CICESE

UNIVERSIDAD AUTÓNOMA METROPOLITANA
Casa abierta al tiempo
Unidad Iztapalapa

Casa abierta al tiempo

Secuenciación masiva y anotación de genes.

Dr. Miguel Ángel del Río Portilla
Cátedra Divisional
"Alejandro Villalobos"

Casa abierta al tiempo

Secuenciación masiva y anotación de genes

- **ORGANIZADORES:**
- Dra. Irene de los Ángeles Barriga Sosa
- Dr. Miguel Angel del Río Portilla

- **PROFESORES**
- Dr. Miguel Angel del Río Portilla.
- mdelrio@cicese.mx, mdelrio@xanum.uam.mx
- M. en C. Eduardo Zúñiga León
- eduardoz@xanum.uam.mx
- Dra. Dra. Erika Magallón Gayón

Casa abierta al tiempo

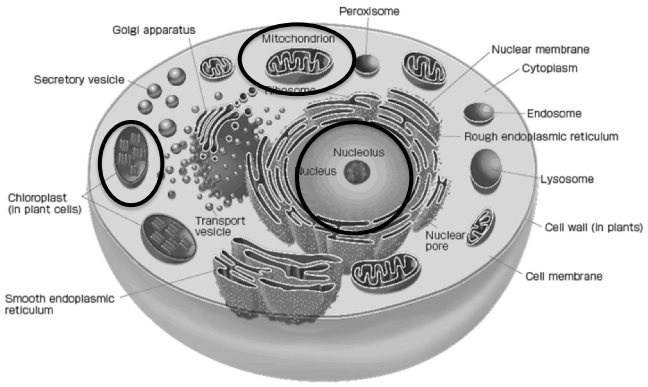
Bioinformática

- Colección, clasificación, almacenamiento y análisis de información bioquímica y biológica utilizando computadoras, especialmente con aplicación a genética molecular y genómica. (https://www.merriam-webster.com/dictionary/bioinformatics)
- La suma de aproximaciones computacionales para analizar, manejar y almacenar datos biológicos

Anotación de genes nucleares y mitocondriales

Casa abierta al tiempo

Célula eucariota



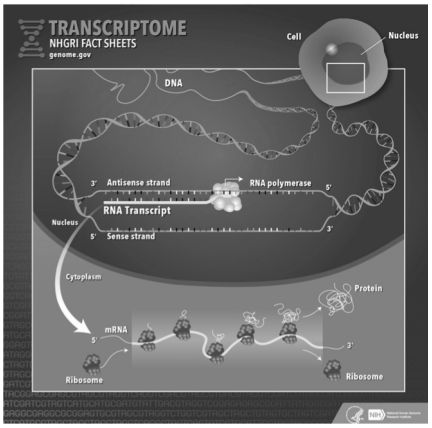
© CSLS/The University of Tokyo

http://csls-text2.c.u-tokyo.ac.jp/inactive/02_05.html

Casa abierta al tiempo

Genoma

- Gen
  - Cromosoma
  - Genoma
- RNA (mRNA transcripto)
  - Transcriptoma
- Proteína
  - Proteoma
- Interacción de genes
  - Epigenética
- Tejido y condición específico



Casa abierta al tiempo

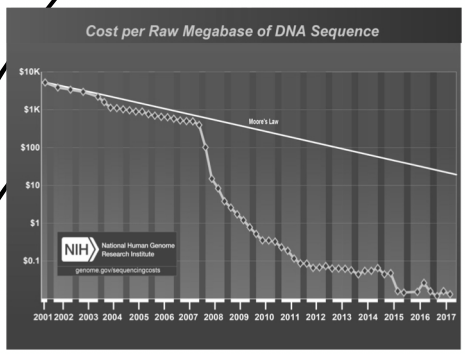
Herramientas Genómicas

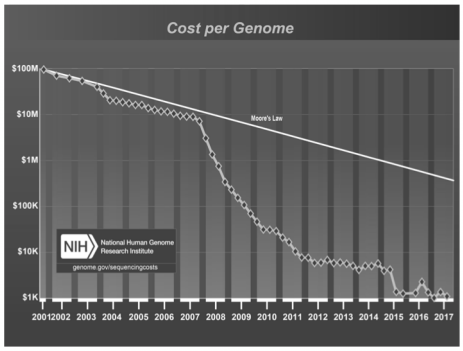
- Proyecto del genoma Humano (1990-2003)
- Secuenciación Sanger
- Desarrollo de nuevas tecnologias
- Next generation Sequencing
- Secuencies de varios millones de fragmentos (lecturas) en una corrida
  - Microsatélites
  - SNP
  - Genomas Mitochondriales

Indispensable para microsatélites y SNPs y análisis de RNA (RNAseq, miRNA…)

Casa abierta al tiempo

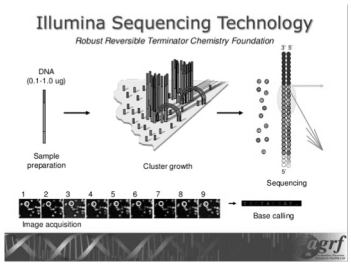
Costos de la secuenciación masiva





Casa abierta al tiempo

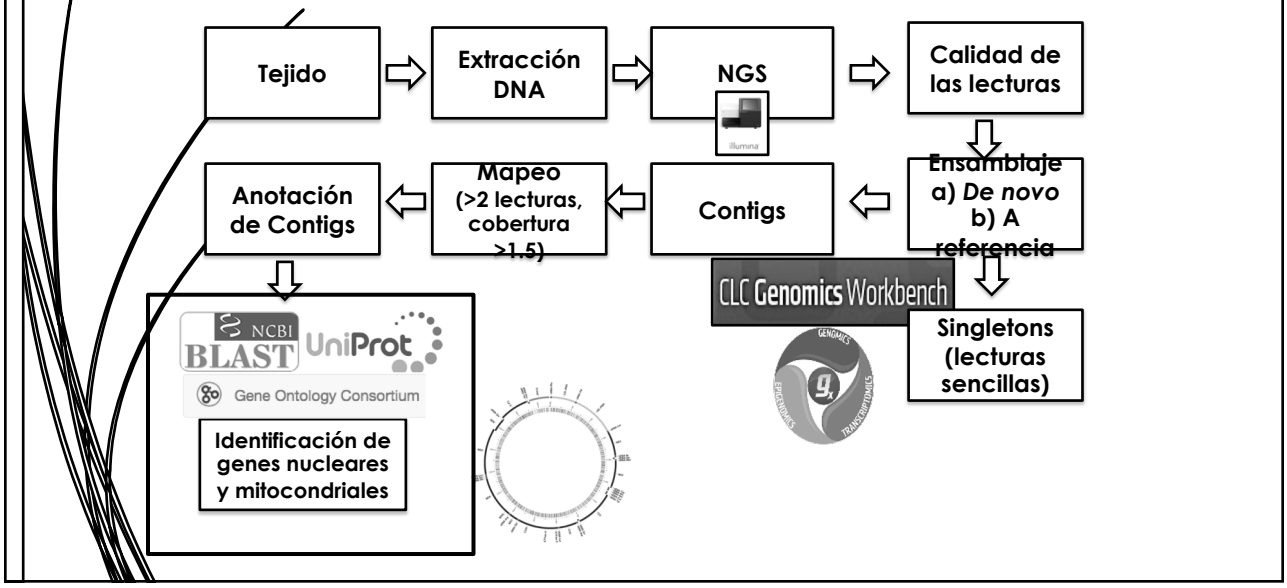
NGS análisis genómico y transcriptómico



- Obtención de genoma
  - Nuclear (identificación del número de genes)
  - Mitocondrial (estructura)
- Microsatélites
  - Genética de poblaciones
  - Identificación de progenitores
  - Localización de caracteres cuantitativos de interés (QTLs)
- SNP
  - Identificación de de progenitores
  - Localización de caracteres cuantitativos de interés (QTLs)
- Secuenciación de mRNA (transcriptoma: tejido/condición)
  - Genes de interés
  - Respuesta a enfermedades

Casa abierta al tiempo

Procedimiento para el análisis genómico



Casa abierta al tiempo

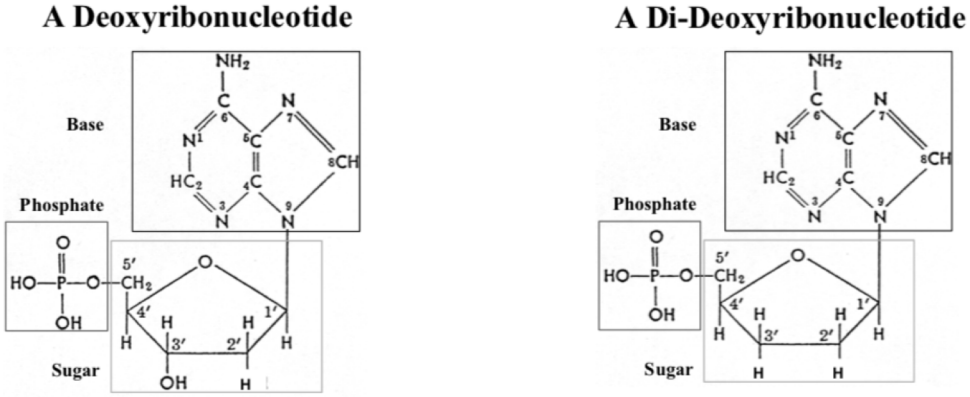
Secuención del DNA

- Método Maxam–Gilbert (1977)
  - Utilizaba fósforo 32 y reactivos para marcar, de manera independiente, las bases nitrogenadas
  - Enzimas de restricción para cortar en sitios específicos
  - Lograron secuenciar un plásmido de 5000 base de longitud
  - Fue muy popular para secuenciar, pero el método de Sanger fue más eficiente
- Sanger (1977)
  - Método de síntesis y terminación
  - Marcaba cada base con P^{32}
  - Electroforesis en gel de agarosa

https://www.future-science.com/btm/news/ju18/04

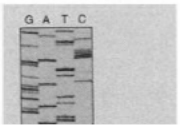
Casa abierta al tiempo

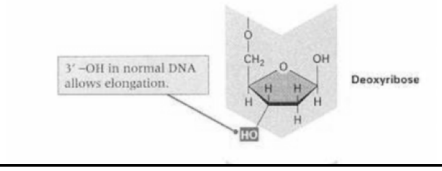
Secuenciación Sanger. Secuenciación por terminación

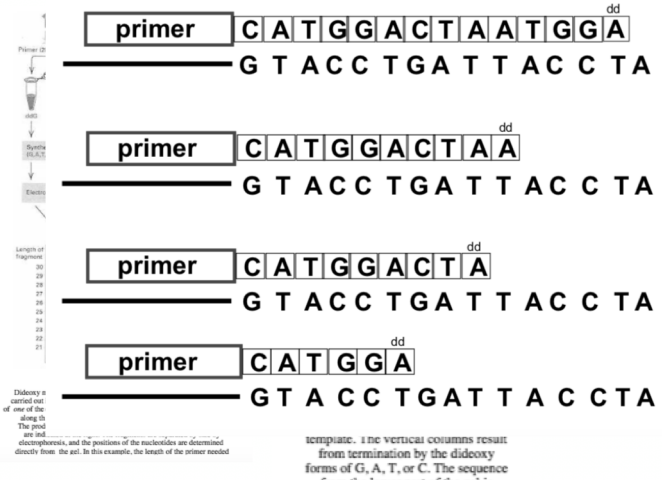


Casa abierta al tiempo

Secuenciación







Hartl, D. L., Jones, E. W. (1998). *Genetics: Principles and analysis*. Jones and Barlett Publishers.

Secuenciación

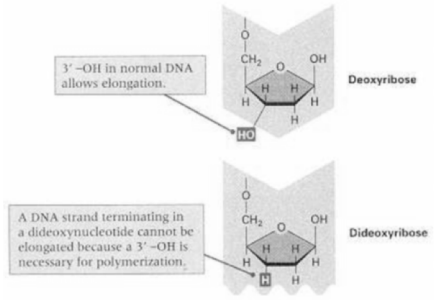


Figure 5.38
Structures of normal deoxyribose and the dideoxyribose sugar used in DNA sequencing. The dideoxyribose has a hydrogen atom (red) attached to the 3′ carbon, in contrast with the hydroxyl group (red) at this position in deoxyribose. Because the 3′ hydroxyl group is essential for the attachment of the next nucleotide in line in a growing DNA strand, the incorporation of a dideoxynucleotide immediately terminates synthesis.

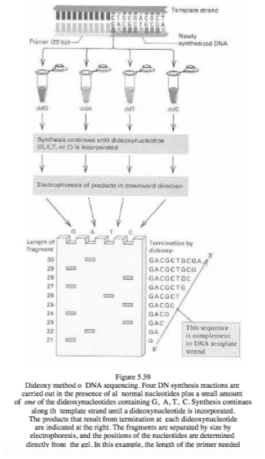
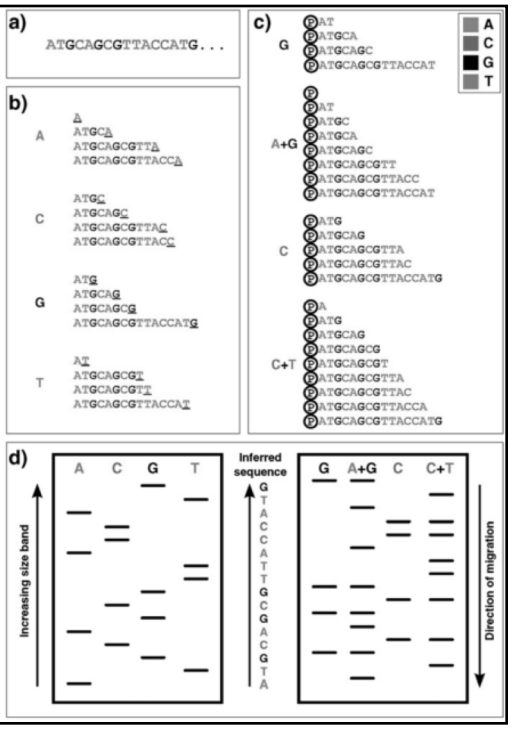


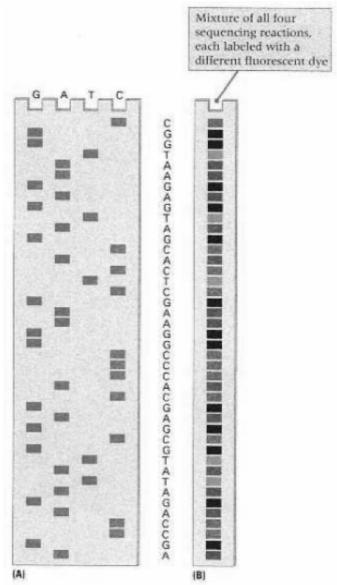
Figure 5.39
Dideoxy method o DNA sequencing. Four DN synthesis reactions are carried out in the presence of al normal nucleotides plus a small amount of one of the dideoxynucleotides containing G, A, T, C. Synthesis continues along th template strand until a dideoxynucleotide is incorporated. The products that result from termination at each dideoxynucleotide are indicated at the right. The fragments are separated by size by electrophoresis, and the positions of the nucleotides are determined directly from the gel. In this example, the length of the primer needed



Hartl, D. L., Jones, E. W. (1998). *Genetics: Principles and analysis.* Jones and Barlett Publishers.

Electroforesis capilar

Casa abierta al tiempo



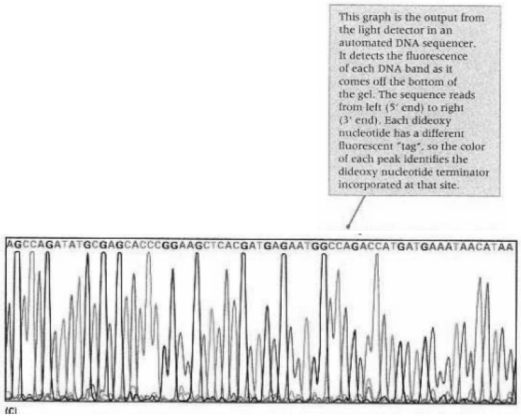
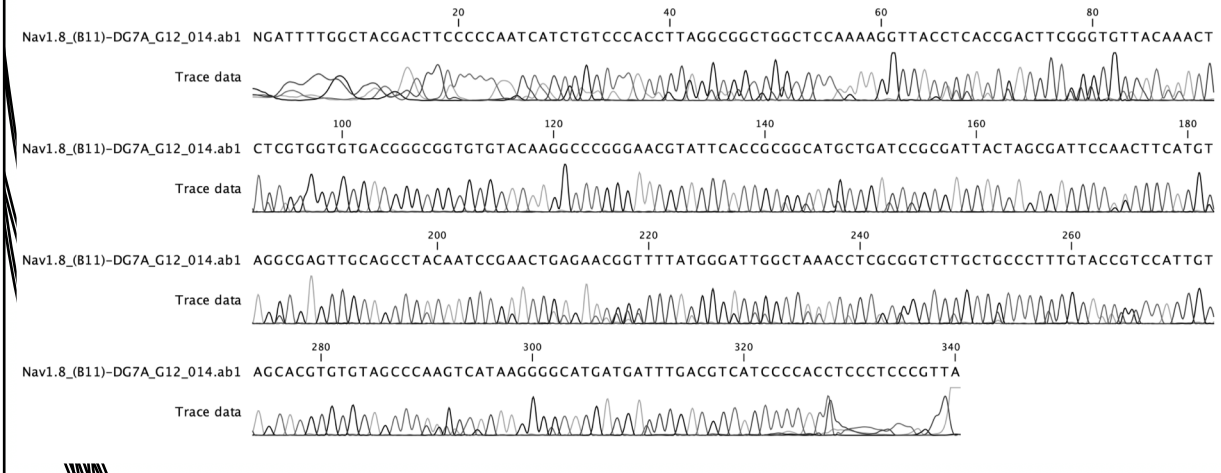


Figure 9.31
Automated DNA sequencing. (A) Conventional sequencing gel obtained from the dideoxy procedure (Section 5.9). The DNA sequence can be determined directly from photographic film according to the positions of the bands. (B) Banding pattern obtained when each of the terminating nucleotides is labeled with a different fluorescent dye and the bands are separated in the same lane of the gel. (C) Trace of the fluorescence pattern obtained from the gel in part B by automated detection of the fluorescence of each band as it comes off the bottom of the gel during continued electrophoresis.

Casa abierta al tiempo

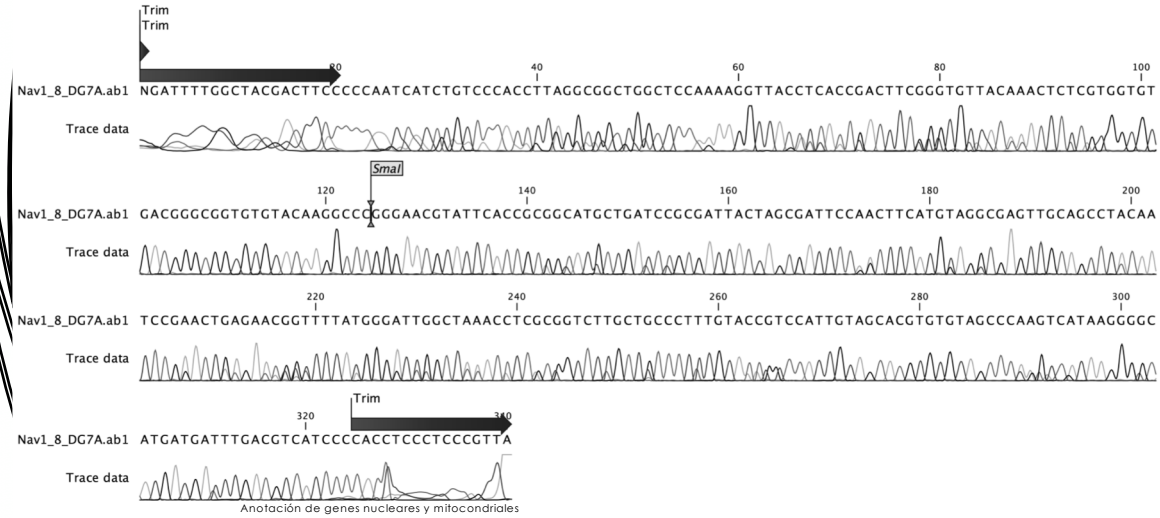
Secuencia, archivo .ab1



Anotación de genes nucleares y mitocondriales

Casa abierta al tiempo

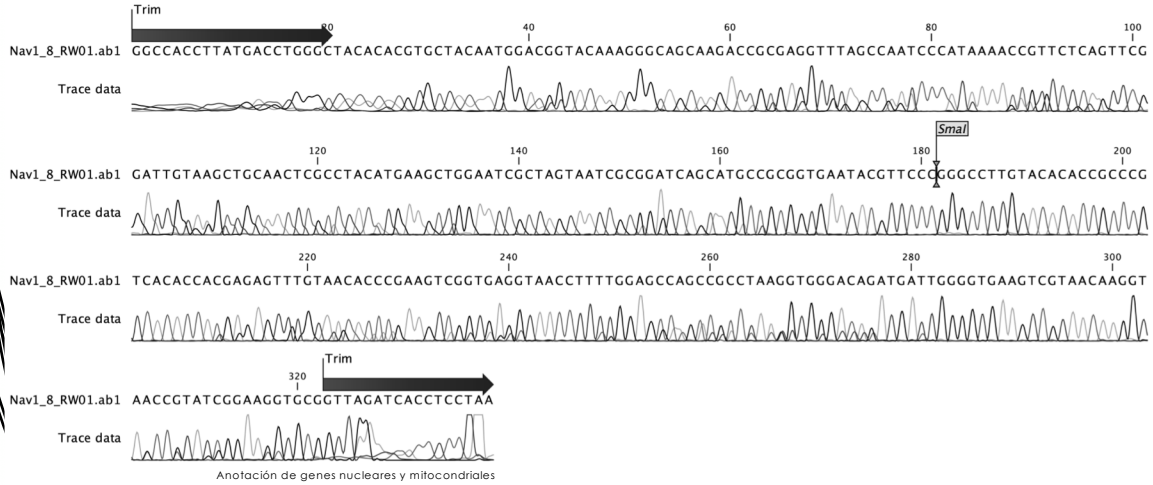
Calidad de la secuencia



Anotación de genes nucleares y mitocondriales

Casa abierta al tiempo

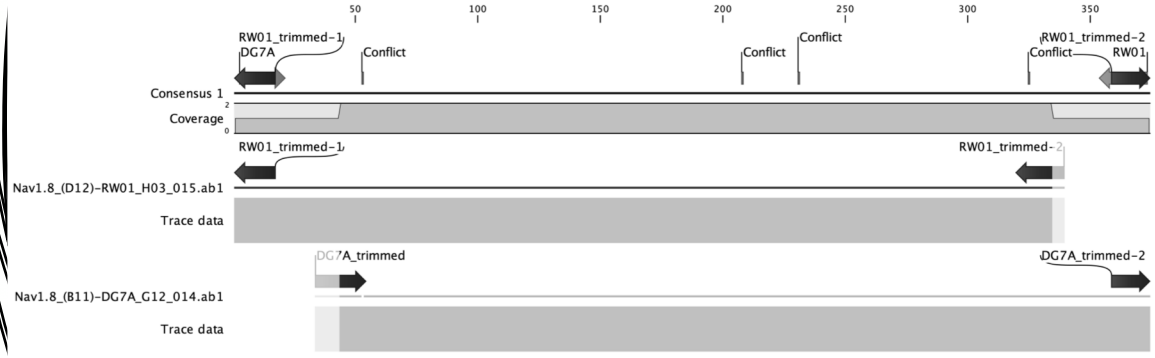
Calidad de la secuencia reverso complementaria



Anotación de genes nucleares y mitocondriales

Casa abierta al tiempo

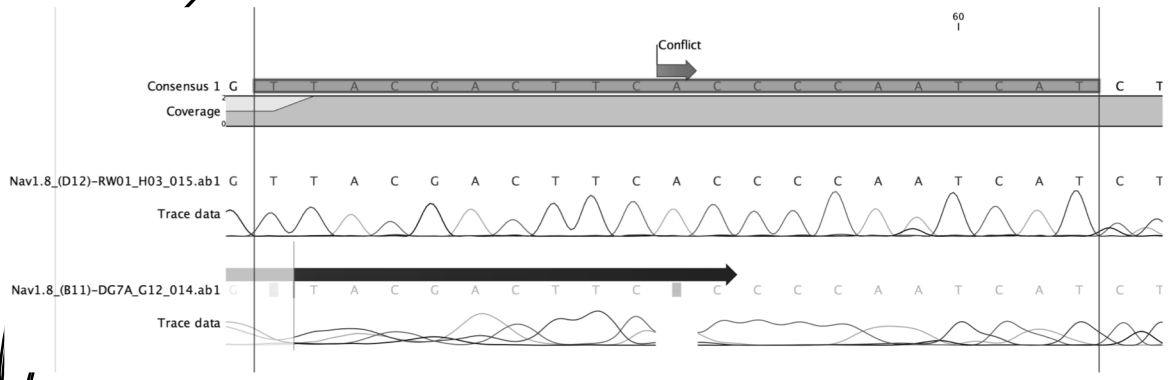
Consenso



Anotación de genes nucleares y mitocondriales

Casa abierta al tiempo

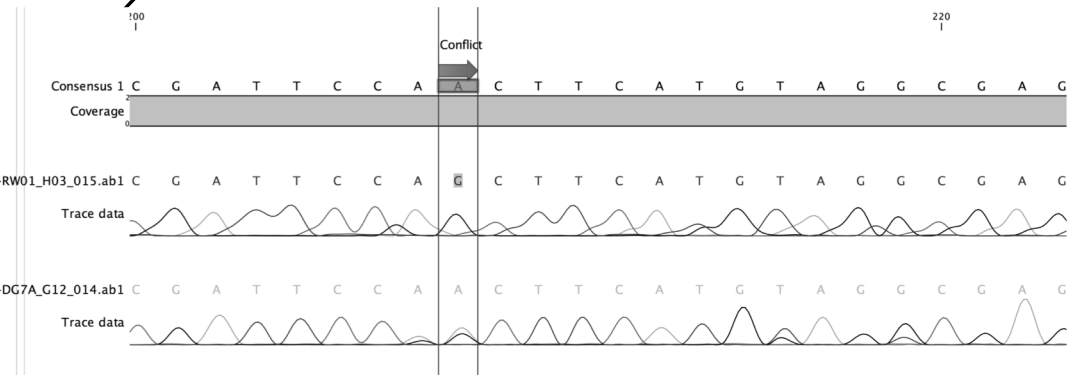
Conflicto 1



Anotación de genes nucleares y mitocondriales

Casa abierta al tiempo

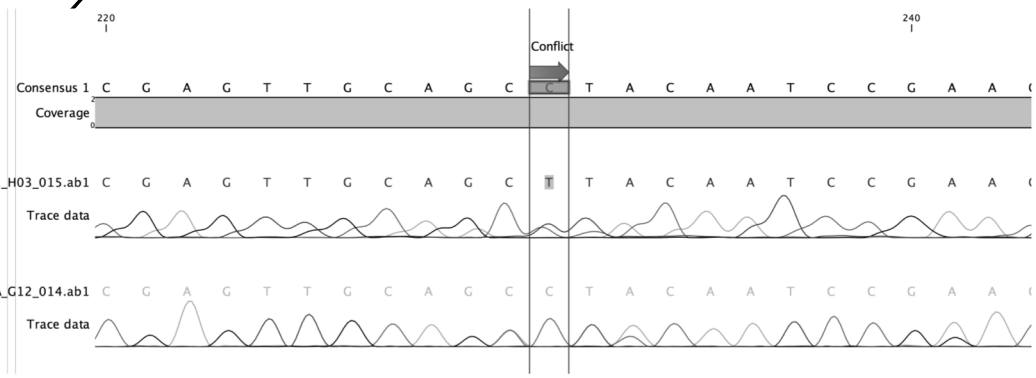
Conflicto 2



Anotación de genes nucleares y mitocondriales

Casa abierta al tiempo

Conflicto 3



Casa abierta al tiempo

Secuenciación de siguiente generación Herramientas Genómicas

- Proyecto del genoma Humano (1990-2003) (~3 Gb)
- Secuenciación Sanger
  (600-800 pb)

Desarrollo de nuevas tecnologias

- Next generation sequencing (NGS)
- Secuencias de varios millones de fragmentos (lecturas) en una corrida
  - Microsatélites
  - SNP
  - Genomas Mitocondriales

Indispensable para microsatélites y SNPs y análisis de RNA (RNAseq, miRNA…)

illumina Roche life technologies PACIFIC BIOSCIENCES Helicos ion torrent Oxford NANOPORE

Casa abierta al tiempo

Secuenciación de siguiente generación (Next generation sequencing)

Pirosecuenciación 454

Cambios en el pH IonTorrent

Illumina

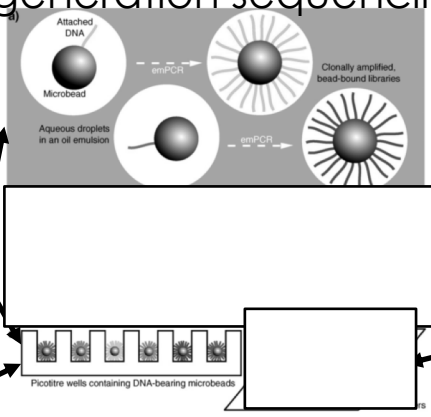


Fig. 2. Second-generation DNA sequencing parallelized amplification. (a): DNA molecules being clonally amplified in an emulsion PCR (emPCR). Adapter ligation and PCR produces DNA libraries with appropriate 5′ and 3′ ends, which can then be made single stranded and immobilized onto individual suitably oligonucleotide-tagged microbeads. Bead–DNA conjugates can then be emulsified using aqueous amplification reagents in oil, ideally producing emulsion droplets containing only one bead (illustrated in the two leftmost droplets, with different molecules indicated in different colours). Clonal amplification then occurs during the emPCR as each template DNA is physically separate from all others, with daughter molecules remaining bound to the microbeads. This is the conceptual basis underlying sequencing in 454, Ion Torrent and polony sequencing protocols. (b): Bridge amplification to produce clusters of clonal DNA populations in a planar solid-phase PCR reaction, as occurs in Solexa/Illumina sequencing. Single-stranded DNA with terminating sequences complementary to the two lawn-oligos will anneal when washed over the flow-cell, and during isothermal PCR will replicate in a confined area, bending over to prime at neighbouring sites, producing a local cluster of identical molecules. (c) and (d) demonstrate how these two different forms of clonally-amplified sequences can then be read in a highly parallelized manner: emPCR-produced microbeads can be washed over a picotiter plate, containing wells large enough to fit only one bead (c). DNA polymerase can then be added to the wells, and each nucleotide can be washed over in turn, and dNTP incorporation monitored (e.g. via pyrophosphate or hydrogen ion release). Flow-cell bound clusters produced via bridge amplification (d) can be visualized by detecting fluorescent reversible-terminator nucleotides at the ends of a proceeding extension reaction, requiring cycle-by-cycle measurements and removal of terminators.

Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. doi:10.1016/j.ygeno.2015.11.003

Casa abierta al tiempo

Next generation sequencing

Roche 454

illumina

Helicos 2 pass

Pac Bio

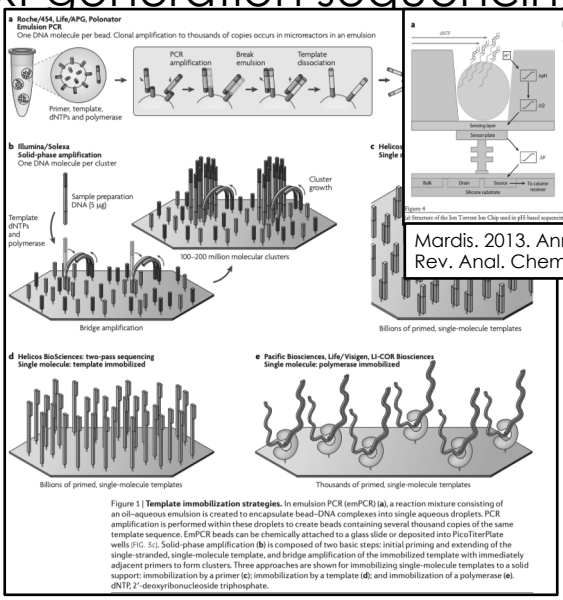


Figure 1 | **Template immobilization strategies.** In emulsion PCR (emPCR) (**a**), a reaction mixture consisting of an oil–aqueous emulsion is created to encapsulate bead–DNA complexes into single aqueous droplets. PCR amplification is performed within these droplets to create beads containing several thousand copies of the same template sequence. EmPCR beads can be chemically attached to a glass slide or deposited into PicoTiterPlate wells (FIG. 3c). Solid-phase amplification (**b**) is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. Three approaches are shown for immobilizing single-molecule templates to a solid support: immobilization by a primer (**c**); immobilization by a template (**d**); and immobilization of a polymerase (**e**). dNTP, 2′-deoxyribonucleoside triphosphate.

Mardis. 2013. Annu. Rev. Anal. Chem.

Metzker 2010. Nature Reviews.

Casa abierta al tiempo

Tercer generación de secuenciación

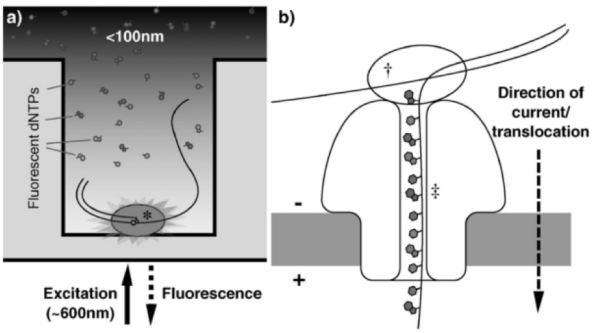


Fig. 3. Third-generation DNA sequencing nucleotide detection. (a): Nucleotide detection in a zero-mode waveguide (ZMW), as featured in PacBio sequencers. DNA polymerase molecules are attached to the bottom of each ZMW (*), and target DNA and fluorescent nucleotides are added. As the diameter is narrower than the excitation light's wavelength, illumination rapidly decays travelling up the ZMW: nucleotides being incorporated during polymerisation at the base of the ZMW provide real-time bursts of fluorescent signal, without undue interference from other labelled dNTPs in solution. (b): Nanopore DNA sequencing as employed in ONT's MinION sequencer. Double stranded DNA gets denatured by a processive enzyme (†) which ratchets one of the strands through a biological nanopore (‡) embedded in a synthetic membrane, across which a voltage is applied. As the ssDNA passes through the nanopore the different bases prevent ionic flow in a distinctive manner, allowing the sequence of the molecule to be inferred by monitoring the current at each channel.

Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. doi:10.1016/j.ygeno.2015.11.003

Casa abierta al tiempo

Comparación

Table 1 | **Summary of NGS platforms**

Platform	Read length (bp)	Throughput	Reads	Runtime	Error profile	Instrument cost (US$)	Cost per Gb (US$, approx.)
Sequencing by ligation							
SOLiD 5500 Wildfire	50 (SE)	80 Gb	~700 M*	6 d*	≤0.1%, AT bias‡	NA§	$130‡
	75 (SE)	120 Gb					
	50 (SE)*	160 Gb*					
SOLiD 5500xl	50 (SE)	160 Gb	~1.4 B*	10 d*	≤0.1%, AT bias‡	$251,000‡	$70‡
	75 (SE)	240 Gb					
	50 (SE)*	320 Gb*					
BGISEQ-500 FCS[155]	50–100 (SE/PE)*	8–40 Gb*	NA‖	24 h*	≤0.1%, AT bias‡	$250 (REF. 155)	NA‖
BGISEQ-500 FCL[155]	50–100 (SE/PE)*	40–200 Gb*	NA‖	24 h*	≤0.1%, AT bias‡	$250,000 (REF. 155)	NA‖

Anotación de genes nucleares y mitocondriales

Casa abierta al tiempo

Comparación

Table 1 | **Summary of NGS platforms**

Platform	Read length (bp)	Throughput	Reads	Runtime	Error profile	Instrument cost (US$)	Cost per Gb (US$, approx.)
Sequencing by synthesis: CRT							
Illumina MiniSeq Mid output	150 (SE)*	2.1–2.4 Gb*	14–16 M*	17 h*	<1%, substitution‡	$50,000 (REF. 118)	$200–300 (REF. 118)
Illumina MiniSeq High output	75 (SE)	1.6–1.8 Gb	22–25 M (SE)*	7 h	<1%, substitution‡	$50,000 (REF. 118)	$200–300 (REF. 118)
	75 (PE)	3.3–3.7 Gb	44–50 M (PE)*	13 h			
	150 (PE)*	6.6–7.5 Gb*		24 h*			
Illumina MiSeq v2	36 (SE)	540–610 Mb	12–15 M (SE)	4 h	0.1%, substitution‡	$99,000‡	~$1,000
	25 (PE)	750–850 Mb	24–30 M (PE)*	5.5 h			$996
	150 (PE)	4.5–5.1 Gb		24 h			$212
	250 (PE)*	7.5–8.5 Gb*		39 h*			$142‡
Illumina MiSeq v3	75 (PE)	3.3–3.8 Gb	44–50 M (PE)*	21–56 h*	0.1%, substitution‡	$99,000‡	$250
	300 (PE)*	13.2–15 Gb*					$110‡
Illumina NextSeq 500/550 Mid output	75 (PE)	16–20 Gb	Up to 260 M (PE)*	15 h	<1%, substitution‡	$250‡	$42
	150 (PE)*	32–40 Gb*		26 h*			$40‡

Casa abierta al tiempo

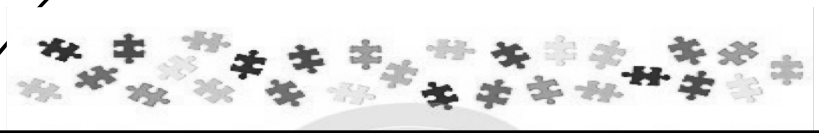
Comparación

Table 1 | **Summary of NGS platforms**

Platform	Read length (bp)	Throughput	Reads	Runtime	Error profile	Instrument cost (US$)	Cost per Gb (US$, approx.)
Sequencing by synthesis: CRT							
Illumina HiSeq2500 v2 Rapid run	36 (SE)	9–11Gb	300 M (SE)*	7 h	0.1%, substitution‡	$690‡	$230
	50 (PE)	25–30 Gb	600 M (PE)*	16 h			$90
	100 (PE)	50–60 Gb		27 h			$52
	150 (PE)	75–90 Gb		40 h			$45
	250 (PE)*	125–150 Gb*		60 h*			$40‡
Illumina HiSeq2500 v3	36 (SE)	47–52 Gb	1.5 B (SE)	2 d	0.1%, substitution‡	$690‡	$180
	50 (PE)	135–150 Gb	3 B (PE)*	5.5 d			$78
	100 (PE)*	270–300 Gb		11 d*			$45‡
Illumina HiSeq2500 v4	36 (SE)	64–72 Gb	2 B (SE)	29 h	0.1%, substitution‡	$690‡	$150
	50 (PE)	180–200 Gb	4 B (PE)*	2.5 d			$58
Illumina HiSeq3000/4000	50 (SE)	105–125 Gb	2.5 B (SE)*	1–3.5 d*	0.1%, substitution‡	$740/$900 (REF. 156)	$50
	75 (PE)	325–375 Gb					$31
	150 (PE)*	650–750 Gb*					$22 (REF. 157)

Casa abierta al tiempo

Ensamblaje

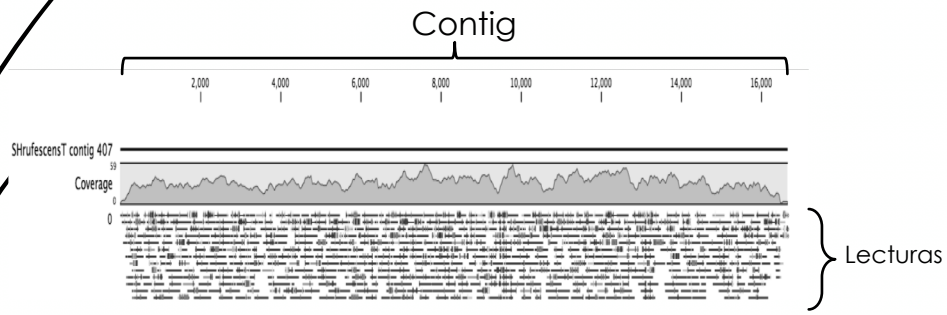


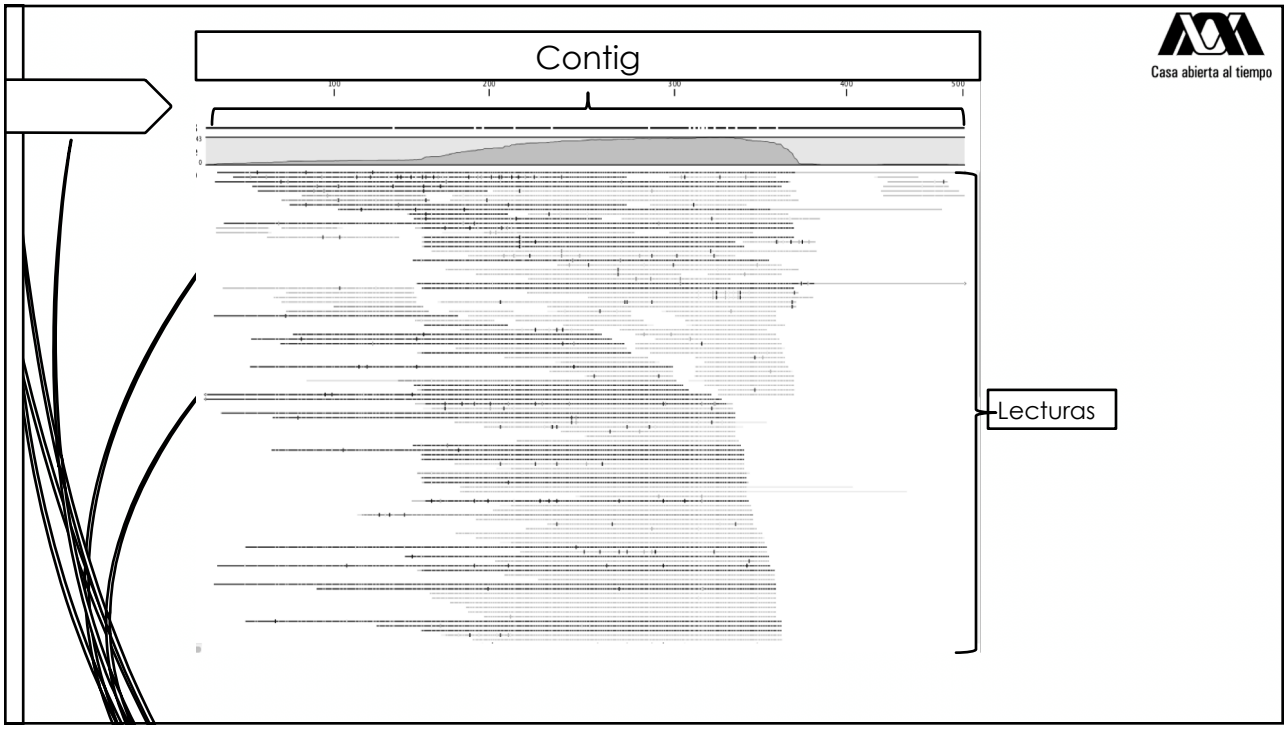
https://es.dreamstime.com/imgenes-de-archivo-libres-de-regalas-espiral-coloreado-de-la-dna-del-rompecabezas-image30690439

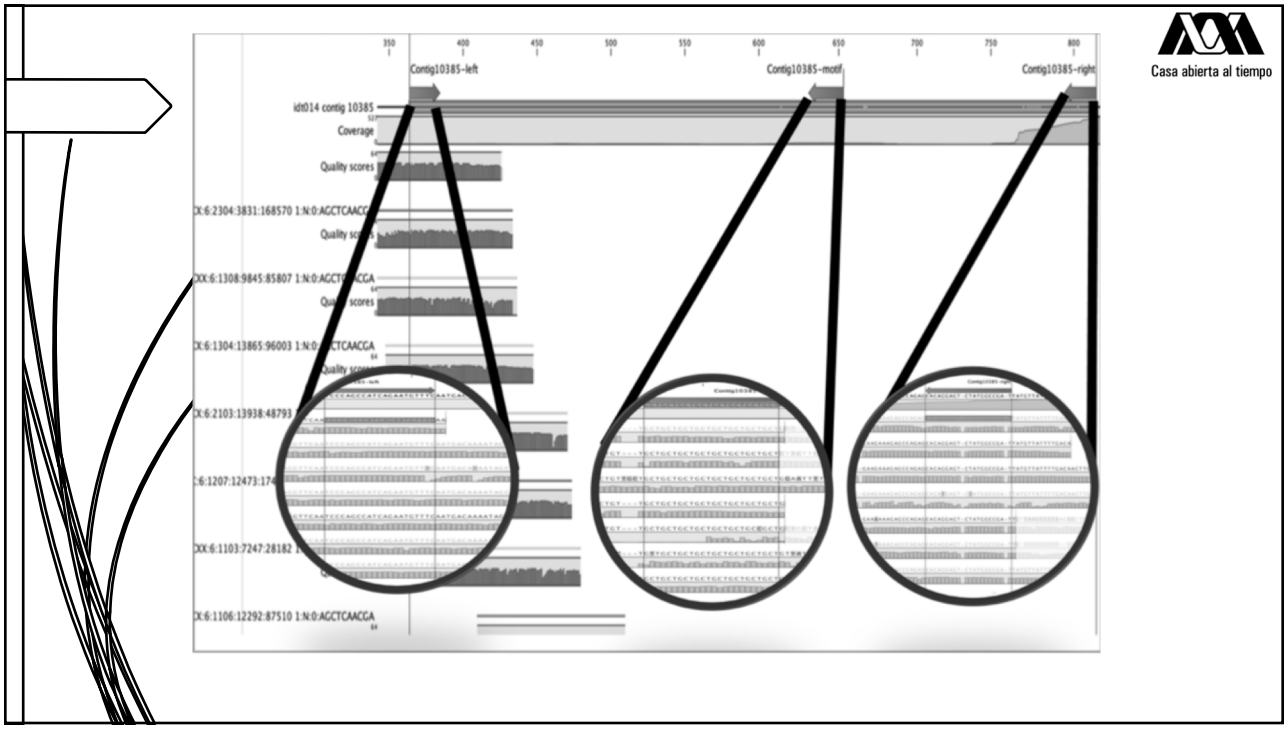
Crédito: American Mathematical Society.
http://laenciclopediagalactica.info/2012/11/13/bioinformatica-interpretando-el-genoma-humano/figura-03-2/

Casa abierta al tiempo

Ensamblaje




Contig
Lecturas
Casa abierta al tiempo


Contig10385-left
Contig10385-motif
Contig10385-right
idt014 contig 10385
Coverage
Quality scores
Casa abierta al tiempo

Casa abierta al tiempo

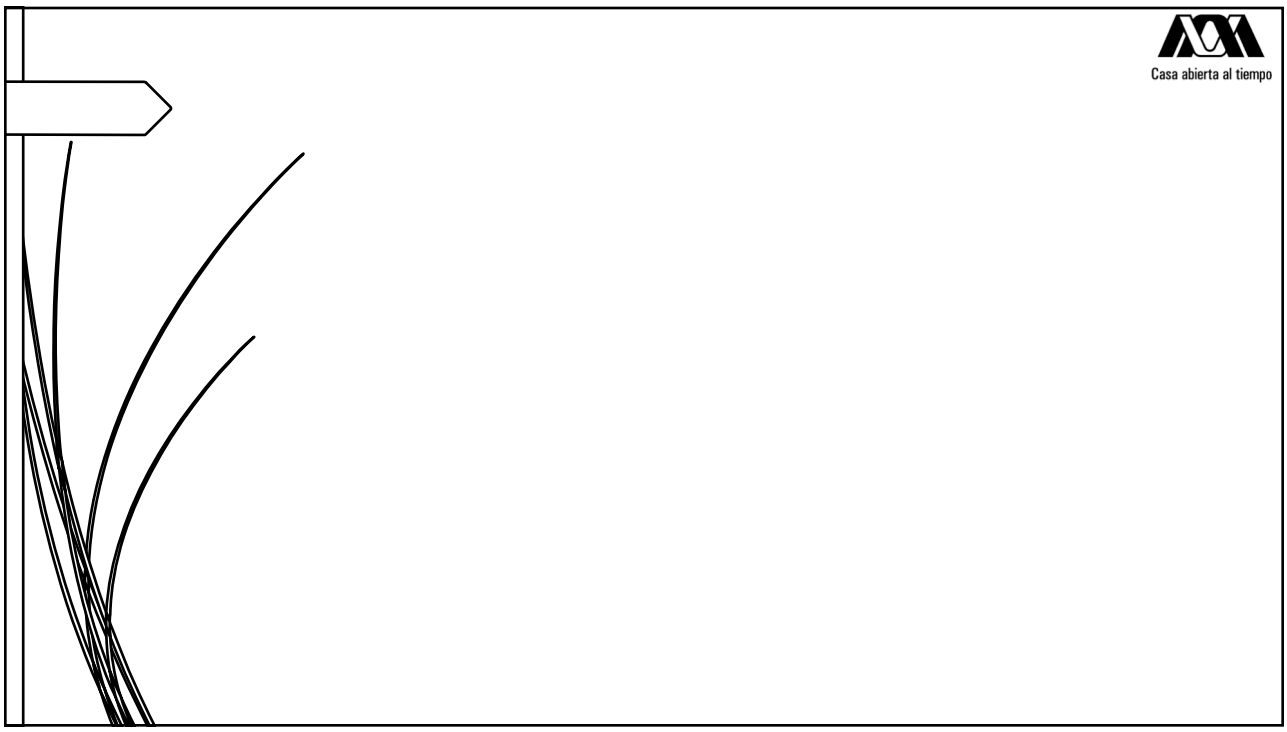
NGS análisis genómico y transcriptómico

- Obtención de genoma
  - Mitocondrial
  - Nuclear
- Secuenciación de mRNA (transcriptoma: tejido/condición)
  - Genes de interés
- Microsatélites
  - Genética de poblaciones
  - Identificación de progenitores
  - Localización de caracteres cuantitativos de interés (QTLs)
- SNP
  - Identificación de de progenitores
  - Localización de caracteres cuantitativos de interés (QTLs)

Casa abierta al tiempo

NGS

- Ventajas
  - Obtención de una gran cantidad de secuencias en una sola corrida a diferencia del método de Sanger
  - Análisis de muchas secuencias a la vez
  - Análisis cuantitativo del transcriptoma (expresión de genes)
  - Búsqueda de marcadores moleculares más fácilmente (microsatélites, Polimorfismo de un solo nucleótido [SNP])
  - Biomas (metagenómica)
  - Costo va reduciéndose cada vez más.
  - Uso en especies no modelo
  - Uso en muestras ambientales
- Desventajas
  - Requiere programas específicos para el análisis de las secuencias (bioinformática)
  - Capacidad de cómputo y almacenamiento


Casa abierta al tiempo

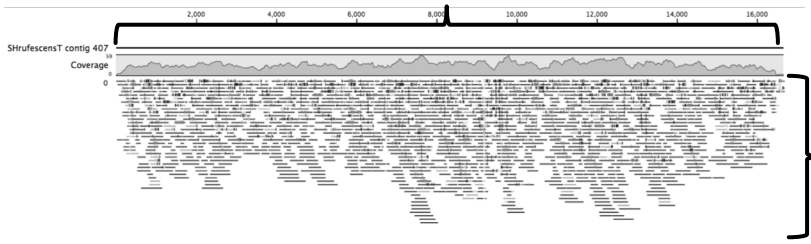

Casa abierta al tiempo
Análisis genómico
Genomas mitocondriales

Casa abierta al tiempo

Resultados del ensamblaje

	amarillo	negro
Bibliotecas	1	1
Lecturas	1,440,342	1,125,142
Contigs	83,213	59,213
Mitogenoma	16,842	16,516

Contig



Lecturas

Casa abierta al tiempo

Genes mitocondriales

Table 1. The 13 protein, two ribosomal RNA and 22 tRNA genes typically found in animal mitochondrial genomes

Protein encoded	Designation for animal mtDNA	Synonym
Cytochrome oxidase subunit I, II, III	*COI, COII, COIII*	*cox1, cox2, cox3*
Cytochrome*b* apoenzyme	*Cytb*	*cob*
NADH dehydrogenase subunits 1-6, 4L	*ND1-6, 4L*	*nad1-6, 4L*
ATP synthase subunits 6, 8	*A6, A8* or *ATP6, ATP8*	*atp6, atp8*
Large ribosomal subunit RNA	*lrRNA*	*rnl*
Small ribosomal subunit RNA	*srRNA*	*rns*
18 Transfer RNAs each specifying a single amino acid	Corresponding one-letter amino acid code	*trnX*
Two transfer RNAs specifying leucine	Differentiated by codon recognized, *L(CUN)* and *L(UUR)*	Differentiated by subscript
Two transfer RNAs specifying serine	Differentiated by codon recognized, *S(AGN)* and *S(UCN)*	Differentiated by subscript

13

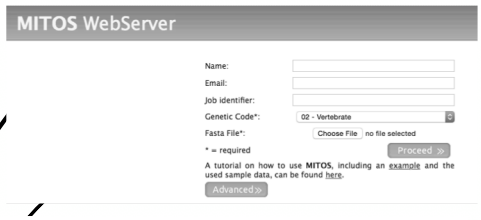
22

For historical reasons these have designations commonly used only for animal systems. To aid comparison with other works, column three lists the synonymous gene labels used for non-animal systems.

Boore, J.L., 1999. Animal mitochondrial genomes. *Nucleic Acids Research*, 27(8), pp.1767–1780

Casa abierta al tiempo

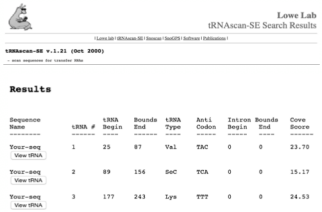
Anotación del Mitogenoma



Citation: *M. Bernt, A. Donath, F. Jühling, F. Externbrink, C. Florentz, G. Fritzsch, J. Pütz, M. Middendorf, P. F. Stadler*
MITOS: Improved de novo Metazoan Mitochondrial Genome Annotation
Molecular Phylogenetics and Evolution 2013, 69(2):313-319 link

Automatic annotation of organellar genomes with DOGMA. Wyman SK, Jansen RK, Boore JL, Bioinformatics 2004 20(17):3252-3255.

Identificación de los tRNA

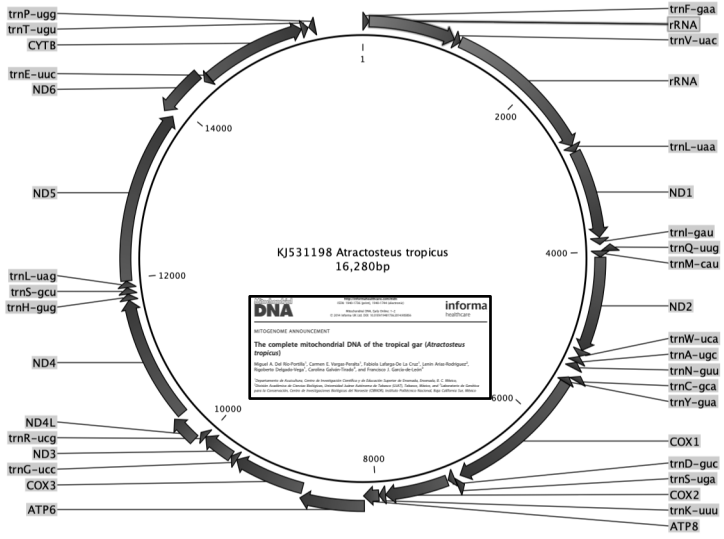


Rfam 12.0: updates to the RNA families database. Eric P. Nawrocki, Sarah W. Burge, Alex Bateman, Jennifer Daub, Ruth Y. Eberhardt, Sean R. Eddy, Evan W. Floden, Paul P. Gardner, Thomas A. Jones, John Tate and Robert D. Finn
Nucleic Acids Research (2014) 10.1093/nar/gku1063

Casa abierta al tiempo

Atractosteus tropicus

pejelagarto



Casa abierta al tiempo

Panopea
Almeja generosa

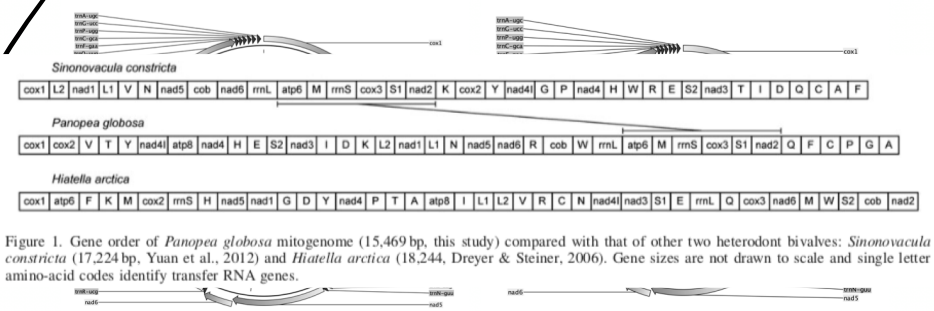


Figure 1. Gene order of *Panopea globosa* mitogenome (15,469 bp, this study) compared with that of other two heterodont bivalves: *Sinonovacula constricta* (17,224 bp, Yuan et al., 2012) and *Hiatella arctica* (18,244, Dreyer & Steiner, 2006). Gene sizes are not drawn to scale and single letter amino-acid codes identify transfer RNA genes.

Casa abierta al tiempo

Análisis genómico

Genes nucleares

Ejemplo

Análisis bioinformático de datos de secuenciación masiva para el diseño de marcadores microsatélites en dos especies de *Octopus* (Cephalopoda: Octopodidae)

Cecilia Daniela Garduño Hernández.
Asesores.
Dra. Irene de los Ángeles Barriga Sosa.
Dr. Miguel A. del Río Portilla

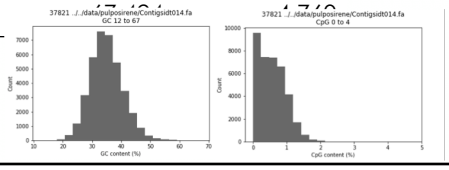
Casa abierta al tiempo

	O. mimus	***Octopus* sp.**	***O. bimaculoides***
Lecturas (Reads)	12,714,080	13,393,466	-
N75 (bp)	269	263	-
N50 (bp)	304	298	-
N25 (bp)	353	348	-
Mínimo (bp)	200	200	-
Máximo (bp)	15,078	10,449	-
Promedio (bp)	312	306	-
Contigs (n)	37,784	21,609	-
Total (bp)	11,778,014 (0.49%)	6,613,744 (0.28%)	2,383,190,000
pb entre microsatélite	3044	3353	1034

Casa abierta al tiempo

O. mimus

	Largo	**GC**	**CpG**
n	37821.000	37821.000	37821.000
Promedio	311.061	35.043	0.573
Desviación estandar	146.922	5.639	0.411
Mínimo	122.000	12.397	0.000
25%	259.000	31.222	0.234
50%	291.000	34.672	0.539
75%	331.000	38.484	0.846
Máximo	15078.000		



Casa abierta al tiempo

Blastn

5818 (15.38%)

Especies	n
cephalopods	5670
bonyfishes	35
bats	16
flatworms	14
even-toedungulates	12
nematodes	12
primates	10
bivalves	9
ants	6
crustaceans	5
otros	29

Casa abierta al tiempo

Microsatélites

- Son repeticiones de segmentos cortos de DNA:

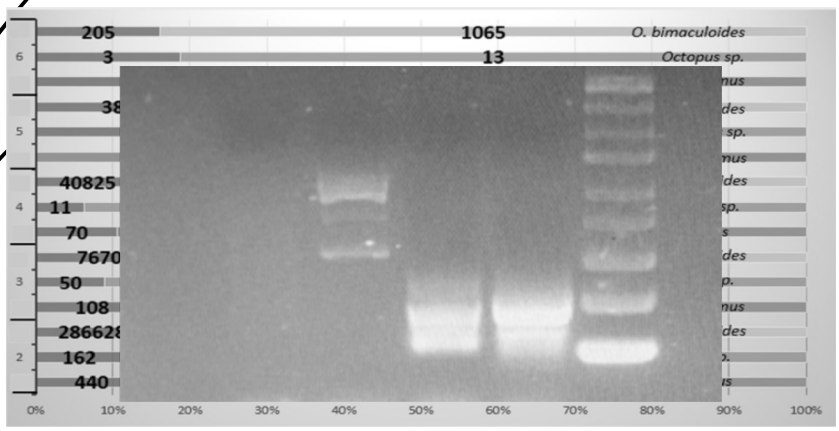
  GAGAGAGA.

- pueden ocurrir en diferentes número de repeticiones : $(GA)_4$, $(GA)_7$, $(GA)_8$, $(GA)_{30}$,
- Cada organismo tendrá dos copias.
- Por ejemplo,
  - ♂ $(GA)_4$ - $(GA)_7$
  - ♀ $(GA)_8$ - $(GA)_{30}$

	♀	
♂	**$(GA)_4$ $(GA)_7$**	
$(GA)_8$	$(GA)_4$-$(GA)_8$	$(GA)_7$-$(GA)_8$
$(GA)_{30}$	$(GA)_4$-$(GA)_{30}$	$(GA)_7$-$(GA)_{30}$

Casa abierta al tiempo

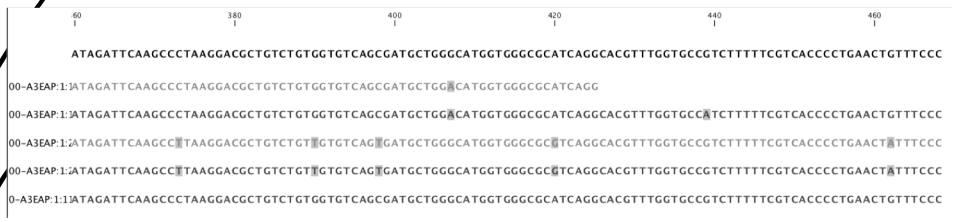
Distribución de microsatélites en *O. mimus*, *Octopus* sp. y *O. bimaculoides*



Casa abierta al tiempo

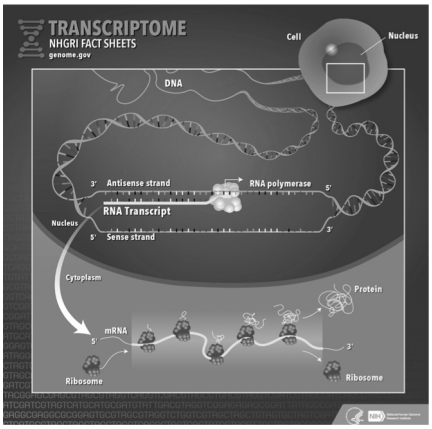
SNP

Polimorfismo de un solo nucleótido (Single Nucleotide Polymorphism)



Análisis transcriptómico

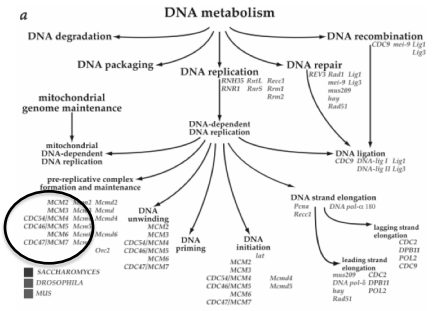
Casa abierta al tiempo



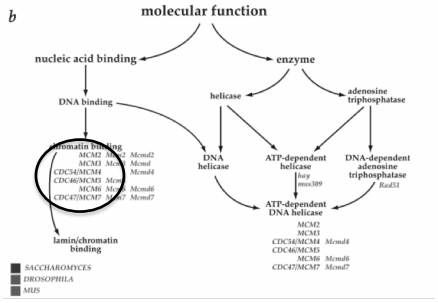
Anotación Ontología genética

Casa abierta al tiempo

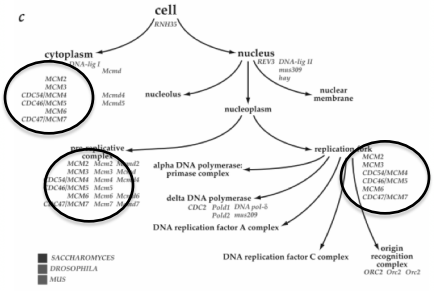
Procesos Biologicos



Función Biologica



Componentes celulares

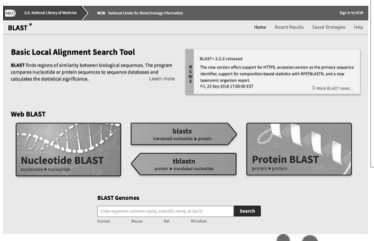


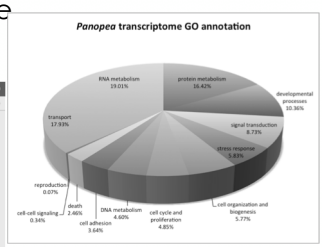
Casa abierta al tiempo

Anotación de los contigs: Blast, Uniprot

Identificar la localización o función de una secuencia de DNA

blast2go





*Fast*Annotator

Home | Analysis | Tutorial | Demo & Benchmark

NCBI BLAST

UniProt

Gene Ontology Consortium

Casa abierta al tiempo

Clasificación de los genes Ontología génica

Procesos Biológico

Función Molecular

Componentes celulares

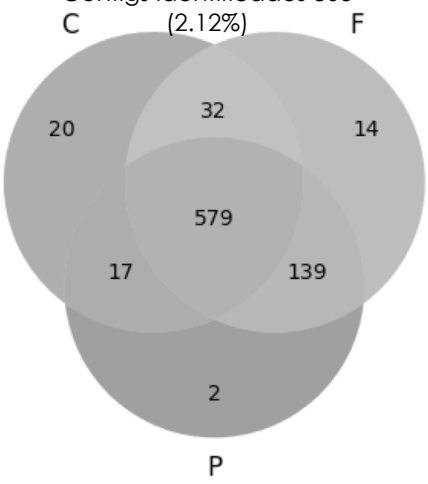
Casa abierta al tiempo

Blastx

Octopus mimus

Especies	n
cephalopods	5670
bonyfishes	35
bats	16
flatworms	14
even-toedungulates	12
nematodes	12
primates	10
bivalves	9
ants	6
crustaceans	5
otros	29

Contigs identificados 803 (2.12%)



Casa abierta al tiempo

Funciones biológicas

F	764
other molecular function	483
nucleic acid binding activity	229
signal transduction activity	19
transporter activity	17
enzyme regulator activity	13
cytoskeletal activity	3

Casa abierta al tiempo

Procesos biológicos

P	737
DNA metabolism	213
RNA metabolism	111
other metabolic processes	92
transport	69
protein metabolism	52
other biological processes	51
signal transduction	39
cell adhesion	30
developmental processes	22
cell organization and biogenesis	20
cell cycle and proliferation	17
stress response	16
death	5

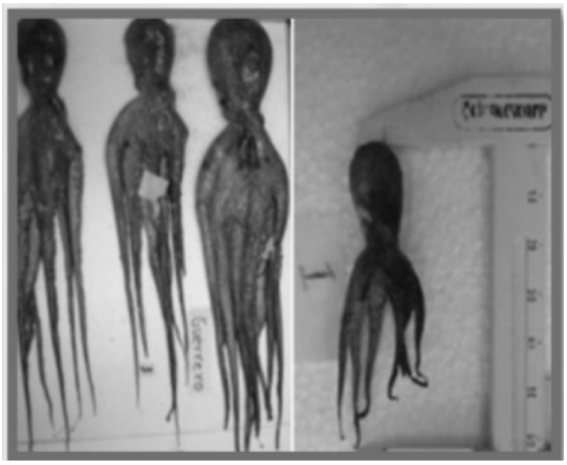
Casa abierta al tiempo

Componentes celulares

C	648
nucleus	313
other cellular component	145
other membranes	129
mitochondrion	26
non-structural extracellular	17
cytosol	4
ER/Golgi	4
other cytoplasmic organelle	4
cytoskeleton	3
plasma membrane	3

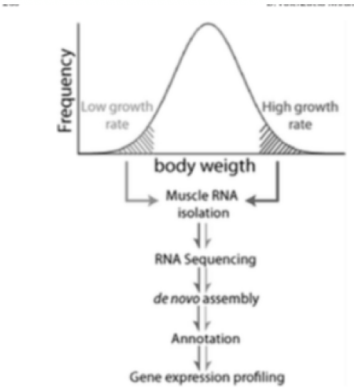
Casa abierta al tiempo

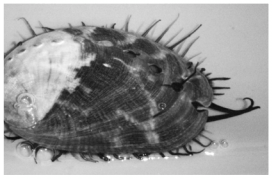
Octopus mimus
Octopus sp.



Casa abierta al tiempo

Anotación genética
Transcriptoma abulón





Marine Genomics

Genomics/Technical resource papers

Characterization of the growth-related transcriptome in California red abalone (*Haliotis rufescens*) through RNA-Seq analysis

Diego Valenzuela-Miranda*, Miguel A. Del Río-Portilla**, Cristian Gallardo-Escárate**

Fig. 1. Representation of experimental design. *H. rufescens* were cultured for six months and classified into low growth rate and high growth rate groups according to body weight. RNA-Seq was conducted in muscle from both groups, and differentially expressed genes were identified.

Casa abierta al tiempo

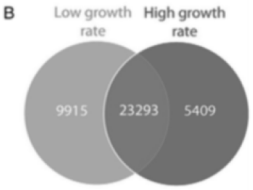
Anotación genética
Transcriptoma c

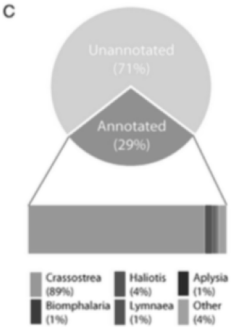
A	Sequencing	
	Low growth reads	783261
	High growth reads	764686
	Total	1547947
	***de novo* assembly**	
	Contigs	44.312
	Matched reads	1346817
	Not matched reads	201130
	N50	910
	Average length	790

Genomics/Technical resource papers

Characterization of the growth-related transcriptome in California red abalone (*Haliotis rufescens*) through RNA-Seq analysis

Diego Valenzuela-Miranda*, Miguel A. Del Río-Portilla**, Cristian Gallardo-Escárate**





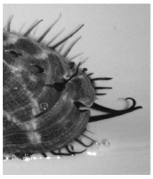
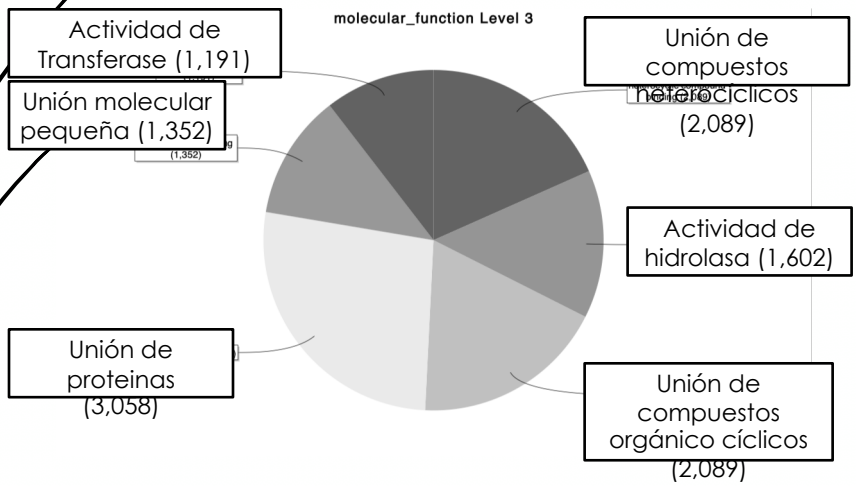


Fig. 2. Sequencing and annotation statistics. A) Number of reads yielded for each group and *de novo* assembly statistics. B) Venn diagram showing the distribution of contigs between low growth rate (Blue) and high growth rate (Red) individuals. C) Annotation statistics, including species distribution of annotated contigs.

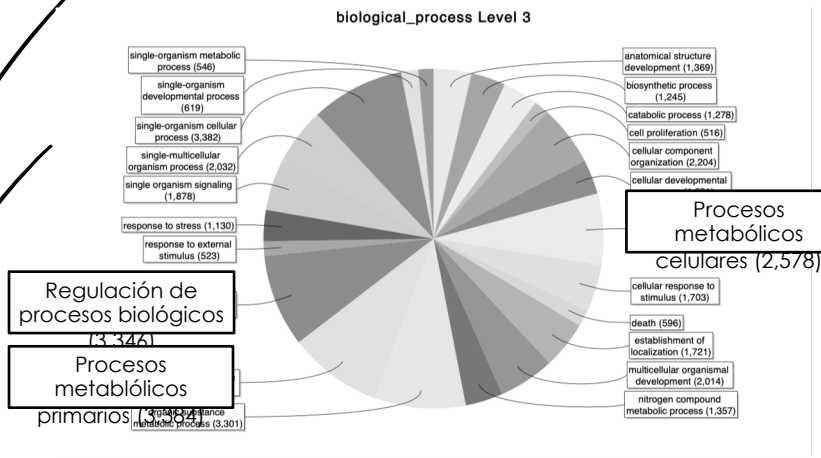
Casa abierta al tiempo

Función Biological



Casa abierta al tiempo

Procesos Biológicos

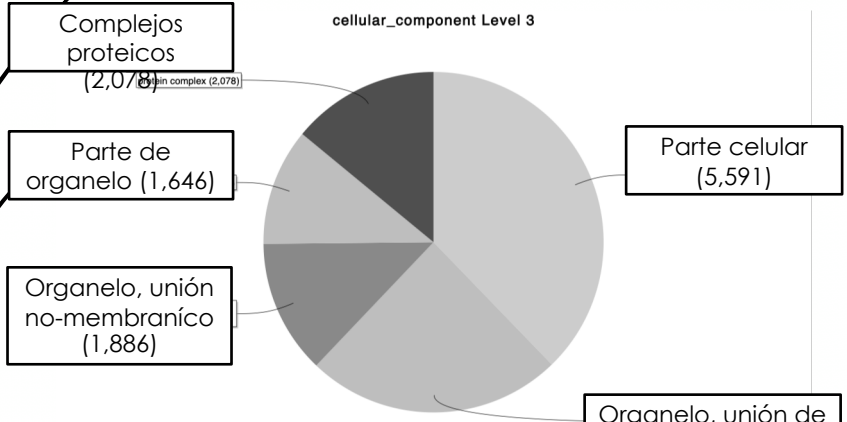


Procesos metabólicos celulares (2,578)

Regulación de procesos biológicos (3,346)

Procesos metablólicos primarios (3,164)

Componentes celulares



Complejos proteicos (2,078)

Parte de organelo (1,646)

Organelo, unión no-membraníco (1,886)

Parte celular (5,591)

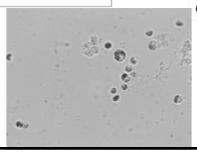
Organelo, unión de membrana (3,604)

Casa abierta al tiempo

Microalgas
Chaetoceros calcitrans

Tabla 1. Estadística básica del ensamblaje *De Novo* obtenidos al secuenciar parcialmente el genoma de *Chaetoceros calcitrans*.

Parámetro	Longitud (bases)
N75	1,737
N50	11,997
N25	48,155
Mínimo	200
Máximo	733,203
Promedio	2,153
Contigs	25,338
Total	55,620,496



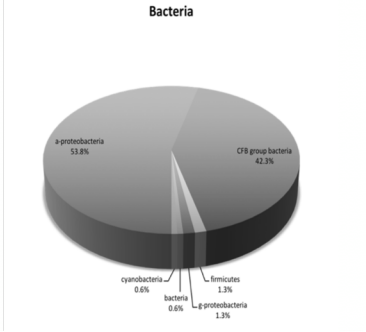


Fig.1 Distribución de la identificación de secuencias nucleotídicas (blastn) de los contigs relacionados con Bacterias resultado de la secuenciación genómica parcial de *C. calcitrans*

Casa abierta al tiempo

Microalgas
Chaetoceros calcitrans

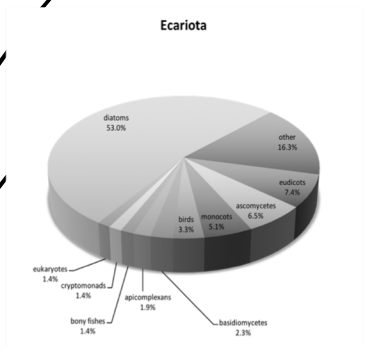


Fig. 2. Distribución de la identificación de secuencias nucleotídicas (blastn) de los contigs relacionados con Eucariotas resultado de la secuenciación genómica parcial de *C. calcitrans*.

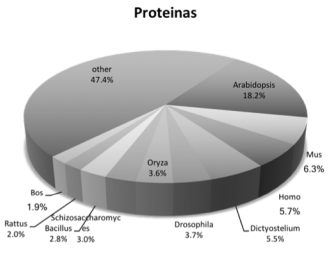
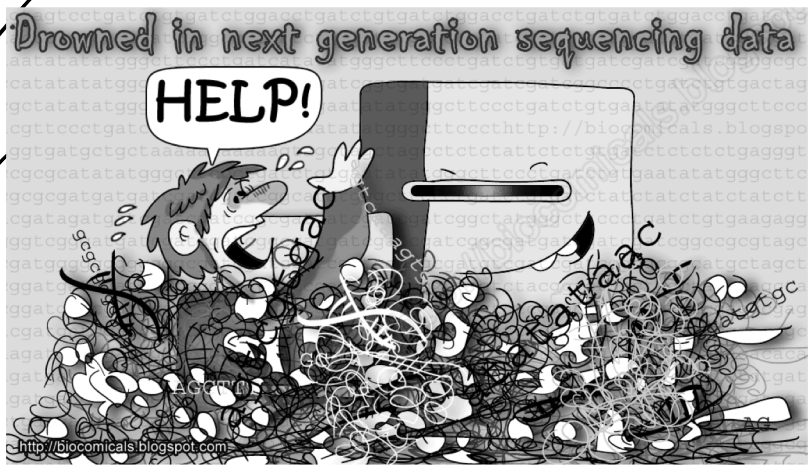


Fig. 3. Resultados de la búsqueda de proteínas (Blastx) de los contigs relacionados con secuencias de Eucariota obtenidos en la secuenciación genómica parcial de *C. calcitrans*.

Casa abierta al tiempo

Se obtiene una gran cantidad de datos



Casa abierta al tiempo

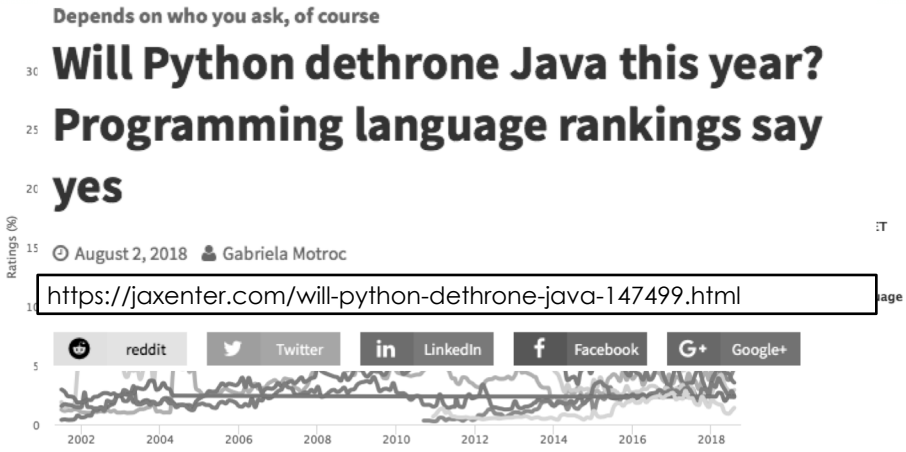
Bash

- Bash (Bourne-again shell) es un programa que sirve para interpretar las instrucciones en una línea de comandos.
- Se originón de palabras (born-again significa "nacido de nuevo") sobre la Bourne shell (sh), que fue uno de los primeros intérpretes importantes de Unix.

https://es.wikipedia.org/wiki/Bash

Casa abierta al tiempo

Lenguajes de programación



https://tiobe.com/tiobe-index

Casa abierta al tiempo

Bibiliotecas

- BioPERL
- BioPython
- BioRuby
- BioJava
- Bio-Conductor (R)

Casa abierta al tiempo

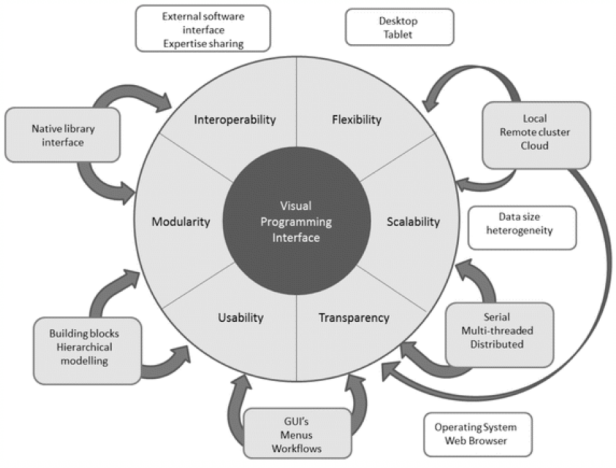
¿Por qué es importante documentar?

- Uso de bitácoras electrónicas
- " In fact, many journals will now require you to publish your source code along with a manuscript." (Dudley and Butte, 2009)

Dudley, J.T., Butte, A.J., 2009. A quick guide for developing effective bioinformatics programming skills. PLoS Comput. Biol. 5, e1000589. doi: 10.1371/journal.pcbi.1000589

Casa abierta al tiempo

Programas para NGS



Milicchio, F., Rose, R., Bian, J., Min, J., Prosperi, M., 2016. Visual programming for next-generation sequencing data analytics. BioData Min. 9, 16. doi:10.1186/s13040-016-0095-3

Casa abierta al tiempo

- · The Importance of Building Your Technology Toolbox
- · The Benefits and Opportunities of Open Source Communities
- · The Importance of UNIX Skills
- · Keeping Projects Documented and Manageable
- · Preserving Your Source Code
- · Embracing Parallel Computing Paradigms
- · Structuring Data for Speed and Scalability
- · Understand the Capabilities of Hardware
- · Embracing Standards and Interoperability
- · Value Your Time

Dudley, J.T., Butte, A.J., 2009. A quick guide for developing effective bioinformatics programming skills. PLoS Comput. Biol. 5, e1000589. doi: 10.1371/journal.pcbi.1000589

Casa abierta al tiempo

- Desarrollo de programas
- "Consequently it is no surprise that many successful bioinformatics apps are written by biologists who lack formal computer science training, · as they un · doubtedly put scientific utility ahead of architectural elegance and completeness."

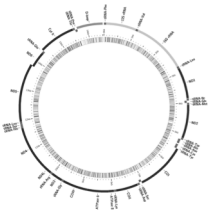
Dudley, J.T., Butte, A.J., 2009. A quick guide for developing effective bioinformatics programming skills. PLoS Comput. Biol. 5, e1000589. doi: 10.1371/journal.pcbi.1000589

Diagrama de flujo NGS

Casa abierta al tiempo

Toma de muestra → Extracción del DNA → Secuenciación NGS → Calidad → Recorte de baja calidad → *Ensamblaje De novo* → Contigs (más de tres lecturas y 1.5 de cobertura → Mapping (coverage) → Anotación de los contigs

Singletons (lecturas solas)



Galaxy

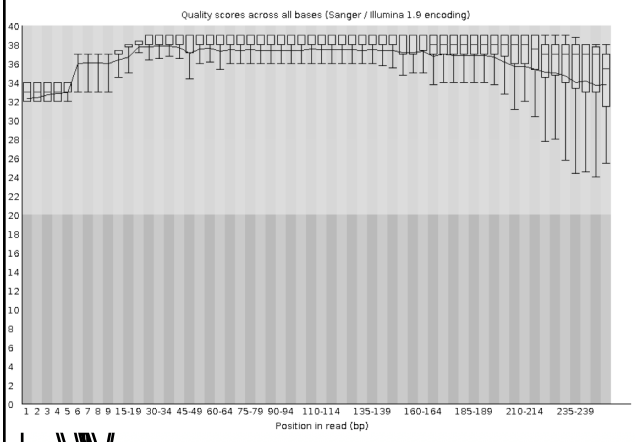
Unicycler

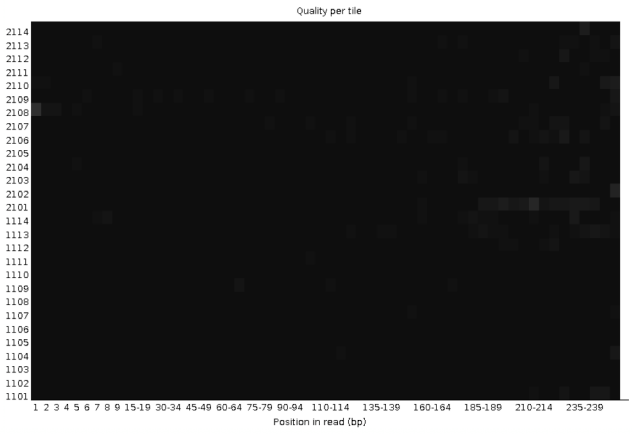
CLC Genomics Workbench

Quality check
De novo assembly
Mapping
CLC bio

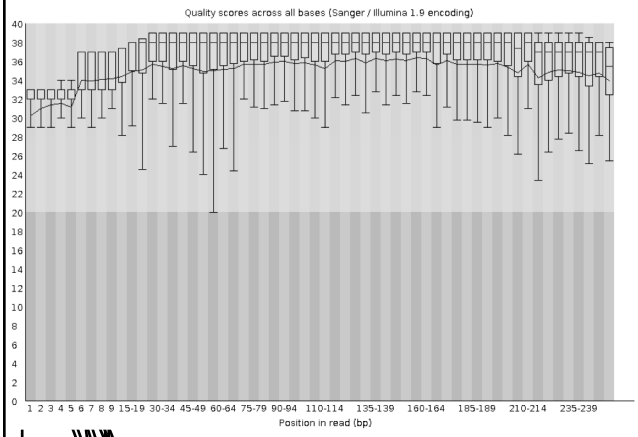
Calidad Fastqc

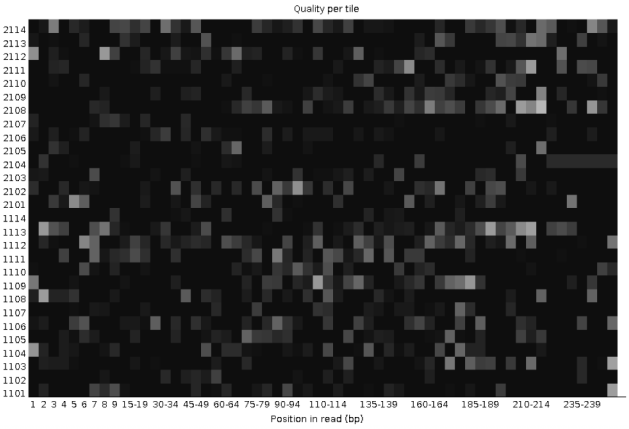
Casa abierta al tiempo



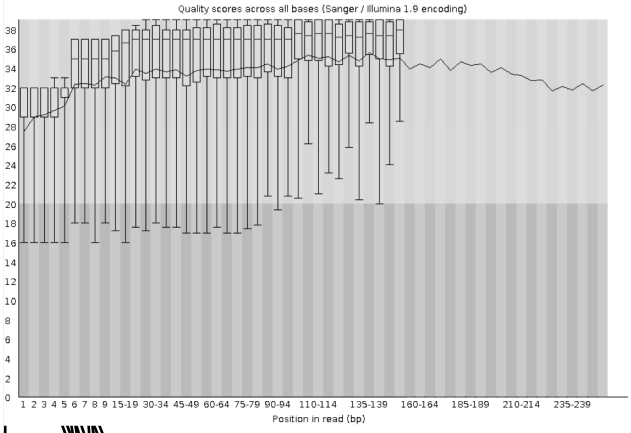


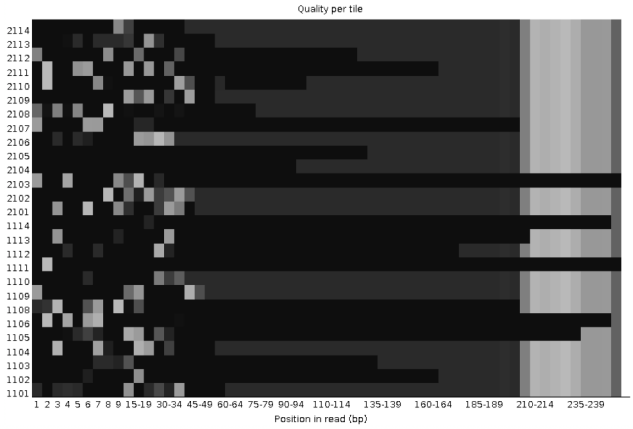
Casa abierta al tiempo



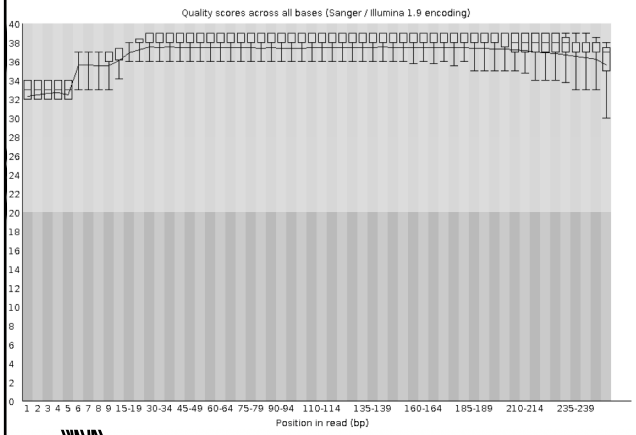


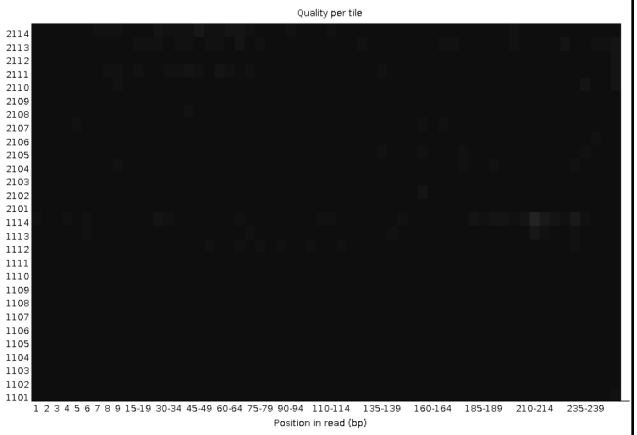
Casa abierta al tiempo





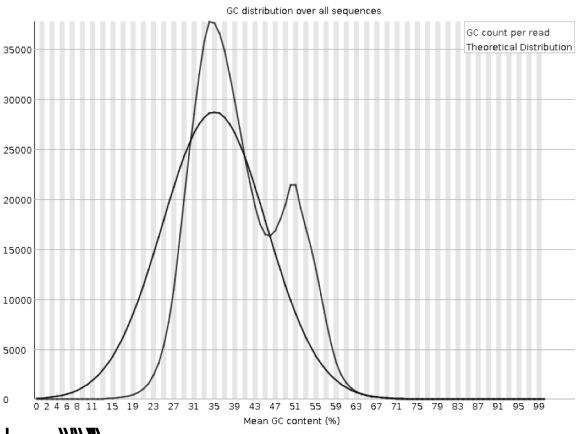
Casa abierta al tiempo

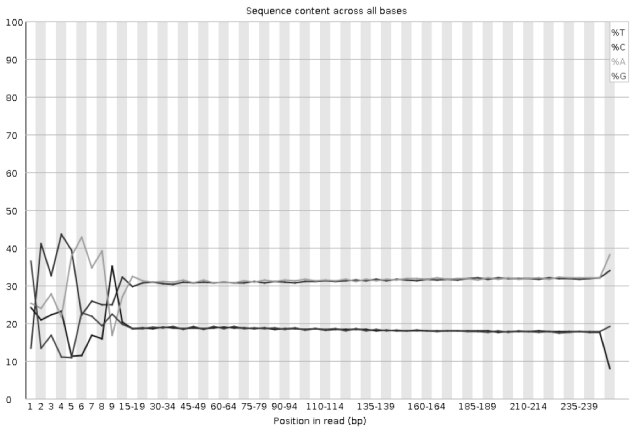




Casa abierta al tiempo

Distribución de GC y por base





Casa abierta al tiempo

Metageómica

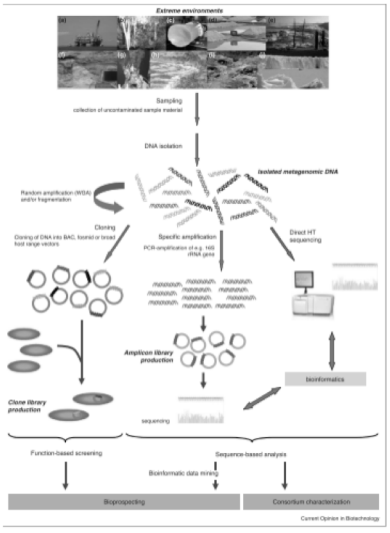
Available online at www.sciencedirect.com
SciVerse ScienceDirect
Current Opinion in Biotechnology

Metagenomics of microbial life in extreme temperature environments

Anna Lewin[1], Alexander Wentzel[2] and Svein Valla[1]

Doi: 10.1016/j.copbio.2012.10.012

Figure 1

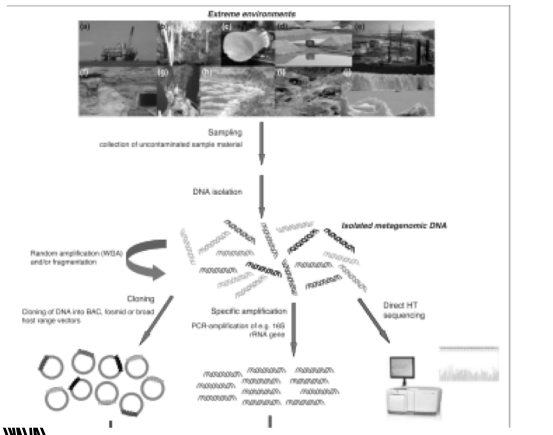


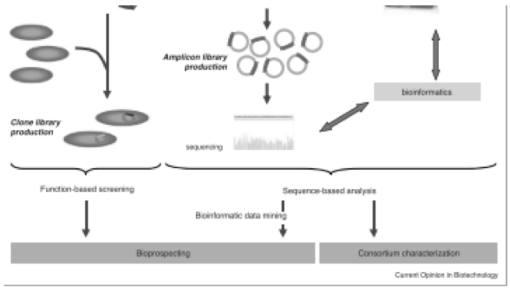
Schematic overview of metagenomic study methodology. As for the scope of this review, the habitats sampled are illustrated by extreme environments: **(a)** offshore oil reservoir (photo: http://en.wikipedia.org/wiki/Oil_platform), **(b)** acidic snottites (photo: NASA), **(c)** ice core (photo: NEEM ice core drilling project; Sepp Kipfstuhl), **(d)** saltern crystalliser ponds (photo: Marisol[1] Sea Salt, Böer & Siebert, Lda.), **(e)** YNP hot spring (photo: TheBrocklehurstGlory (http://en.wikipedia.org/wiki/File: Dead_trees_at_Mammoth_Hot_Springs.jpg)), **(f)** mine (photo: http://en.wikipedia.org/wiki/File: UraniumMineUtah.JPG), **(g)** black smoker chimney (photo: WHOI), **(h)** the Dead Sea (photo: http://en.wikipedia.org/wiki/File:DeadSea3667.jpg), **(i)** AMD (photo: NASA) and **(j)** glacier ice (photo: ©: Secretariat of Tourism Buenos Aires). Sequencing is illustrated by a Roche 454 Gene sequencer (photo: 454 Sequencing©: Roche Diagnostics) and/or a sequencing chromatogram (picture University of Utah (http://www.cores.utah.edu)).

Casa abierta al tiempo

Metageómica

Figure 1



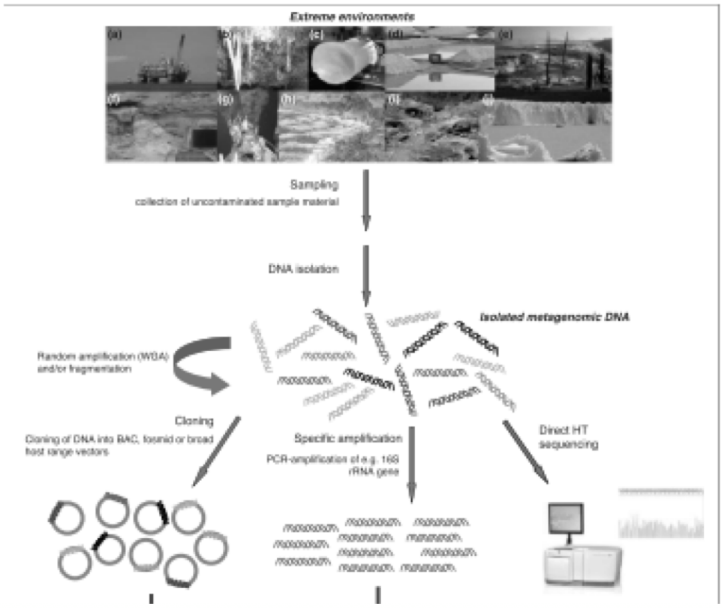


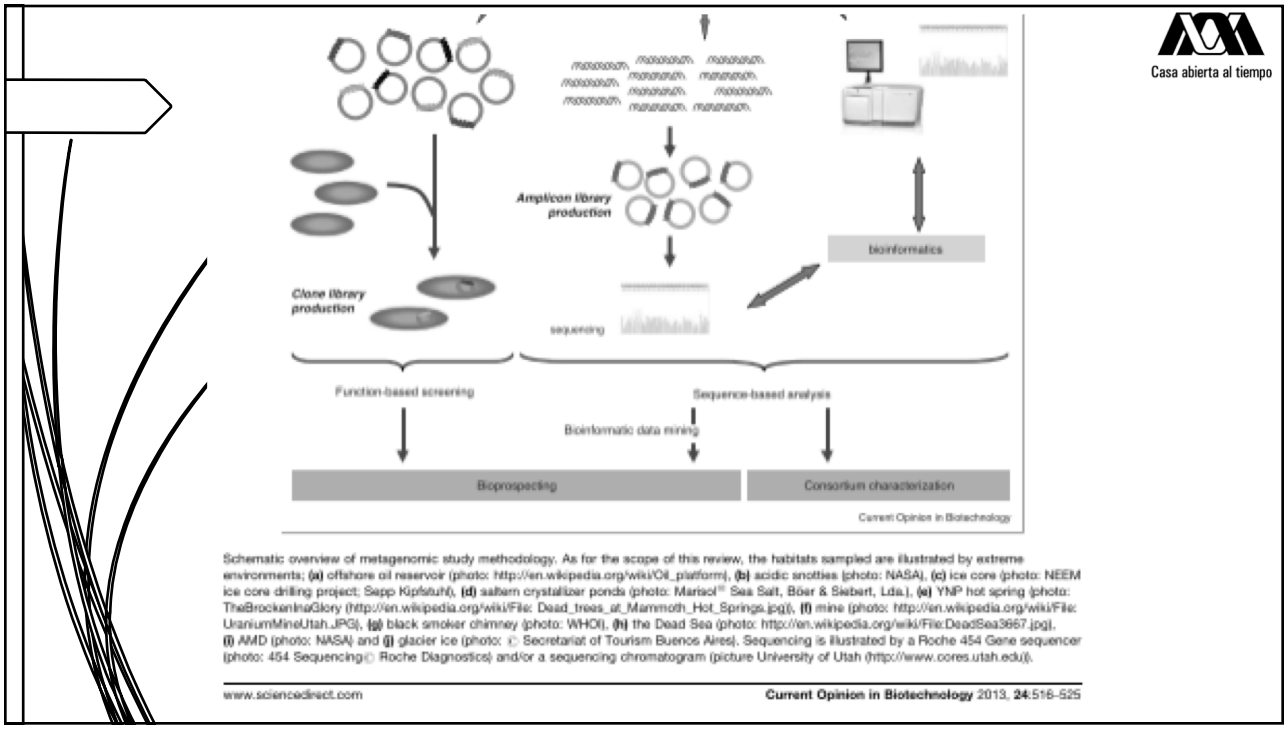
Schematic overview of metagenomic study methodology. As for the scope of this review, the habitats sampled are illustrated by extreme environments; **(a)** offshore oil reservoir (photo: http://en.wikipedia.org/wiki/Oil_platform), **(b)** acidic snottites (photo: NASA), **(c)** ice core (photo: NEEM ice core drilling project; Sepp Kipfstuhl), **(d)** saltern crystallizer ponds (photo: Marisol[©] Sea Salt, Böer & Siebert, Ldta.), **(e)** YNP hot spring (photo: TheBrockenInaGlory (http://en.wikipedia.org/wiki/File: Dead_trees_at_Mammoth_Hot_Springs.jpg)), **(f)** mine (photo: http://en.wikipedia.org/wiki/File: UraniumMineUtah.JPG), **(g)** black smoker chimney (photo: WHOI), **(h)** the Dead Sea (photo: http://en.wikipedia.org/wiki/File:DeadSea3667.jpg), **(i)** AMD (photo: NASA) and **(j)** glacier ice (photo: © Secretariat of Tourism Buenos Aires). Sequencing is illustrated by a Roche 454 Gene sequencer (photo: 454 Sequencing© Roche Diagnostics) and/or a sequencing chromatogram (picture University of Utah (http://www.cores.utah.edu)).

Casa abierta al tiempo

Metagenomics of microbial life in extreme temperature environments Lewin, Wentzel and Valla 517

Figure 1





Schematic overview of metagenomic study methodology. As for the scope of this review, the habitats sampled are illustrated by extreme environments; **(a)** offshore oil reservoir (photo: http://en.wikipedia.org/wiki/Oil_platform), **(b)** acidic snottites (photo: NASA), **(c)** ice core (photo: NEEM ice core drilling project; Sepp Kipfstuhl), **(d)** saltern crystallizer ponds (photo: Marisol® Sea Salt, Böer & Siebert, Lda.), **(e)** YNP hot spring (photo: TheBrockenInaGlory (http://en.wikipedia.org/wiki/File: Dead_trees_at_Mammoth_Hot_Springs.jpg)), **(f)** mine (photo: http://en.wikipedia.org/wiki/File: UraniumMineUtah.JPG), **(g)** black smoker chimney (photo: WHOI), **(h)** the Dead Sea (photo: http://en.wikipedia.org/wiki/File:DeadSea3667.jpg), **(i)** AMD (photo: NASA) and **(j)** glacier ice (photo: © Secretariat of Tourism Buenos Aires). Sequencing is illustrated by a Roche 454 Gene sequencer (photo: 454 Sequencing© Roche Diagnostics) and/or a sequencing chromatogram (picture University of Utah (http://www.cores.utah.edu)).

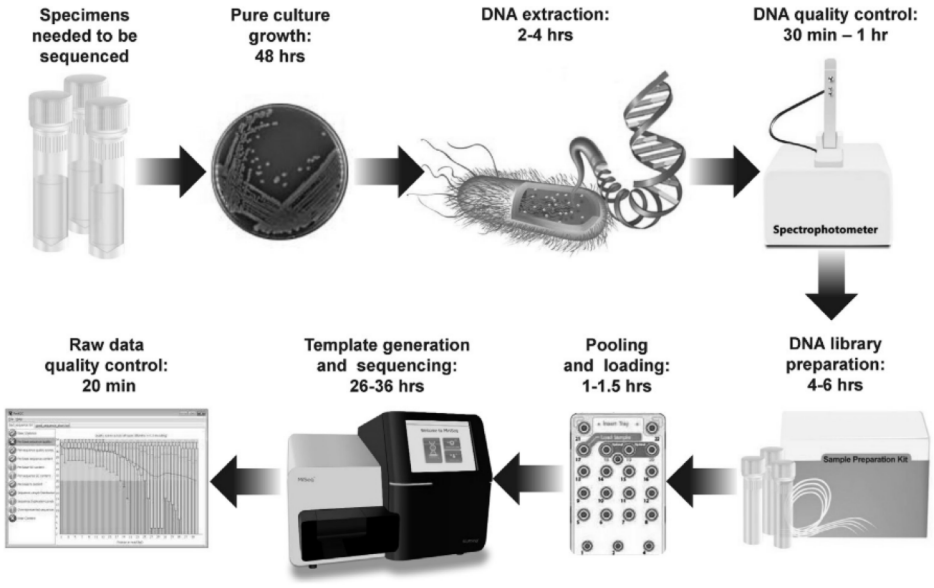


Fig. 2. Typical whole genome sequencing workflow in a clinical or public health laboratory.

J. Besser et al. / Clinical Microbiology and Infection 24 (2018) 335–341

Table 2
Characteristics, strengths and weaknesses of commonly used sequencing platforms

Platform \ Instrument	Throughput range (Gb)[a]	Read length (bp)	Strength	Weakness
Sanger sequencing				
ABI 3500/3730	0.0003	Up to 1 kb	Read accuracy and length	Cost and throughput
Illumina				
MiniSeq	1.7–7.5	1×75 to ×150	Low initial investment	Run and read length
MiSeq	0.3–15	1×36 to 2×300	Read length, scalability	Run length
NextSeq	10–120	1×75 to 2×150	Throughput	Run and read length
HiSeq (2500)	10–1000	×50 to ×250	Read accuracy, throughput,	High initial investment, run
NovaSeq 5000/6000	2000–6000	2×50 to ×150	Read accuracy, throughput	High initial investment, run
IonTorrent				
PGM	0.08–2	Up to 400	Read length, speed	Throughput, homopolymers[c]
S5	0.6–15	Up to 400	Read length, speed,	Homopolymers[c]
Proton	10–15	Up to 200	Speed, throughput	Homopolymers[c]
Pacific BioSciences				
PacBio RSII	0.5–1[b]	Up to 60 kb	Read length, speed (Average 10 kb, N50 20 kb)	High error rate and initial
Sequel	5–10[b]	Up to 60 kb	Read length, speed (Average 10 kb, N50 20 kb)	High error rate
Oxford Nanopore				
MInION	0.1–1	Up to 100 kb	Read length, portability	High error rate, run length,

[a] The throughput ranges are determined by available kits and run modes on a per run basis. As an example of a 15-GB throughput, thirty-five 5-MB genomes can be sequenced to a minimum coverage of 40× on the Illumina MiSeq using the v3 600 cycle chemistry.
[b] Per one single-molecule real-time cell.
[c] Results in increased error rate (increased proportion of reads containing errors among all reads) which in turn results in false-positive variant calling.

Doi: 10.1016/j.cmi.2017.10.013