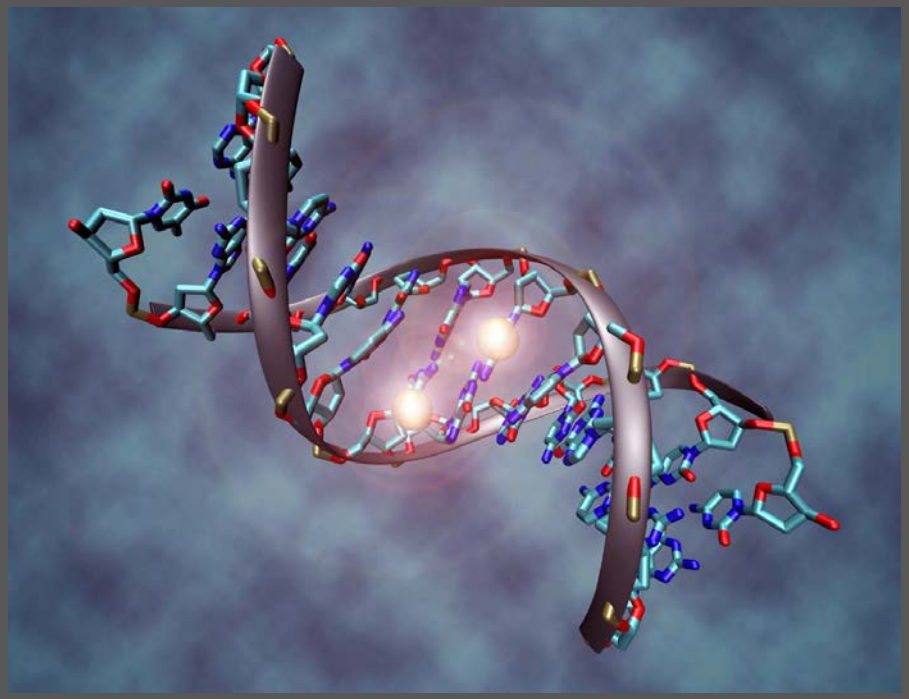
Predicción computacional de islas CpG

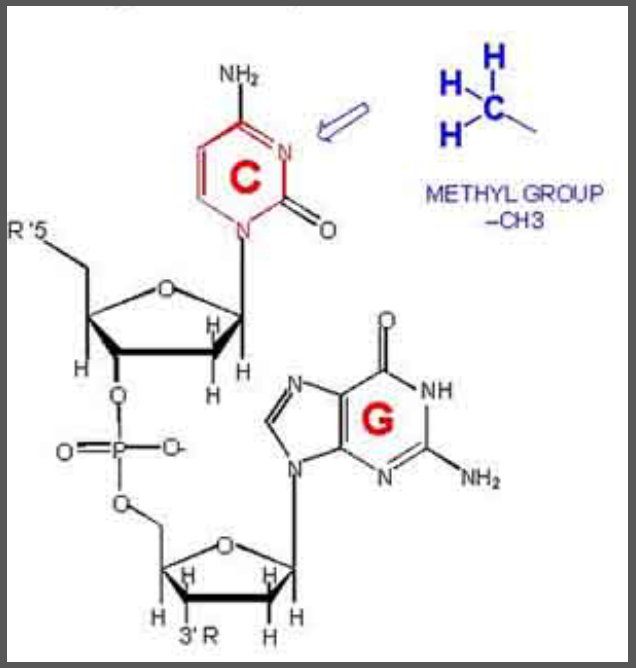
- M. Hackenberg, C. Previti, P.L. Luque-Escamilla, P. Carpena, J. Martinez-Aroza and J.L. Oliver. 2006. **CpGcluster**: A distance-based algorithm for CpG-island searching
  BMC Bioinformatics 7: 446

- M. Hackenberg, P. Carpena, P. Bernaola-Galván, G. Barturen, A.M. Alganza and J.L. Oliver. 2011. **WordCluster**: detecting clusters of DNA words and genomic elements
  Algorithms for Molecular Biology **6:**2

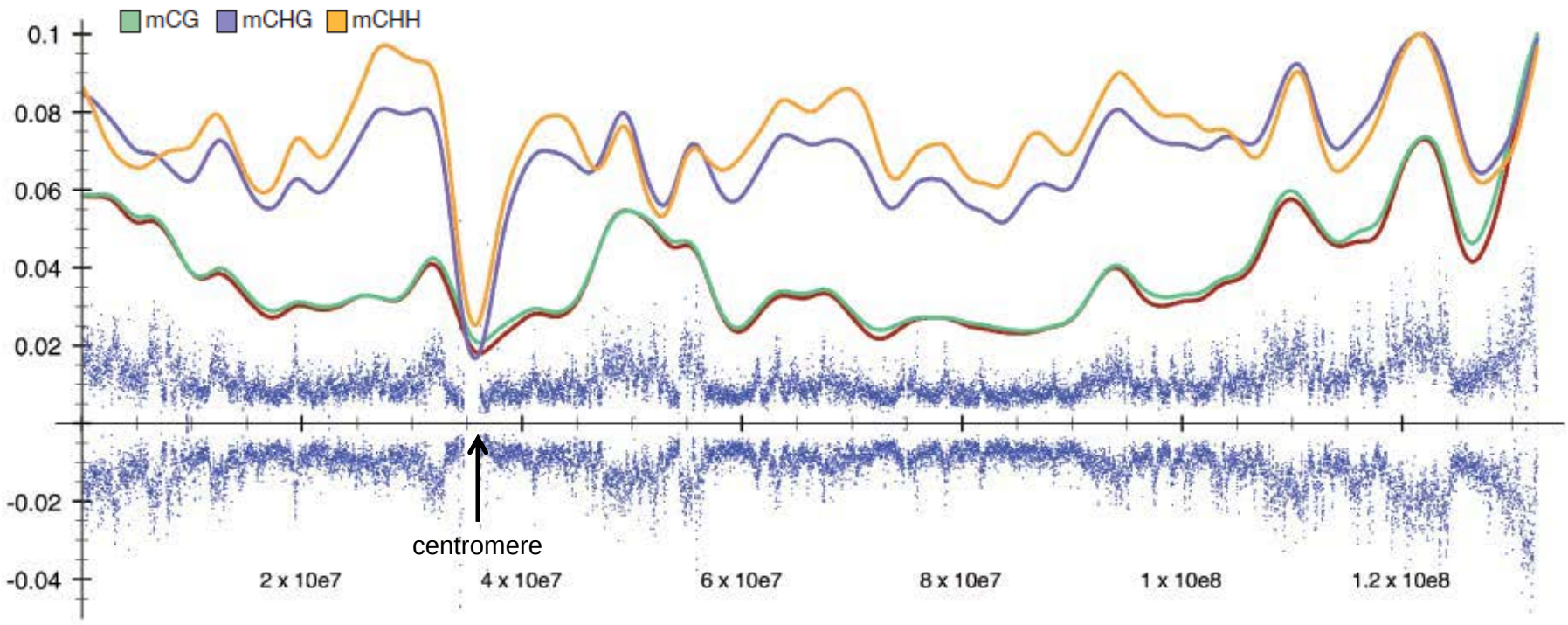
Metilación del ADN



- C → T
- Baja frecuencia de dinucleótidos CpG en vertebrados
  - Por azar se espera un 6.25%
  - Pero sólo se observa el 1% (Bird 1986)



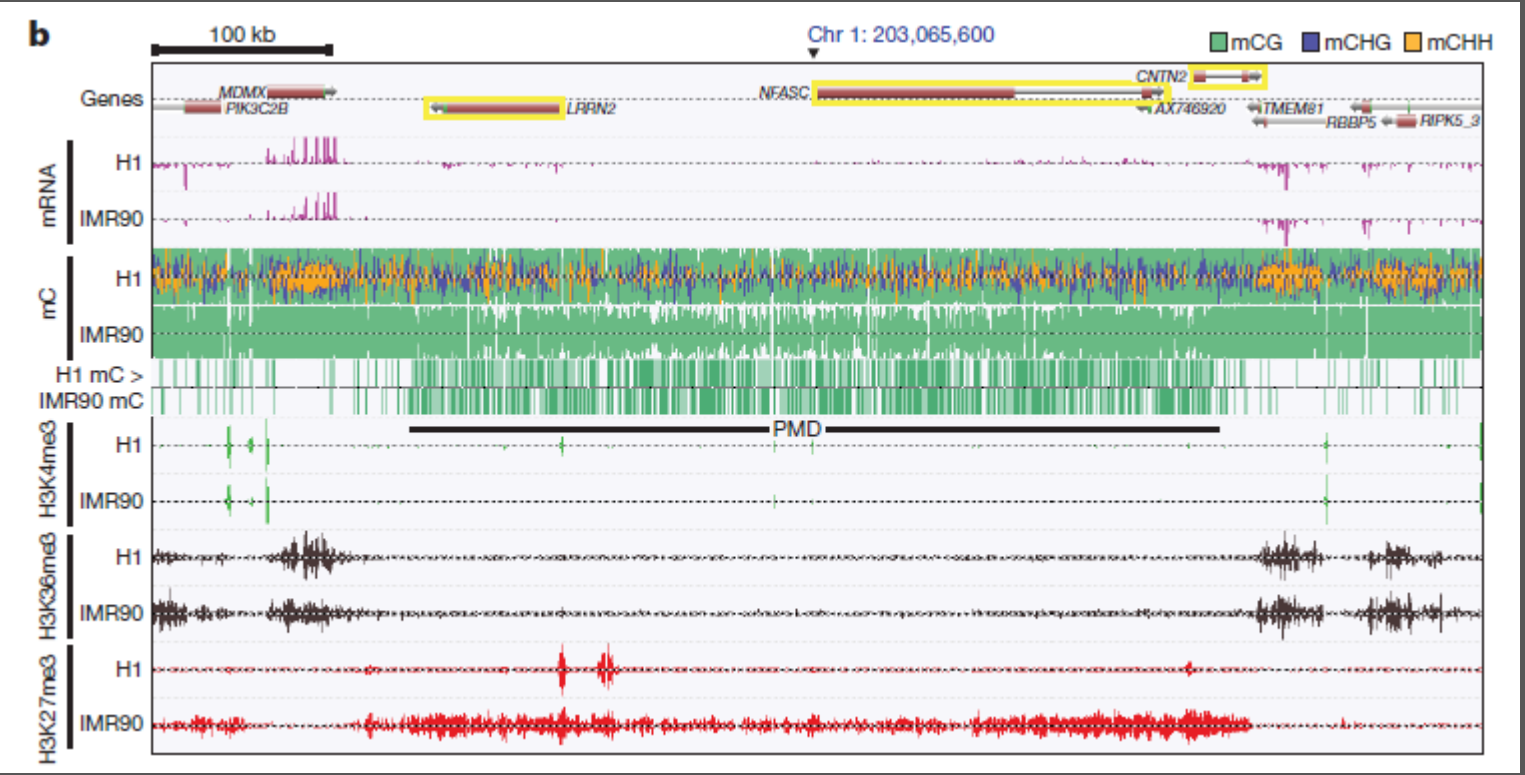
Global trends of human DNA methylomes



Lister et al. NATURE, Vol 462, 19 November 2009

Blue dots indicate methylcytosine density in H1 cells in 10-kb windows throughout chromosome 12. Smoothed lines represent the methylcytosine density in each context in H1 and IMR90 cells. Black triangles indicate various regions of contrasting trends in CG and non-CG methylation. mC, methylcytosine.

Cell-type variation in DNA methylation



Lister et al. NATURE, Vol 462, 19 November 2009

DNA methylation, mRNA and histone modifications in H1 and IMR90 cells associated with a PMD (partially methylated domain). Vertical lines above and below the dotted central line in DNA methylation tracks indicate methylcytosines on the Watson and Crick strands, respectively. Line vertical height indicates the methylation level. The H1mC.IMR90mCtrack indicates methylcytosines significantly more methylated in H1 than IMR90 at a 5% FDR (Fisher's exact test).

Cell-type variation in DNA methylation

NGSmethDB v2

A database for NGS single-cytosine-resolution DNA methylation data

http://bioinfo2.ugr.es/NGSmethDB/index.php

Stefanie Geisen, Guillermo Barturen, Ángel M. Alganza, Michael Hackenberg and José L. Oliver. 2014. Nucleic Acids Research, Vol. 42, Database issue D53–D59

Islas CpG

En el genoma humano, la frecuencia observada de CpGs es 5 veces más baja que la esperada, teniendo en cuenta el %GC. Se debe a la alta tasa de mutación de los CpGs metilados (el 70-80% de los CpGs del genoma humano están metilados.

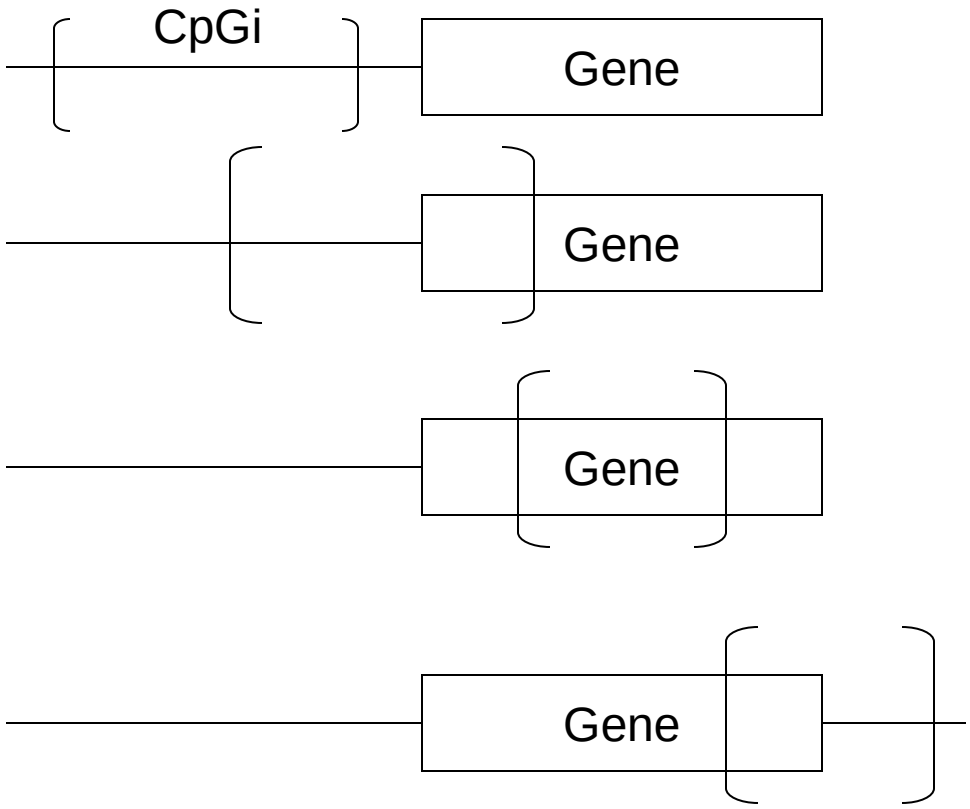
→"regions of DNA with a high G + C content and a high frequency of CpG dinucleotides relative to the bulk genome" Gardiner-Garden and Frommer (1987)

Propiedades:

- El 50-70% de los genes tienen islas CpG en sus promotores
- Casi todos los genes domésticos tienen islas CpG
- Las islas CpG suelen estar hipometiladas. Cuando se metilan pueden dar lugar a cambios epigenéticos que conducen al desarrollo de cáncer y otras enfermedades (Alzheimer, etc).

CpG islands & Genes

5’ end



Promoter CpG islands

CpG islands in body

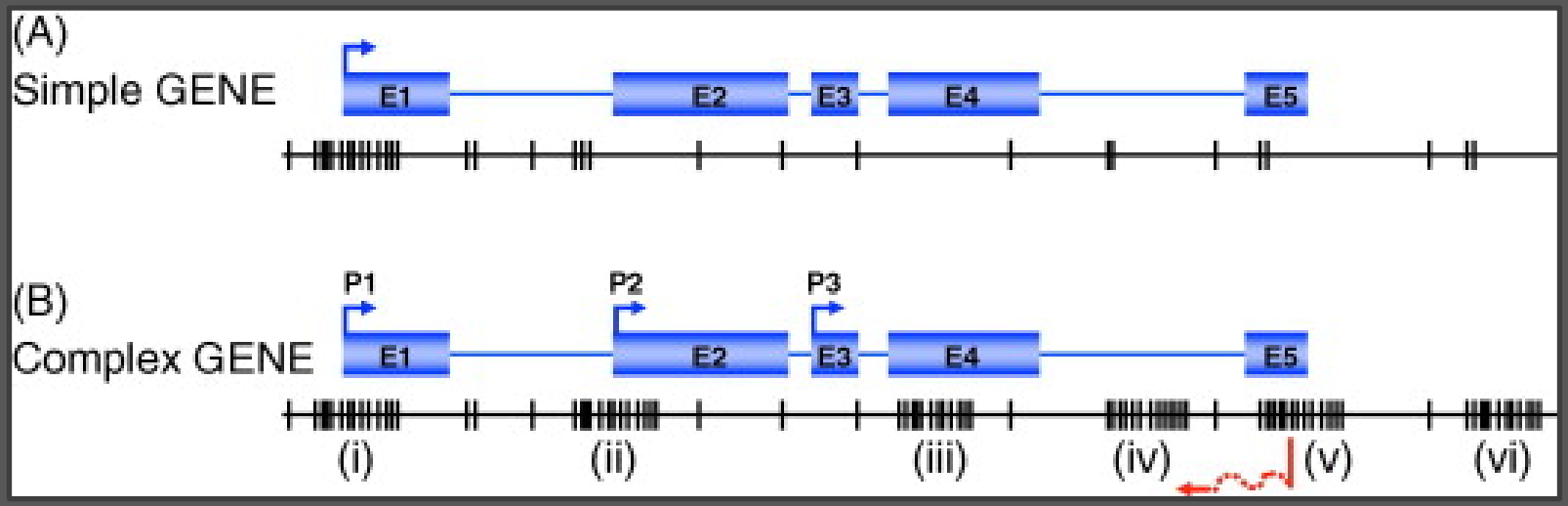
3’ end CpG islands

CLASIFICACIÓN Y LOCALIZACIÓN DE LAS ISLAS CpG RELATIVA A LOS GENES

→**Constitutivamente no-metiladas (asociadas a "Housekeeping genes")**
~100% de los genes domésticos tienen alguna isla asociada

→**Diferencialmente metiladas (genes tejido-específicos)**
~50% "Tejido específicos" isla asociada

→**Parcialmente metiladas (genes improntados)**

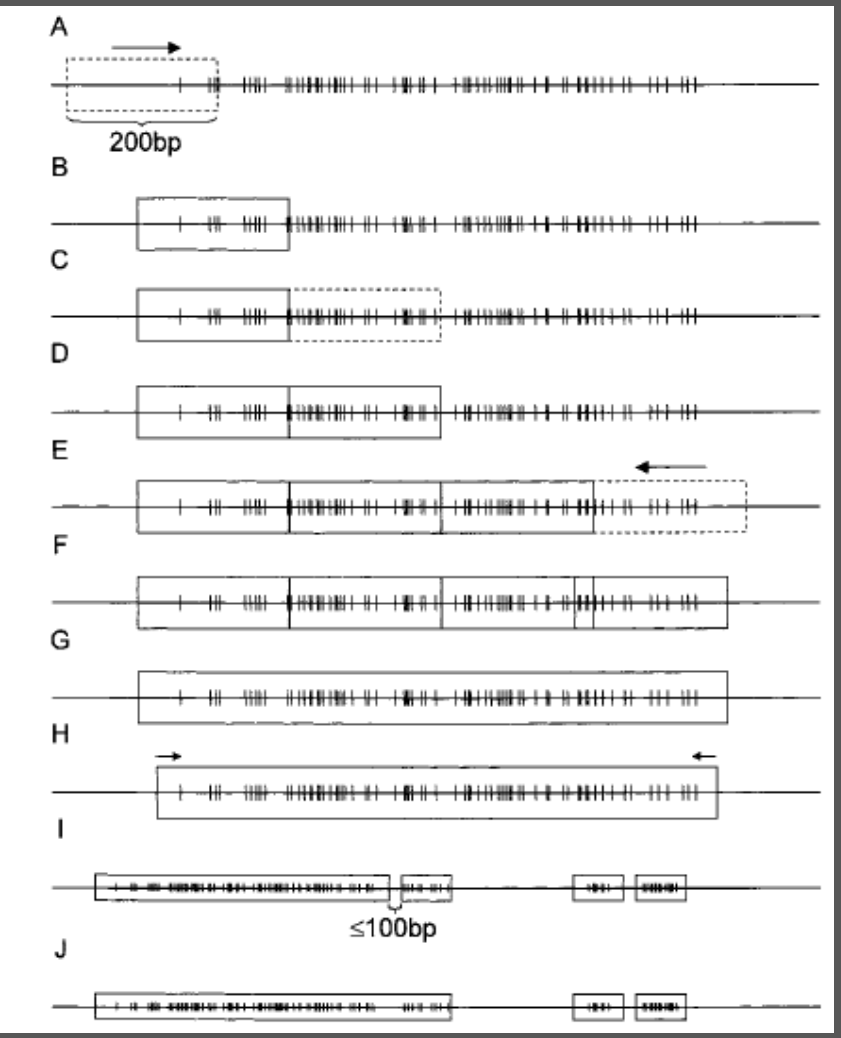


PROPIEDADES DE LAS ISLAS CpGs

- Ricas en G+C y dinucleótidos CpG con longitudes alrededor de 1 kb
- Ratios O/E (número de CpG observado / esperado) altos
- Regiones libres de metilación en algún tejido o condición
- Frecuentemente localizadas en la región promotora de los genes

Algoritmos tradicionales (ejemplo: *CpGplot*)

Ventanas móviles y umbrales:



Espacio paramétrico muy alto:

Umbrales arbitrarios para:

- Proporción CpGs obs/esp
- %GC
- Longitud

Otros parámetros arbitrarios:

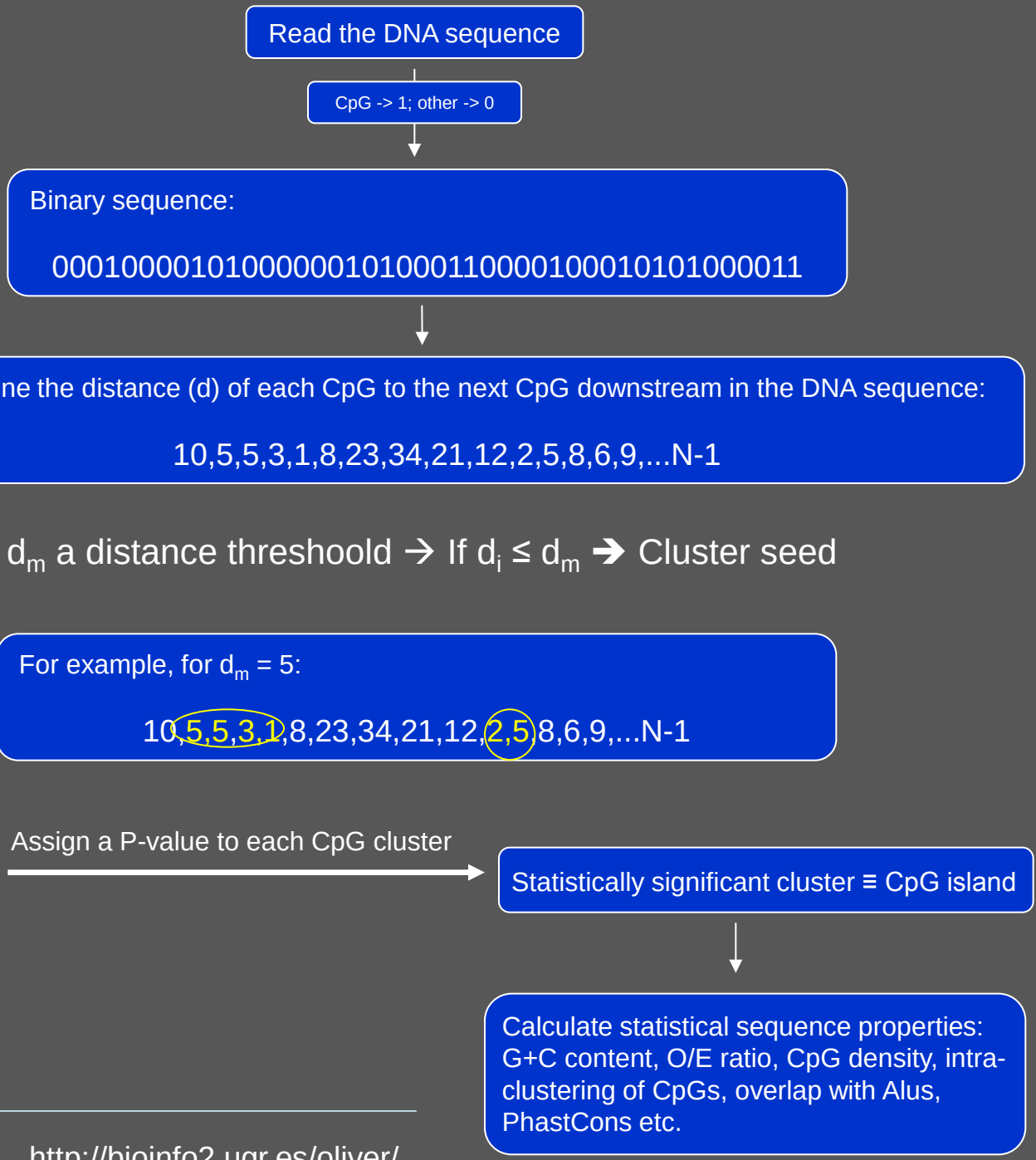
- Longitud de ventana
- Salto
- Distancia para fusionar proto-islas

Principal inconveniente: hay que ajustar los umbrales para filtrar las Alus

From Takai and Jones (2002)

CpGcluster

Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, Oliver JL. 2006. BMC Bioinformatics 7: 446



→ Si se distribuyeran al azar, las distancias entre CpGs seguirían la distribución geométrica:

$$P(d) = (1-p)^{d-1}p$$

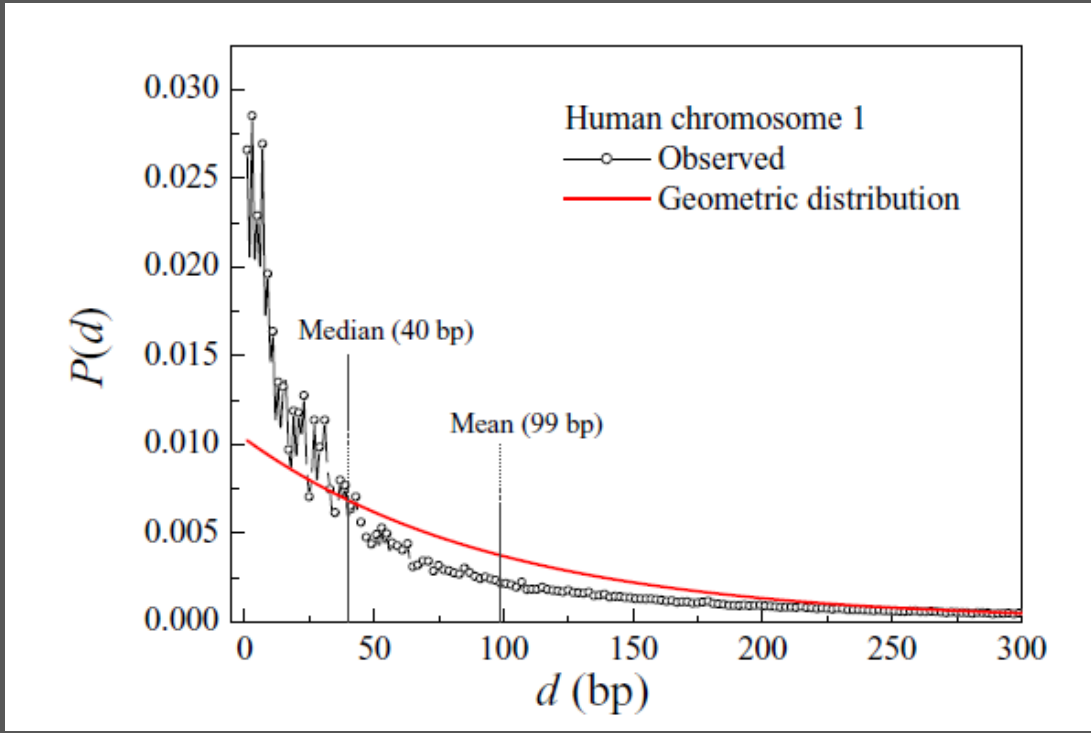
P(d), probabilidad de encontrar una distancia *d* entre CpGs adyacentes y *p* la probabilidad de encontrar un CpG en la secuencia.

→ Las distancias cortas observadas se encuentran sobre-representadas en el genoma, por encima de lo esperado (Existen "Clusters de CpGs").

→ El cruce entre observada y esperada se utiliza como distancia para agrupar CpGs.

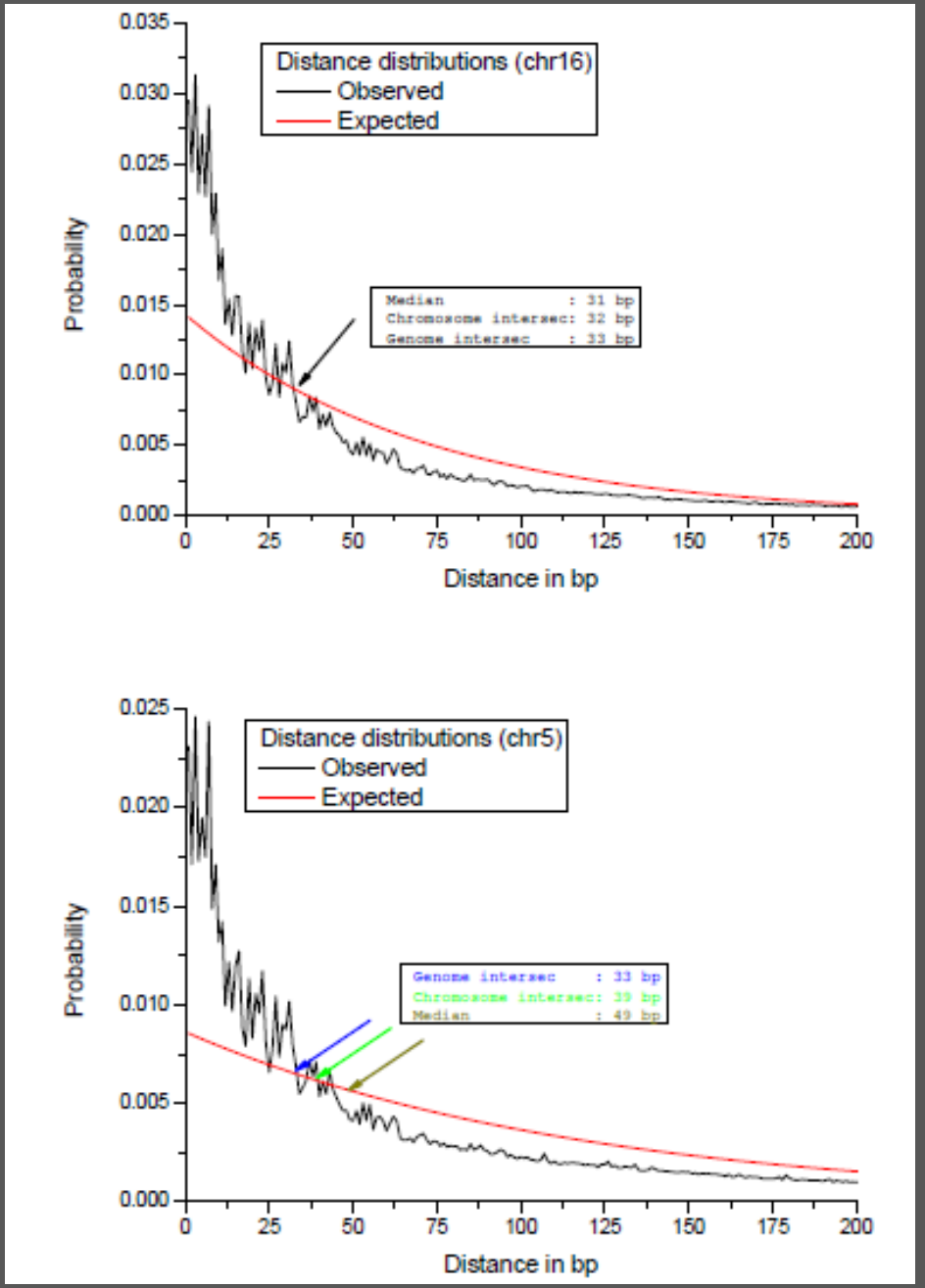
The intersection point as the distance threshold

Probability density function of distances between neighboring CpGs. Distribution of distances between neighboring CpG dinucleotides in the human chromosome 1. The observed distribution is represented in symbols, while the random expectation corresponding to the geometric distribution is represented in a solid line. Note that, in a good approximation, the median separates over-represented distances from under-represented ones.



CpGcluster - Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, Oliver JL. 2006. BMC Bioinformatics 7: 446

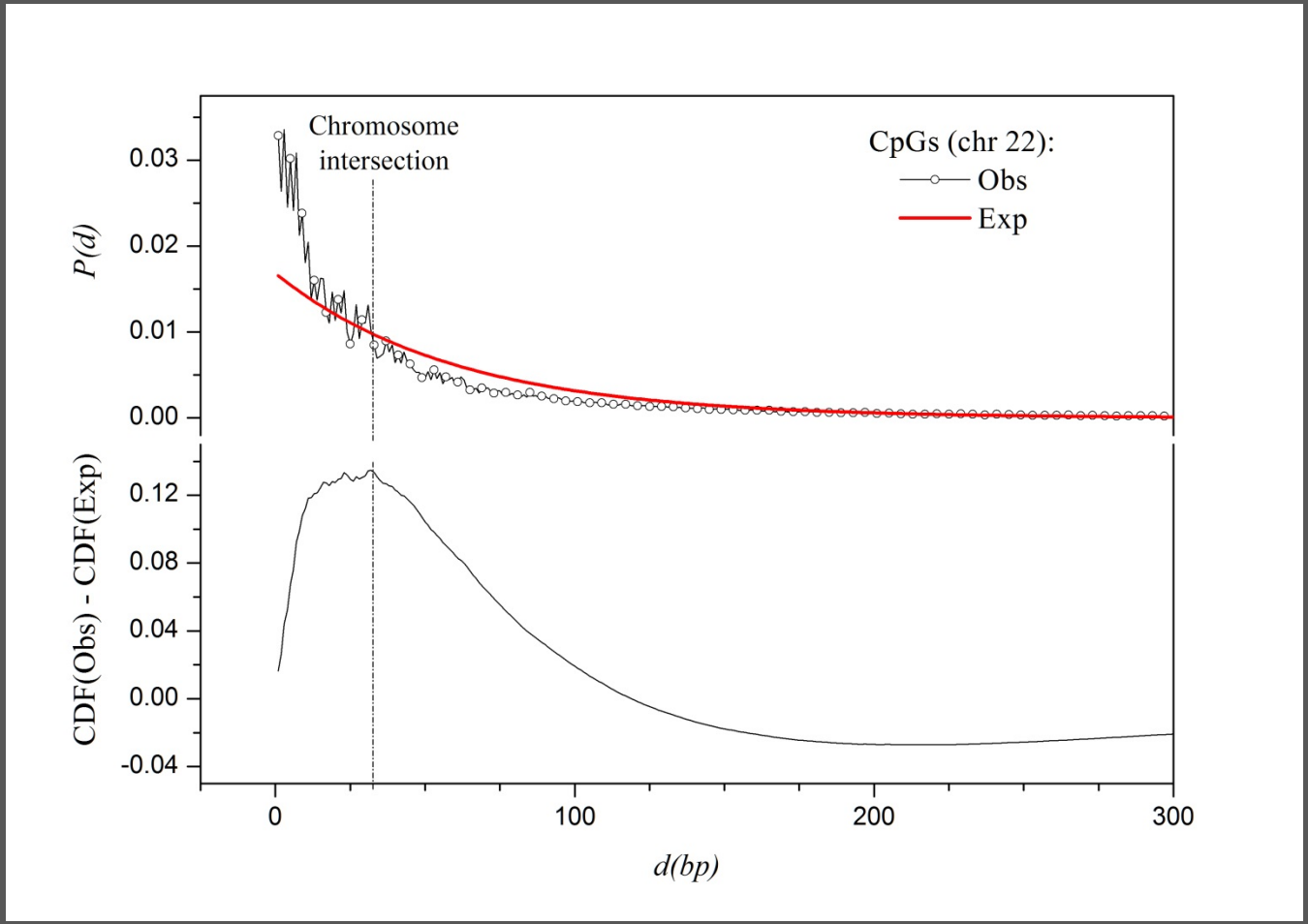
Distance distributions. Expected and observed distance distributions for human chromosomes 16 (above) and 5 (below). It can be seen that for chr16 the median, the chromosome intersection and the genome intersection are very close (within 1 bp), while for chromosome 5 notable differences exist (from 33 bp to 49 bp).



WordCluster - Michael Hackenberg, Pedro Carpena, Pedro Bernaola-Galván, Guillermo Barturen, Ángel M. Alganza and José L. Oliver. 2011. **Algorithms for Molecular Biology 6:**2

Computing the genome (or chromosome) intersection point

Observed and expected distance distributions for the CpGs of chr22 (hg19). Note that short distances are overrepresented and the large ones underrepresented as compared to the expected distances (geometric distribution).The first cross between both curves separates both regimes. Bottom: The intersection between both curves (called the chromosome intersection) can be precisely computed as the maximum difference between the observed and expected cumulative density functions (CDFs).

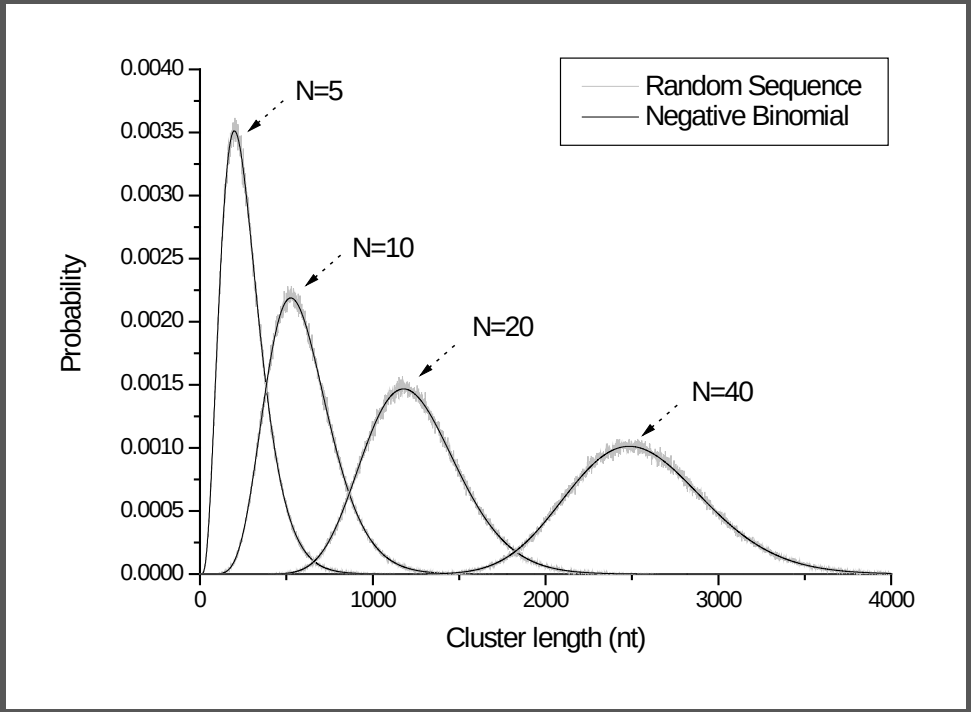


WordCluster - Michael Hackenberg, Pedro Carpena, Pedro Bernaola-Galván, Guillermo Barturen, Ángel M. Alganza and José L. Oliver. 2011. **Algorithms for Molecular Biology 6:**2

Statistical significance of genome clusters

What is the probability to find a cluster with X CpGs and length N in a random distribution?

Negative Binomial Distribution:
Probability to get r failures (non-CpGs) when the number of successes (CpGs) is fixed in advance



WordCluster - Michael Hackenberg, Pedro Carpena, Pedro Bernaola-Galván, Guillermo Barturen, Ángel M. Alganza and José L. Oliver. 2011. **Algorithms for Molecular Biology 6:**2

Benchmark test

CpGcluster - Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, Oliver JL. 2006. BMC Bioinformatics 7: 446

Compare predictions to five commonly used algorithms by means of a prepared set of test sequences (400 experimental CpG islands embedded into a random background):

- CpGcluster shows moderate Sensitivity (Sn) but reaches highest values for Specificity (Sp) and Correlation (between Sn and Sp)
- CpGcluster "hits" more islands than the other finder (at least the "core" of the island gets predicted)

→ High specificity and correlation together with highest hit percentage seems to indicate an advantage of CpGcluster

CpGcluster parameters: Median distance and P-value 1E-5

Program	Sn ± SD	Sp ± SD	CC ± SD	Hit* [%] ± SD
Newcpgreport	0.545 ± 0.002	0.973 ± 0.002	0.725 ± 0.005	87.000 ± 0.540
CpGProD	0.918 ± 0.003	0.657 ± 0.003	0.772 ± 0.006	94.675 ± 0.808
CpGIS	0.832 ± 0.003	0.756 ± 0.007	0.789 ± 0.013	86.675 ± 1.528
CpGIE	0.910 ± 0.002	0.667 ± 0.003	0.775 ± 0.006	94.650 ± 0.810
CpGED	0.819 ± 0.013	0.584 ± 0.004	0.685 ± 0.005	84.075 ± 1.191
CpGcluster (d_t = median, or 44 bp)	0.655 ± 0.003	0.976 ± 0.005	0.797 ± 0.009	95.475 ± 0.870
CpGcluster (d_t = 75th percentile, or 94 bp)	0.866 ± 0.006	0.832 ± 0.009	0.846 ± 0.006	95.050 ± 0.643

Statistical properties compared to traditional finders

- Higher mean G+C content, O/E ratios and CpG densities
- Lower mean and maximum length
- Lower overlap with spurious Alu elements
- Higher overlap with conserved phylogenetic elements (PhastCons)
- Detects short CpG islands in tissue specific genes which are normally missed by other finders

	CpGcluster	CpGproD
Genome length (without N-runs, bp)	2.85E+09	2.85E+09
Total number of CpGs	28,073,991	28,073,991
CpG-dinucleotides in CpG-islands (%)	4,489,575 (15.99)	4,323,799 (15.40)
Number of islands predicted	197,727	76,793
*Island coverage (%)	1.90	2.81
Island length (bp):		
Average	273.5 ± 246.7	1043.8 ± 761.7
Minimum	8	500
Maximum	7,774	42,276
Average island GC-content (%)	63.76 ± 7.51	54.58 ± 6.12
Average CpG O/E ratio	0.855 ± 0.265	0.636 ± 0.089
Average CpG-density	0.087 ± 0.041	0.047 ± 0.016

	Overlap with TSS of MAGE genes		% of overlap with	
Program	#CGI	Average length ± SD	Alus	PhastCons
newcpgreport	2	271.0 ± 18.4	19.49	23.73
CpGProD	3	1,314.3 ± 525.1	23.40	13.31
CpGIS	3	800.0 ± 243.3	10.52	20.59
CpGIE	3	1,093.0 ± 476.1	23.99	14.00
CpGED	2	730.5 ± 320.3	15.32	15.82
CpGcluster	8	258.3 ± 100.8	6.79	28.53

CpGcluster - Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, Oliver JL. 2006. BMC Bioinformatics 7: 446

Many other genome elements are also clustered…

Table 1. Clusters of genome elements pertaining to ten different categories in the human genome (hg19)

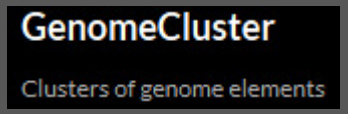
Genome entity	Number of elements	Number of clusters	Elements forming clusters (%)	Mean cluster length (bp) ± SD	Mean number of elements by cluster ± SD
Genes	19152	206	4408 (23%)	441645 ± 294208	21 ± 13
Exons	198933	5089	78425 (39%)	16035 ± 10234	15 ± 11
Introns	179781	5178	73506 (41%)	14692 ± 9778	14 ± 10
CpG islands	204834	5563	44408 (22%)	3384 ± 3143	8 ± 6
TFBSs	4380444	160519	2707380 (62%)	230 ± 158	17 ± 15
Enhancers	318454	25944	176925 (56%)	649 ± 672	7 ± 3
DNase sites	1281988	5838	121274 (9%)	10214 ± 3914	21 ± 7
Alus	1175329	8020	158121 (13%)	8140 ± 4593	20 ± 10
LINE1	1480420	898	13860 (1%)	5226 ± 2522	15 ± 6
SNPs	55448579	206858	2146020 (4%)	25 ± 84	10 ± 61

GenomeCluster - Dios F., G. Barturen, R. Lebrón, A. Rueda, M. Hackenberg and J.L. Oliver. 2014. **Computational Biology and Chemistry** (in press).

http://bioinfo2.ugr.es/CpGcluster/

http://bioinfo2.ugr.es/wordCluster/wordCluster.php

http://bioinfo2.ugr.es/CpGislands/



http://bioinfo2.ugr.es/GenomeCluster/