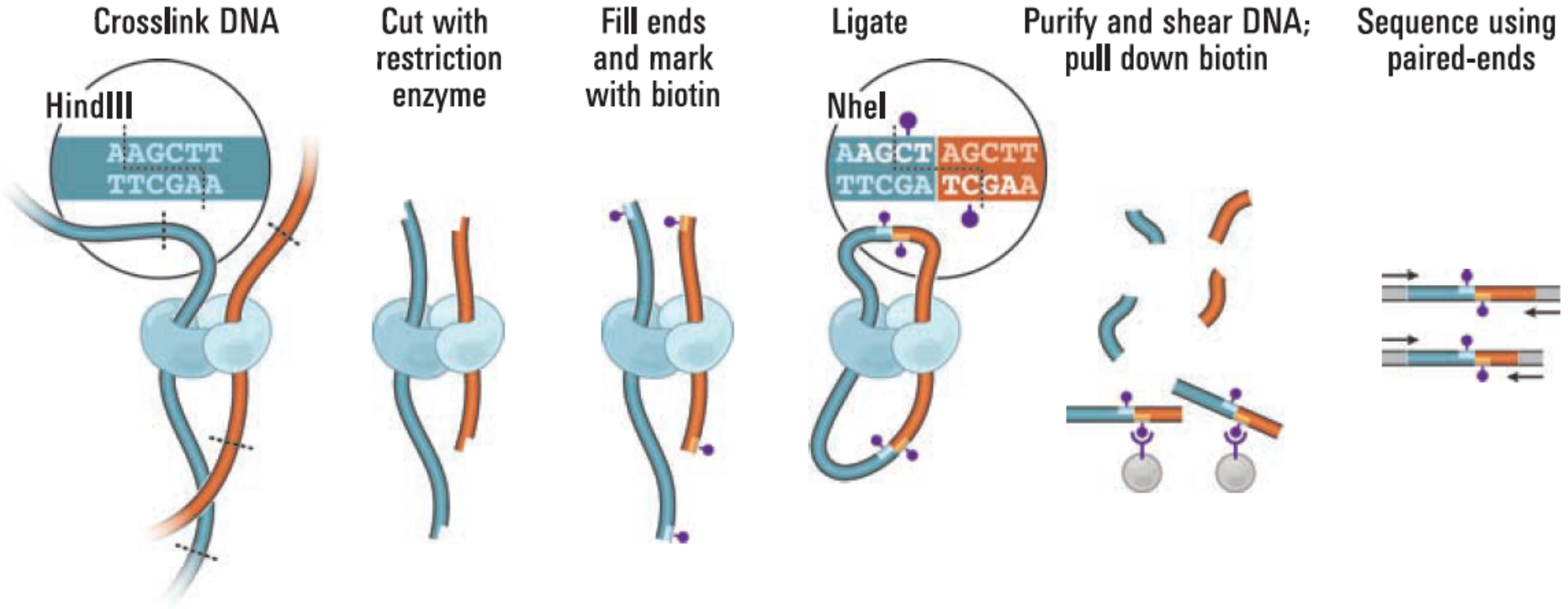
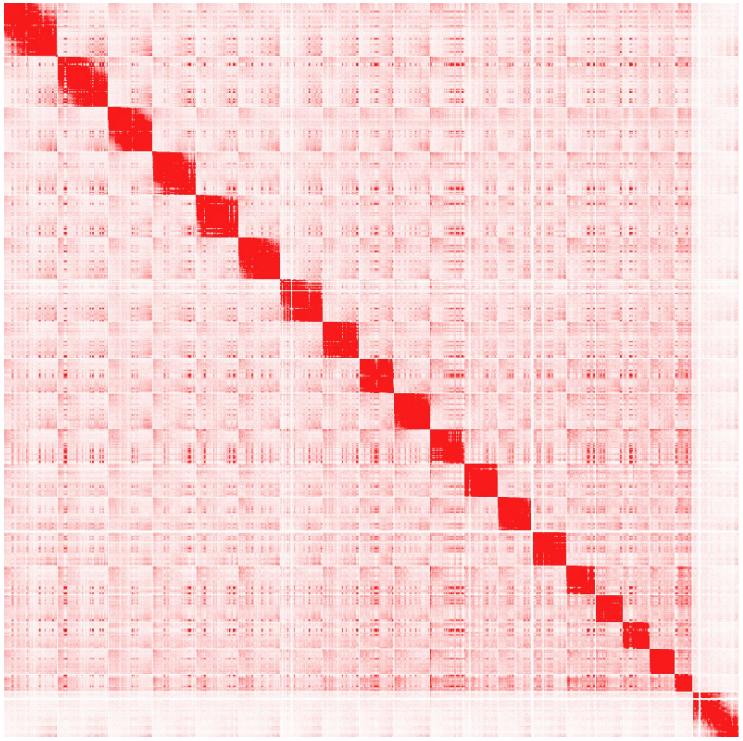
Сравнение алгоритмов поиска ТАДов в данных Hi-C

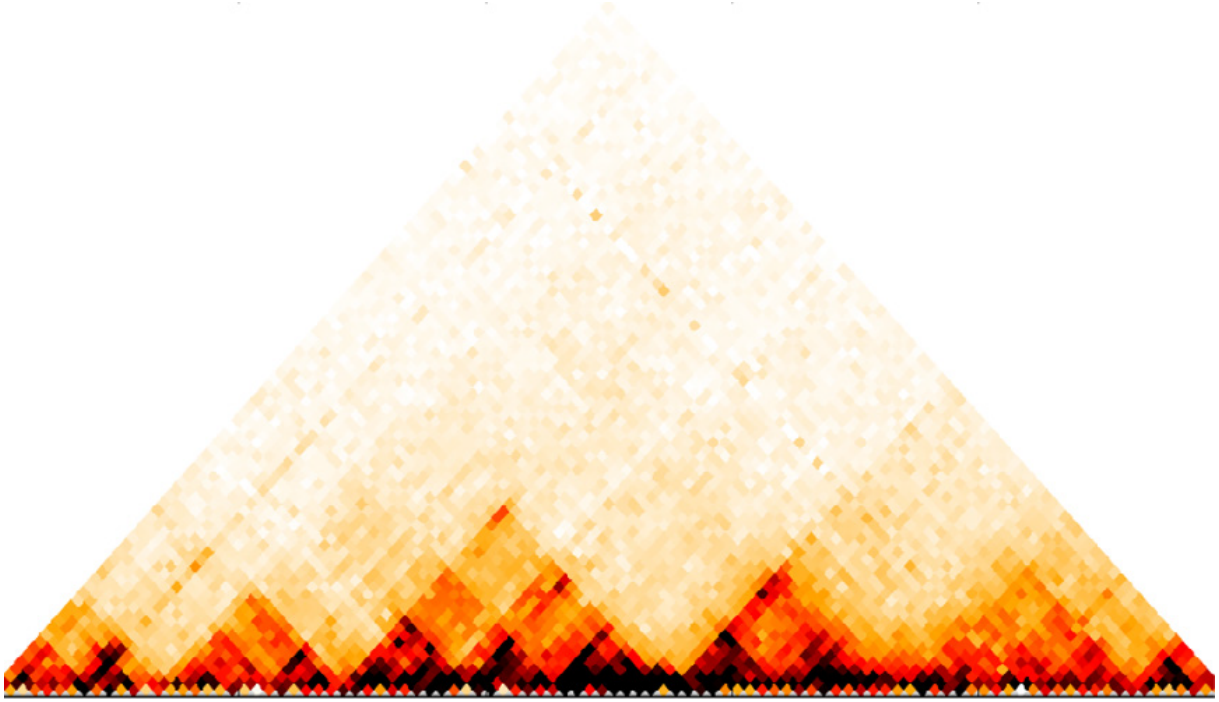
Мыларщиков Дмитрий
Галицына Александра

Hi-C – определение хроматиновых контактов



Топологически ассоциированные домены





Как выглядят данные?

5	2	1	1	.	.	0	0	0	0
2	2	4	3						0
1	4	10	6						0
1	3	6	.						0
.				.					.
.					.				.
0						.			.
0							7	3	4
0							3	20	12
0	0	0	0	.	.	.	4	12	1

F(y)

Chr	Start	End
1	260000	500000
1	500000	750000
…	…	…

Сравнение алгоритмов поиска ТАДов

Comparison of computational methods for Hi-C data analysis

Mattia Forcato[1], Chiara Nicoletti[1], Koustav Pal[2], Carmen Maria Livi[2], Francesco Ferrari[2–4] & Silvio Bicciato[1,4]

Hi-C is a genome-wide sequencing technique used to investigate 3D chromatin conformation inside the nucleus. Computational methods are required to analyze Hi-C data and identify chromatin interactions and topologically associating domains (TADs) from genome-wide contact probability maps. We quantitatively compared the performance of 13 algorithms in their analyses of Hi-C data from six landmark studies and simulations. This comparison revealed differences in the performance of methods for chromatin interaction identification, but more comparable results for TAD detection between algorithms.

The identification of the 3D structure of chromatin inside the nucleus is crucial for deciphering how the spatial organization of DNA affects genome functionality and transcription. Methods

We also addressed elements of tool usability including running time and computational requirements. In general, we see that, depending on the tool, identified structures vary in terms of quantity and characteristics and are more reproducible for TADs than for interactions.

RESULTS

Tools and data preprocessing

We compared 13 methods for the analysis of Hi-C data (**Table 1**; **Supplementary Notes 1** and **2**) using experimental and simulated data. Experimental data were obtained from six landmark studies[2,5,7–9,25], from which we selected nine data sets for a total of 41 samples covering multiple protocol variations, data resolutions, and cell types (**Table 2** and **Supplementary Table 1**). We generated simulated data with a modified version of the model

Алгоритмы поиска ТАДов

Метод в статье	Язык	Метод	Пакет
Armatus	C++	Armatus	Lavaburst
HiCseg	R	Modularity	Lavaburst
DomainCaller	Matlab, Perl	Corner	Lavaburst
InsulationScore	Perl	Variance	Lavaburst
Arrowhead	Java	Potts	Lavaburst
TADtree	Python	Insulation score	TADtool
TADbit	Python	Directionality index	TADtool

План исследования

1. Симулированные данные
    1. Подбор параметров – proof-of-concept
    2. Сравнение двух Armatus
    3. Анализ поведения алгоритмов
2. Реальные данные
    1. Подбор параметров
    2. Анализ результатов при лучших параметрах

Симулированные данные

Как выглядят данные?

5	2	1	1	.	.	0	0	0	0
2	2	4	3						0
1	4	10	6						0
1	3	6	.						0
.				.					.
.					.				.
0						.			.
0							7	3	4
0							3	20	12
0	0	0	0	.	.	.	4	12	1

Chr	Start	End
1	260000	500000
1	500000	750000
…	…	…

Метрики сравнения

$$TPR = \frac{TP}{TP + FN}$$

$$JI = \frac{n\,A \cap B}{n\,A \cup B}$$

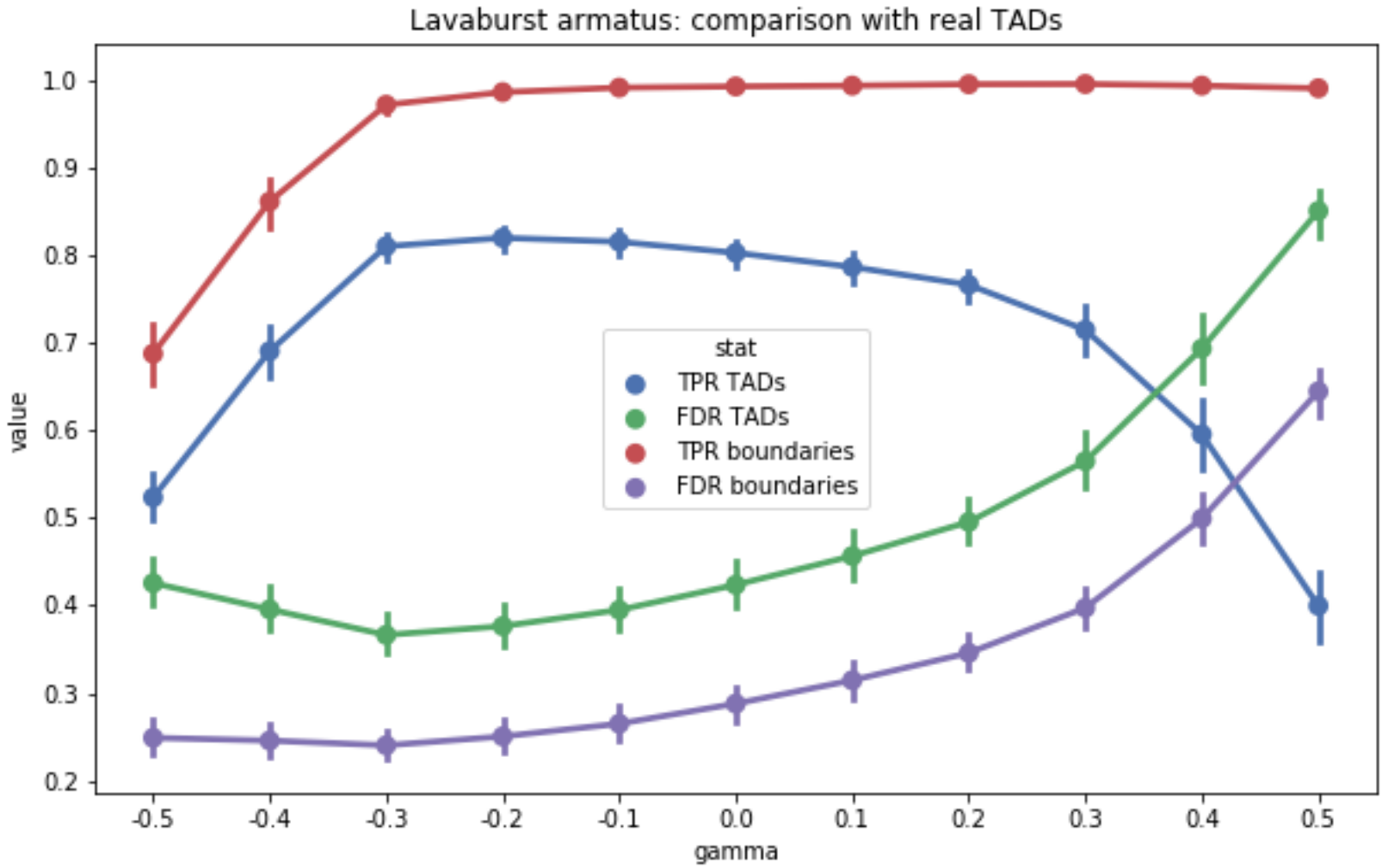
$$FDR = \frac{FP}{TP + FP}$$

$$OC = \frac{n\,A \cup B}{min\left(|A|, |B|\right)}$$

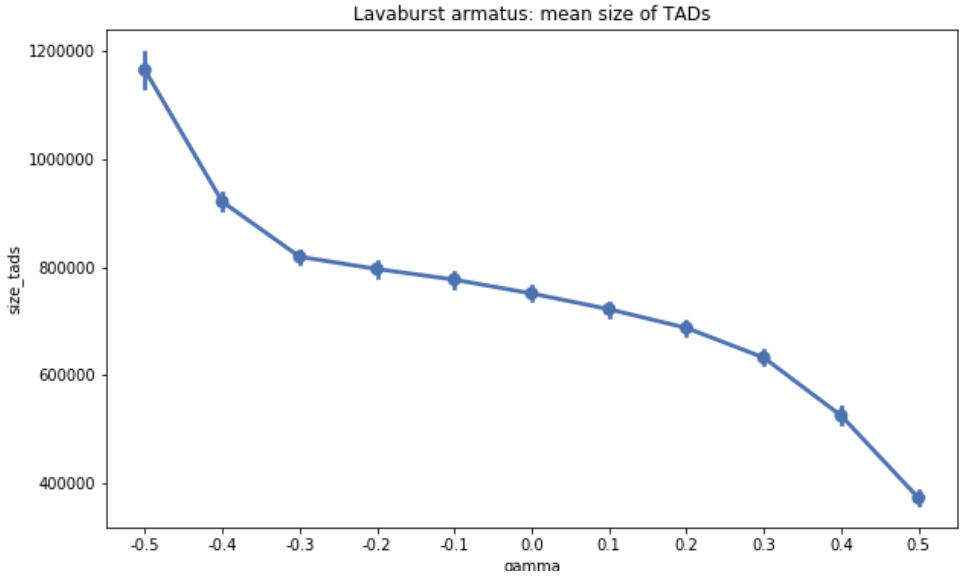
Алгоритмы: свойства

1 параметр	2 параметра	~0 параметров
Armatus	Insulation score	HiCseg
Modularity	Directionality index	
Corner		
Potts		
Variance		

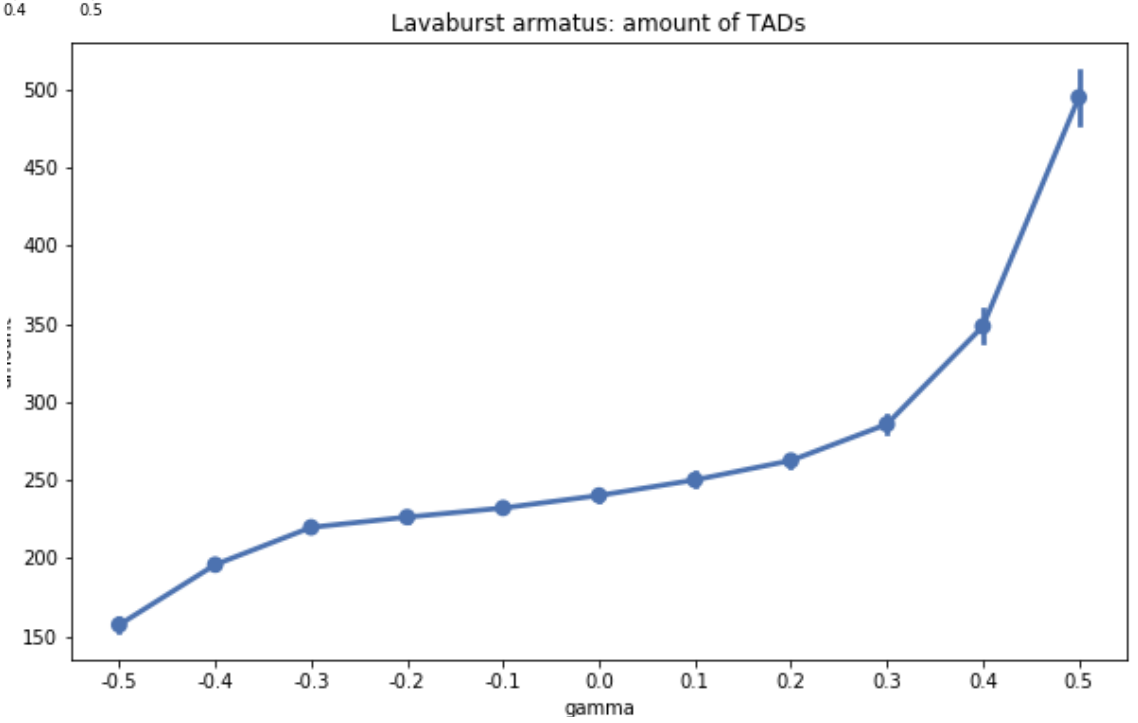
Подбор параметра



Подбор параметра



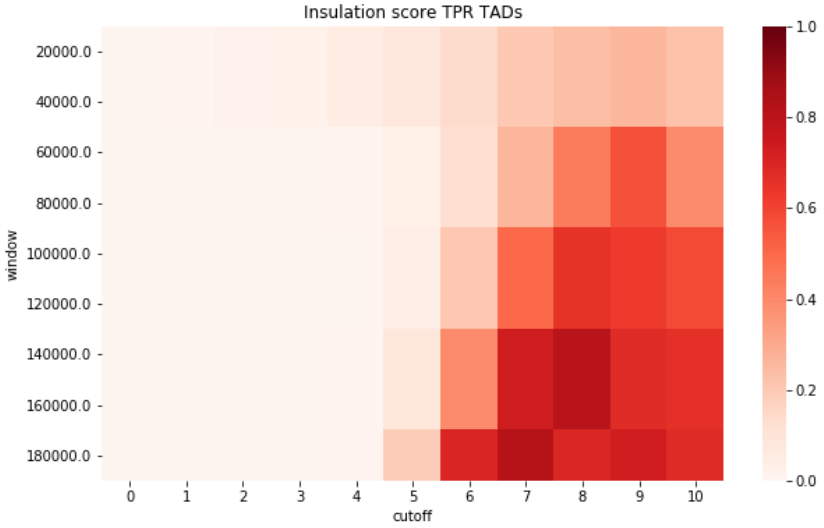
Средний размер ТАДов



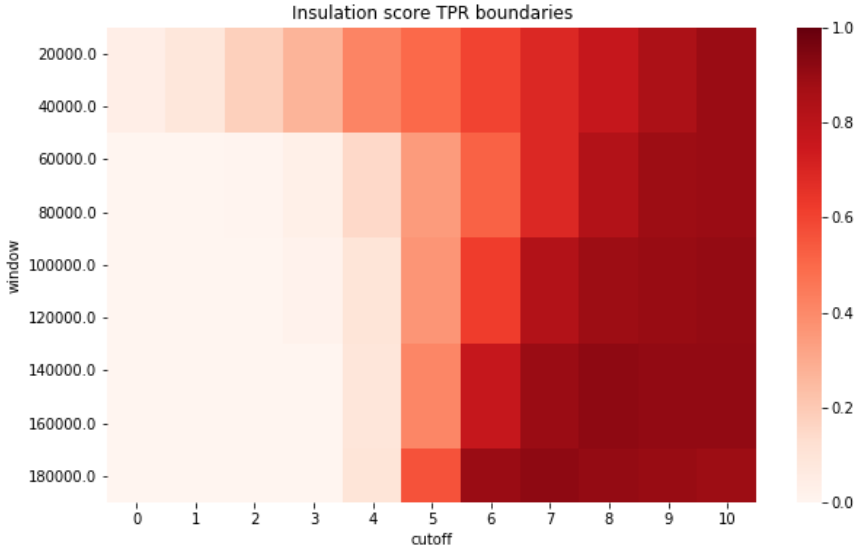
Число ТАДов

Подбор параметров

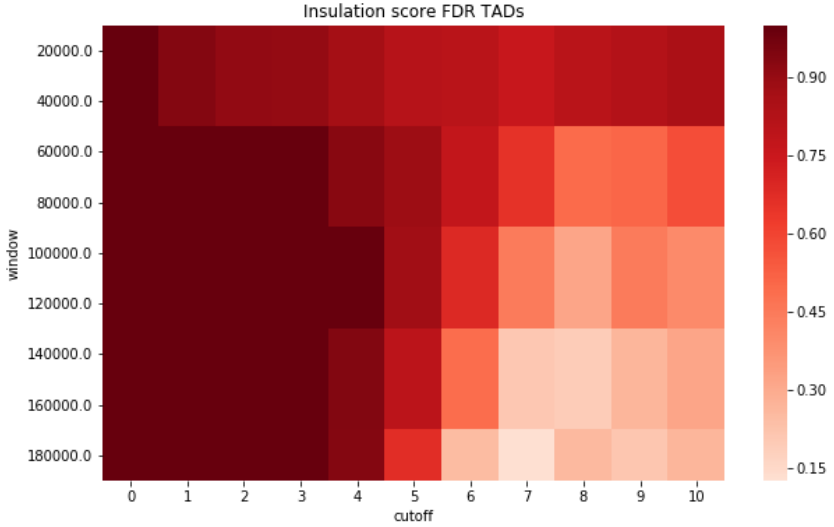
TPR TADs



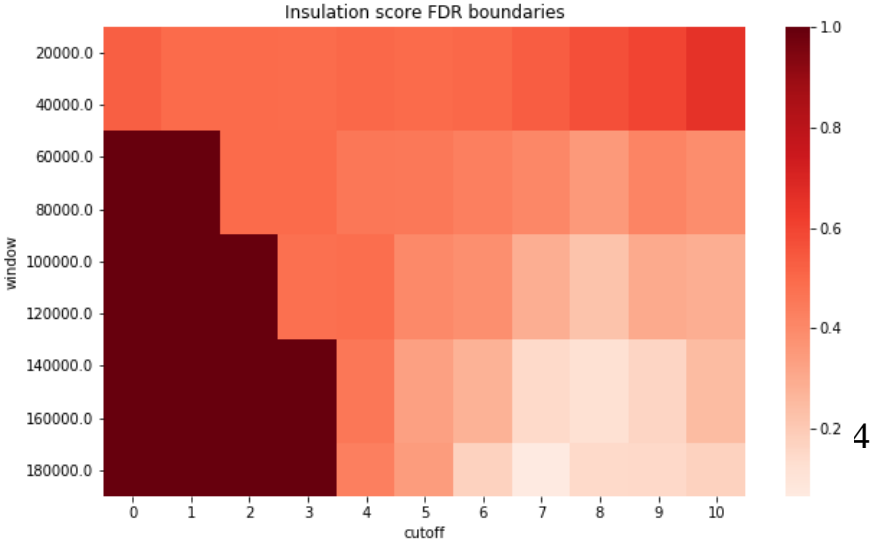
TPR boundaries



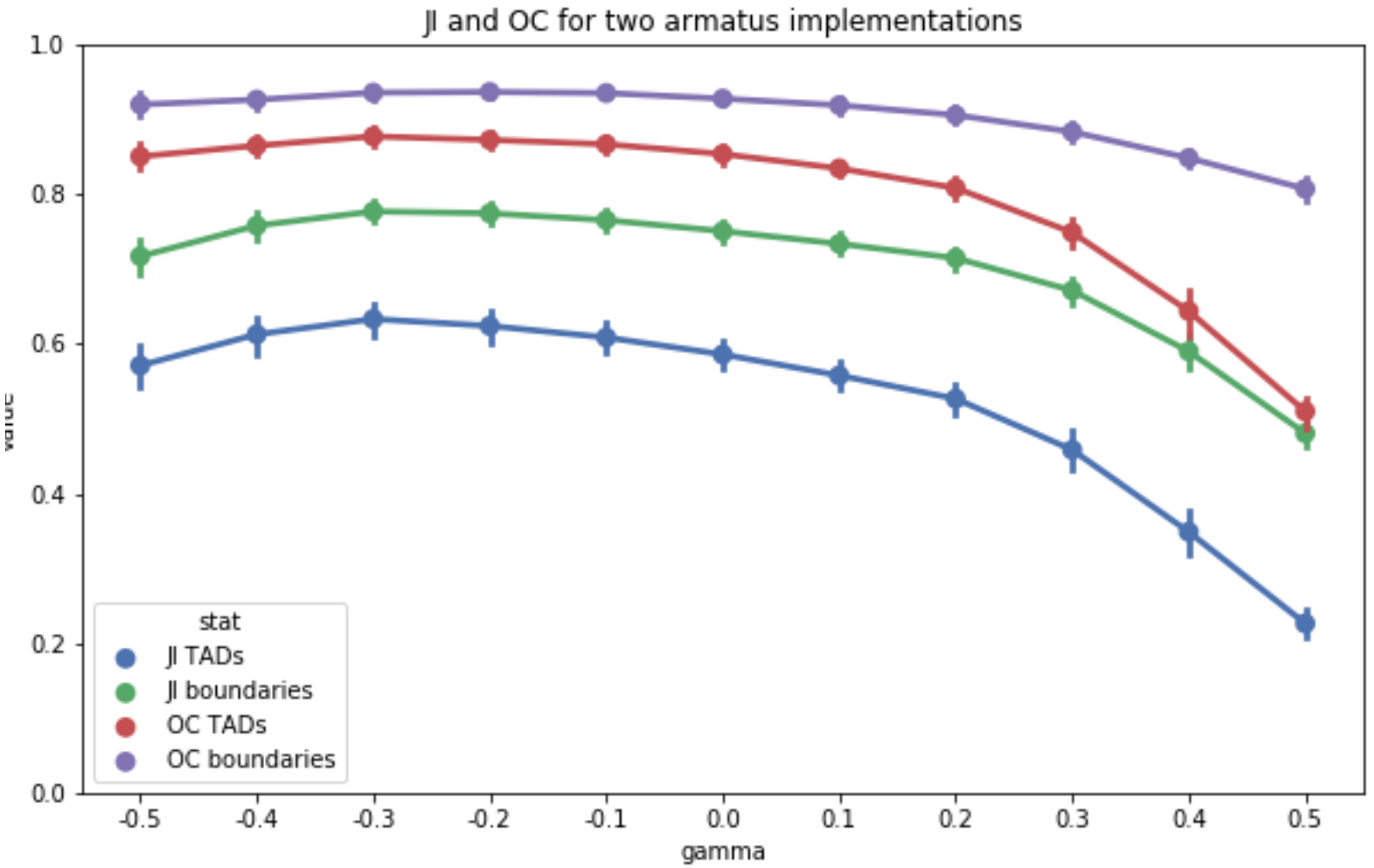
FDR TADs



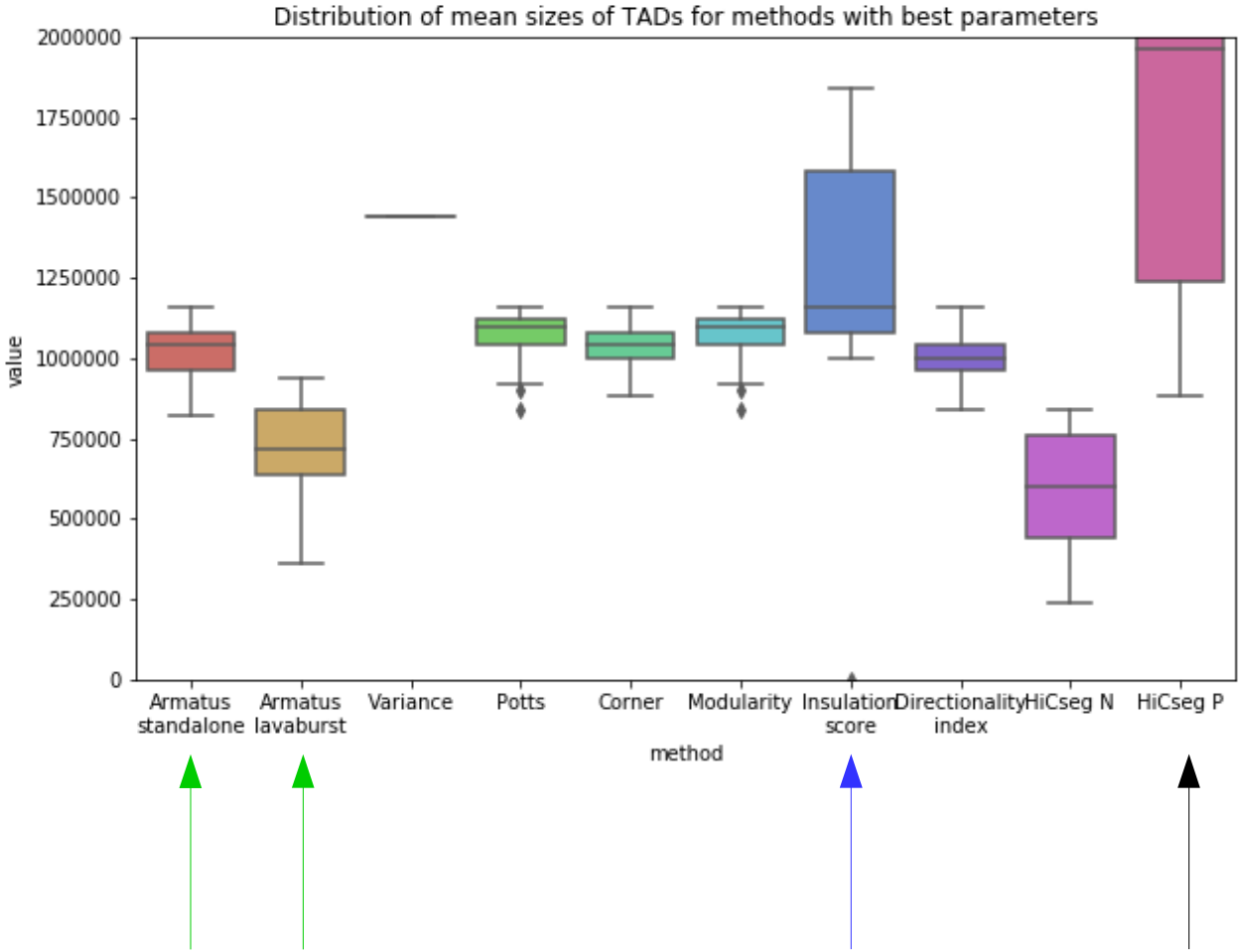
FDR boundaries



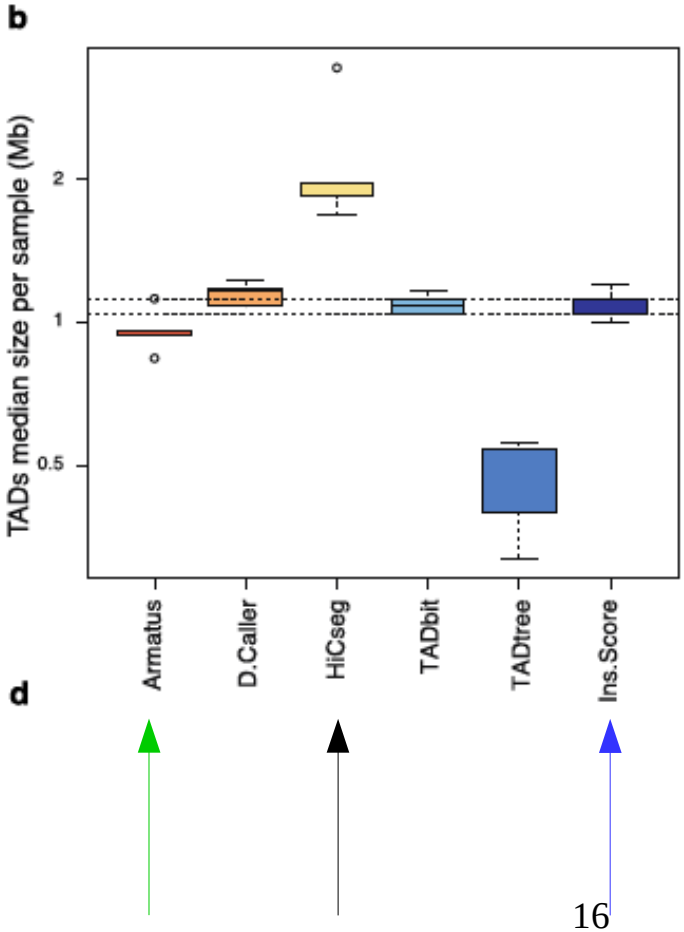
Сравнение двух Armatus



Средняя длина ТАДов

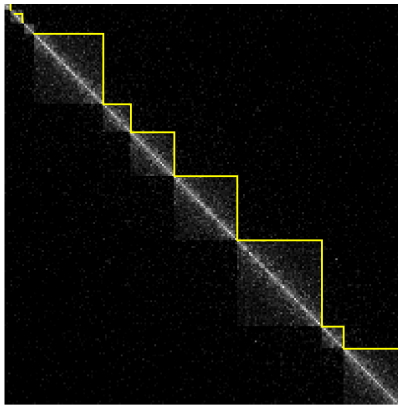


В статье

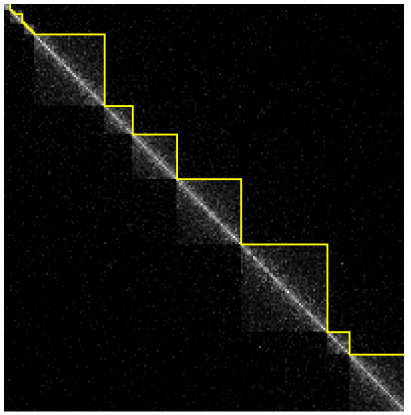


ТАДы

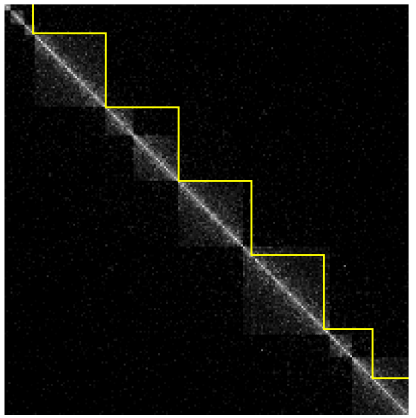
Armatus C++



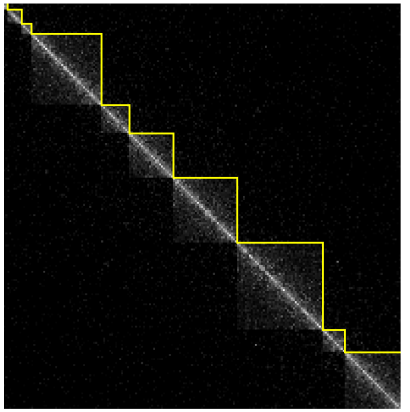
Armatus lava



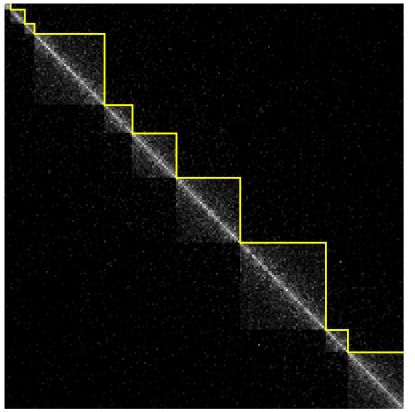
Variance



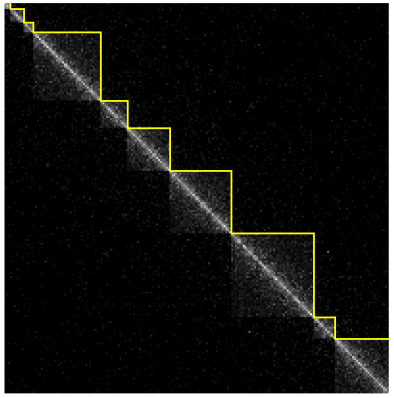
Modularity



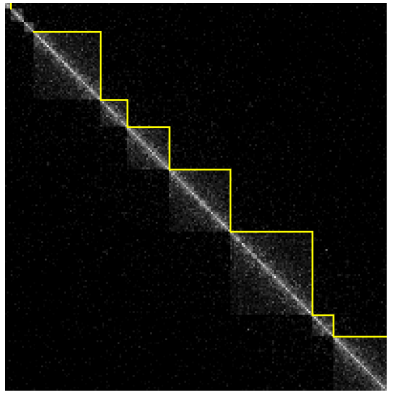
Potts



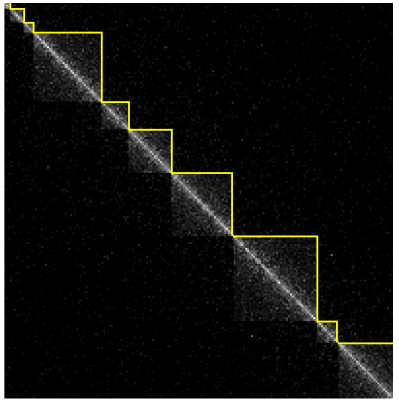
Corner



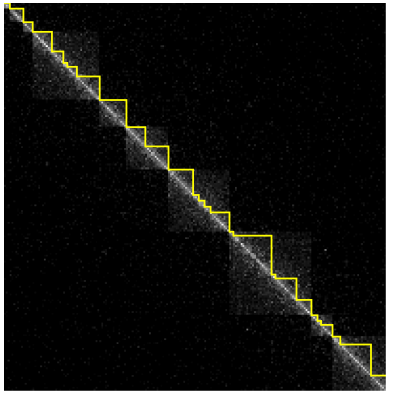
Insulation



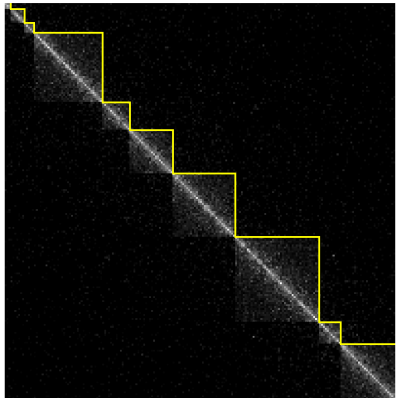
Directionality



HiCseg



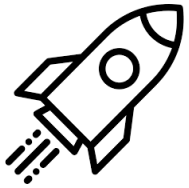
TRUE

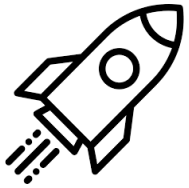


Итоги по симулированным данным

Armatus C++ 

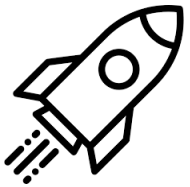
Armatus python

Modularity 

Potts 

Corner

Variance 

Insulation score  

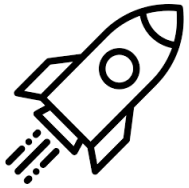
Directionality index 

HiCseg  

Итоги по симулированным данным

Armatus C++ 

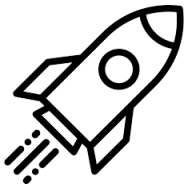
Armatus python

Modularity 

Potts

Corner

Variance

Insulation score  

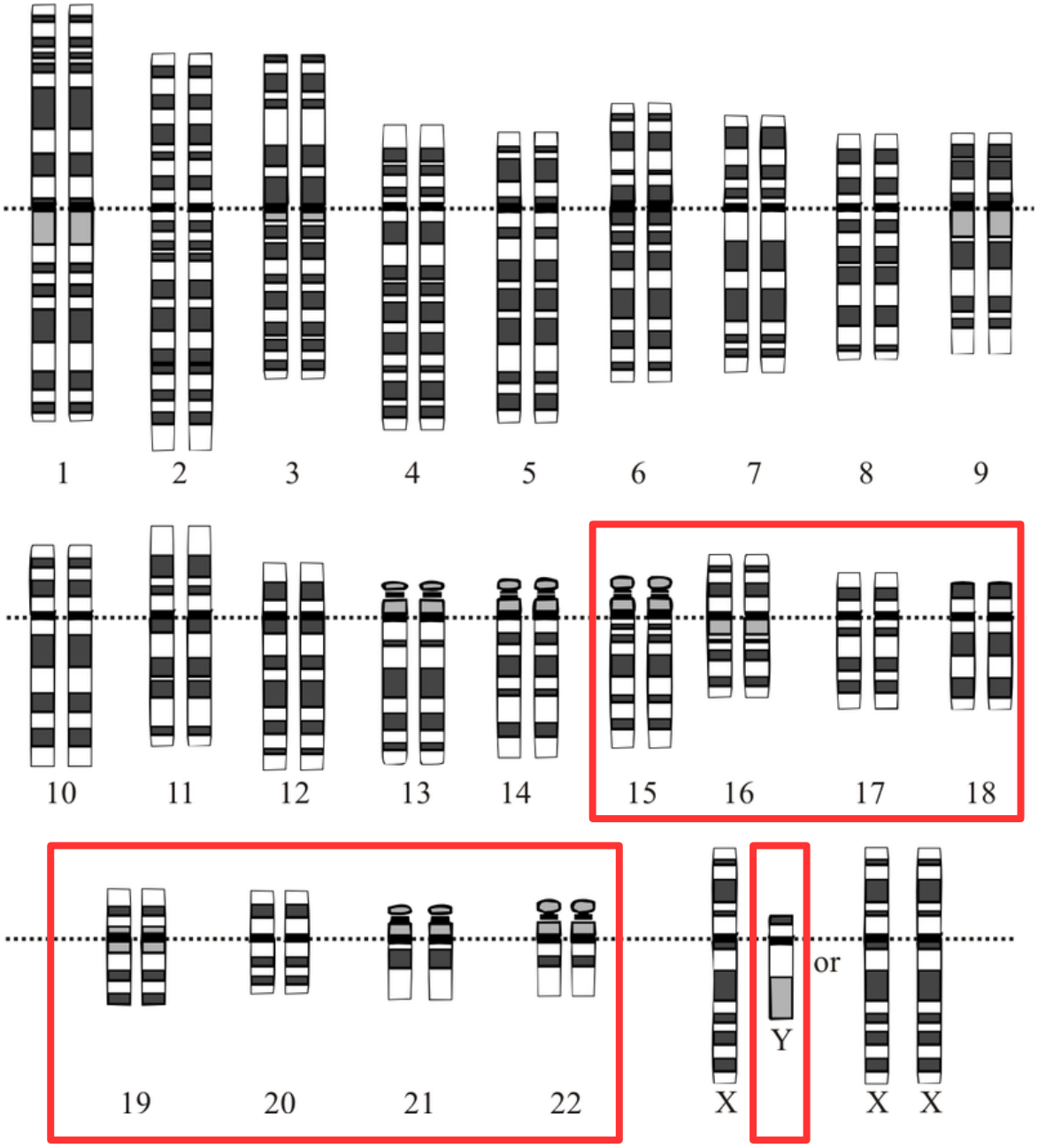
Directionality index 

HiCseg  

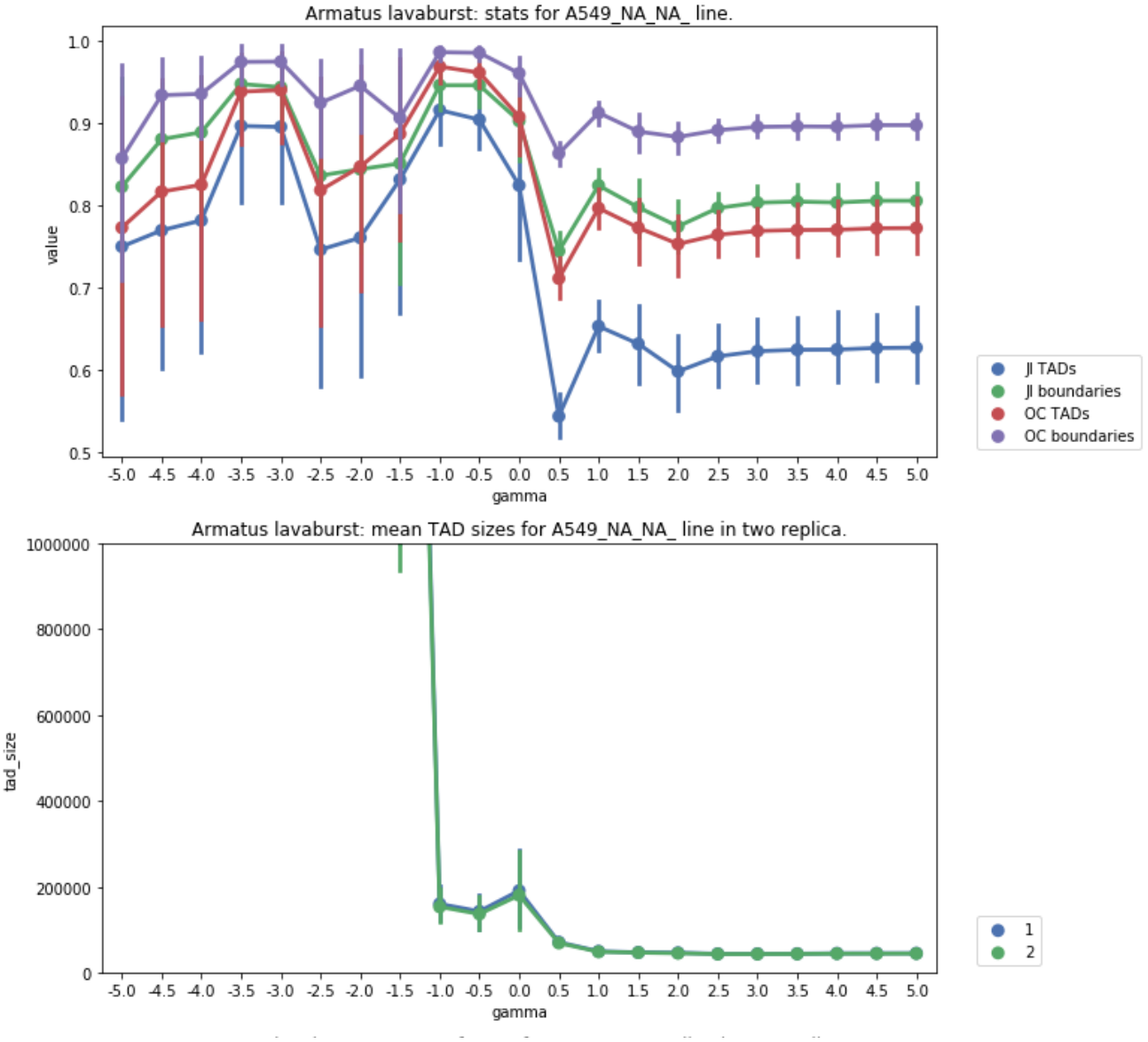
Реальные данные

Клеточные линии

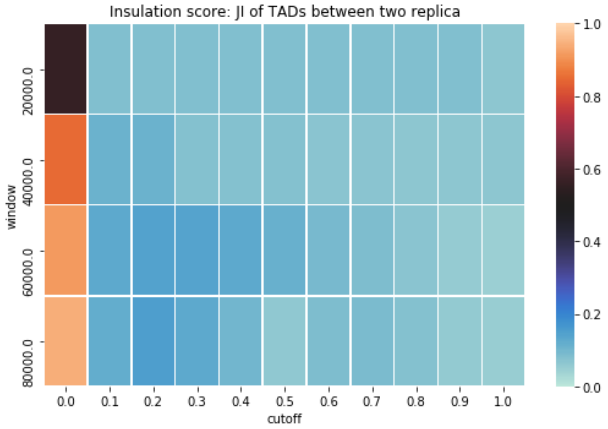
- A549
- HEK293
- HepG2
- RAD21cv

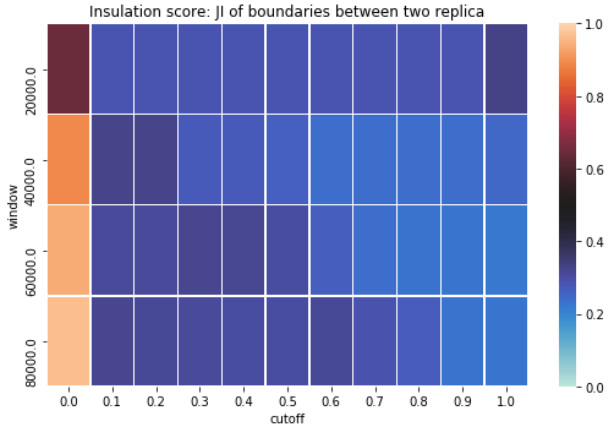


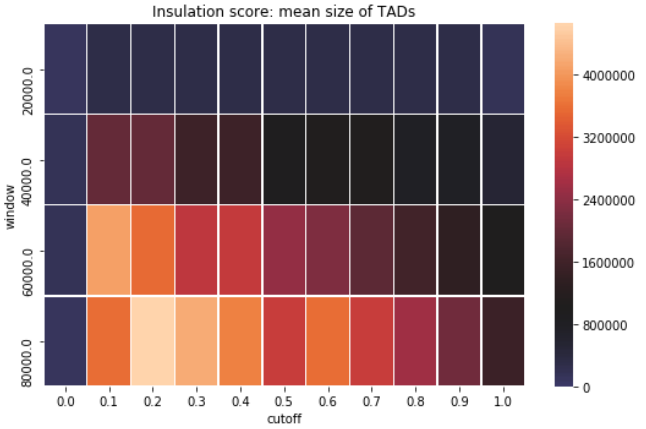
Подбор параметра: максимизация JI и OC между репликами

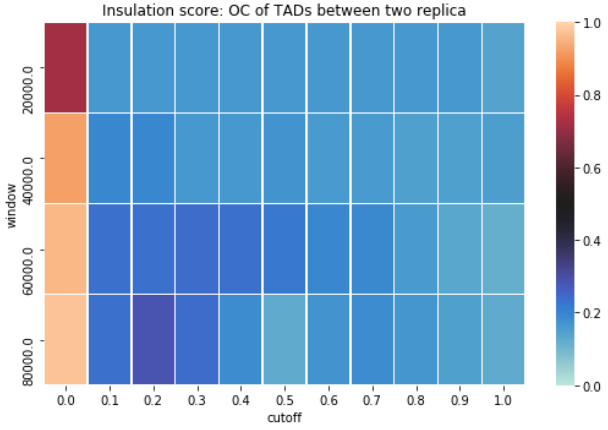


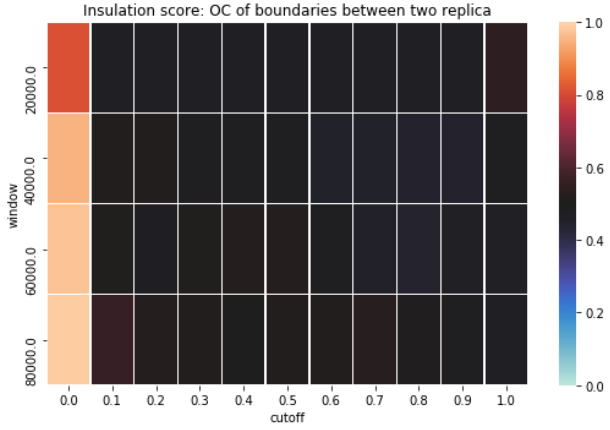
Плохие алгоритмы

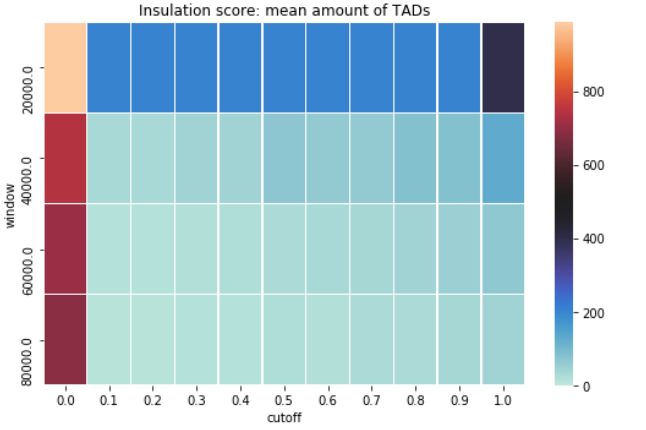




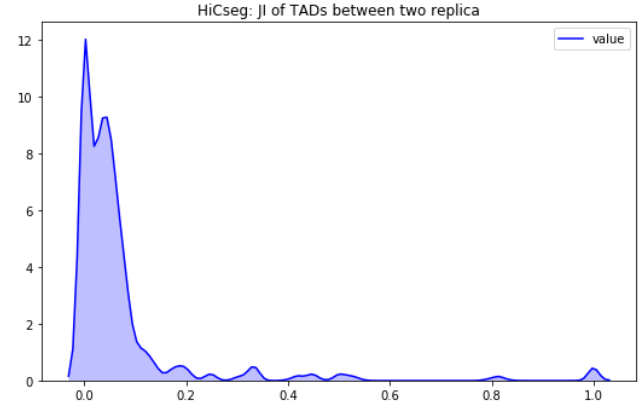


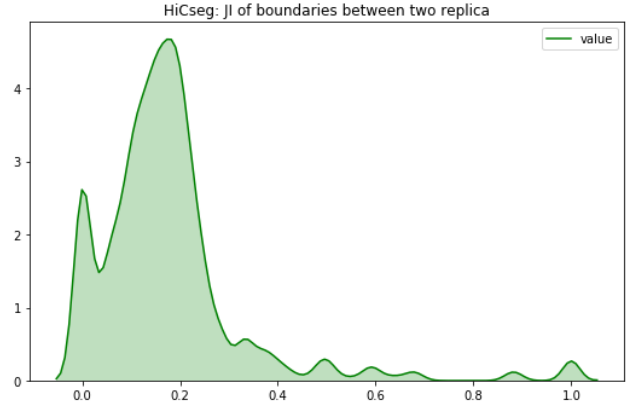


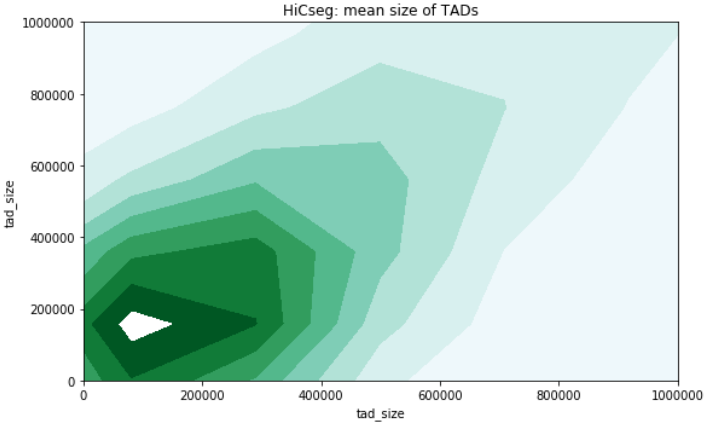


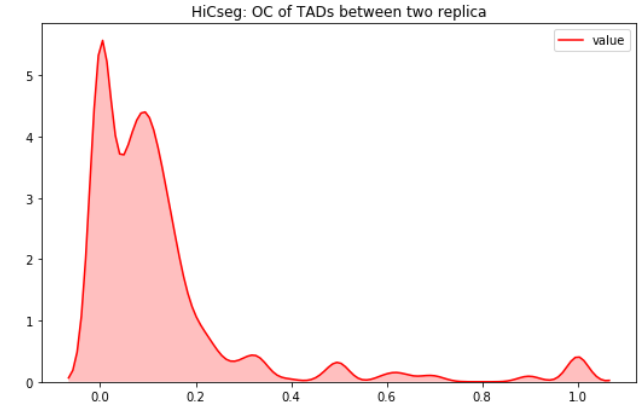


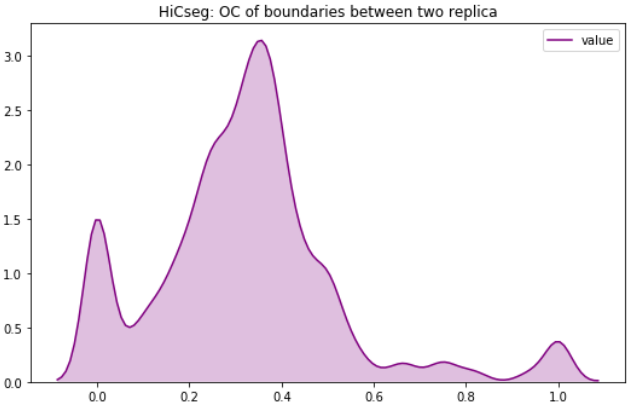
Плохие алгоритмы

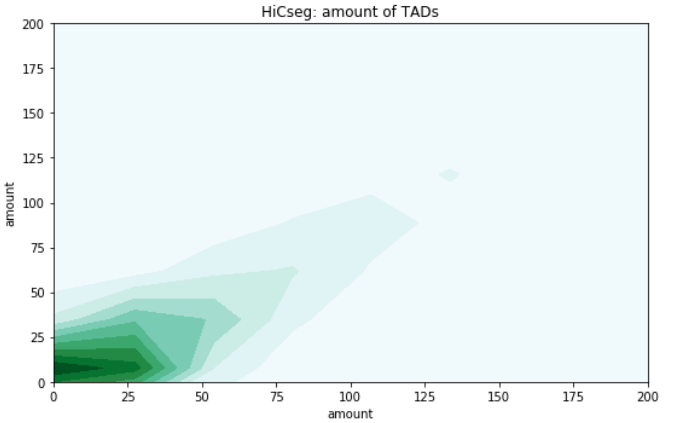




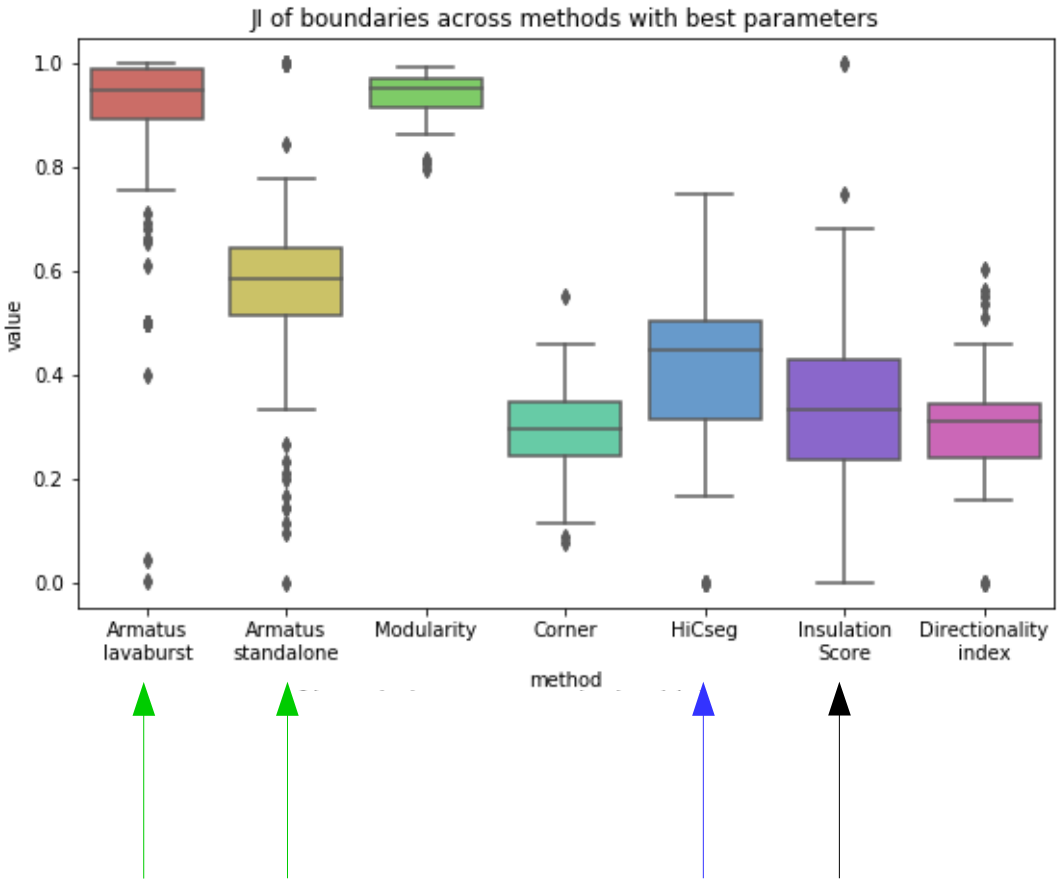


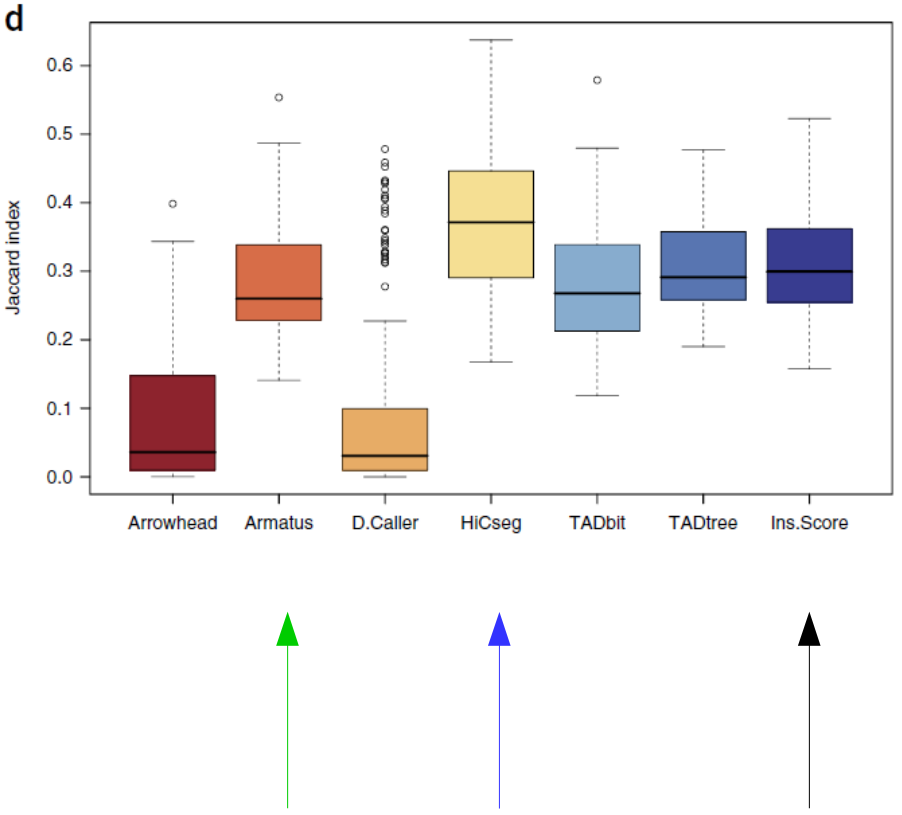






Статистика по репликам



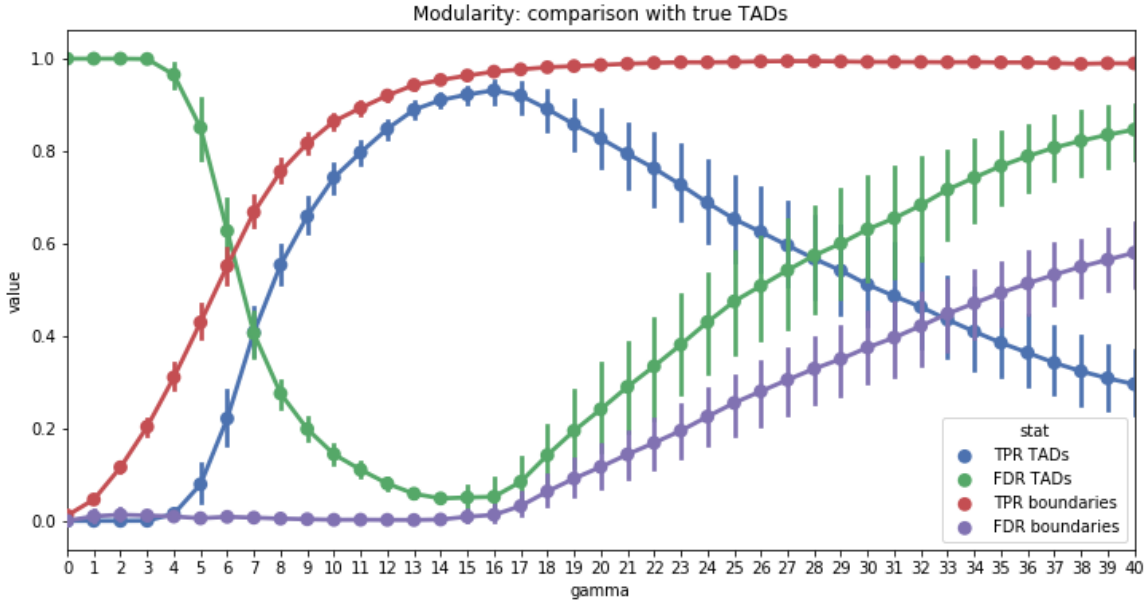


Планы на будущее

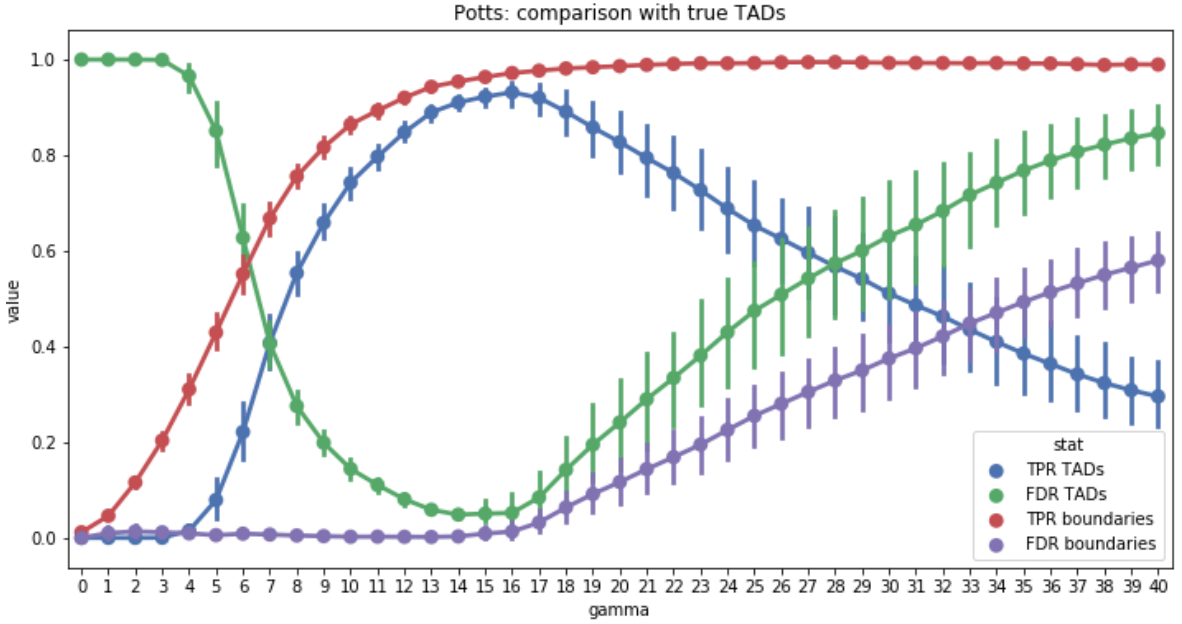
- Больше алгоритмов
- Точнее подбор параметров
- Больше линий
- Биология?

Приложение

Близнецы?

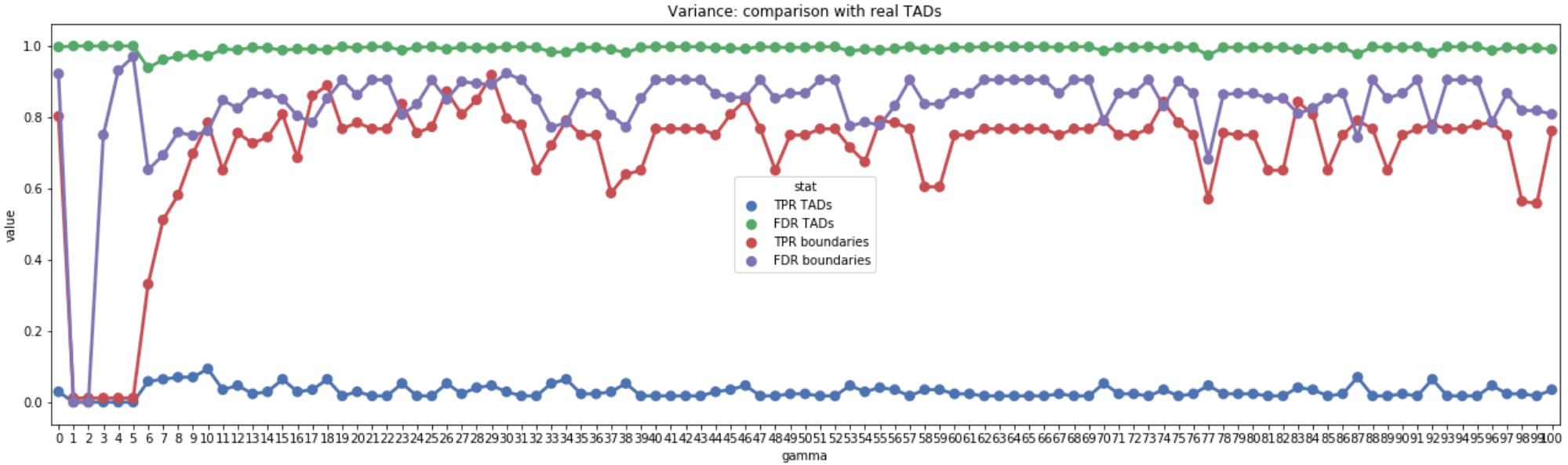


Modularity

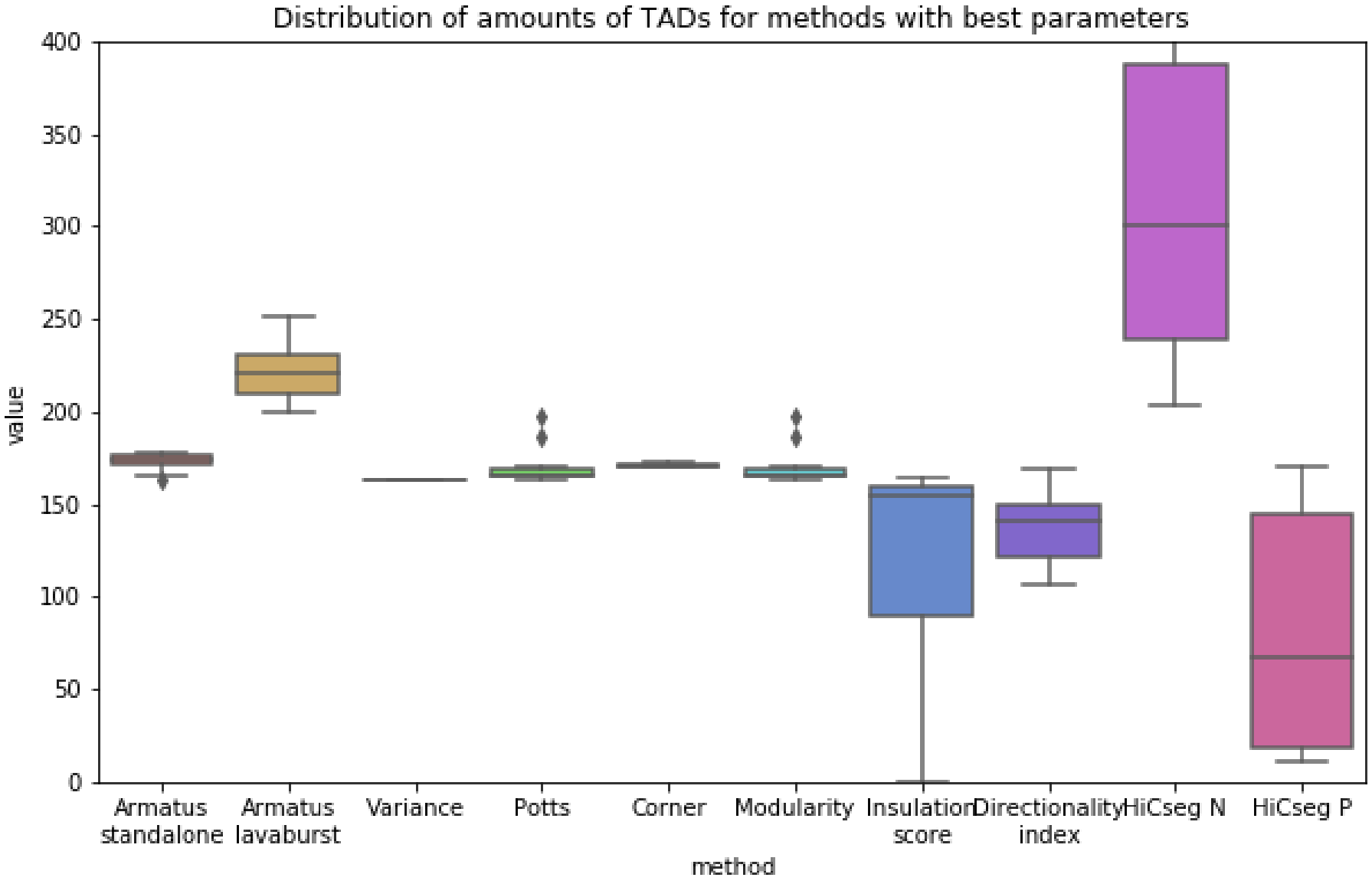


Potts

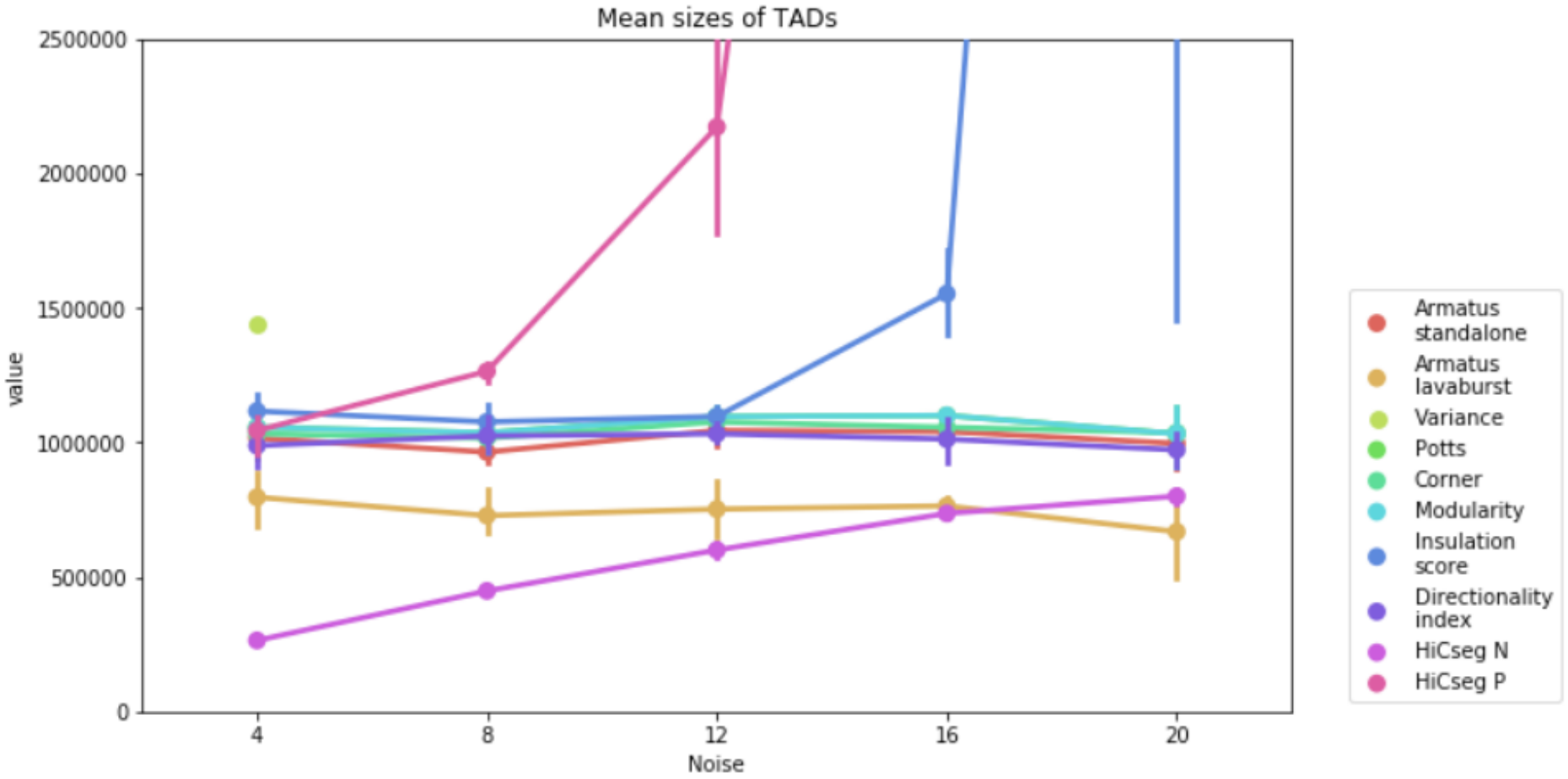
Сумасшедший



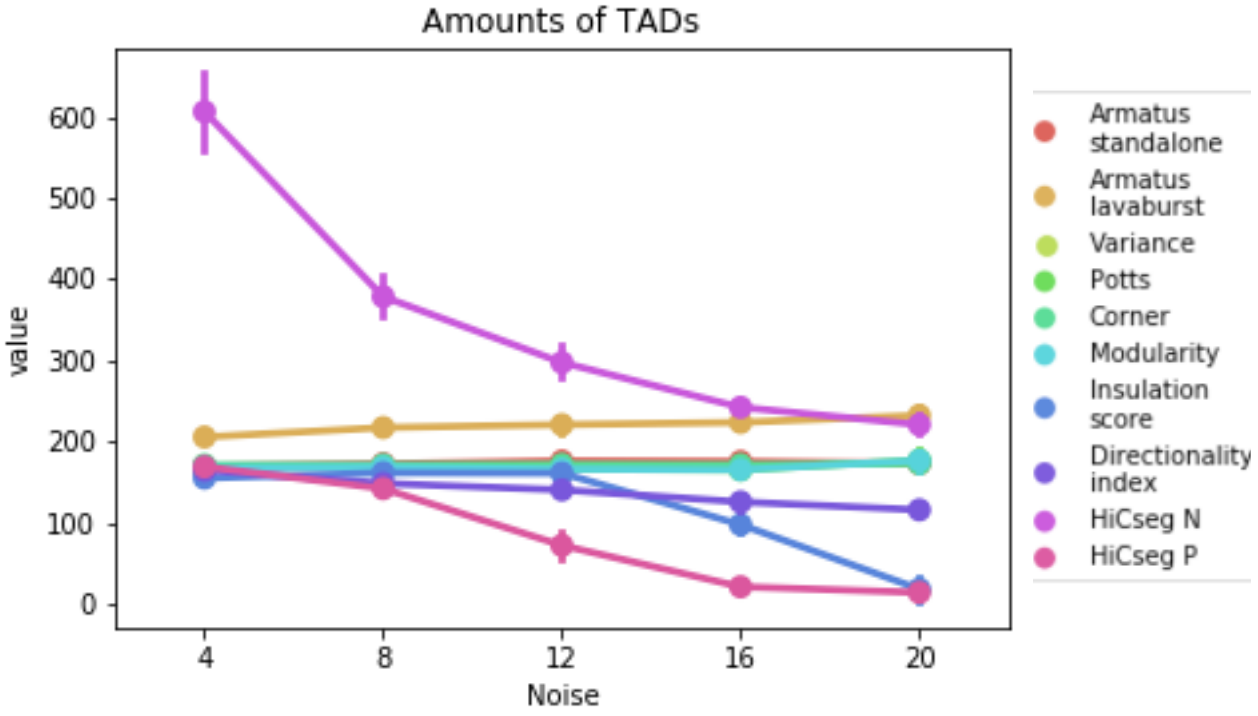
Сравнение алгоритмов с лучшими параметрами

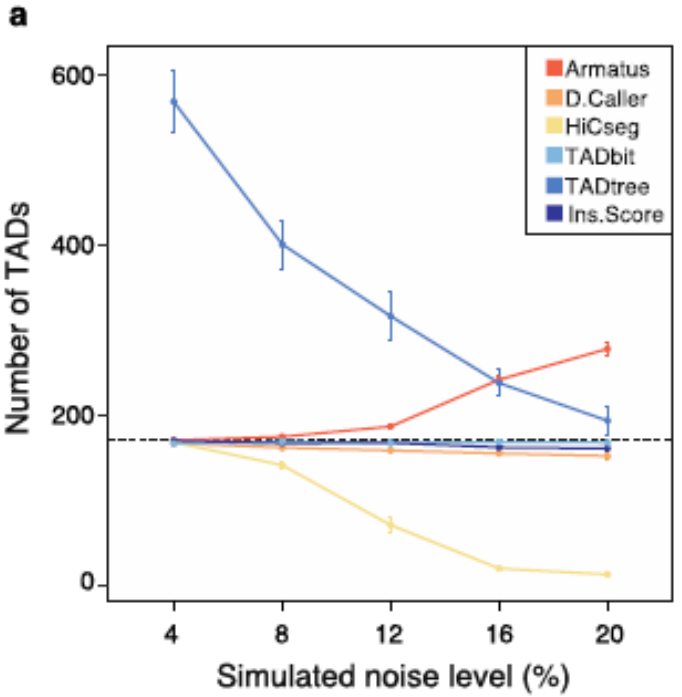


Размеров ТАДов от шума

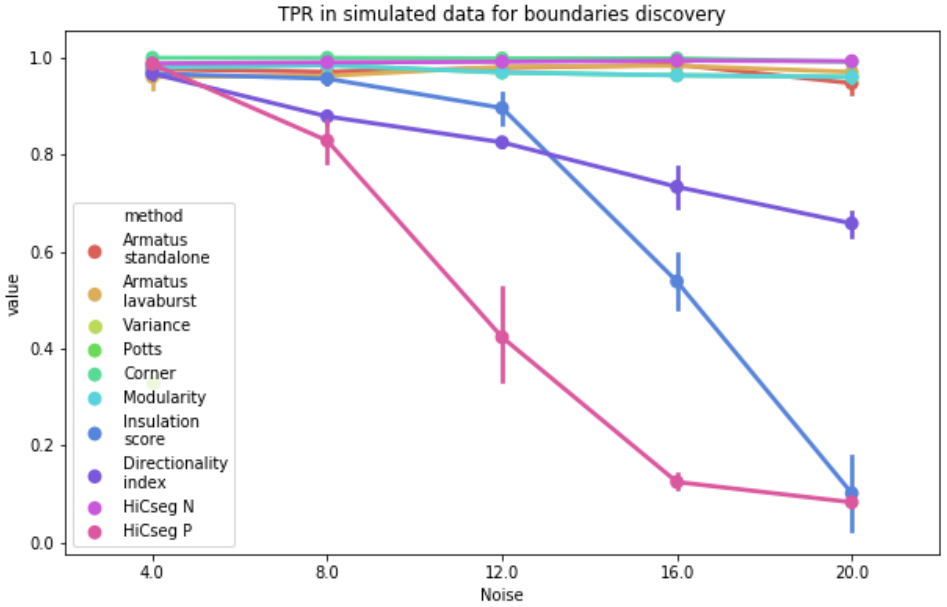


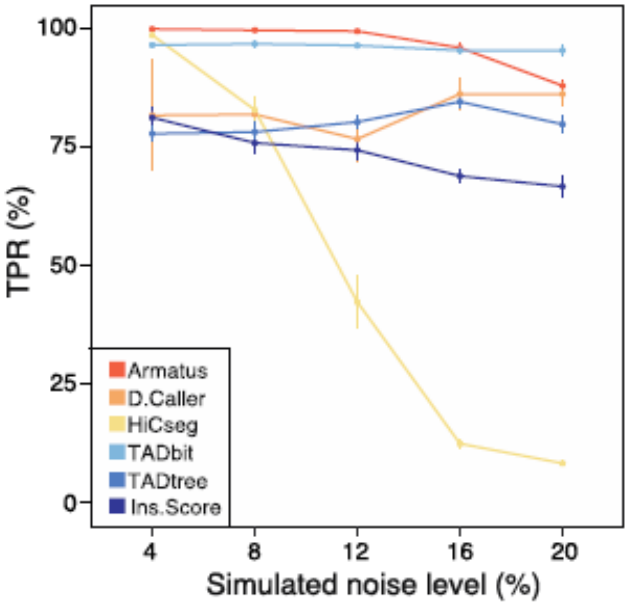
Число ТАДов от шума

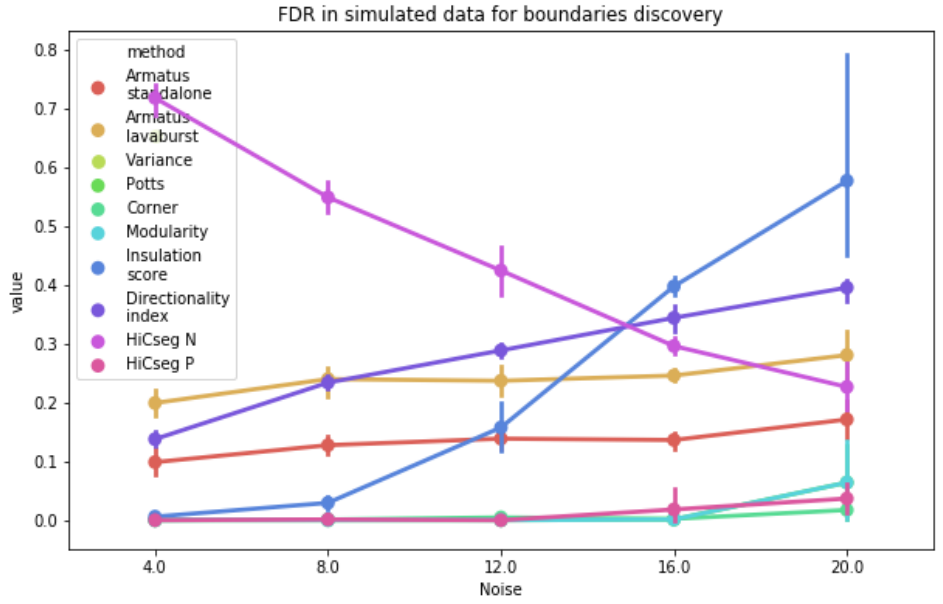


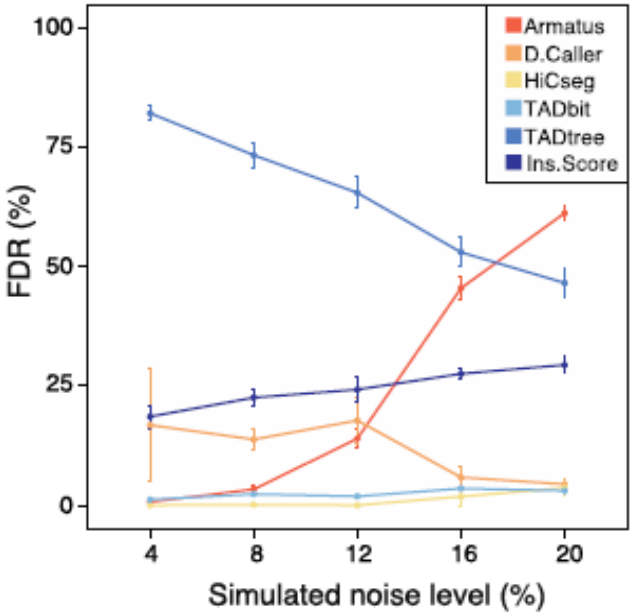


Границы ТАДов от шума









Статистика по репликам