

Metilación & Epigenética

Master de Genética y Evolución 2011/2012

Análisis de Secuencias

Michael Hackenberg (mlhack@gmail.com)

<http://bioinfo2.ugr.es/secuencias/>

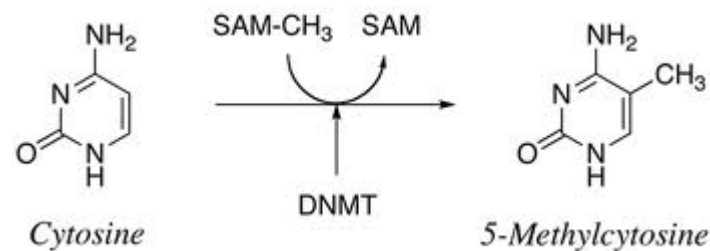
Resumen sesión 5: Elementos genómicos

- Galaxy es una colección de herramientas para el análisis genómico
- Galaxy permite almacenar los resultados y es accesible mediante cualquier navegador
- Los genomas de muchos mamíferos son altamente estructurados respecto al contenido en G+C (isocoras)
- Las isocoras guardan relación con muchas propiedades como la densidad y estructura génica, tasas de recombinación, etc.
- Casi la mitad del genoma humano corresponde a DNA repetido

Introducción

Definición:

La metilación es la adición de un grupo metilo (-CH₃) a una molécula. La metilación se observa tanto en el ADN como en proteínas.



- La metilación se establece mediante diferentes metiltransferasas
- En eucariontes se metila el carbono C-5 de las citosinas
- En genomas de mamíferos, el dinucleótido CpG es la principal diana de metilación

De novo metiltransferasa (DNMT3)

Metila citosinas previamente no-metiladas (o hemi-metiladas)

Metiltransferasa de mantenimiento (DNMT1)

Actúa principalmente sobre citosinas hemi-metiladas (metiladas en una hebra) para mantener el estado de metilación después de la replicación del DNA

Función

- La metilación es clave en el desarrollo embrionario incluyendo la desmetilación y metilación *de novo* (reprogramación) de las células germinales.
- La metilación de DNA es clave en la impronta génica y en la desactivación del cromosoma X
- Los grados de metilación de la región promotor de un gen y en el cuerpo génico influye en los niveles de expresión.

Metilación de DNA, estado de cromatina y expresión génica

- Hay proteínas que se unen específicamente al DNA metilado
- Hay proteínas que se unen específicamente a islas CpG no-metilados

La interacción entre la metilación de DNA y varias proteínas determinan el estado de la cromatina y están así implicados en la regulación génica

Publicaciones

Deaton AM, Bird A.(2011). CpG islands and the regulation of transcription. Genes Dev. 25(10):1010-22.

Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James KD, Turner DJ, Illingworth R, Bird A. (2010).**CpG islands influence chromatin structure via the CpG-binding protein Cfp1.** Nature. 464(7291):1082-6.

Suzuki, M.M. and Bird, A. (2007). **DNA methylation landscapes: provocative insights from epigenomics.** Nat. Rev. Genet. 9(6):465-76.

Manifestación en el genoma: las islas CpG

En el genoma humano solo se encuentra el 20% del 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

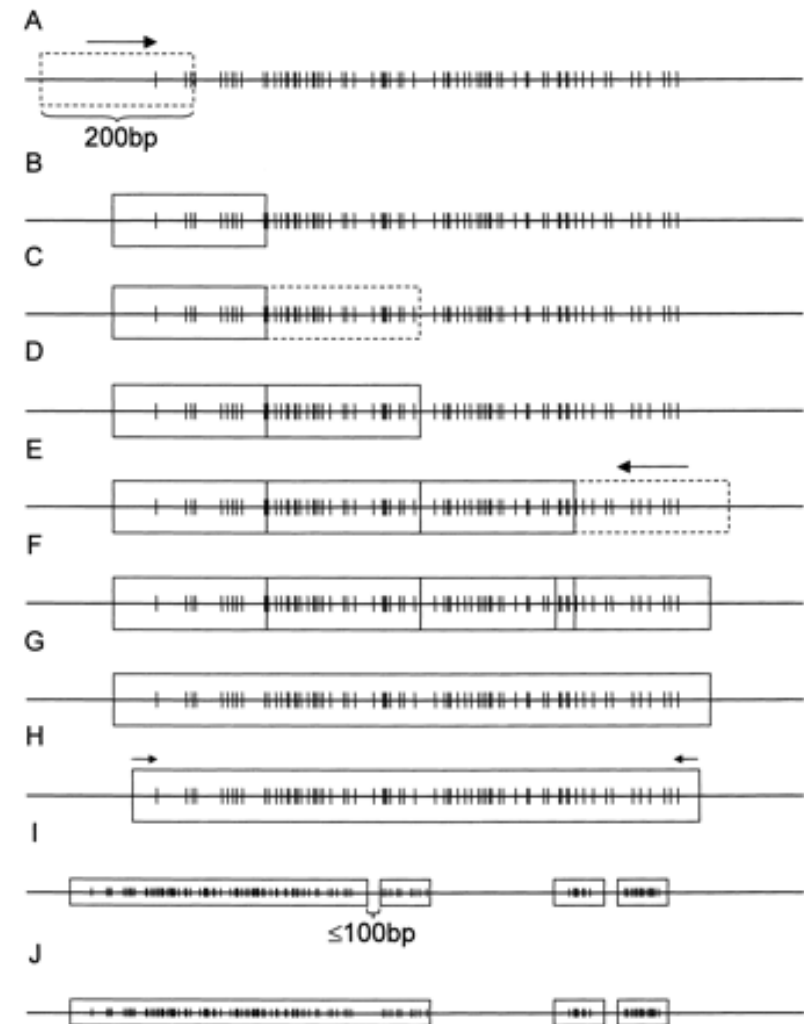
Comprehensive analysis of CpG islands in human chromosomes 21 and 22

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)

Detectar la metilación de DNA

El problema:

- 1) Hibridación es insensible frente a la metilación → no se puede usar chips de DNA
 - 2) La PCR elimina la información acerca del estado de metilación
- Pre-tratamiento del DNA & lectura del estado de metilación

FROM THE FOLLOWING ARTICLE:

Principles and challenges of genome-wide DNA methylation analysis

Peter W. Laird

Nature Reviews Genetics **11**, 191-203

doi:10.1038/nrg2732

[< Back to article](#) | [< Back to figures and tables](#) | [Next table >](#)

Pretreatment	Analytical step			
	Locus-specific analysis	Gel-based analysis	Array-based analysis	NGS-based analysis
Enzyme digestion	• <i>Hpa</i>II-PCR	• Southern blot • RLGS • MS-AP-PCR • AIMS	• DMH • MCAM • HELP • MethylScope • CHARM • MMASS	• Methyl-seq • MCA-seq • HELP-seq • MSCC
Affinity enrichment	• MeDIP-PCR		• MeDIP • mDIP • mCIP • MIRA	• MeDIP-seq • MIRA-seq
Sodium bisulphite	• MethyLight • EpiTYPER • Pyrosequencing	• Sanger BS • MSP • MS-SNuPE • COBRA	• BiMP • GoldenGate • Infinium	• RRBS • BC-seq • BSPP • WGSBS

AIMS, amplification of inter-methylated sites; BC-seq, bisulphite conversion followed by capture and sequencing; BiMP, bisulphite methylation profiling; BS, bisulphite sequencing; BSPP, bisulphite padlock probes; CHARM, comprehensive high-throughput arrays for relative methylation; COBRA, combined bisulphite restriction analysis; DMH, differential methylation hybridization; HELP, *Hpa*II tiny fragment enrichment by ligation-mediated PCR; MCA, methylated CpG island amplification; MCAM, MCA with microarray hybridization; MeDIP, mDIP and mCIP, methylated DNA immunoprecipitation; MIRA, methylated CpG island recovery assay; MMASS, microarray-based methylation assessment of single samples; MS-AP-PCR, methylation-sensitive arbitrarily primed PCR; MSCC, methylation-sensitive cut counting; MSP, methylation-specific PCR; MS-SNuPE, methylation-sensitive single nucleotide primer extension; NGS, next-generation sequencing; RLGS, restriction landmark genome scanning; RRBS, reduced representation bisulphite sequencing; -seq, followed by sequencing; WGSBS, whole-genome shotgun bisulphite sequencing.

Bisulfito sódico & Secuenciación

El protocolo:

1) Tratar el DNA con Bisulfito sódico

→ una citosina no-metilada se desamina: citosina → uracilo → timina (PCR)

→ una citosina metilada se mantiene

2) Secuenciar el DNA

Ventaja:

Se obtiene información de metilación para cada citosina y no solo valores medios para una región como ocurre con muchos otros métodos

Se puede detectar la metilación en todos los contextos y no solo CpG

Reto:

Re-secuenciar un genoma entero

Alinear miles de millones de secuencias cortas (reads)

Tratamiento con Bisulfito

Genomic
Fragment

5' ...TCGTAAGCTGCGATGT CAGCGTAGTTCCGATG... 3'
3' ...AGCATTTCGACGCTACAGTTCGCATCAAGGCTAC... 5'

Denaturation
&
Bisulfite
modification

BSW
BSC

5' ...TUGTAAGCTGUGATGT CAGUGTAGTT CUGATG... 3'
3' ...AGUATTTCGACGUTACAGTCGUATCAAGGUTAC... 5'

A

MethylC-Seq

PCR I

5' ...TTGTAAGCTGTGATGT CAGTGTAGTTCTGATG... 3' BSW
3' ...AGTATTTCGACGTTACAGTCGTATCAAGGTTAC... 5' BSC

Secuenciación

Métodos actuales → Second Generation Sequencing
(Secuenciación masiva)



454
Pyrosequencing (PS)

454
SEQUENCING

illumina

Illumina
Reversible Termination (RT)



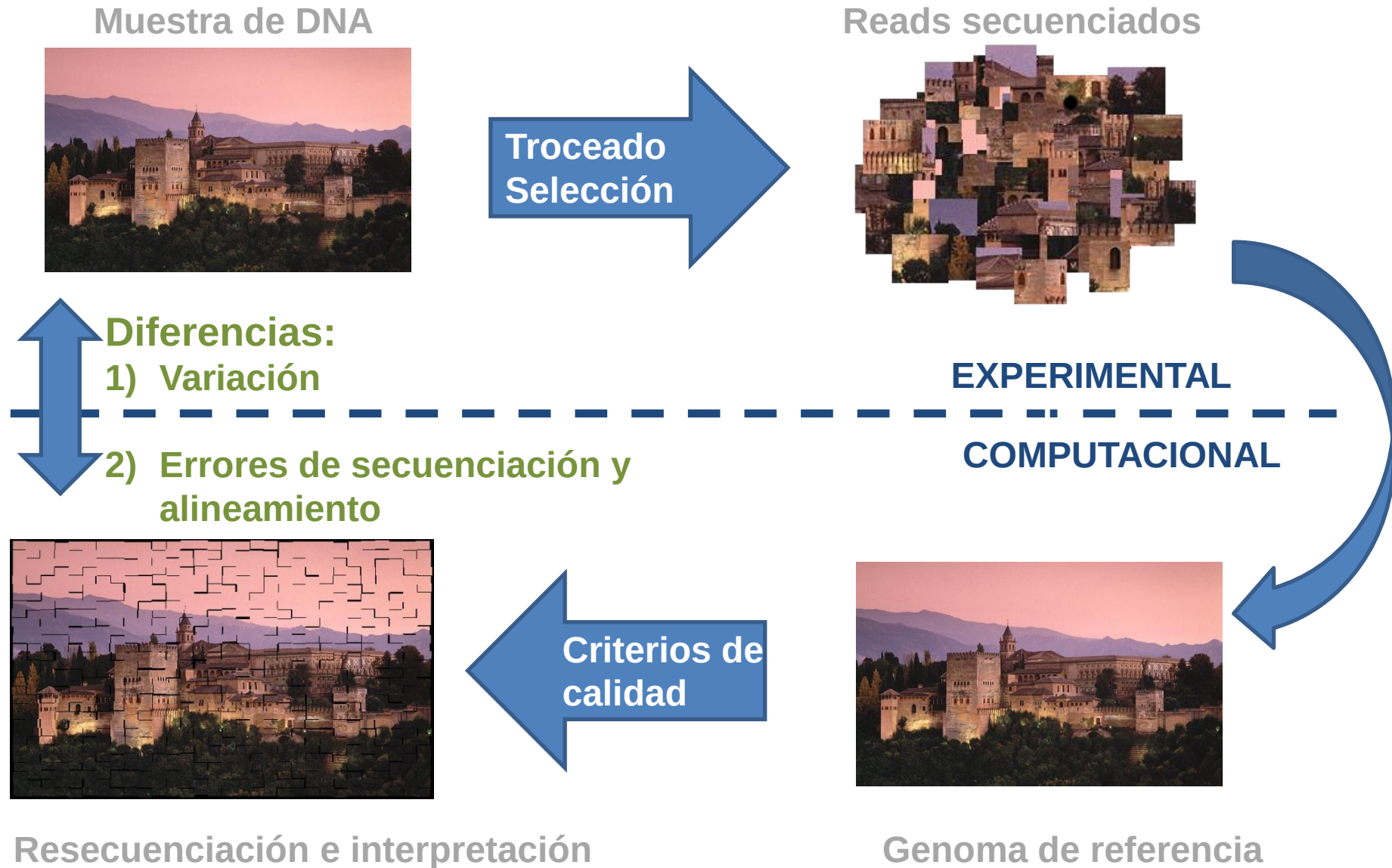
life
technologies[®] | AB applied
biosystems[®]

SOLID
Sequencing by Ligation (SBL)

Bisulfito & Secuenciación

	SECUENCIACIÓN MASIVA			
	SANGER			
	Di-deoxy terminator	Roche 454 GS FLX (PS)	Illumina HiSeq 2000 (RT)	SOLID V4 (SBL)
Salida por proceso	1.6 Mb	600 Mb	200 GB	100 GB
Tiempo/Proceso	1h	10 h	9 d	11 d
Longitud media "reads"	800 pbs	400 pb	100 pb	75 pb
Salida por día	38.4 Mb	1.44 GB	22.2 GB	9 GB
Usos frecuentes	-	Secuenciación de novo Captura de exones	Resecuenciación Captura de exones Metagenómica	Resecuenciación Captura de exones Metagenómica

Resecuenciación con NGS



Obtener niveles de metilación

Distinguir entre la acción del bisulfito y

- 1) Errores de secuenciación
- 2) SNV (Single Nucleotide Variation) – un polimorfismo con los alelos C/T sería detectado como una citosina no-metilado!!

Watson mapped reads

Watson reference
Crick reference

Crick mapped reads

```

TCTAACGTTCTGATATCAGCGTAGTTTCCGATC
TTGTAAGCTGTGATATCAGCGTAGTTTGGATG
TTGTAAGTTGTGATATCAGCGTAGTTTCCGATG
TTGTAAGCTGTGATATCAGCGTAGTTTCCGATT
TCGTAAGCTGCGATTTTAGCGTAGTTCCGATG
TTGTAAGTTGTGATATTAGCGTAGTTTCCATG
TTGTAAGTTGTGATATTACCGTAGTTTGGATA
TTGTAAGCTGTGATATTAGCGTAGTTTCCGATG
TTGTAAGCTGCGATATTAGCGTAGTTTCCGATG
TTGTAAGCTGTGATATTAGCGTAGTTCCGATG

5' ...TCGTAAGCTGCGATGTAGCGTAGTTCCGATG... 3'
3' ...AGCATTGACGCTACAGTCGCATCAAGGCTAC... 5'

AATATTTGACGTTATAGTTGCATTAAGGCTAT
AATATTTGATGTTATAGTTGCATTAAGGCTAT
AACATTTGACGTTATAGTTGTATTAAGGCTAT
AATATTTGACGCTATAGTTGCATTAAGGCTAT
AATATTTGACGTTATAGTTGCATCAAGGCTAT
AGTATTTGACGTTATAGTTGCATTAAGGCTAT
AATATTTGACGTTATAGTTGCATTAAGGCTAT
AACATTTGACGTTATAGTTGCATTAAGGCTAT
AATATTTGACGTTATAGTTGCATTAAGGCTAT
AATATTTGACGTTATAGTTGCATTAAGGCTAT

```

Context	Start	Watson Level	Crick level	Merged Level	Watson SNV fraction	Crick SNV fraction	Result
CG (+/-)	2	0.1	0.2	0.15	0	0.9	rejected
CTT (-)	4	-	0	-	0	-	unMeth
CWG (+/-)	8	0.6	0.9	0.75	0	0	interMeth
CG (+/-)	11	0.2	0.1	0.15	0	0	unMeth
CAT (-)	13	-	0	0	1	-	rejected
CWG (+/-)	17	0.3	0.1	0.2	0.2	0	unMeth
CG (+/-)	20	1	0.9	0.95	0	0	meth
CTA (-)	22	-	0.1	-	0	-	unMeth
CCG (+)	27	0.2	-	-	-	0.2	unMeth
CG (+/-)	28	0.8	0.9	0.85	0.1	0	meth
CAT (-)	30	-	0	-	0.2	0	unMeth

Resultados recientes y perspectivas

En células no diferenciados, hasta un 25% de las citosinas fuera del contexto de CpG están metiladas

Entre el 30 y 40% del genoma de células diferenciadas muestran metilación parcial (niveles de metilación entre el 0.1 y 0.6)

Lister et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315-322

La metilación de DNA en el cuerpo génico es mas alto en genes activos

Ball, M. P. et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotech.* 27, 361–368 (2009)

La metilación del contexto CpHpG es crucial para silenciar transposones en plantas

Miura, A. et al. An Arabidopsis JmjC domain protein protects transcribed genes from DNA methylation at CHG sites. *EMBO J.* 28, 1078–1086 (2009).

El futuro

Establecer el enlace definitivo entre cromatina y metilación de DNA, caracterizar los factores de transcripción y proteínas que se unen al DNA metilado y no-metilado

Stadler et al. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 10.1038/nature10716

Propiedades de la base de datos NGSmethDB

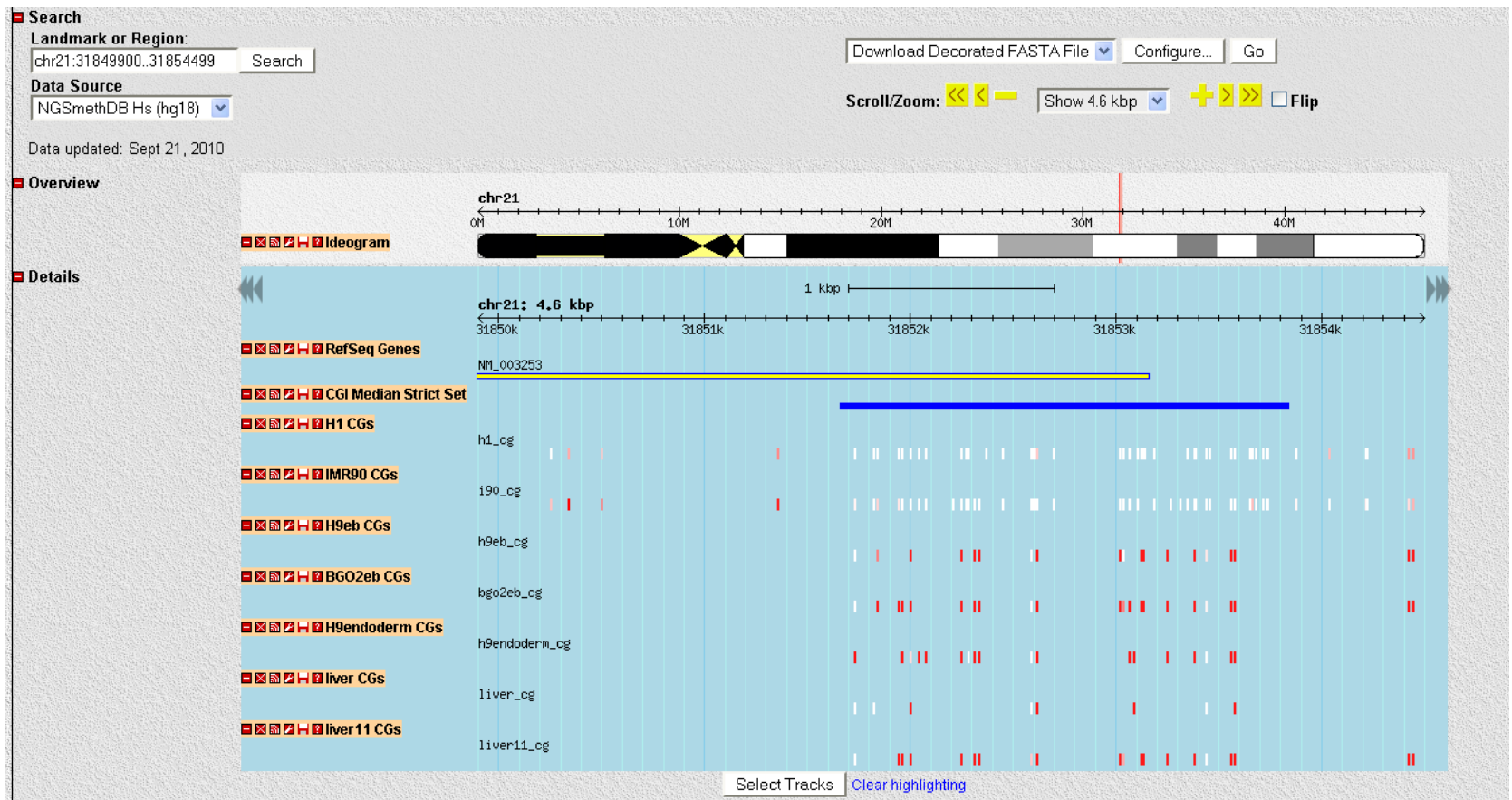
(<http://bioinfo2.ugr.es/NGSmethDB/gbrowse>)

- La base de datos NGSmethDB almacena datos de metilación a nivel de la citosina, generados por secuenciación masiva y conversión con bisulfito
- Actualmente contiene datos para 3 especies (humano, ratón y Arabidopsis) y un total de 22 tejidos únicos.
- Todos los datos se pueden bajar en formato texto.
- No solo tiene datos del contexto CpG sino también de CWG (CAG/CTG)
- El usuario puede elegir entre diferentes cuberturas (las veces que se ha secuenciado una posición)
- Para humano hay 1166274639, para ratón 69459481y para Arabidopsis 16321229 datos de metilación.
- La base de datos cuenta con dos modos para obtener y analizar los datos. Primero, un navegador (Genome Browser) para mostrar la metilación en el contexto de otras anotaciones como genes, islas CpG, contenido en G+C, etc. Segundo, varias pequeñas herramientas (php/Java) están disponibles para 1) buscar citosinas no metiladas, 2) detectar citosinas con metilación diferencial en un conjunto de tejidos, 3) obtener los niveles de metilación en una región dada, 4) analizar los niveles de metilación en la región promotora de los genes.

NGSmethDB

El navegador de NGSmethDB

La metilación se muestra mediante un gradiente de color: de blanco (no metilado) a rojo intenso (totalmente metilado)



Interfaz a las herramientas de NGSmethDB

Los valores de metilación pueden tomar valores entre 0 (no metilado) y 1 (totalmente metilado)

Overview and dump download

See content of DB: [here](#)

This section gives an overview on the DB content and allows to download the data as gzipped text files

Retrieve unmethylated contexts

Retrieve Methylation data for Species:

Retrieve all unmethylated contexts within the a set of user-selected tissues. The user can specify the methylation threshold, the methylation context and the tissues. Just contexts will be reported which are below the threshold in ALL tissues.

Retrieve differentially methylated contexts

Retrieve Methylation data for Species:

Retrieve all differentially methylated contexts for a set of tissues. The user can specify the methylation thresholds, the methylation context and the tissues. Just contexts will be reported which are in at least one tissue methylated and in another tissue unmethylated. For all tissues the methylation state must be either methylated or unmethylated, i.e. contexts which show intermediate methylation in one tissue will not be reported.

Get methylation states of promoter regions (RefGene)

Retrieve promoter methylation data for Species:

Retrieve the methylation in a promoter region [TSS-1.5 kb, TSS+500 bp] for all tissues in the database.

Retrieve methylation data for chromosome region

Retrieve data for Species:

Retrieve all contexts in a given region for a set of tissues.