ПРОБЛЕМА ВЫДЕЛЕНИЯ CpG-ОСТРОВОВ И ГОМОЛОГИЧНЫХ ИМ СТРУКТУР В РАЗНЫХ ГРУППАХ ОРГАНИЗМОВ.

Литературный обзор

CpG-islands in vertebrate genomes.

M.Gardiner-Garden, M.Frommer
(J.Mol.Biol.,1987, v.196. pp.261-282)

- **Observed/Expected ratio > 0.60**
- **Percent C + Percent G > 50.00**
- **Length > 200**
- Теперь эти параметры обычно ужесточают – но это не всегда правильно!
- Обычно неметилированы или слабо метилированы
- Отмечалось также наличие последовательностей GGGCGG, особенно в промоторной области гена и вниз по течению вдали от 5’конца. В промоторах это сайт связывания фактора Sp1 (единственный, связывающийся независимо от метилирования CpG-пары и способный разметилировать её).

Варианты расположения CpG-островов

- Связанные с промотором и первым экзоном
- Внутригенные, не захватывающие старт транскрипции
- Захватывающие последний экзон и 3’UTR
- Межгенные

- Некоторые острова могут захватывать практически весь ген

Assessment of clusters of transcription factor binding sites in relationship to human promoter, CpG islands and gene expression

(Katsuhiko Murakami, Toshio Kojima and Yoshiyuki Sakaki, 2004, BBC Genomics, 5:16)

- **Some CGI-related PWMs**
- Анализ коэффициентов корреляции показал, что выделяется группа факторов, скоррелированных с CpG-островами, а уже за счёт этого – с промотором.

Dualism of gene GC-content and CpG pattern in regards to expression in the human genome: magnitude versus breadth

A.E.Vinogradov
(Genome Analysis, 2005)

- %GC **для гена в целом** связан с максимальным уровнем экспрессии в тканях

- Наличие 5’-концевого острова и %CpG **для гена в целом** коррелирует с шириной спектра экспрессии в разных тканях. Первое подтверждено также в статье Loïc Ponger, Laurent Duret and Dominique Mouchiroud (см. далее).

Наличие 5’-острова коррелирует с шириной спектра экспрессии и для растений.

- Gene-associated CpG Islands and the Expression Pattern of Genes
- in Rice
- Ikuo Ashikawa.
- DNA Research, 2002, v.9, pp.131–134
- Экспрессия в двух или более тканях или экспрессия в каллусе коррелирует с наличием 5’-концевого острова.

Comprehensive analysis of the base composition around the transcription start site in Metazoa

Stein Aerts, Gert Thijs, Michal Dabrowski, Yves Moreau, Bart De Moor (*BMC Genomics* 2004, **5**:34)

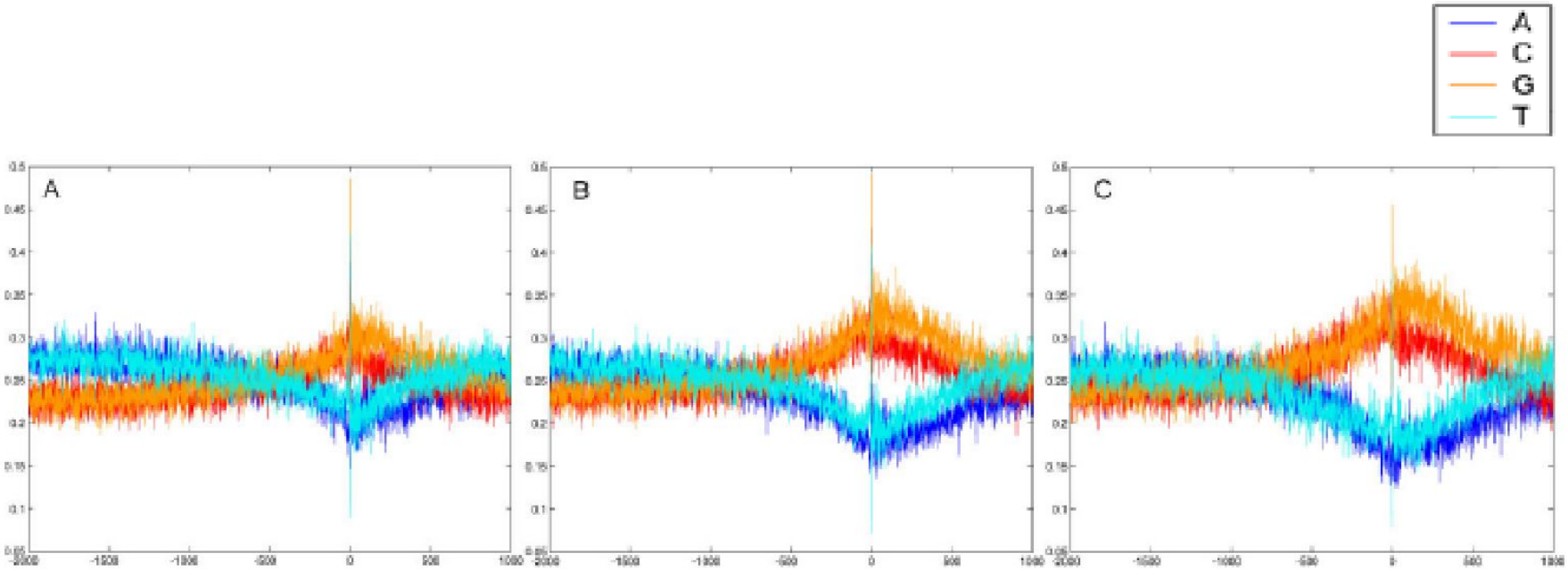


Figure 5
Nucleotide frequencies of three human gene groups: genes with a narrow expression pattern (A), a medium pattern (B), and a wide pattern(C).

Зачем нужны внутригенные острова?

- Genes and Transposons Are Differentially
- Methylated in Plants, but Not in Mammals
- Pablo D. Rabinowicz, Lance E. Palmer, Bruce P. May, Michael T. Hemann,
- Scott W. Lowe, W. Richard McCombie, and Robert A. Martienssen
- Genome Research, 2003, v.13, pp.2658–2664
- Внутренние экзоны генов млекопитающих обычно метилированы, в то время как внутренние экзоны генов высших растений – нет.

Distribution and Characterization of Regulatory Elements in the Human Genome

(Jacek Majewski and Jurg Ott
2002, Genome Research, 12:1827-1836)

GGG- энхансер сплайсинга

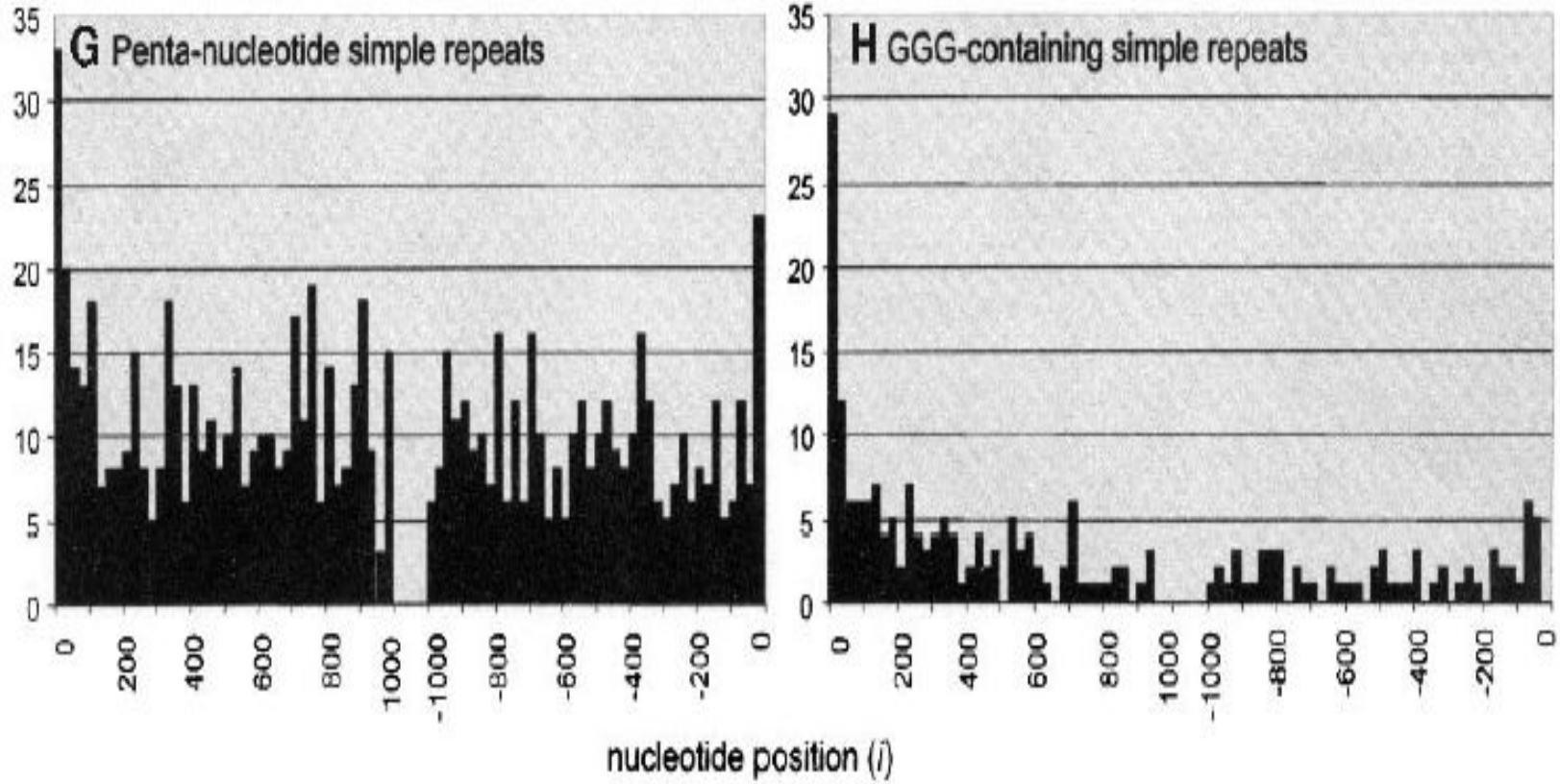


Figure 4 Distribution of simple tandem repeats (1–7-nucleotide repeat unit) and low-complexity regions in introns. Panel (*A*) shows the combined distribution of all such regions, whereas the remaining panels identify the particular types of sequences that most significantly contribute to the overrepresentation near splice sites. Such sequences are most likely involved in splicing recognition and control.

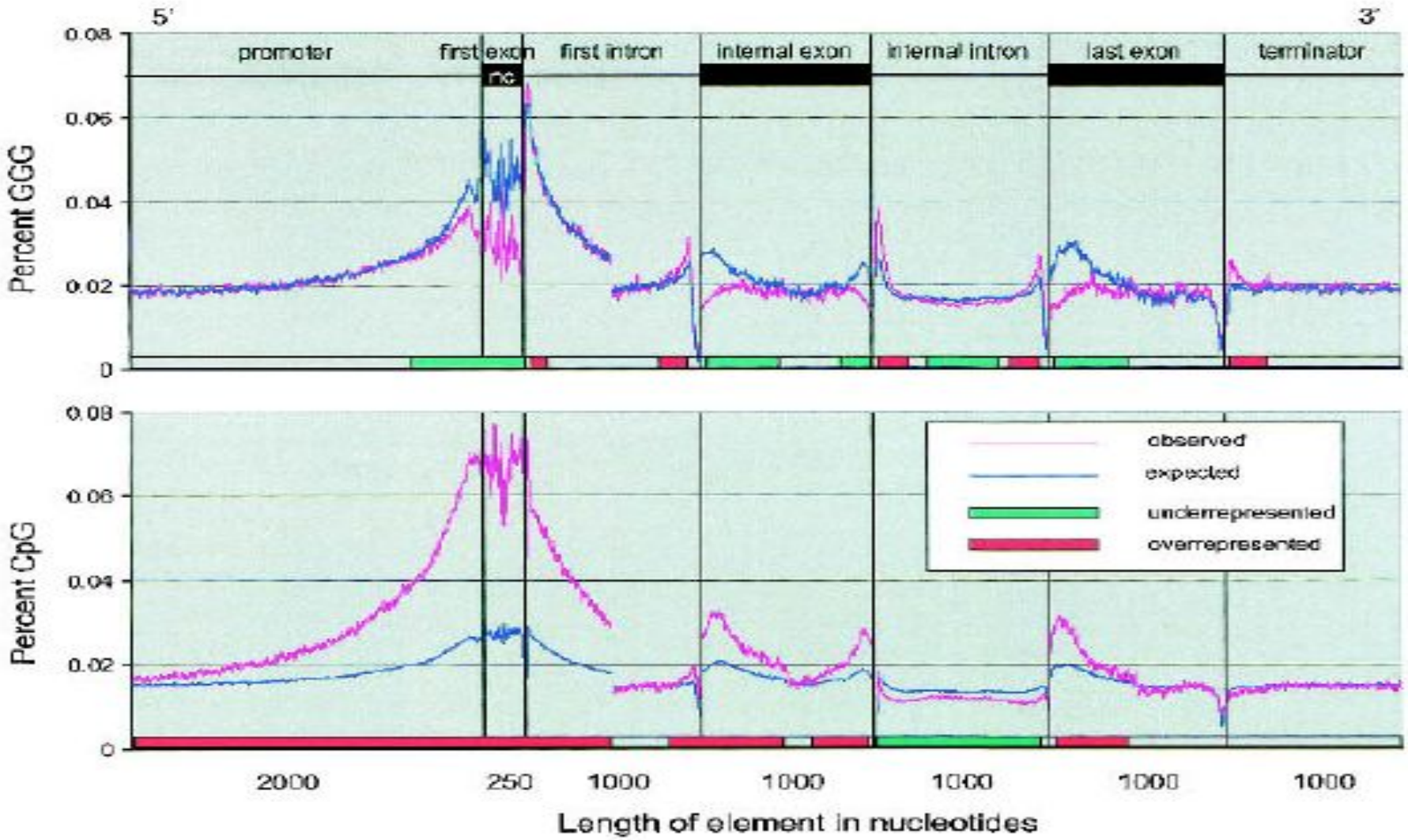


Figure 5 Observed and expected frequencies of common regulatory elements, the GGG trinucleotide and the CpG dinucleotide, in a typical human gene. The model gene contains information combined from all known genes in the entire human genome. The expected frequencies are calculated from local individual nucleotide frequencies, corrected for biases in occurrence of GGG and CpG in noncoding genomic regions. The slider bar at the bottom of each graph indicates the regions of relative over- and underrepresentation of each motif. Excess of a given motif over the expected frequency indicates a possible regulatory function.

Маевский и Отт – возможная роль CpG в сплайсинге, отсюда и падение в 3’-конце последнего экзона, который не сплайсируется. Вообще в интронах CpG-острова мало представлены (см. презентацию Ю. Медведевой), но GGG-тринуклеотиды, характерные для CpG-островов – на самой границе интронов и экзонов.
%CpG для гена в целом в работе Виноградова, который коррелировал с шириной экспрессии, это, по существу, CpG внутригенных островов.

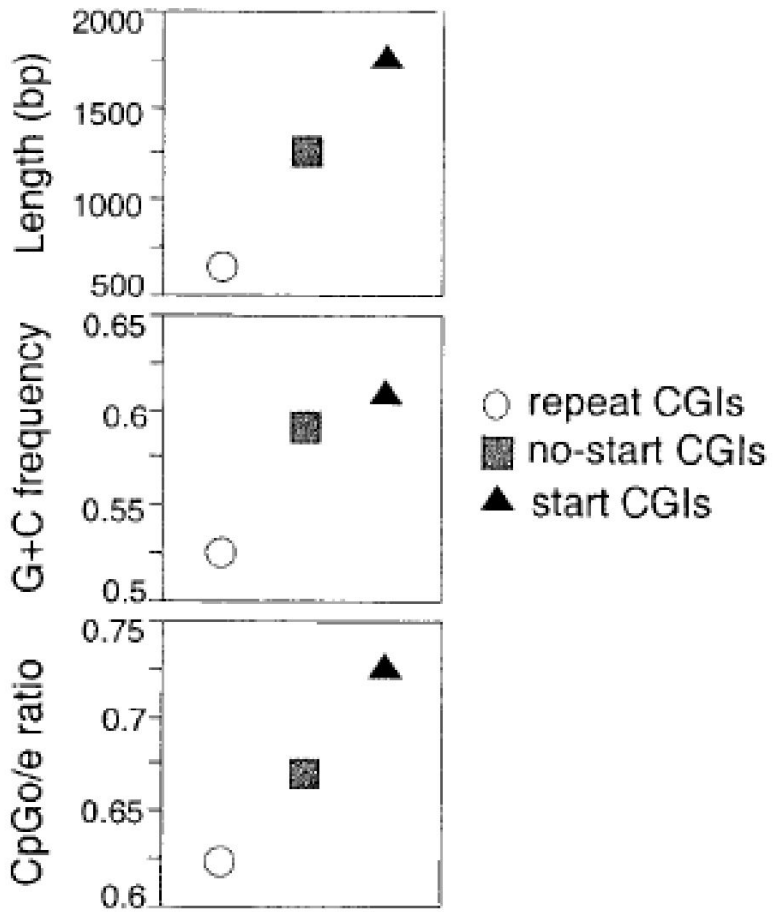
PEG3 (ZIM2)

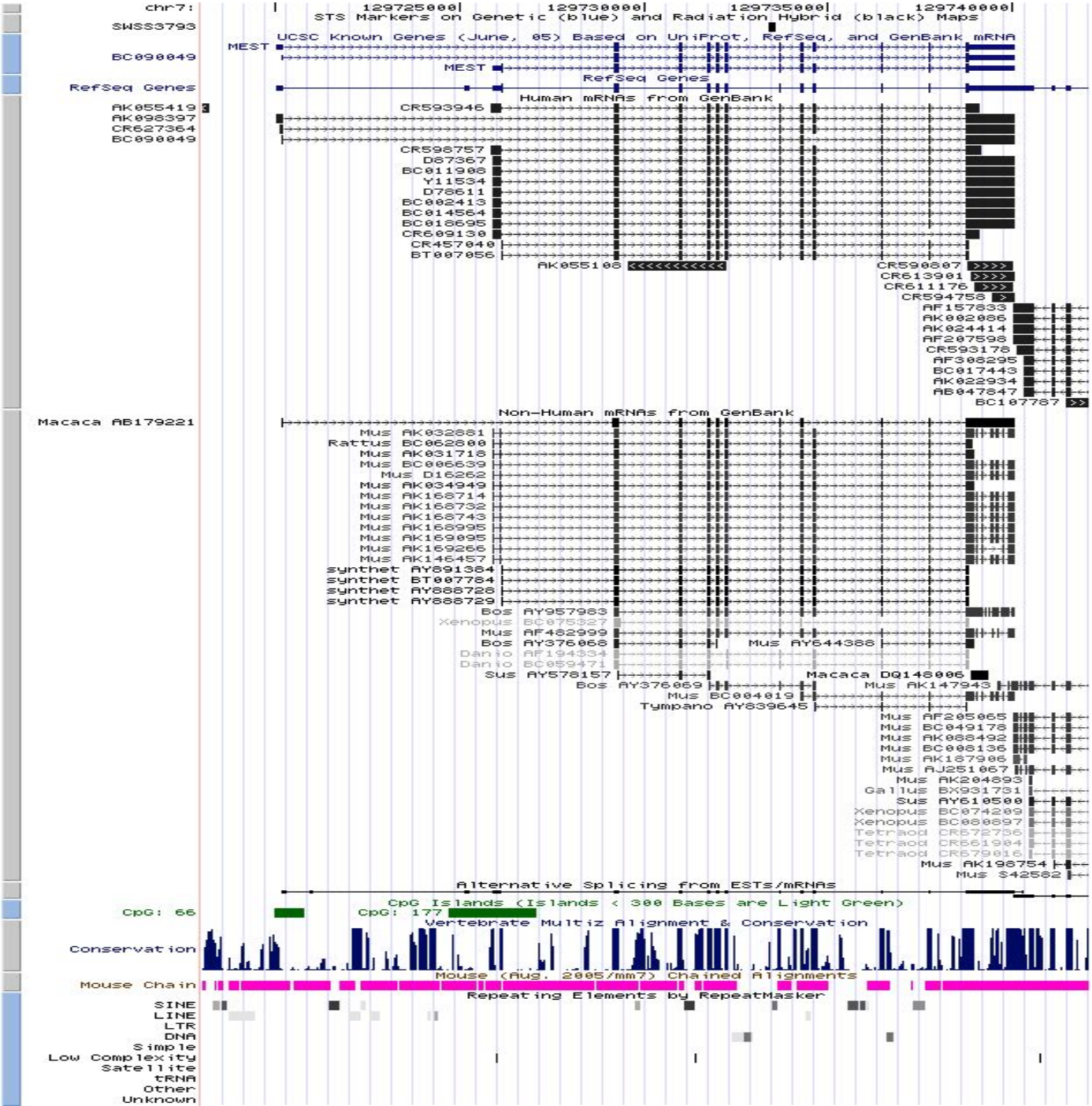
- Импринтируемый ген.
- Обе изоформы – отцовские, но тканеспецифичность разная
- PEG3 - 7 первых экзонов и 2 следующих за ними
- ZIM2 – 7 первых экзонов и 4 (5), идущих после экзонов гена PEG3
- Между экзонами генов PEG3 и ZIM2 – небольшой остров, частично метилированный. Влияет на сплайсинг, регулируя скорость транскрипции?

Determinants of CpG Islands: Expression in Early Embryo and Isochore Structure

Loïc Ponger, Laurent Duret and Dominique Mouchiroud

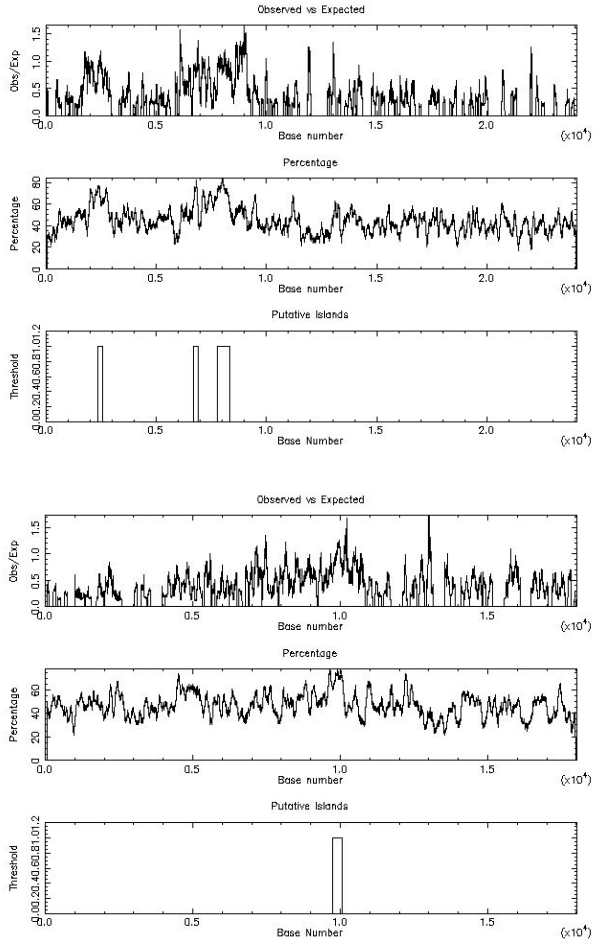
(*Genome Res.* 2001 11: 1854-1860)





- **MEST-импринтированный ген**
- Одна изоформа с отцовской экспрессией,
- другая – с биаллельной
- CpG-острова по всей видимости функциональны

CpG-острова в гене MEST (PEG1) у человека и мыши.



- У человека начало фрагмента в 2 т.п.н. выше по течению от первого экзона, у мыши в 10 т.п.н. выше по течению от первого экзона.
- Параметры поиска
- **Observed/Expected ratio > 0.70**
  Percent C + Percent G > 60.00
  Length > 200
- У человека оба острова находятся и при более жёстких параметрах
- **Observed/Expected ratio > 0.80**
- У мыши один из островов находится при тех же параметрах, другого не видно и при более мягких.

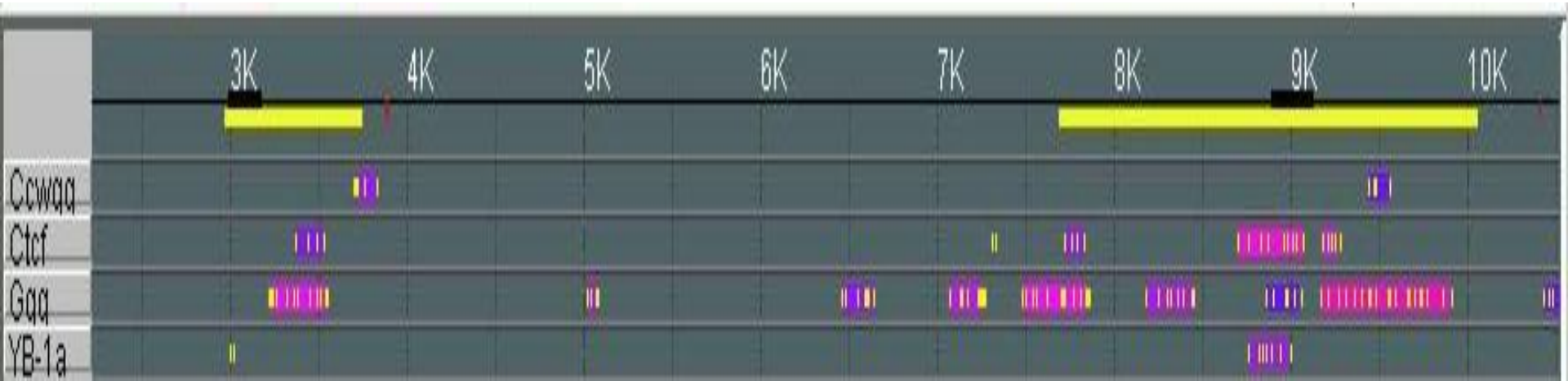
Выводы?

- Ужесточение критериев поиска не всегда целесообразно даже для 5’-концевых островов, хотя и позволяет отбросить повторы. Искать не только по статистическим критериям, но, например, по консервативности и характерным мотивам?
- Дополнительный старт – дополнительный CpG остров? В презентации Н.Опариной увидим, что так бывает, но сравнительно редко.

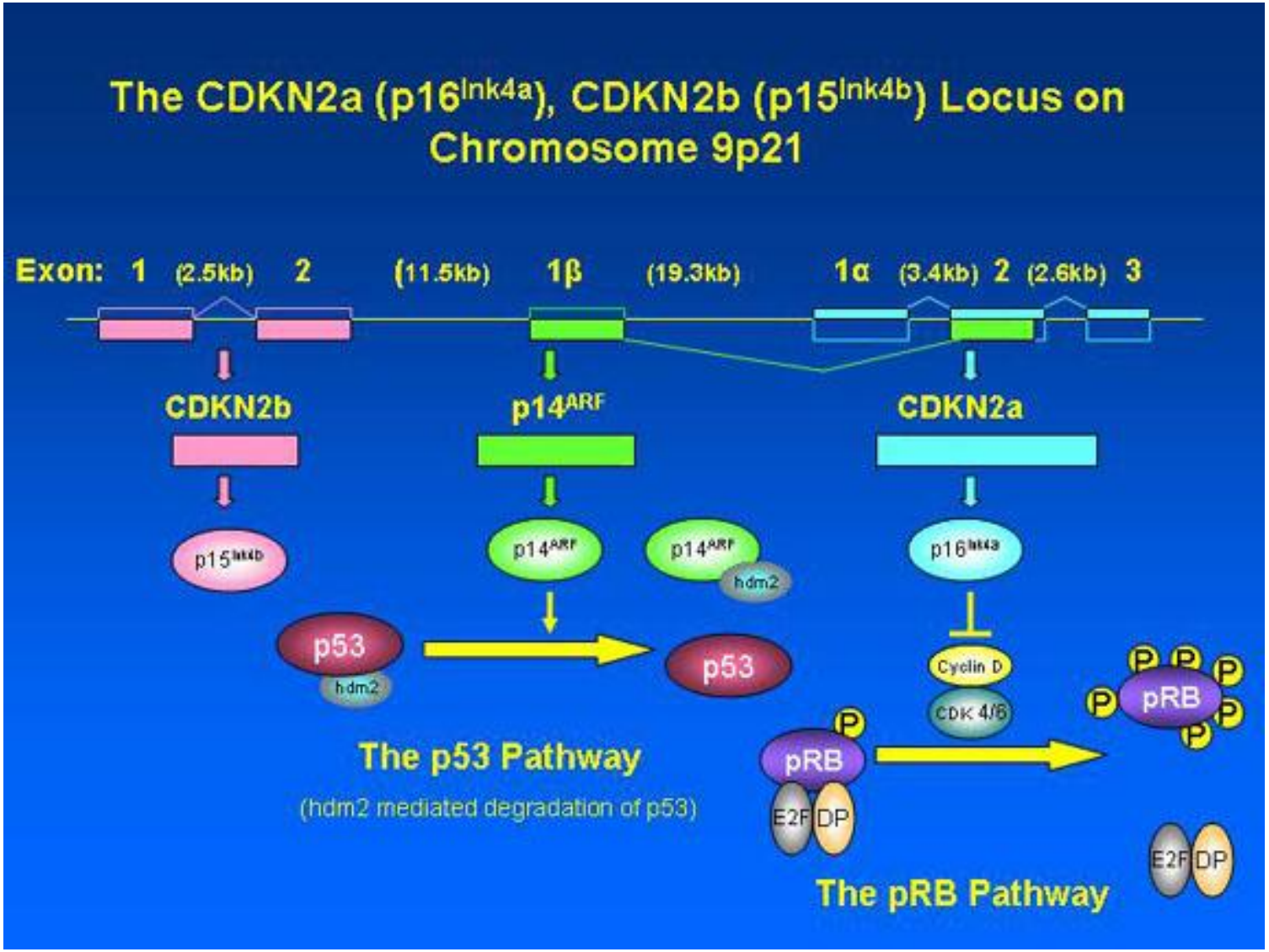
Структура CpG-островов и их флангов, связанных с альтернативными стартами, в гене MEST (PEG1)

Скопления мотивов (локализовано программой А.Миронова).

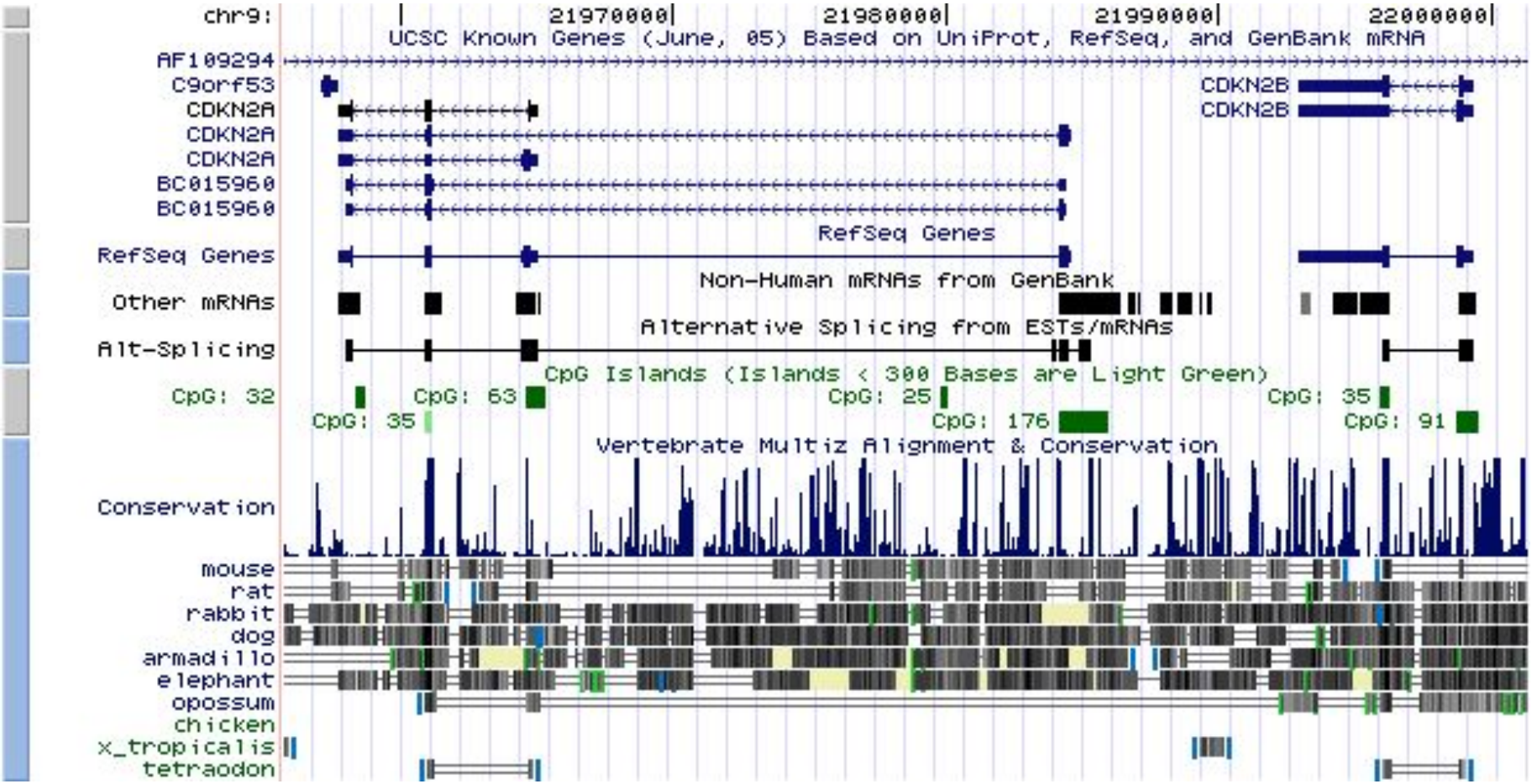
- CCWGG – фланкирует острова, метилируется
- CTCF-консенсус – в неметилированном виде связывает CTCF
- GGG – энхансер сплайсинга
- YB-1-консенсус для YB-1-белка холодового шока. Это транскрипционный фактор, связывание которого блокируется CTCF. Связывается с экзонными энхансерами. Направляет альтернативный сплайсинг и выбор первого экзона. Входит в состав м-РНП, связывается с 5’-кэпом, влияет на трансляцию и время жизни РНК



Биологическая роль продуктов локусов CDKN2a и CDKN2b



Гены CDKN2a и CDKN2b



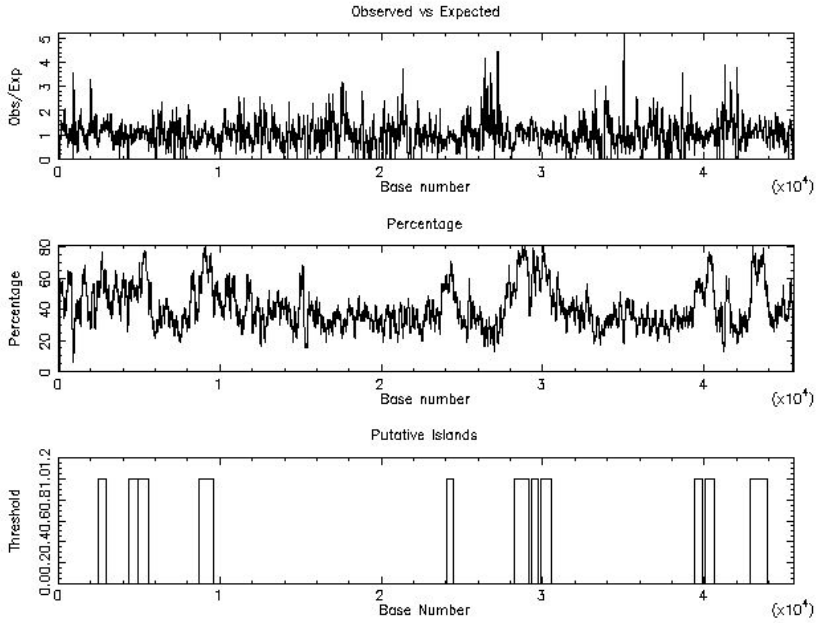
p14 и p16 считываются с альтернативных стартов в разных рамках. CpG-острова, перекрывающиеся со стартом для p14, стартом для p16 и в экзоне – «хорошие» . Показано их связывание с MeCP2 при метилировании и независимое влияние метилирования каждого из них на транскрипцию соответствующего продукта в разных видах опухолей.

Тем не менее, и они выделяются только при использовании достаточно «мягкого» критерия. Соответствующие параметры, однако, совпадают для разных видов млекопитающих, у которых есть оба эти продукта.

Структура флангов CpG-острова, перекрывающегося с 1 α-экзоном гена CDKN2a

- **Наиболее интересен «старый» остров, перекрывающийся со стартом p16 (более «старым»)**
- CCWGG – фланкирует острова, метилируется
- CTCF-консенсус – в неметилированном виде связывает CTCF
- GGG – энхансер сплайсинга
- YB-1-консенсус – связывает YB-1
- **5% альтернативно сплайсируемых генов имеют ARF. Гомологичны обычно у приматов, иногда в пределах млекопитающих. Позволяет исключить вопрос о функциональной значимости продукта. Хорошая модель для изучения альтернативных стартов?**

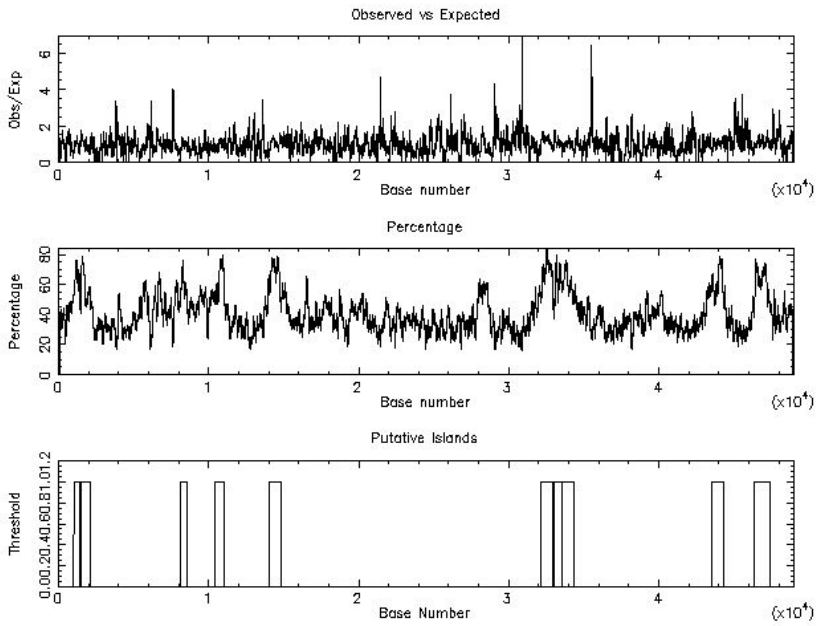
Структура CpG-островов локуса p15-p16 у человека, параметры поиска.



- комплемент
- **Observed/Expected ratio > 0.60**
  Percent C + Percent G > 50.00
- **Length > 400**
- «старый» остров при P16 находится лишь при наиболее «мягких» критериях

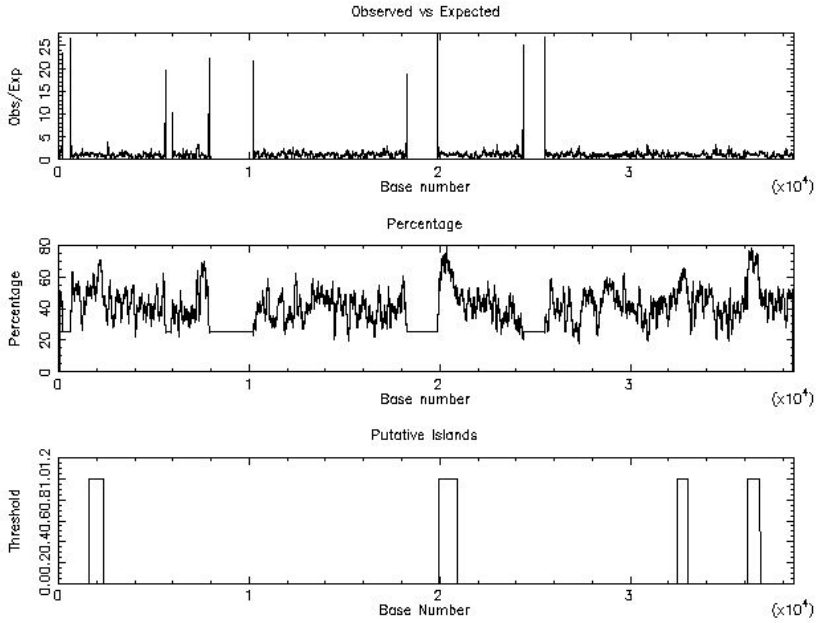
- Length 540 (2457..2996)
- Length 556 (4363..4918)
- Length 666 (4921..5586)
- Length 897 (8684..9580)
- Length 469 (24036..24504)
- Length 897 (28246..29142)
- Length 470 (29283..29752)
- Length 663 (29844..30506)
- Length 497 (39417..39913)
- Length 549 (40057..40605)
- Length 1011 (42889..43899)

Структура CpG-островов локуса p15-p16 у быка, параметры поиска.



- Комплемент
- **Observed/Expected ratio > 0.60**
- **Percent C + Percent G > 50.00**
- **Length > 400**
- Length 420 (1017..1436)
- Length 604 (1473..2076)
- Length 446 (8133..8578)
- Length 662 (10415..11076)
- Length 747 (14064..14810)
- Length 828 (32117..32944)
- Length 500 (33026..33525)
- Length 733 (33583..34315)
- Length 773 (43552..44324)
- Length 1027 (46366..47392)

Структура CpG-островов локуса p15-p16 у крысы, параметры поиска.



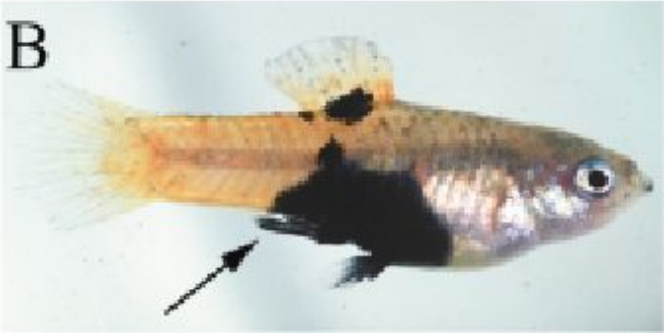
- Комплемент
- **Observed/Expected ratio > 0.60**
  Percent C + Percent G > 50.00
- **Length > 400**
- у крысы на месте CpG-острова для p16 сборка плохая (у мыши ещё хуже)
- Length 707 (1624..2330)
- Length 1010 (19932..20941)
- Length 568 (32464..33031)
- Length 683 (36115..36797

Тут опять ужесточение критериев поиска помешало бы увидеть биологически значимый, эволюционно более древний и консервативный CpG-остров, содержащий биологически значимые мотивы.

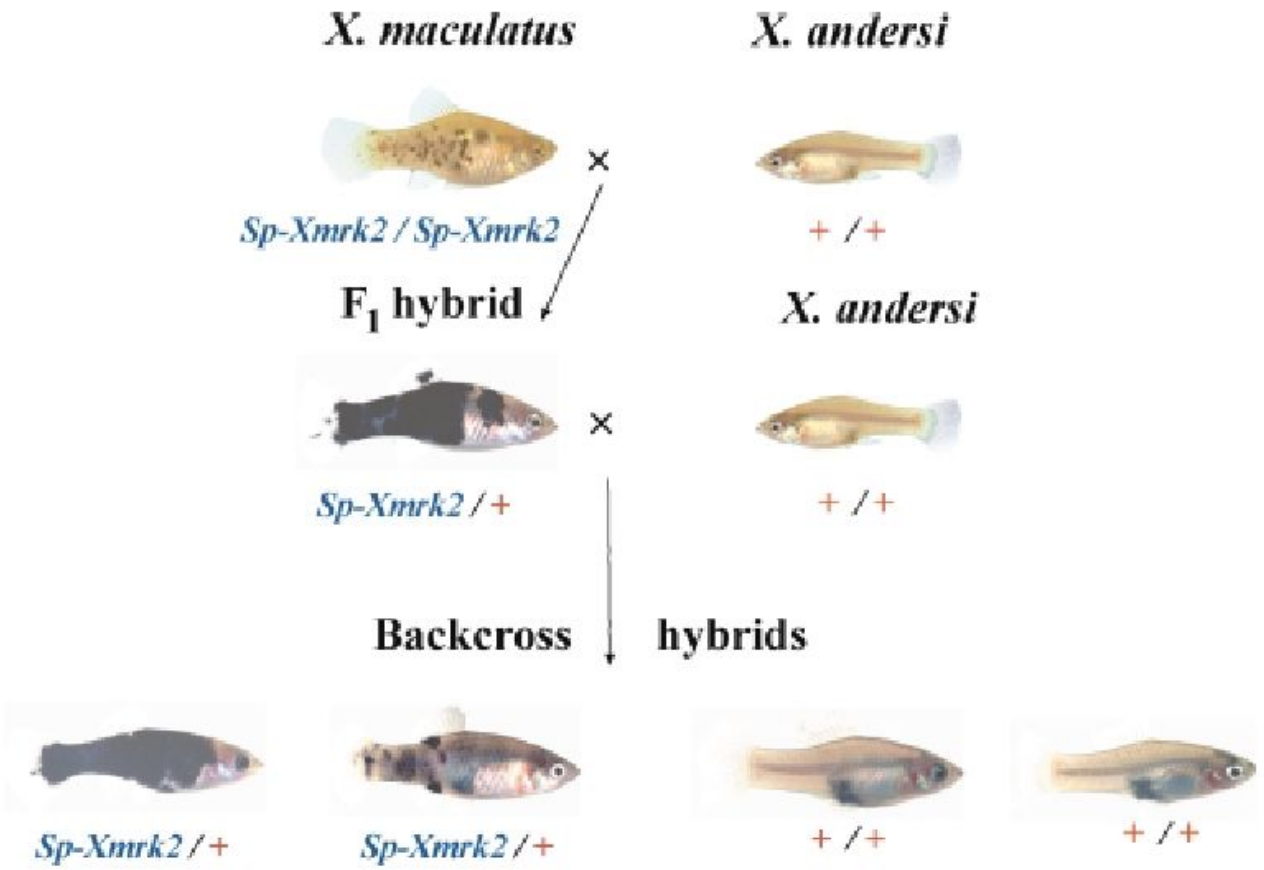
Обращает на себя внимание близость параметров, при которых обнаруживается аналогичная структура островов у разных млекопитающих.

Меланома у меченосцев

связана с геном CDKNX, дупликация которого дала p15 и p16. По функциям ближе к p16.



Провоцируется межлинейными и межвидовыми скрещиваниями



Характеристики CpG-островов гена CDKNX у *Xiphophorus helleri*

Вновь видим острова, перекрывающиеся со стартом p16 и с его экзоном. Они обнаруживаются при более жестких параметрах поиска.
(по работе S.Kazianis, L.Della Coletta, D.C.Morizot, D.A.Johnston, E.A.Osterndorff, R.S.Nairn, 2000, Carcinogenesis, v.21, n.4, pp.599-605)

- Первый CpG-остров
- Захватывает 5’-конец гена и первый экзон
- CpG/GpC ratio = 0,98
- Percent C + Percent G = 51,9%
- Length = 538bp
- 33 CpG
- Неметилирован в норме, слабо метилирован в меланомах

- Второй CpG-остров
- Расположен ниже по течению
- CpG/GpC ratio = 0,99
- Percent C + Percent G = 57,5%
- Length = 317bp
- 26 CpG

- Однако РНК, считанная с соответствующего гена, в меланомах экспрессируется **избыточно**

Особенности CpG-островов у рыб

- **Primordial stage?**
- **Острова короче, а повышения содержания G+C нет, хотя есть повышение содержания CpG и внутри острова нет метилирования.**
- Это показано также, например, для MTF-1 гена Fugu.
- **5’UTR у рыб короче, а кодирующий район начинается непосредственно после старта транскрипции. Небольшая длина островов связана с этим? А может быть самые древние, ещё рыбьи острова – самые короткие?**
- MEST тоже есть у рыб – старт трансляции соответствует неимпринтированной изоформе, а соответствующий CpG-остров метилирован для обоих аллелей.

Возможна ли классификация островов по метилированности?

CpG Island Methylation in Human Lymphocytes Is Highly Correlated with DNA Sequence, Repeats, and Predicted DNA Structure

(Christoph Bock, Martina Paulsen, Sascha Tierling, Thomas Mikeska, Thomas Lengauer, Jorn Walter
PloS Genetics, 2006, v.2, i.3, e.26)

влияет также структура ДНК на флангах в 20 т.п.н.

Table 1. DNA-Related Attributes Differ Significantly between Methylated and Unmethylated CpG Islands

Rank	Attribute Name	Attribute Description	Attribute Class	Higher Value for	Single Test Significance
1	SAI_len	Total length of self-alignments (alignments of the human genome against itself)	(2)	Methylated CpG Islands	2.62×10^{-11}
2	SAI_no	Total number of self-alignments	(2)	Methylated CpG Islands	3.23×10^{-10}
3	Pat_CCGC	Chromosome plus-strand pattern frequency of CCGC	(1)	Unmethylated CpG Islands	5.18×10^{-10}
4	Pat_CCCC	Chromosome plus-strand pattern frequency of CCCC	(1)	Unmethylated CpG Islands	1.39×10^{-9}
5	SAI_std	Standard deviation of self-alignment lengths	(2)	Methylated CpG Islands	1.96×10^{-9}
6	Uni_AAAG	Non-strand-specific pattern frequency of AAAG/CTTT	(1)	Unmethylated CpG Islands	8.87×10^{-9}
7	fC_std	Standard deviation of C content distribution	(1)	Unmethylated CpG Islands	9.13×10^{-9}
8	Rise_avg	Average DNA structure rise (as predicted from sequence)	(4)	Methylated CpG Islands	3.82×10^{-8}
9	Pat_CGCC	Chromosome plus-strand pattern frequency of CGCC	(1)	Unmethylated CpG Islands	5.05×10^{-8}
10	Pat_AAAG	Chromosome plus-strand pattern frequency of AAAG	(1)	Unmethylated CpG Islands	7.72×10^{-8}

Есть ли в других группах метилирование CpG и структуры, аналогичные CpG-островам?
Если есть метилирование, то должно быть и смещение частоты CpG (часть метилированных C мутирует в T).

Относительные частоты динуклеотида CpG у разных организмов.

(Andrew J. Gentles and Samuel Karlin
2001 *Genome Res* 11: 540-546 .)

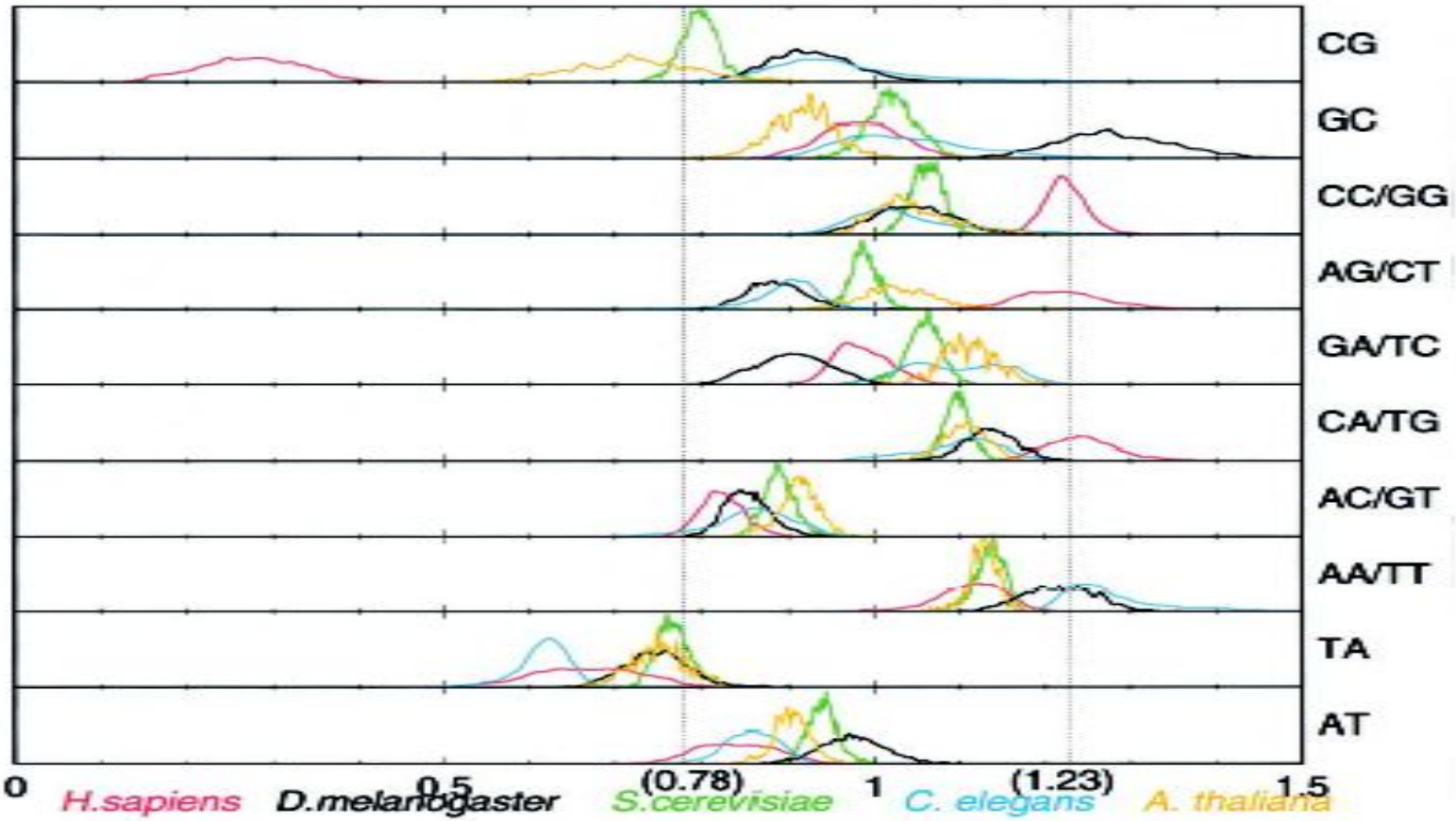


Figure 1 Distribution of ρ^* values for all 50-kb samples from human (red), *Drosophila melanogaster* (black), *Saccharomyces cerevisiae* (green), *Caenorhabditis elegans* (blue), and *Arabidopsis thaliana* (orange).

Метилирование у *Drosophila melanogaster*

- Есть высококонсервативная метилаза Dnmt2. Слабоактивна, метилирует в большей степени CpT и CpA, чем CpG (Dnmt3 млекопитающих тоже может метилировать CpT и CpA, хотя и предпочитает CpG). Мутанты жизнеспособны.
- Есть метилсвязывающий белок MBD2/3. В большей степени связывает CpT и CpA, чем CpG, малоактивен. Мутанты жизнеспособны, нарушено расхождение хромосом. Может действовать как репрессор транскрипции. Может взаимодействовать с белками, участвующими в ремоделинге.
- Есть метилированый C, но его доля 0,05-0,1% у взрослых мух (у млекопитающих 2-10%). Обычно в составе CpT и CpA, но есть и CpG.
- Показано наличие CpG-метилирования и его роль в регуляции активности гена Rbf (Ferres-Marco D., Gutierrez-Garsia I., Vallejo D., Bolivar J., Gutieres-Avino J., Dominguez M. Nature, 2006, v.439, pp.430-434)

Метилирование у других насекомых.

- Метилаза комара *Anopheles gambiae* очень похожа на таковую у дрозофилы.
- У молей *Mamestra brassicae* уровень метилирования CpG достигает такового у млекопитающих, но CpG-островов как скоплений CpG нет.
- У тлей *Myzus persicae* наблюдаются скопления CpG. Однако для амплифицированного эстеразного гена тлей их метилирование коррелирует скорее с активацией гена.
- У кокцид корреляция инактивации одного из хромосомных наборов (отцовского) с метилированием хромосом тоже скорее обратная.
- (по работам F.Lyko и других авторов)

Comprehensive analysis of the base composition around the transcription start site in Metazoa

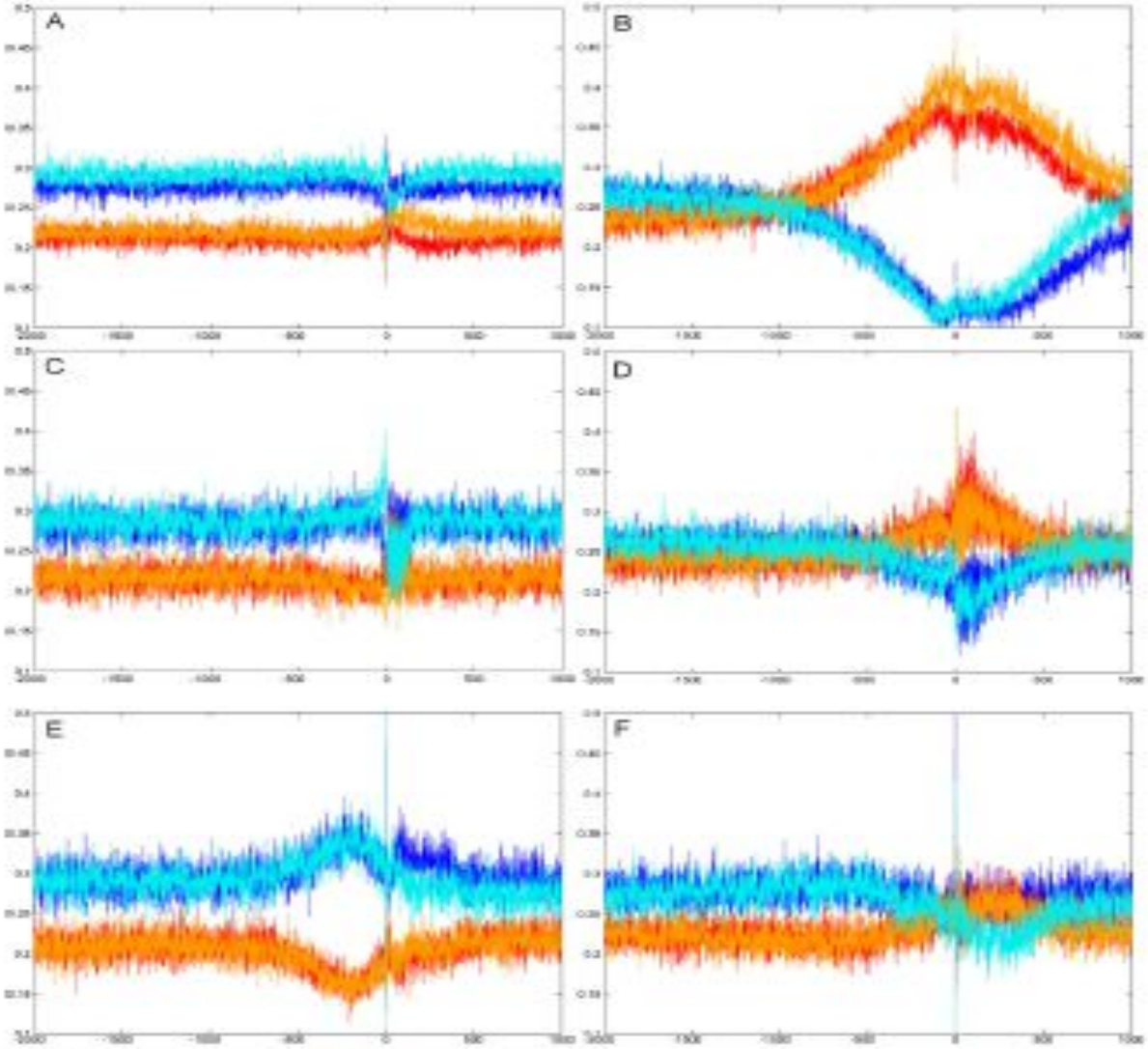


Figure 4
Nucleotide frequencies of several gene classes, separated according to the concentration of a dinucleotide in the [-400, 400] region around the TSS. A. Human genes with few CpG doublets. B. Human genes with many CpG doublets. C. Fugu genes with few CpGs. D. Fugu genes with many CpGs. E. Fly genes with many ApTs. F. Fly genes with few ApTs.

- У дрозофилы в регуляции транскрипции участвует целое семейство Sp-факторов. Их консенсус GC-богат

Особенности метилирования у растений.

- Есть не только ферменты, которые метилируют CpG, но также метилирующие CpNpG и CpX.
- Не только CpG, но и CpNpG острова существуют как отдельные структуры.
- Есть аналог геномного импринтинга у млекопитающих, хотя он связан скорее с разметилированием материнских аллелей.
- Внутренние экзоны генов обычно неметилированы. Тем не менее, видим те же варианты расположения островов.
- Малая длина островов: 0,2-1.0 kb vs 0,8-1,2 kb у млекопитающих (мы видели, что бывает и меньше). Остальные критерии сходны: наблюдается как повышение %GC, так и повышение O/E. Острова неметилированы.

Gene-associated CpG islands in plants as revealed by analyses of genomic sequences

Ikuo Ashikawa

The Plant Journal (2001) **26**(6), 617–625

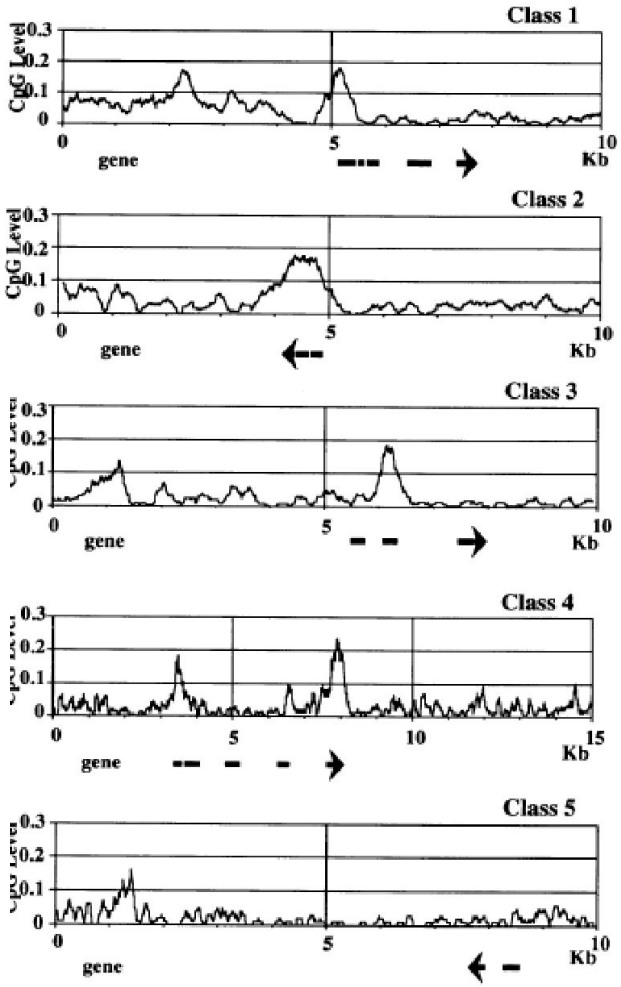


Figure 2. The CpG frequency graphs of five classes of plant genes, grouped according to the position of the associated CpG-rich cluster.
In class 1 genes, the CpG-rich cluster is located only at the 5′-end of the gene. The CpG-rich cluster covers the whole gene region of class 2 genes. Class 3 genes comprise those in which the CpG-rich cluster occurs at considerable distance downstream of the 5′-end of the gene. Class 4 genes contain a CpG-rich cluster at the 5′-end of the gene and another cluster downstream of the 5′-end of the gene. Class 5 comprises genes that lack CpG clusters. Genes representative of each class were selected from the analysed rice genomic sequences as examples and are shown with their CpG distribution patterns.

Table 2. CpG methylation status in restriction sites located in class 1 and class 2 CpG clusters and in the DNA regions outside CpG clusters

DNA region	CpG methylation level[a]	Class[b]
2487-1	0/3	1
2816-1	0/3	1
1550-1	1/3	1
969-4	4/5	1
969-3	0/4	1
1081-10	0/3	1
815-2	0/2	1
815-3	1/3	1
1080-15	0/1	1
815-17	0/3	2
836-11	2/5	2
969-9	3/3	2
1073-1	1/3	2
1080-5	3/3	2
1081-16	2/5	2
1081-17	1/4	2
1083-2	4/4	2
1366-2	5/5	2
815-2	3/3	D
969-1	2/3	D
1072-1	1/1	D
1080-2	2/3	D
1081-5-1	3/3	D
1366-1	1/1	D
1383-1	1/2	D

[a]CpG methylation level = number of methylated CpGs/total number of CpGs examined.
[b]D, DNA regions depleted of CpG dinucleotides.

Для выделения CpG-островов у растений необходим пересчёт параметров для типичного для них %GC и отношения CpG/GpC.

Важно использовать также критерий консерватизма выделяемых островов у разных видов.

В работе Ashikawa

это было сделано удачно. В цитируемой ниже работе из-за отсутствия подобного подхода получены лишь общие оценки.

По I.Ashikawa

Table 1. Overall G + C content and the frequencies of CpC, CpG, GpC and GpG calculated from the rice, *A. thaliana*, sorghum, maize, barley and human DNA sequences[a]

	C + G	CpC	CpG	GpC	GpG
Rice	0.43	0.041	0.039	0.051	0.040
A. thaliana	0.36	0.029	0.022	0.030	0.029
Sorghum	0.43	0.042	0.035	0.051	0.040
Maize	0.45	0.043	0.039	0.055	0.044
Barley	0.46	0.048	0.041	0.052	0.044
Human	0.51	0.057	0.022	0.062	0.059

Computational Approaches to Identify Promoters and cis-Regulatory Elements in Plant Genomes[1]

Stephane Rombauts[2], Kobe Florquin[2], Magali Lescot, Kathleen Marchal, Pierre Rouzé*, and Yves Van de Peer

- **Plant Physiol, 2003, v.132, pp.1162-1176**

- **Авторы использовали длину окна в 200 нуклеотидов.**

Table II. *Percentage of genes (out of 5,025) containing CpG (top) or CpNpG (bottom) islands, for a few different parameter settings*

Additional values for other parameter settings can be found at http://www.psb.rug.ac.be/bioinformatics/.

CG% Threshold	o/e Threshold	With Island in Promoter	With Island in at Least One Exon[a]	With Island in Promoter AND in Exons	With Island in Promoter ONLY	With Island in Exons ONLY
CpG						
39	0.6	85.75	99.42	85.73	0.02	13.69
42	1.6	0.82	98.95	0.82	0.00	98.13
CpNpG						
39	0.6	64.98	99.38	64.84	0.14	34.47
45	1.6	0.02	98.73	0.02	0.00	98.71

[a]Refer to exons either in coding regions or in the UTR region.