Γονιδιωματική

ΜΠΣ2

Γρ. Αμούτζιας

Οι τεχνολογίες

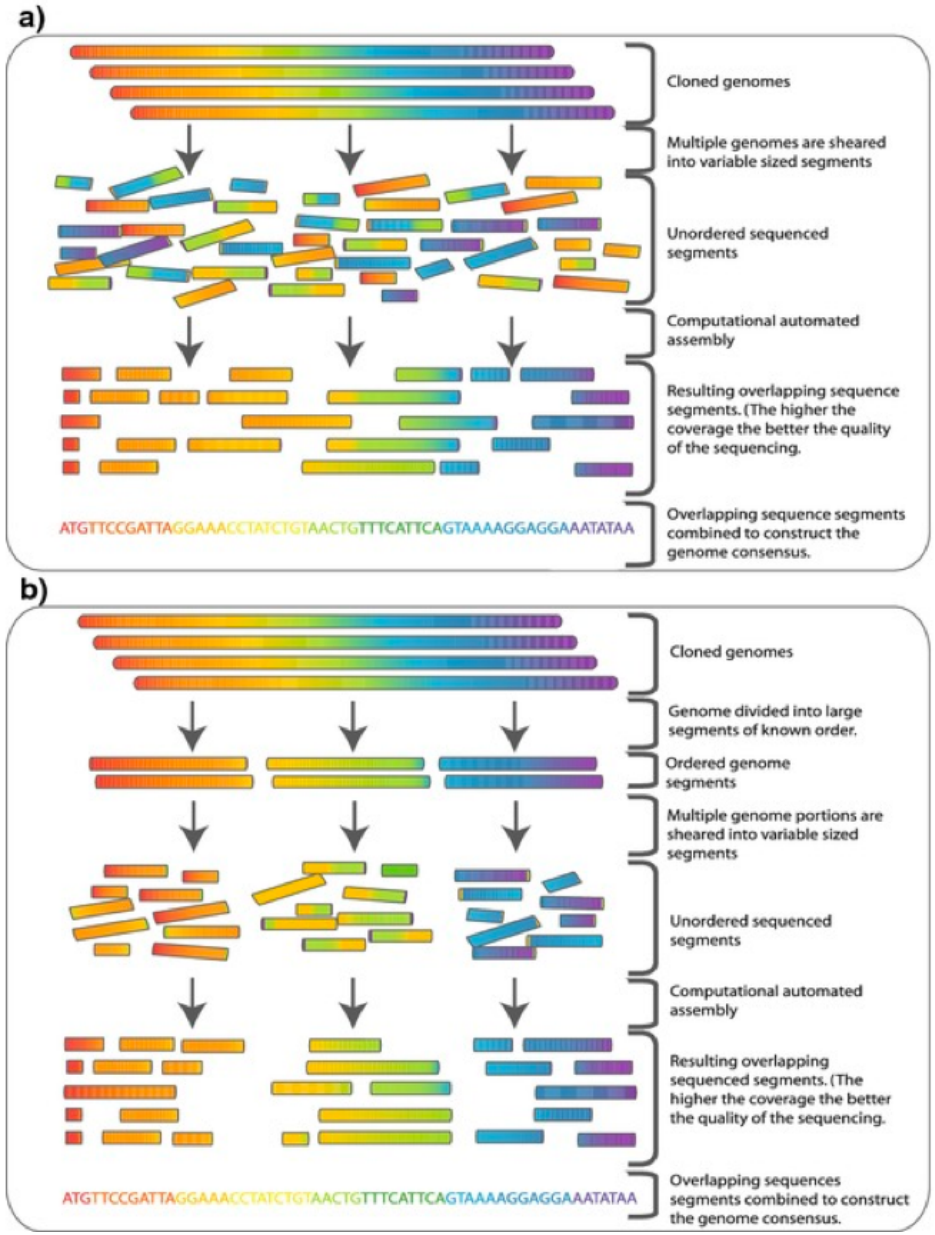
Κυριότερες τεχνολογίες

- Sanger

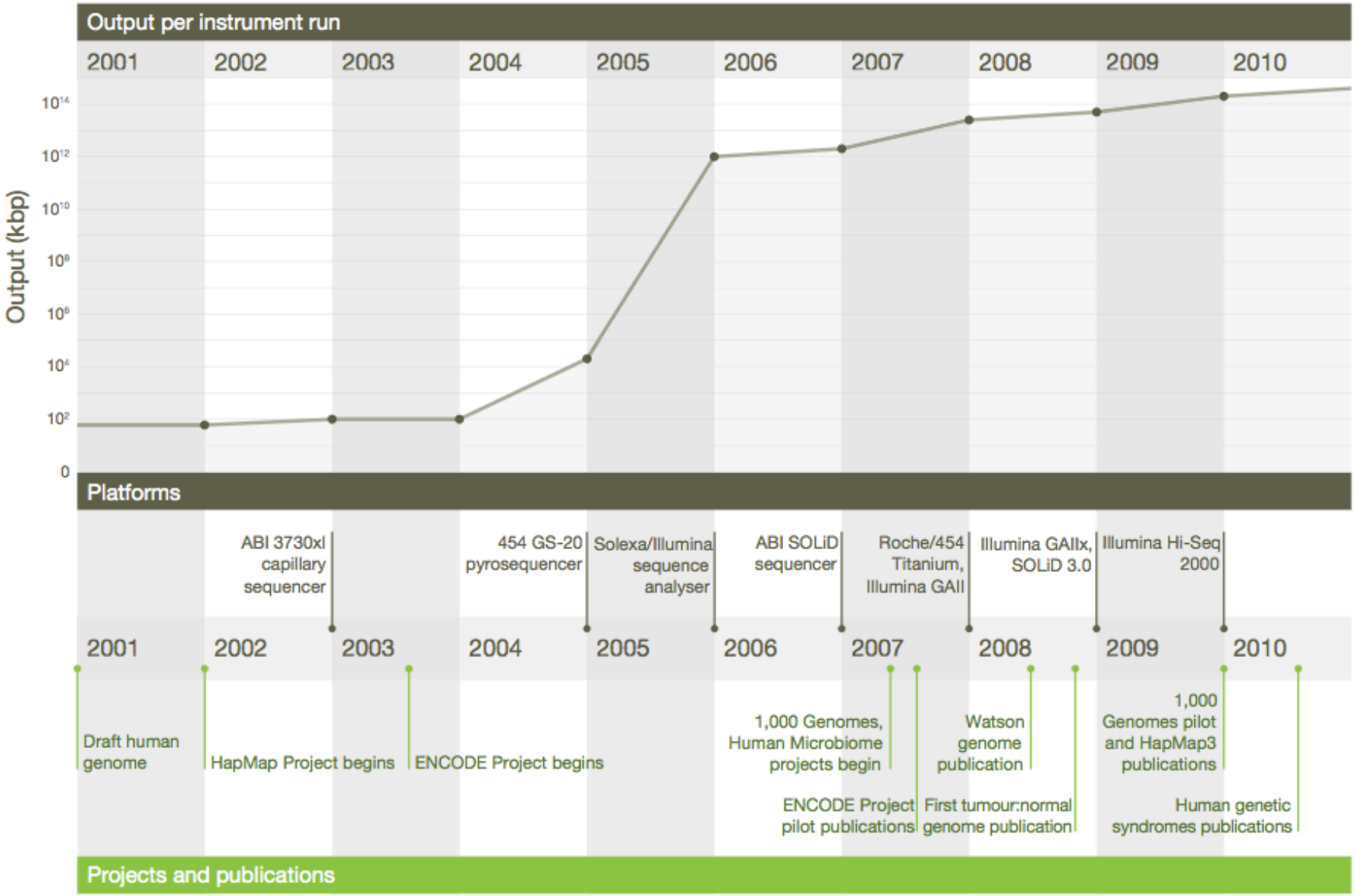
- 454 pyrosequencing
- Solid
- Illumina
- Pacific Biosciences

- Ion torrent / Ion proton
- Oxford Nanopore

Shotgun sequencing



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



Sequencing technologies

- Illumina:
  - χαμηλότερη ακρίβεια στην αναγνώριση βάσεων
- Solid:
  - πολλά reads δεν ταιριάζουν πουθενά στο γονιδίωμα!
- Roche 454 pyrosequencing
  - λάθη στον αριθμό των βάσεων εντός μιας περιοχής ομοπολυμερών (π.χ. AAAAAAAAAAAAAAA)
- Sanger:
  - χρειάζεται σχετικά μεγάλες ποσότητες DNA

Reads

- Sanger: μήκος: 1000-2000 bp
- 454: 450Mbp/run - μήκος: ~330bp
- Illumina: 18-35 Gbp/run - μήκος: ~75-100bp
- SOLID: 30-50 Gbp/run - μήκος: 50bp

Reviews στο Next Generation Sequencing

Anal Chem. 2011 Jun 15;83(12):4327-41. Epub 2011 May 25.

Landscape of next-generation sequencing technologies.

Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE.

Department of Chemical Engineering, Stanford University, Palo Alto, California, USA.

PMID: 21612267 [PubMed - indexed for MEDLINE] PMCID: PMC3437308 **Free PMC Article**

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

Nat Rev Genet. 2010 Jan;11(1):31-46. Epub 2009 Dec 8.

Sequencing technologies - the next generation.

Metzker ML.

Human Genome Sequencing Center and Department of Molecular & Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA. mmetzker@bcm.edu

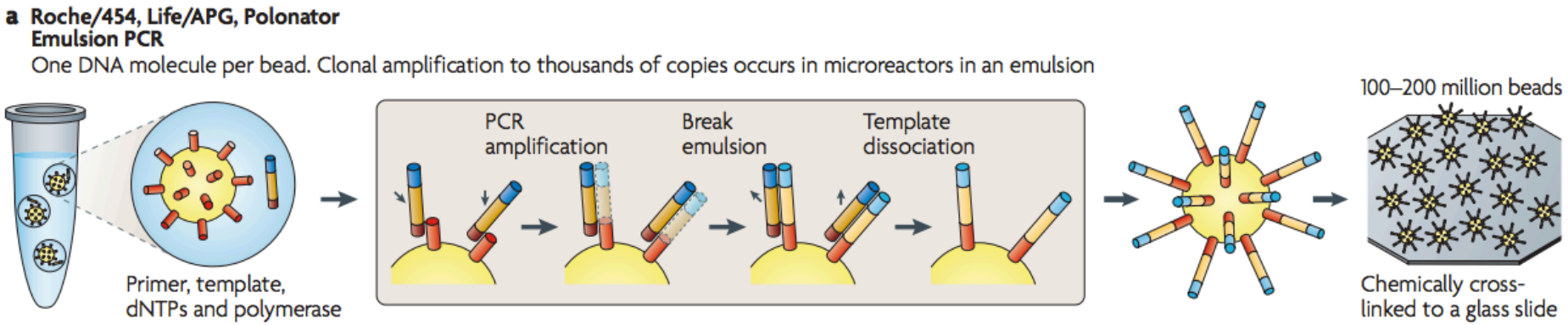
Abstract

Demand has never been greater for revolutionary technologies that deliver fast, inexpensive and accurate genome information. This challenge has catalysed the development of next-generation sequencing (NGS) technologies. The inexpensive production of large volumes of sequence data is the primary advantage over conventional methods. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly approaches, and recent advances in current and near-term commercially available NGS instruments. I also outline the broad range of applications for NGS technologies, in addition to providing guidelines for platform selection to address biological questions of interest.

PMID: 19997069 [PubMed - indexed for MEDLINE]

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

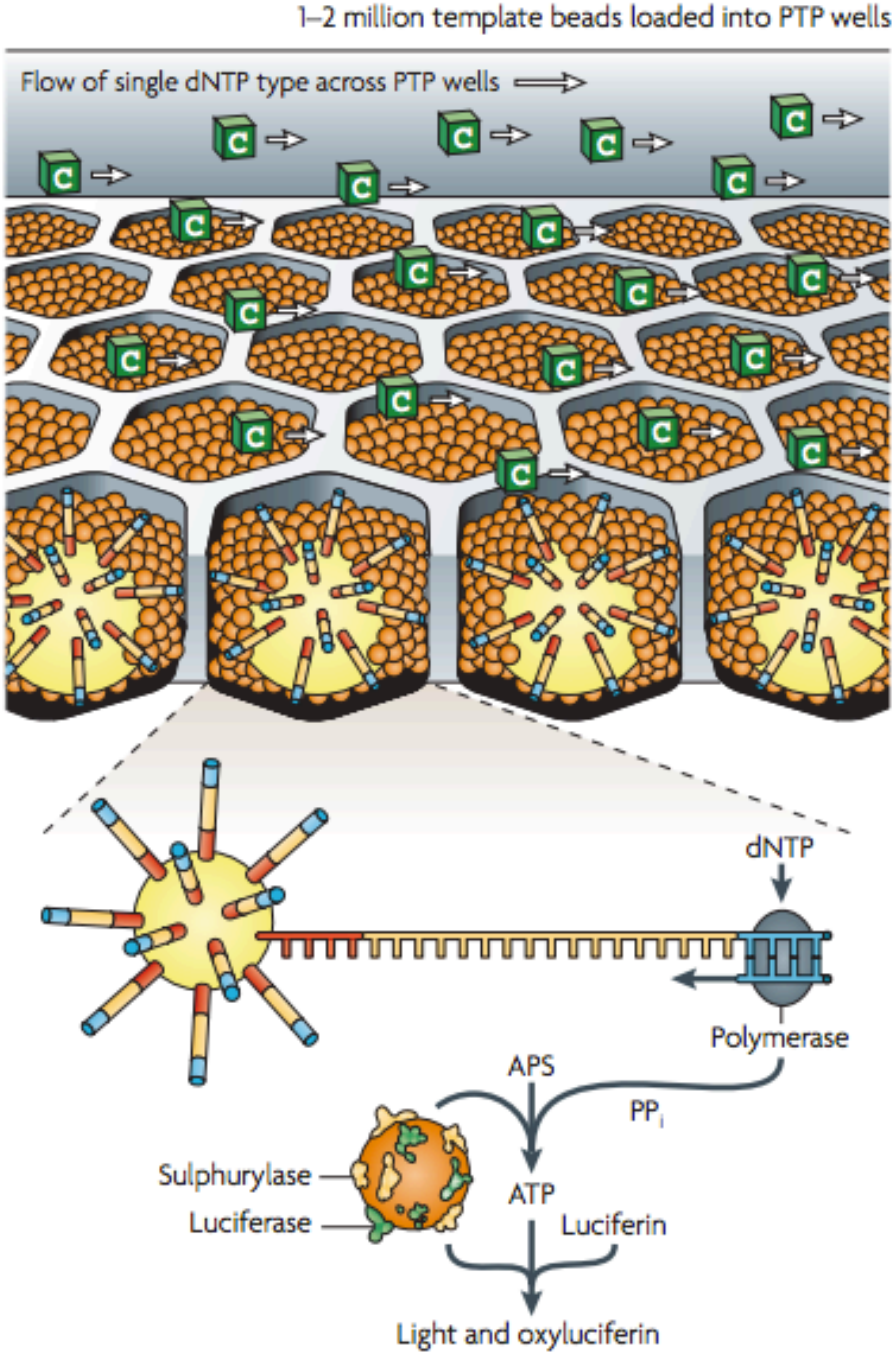
Pyrosequencing



http://www.youtube.com/watch?v=nFfgWGFe0aA

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

Roche/454 — Pyrosequencing



Pacific Biosciences

Analytical Chemistry REVIEW

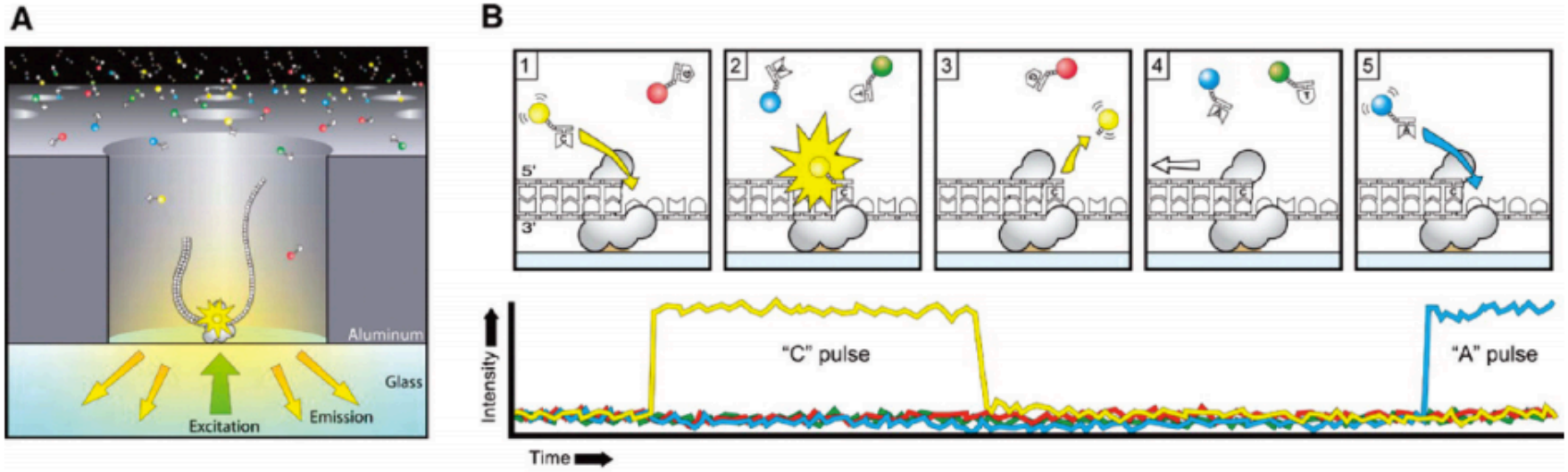


Figure 2. Schematic of PacBio's real-time single molecule sequencing. (A) The side view of a single ZMW nanostructure containing a single DNA polymerase (Φ29) bound to the bottom glass surface. The ZMW and the confocal imaging system allow fluorescence detection only at the bottom surface of each ZMW.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

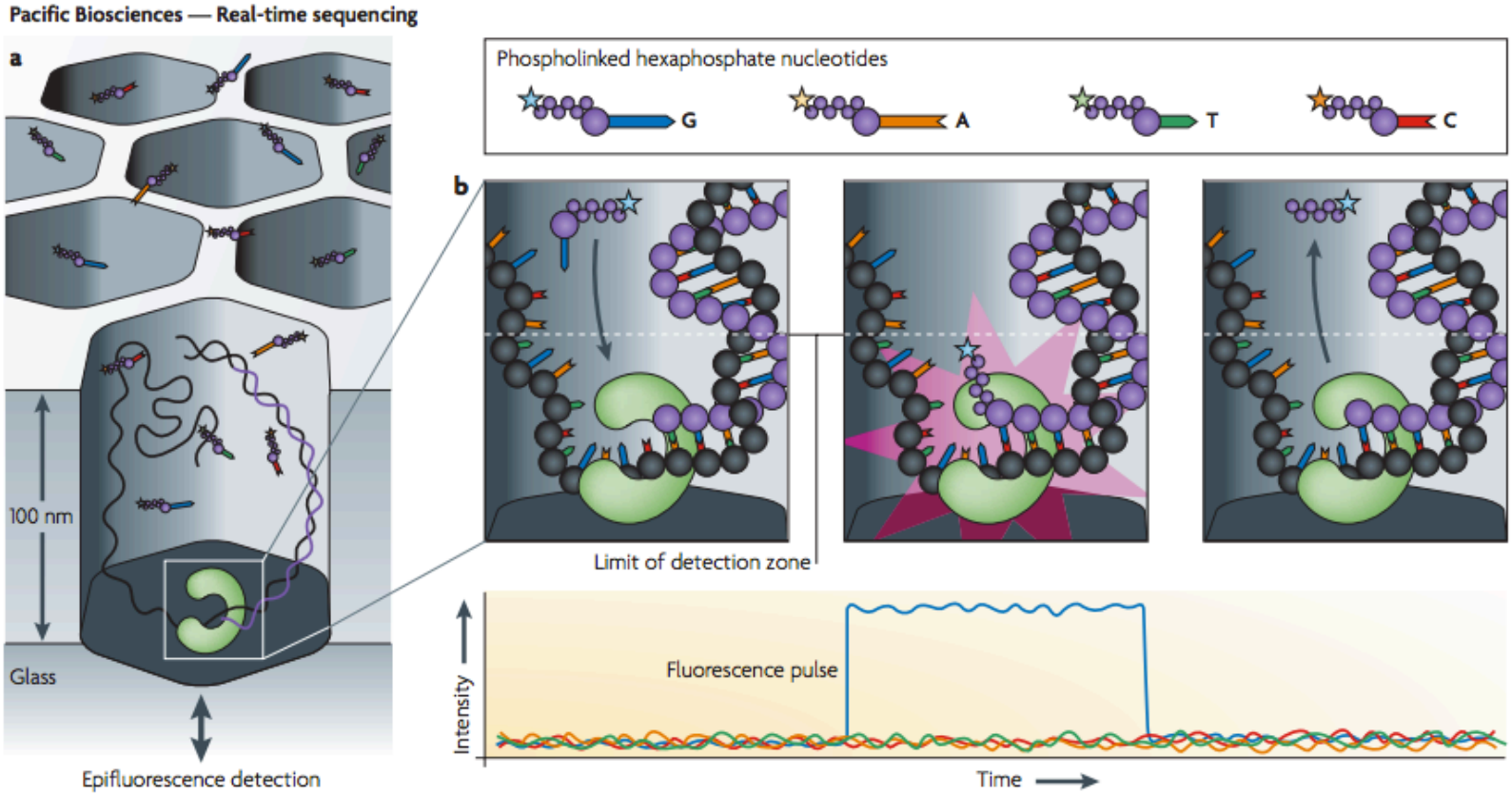
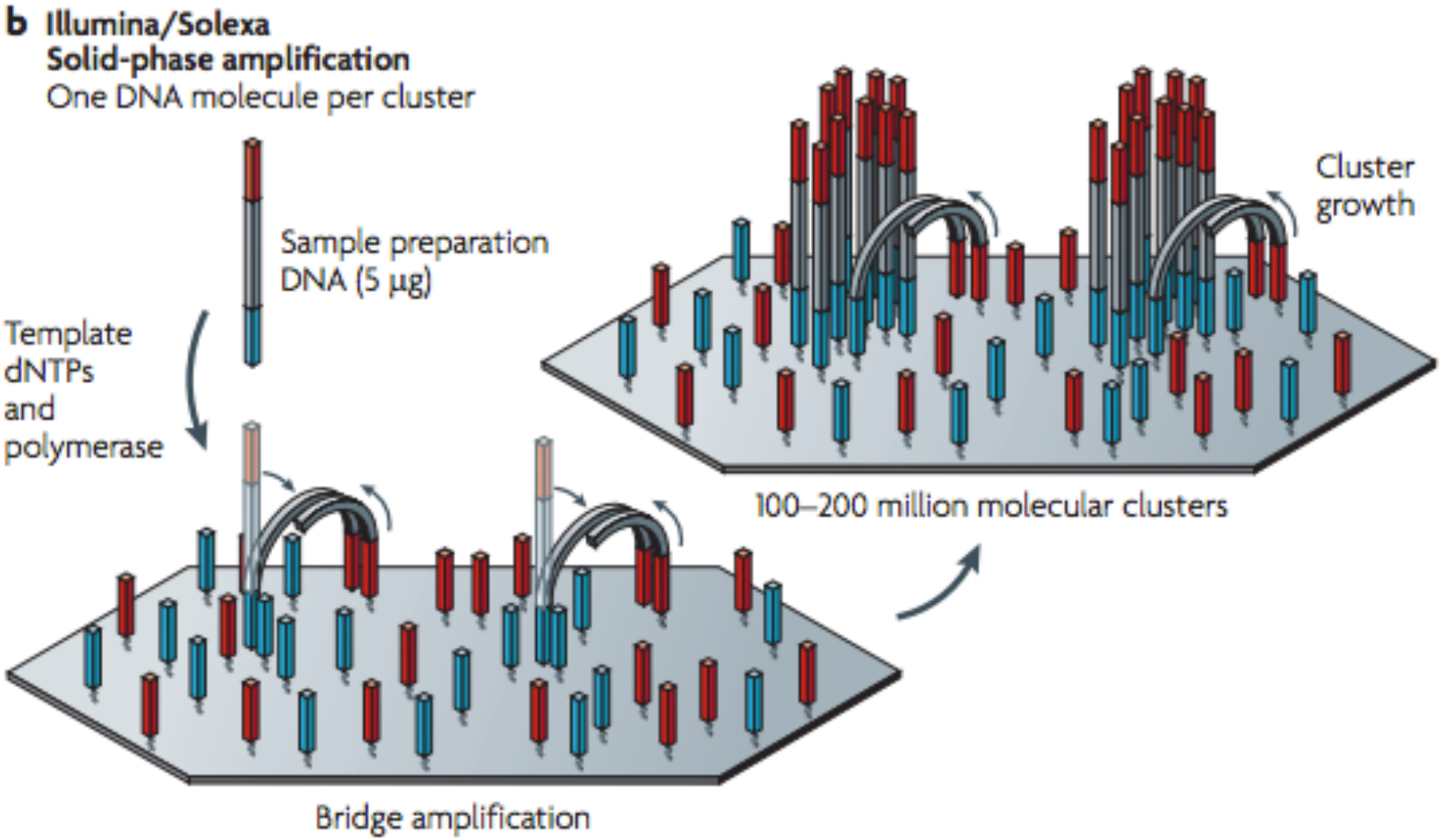


Figure 4 | **Real-time sequencing.** Pacific Biosciences' four-colour real-time sequencing method is shown. **a** | The zero-mode waveguide (ZMW) design reduces the observation volume, therefore reducing the number of stray fluorescently labelled molecules that enter the detection layer for a given period. 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.

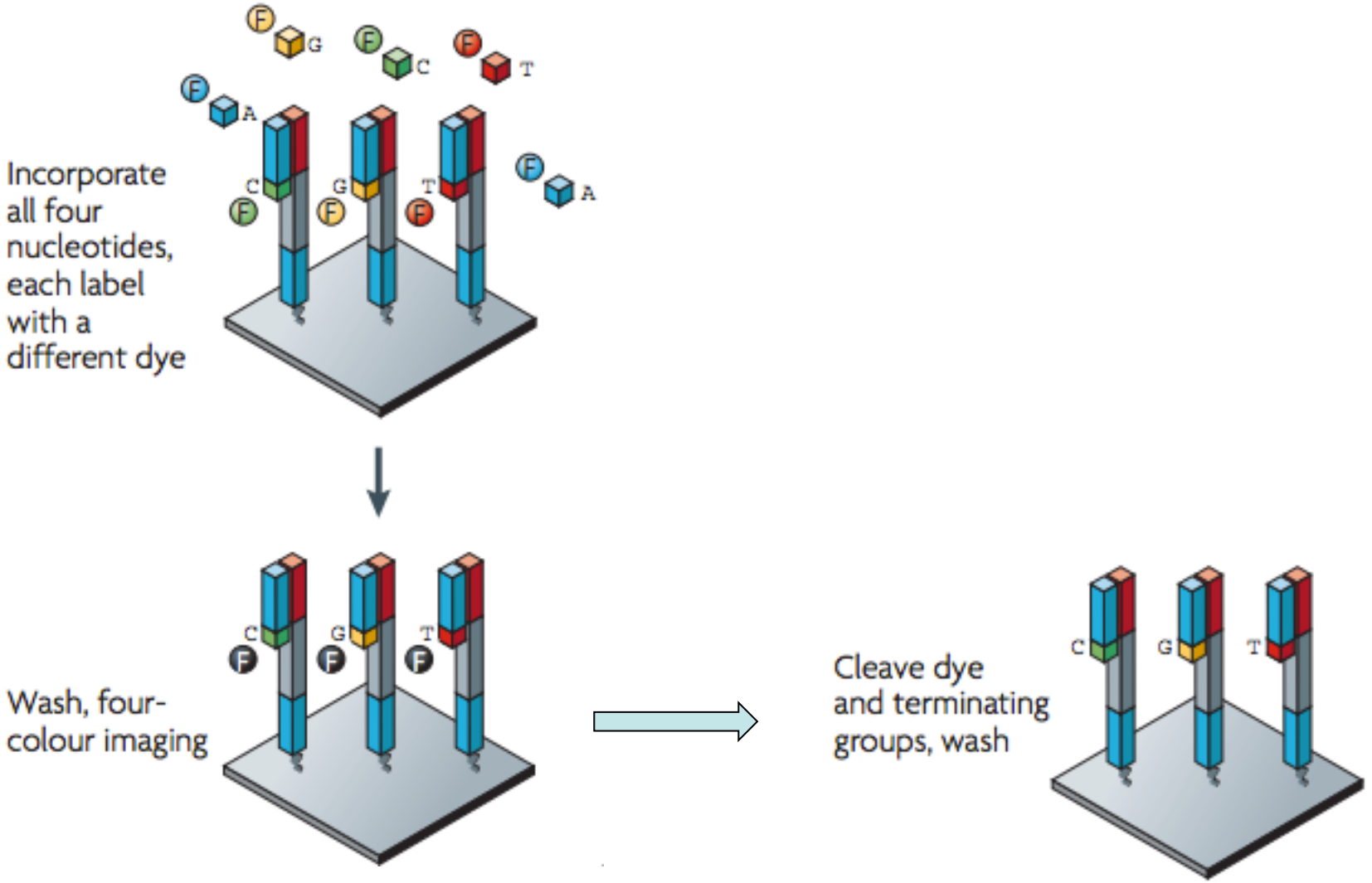
Illumina

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

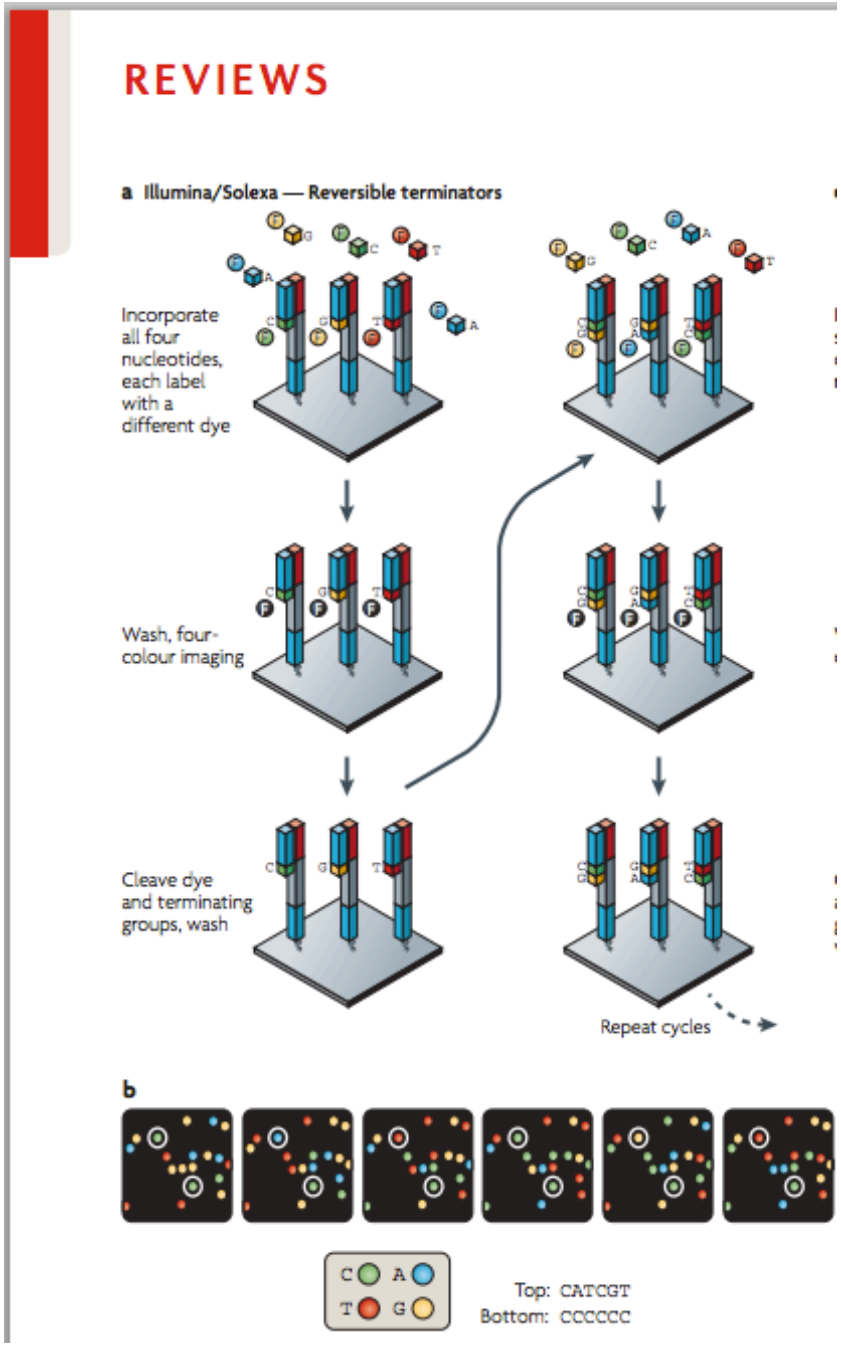
http://www.youtube.com/watch?v=l99aKKHcxC4



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



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



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

Table 1 | **Comparison of next-generation sequencing platforms**

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.

Ion Proton

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

Press Releases

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

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

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

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

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

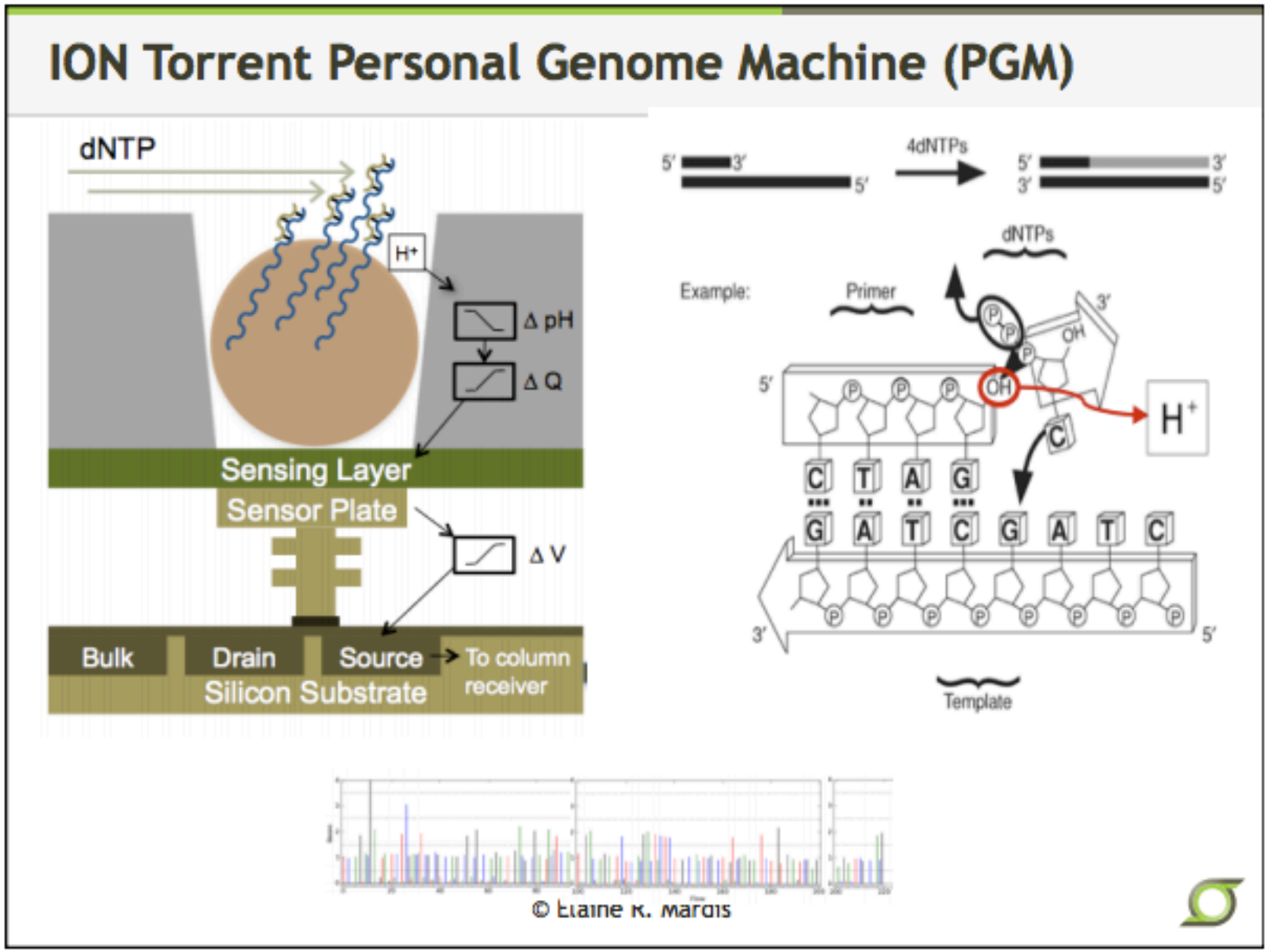
Up to now, it has taken weeks or months to sequence a human genome at a cost of $5,000 to $10,000 using optical-based sequencing technologies. The slow pace and the high instrument cost of $500,000 to $750,000 have limited human genome sequencing to relatively few research labs.

Ion Proton

Ion torrent chemistry

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

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

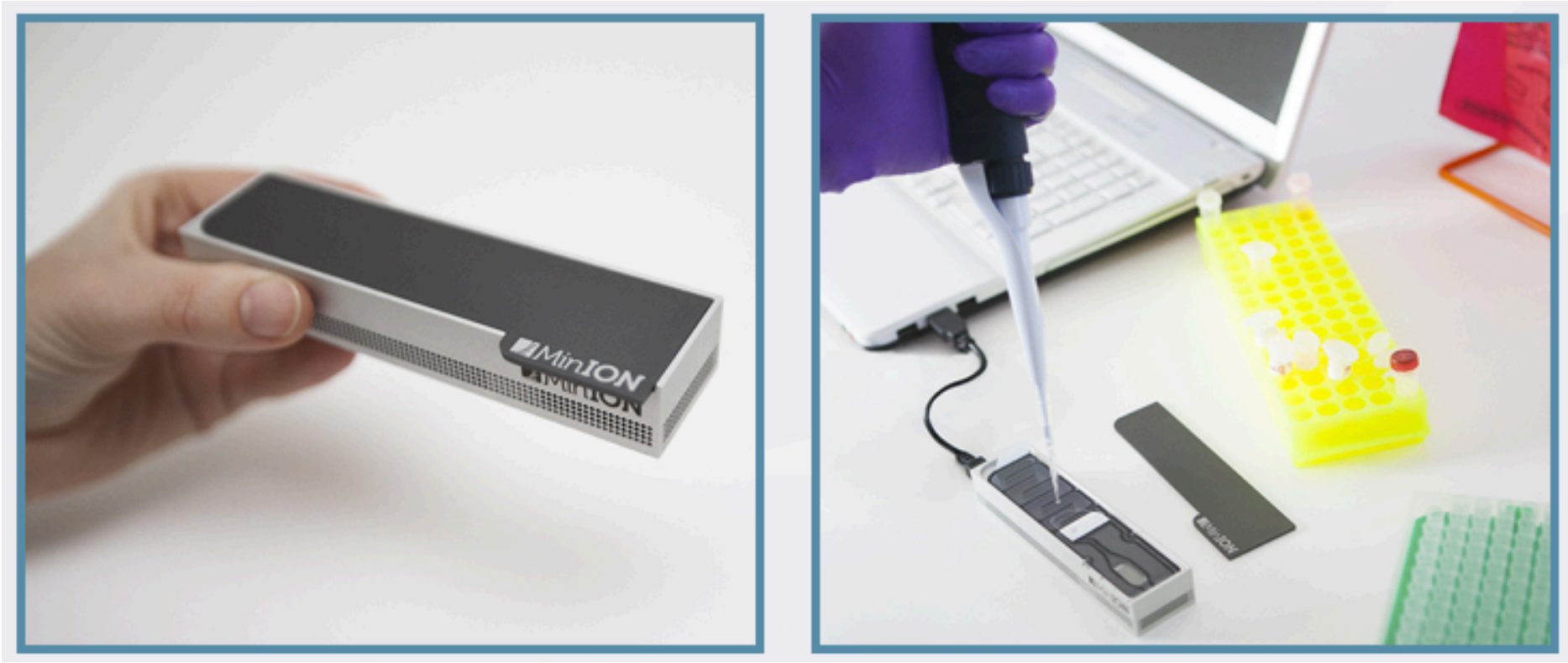


Εικόνα Από Elaine Mardis

Oxford Nanopore

(Στο εγγύς μέλλον;)

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



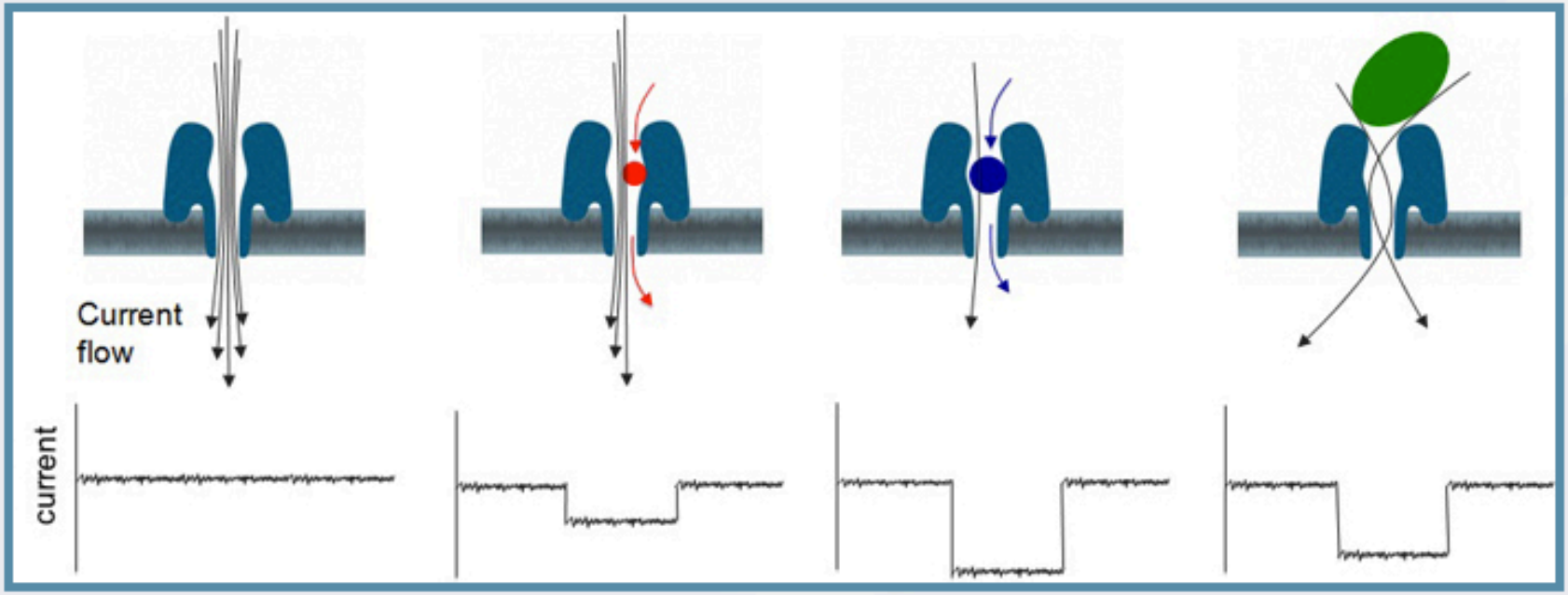
http://www.nanoporetech.com/technology/minion-a-miniaturised-sensing-instrument

Biological Nanopore

(Στο εγγύς μέλλον;)

Nanopore sensing

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



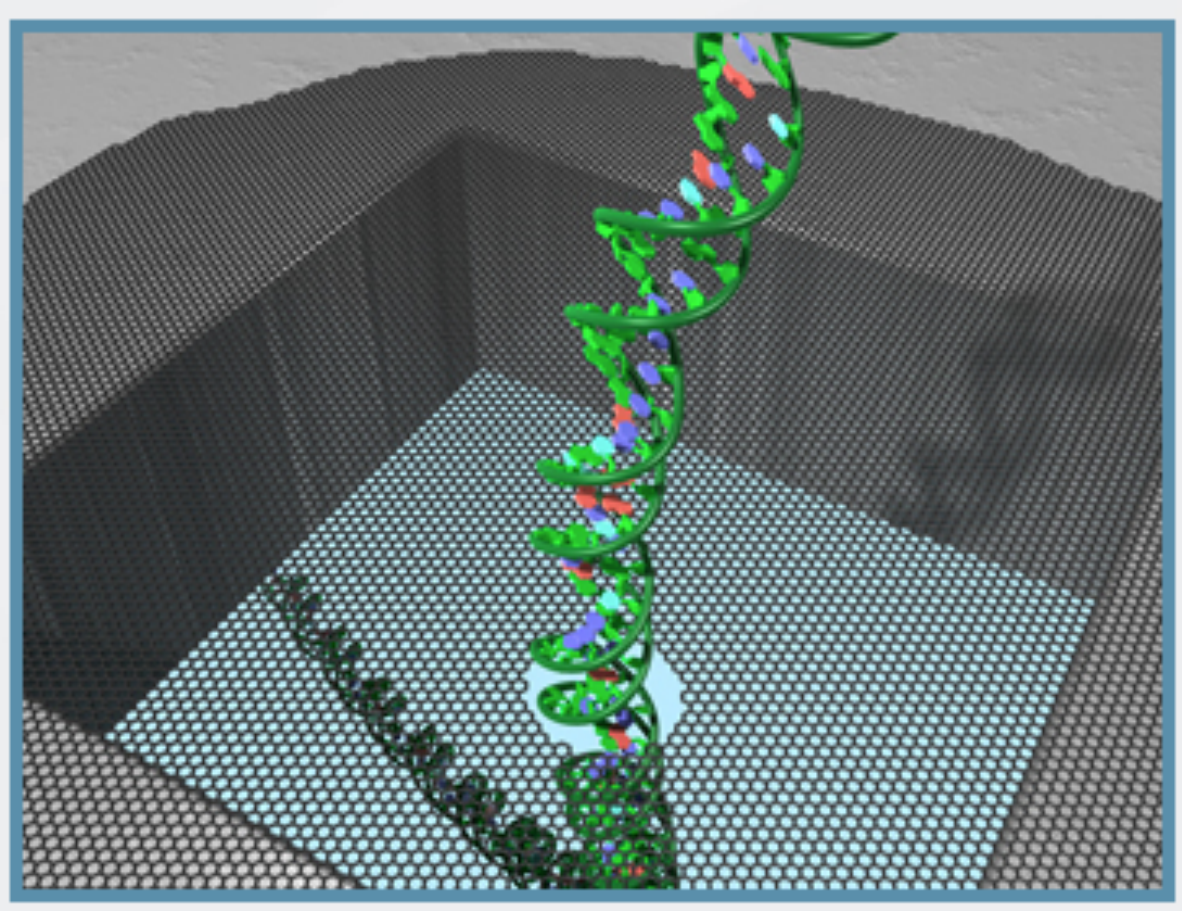
This diagram shows a protein nanopore set in an electrically resistant membrane bilayer. An ionic current is passed through the nanopore by setting a voltage across this membrane.

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

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

Solid state (Graphene) Nanopore

(Στο εγγύς μέλλον;)



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

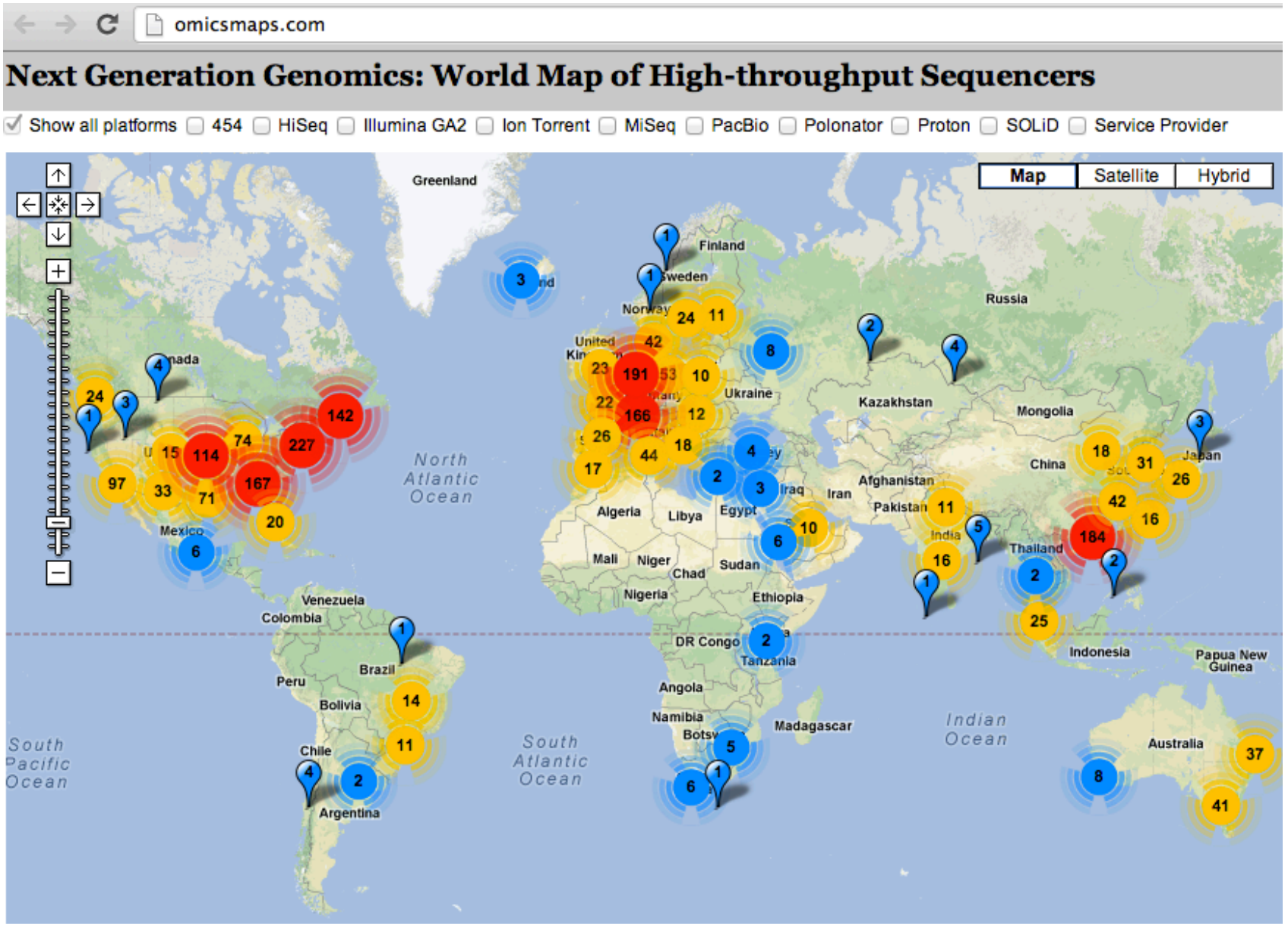
The sequence read archive: explosive growth of sequencing data

- http://nar.oxfordjournals.org/content/40/D1/D54.full

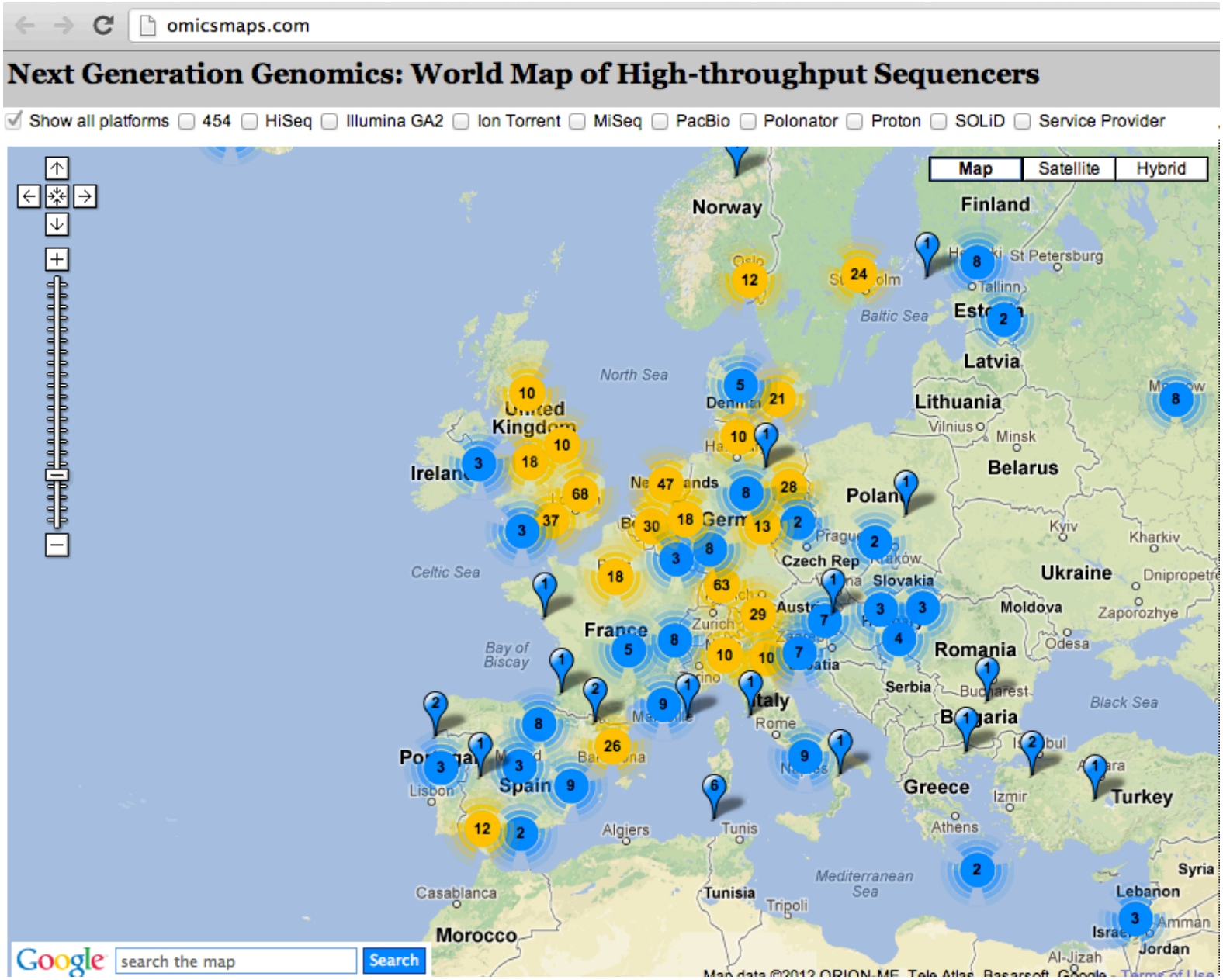
- Illumina™ platform comprises 84% of sequenced bases, with SOLiD™ and Roche/454™ platforms accounting for 12% and 2%, respectively.

- The most active SRA submitters in terms of submitted bases are the **Broad Institute**, the **Wellcome Trust Sanger Institute** and **Baylor College of Medicine** with 31, 13 and 11%, respectively. The largest individual global project generating next-generation sequence is the 1000 Genomes project which has contributed nearly one third of all bases.

http://omicsmaps.com/



http://omicsmaps.com/

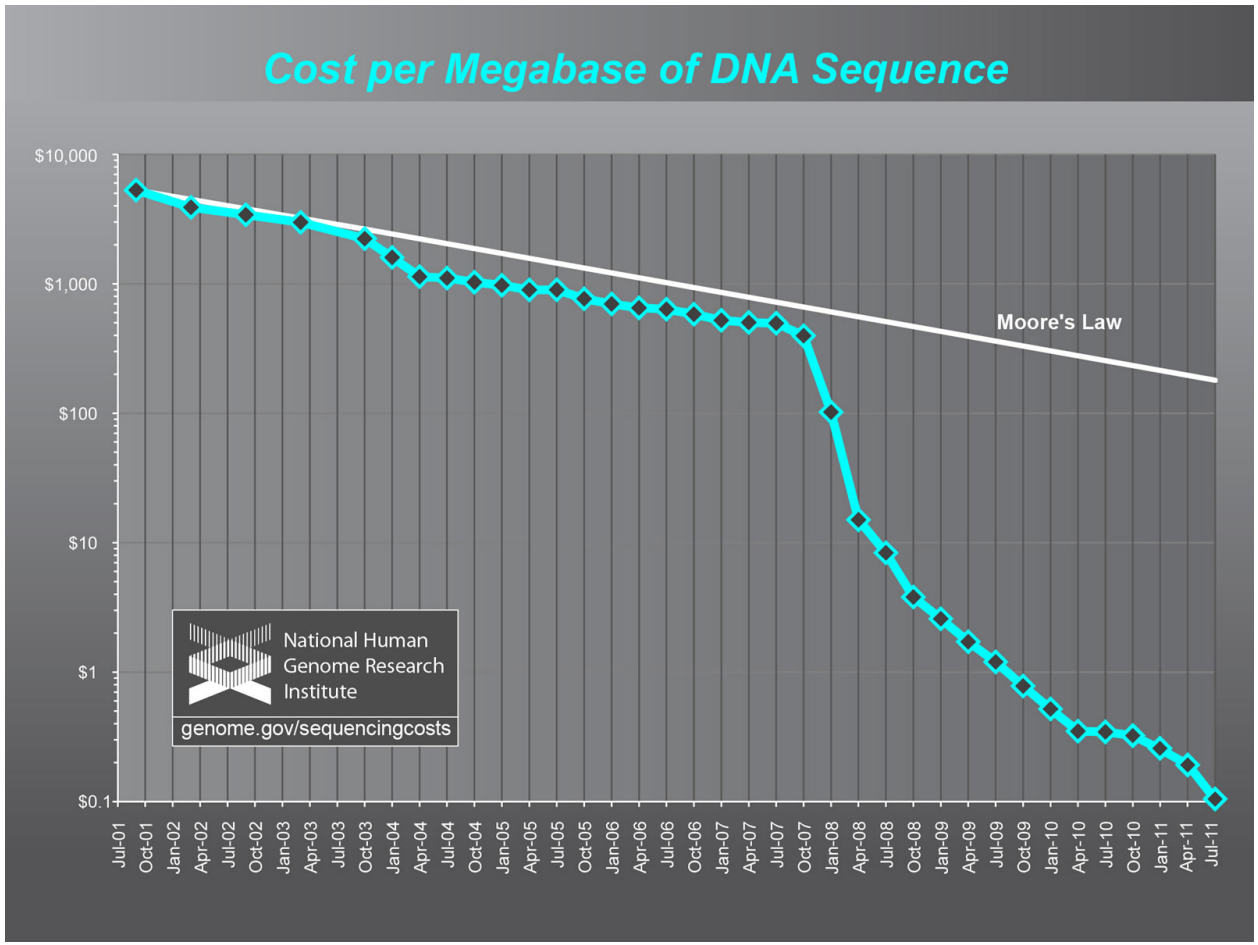


Χαμηλό κόστος γενωμικών τεχνολογιών θα οδηγήσει σε καθημερινές εφαρμογές.

- Κόστος αλληλούχισης πέφτει διαρκώς.
  - Illumina -> 1 lane: 19GBp, ~ €3000, 10 βακτηριακά γενώματα.
- Τα δείγματα αποστέλλονται σε κέντρα με μεγάλες εγκαταστάσεις και χαμηλό κόστος λειτουργίας (οικονομία κλίμακας). Η ανάλυση των δεδομένων όμως δεν υπόκειται σε όρους οικονομίας κλίμακας.
- Πλέον, ένα σημαντικό μέρος του ολικού κόστους είναι η βιοπληροφορική ανάλυση.
- Μηχανήματα αλληλούχισης ακριβά (Illumina ~ €600.000) - service φτηνό.
- Μισθός ακριβός (ίσως ένα νέο μοντέλο συμβουλευτικής?)
- Υπολογιστής φτηνός (€3-5.000), εφόσον πρόκειται για μικρά γονιδιώματα (de novo assembly), ή για re-sequencing.

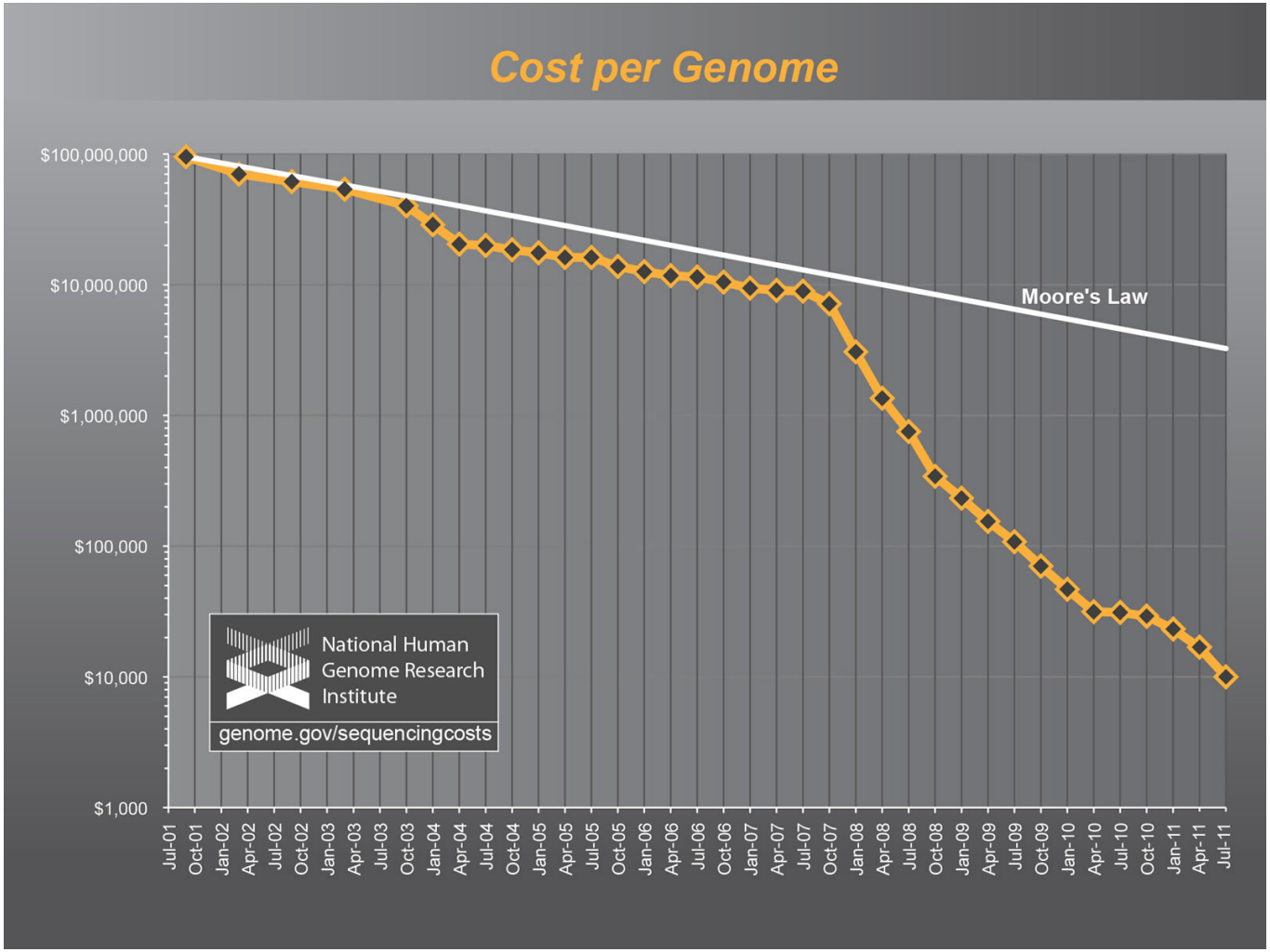
Χαμηλό κόστος γενωμικών τεχνολογιών θα οδηγήσει σε καθημερινές εφαρμογές

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



Χαμηλό κόστος γενωμικών τεχνολογιών θα οδηγήσει σε καθημερινές εφαρμογές

- Κόστος αλληλούχισης
  - http://www.genome.gov/sequencingcosts/



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]

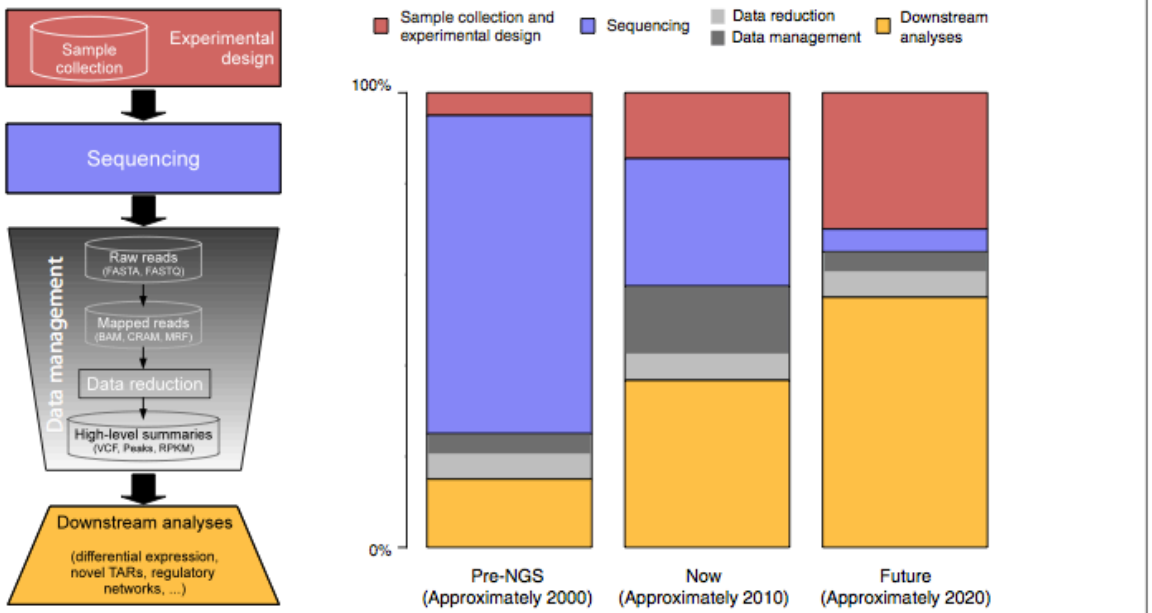
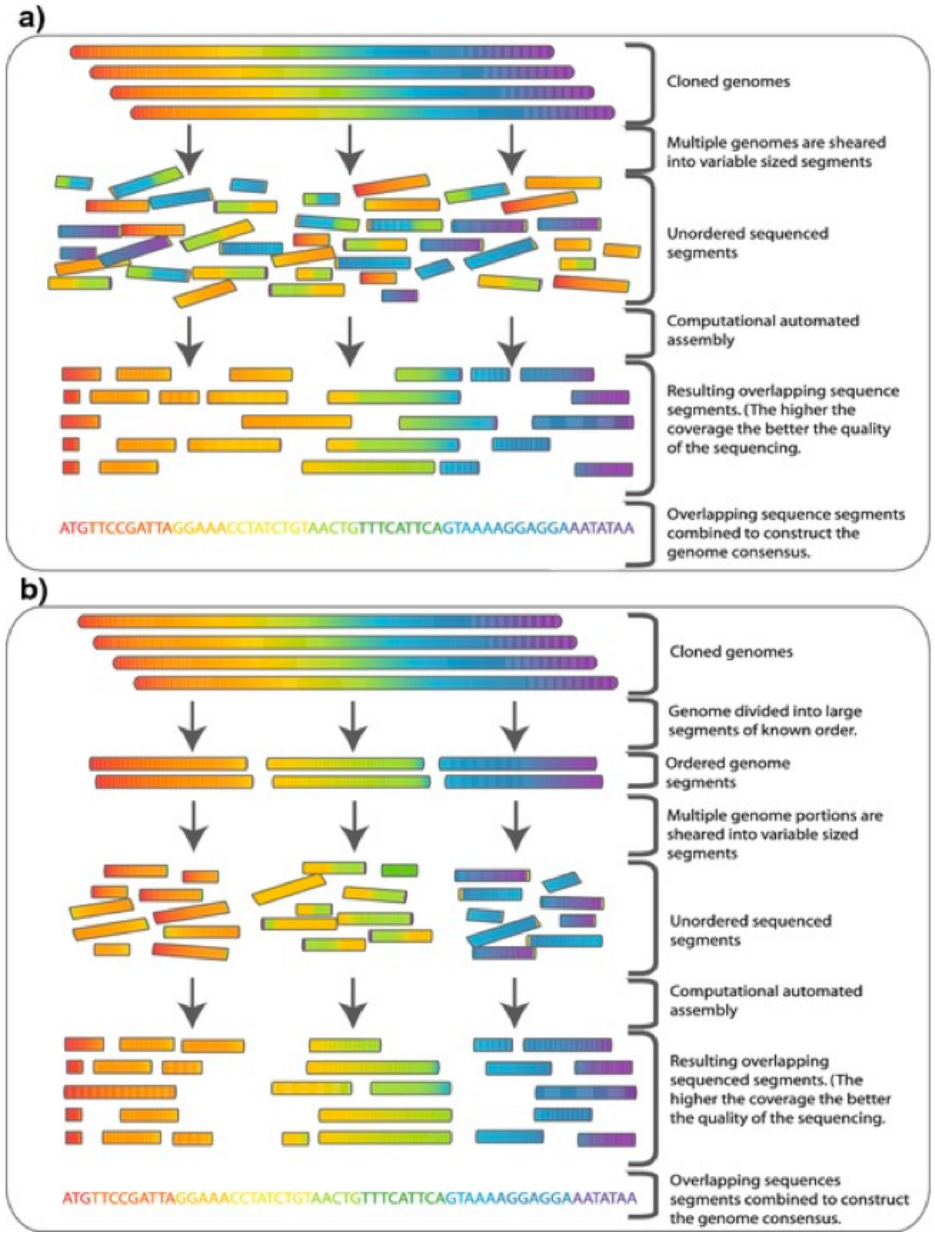


Figure 1. Contribution of different factors to the overall cost of a sequencing project across time. Left, the four-step process: (i) experimental design and sample collection, (ii) sequencing, (iii) data reduction and management, and (iv) downstream analysis. Right, the changes over time of relative 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



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

Ποιότητα των 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

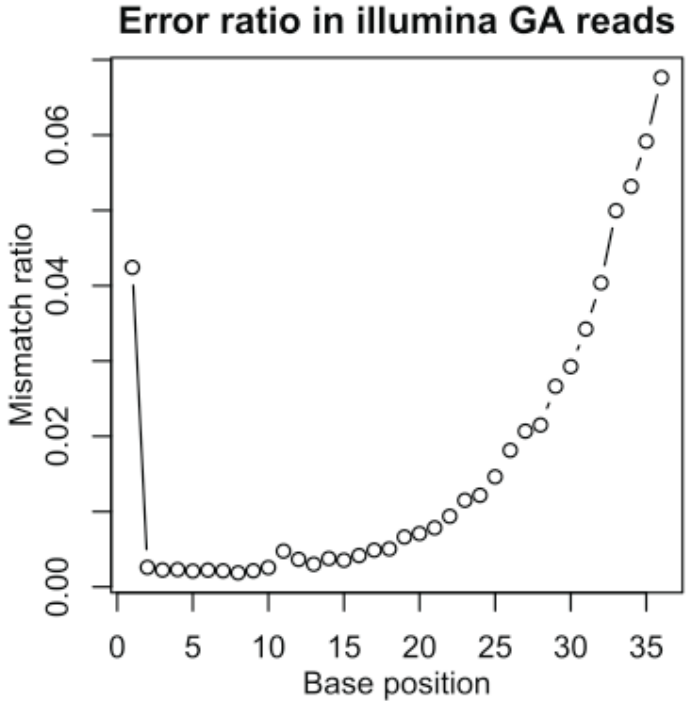


Figure 1. Error ratio in GA reads depending on the base position of the read. Ratio of mismatch between mapped reads and reference sequence to the total number of mapped reads was plotted against base position in the reads. The mismatch ratio increases along with the base position indicating decrease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

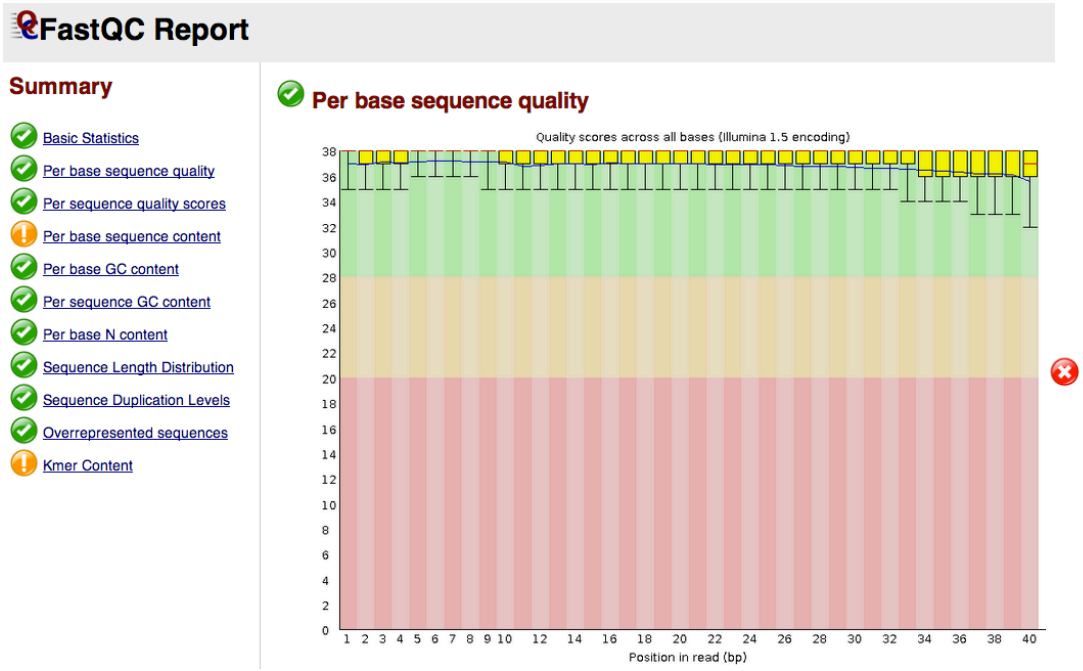
SOLID: ~50% reads δεν στοιχίζονται στο γονιδίωμα, από το οποίο έγινε το Sequencing!

Πρόβλημα στις χημικές αντιδράσεις μάλλον.

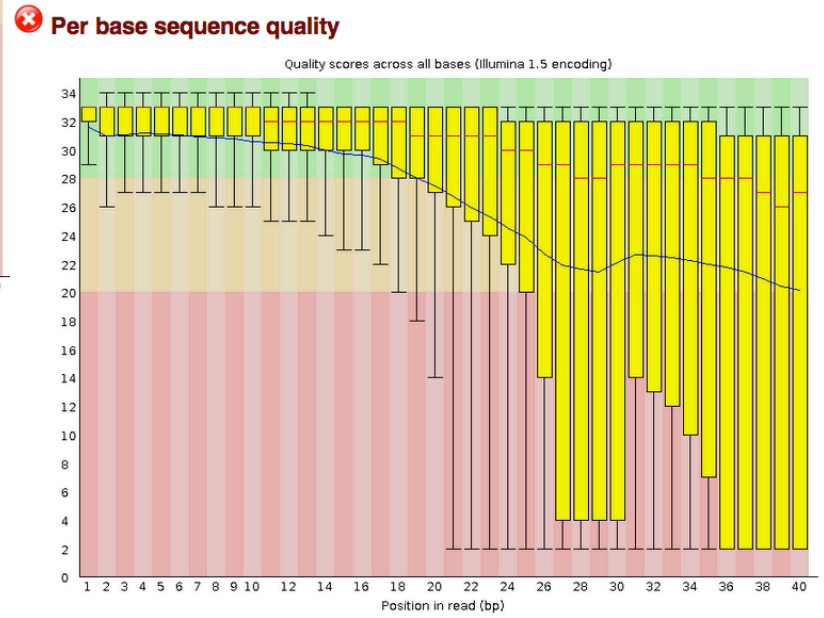
Εδώ, το πρόβλημα εντοπίζεται στην συσσώρευση λαθών κατά την ενσωμάτωση φθοριζόντων dNTPs.

Sequence reads – Έλεγχος ποιότητας δεδομένων (quality control)

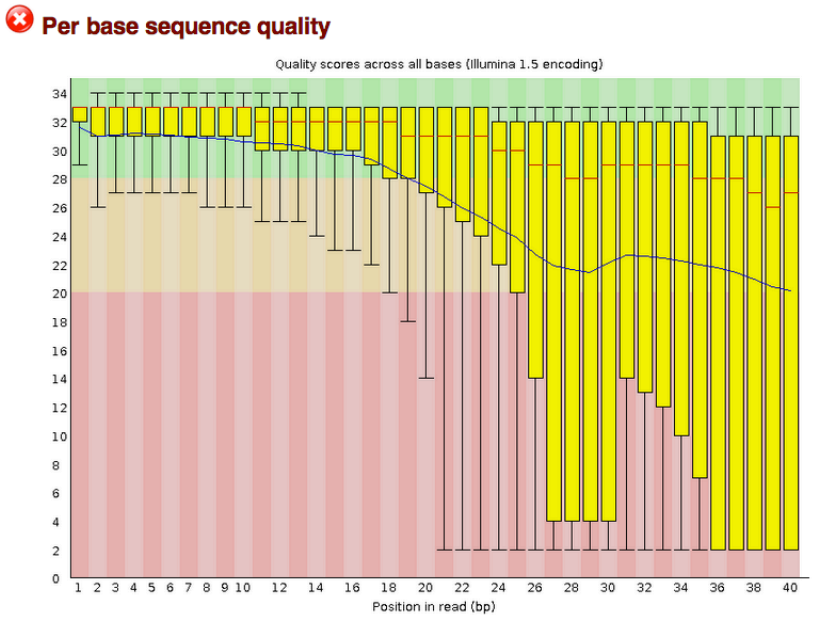
Πολύ υψηλής ποιότητας δεδομένα.



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.

Road map

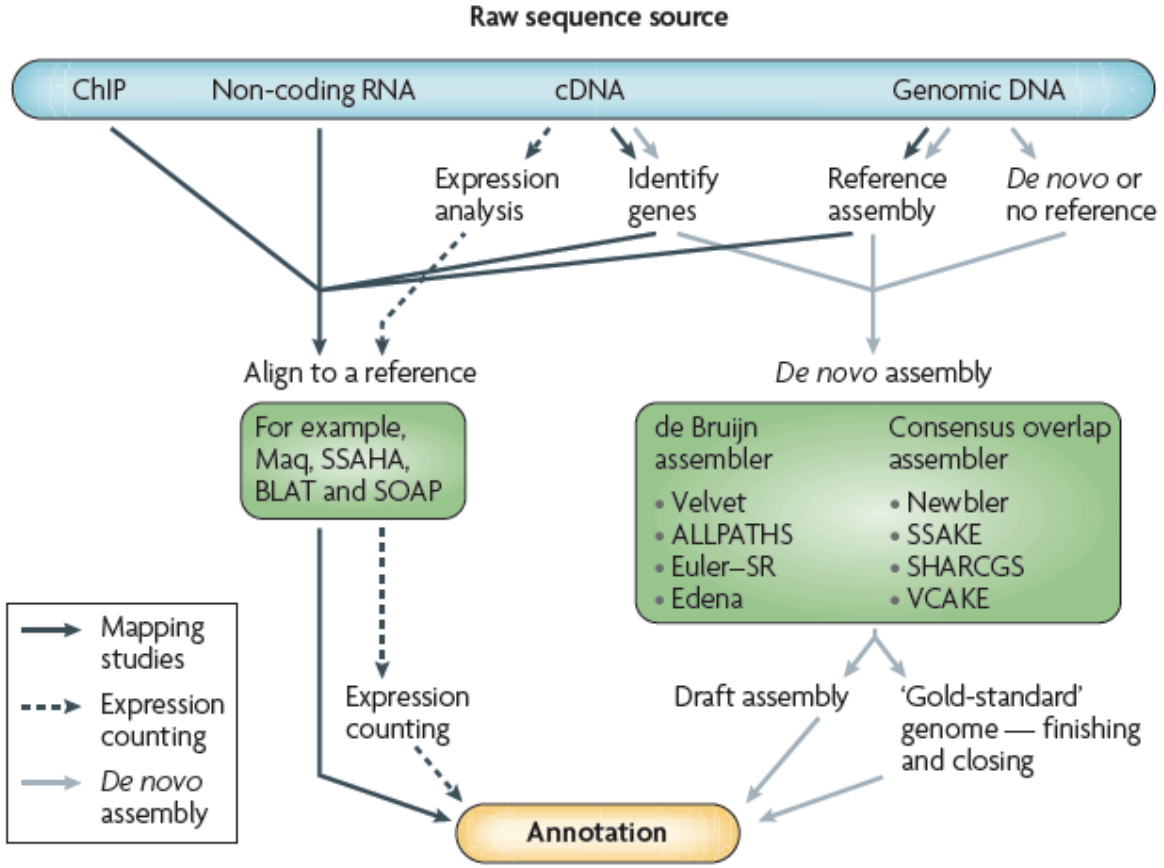
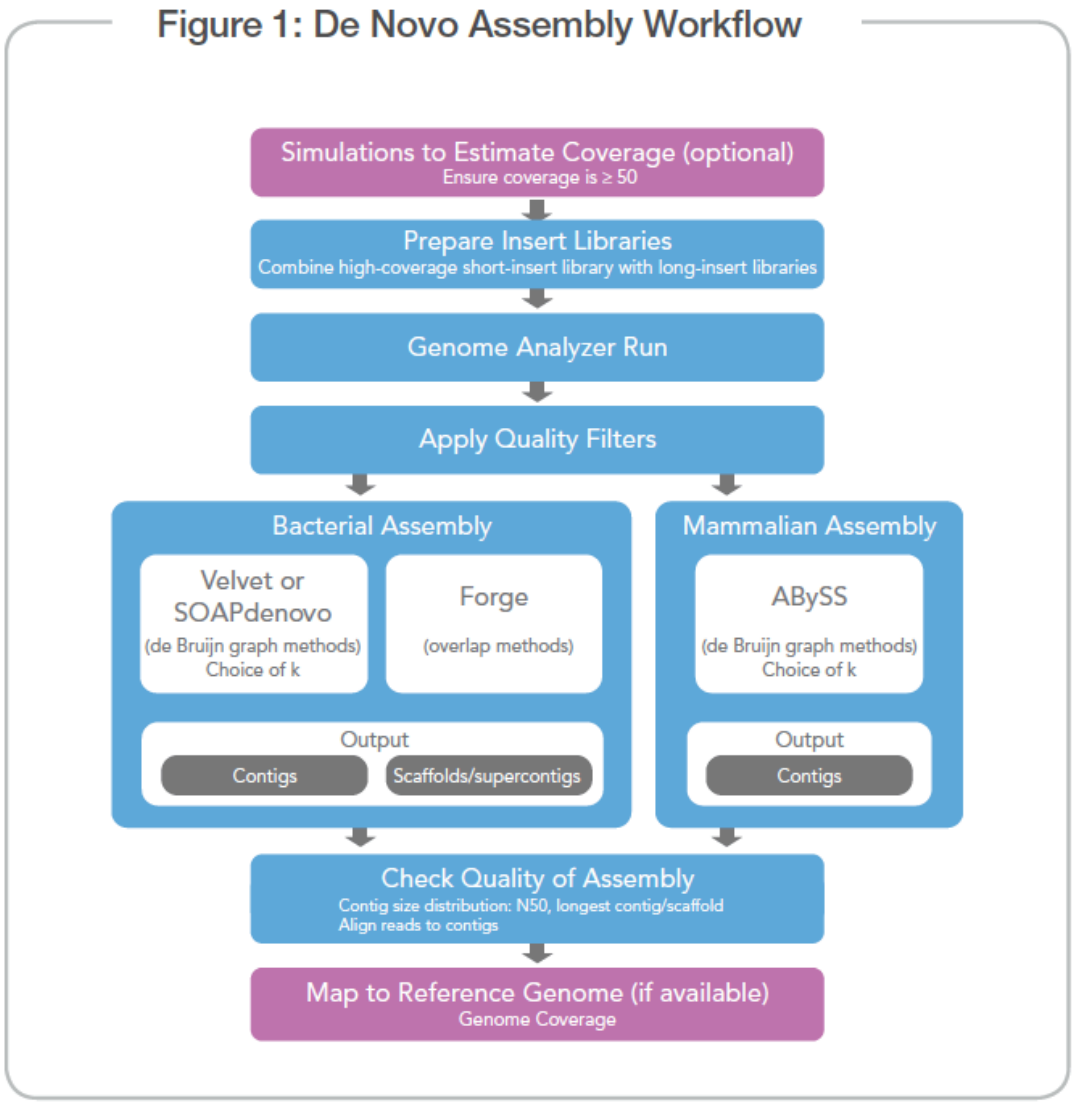


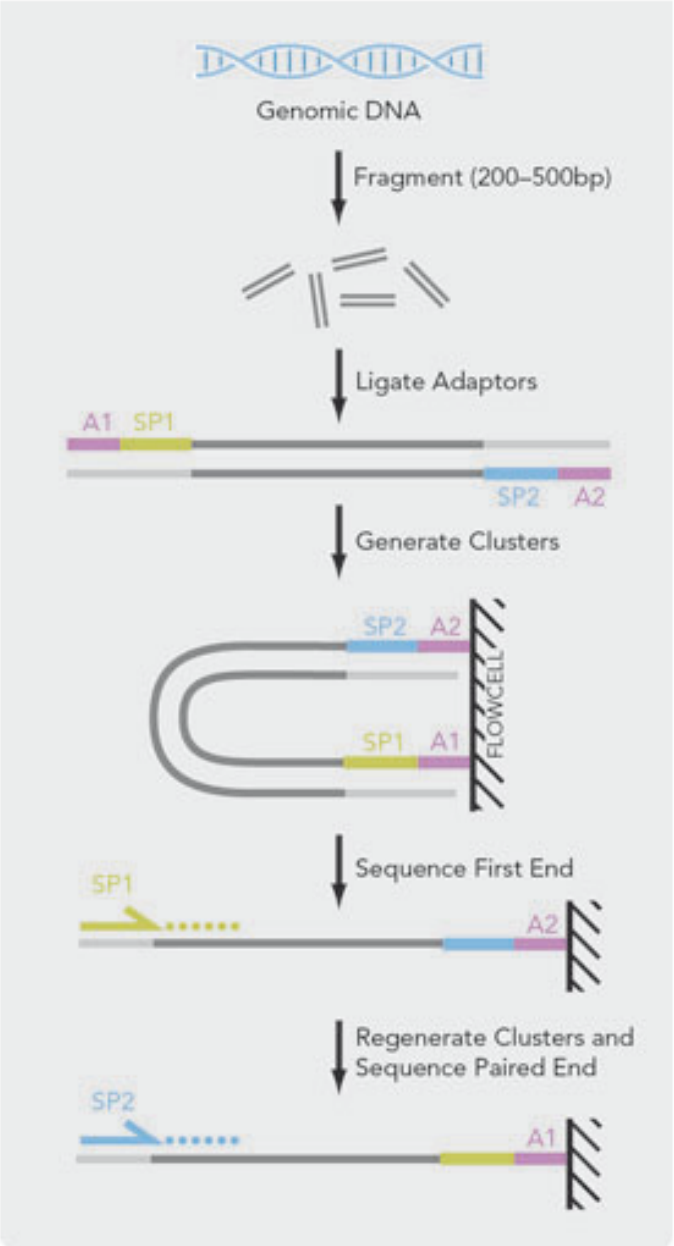
Figure 3 | **Road map for planning software solutions for experiments with different data sources and different goals.** Sequence reads derived from ChIP, non-coding RNA and cDNA sequencing experiments are aligned to a reference sequence before expression counting and final annotation. Sometimes, a cDNA sequence can be assembled *de novo* before these steps. Genome sequence reads may be aligned if a reference is available, but if not assembly *de novo* can still be carried out.

De novo Sequencing assembly workflow

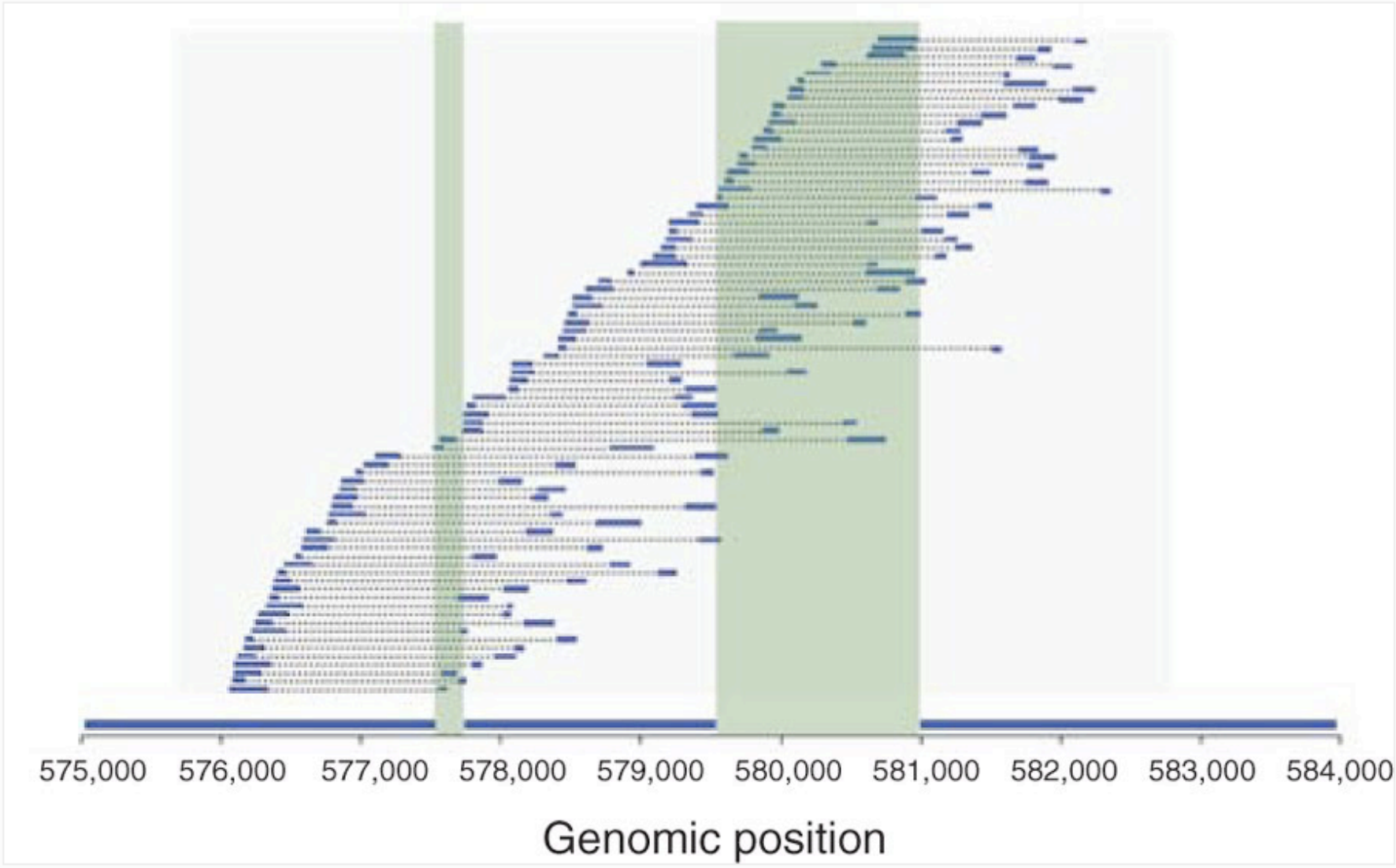


Sequencing

- Single end reads
- Paired end reads



Sequencing - paired end reads

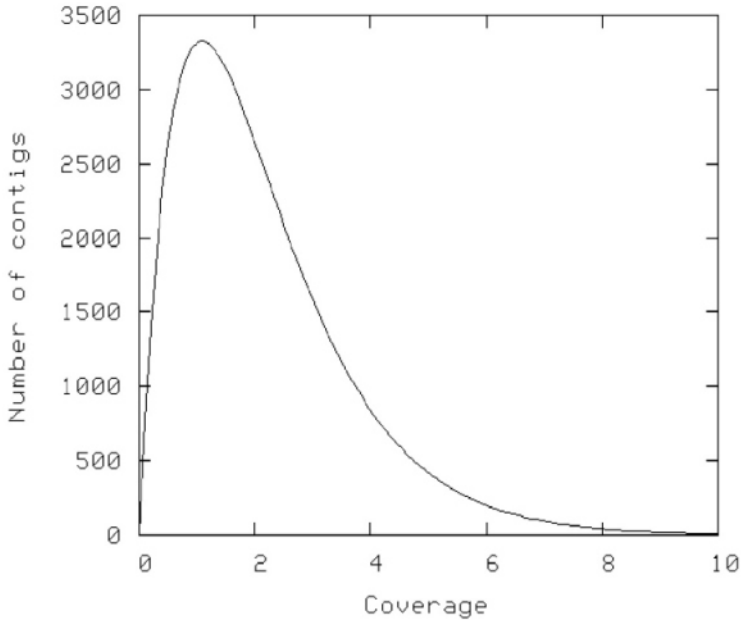


A region of the *de novo* assembly of *E. coli* K-12, with the *de novo*–assembled contigs covering the region shown in blue along the bottom axis. The paired-end reads generated with this protocol are capable of bridging the 0.2-kb and 1.5-kb gaps between the contigs, highlighted in green.

Lander - Waterman

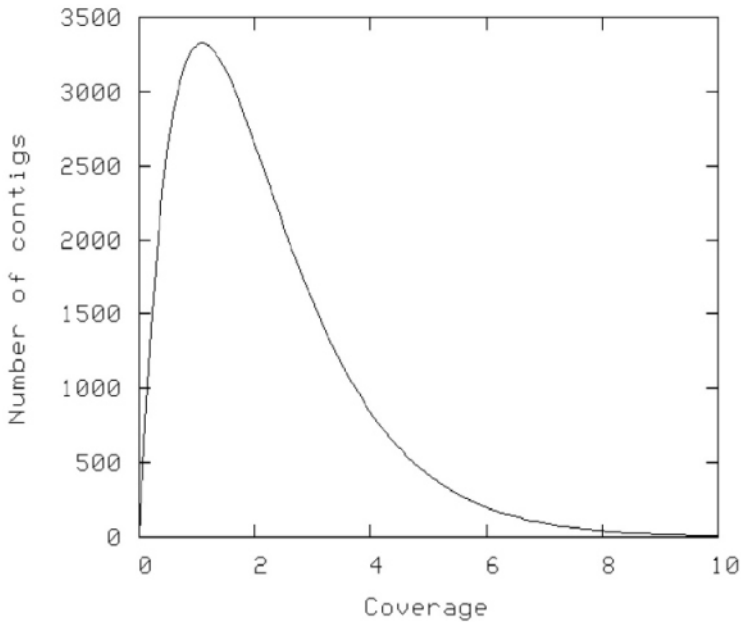
- Πόσο sequencing coverage απαιτείται για να μπορεί να συναρμολογηθεί ένα γονιδίωμα?
- Τουλάχιστον 8-10X

- 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).

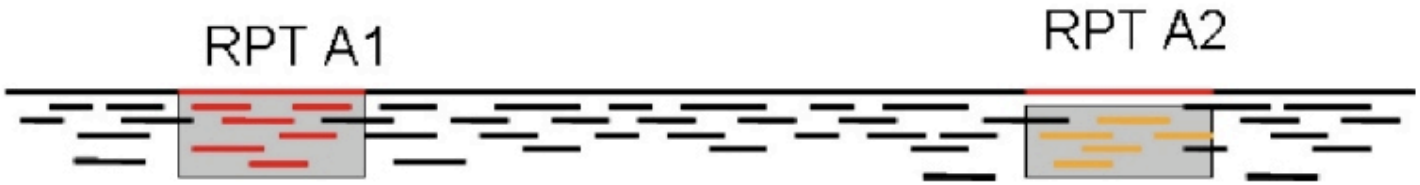


Figure 5. Two copies of a repeat along a genome. The reads colored in red and those colored in yellow appear identical to the assembly program.

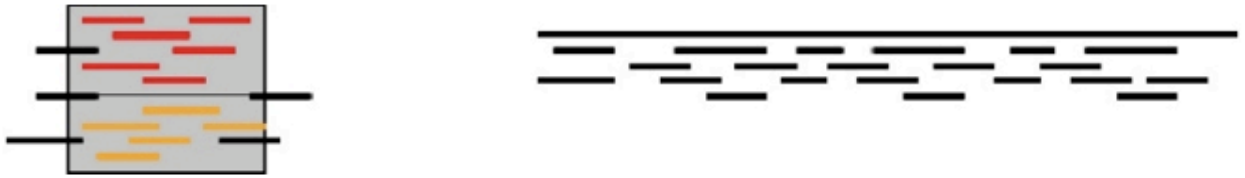


Figure 6. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs

Προβλήματα συναρμολόγισης 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

Κενά μετά την συναρμολόγιση

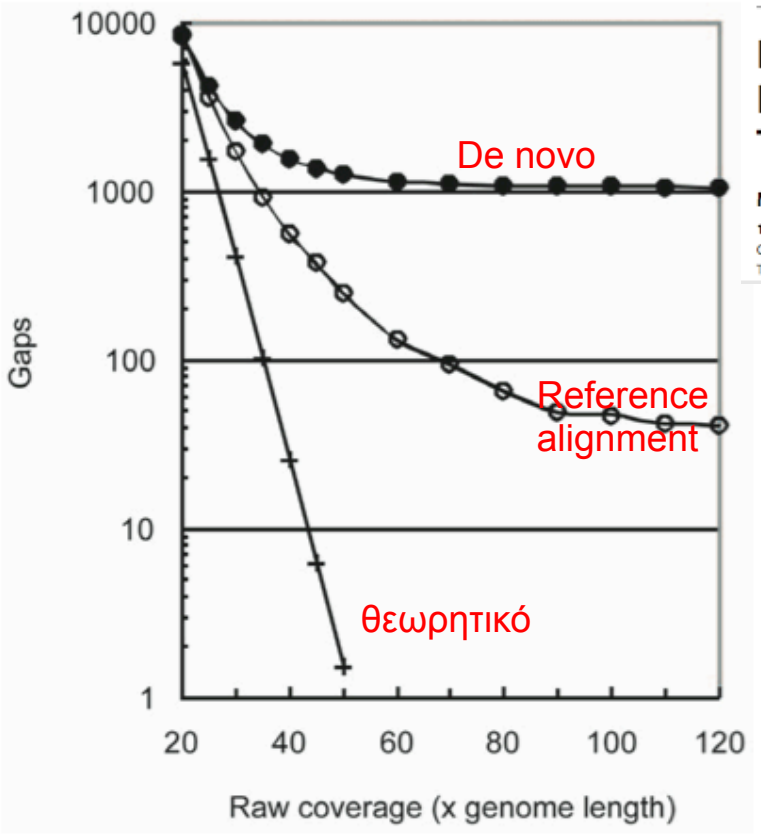


Figure 1. Assessing the cause of gaps in an assembly of 36nt reads. The predicted number of sequence gaps based on the Lander-Waterman model (+) is presented along with the actual number of sequence gaps in sets of 36nt Illumina reads (○). This was determined by aligning the reads in each set to the 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

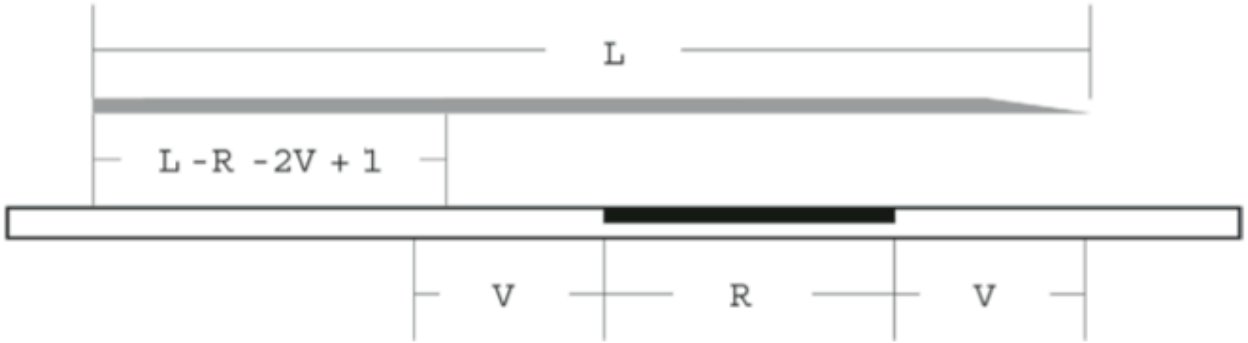


Figure 2. A model of repeat assembly. To unambiguously assemble a repeat (black rectangle), a read must encompass the entirety of the repeat and extend, in both directions, into unique sequence. If the repeat has a length of R nt, and the adjacent unique sequence must be at least V nt, then resolution 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

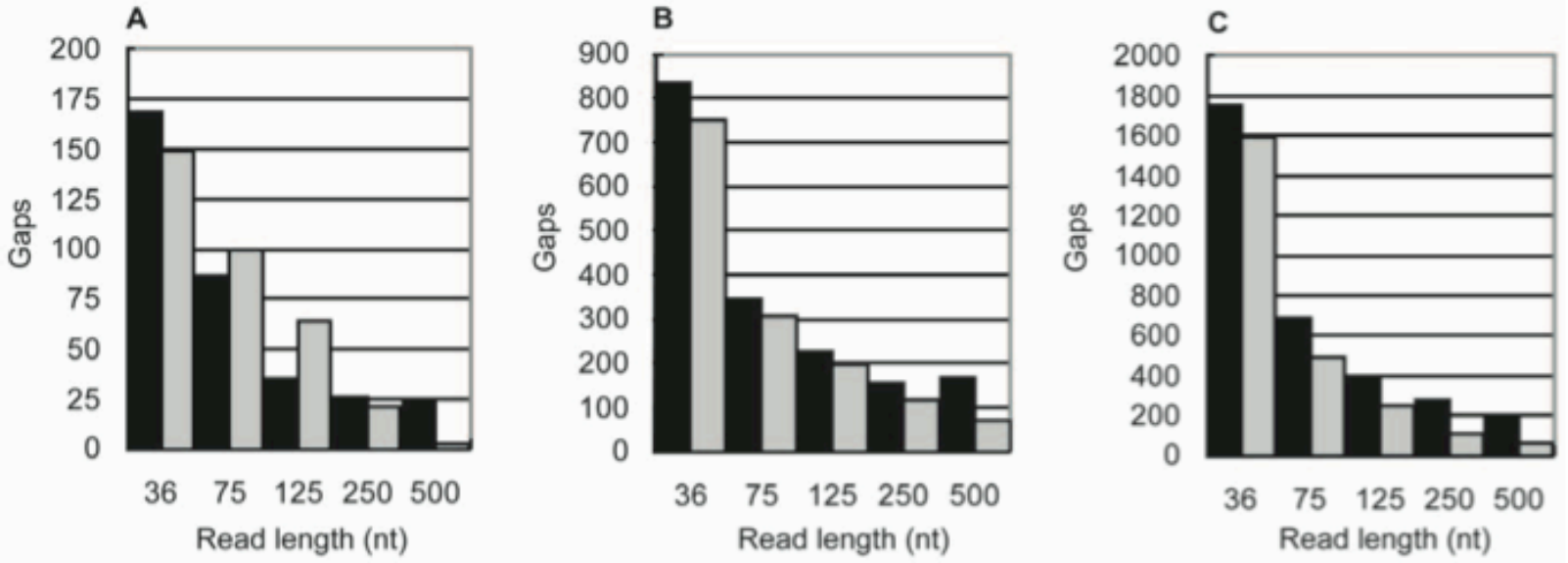


Figure 3. Assessing the accuracy of the algorithm. The number of repeat-induced gaps predicted by the algorithm (grey bars) compared to the number of gaps observed (black bars) in actual assemblies of 36, 75, 125, 250, and 500nt simulated reads from **A**) *M. genitalium*, **B**) *E. coli* and **C**) *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

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

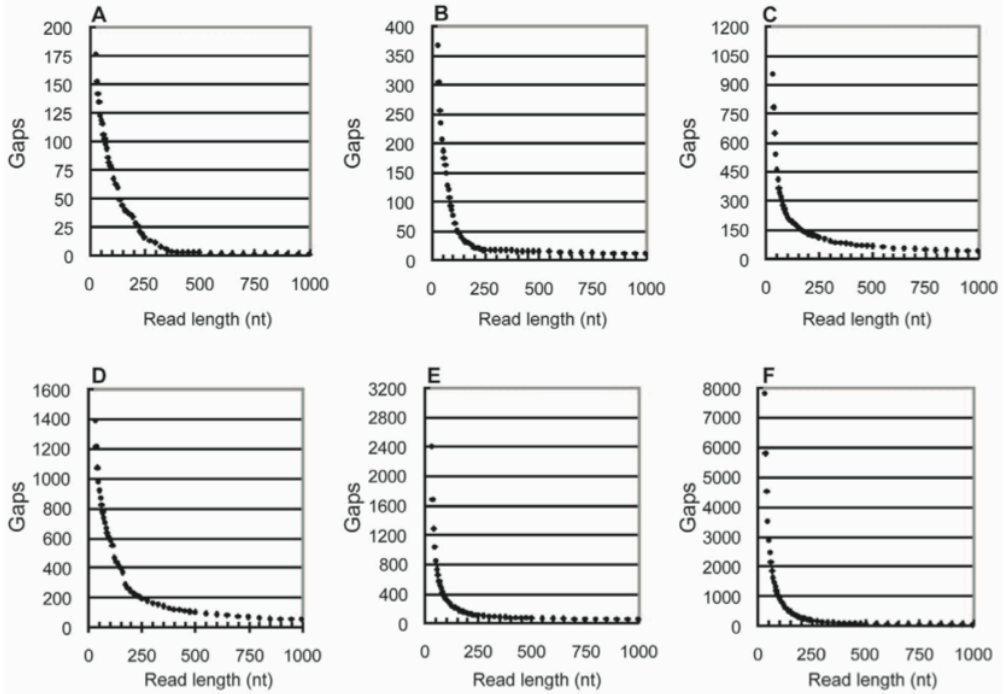


Figure 5. Read length and repeat resolution in 6 genomes. The algorithm was used to predict the occurrence of repeat-induced gaps in assemblies of six bacterial genomes from a range of read lengths. A raw coverage of 100× was used for all genome/read length pairings. Assembly results were predicted for 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 τον οργανισμό

36nt reads

125nt reads

500nt reads

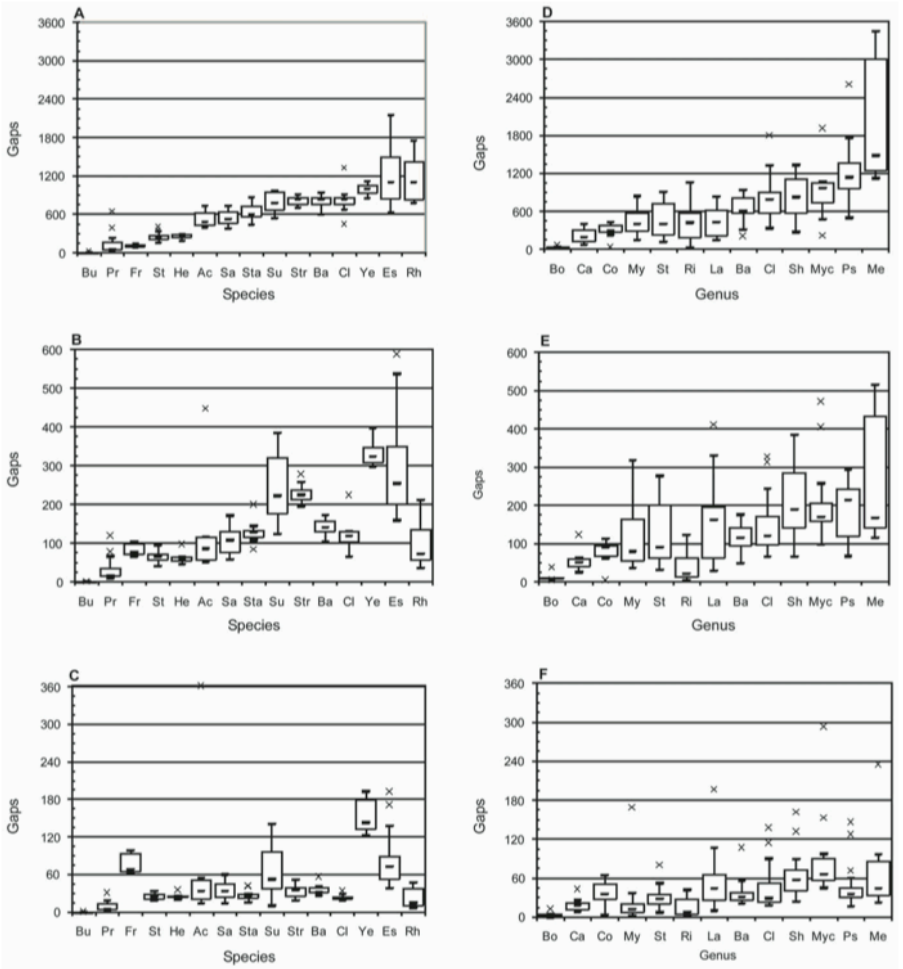


Figure 6. Variation in assembly results within taxa. The median number of repeat-induced gaps for all members of a group is represented by (−). The lower and upper bounds of the hollow rectangle correspond to the first and third quartile, and the range is indicated by the whiskers. Any outliers are plotted 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

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

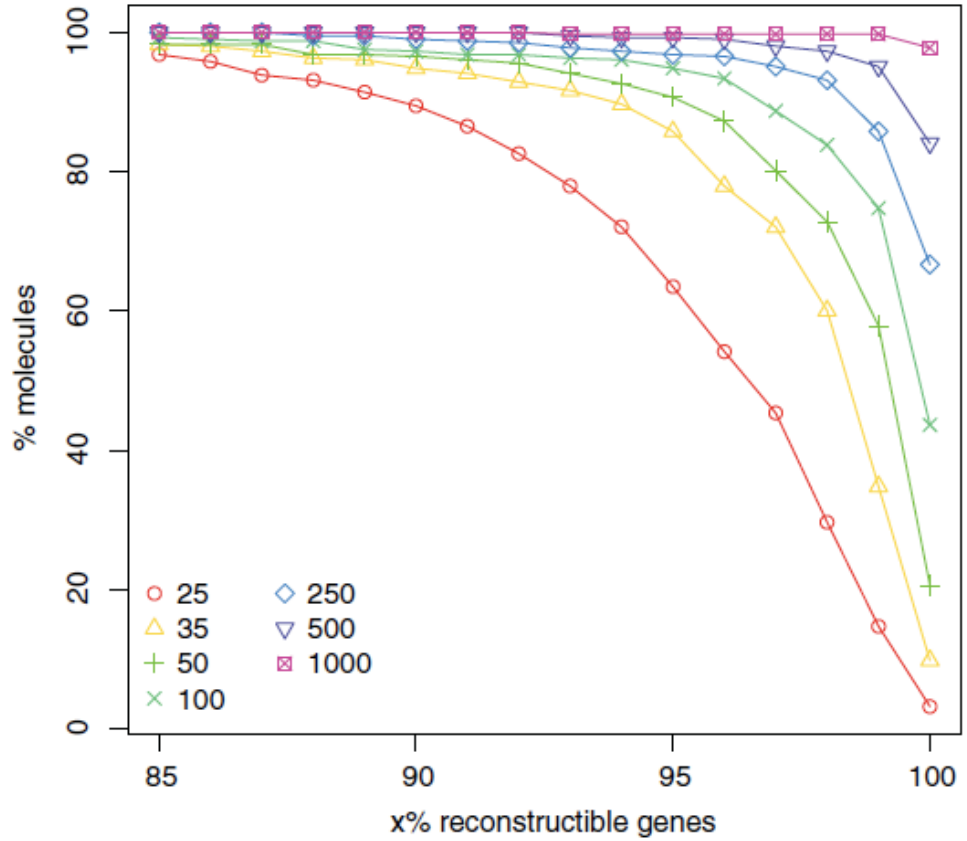


Figure 4 Fraction of reconstructible genes. Cumulative histogram plotting a percentage of genes (x-axis) against the percentage of chromosomes for which at least that many genes can be reconstructed. When $k = 1000$, nearly all chromosomes have all their genes completely reconstructed. When $k = 25$, 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.

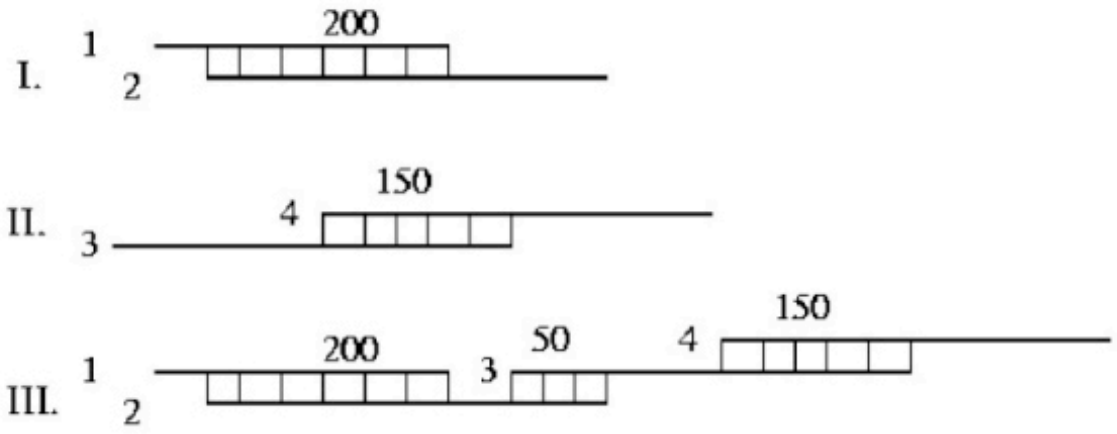


Figure 8. Greedy assembly of four reads.

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.

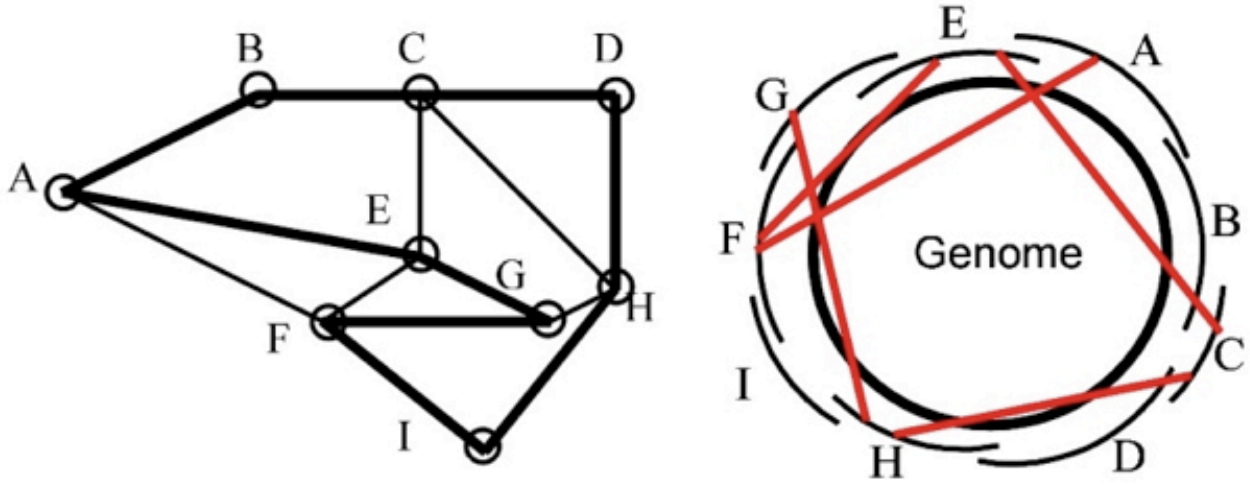
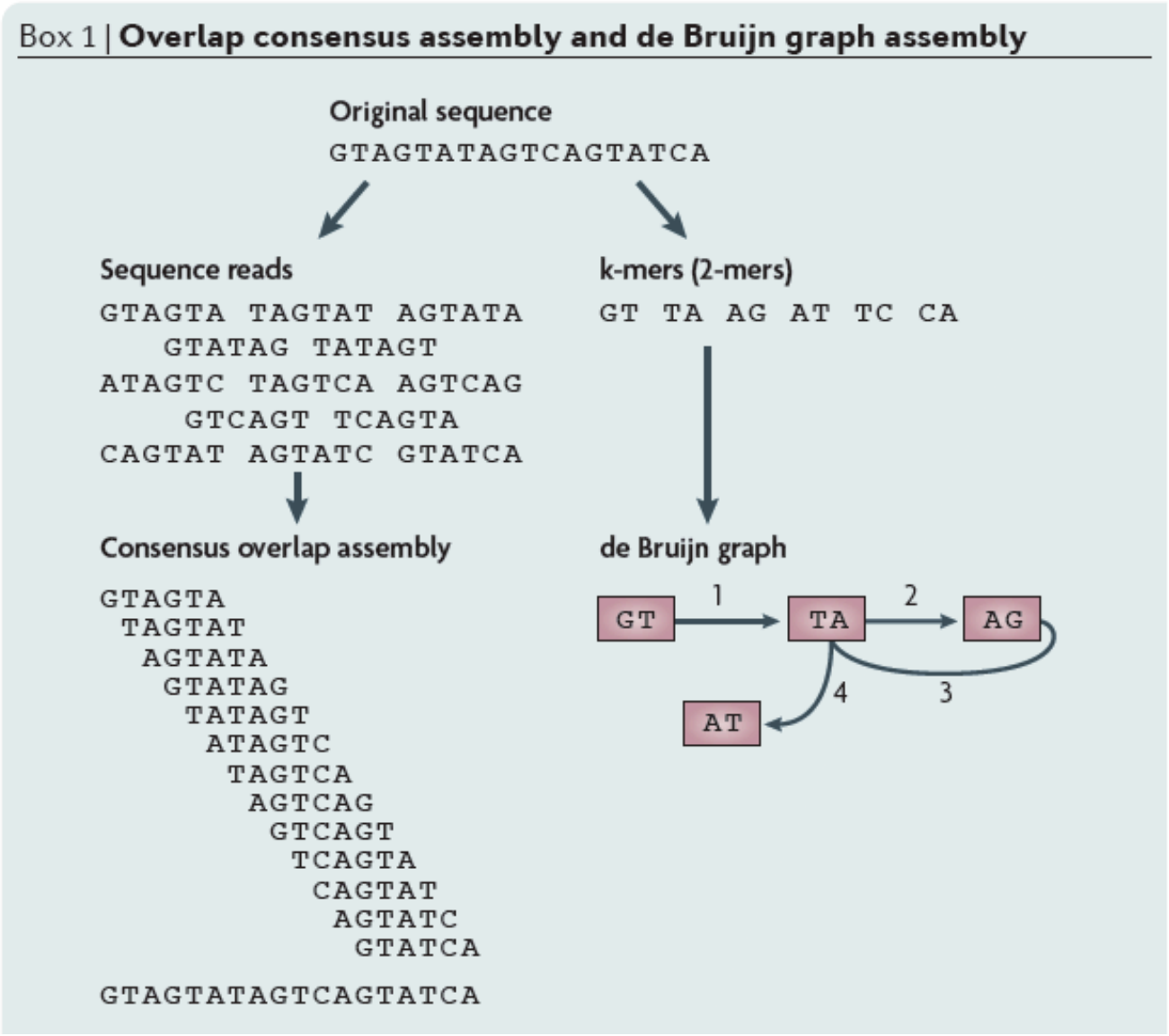


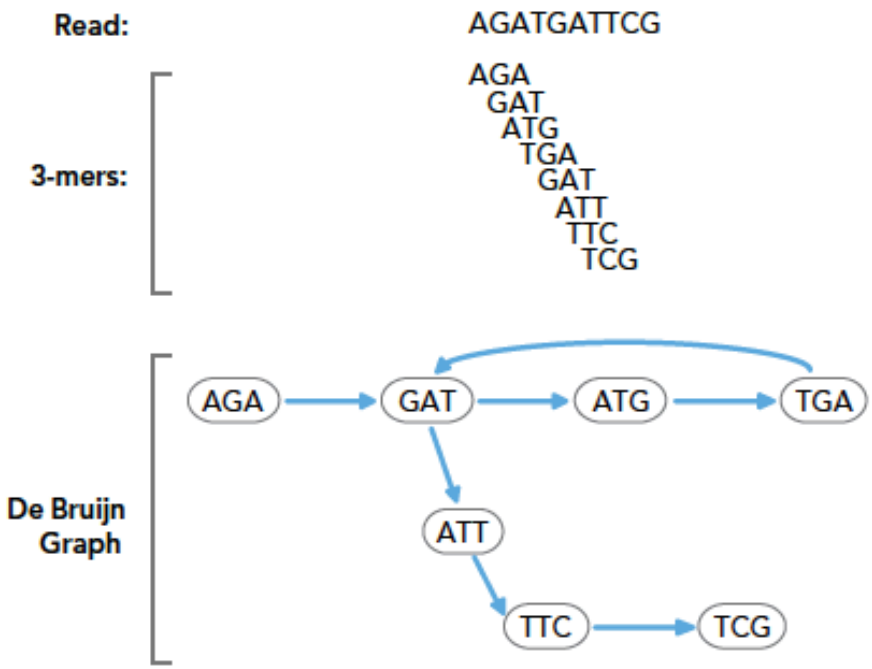
Figure 9. Overlap graph for a bacterial genome. The thick edges in the picture on the left (a Hamiltonian cycle) correspond to the correct layout of the reads along the genome (figure on the right). The remaining edges represent false overlaps induced by repeats (exemplified by the red lines in the figure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

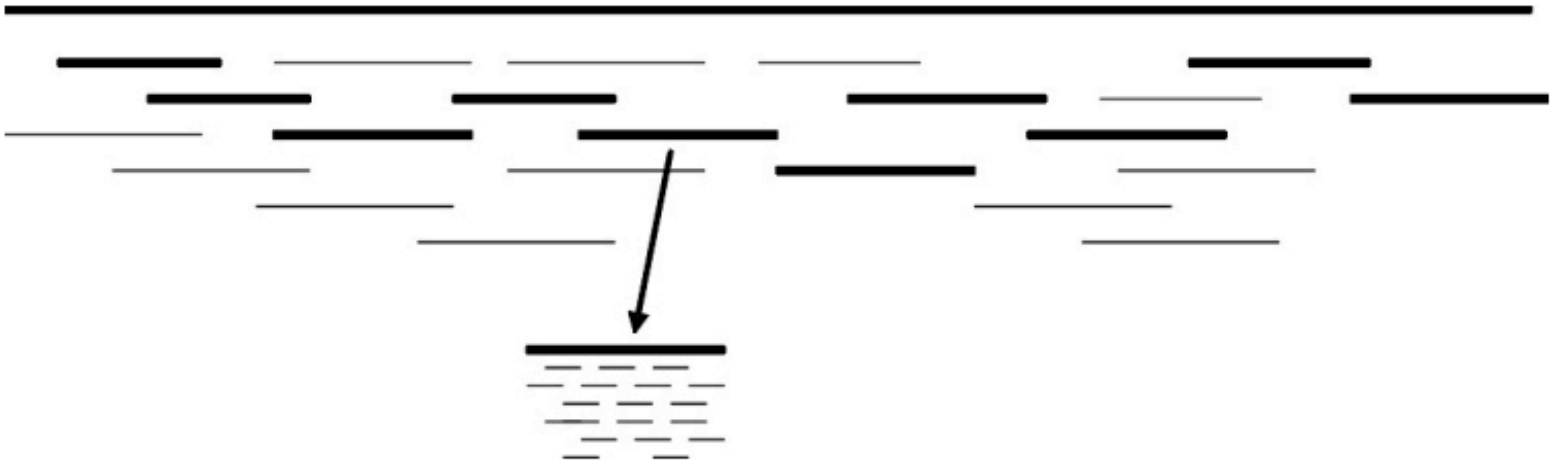
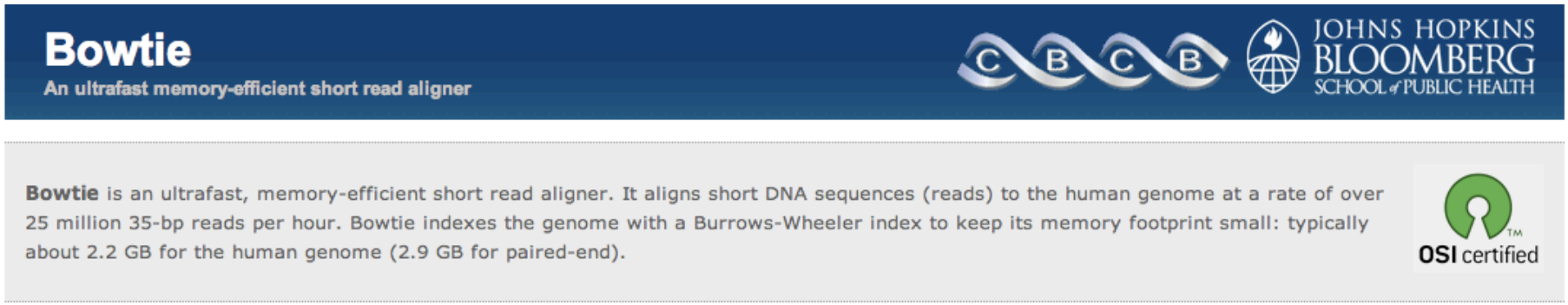


Figure 10. BAC-by-BAC approach. The long lines represent individual BACs. The minimal tiling path is represented by thick lines. Each BAC in the tiling path is then sequenced through the shotgun method.

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.

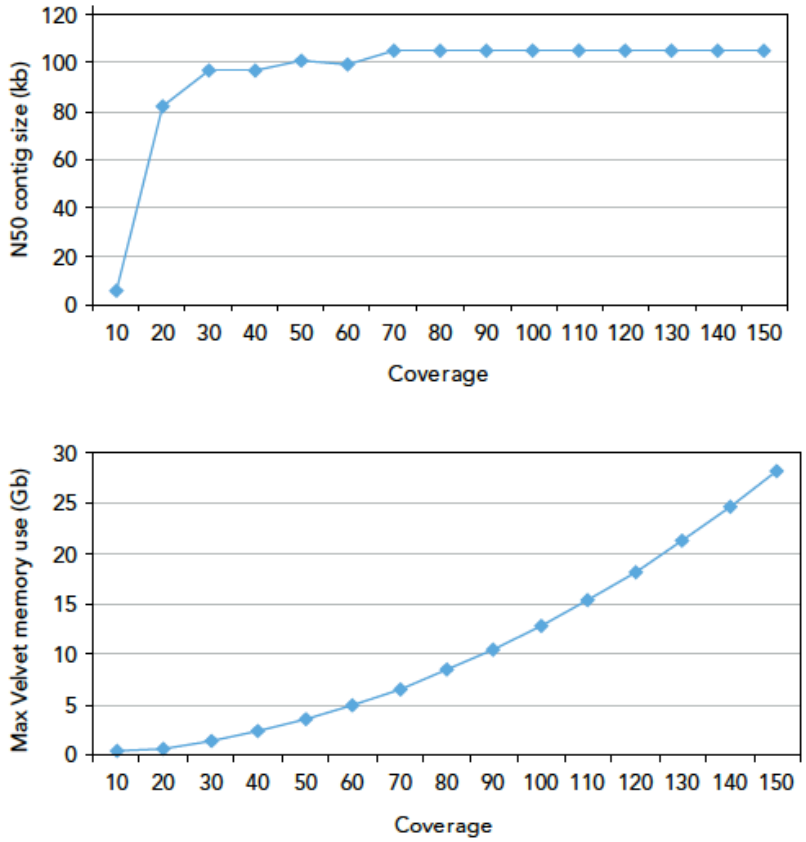
- Μεγάλη τιμή του N50 σημαίνει ότι το μεγαλύτερο μέρος του γονιδιώματος βρίσκεται σε λίγα και μεγάλα contigs.

- 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.

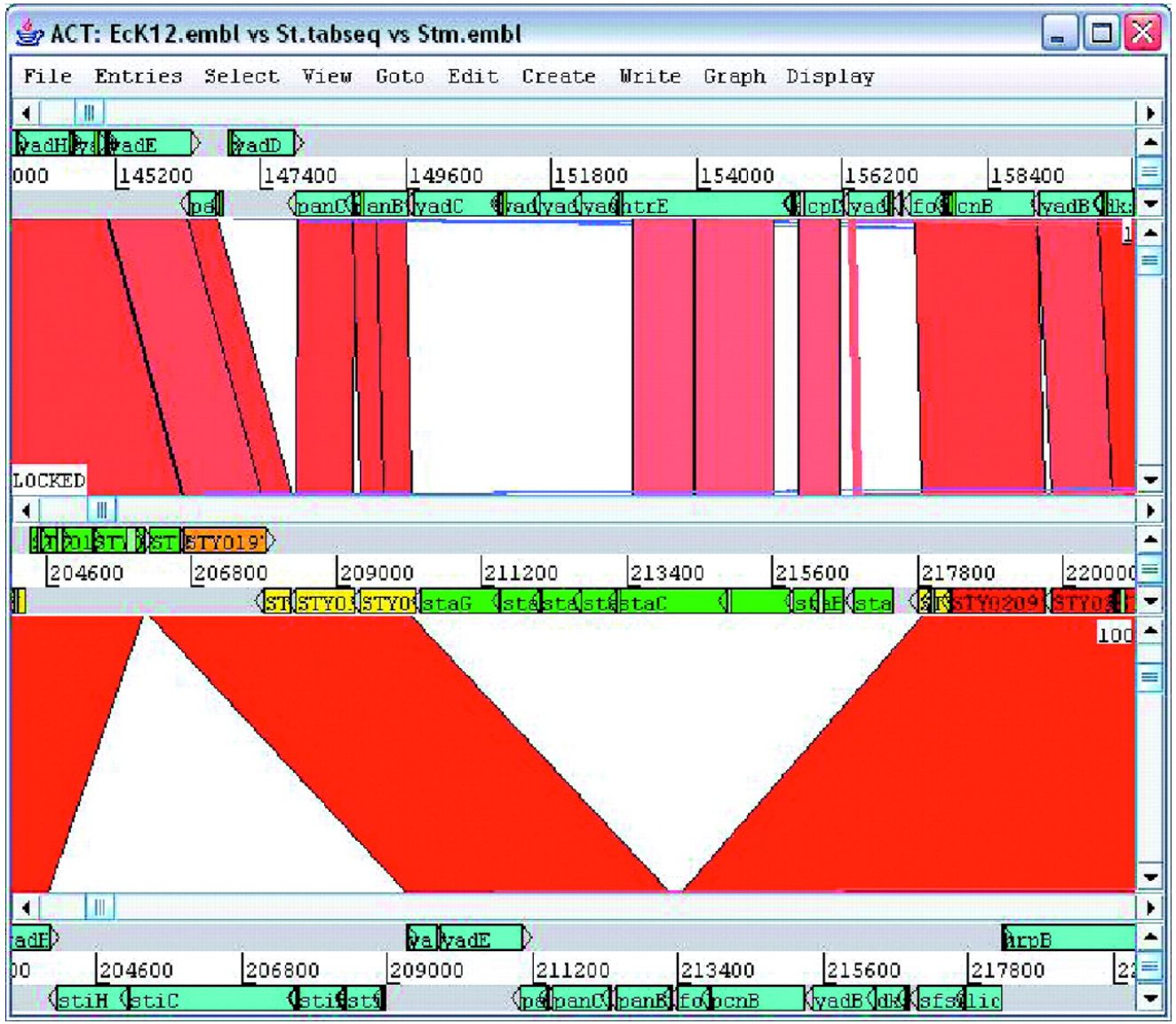
Εφαρμογές

‘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).

Επιδημία χολέρας στην Αϊτή 2010

•Αλληλούχιση του γονιδιώματος:
- •2 κλινικών στελεχών από την τωρινή επιδημία στην Αϊτή.
- •1 κλινικό στέλεχος από την επιδημία του 1991 στη Νότια Αμερική.
- •2 στέλεχη που απομονώθηκαν στη Νότια Ασία το 2002 και 2008.

•Επίσης χρησιμοποιήθηκαν οι μερικές αλληλουχίες από 23 άλλα στελέχη ανά την υφήλιο (τα τελευταία 98 χρόνια).
•1588 συντηρημένα ορθόλογα γονίδια χρησιμοποιήθηκαν από το κάθε στέλεχος, για να γίνει το φυλογενετικό δένδρο.

Επιδημία χολέρας στην Αϊτή 2010

ORIGIN OF CHOLERA OUTBREAK STRAIN IN HAITI

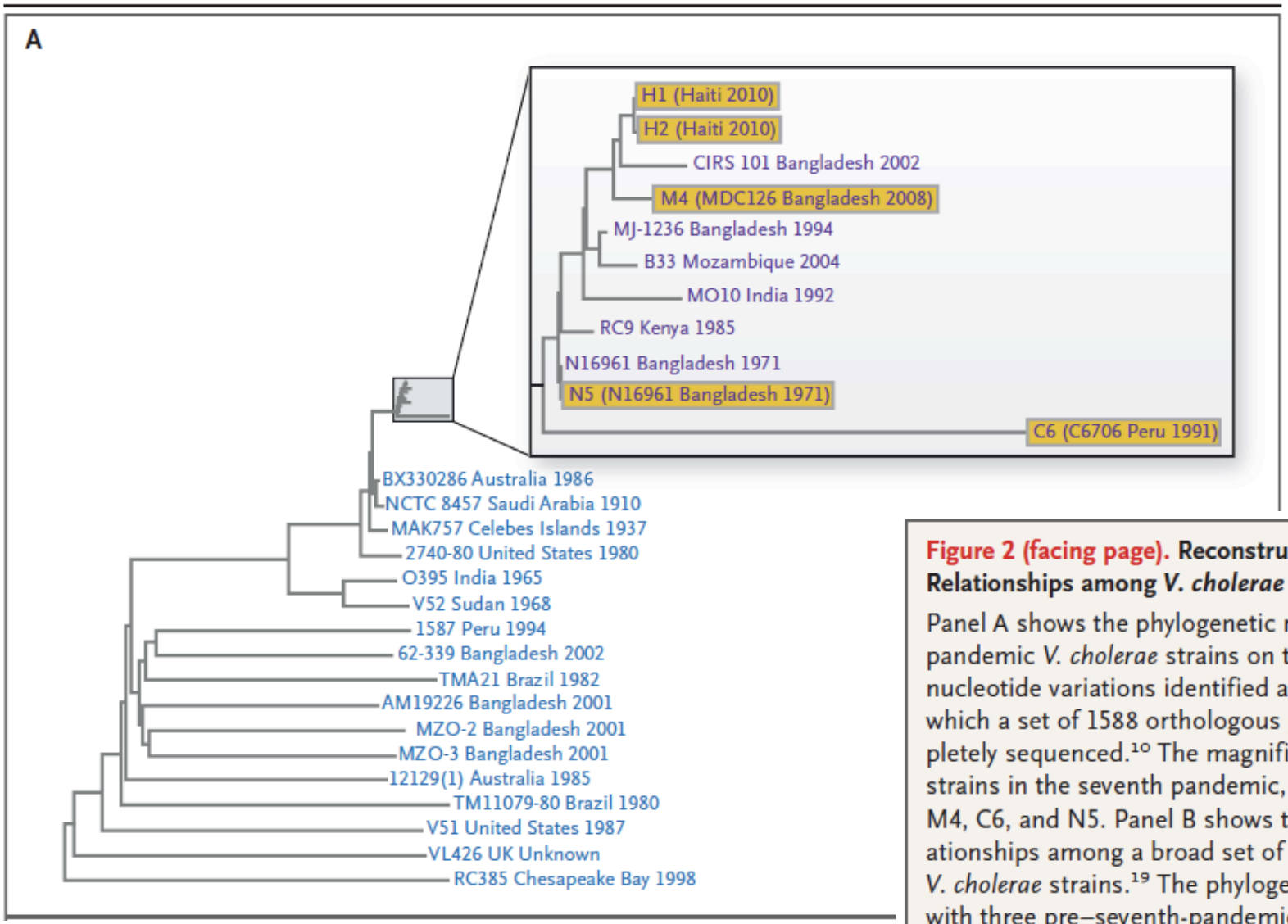


Figure 2 (facing page). Reconstructing Phylogenetic Relationships among *V. cholerae* Strains.

Panel A shows the phylogenetic relationships among pandemic *V. cholerae* strains on the basis of single-nucleotide variations identified among all strains for which a set of 1588 orthologous genes has been completely sequenced.[10] The magnified inset represents strains in the seventh pandemic, including H1, H2, M4, C6, and N5. Panel B shows the phylogenetic relationships among a broad set of seventh-pandemic *V. cholerae* strains.[19] The phylogenetic tree is rooted with three pre–seventh-pandemic strains.

Οι ανθρώπινοι εντερότυποι

ARTICLE

doi:10.1038/nature09944

Enterotypes of the human gut microbiome

Manimozhiyan Arumugam[1]*, Jeroen Raes[1,2]*, Eric Pelletier[3,4,5], Denis Le Paslier[3,4,5], Takuji Yamada[1], Daniel R. Mende[1], Gabriel R. Fernandes[1,6], Julien Tap[1,7], Thomas Bruls[3,4,5], Jean-Michel Batto[7], Marcelo Bertalan[8], Natalia Borruel[9], Francesc Casellas[9], Leyden Fernandez[10], Laurent Gautier[8], Torben Hansen[11,12], Masahira Hattori[13], Tetsuya Hayashi[14], Michiel Kleerebezem[15], Ken Kurokawa[16], Marion Leclerc[7], Florence Levenez[7], Chaysavanh Manichanh[9], H. Bjørn Nielsen[8], Trine Nielsen[11], Nicolas Pons[7], Julie Poulain[3], Junjie Qin[17], Thomas Sicheritz-Ponten[8,18], Sebastian Tims[15], David Torrents[10,19], Edgardo Ugarte[3], Erwin G. Zoetendal[15], Jun Wang[17,20], Francisco Guarner[9], Oluf Pedersen[11,21,22,23], Willem M. de Vos[15,24], Søren Brunak[8], Joel Doré[7], MetaHIT Consortium†, Jean Weissenbach[3,4,5], S. Dusko Ehrlich[7] & Peer Bork[1,25]

Our knowledge of species and functional composition of the human gut microbiome is rapidly increasing, but it is still based on very few cohorts and little is known about variation across the world. By combining 22 newly sequenced faecal metagenomes of individuals from four countries with previously published data sets, here we identify three robust clusters (referred to as enterotypes hereafter) that are not nation or continent specific. We also confirmed the enterotypes in two published, larger cohorts, indicating that intestinal microbiota variation is generally stratified, not continuous. This indicates further the existence of a limited number of well-balanced host–microbial symbiotic states that might respond differently to diet and drug intake. The enterotypes are mostly driven by species composition, but abundant molecular functions are not necessarily provided by abundant species, highlighting the importance of a functional analysis to understand microbial communities. Although individual host properties such as body mass index, age, or gender cannot explain the observed enterotypes, data-driven marker genes or functional modules can be identified for each of these host properties. For example, twelve genes significantly correlate with age and three functional modules with the body mass index, hinting at a diagnostic potential of microbial markers.

http://www.nature.com/nature/journal/v473/n7346/full/nature09944.html

Οι ανθρώπινοι εντερότυποι

- Χρησιμοποιήθηκαν 22 μεταγενώματα κοπράνων, μαζί με προηγούμενα δημοσιευμένα δεδομένα (13+2+2), σύνολο 39.
- Δείγματα από 4 κράτη (Δανία, Γαλλία, Ιταλία, Ισπανία).
- Από προηγούμενες έρευνες, δείγματα από Ιαπωνία, Αμερική

- Εντοπίστηκαν 3 βασικοί εντερότυποι.
- 12 γονίδια συσχετίζονται με την ηλικία.
- 3 λειτουργικές ομάδες (functional modules) συσχετίζονται με τον δείκτη μάζας σώματος.

Μέγεθος μικροβιακού γονιδιώματος

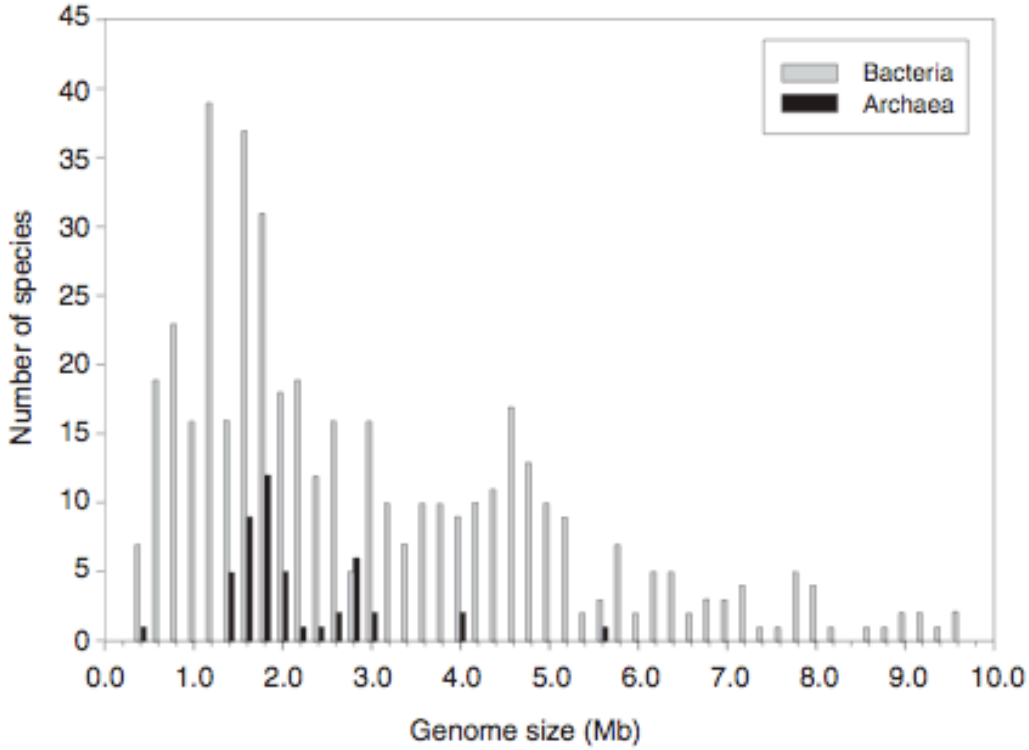


FIGURE 10.12 The distribution of genome size variation among prokaryotes based on complete genome sequencing and pulse-field gel electrophoresis (PFGE) estimates. This includes 18 complete sequences and 29 PFGE measurements for Archaea (black bars), and 125 complete sequences and 323 PFGE estimates for 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

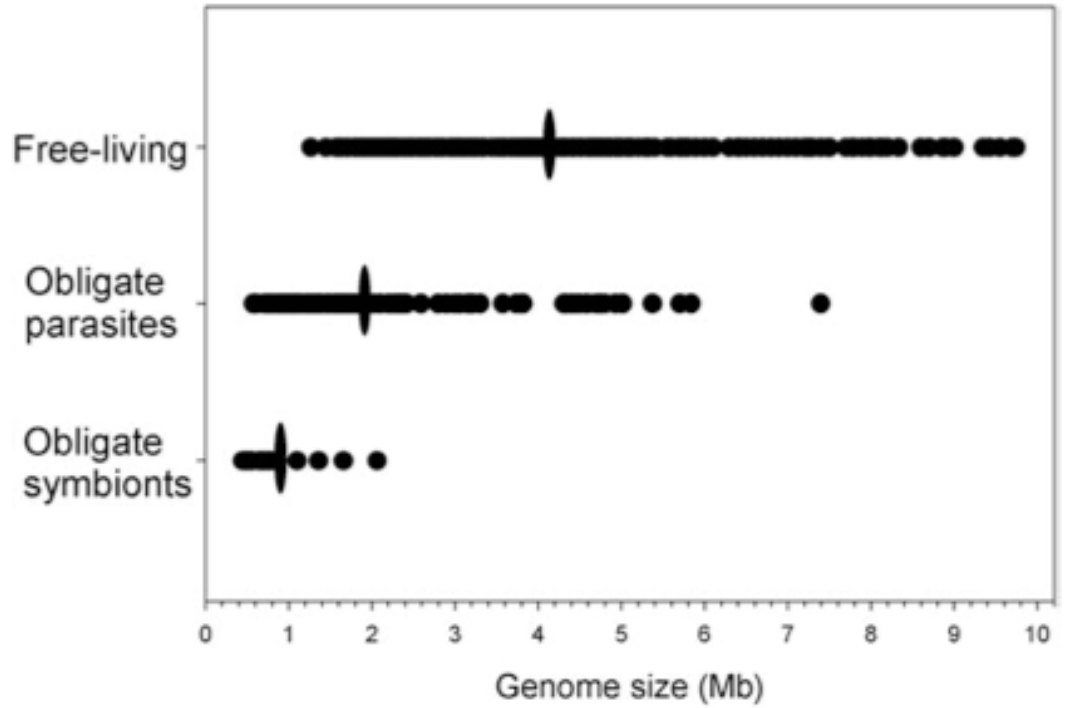


FIGURE 10.13 The distribution of genome sizes according to lifestyle in the Bacteria. Each point represents the genome size (measured by pulse-field gel electrophoresis) of one species or strain of bacteria categorized as either free-living (n = 398), obligately parasitic (n = 227), or obligately symbiotic (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

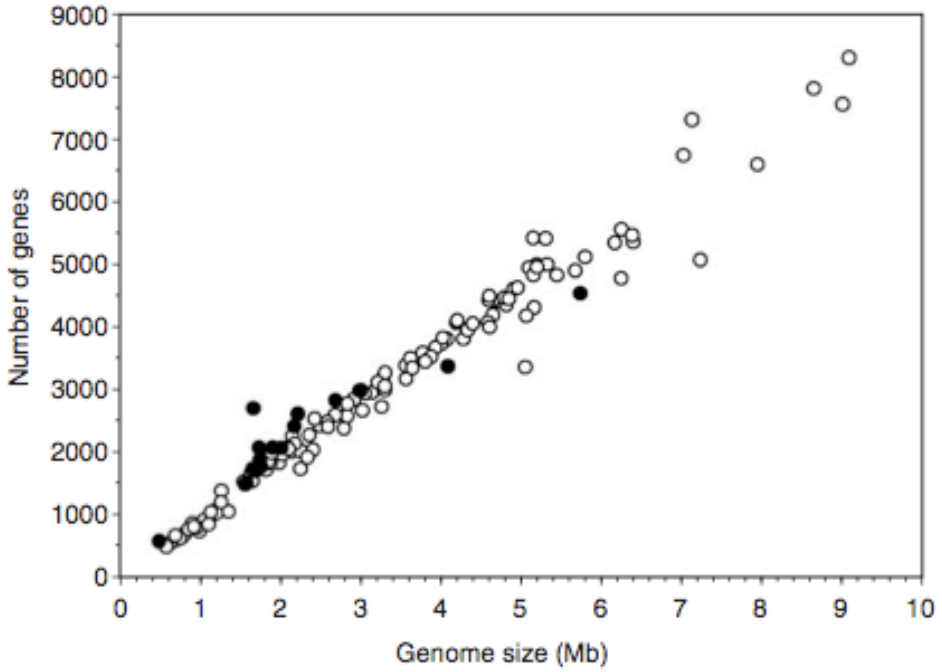


FIGURE 10.11 The relationship between gene (i.e., open reading frame) number and genome size in prokaryotes, as revealed by data from 140 completely sequenced genomes. Unlike in eukaryotes, gene number is strongly positively correlated with genome size in both Archaea (●) and Bacteria (○). The regression 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 C.

Προφάγοι στο γονιδίωμα

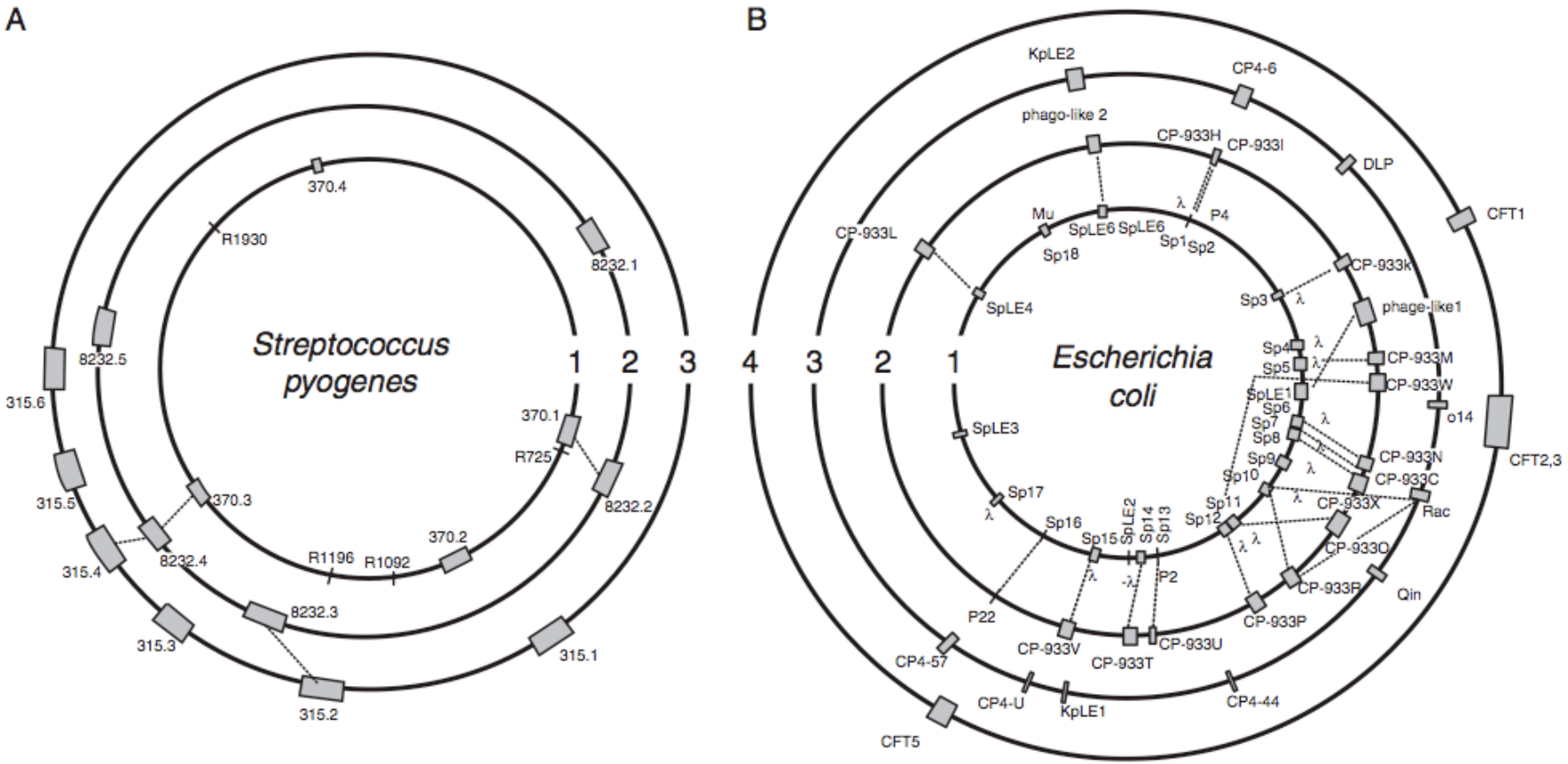


FIGURE 10.9 Prophage content and locations (gray boxes) in several strains of two species of bacteria. **(A)** *Streptococcus pyogenes*, also known as "group A *Streptococcus*" (GAS), which causes a wide range of infections. The numbered rings represent the genomes of three different serotypes: (1) M1, (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.

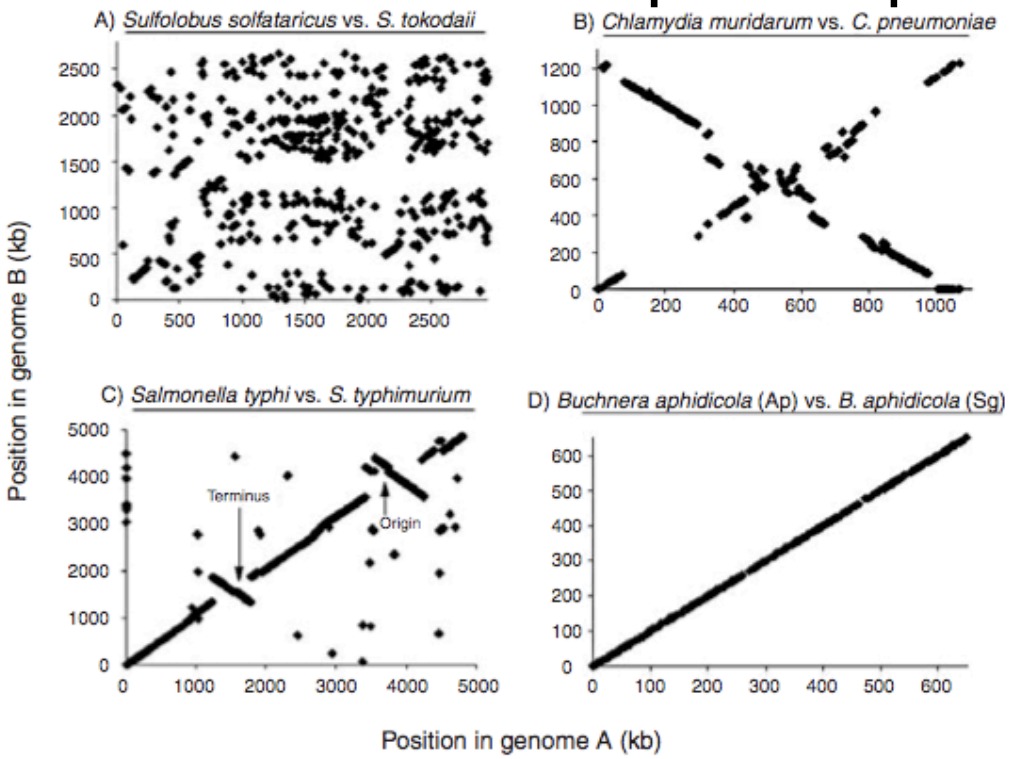


FIGURE 10.6 Gene position plots showing examples of both plasticity and stability in gene order between closely related species of prokaryotes. In these plots, the location of a given gene, measured as its distance from a given starting point in kilobases (kb), is plotted on one axis each for the two 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.