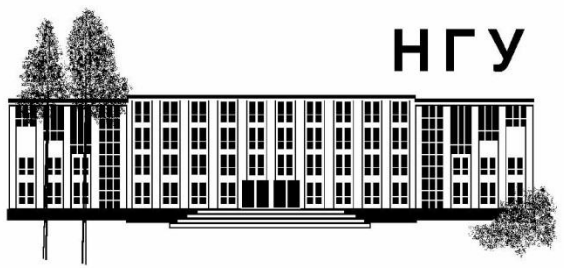
Компьютерный анализ и обработка экспрессионных данных

А. М. Спицина, ММФ НГУ, ИЦиГ СО РАН

Научный руководитель – д.б.н. Ю. Л. Орлов



Цели и задачи

Цели:

- Улучшение имеющейся компьютерной программы для анализа экспрессии генов на микрочипах
- Компьютерное исследование экспрессии генов

Задачи:

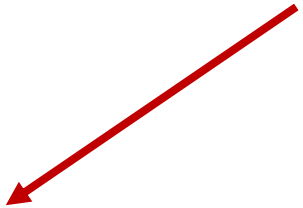
- Обработка данных Affymetrix (по базе данных BioGPS)
- Анализ и визуализация генных сетей

Данные

- Использовались микрочипы affymetrix
- Данные взяты из баз данных BioGPS и GEO NCBI
- Организмы:
  - Мышь *Mus Musculus* (MOE430A, 96 тканей и органов)
  - Крыса *Rattus Norvegicus* (RG_U34A, 30 тканей для нескольких видов крыс)
  - Человек *Homo sapiens* (U133A, 82 ткани)

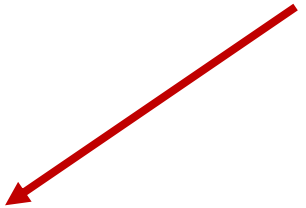
Таблица данных

Данные экспрессии



Ratlas	frontal cortex wistar	amygdala wistar	hippocampus wistar	cerebellum wistar
A01157cds_s_at	5.840344578	4.846337769	4.733836781	4.99912415
A03913cds_s_at	113.7079439	168.6095054	132.6478154	105.569821
A04674cds_s_at	5.611282799	5.177875408	4.950257838	5.237459893
A07543cds_s_at	3.404399277	3.064598414	3.062462665	3.183521622
A09811cds_s_at	67.76096038	118.8854593	210.0054335	81.37296803
A16585cds_s_at	6.784348662	6.068279729	5.750635961	6.249794391

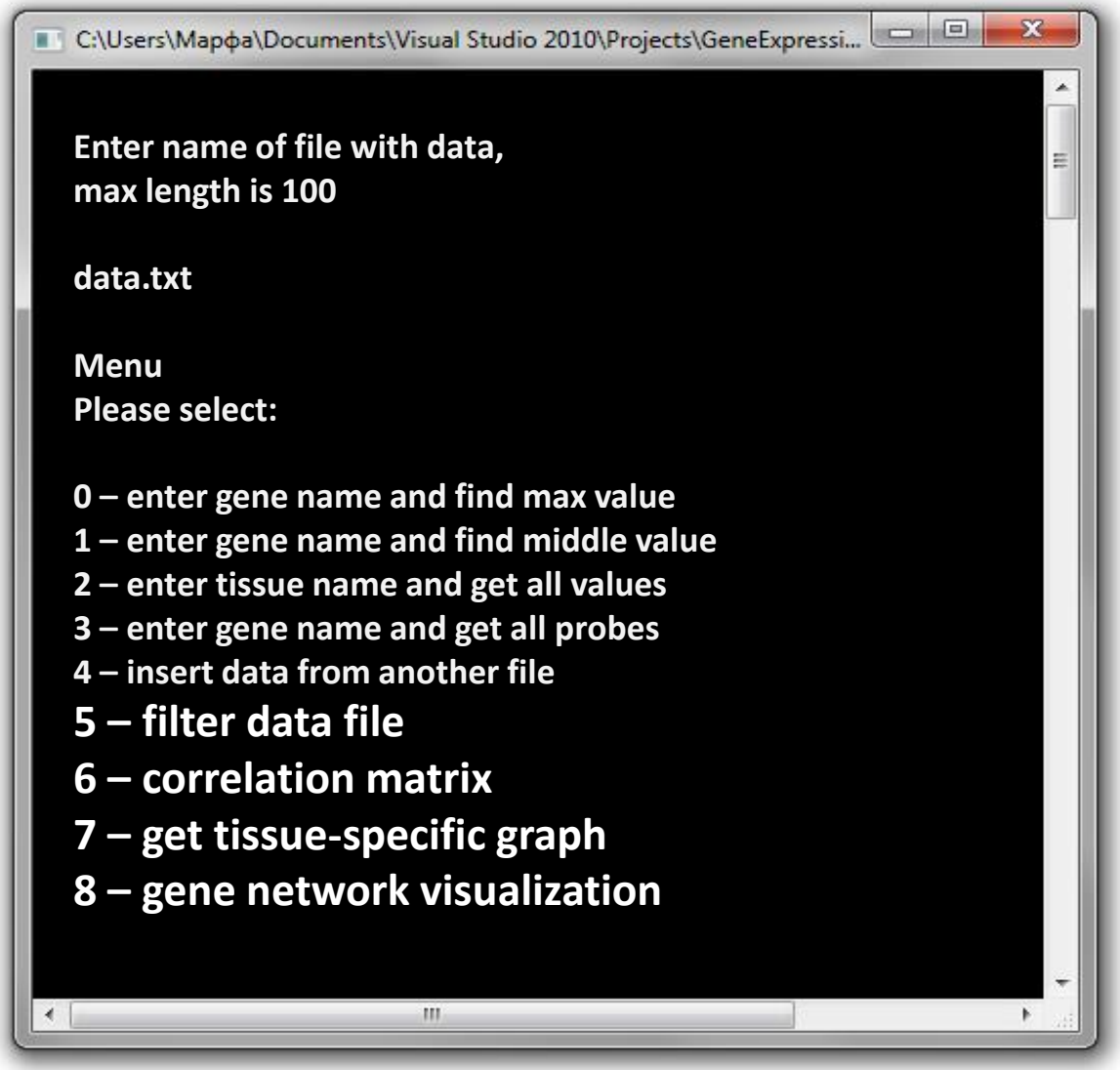
Аннотация проб

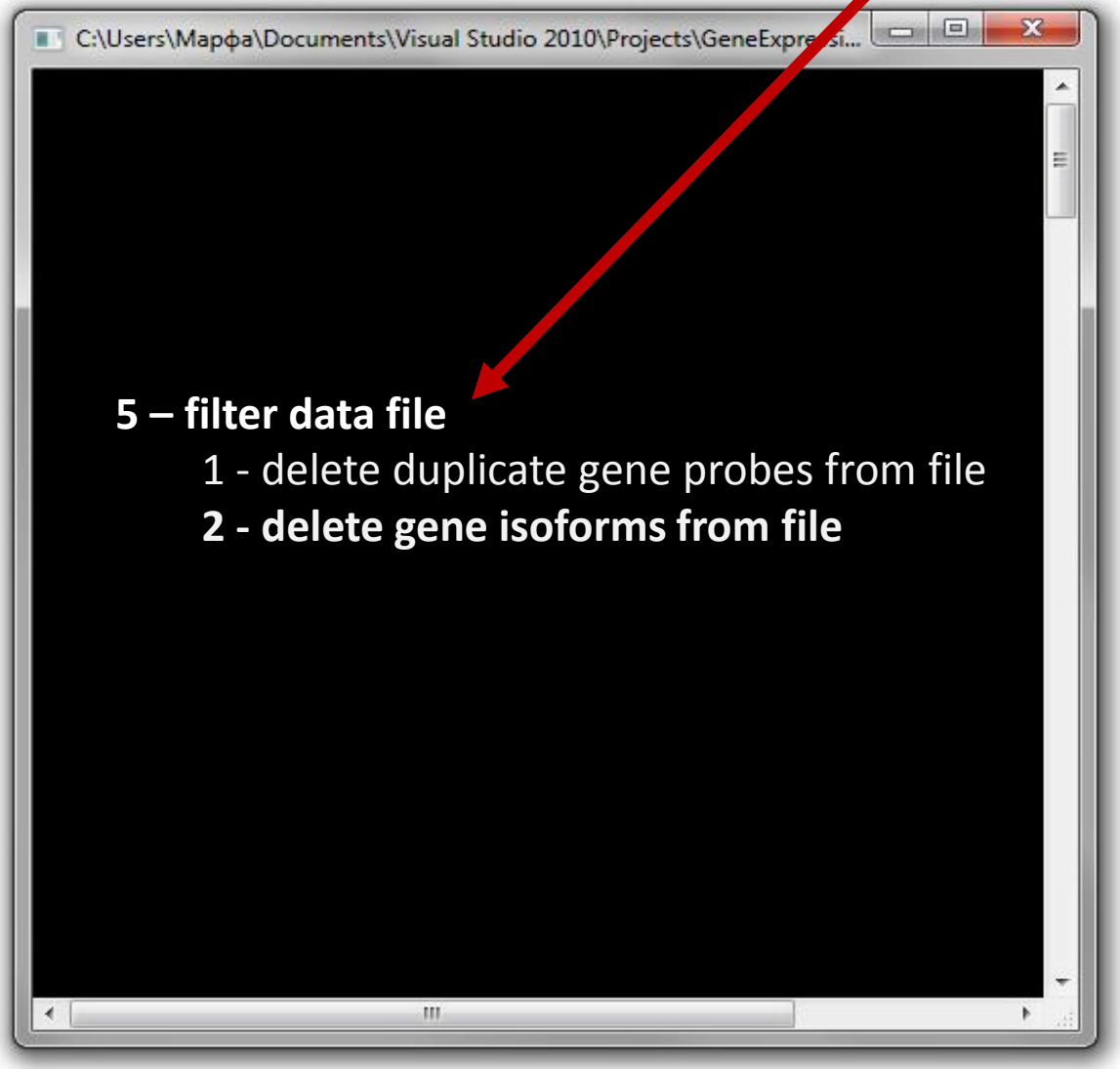


Probe Set ID	Alignments	Gene Title	Gene Symbol	Gene Ontology Biological Process
A01157cds_s_at	chr1:259510964-259525875 (+)	lipase, gastric	Lipf	0006108 // malate metabolic process // inferred from direct assay
A03913cds_s_at	chr9:85312884-85339142 (-)	serpin peptidase inhibitor	Serpine2	0006508 // proteolysis // inferred from electronic annotation
A04674cds_s_at	chr19:35438837-35442718 (+)	mitochondrial brown fat uncoupling protein	Ucp1	0006091 // generation of precursor metabolites and energy // traceable author statement
A07543cds_s_at	chr14:21398247-21521012 (-)	variable coding sequence A1	Vcsa1	0010466 // negative regulation of peptidase activity // inferred from direct
A09811cds_s_at	chr9:79888781-79914887 (+)	insulin-like growth factor binding protein 2	Igfbp2	0001558 // regulation of cell growth // inferred from electronic annotation
A16585cds_s_at	chr1:254731775-254734445 (-)	relaxin 1	Rln1	0007188 // adenylate cyclase-modulating G-protein coupled receptor signaling
AB001982_at	chr2:132784868-132788245 (+)	growth hormone secretagogue receptor	Ghsr	0007165 // signal transduction // inferred from electronic annotation

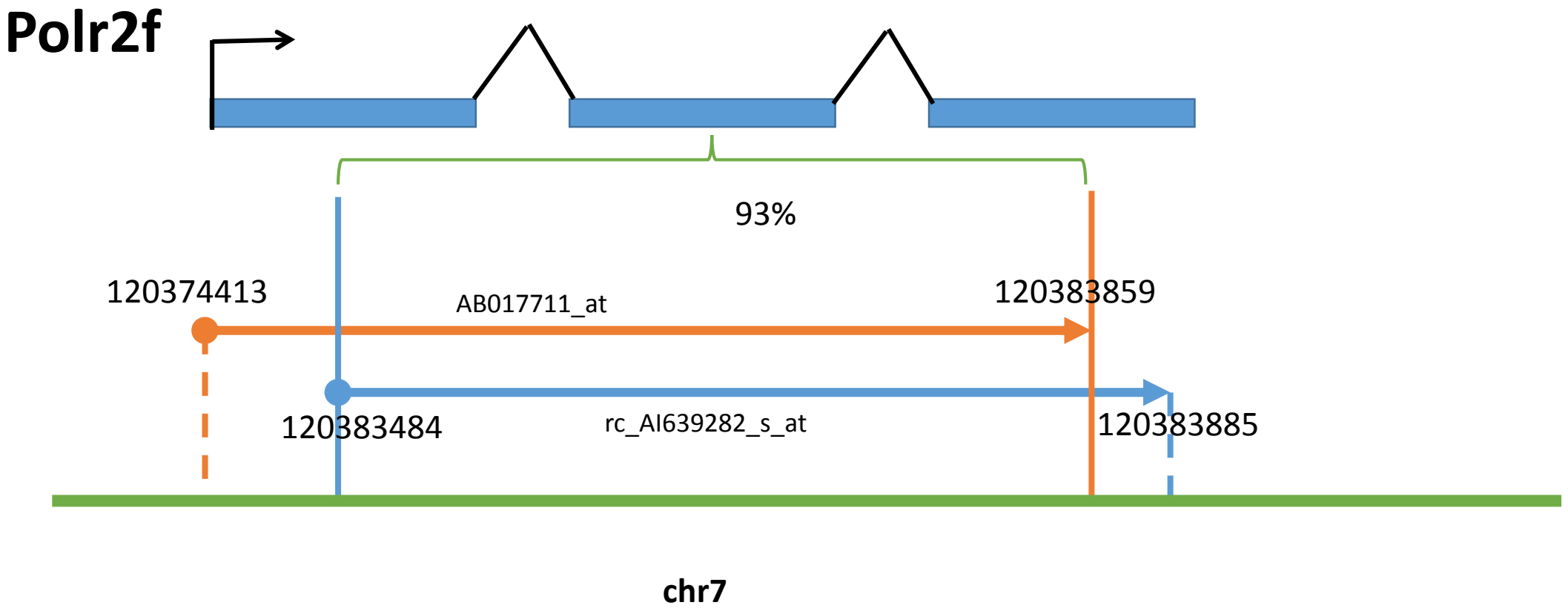
Разработанная программа на языке C++, скрипт на JavaScript

Фильтрация базы данных: удаление дублей и изоформ





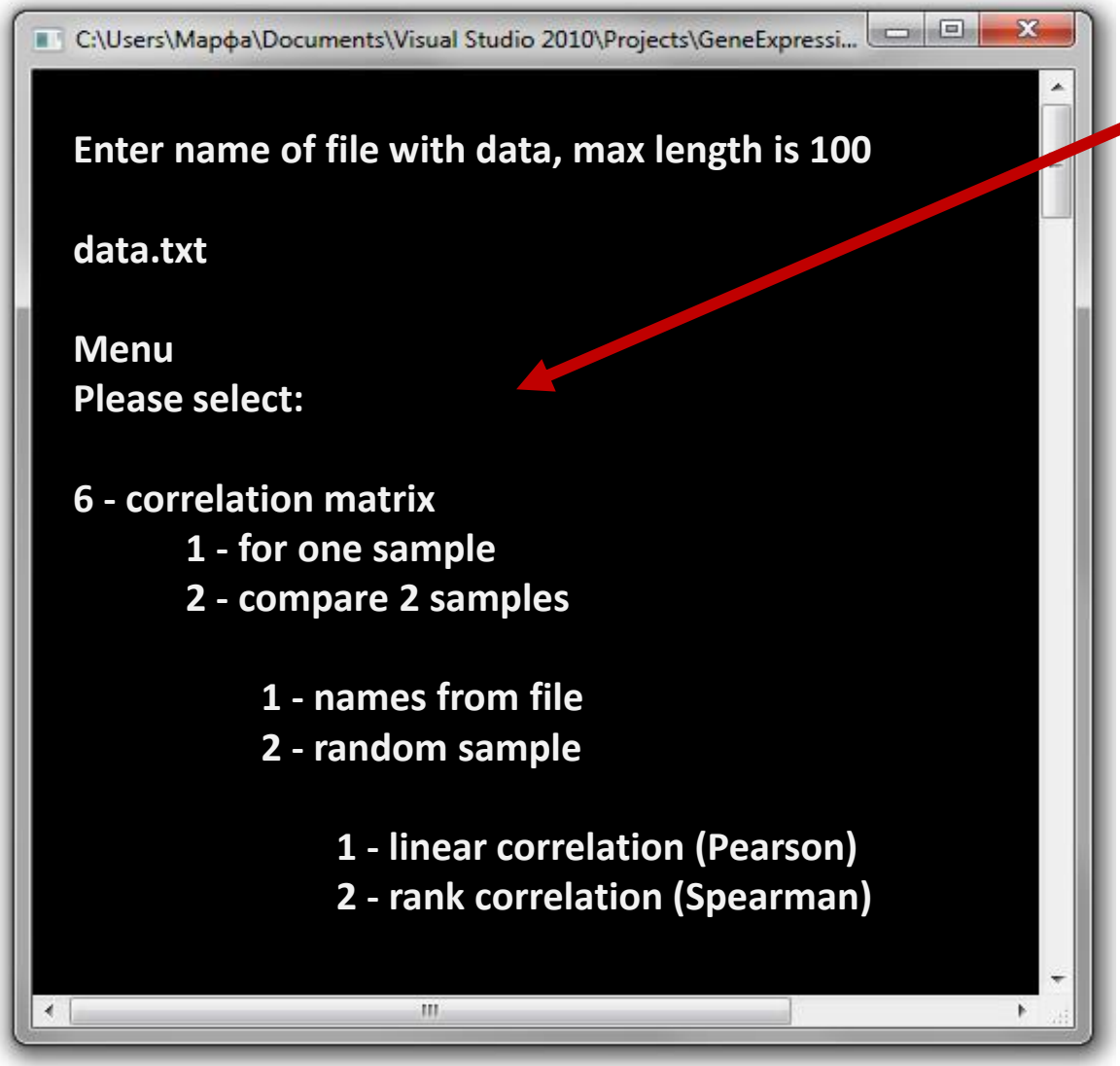
Фильтрация базы данных: удаление изоформ



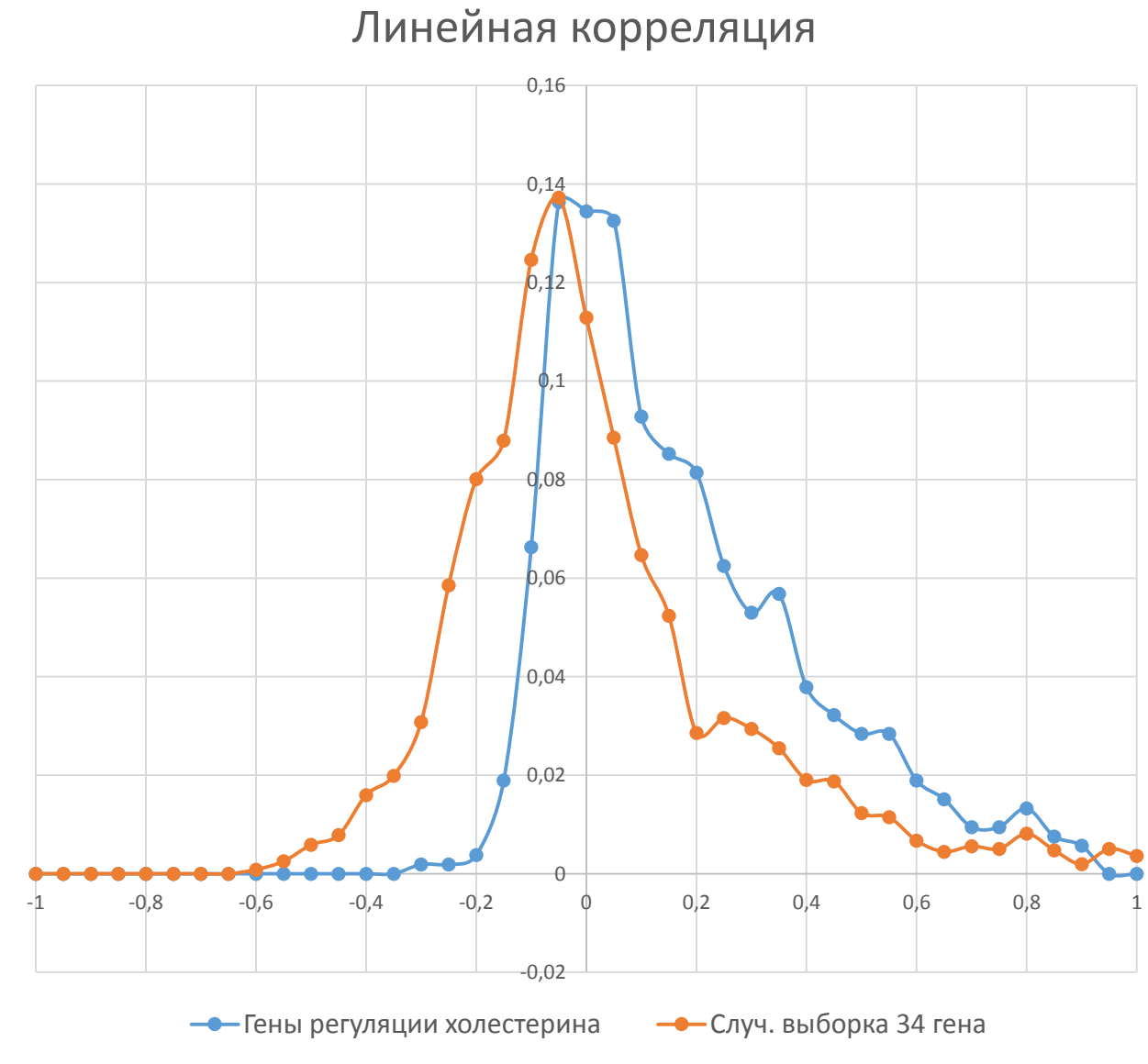
Задача выбора пробы, соответствующей гену, из набора проб микрочипа

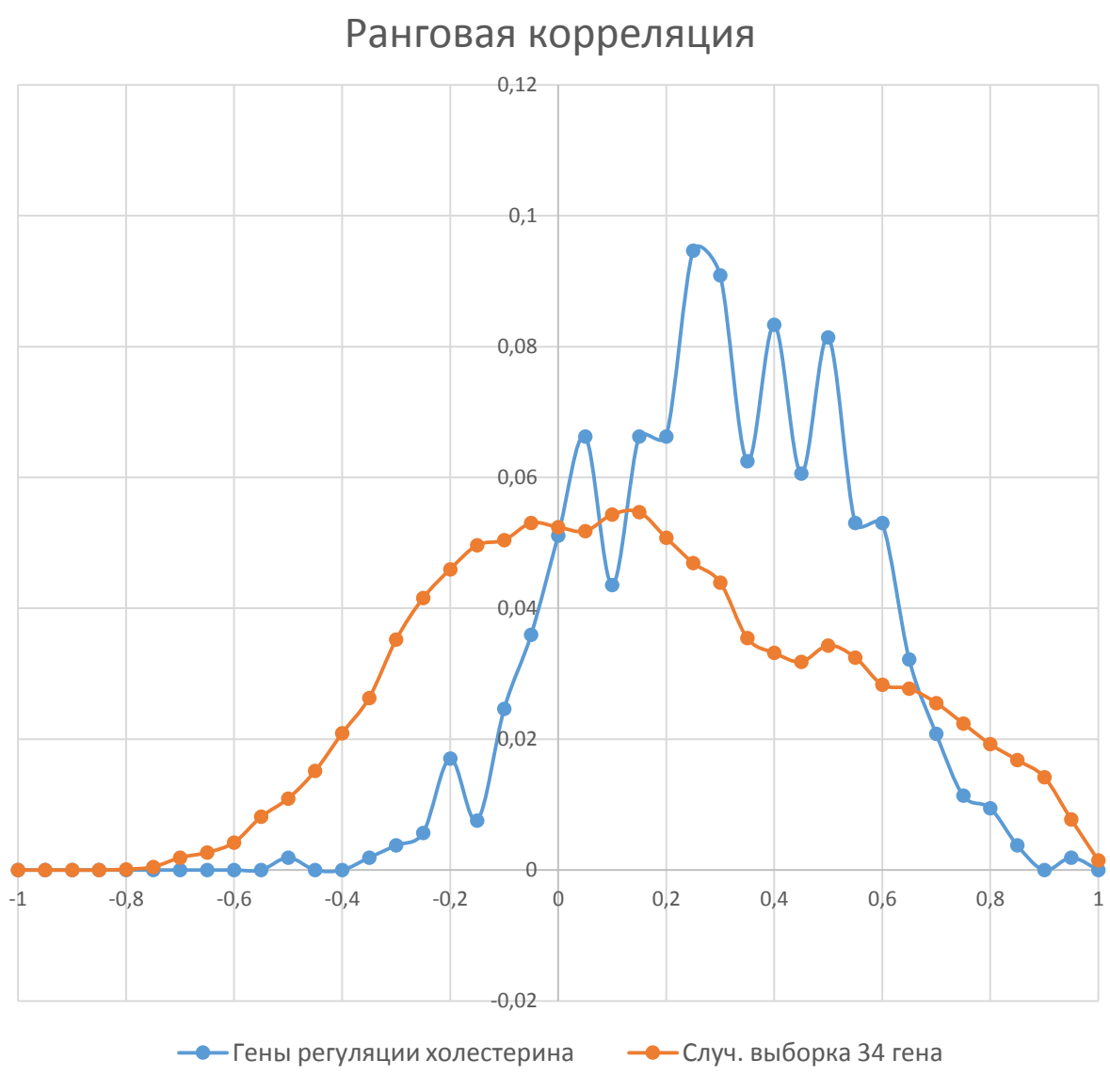
Разработанная программа

Матрица корреляций

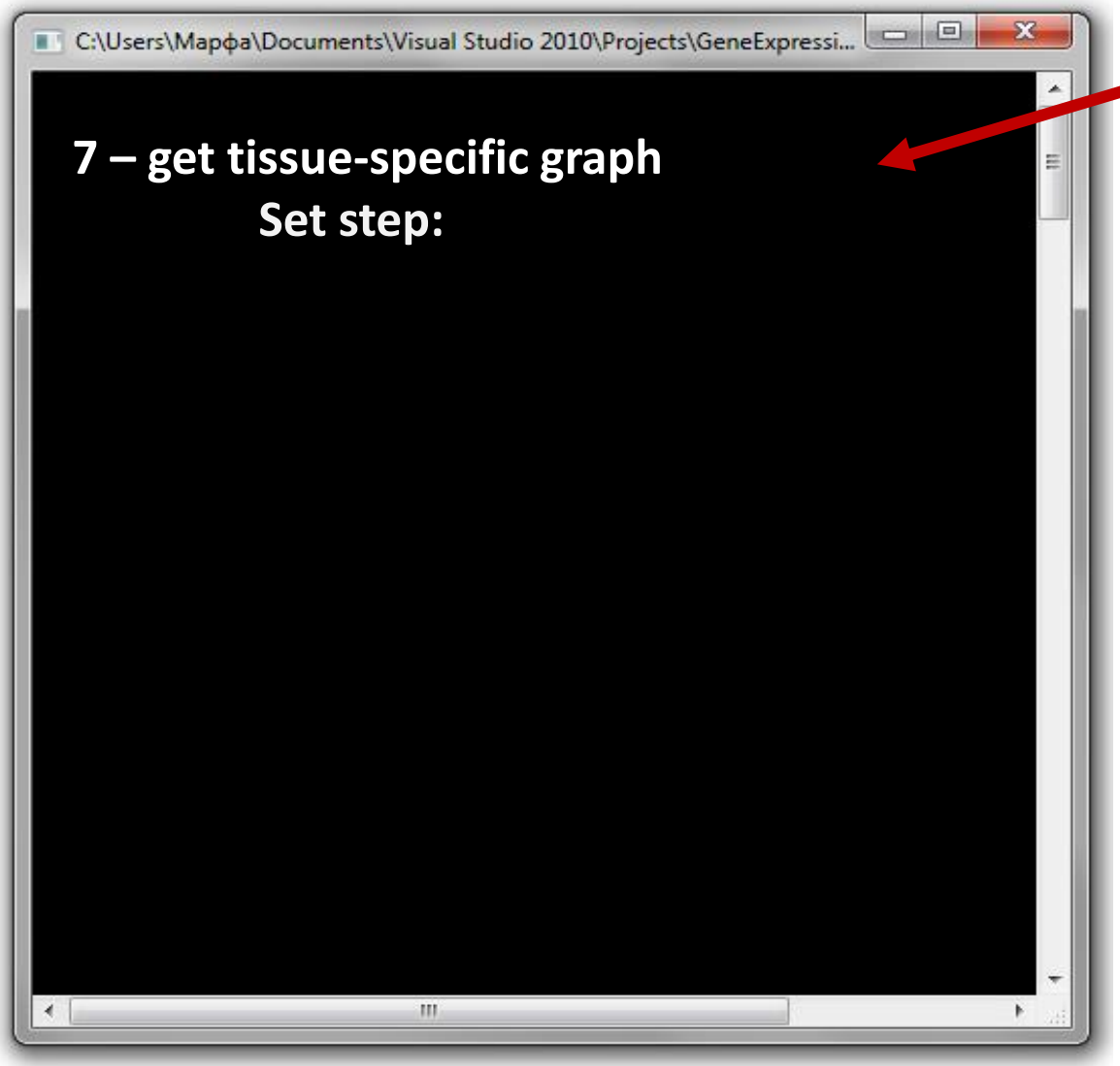


Регуляция холестерина





Разработанная программа

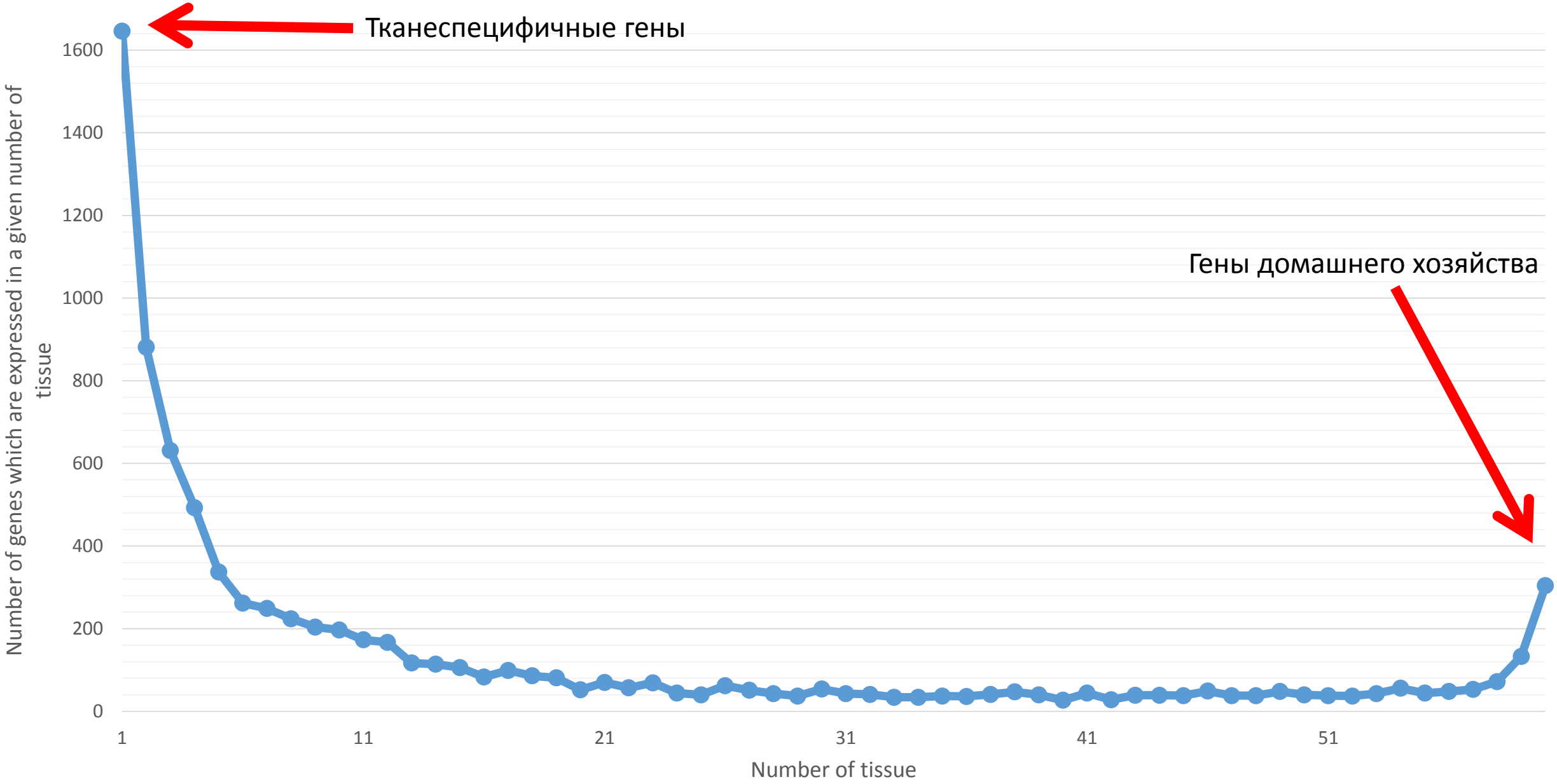


Расчет профилей тканеспецифичности

$$P = Aver + \lambda * (Max - Aver) =$$

$$= \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} G_{IJ}}{N * M} + \lambda * \left(\max_{I,J} G_{IJ} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} G_{IJ}}{N * M} \right)$$

Профиль тканеспецифичности

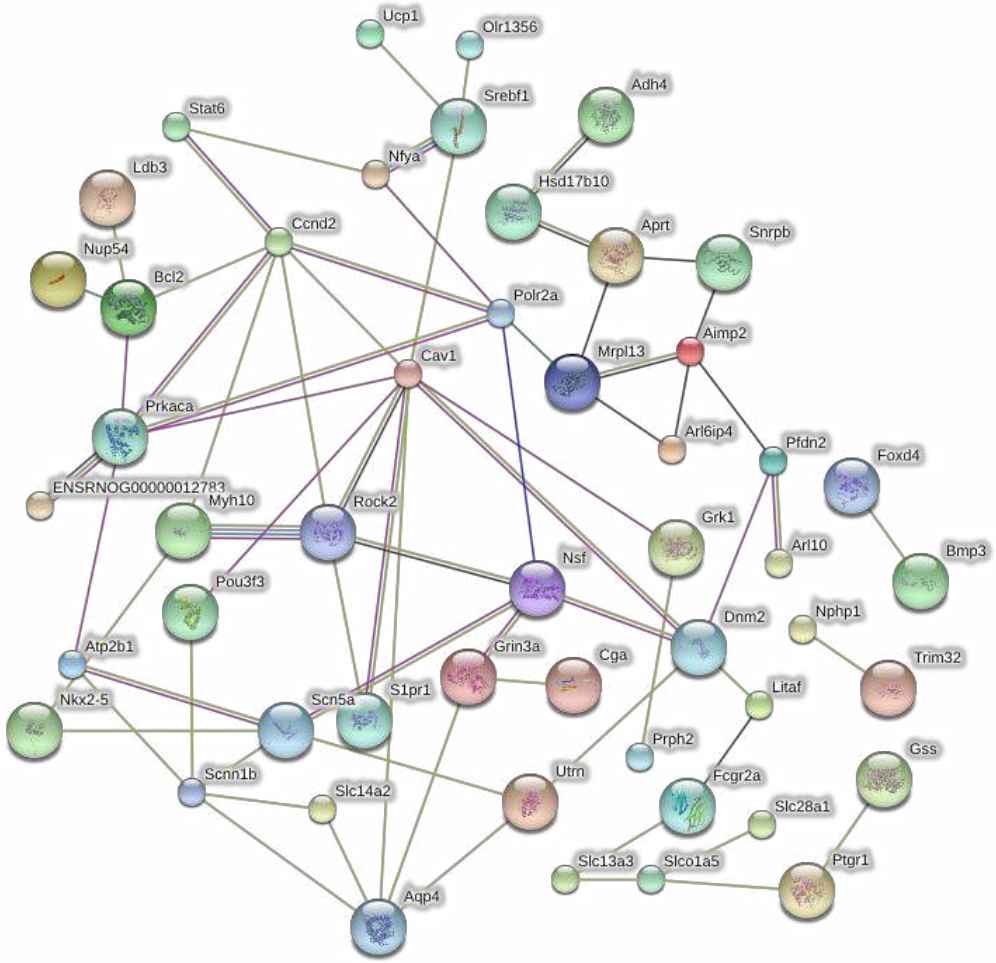


Визуализация генных сетей

	Grin1	Kras	Ldb3	Lsamp	Mybpc1	Pdk4	Syn2	Tnnc2
Grin1	1	0	0	0	0	0	6	0
Kras	0	1	0	0	0	0	0	-8
Ldb3	0	0	1	0	7	-6	0	8
Lsamp	0	0	0	1	0	0	0	0
Mybpc1	0	0	7	0	1	-7	0	7
Pdk4	0	0	-6	0	-7	1	0	6
Syn2	6	0	0	0	0	0	1	0
Tnnc2	0	-8	8	0	7	6	0	1

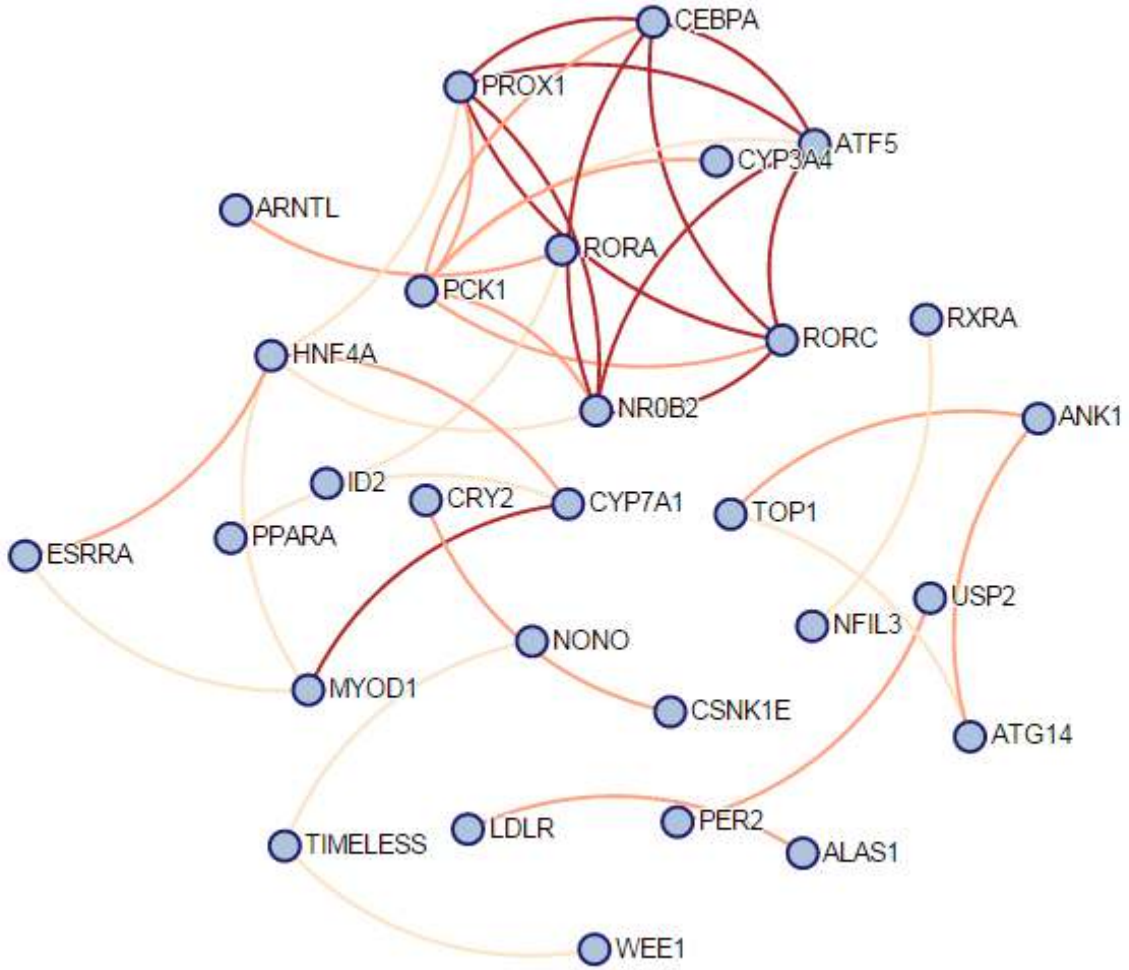
Матрица смежности S(i,j):
S[i][j] = -8 – сильная отрицательная корреляция
S[i][j] = -7 – средняя отрицательная корреляция
S[i][j] = -6 – слабая отрицательная корреляция
S[i][j] = 0 – нет корреляции
S[i][j] = 6 – слабая положительная корреляция
S[i][j] = 7 – средняя положительная корреляция
S[i][j] = 8 – сильная положительная корреляция

Визуализация генных сетей

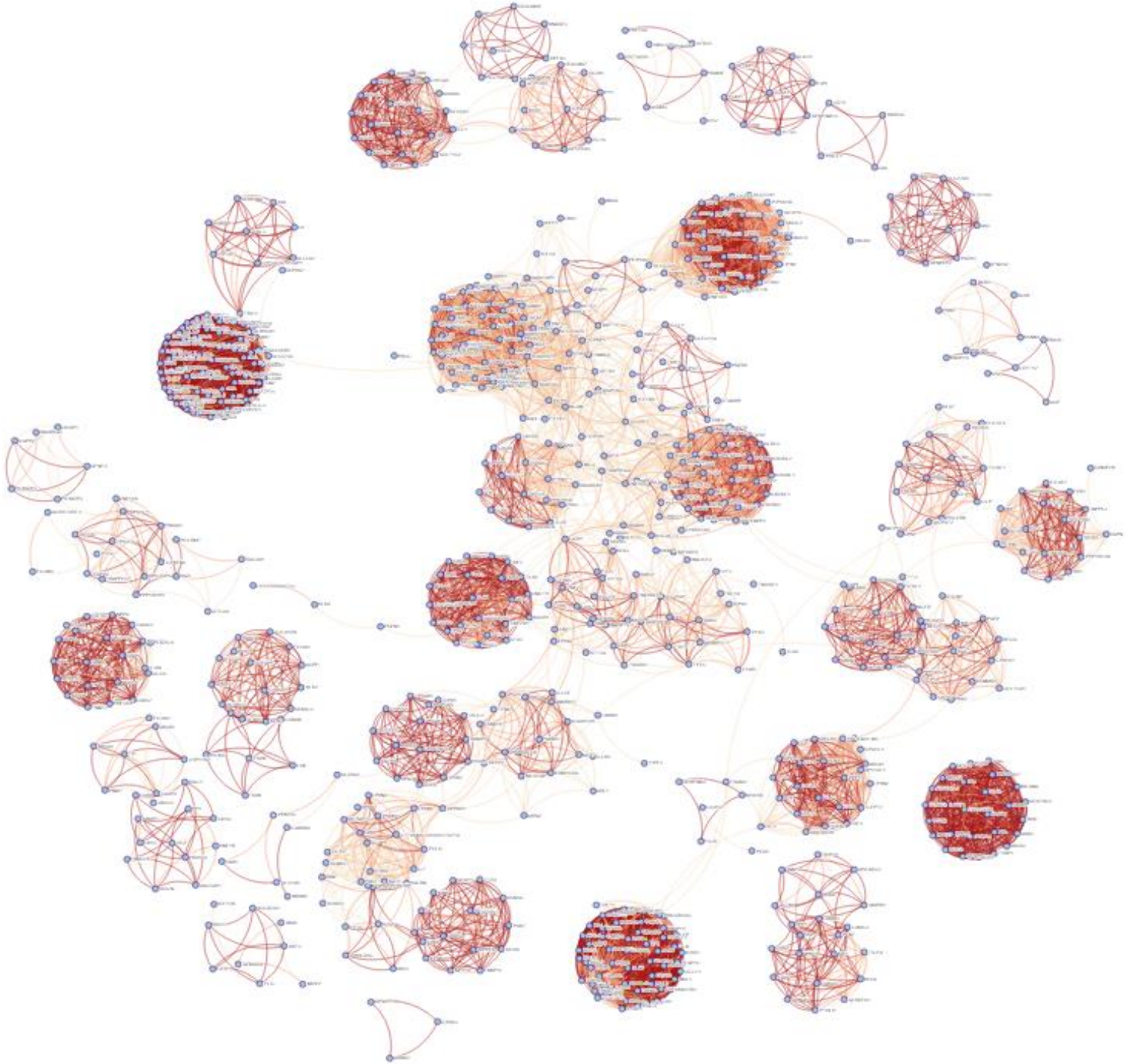


http://string-db.org/

http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/



Собственная визуализация: генная сеть регуляции холестерина



Тканеспецифичные гены

Результаты

Имеющаяся программа была улучшена, и с ее помощью:

- Проведен сравнительный анализ экспрессионных данных (на микрочипах)
- Обработаны списки генов «агрессивности», полученных в экспериментах RNA-seq
- Подготовлены научные публикации
- Проведено исследование генных сетей
  - Рассчитаны профили тканеспецифичности и гистограммы корреляций
  - Визуализация генных сетей по заданному списку генов

Тезисы и публикации

1. **A.M.Spitsina**, Y.L.Orlov, V.M.Efimov, V.N.Babenko. Computer analysis of human gene expression data using BioGPS database of microarray Affymetrix U133. // Abstracts of the Youth Scientific School «Molecular and cellular basis of the early evolution of life". – 2014

1. Орлов Ю.Л., Кулакова Е.В, **Спицина А.М.,** Дергилев А.И., Свичкарев А.В., Афонников Д.А., Чен М., Ли Г., Руан Й., Колчанов Н.А. (2014) Интеграция геномных и транскриптомных данных о хромосомных контактах в геноме человека, полученных по методу ChIA-PET // «Постгеномные методы анализа в биологии, лабораторной и клинической медицине» Постгеном-2014, 29 октября – 1 ноября 2014, Казань. Издательство Казанского (Приволжского) Федерального Университета (ISBN 987-5-00019-293-1) S05-24, С.150.

2. Орлов Ю.Л., Кулакова Е.В., **Спицина А.М.,** Дергилев А.И., Свичкарев А.В., Чен М., Ли Г., Кучин Н.В., Подколодный Н.Л., Черных И.Г., Глинский Б.М. (2014) Суперкомпьютерный анализ геномных и транскриптомных данных, полученных с помощью технологий высокопроизводительного секвенирования ДНК // Национальный Суперкомпьютерный форум-2014 (НСКФ-2014), Переславль-Залесский, ИПС имени А.К. Айламазяна РАН, 25-27 ноября 2014 года.
3. **А. М. Спицина,** Ю. Л. Орлов и др. «Суперкомпьютерный анализ геномных и транскриптомных данных, полученных с помощью технологий высокопроизводительного секвенирования ДНК», Программные системы: теория и приложения, 2015, 6:1(23), с. 157–174.

4. Медведева И.В., Вишневский О.В., Сафронова Н.С., Кожевникова О.С., Суслов В.В., Кулакова Е.В., **Спицина А.М.,** Афонников Д.А., Кочетов А.В., Орлов Ю.Л. Геномная организация и контекстные характеристики генов с повышенной экспрессией в клетках мозга // XVI Всероссийская научно-техническая конференция «Нейроинформатика-2014». Сборник научных трудов. Часть 2. М.: НИЯУ МИФИ. – 2014. С. 32-42.

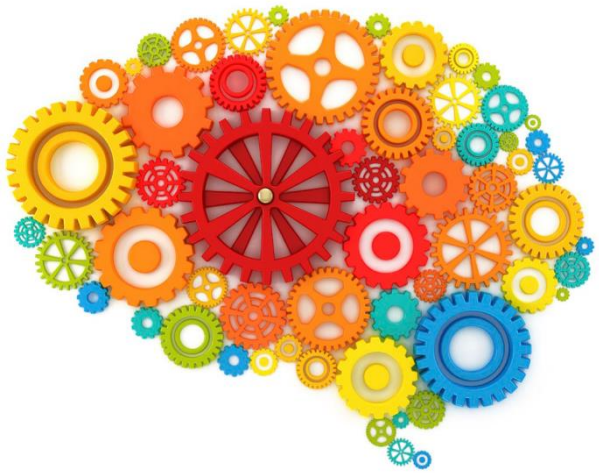
5. **Spitsina A.M.,** Orlov Y.L., Efimov V.M., Babenko V. Computer analysis of human gene expression data using BioGPS database of microarray Affymetrix U133 // // In: Abstracts of the Ninth International Conference on Bioinformatics of Genome Regulation and Structure\ Systems Biology BGRS\SB-2014. – 2014. P. 154. Издательство СО РАН.

6. Vasiliev G.V., Gubanova N.V., **Spitsina A.M.,** Safronova N.S., Orlov Y.L. Computer and experimental analysis of molecular mechanisms of gene expression regulation in brain tumor cells // Abstracts of the Ninth International Conference on Bioinformatics of Genome Regulation and Structure\ Systems Biology BGRS\SB-2014. – 2014. P. 164. Издательство СО РАН.

7. Orlov Y.L., **Spitsina A.M.**, Medvedeva I.V., Bragin A.O., Anikeev A.V., Galyamina A.G., Kozhemyakina R.V., Safronova N.S., Kovalenko I.L., Konoshenko M.I., Moreva T.A., Kudryavtseva N.N., Markel A.L. Computer study of gene expression related to aggressive and tolerant behaviors on laboratory animals // Abstracts of the Ninth International Conference on Bioinformatics of Genome Regulation and Structure\ Systems Biology BGRS\SB-2014. – 2014. P. 188. Издательство СО РАН.

8. **Spitsina A.**, Kulakova E.V., Safronova N., Orlova N.G. Statistical analysis of gene expression data by rank correlation coefficients // Proceedings of Young Scientists School "System biology and Bioinformatics" SBB-2014. – 2014. Издательство СО РАН.

Спасибо за внимание!



Матрица корреляций

- Ранговая корреляция (Spearman):

$$C[I][J] = 1 - \left(\frac{6\sum_{i=0}^{L-1} d_i^2}{L(L^2-1)}\right), d_i = \mathrm{Rank}(I_i) - \mathrm{Rank}(J_i)$$

- Линейная корреляция (Pearson):

$$C[I][J] = \frac{\sum_{i=0}^{L-1}(I_i - \bar{I})(J_i - \bar{J})}{\sqrt{\sum_{i=0}^{L-1}(I_i - \bar{I})^2 \sum_{i=0}^{L-1}(J_i - \bar{J})^2}}, \bar{I} = \frac{\sum_{i=0}^{L-1} I_i}{L}$$

Пример расчета коэффициента для пары генов - PRPF8 RPL35

					0					L-1
I_i	200000_s_at	PRPF8	NM_006445	chr17	17.01	92.4	85.4	66.15	117.6	203.05
Rank(I_i)					1	4	3	2	5	6
J_i	200002_at	RPL35	NM_007209	chr9	2100.95	1491.1	2372.7	690.9	756.2	1263.7
Rank(J_i)					5	4	6	1	2	3

Проблема агрессивности

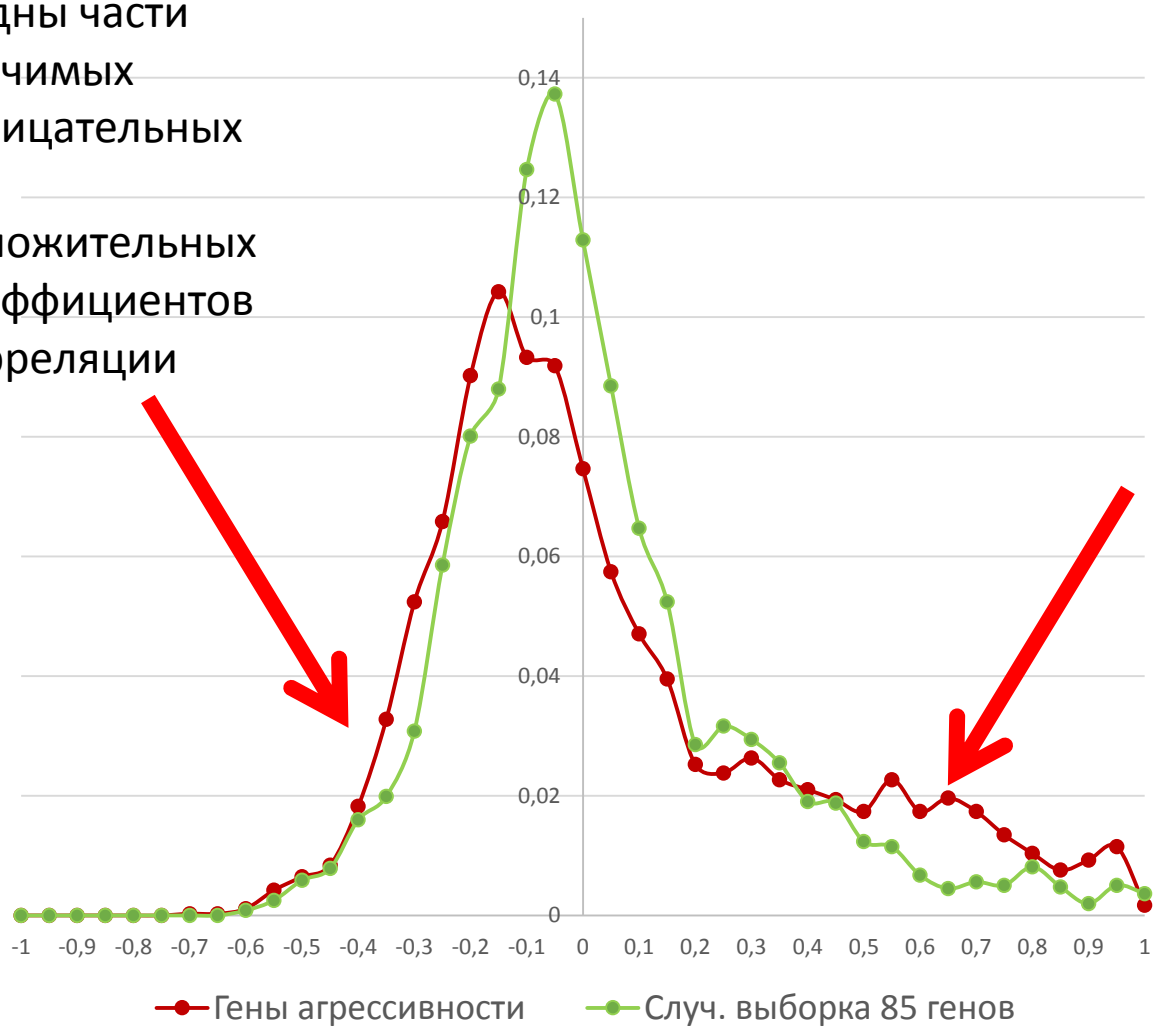
Работа поддержана грантом РНФ 14-14-00269

- С помощью полногеномного секвенирования (RNA-Seq) выявлены дифференциально экспрессирующиеся гены в отделах мозга у агрессивных и контрольных животных (крыс).

- Исследованы следующие отделы мозга: гипоталамус, район покрышки среднего мозга (tegmentum), дорзальные ядра шва (nucleus raphe dorsalis) и центральное серое вещество (substantia grisea centralis) среднего мозга.

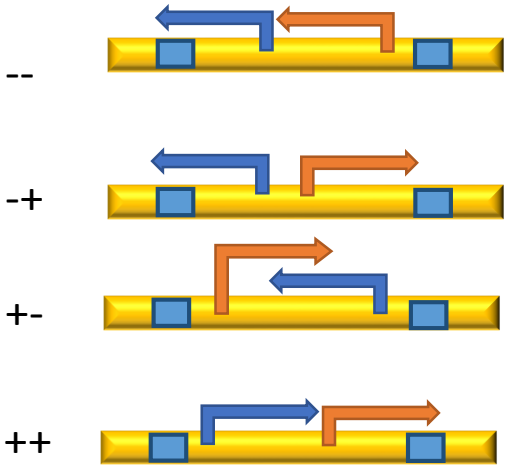
Распределение коэффициентов корреляции всех генов крысы и генов «агрессивности»

Видны части значимых отрицательных и положительных коэффициентов корреляции



Данные UCSC (ChIA-PET) расположение генов относительно сайтов связывания ТФ CTCF

Два гена в одной петле (1540011 пар)

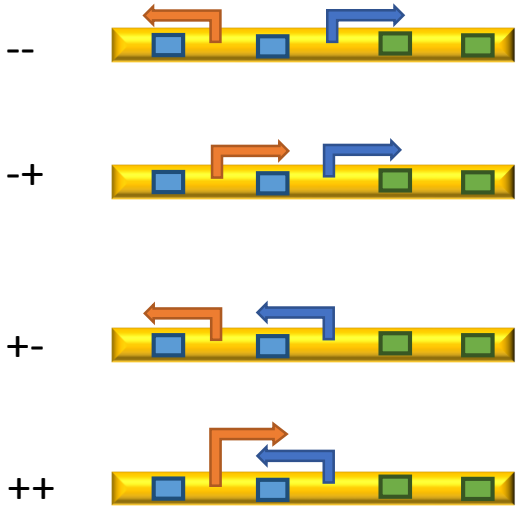


29,4% (453662 пары)

36,7% (566548 пар)

33,7% (519801 пара)

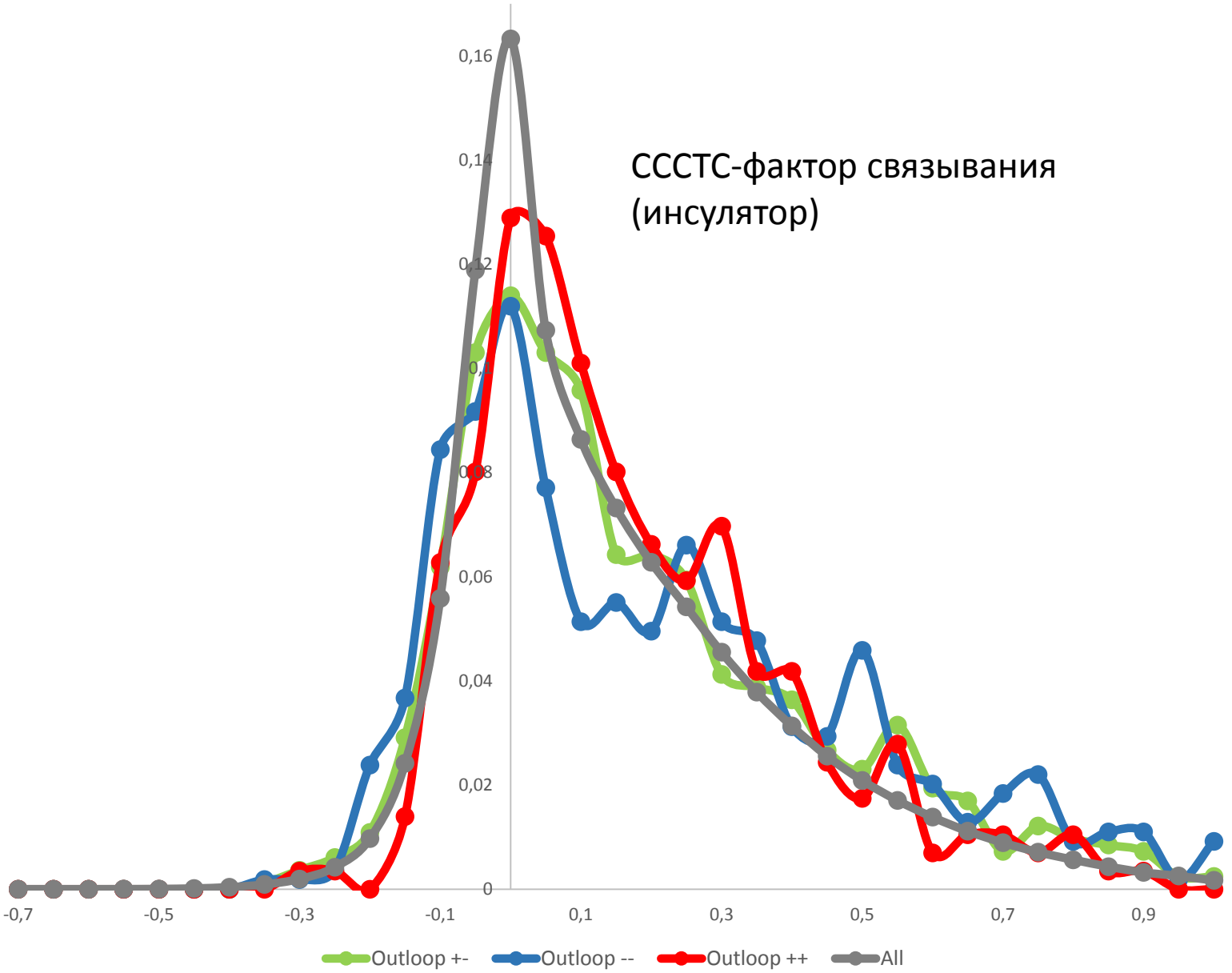
Один ген в петле, другой вне: 2219 пар



38,9% (864 пары)

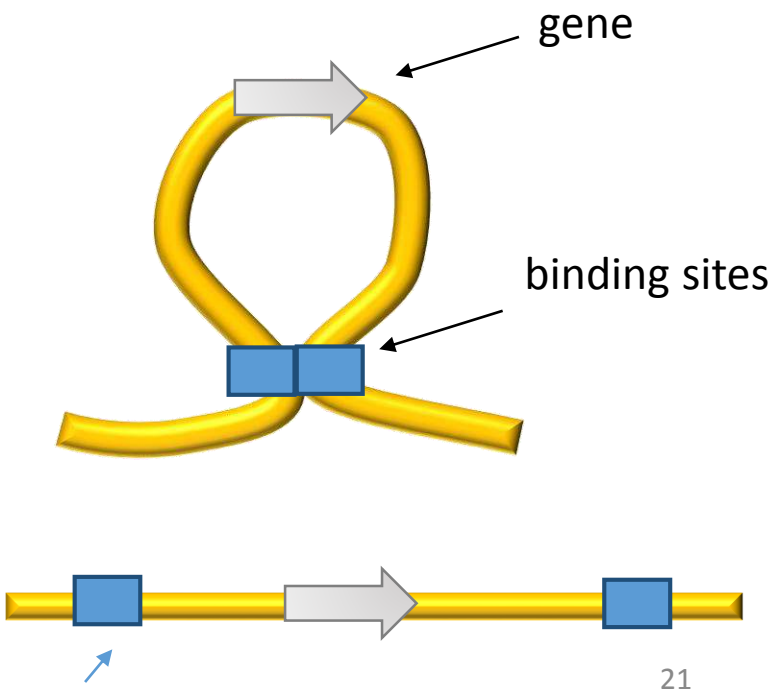
47,9% (1065 пар)

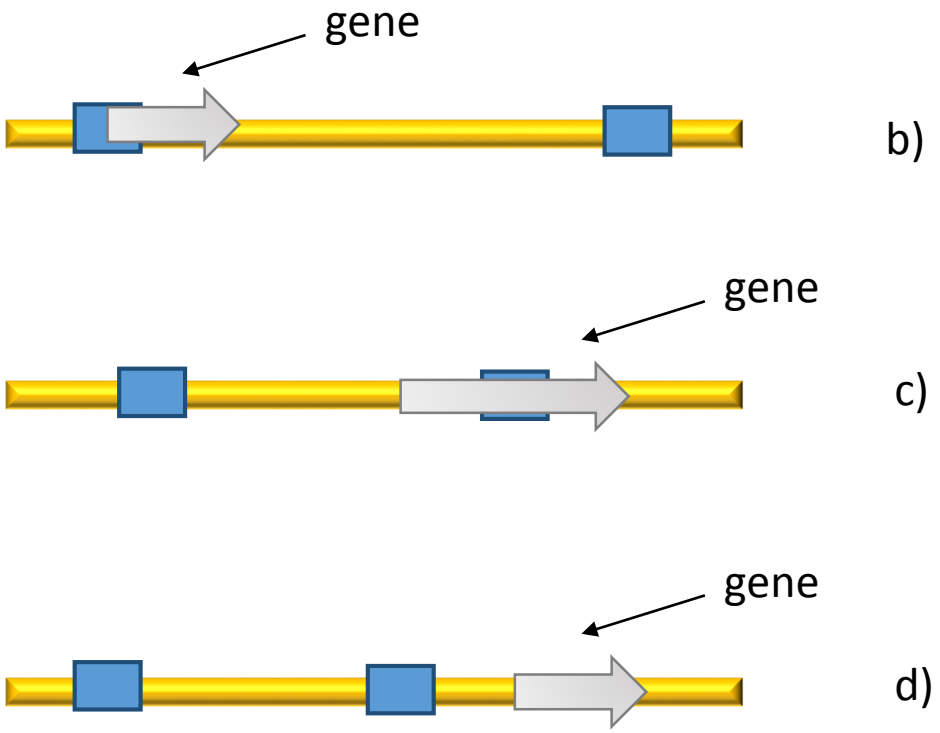
13% (290 пар)



Data CTCF (ChIA-PET) & Location genes relative binding sites

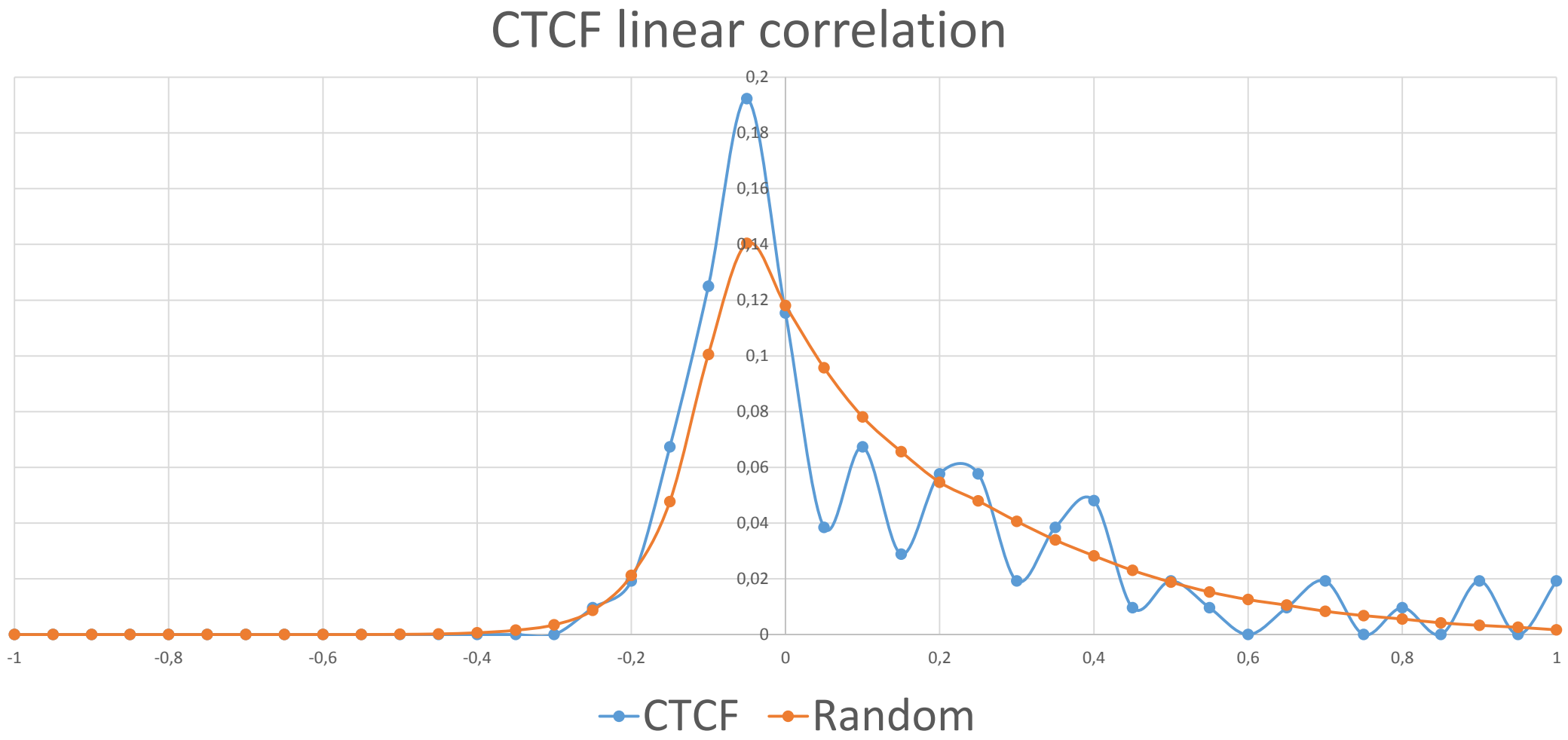
chrom_left	start_left	end_left	chrom_right	start_right	end_right	pet_count_between_anchors	p-value	q-value
chr1	805132	805781	chr1	863968	864858	2	2,69E-10	1,64E-07
chr1	805220	805777	chr1	839523	840175	2	4,92E-09	2,09E-06
chr1	839696	840654	chr1	855782	856703	3	9,77E-13	9,50E-10
chr1	839716	840950	chr1	886489	887404	8	4,01E-33	1,52E-29
chr1	839734	840300	chr1	976162	976774	2	3,93E-09	1,73E-06
chr1	839808	840678	chr1	878347	879101	2	2,16E-09	1,03E-06





(c) Кулакова Е.В.

Data CTCF


CTCF linear correlation
0,2
0,18
0,16
0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
-1
-0,8
-0,6
-0,4
-0,2
0
0,2
0,4
0,6
0,8
1
CTCF
Random