

Resumen sesión 4: Anotación de una secuencia

- Con la herramienta BLAT podemos buscar las coordenadas cromosómicas de una secuencia anónima dentro de un ensamblado de referencia
- ¡Las coordenadas cromosómicas carecen de sentido sin la información a que ensamblado pertenecen!
- Un ensamblado es una versión concreta de un genoma.
- Las coordenadas por regla general cambiarán de un ensamblado a otro → ¡la misma secuencia tendrá coordenadas cromosómicas diferentes!

Contenido sesión 5

Estructura composicional de los genomas: isocoras

- Detección experimental y definición
- Detección computacional
- Relevancia biológica

Contenido del genoma

- Elementos funcionales: genes, TFBs, RNA no codificante (microRNAs, siRNAs), islas CpG
- DNA Repetido: Transposones, LTR, DNA satélite

Variación en el genoma

- SNPs (Polimorfismo de una sola base)
- CNVs (Copy number variation)
- Indels (inserciones o deleciones cortas)

Estructura composicional: las isocoras

G. Bernardi / *Gene* 241 (2000) 3–17

Estructura composicional: propiedades derivadas de la secuencia. No se refiere a propiedades citogenéticas

Isocoras:

tramos largos de ADN ($\gg 300$ kb)
'relativamente' homogéneos
composicionalmente.

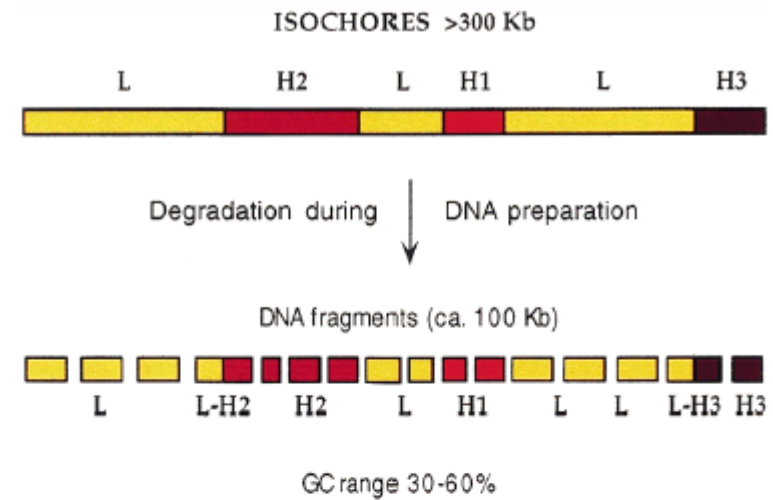
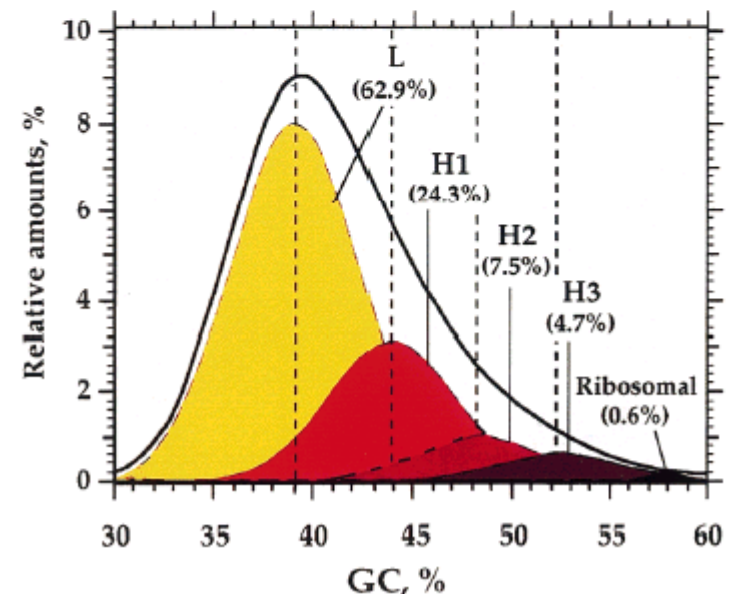


Fig. 1. (Top) Scheme of the isochore organization of the human genome. This genome, which is typical of the genome of most mammals, is a mosaic of large ($\gg 300$ kb, on average) DNA segments, the isochores, which are compositionally homogeneous (above a size of 3 kb) and can be partitioned into a number of families. Isochores are degraded during routine DNA preparations to fragments of approx. 100 kb in size. The GC-range of the isochores from the human genome is 30–60% (from Bernardi, 1995). (Bottom) The CsCl profile of human DNA is resolved into its major DNA components, namely the families of DNA fragments derived from isochore families L (i.e., L1+L2), H1, H2, H3. Modal GC levels of isochore families are indicated on the abscissa (broken vertical lines). The relative amounts of major DNA components are indicated. Satellite DNAs are not represented (from Zoubak et al., 1996).



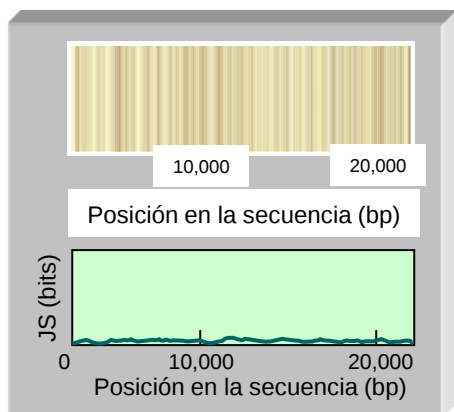
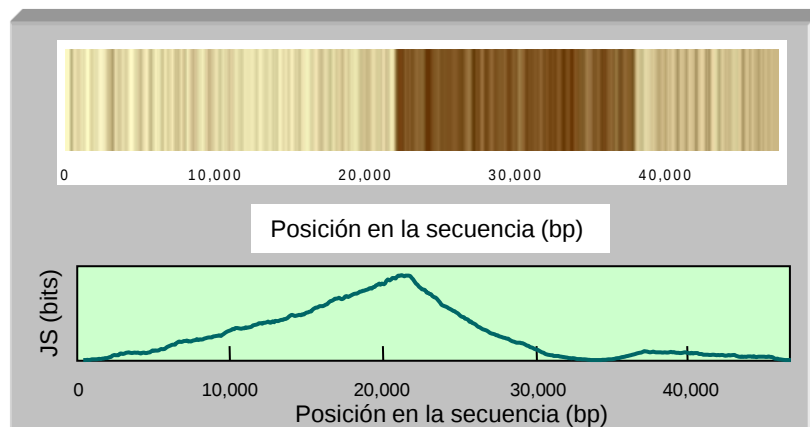
Clasificación

Existen diferentes clasificaciones de isocoras, pero
En todas se distingue entre isocoras L (light) y H (heavy)

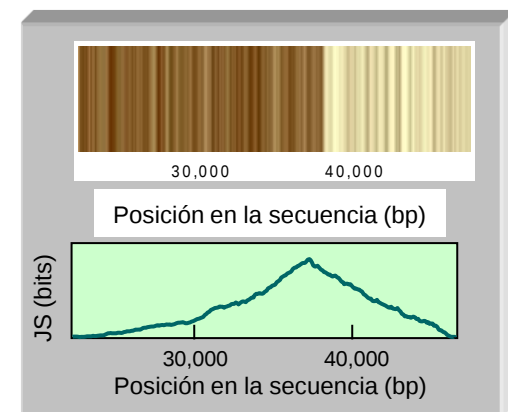
Tabla 1-1: Clasificación de isocoras utilizada por Pavlicek *et al.* (2002)

Isocora	Rangos de G+C
L1	$G+C < 37\%$
L2	$37\% \leq G+C < 41\%$
H1	$41\% \leq G+C < 46\%$
H2	$46\% \leq G+C < 52\%$
H3	$G+C > 52\%$

Detección computacional de las isocoras – segmentación



Nivel de Significación
Probabilidad de que
el corte **NO** se
deba al azar



Resultado final



IsoFinder: chr22

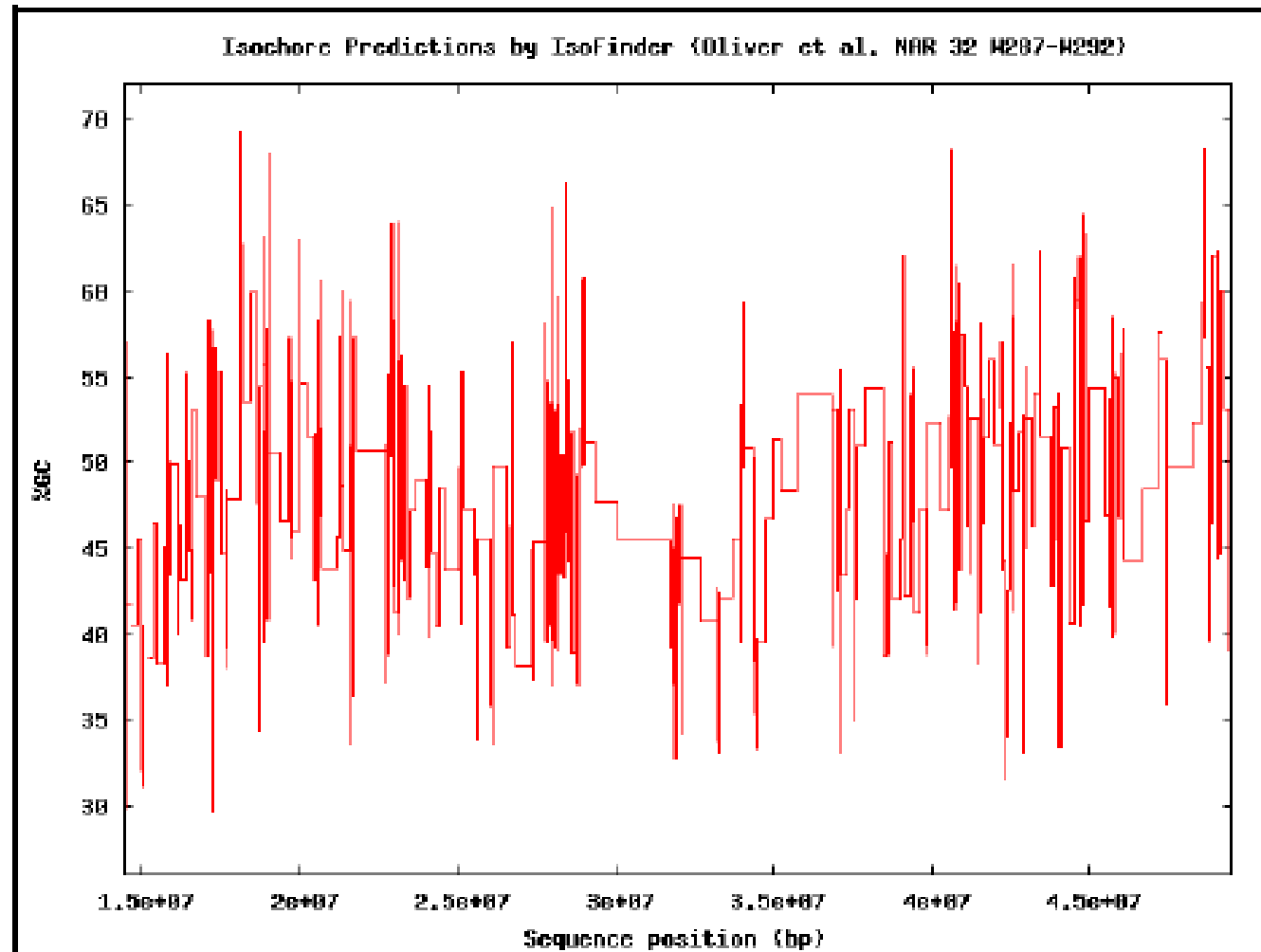


Figura 1-5: Representación gráfica de la segmentación del cromosoma 22 generada mediante la interfaz web del programa IsoFinder (Oliver *et al.*, 2004). Se puede apreciar la variación pronunciada del contenido en G+C entre aproximadamente el 35 y el 65%.

Relevancia biológica de las isocoras

- Correlación con las bandas cromosómicas G(rico en A+T)/R (rico en G+C)
- Densidad de genes: mas alto en isocoras H
- Densidad de islas CpG y transposones (Alus, LINEs)
- Sesgo del uso de codones
- Sesgo mutacional
- Replicación (momento de replicación depende del G+C)
- Frecuencia de recombinación
- Estructura génica (longitudes de genes e intrones)
- Observan principalmente en animales de sangre caliente
- ¿Patrones de metilación?
- ¿Especificidad tisular de la expresión génica?

Como se generan, como se mantienen

Se puede asumir que una estructura composicional como las isocoras no se deben al azar

- Uno o mas mecanismos moleculares y/o evolutivos dejan su impronta en la composición del genoma
- ¿Cómo se generan y como se mantienen las isocoras frente a la presión mutacional hacia la riqueza en A+T?

Sesgo mutacional

- La probabilidad de una incorporación errónea de una base durante la replicación depende de la concentración de nucleótidos libres
- La disponibilidad de nucleótidos libres varia a lo largo del ciclo celular
- Algunas partes del genoma se replican temprano y otras tarde
- → regiones que se replican en momentos diferentes tendrán patrones de mutaciones diferentes
- Desemparejamientos están preferentemente reparados hacia GC

Como se generan, como se mantienen

Selección

- Isocoras ocurren principalmente en animales de sangre caliente
 - enlaces de G+C son termodinámicamente mas estables
 - Los genes se encuentran fuertemente acumulados en isocoras H (ricas en G+C)
 - Selección a favor de mantener el G+C

Conversión génica sesgada

- Durante la recombinación homóloga se puede formar un desemparejamiento en el heterodúplex debido a la heterocigocidad de un locus
- En mamíferos, los desemparejamientos se reparan sesgadamente hacia GC
- También existe un sesgo en las tasas de recombinación en función del G+C
- → se podrían formar las isocoras

Publicaciones relevantes acerca de las isocoras

Bernardi G. Misunderstandings about isochores. Part 1. Gene. 2001 Oct 3;276(1-2):3-13. (<http://www.ncbi.nlm.nih.gov/pubmed/11591466>)

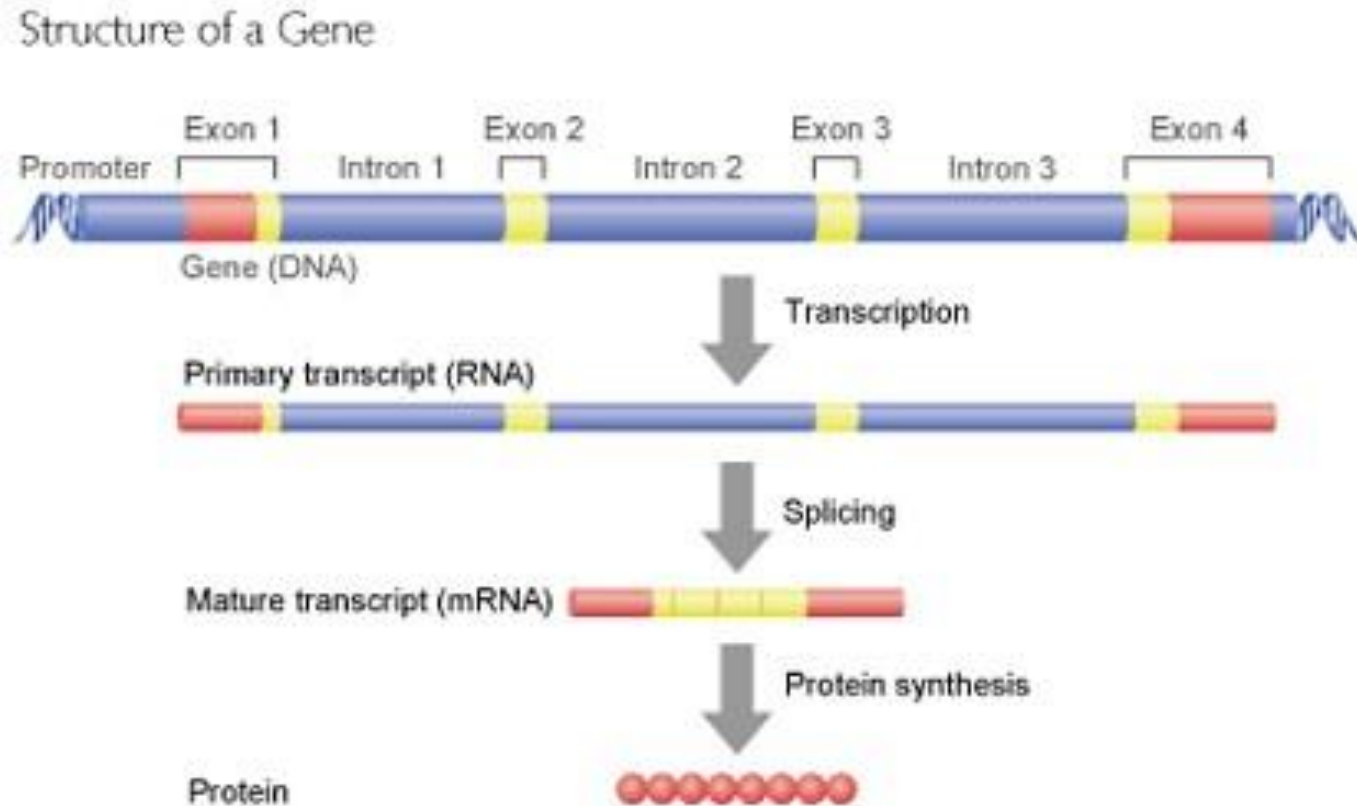
Eyre-Walker A, Hurst LD. The evolution of isochores. Nat Rev Genet. 2001 Jul;2(7):549-55. (<http://www.ncbi.nlm.nih.gov/pubmed/11433361>)

Wolfe, K. H., P. M. Sharp, and W.-H. Li. 1989. Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Maria Costantini and Giorgio Bernardi. Replication timing, chromosomal bands, and isochores. PNAS March 4, 2008 vol. 105 no. 9 3433-3437

Elementos genómicos: los genes

La estructura de un gen



Elementos genómicos: los genes

Existen muchas anotaciones diferentes de genes

NCBI ([National Center for Biotechnology Information](#)): [RefSeq](#) genes: Una colección no redundante de secuencias genómicas, transcritos y proteínas.

Por ejemplo, para el gen con simbolo oficial (HGNC – HUGO Gene Nomenclature Commite) TP53 (tumor protein p53) tenemos:

genómico (el gen y todos sus isoformas) [NG 017013](#)

mRNA(transcrito): [NM 000546.4](#) corresponde a la proteína: [NP 000537.3](#)

Ensembl: basado tanto [en anotación automática como en curación manual](#)

El identificador de TP53 en Ensembl es [ENSG00000141510](#)

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
TP53-001	ENST00000269305	2579	ENSP00000269305	393	Protein coding	CCDS11118
TP53-002	ENST00000445888	2506	ENSP00000391478	393	Protein coding	CCDS11118
TP53-003	ENST00000503591	565	ENSP00000426252	128	Protein coding	-
TP53-004	ENST00000455263	2580	ENSP00000398846	346	Protein coding	CCDS45605

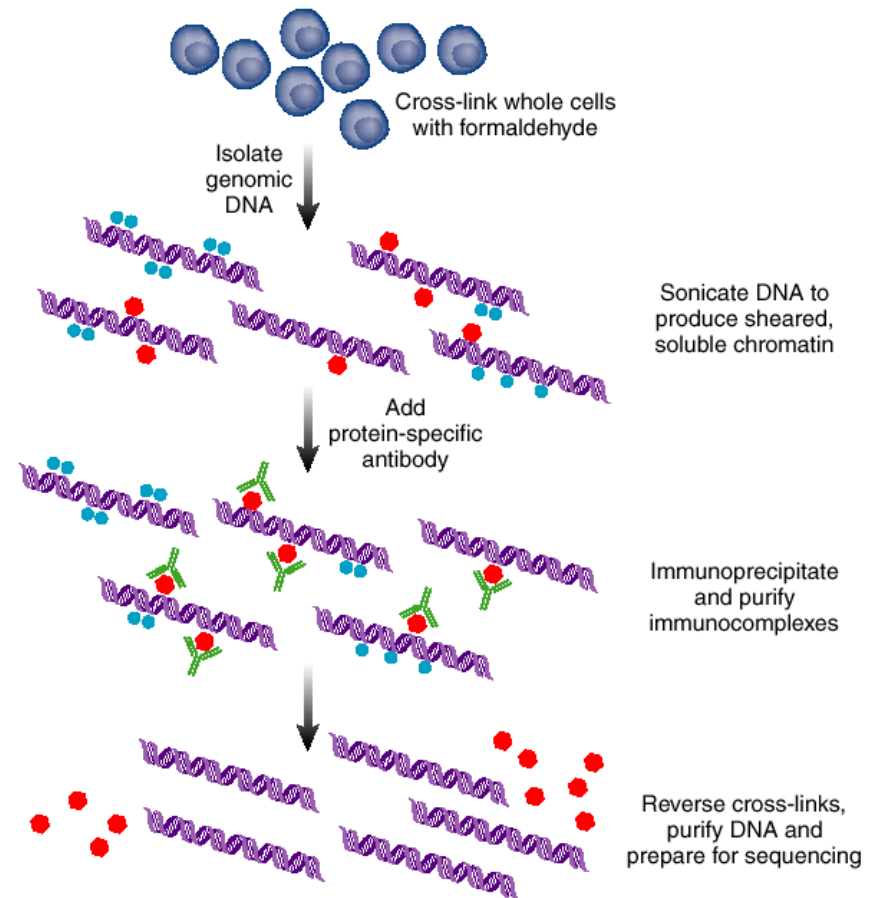
Elementos genómicos: los TFBS

TFBS: Transcription factor binding side (sitios de unión de un factor de transcripción)

- Los factores de transcripción se unen a una secuencia específica
- Regulan de expresión génica
- Implicados en la de metilación de DNA

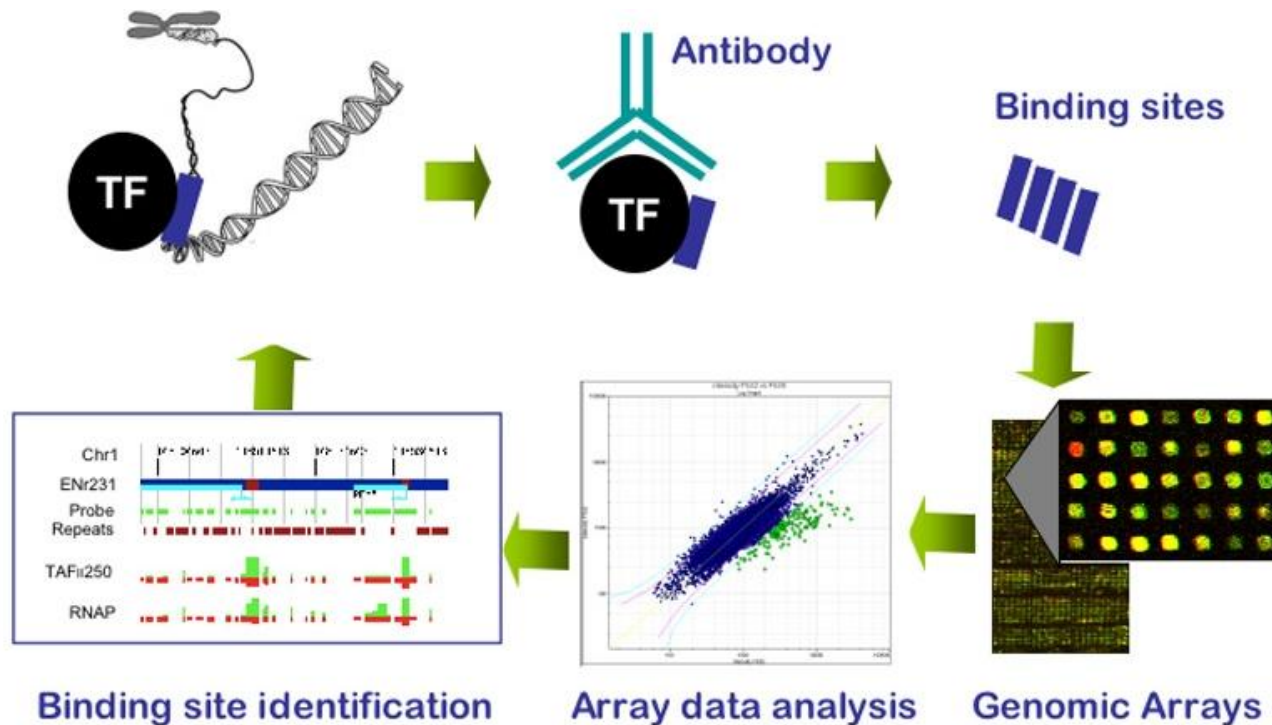
Detección experimental

1. chromatin immunoprecipitation (extraer /enriquecer secuencias de DNA a la que se une la proteína)
2. Detectar las secuencias a las que se une la proteína en el genoma



ChIP-on-Chip: combina ChIP (chromatin immunoprecipitation) & chip (microarrays)

Aparicio et al. (2004) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr Protoc Cell Biol*.



TFBS: ChIP-seq

ChIP-seq:

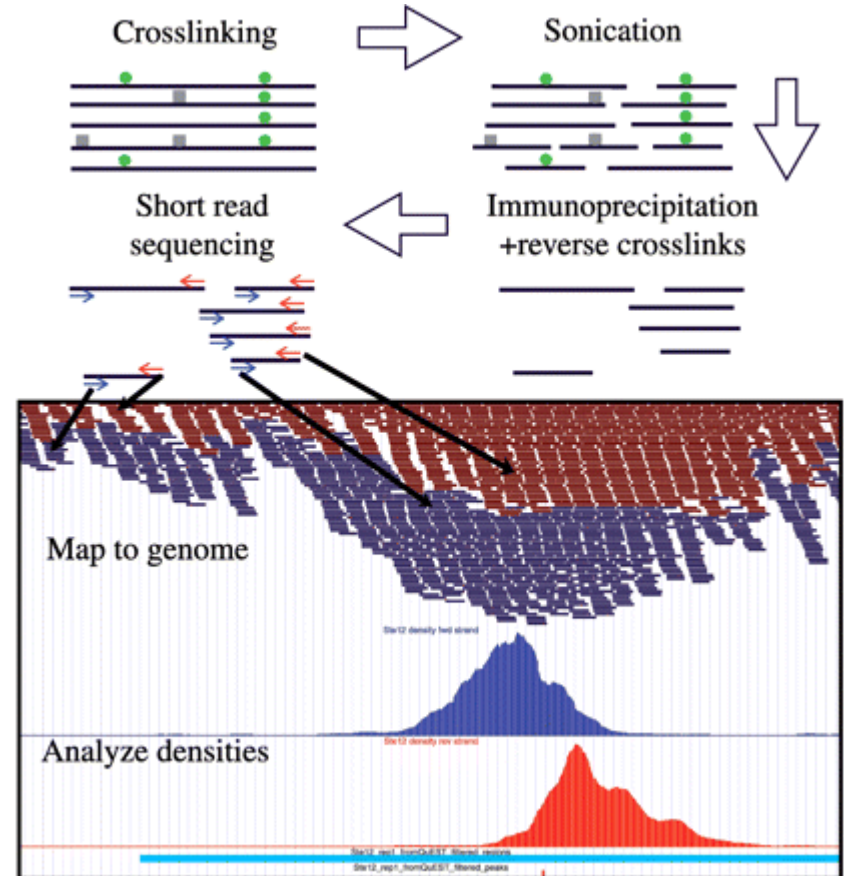
combina ChIP (chromatin immunoprecipitation) & secuenciación masiva

Elaine R Mardis

ChIP-seq: welcome to the new frontier

Nature Methods - 4, 613 - 614
(2007)

doi:10.1038/nmeth0807-613



TFBS: Representación y bases de datos

Representación

- Coordinadas cromosómicas de las regiones detectadas
- Secuencia consenso del motif (GCGANRAA)
- Matriz de peso por posición (Position weight matrix)

Bases de datos

Jasper: <http://jaspar.cgb.ki.se/>: contiene un conjunto no-redundante de motifs y matrices

	ID	name	species	class	family	Sequence logo
☒	MA0008.1	HAT5	<i>Arabidopsis thaliana</i>	Helix-Turn-Helix	Homeo	

cisRED: (www.cisred.org). Elementos regulatorios conservados en secuencias de los promotores (coordinadas cromosómicas)

ORegAnno (<http://www.oreganno.org/oreganno/>): base de datos curado que contiene las coordenadas de elementos regulatorios obtenidos de la literatura y de experimentos

Elementos genómicos: las islas CpG

En el genoma humano solo se encuentra el 20% de los dinucleótidos CpG esperados

→ El 80% de los CpGs están metilados → desaminación de citosinas → CpG muta a TpG

→ Los CpGs permanecen solo en sitios que no se metilan → las islas CpG

Propiedades de las islas CpG

- Ricas en G+C y CpG con longitudes alrededor de 1 kb
- Ratios O/E (número de CpG observado / esperado) altos
- Frecuentemente localizados en la región promotora de los genes → alrededor del 70% de todos los genes tienen una isla CpG asociada
- Prácticamente todos los genes del tipo “*housekeeping*” (se expresan en todos los tejidos), pero solo aprox. la mitad de los genes específicos de un tejido tienen una isla CpG

Existen tanto métodos experimentales como computacionales para detectar islas CpG

islas CpG: ventanas deslizantes

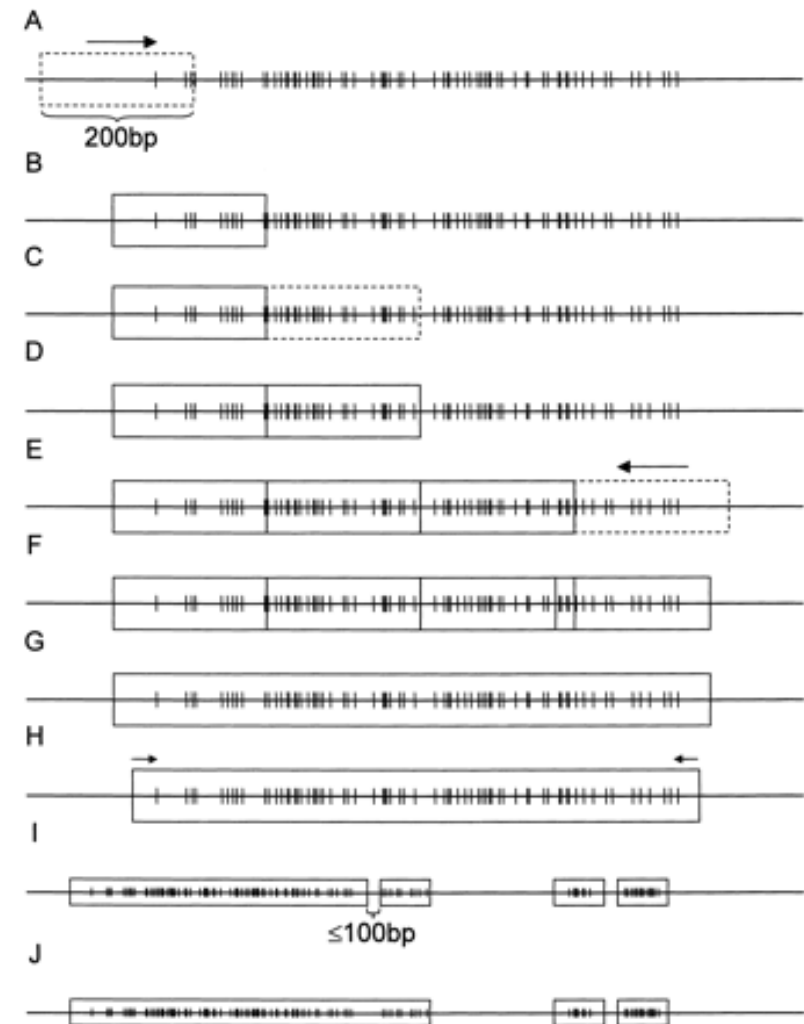
Comprehensive analysis of CpG islands in human chromosomes 21 and 22

Daiya Takai* and Peter A. Jones

Department of Biochemistry and Molecular Biology, University of Southern California/Norris Comprehensive Cancer Center, Keck School of Medicine of the University of Southern California, 1441 Eastlake Avenue, Los Angeles, CA 90033

Edited by Albert de la Chapelle, Ohio State University, Columbus, OH, and approved January 15, 2002 (received for review August 3, 2001)

Figure 1 Schematics for the algorithms for CpG island extraction from human genome sequences. (A) Set a 200-base window in the beginning of a contig, compute %GC and ObsCpG/ExpCpG. Shift the window 1 bp after evaluation until the window meets the criteria of a CpG island. (B) If the window meets the criteria, shift the window 200 bp and then evaluate again. (C and D) Repeat these 200-bp shifts until the window does not meet the criteria. (E) Shift the last window 1 bp toward the 5' end until it meets the criteria. (F) Evaluate total %GC and ObsCpG/ExpCpG. (G) If this large CpG island does not meet the criteria, trim 1 bp from each side until it meets the criteria. (H) Two individual CpG islands were connected if they were separated by less than 100 bp. (I) Values for ObsCpG/ExpCpG and %GC were recalculated to remain within the criteria.



CpGcluster: basado en distancias

Read the DNA sequence

Binary sequence:

0001000010100000010100010100001000101010000101

Determine the distance (d) of each CpG to the next CpG downstream in the DNA sequence:

10,5,5,3,1,8,23,34,21,12,2,5,8,6,9,...N-1

For example, for $d_m = 5$:

10,5,5,3,1,8,23,34,21,12,2,5,8,6,9,...N-1

List of CpG clusters with coordinates, length and number of CpGs

Statistically significant cluster $\equiv$ CpG island

Calculate statistical and sequence properties: G+C content, O/E ratio, CpG density, intra-clustering of CpGs, overlap with Alus, PhastCons etc.

Publicaciones sobre islas CpG

M. Gardiner-Garden and M. Frommer, CpG islands in vertebrate genomes. J. Mol. Biol., 196 (1987), pp. 261–282.

D. Takai and P.A. Jones, Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 3740–3745.

M. Hackenberg, C. Previti, P.L. Luque-Escamilla, P. Carpena, J. Martinez-Aroza and J.L. Oliver, CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinformatics, 7 (2006), p. 446.

Robert S. Illingworth, Adrian P. Bird CpG islands – ‘A rough guide’. FEBS letters (2009)

DNA repetido

Definición simple:

ADN repetido se llama cualquier secuencia que cuenta con varias copias en el genoma

Clasificación:

Básicamente se usa tres clasificaciones: 1) por número de copias, 2) por la distribución en el genoma y 3) por mecanismo de propagación (duplicación o inserción)

Clasificación por frecuencia (cinética de reasociación):

- Altamente repetitivo: aprox. 10-15% de los genomas mamíferos (repeticiones en tandem)
- Moderadamente repetitivo: aprox. 25-40% (incluye la mayoría de los (retro)transposones)
- Elemento único o número bajo: aprox. 50-60% de los genomas

Clasificación por colocación en la secuencia

- 1) Repeticiones en tandem (ADN satélite, pseudogenes)
- 2) Elementos dispersos (SINE, LINE, LTR, transposones de ADN, pseudogenes)

DNA repetido

Clasificación por modo de propagación

- “cortar y pegar”: los elementos cambian su posición en el genoma pero no cambia el número de copias (transposones de ADN)
- “copiar y pegar”: una copia del elemento se inserta en otra posición (retrotransposones)

Retrotransposones

transcriben a RNA y se reinsertan usando transcriptasa inversa

LTR retrotransposones

Caracterizados por sus terminaciones con repeticiones largas (Long Terminal Repeats) como huella dejada por la inserción de retrovirus (aprox. el 8% del genoma humano).

Non-LTR retrotransposones

SINEs (Short interspersed repetitive elements): secuencias cortas (< 500 nt), no-autónomas ya que no codifican por ninguna proteína. Miembros importantes son las Alus (primates) y B1, B2 (roedores).

LINEs (Long interspersed repetitive elements): LINE1 es el miembro más importante. Tiene una longitud de aprox. 7 kbp aunque en el genoma - con excepción de los activos - se encuentran fuertemente recortados. Tienen dos ORF correspondiendo uno de ellos a una transcriptasa inversa.

Las Alus

- Cerca del 11% del genoma humano esta compuesto por este retrotransposon
- La secuencia consenso tiene unos 300 nt y es rica en G+C y CpG.
- Las Alus evolucionaban a partir de la duplicación de un gen 7SL RNA
- Se clasifican en tres familias (AluJ, AluS y AluY) y múltiples subfamilias.
- Se emplea los polimorfismos de inserción de Alus en estudios evolutivos y de poblaciones
- Se estima que aprox. el 0.5% de enfermedades genéticas en humano tiene su origen en una inserción de Alu (neurofibromatosis, hemofilia A y B, Huntington disease y Aspert syndrome, etc.)

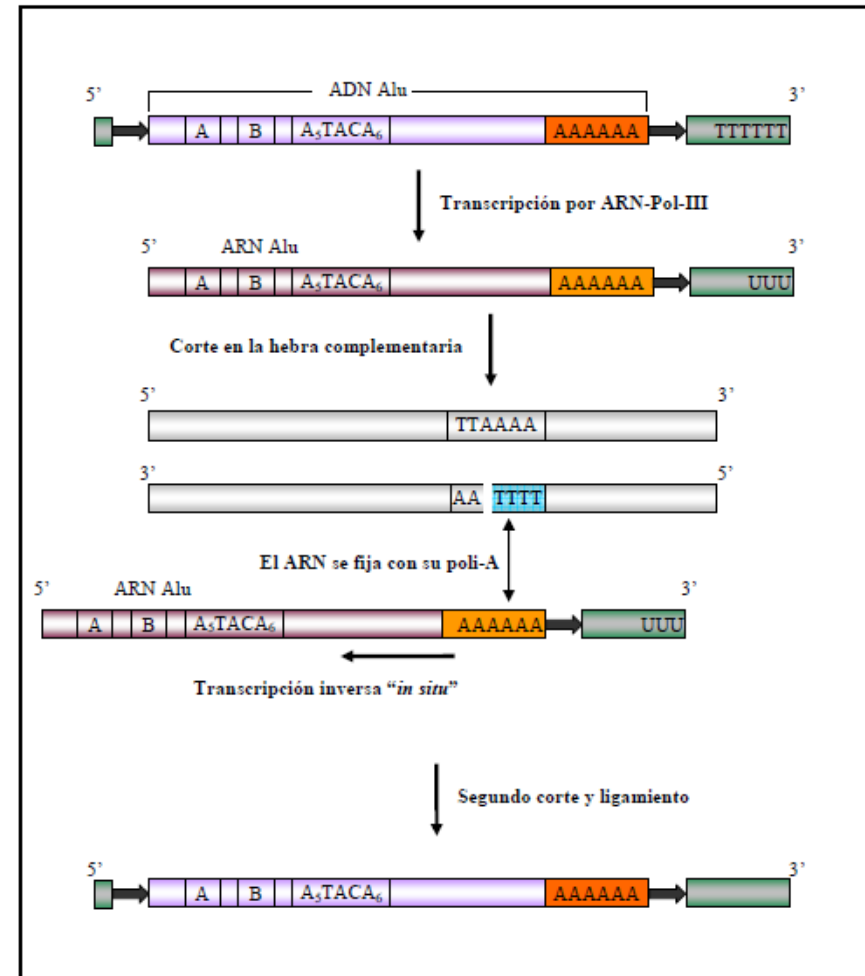


Figura 1-1: Un esquema del proceso de retrotransposición de las Alus. Aparte del primer paso, que consiste en la transcripción de las Alus mediante ARN-pol-III (transcripción mediante ARN-pol-II en el caso de los LINE1), este modelo se puede aplicar también a los elementos LINE1. El segundo paso, muestra la transcripción reversa "in situ" (TPRT: Target-Primed Reverse Transcription; Lin y Levin, 1997). La endonucleasa corta la secuencia (preferentemente en 3'-AA/TTTT-5') y el ARNm del elemento se fija con su cola de poli-A en el extremo 3' del corte abierto, el cual debe de ser rico en T. Después de la transcripción reversa, se produce un segundo corte en la hebra contraria y el ligamiento que cierra el ADN (ilustración adaptada de Batzer y Deininger, 2002).

Evolución de las Alus

El origen de las Alus

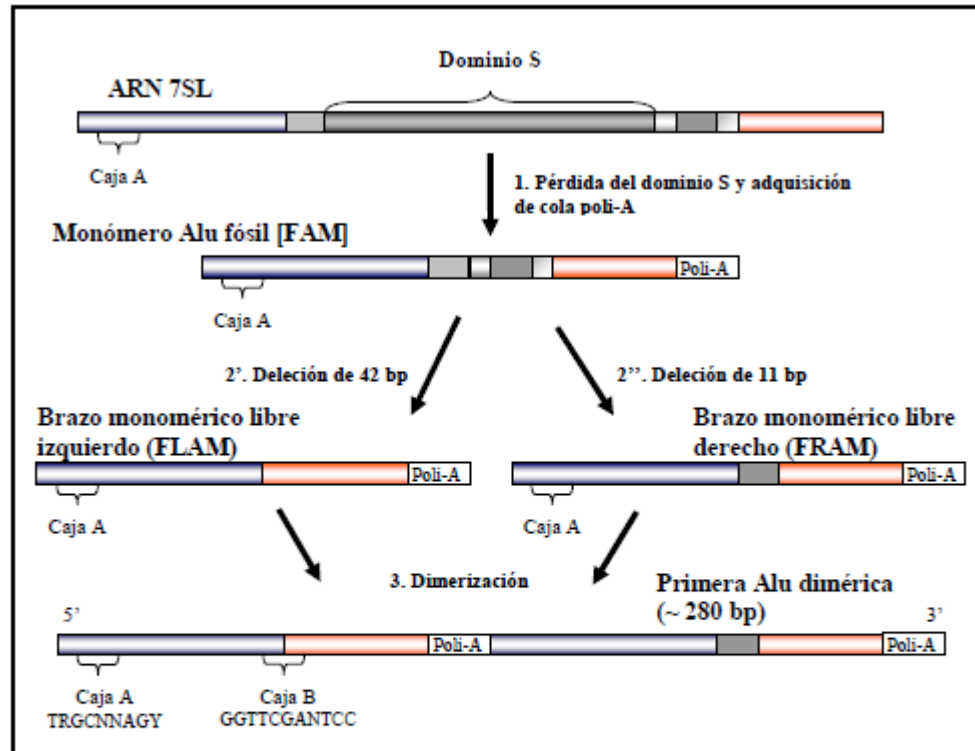


Figura 1-2: Modelo de evolución del gen 7SL hacia la primera Alu dimérica. La evolución ocurrió en tres pasos: 1) pérdida del dominio S del gen 7SL y adquisición de una cola de poli-A, 2) delecciones de 42 bp y 11 bp, respectivamente, que dan lugar a los precursores monoméricos de las Alus, FLAM y FRAM, y 3) la dimerización de FLAM y FRAM. La caja A de las Alus se hereda directamente del gen 7SL, el cual se mantuvo solo en el FLAM y perdió su funcionalidad en el FRAM. El FLAM además adquirió una caja B funcional.

Evolución a partir de un gen maestro

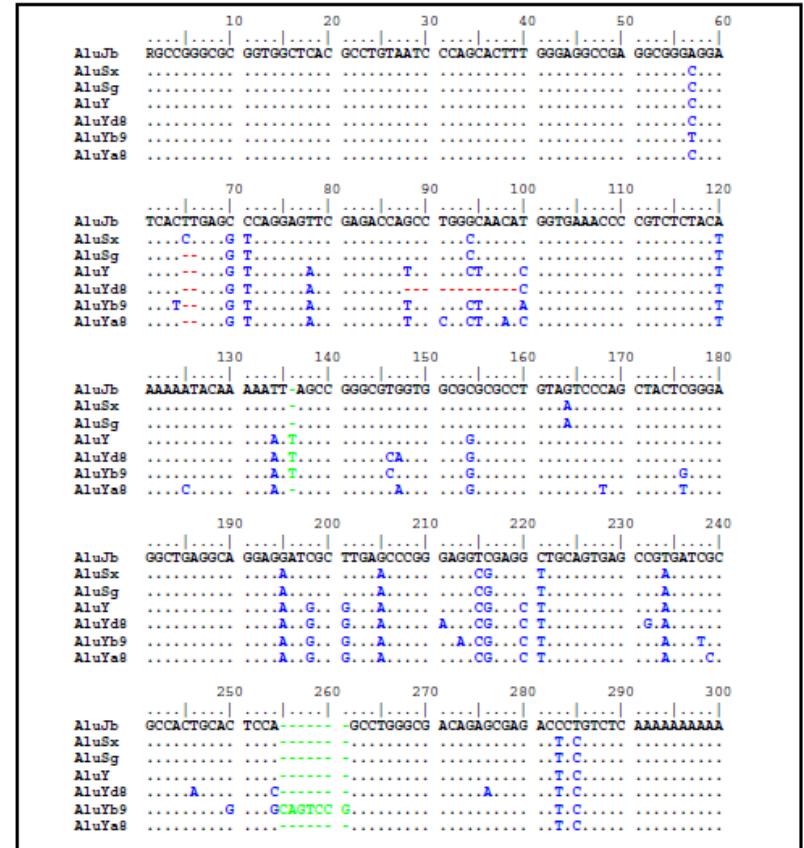


Figura 1-3: Alineamiento múltiple de varias subfamilias de Alus. Se puede observar la evolución de la secuencia AluJo (la más antigua) hacia subfamilias más recientes mediante mutaciones diagnósticas (azul), inserciones (guiones y nucleótidos verdes) y delecciones (guiones rojos). Existen mutaciones que se observan en todas las subfamilias descendientes. No obstante, algunas mutaciones se limitan a determinadas subfamilias. La subfamilia AluYd8 es más antigua que Yb9 e Ya8, pero muestra una delección de 12 bp que no está presente en Yb9 e Ya8. Esto demuestra que la evolución de los genes maestros no es lineal, sino que se forman ramas laterales, indicando así que puede haber más de una subfamilia activa.

Implicaciones funcionales de DNA repetido

Durante muchos años el DNA repetido ha sido considerado como junk DNA (DNA basura), termino acuñado por [Susumu Ohno en 1972](#) aunque este autor no pretendía decir que el DNA repetido sea de ninguna utilidad.

Se cree que la mayoría del DNA repetido no tiene función, incluso pueden tener efectos deletéreos (Alus estan empobrecidos en la región promotora), pero se conoce algunos casos donde un (retro)transposon ha sido “domesticado”:

Carl W. Schmid. Does SINE evolution preclude Alu function? NAR (1998)

AluGene: a database of Alu elements incorporated within protein-coding genes

Lin et al. Diverse Splicing Patterns of Exonized Alu Elements in Human Tissues. PLoS Genetics (2008)

“Our results imply that some Alu exonization events may have acquired adaptive benefits during the evolution of primate transcriptomes.”