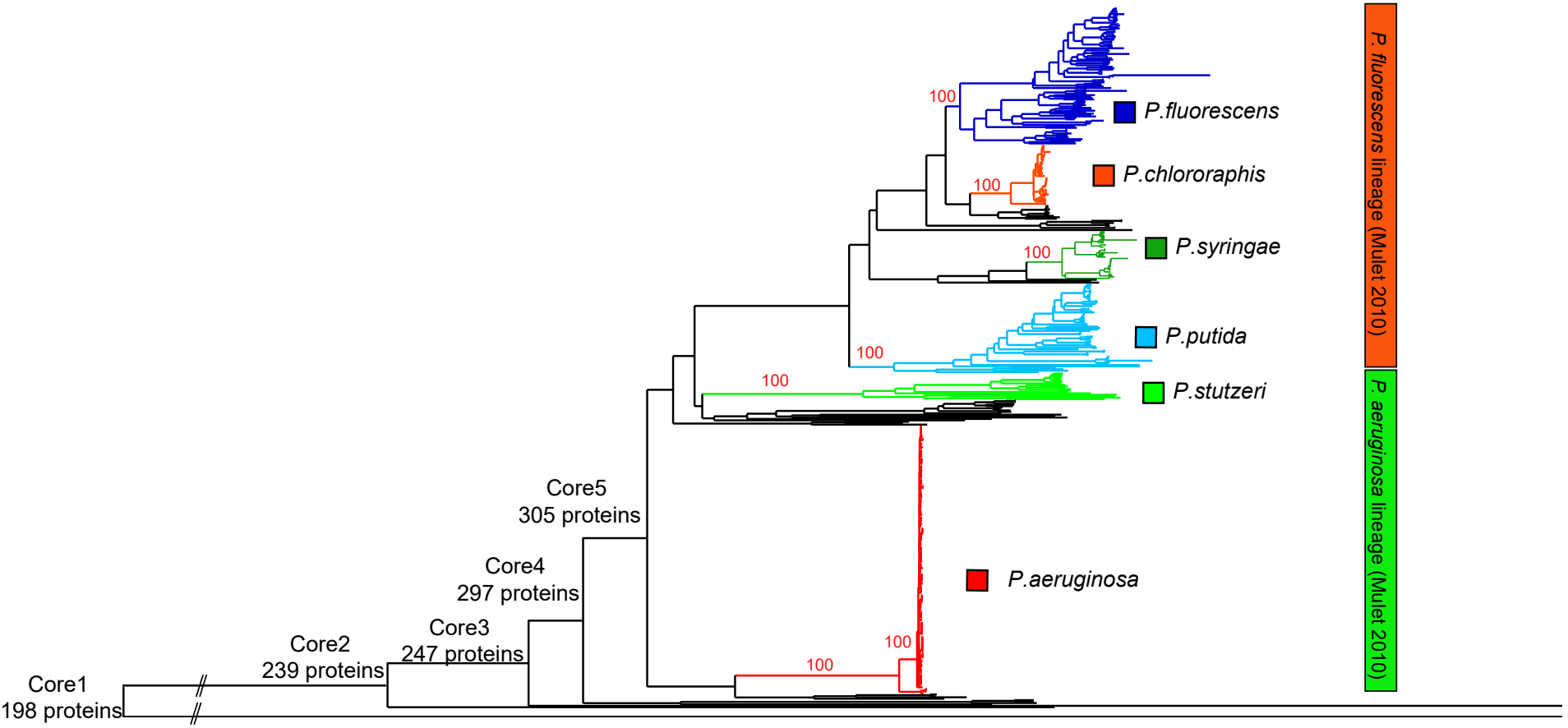
Core/accessory genome

		GeneA	GeneB	GeneC	GeneD	GeneE	GeneF	GeneG	GeneH	GeneI	GeneJ	GeneK	GeneL
Spec1	strain1												
Spec1	strain2												
Spec1	strain3												
Spec1	strain4												
Spec1	strain5												
Spec2	strain6												
Spec2	strain7												
Spec2	strain8												
Spec2	strain9												
Spec2	strain10												

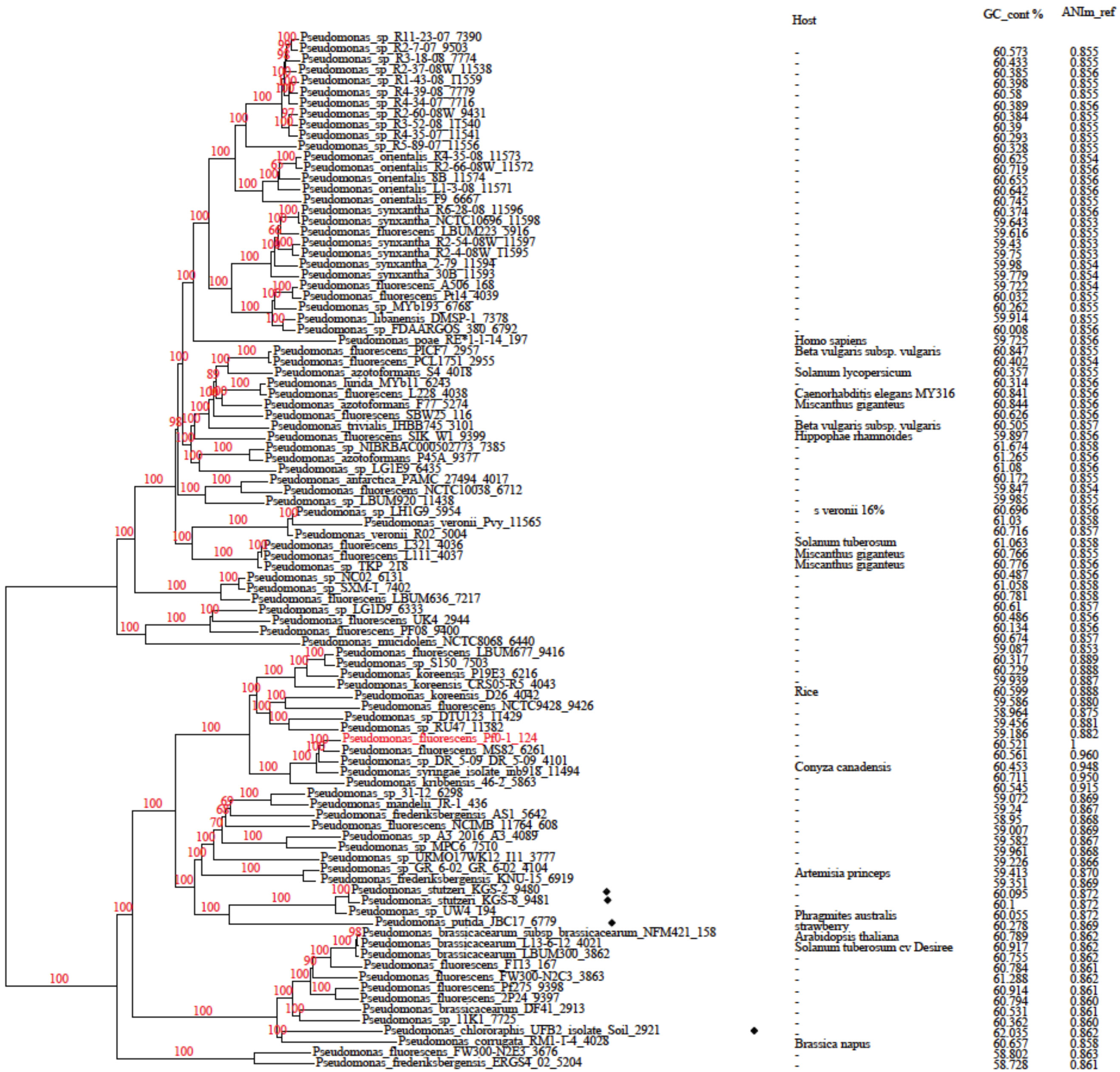
Genus core genes
Species core genes
Species fingerprints
Other

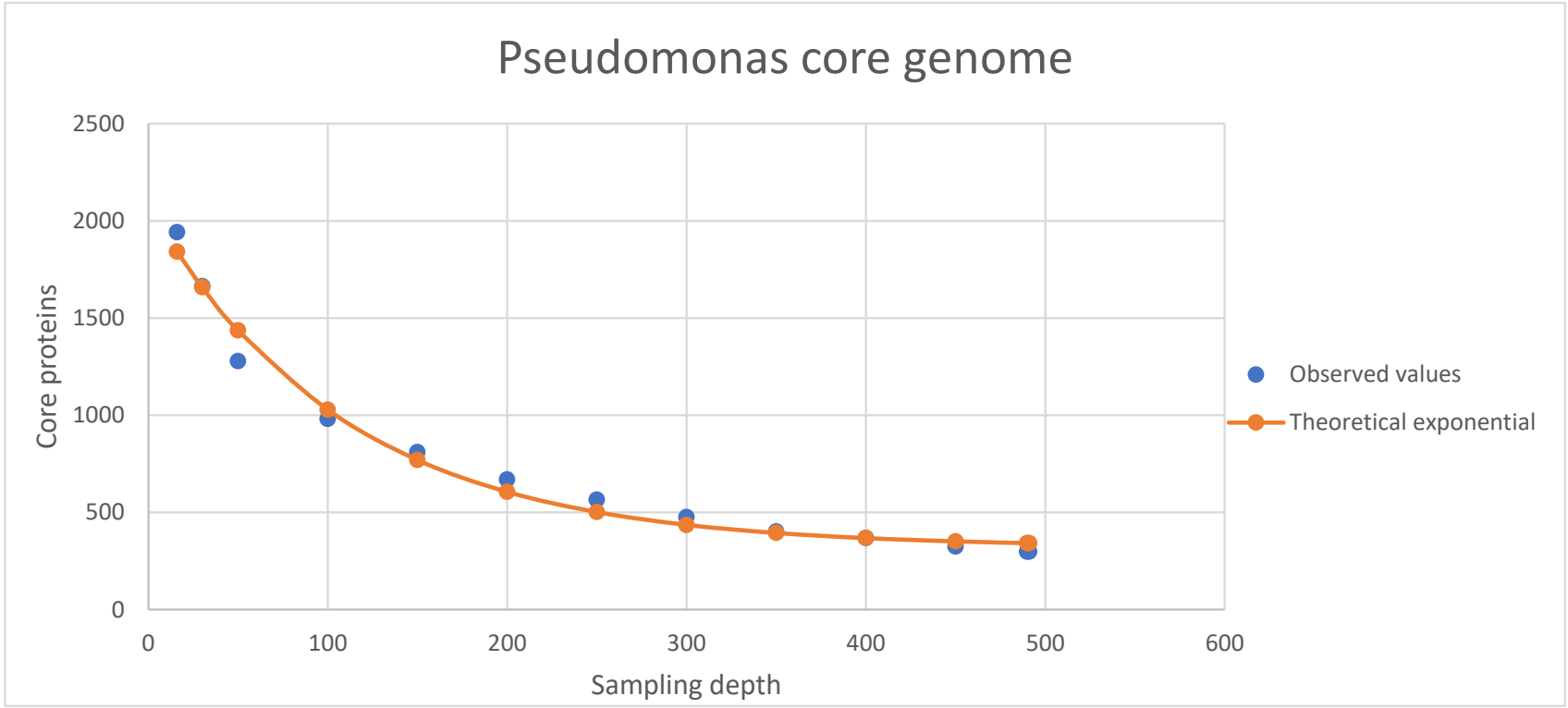

Ref. *P. aeruginosa PAO1*
Reciprocal Blast vs 494 Pseudomonas proteomes
E-value < 1e-5
%identity between 2 orthologs > mean value – 2 st.d.s
Matrix of orthologous proteins
198 core proteins
Muscle multiple alignment + gblocks
Phylogenomic tree NJ + Kimura + 500 bootstraps
ANIm
6 major phylogenetic groups with sufficient number of proteomes
Removed 3 proteomes as genus outgroups
Ref. *P. aeruginosa PAO1*
Ref. *P. aeruginosa PAO1*
Ref. *P. stutzeri 28a24*
Ref. *P. putida PP112420*
Ref. *P. syringae tomato DC3000*
Ref. *P. chlororaphis aurantiaca 449*
Ref. *P. fluorescens Pf0-1 124*
Reciprocal Blast vs 491 Pseudomonas proteomes (removed 3 proteomes)
E-value < 1e-5
%identity between two orthologs > mean value - 2st.d.s
Pseudomonas genus:
491 strains;
297 core proteins;
P. aeruginosa species:
198 strains;
1811 core proteins;
41 group specific proteins
P. stutzeri group:
19 strains;
2080 core proteins;
7 group specific proteins
P. putida group:
63 strains
1724 core proteins;
1 group specific protein
P. syringae group:
34 strains
2944 core proteins;
No group specific proteins
P. chlororaphis species:
43 strains;
3587 core proteins;
11 group specific proteins
P. fluorescens group:
96 strains
1396 core proteins;
No group specific proteins


(A) Concatenation Method
Gene 1
Gene 2
Gene 3
Species A
Species B
Species C
Create
Super-gene sequences
Aligned & Build tree
Phylogenetic Tree


100
100
100
100
100
100
100
P.fluorescens
P.chlororaphis
P.syringae
P.putida
P.stutzeri
P.aeruginosa
P. fluorescens lineage (Mulet 2010)
P. aeruginosa lineage (Mulet 2010)
Core5
305 proteins
Core4
297 proteins
Core3
247 proteins
Core2
239 proteins
Core1
198 proteins

P. fluorescens is not a species + miss-annotations




Pseudomonas core genome
Core proteins
2500
2000
1500
1000
500
0
0
100
200
300
400
500
600
Sampling depth
Observed values
Theoretical exponential

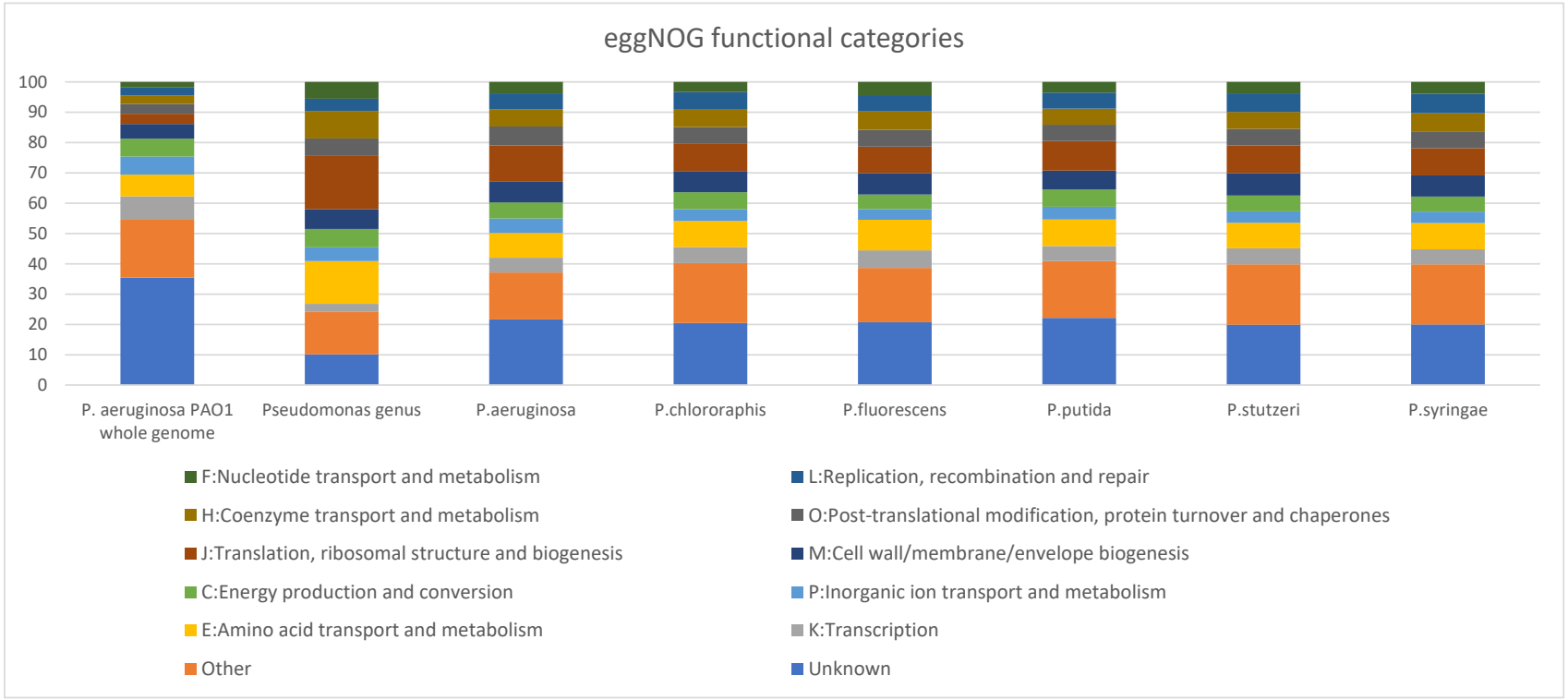

eggNOG functional categories
100
90
80
70
60
50
40
30
20
10
0
P. aeruginosa PAO1 whole genome
Pseudomonas genus
P.aeruginosa
P.chlororaphis
P.fluorescens
P.putida
P.stutzeri
P.syringae
F:Nucleotide transport and metabolism
L:Replication, recombination and repair
H:Coenzyme transport and metabolism
O:Post-translational modification, protein turnover and chaperones
J:Translation, ribosomal structure and biogenesis
M:Cell wall/membrane/envelope biogenesis
C:Energy production and conversion
P:Inorganic ion transport and metabolism
E:Amino acid transport and metabolism
K:Transcription
Other
Unknown

Table 1. Summary table of the core proteomes and their functional categories for the *Pseudomonas* genus and the six evolutionary groups. Note that, for the phylogenomic analysis of the entire genus (core4), we used the alignment produced by core1, that also includes 3 outgroups.

	Genus Core4	***Pseudomonas aeruginosa***	***Pseudomonas chlororaphis***	***Pseudomonas stutzeri***	***Pseudomonas putida***	***Pseudomonas fluorescens***	***Pseudomonas syringae***
Number_of_Strains	491	189	43	19	63	96	34
Amino acids in phylogenomic alignment	Core1:27,997	31,145	103,483	193,427	155,470	115,099	110,643
Core proteins	297	1811	3587	2080	1724	1396	2944
% core proteins with significant presence in other groups (≥90% presence)	100.0	44.3	38.7	70.1	70.5	62.7	50.1
Group-specific core proteins	-	41	11	7	1	0	0
Relaxed group-specific core proteins (10% in others)	-	84	32	32	4	0	61
Relaxed group-specific core proteins (20% in others)	-	116	61	51	4	0	87
%core—Unknown	19.9	33.4	30.4	27.1	25.4	26.6	31.3
%core—Other	24.9	16.6	20.1	19.9	17.9	17.1	18.8
%core—K:Transcription	4.7	7.3	7.8	4.6	5.4	6.5	5.9
%core—E:Amino acid transport and metabolism	11.4	7.2	8.6	7.2	9.5	10.7	8.2
%core—P:Inorganic ion transport and metabolism	4.4	5.9	5.5	4.5	5.5	4.7	5.8
%core—C:Energy production and conversion	4.4	6.5	5.8	6.3	7.0	6.1	4.8
%core—M:Cell wall/membrane/envelope biogenesis	6.7	4.4	5.5	6.1	5.3	5.2	5.7
%core—J:Translation, ribosomal structure and biogenesis	13.1	5.6	4.2	6.9	7.1	6.0	5.2
%core—H:Coenzyme transport and metabolism	7	3.4	3.5	4.6	5.0	5.1	4.2
%core—L:Replication, recombination and repair	3.4	3.0	2.8	5.1	4.4	4.0	3.7

Table 2. Summary of important group-specific and related proteins in *P. aeruginosa* and *P. chlororaphis*.

	Orthologue				Closest Paralogue in Reference Strain				Closest Homologue Outside of Group			
Organism	Name	General Function	Presence within Group	Presence outside of Group	Accession	Name	Annotation	% Identity	Organism	Name	Annotation	% Identity
P. aeruginosa	cntO	pseudopalin production	185/189	234/305	PA0151	-	probable TonB-dependent receptor	34.84	*Pseudomonas furukawaii*	KF707C_RS00865	Ton-B-dependent receptor	34.56
P. aeruginosa	cntL	pseudopalin production	189/189	0/305	-	-	-	-	-	-	-	-
P. aeruginosa	cntM	pseudopalin production	189/189	0/305	-	-	-	-	-	-	-	-
P. aeruginosa	cntI	pseudopalin production	187/189	2/305	PA2628	-	hypothetical protein	34.17	*Pseudomonas mendocina*	EL191_RS11685	DMT family transporter	34.52
P. aeruginosa	pvdL	pyoverdine biosynthesis	171/189	255/305	PA2400	pvdJ	pvdJ	47.28	*Pseudomonas fulva*	CJ462_RS07705	amino acid adenylation domain-containing protein	39.98
P. aeruginosa	pchF	pyochelin biosynthesis	176/189	58/305	PA4226	pchE	dihydroaeruginoic acid synthetase	39.57	*Pseudomonas* sp. R2-7-07 9503	C4J86_RS18060	amino acid adenylation domain-containing protein	26
P. aeruginosa	plcB	motility	189/189	0/305	-	-	-	-		-	-	-
P. aeruginosa	acp1	motility	189/189	0/305	PA2966	acpP	acyl carrier protein	53.4	*Pseudomonas mesoacidophila*	B7P44_RS05820	acyl carrier protein	61.11
P. aeruginosa	mucE	mucoidy	189/189	0/305	-	-	-	-	-	-	-	-
P. aeruginosa	srfA	mucoidy	189/189	0/305	-	-	-	-	-	-	-	-
P. aeruginosa	tse1	toxin/antitoxin operon	189/189	0/305	-	-	-	-	-	-	-	-
P. aeruginosa	tsi1	toxin/antitoxin operon	189/189	1/305	-	-	-	-	-	-	-	-
P. aeruginosa	tse2	toxin/antitoxin operon	187/189	0/305	-	-	-	-	-	-	-	-
P. aeruginosa	tsi2	toxin/antitoxin operon	189/189	0/305	-	-	-	-	-	-	-	-
P. aeruginosa	tse3	toxin/antitoxin operon	189/189	0/305	-	-	-	-	-	-	-	-
P. aeruginosa	tsi3	toxin/antitoxin operon	186/189	0/305	-	-	-	-	-	-	-	-

Table 2. *Cont.*

Orthologue					Closest Paralogue in Reference Strain				Closest Homologue Outside of Group			
P. aeruginosa	esrC	multidrug efflux	189/189	0/305	PA4600	nfxB	transcriptional regulator NfxB	60.71	*Pseudomonas chlororaphis*	C4K30_RS15575	TetR/AcrR family transcriptional regulator	66.47
P. aeruginosa	mexC	multidrug efflux	186/189	240/305	PA0425	mexA	resistance-nodulation cell division (RND) multidrug efflux membrane fusion protein MexA precursor	44.76	*P.* sp URMO17WK12 I11	PSHI_RS20595	hemolysin secretion protein D	27.78
P. aeruginosa	mexD	multidrug efflux	188/189	264/305	PA2018	mexY	resistance nodulation cell division (RND) multidrug efflux transporter MexY	52	*Pseudomonas alcaligenes*	A0T30_RS19090	multidrug efflux RND transporter permease subunit	53.09
P. aeruginosa	lecB	biofilm formation	188/189	0/305	-	-	-	-	-	-	-	-
P. chlororaphis	C4K22 RS23595	holin family	43/43	0/451	-	-	-	-	-	-	-	-
P. chlororaphis	C4K22 RS23415	mitomycin biosynthesis	42/43	0/451	-	-	-	-	-	-	-	-