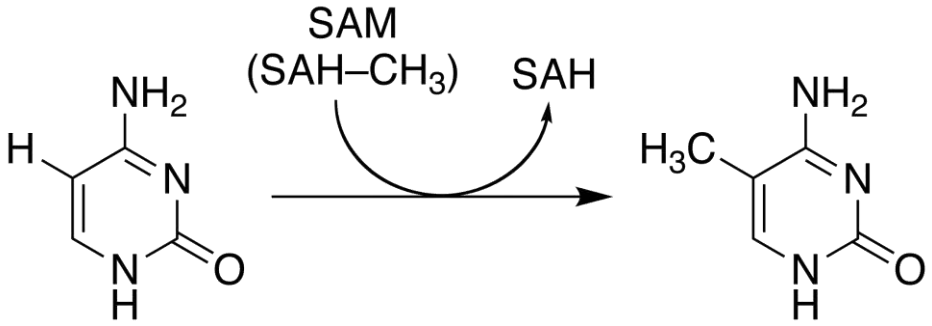
CpG islands

Антон Брагин

Научный руководитель: Николай Вяххи, СПбАУ РАН

CpG островки: структура и функция

5’-mCG-3’



CpG:

- 70 – 80% CpG в геноме человека метилированы
- Репрессия генов (10^6 у позвоночных vs 10^3 у бактерий)
- Активация генов (метилирование инсуляторов)
- Импринтинг генов

CpG islands:

- ~20000 в геноме позвоночных
- Часто локализованы в промоторах генов домашнего хозяйства

CpG островки: определение и поиск

A CGI is defined as a region of at least 200 bp, with the proportion of Gs or Cs, referred to as “GC content,” greater than 50%, and observed to expected CpG ratio (O/E) greater than 0.6

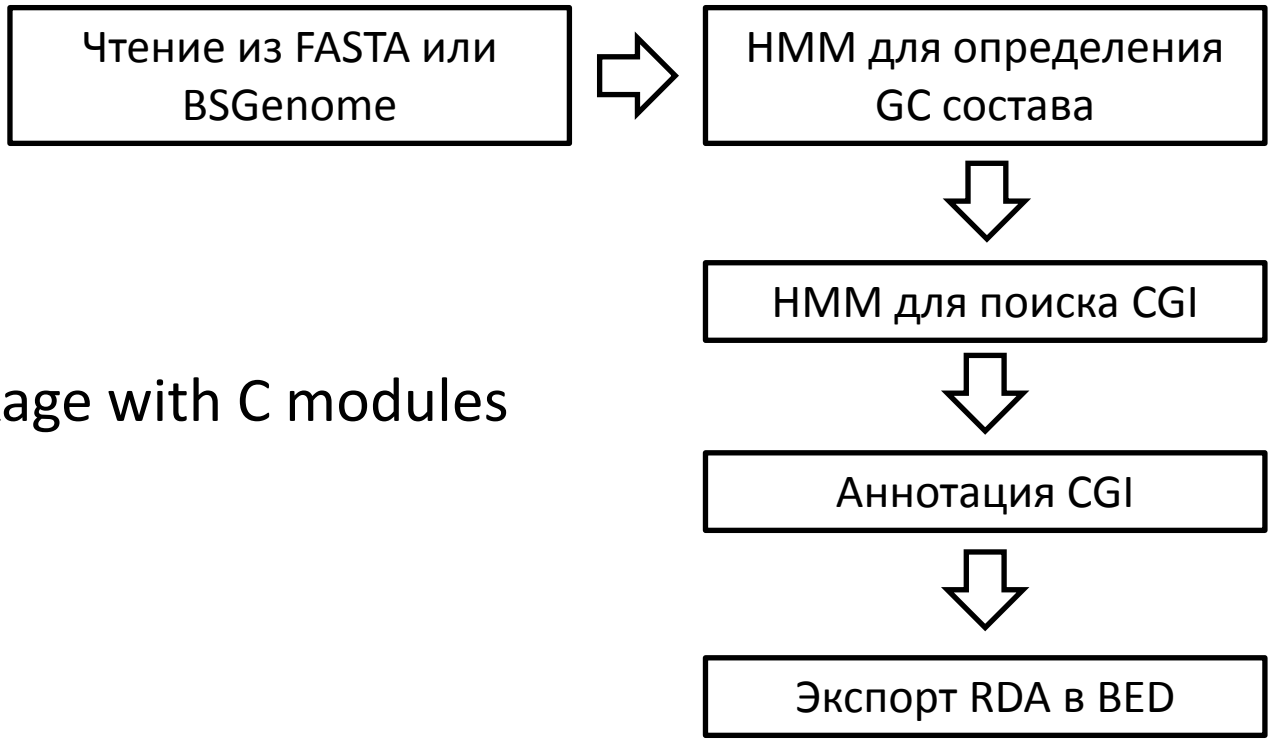
Gardiner-Garden and Frommer, 1987

Задача: поиск CpG островков в геномах позвоночных

Минусы:

- Получены эмпирически без строгого биологического или статистического обоснования
- Полученные островки могут не обладать ожидаемыми функциями (Alu)

makeCGI – инструмент для поиска CpG, основанный на HMM

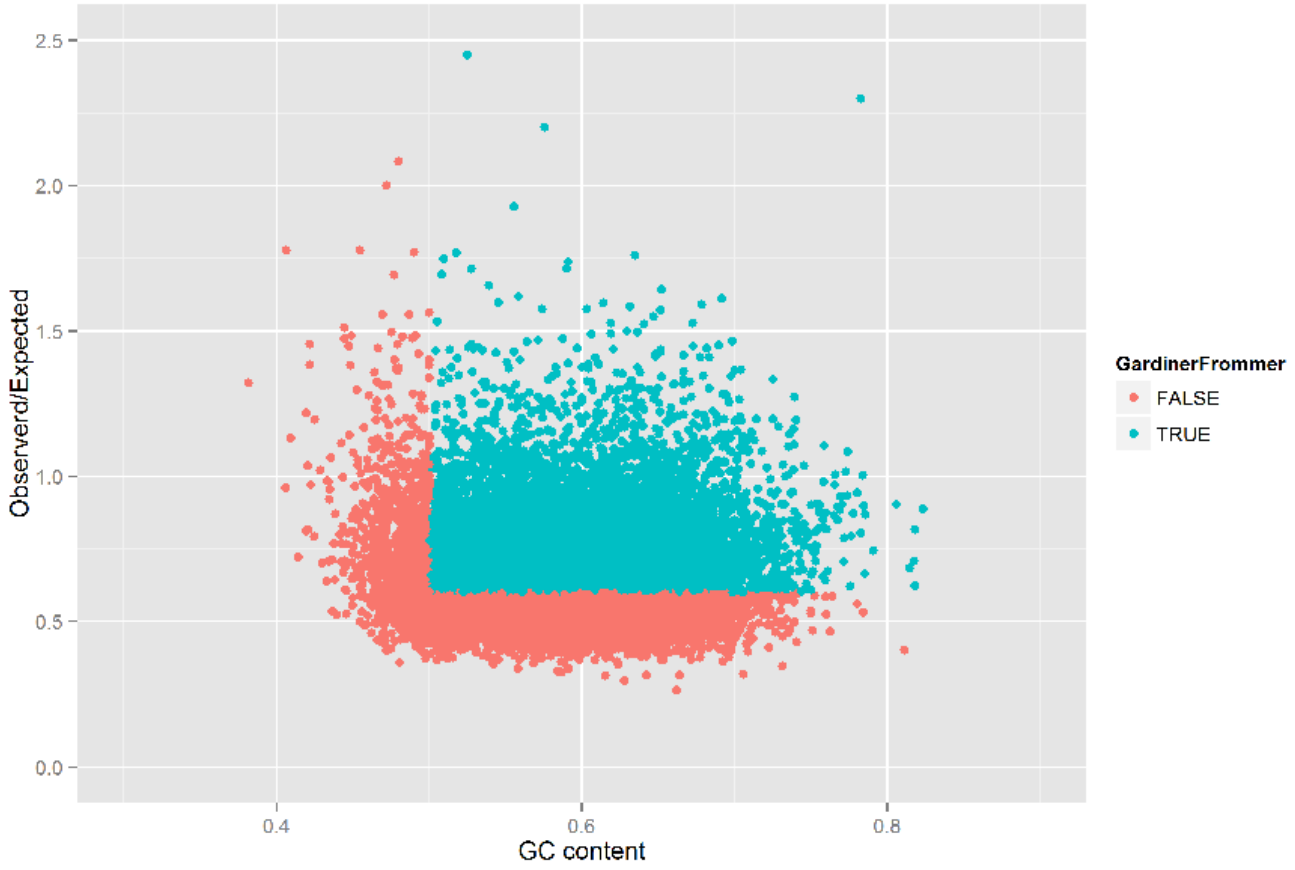


R package with C modules

(cheetah genome + makeCGI)/ Intel Xeon 2620 = 152,5 часа

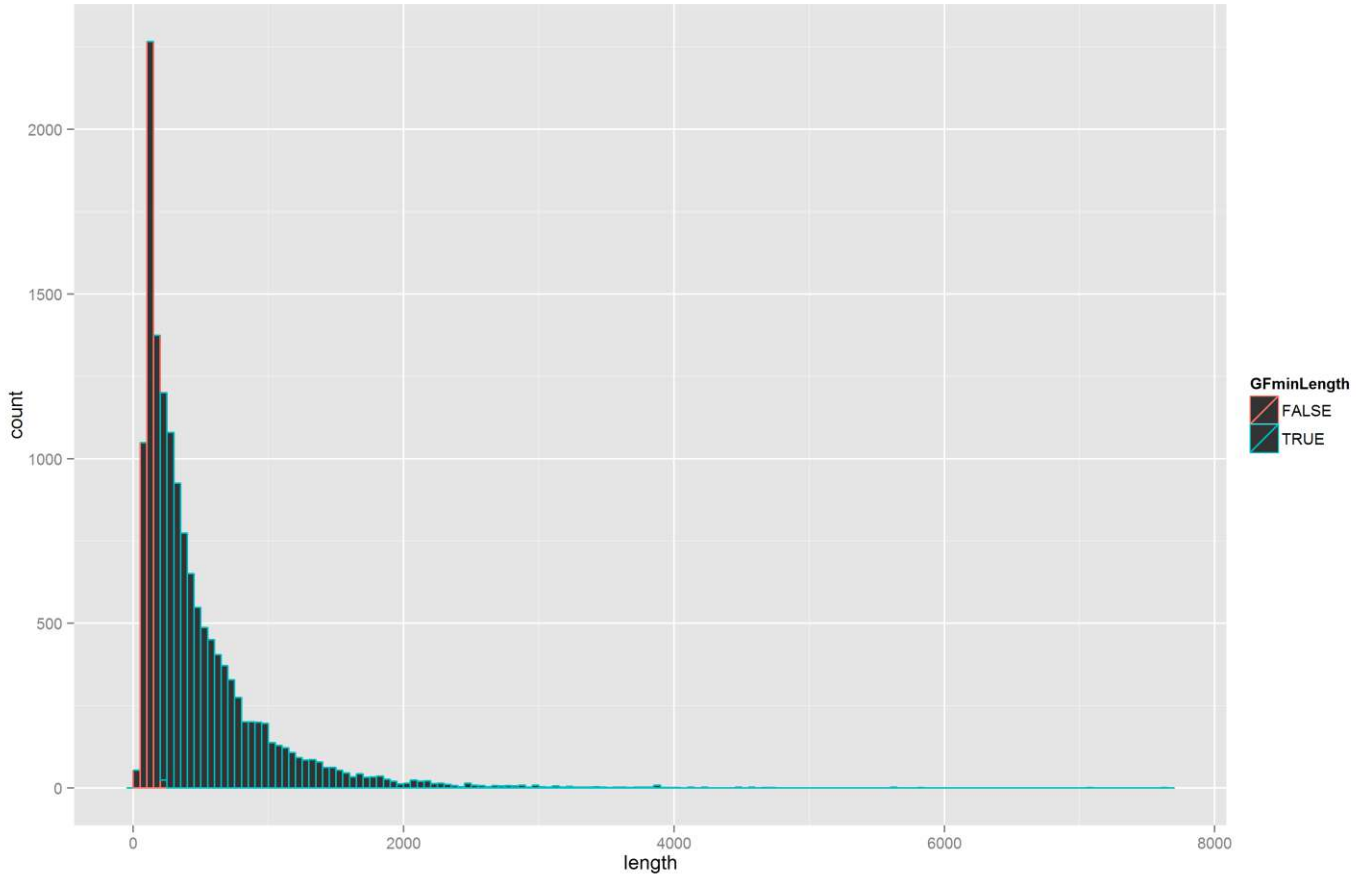
По Wu et al., 2010

CGI в геноме гепарда



Хромосома D1. Красным обозначены те CGI, которые не детектируются с использованием классического определения Gardiner-Garden and Frommer (GC состав > 0.5, observed/expected > 0.6).

CGI в геноме гепарда

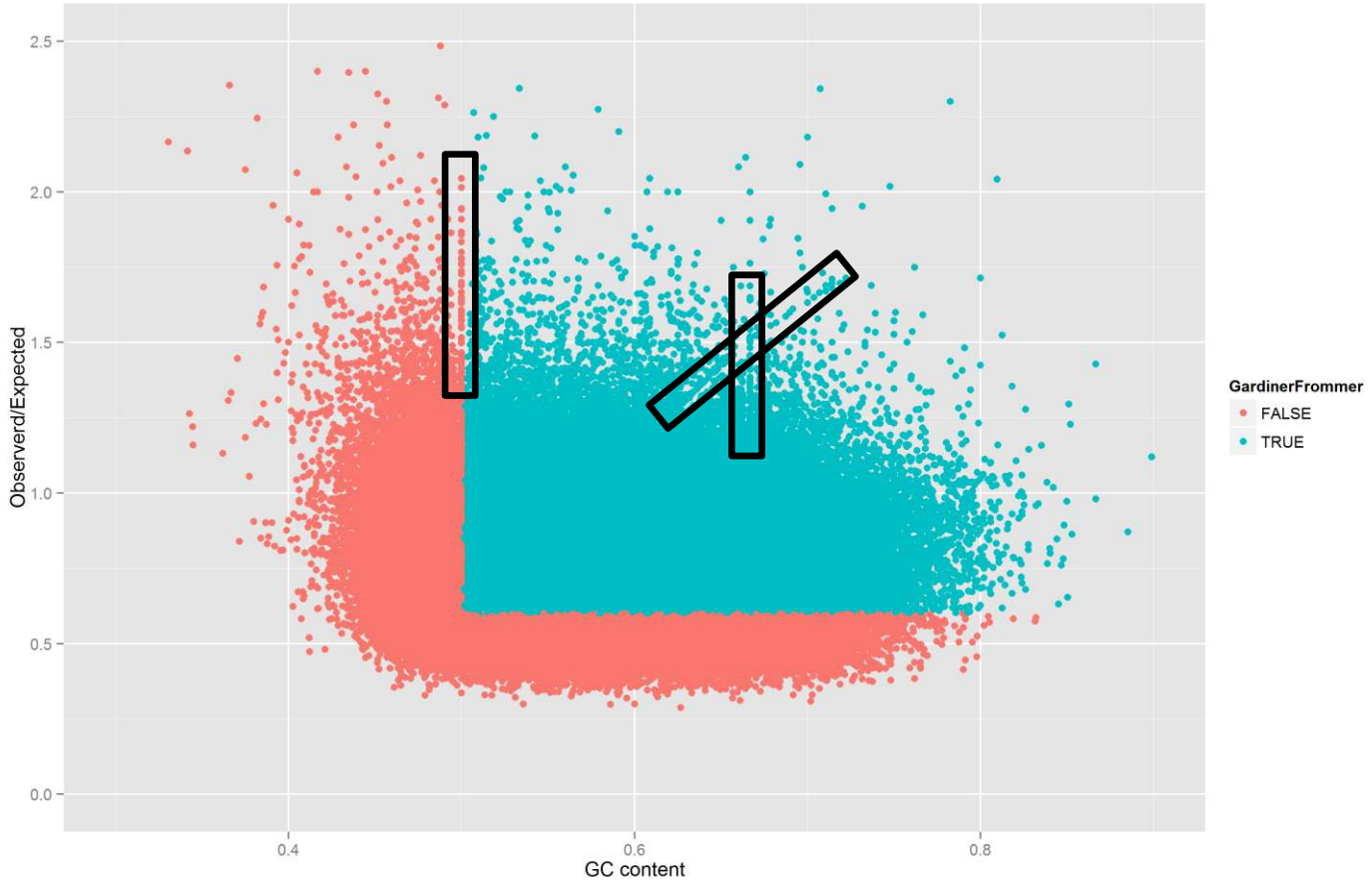


Хромосома D1. Красным обозначены те CGI, которые не детектируются с использованием классического определения Gardiner-Garden and Frommer (длина CGI > 200 п.о.).

CGI в геноме гепарда

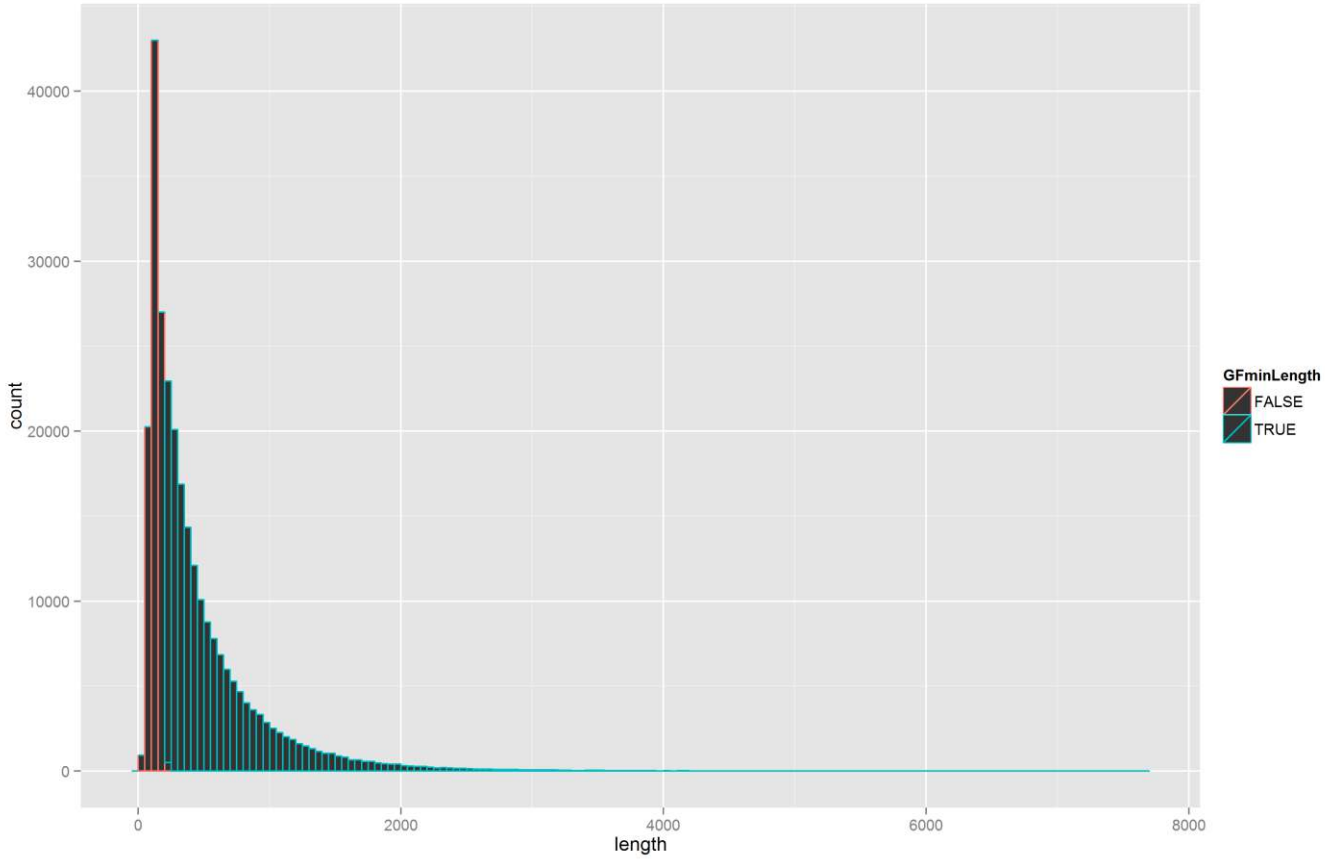
Хромосома D1. Красным обозначены те CGI, которые не детектируются с использованием классического определения Gardiner-Garden and Frommer (GC состав > 0.5, observed/expected > 0.6, длина CGI > 200 п.о.).

CGI в геноме гепарда



Геном гепарда. Красным обозначены те CGI, которые не детектируются с использованием классического определения Gardiner-Garden and Frommer (GC состав > 0.5, observed/expected > 0.6).

CGI в геноме гепарда



Геном гепарда. Красным обозначены те CGI, которые не детектируются с использованием классического определения Gardiner-Garden and Frommer (длина CGI > 200 п.о.).

CGI в геноме гепарда

Геном гепарда. Красным обозначены те CGI, которые не детектируются с использованием классического определения Gardiner-Garden and Frommer (GC состав > 0.5, observed/expected > 0.6, длина CGI > 200 п.о.).

CGI в геноме гепарда

	row.names	chr	start	end	length	CpGcount	GCcontent	pctGC	obsExp
1	1	chrD1	63848	64606	759	39	432	0.5691700	0.6411028
2	2	chrD1	65834	66193	360	19	201	0.5583333	0.6772277
3	3	chrD1	67898	68028	131	13	91	0.6946565	0.8897597
4	4	chrD1	68224	68413	190	15	128	0.6736842	0.7132132
5	5	chrD1	72833	73213	381	20	218	0.5721785	0.6422250
6	6	chrD1	78525	79084	560	19	292	0.5214286	0.5243446
7	7	chrD1	83362	83739	378	17	207	0.5476190	0.6324803
8	8	chrD1	86708	87219	512	36	293	0.5722656	0.8632447
9	9	chrD1	89511	90486	976	32	493	0.5051230	0.5140224
10	10	chrD1	92431	92536	106	7	64	0.6037736	0.7310345
11	11	chrD1	95727	96028	302	22	221	0.7317881	0.6386005
12	12	chrD1	97791	98222	432	40	293	0.6782407	0.8058950
13	13	chrD1	100036	100383	348	15	192	0.5517241	0.5805806
14	14	chrD1	100865	101639	775	21	383	0.4941935	0.4561379
15	15	chrD1	128460	128565	106	8	59	0.5566038	0.9883450
16	16	chrD1	140313	140674	362	22	249	0.6878453	0.5182197
17	17	chrD1	144994	146214	1221	45	652	0.5339885	0.5170807
18	18	chrD1	148083	148295	213	10	117	0.5492958	0.7304527
19	19	chrD1	148596	148846	251	11	132	0.5258964	0.6519481
20	20	chrD1	149820	149951	132	6	65	0.4924242	0.7719298
21	21	chrD1	153454	153575	122	15	84	0.6885246	1.0397727
22	22	chrD1	159520	159897	378	23	203	0.5370370	0.8513514
23	23	chrD1	160223	160328	106	6	53	0.5000000	0.9325513
24	24	chrD1	162280	163137	858	50	514	0.5990676	0.6584905
25	25	chrD1	163604	163692	89	6	52	0.5842697	0.7899408
26	26	chrD1	165629	165680	52	5	36	0.6923077	0.8049536
27	27	chrD1	166000	166349	350	14	201	0.5742857	0.5457786

CGI в геноме гепарда

Хромосома	Длина, Mbp	GC-состав, %	CGI	CGI GGF
A1	130	39.7	18327	6148
A2	93	41.6	18964	6800
A3	78	42.6	19428	7022
B1	116	38.9	13081	4715
…	…	…	…	…
E2	31	44.9	11993	4687
E3	20	47.5	12508	4624
F1	38	42.6	12540	4525
F2	47	40.4	8938	2885
Всего	1260	41.36	267179	96050 (35,9%)

Заключение

1. Использованием makeCGI HMM позволило определить CGI в геноме гепарда.
2. Полученные результаты конвертированы в BED.
3. Сформулированы направления дельнейшего анализа полученных данных.