Анализ данных полноэкзомного секвенирования

Юрий Барбитов

Институт биоинформатики

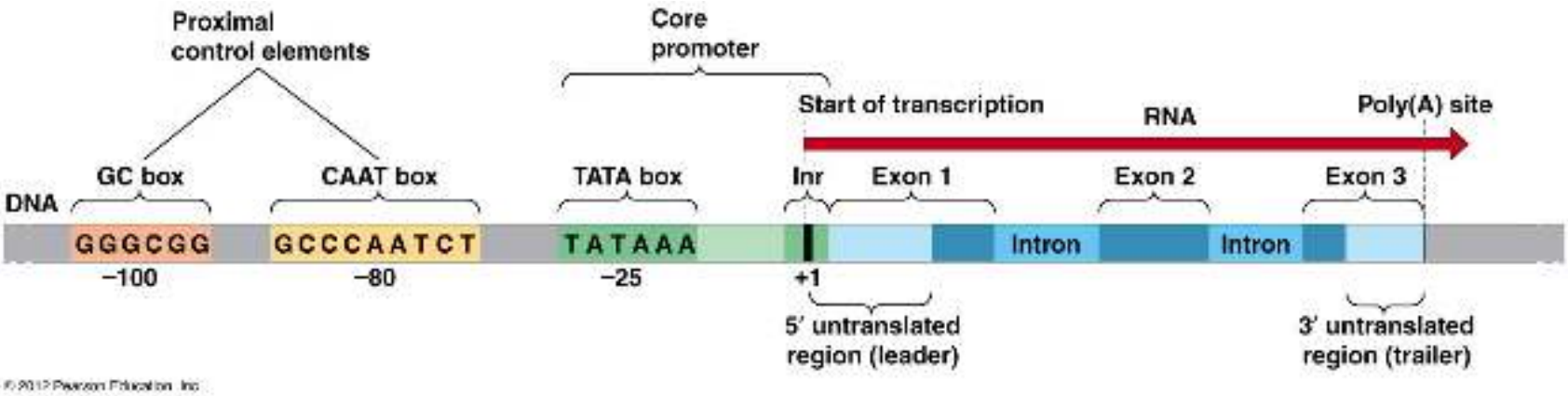
28 мая 2016 г.

Итоговые презентации проектов в Институте Биоинформатики

Руководитель: А. С. Глотов (РЦ "Биобанк" Научного Парка СПбГУ)
Александр Предеус (Институт биоинформатики)

Введение

Типичный человеческий ген имеет интрон-экзонную структуру

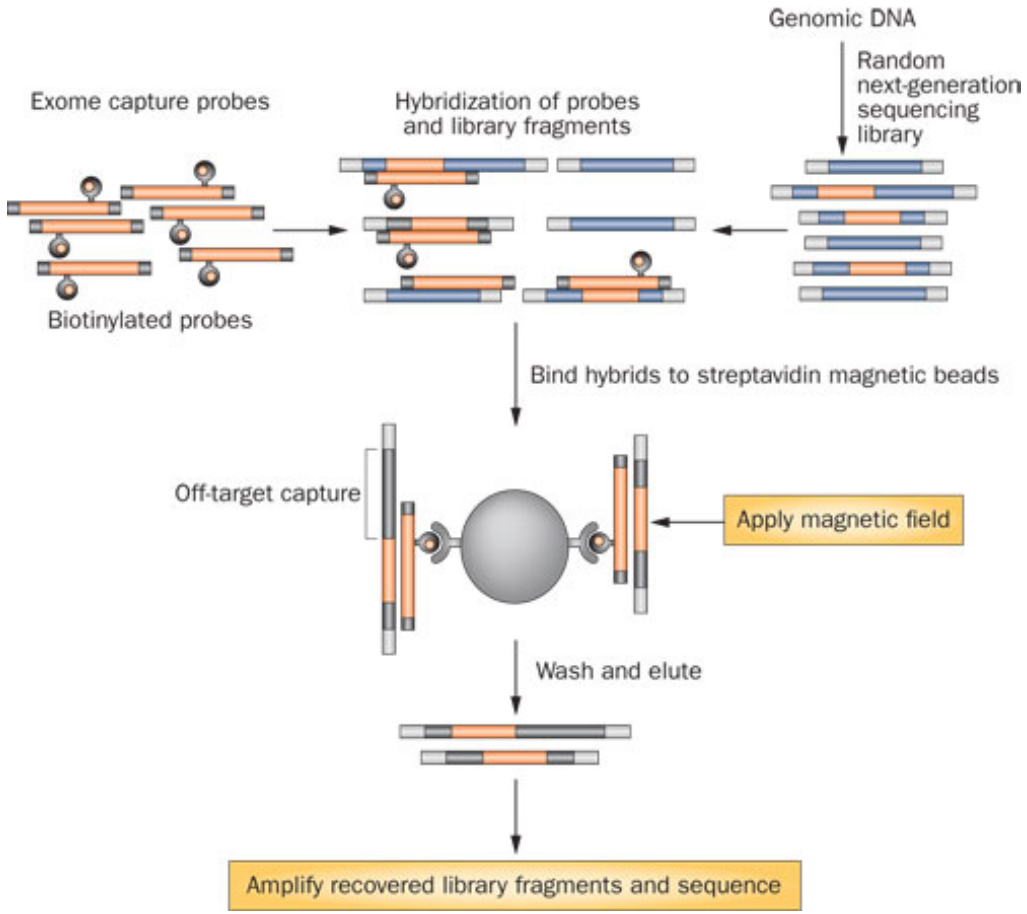


Кодирующие участки генома (экзоны) составляют всего около 1-2% всей его длины.

Методология

Два основных подхода к экзомному обогащению:

- Твердофазный (on-chip) - старая технология
- "В растворе" (solution-based) - все современные методы



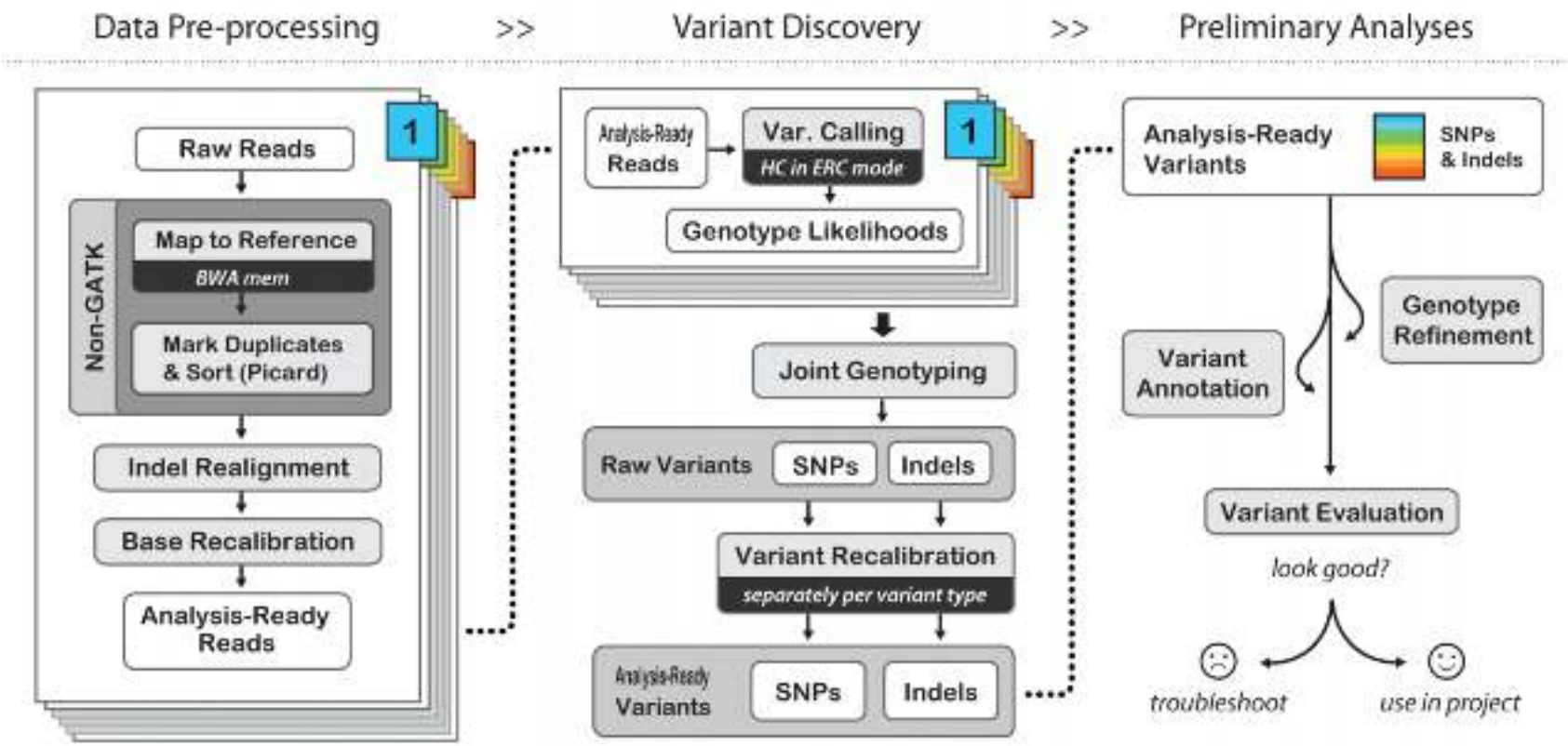
Цель

Цель проекта - наладить процесс анализа данных полноэкзомного секвенирования.

Первоначальные задачи

1. Сравнить различные технологии приготовления экзомных библиотек
2. Построить пайплайн анализа данных и определения вариантов
3. Составить локальную базу вариации и "локальный референсный экзом"

Пайплайн анализа данных



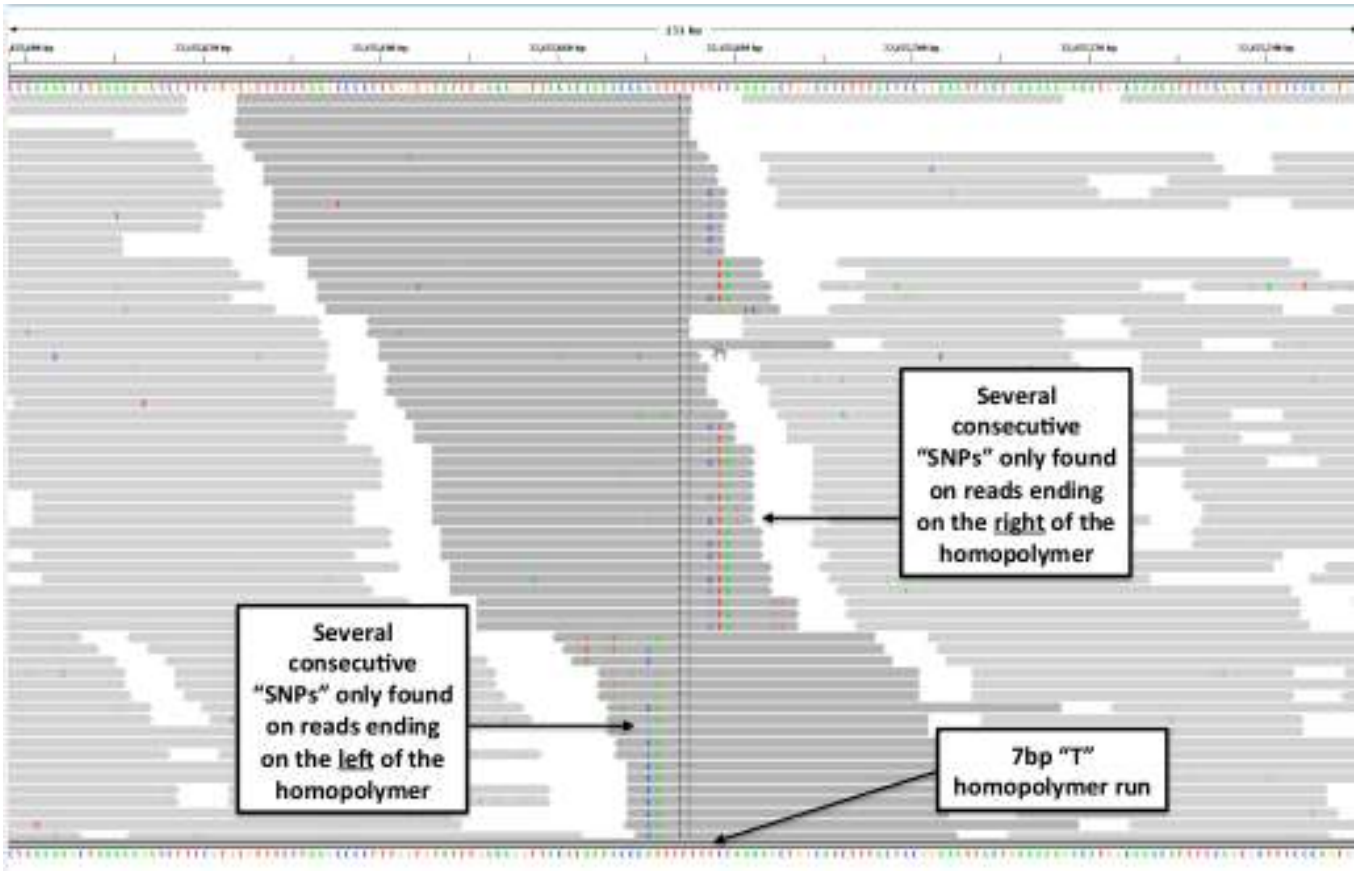
Маркировка дупликатов

Showing duplicate reads

Hiding duplicate reads

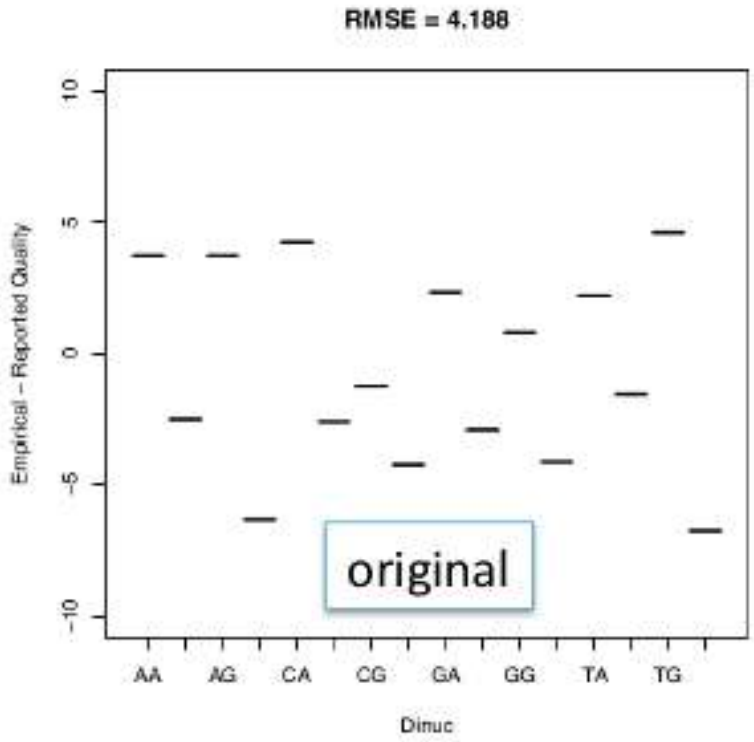
Перевыравнивание инделов

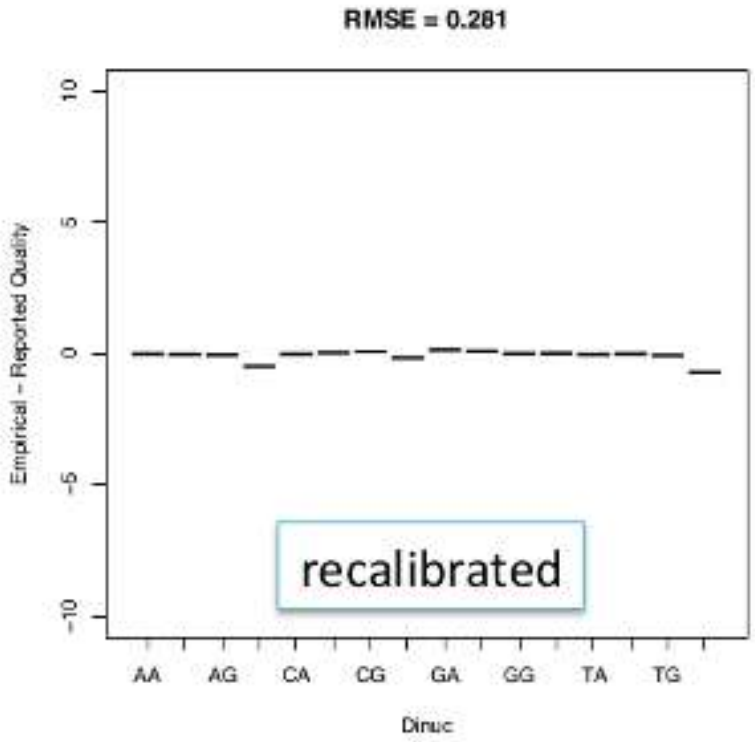
GATK IndelRealigner избавляет от ложных снипов около инделов.



BQSR

Приборы часто врут при оценке качества оснований!
Перекалибровка качества убирает bias.



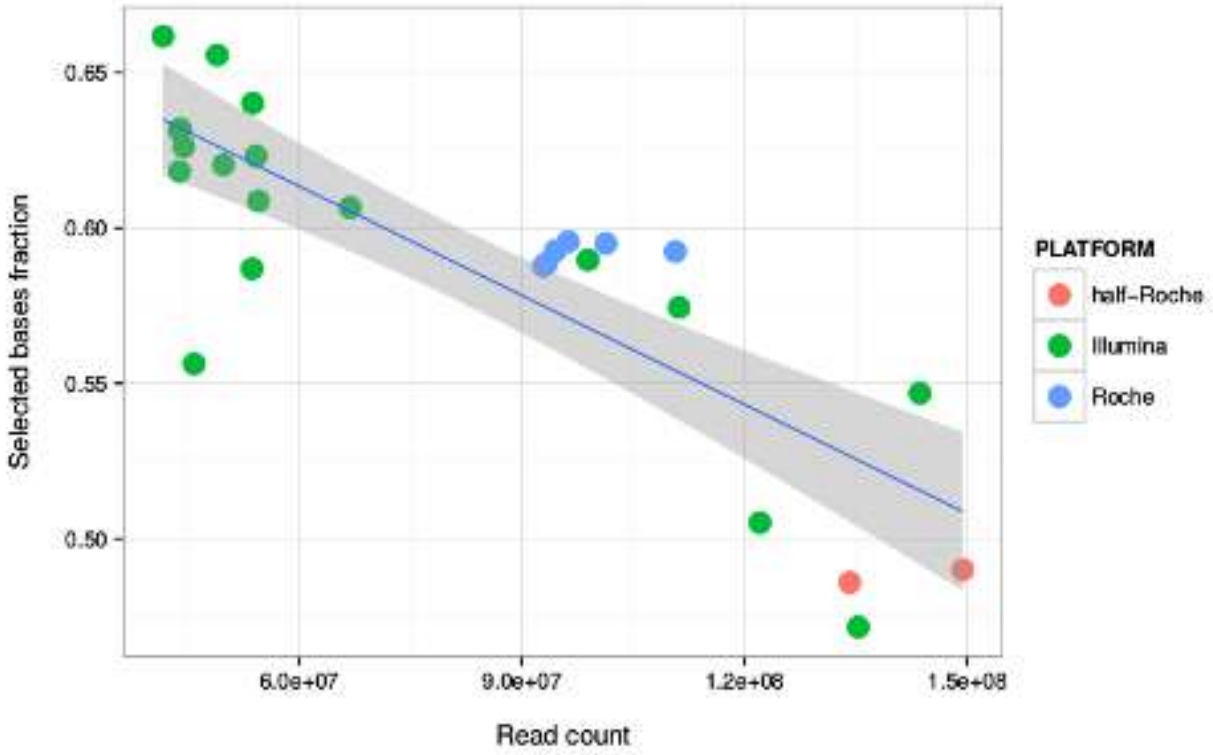


Исходный материал

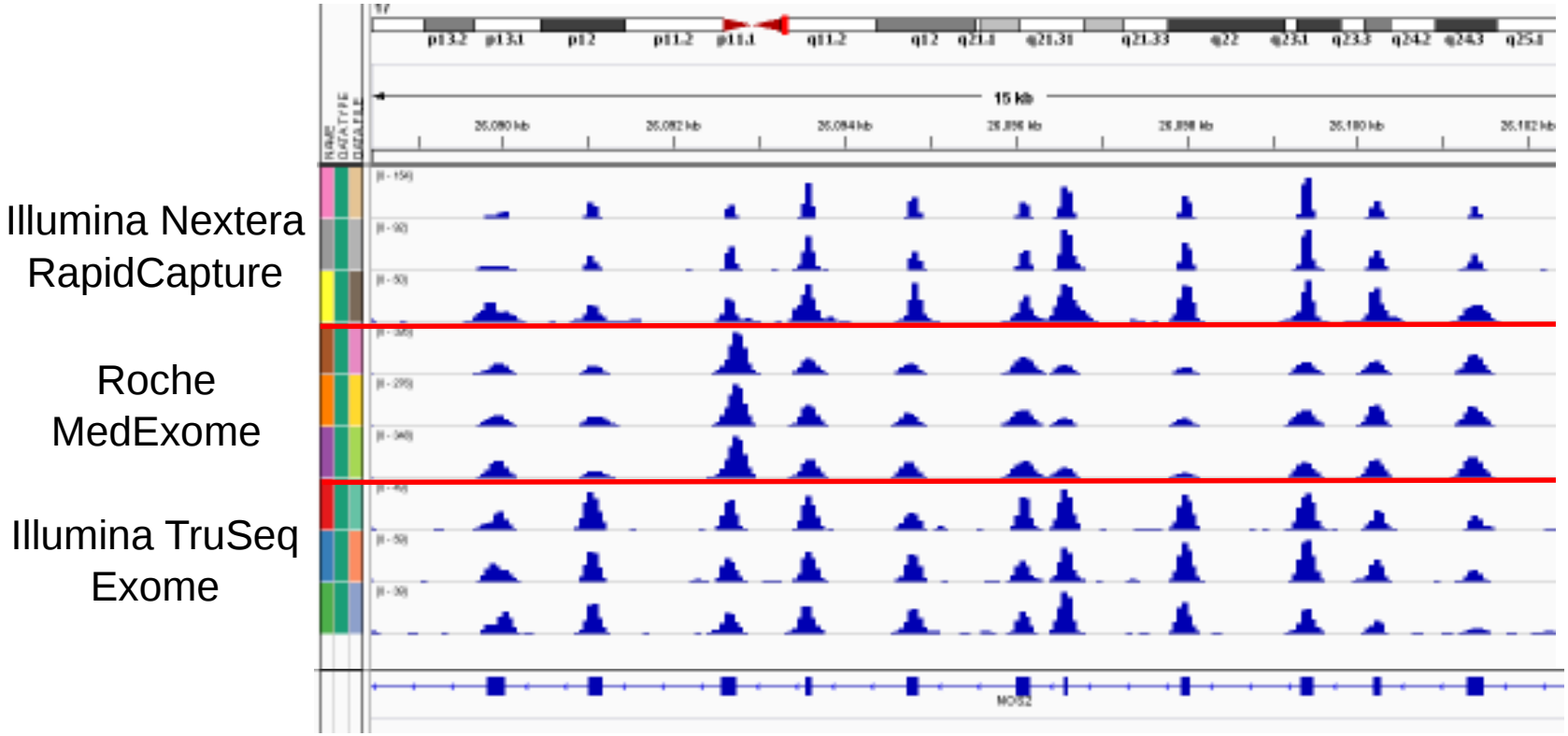
- 110 клинических образцов пациентов.
- Три используемые технологии - Illumina Nextera RapidCapture, Roche MedExome и Illumina TruSeq Exome.
  Illumina vs. NimbleGen
- По технологиям - 48 образцов сделаны набором TruSeq, 30 - Roche MedExome, 32 - Illumina Nextera. Четыре образца сделаны по двум протоколам.

Эффективность обогащения

Эффективность обогащения значимо зависит от глубины библиотеки, и незначительно - от технологии

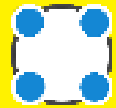


Покрытие экзома - пример

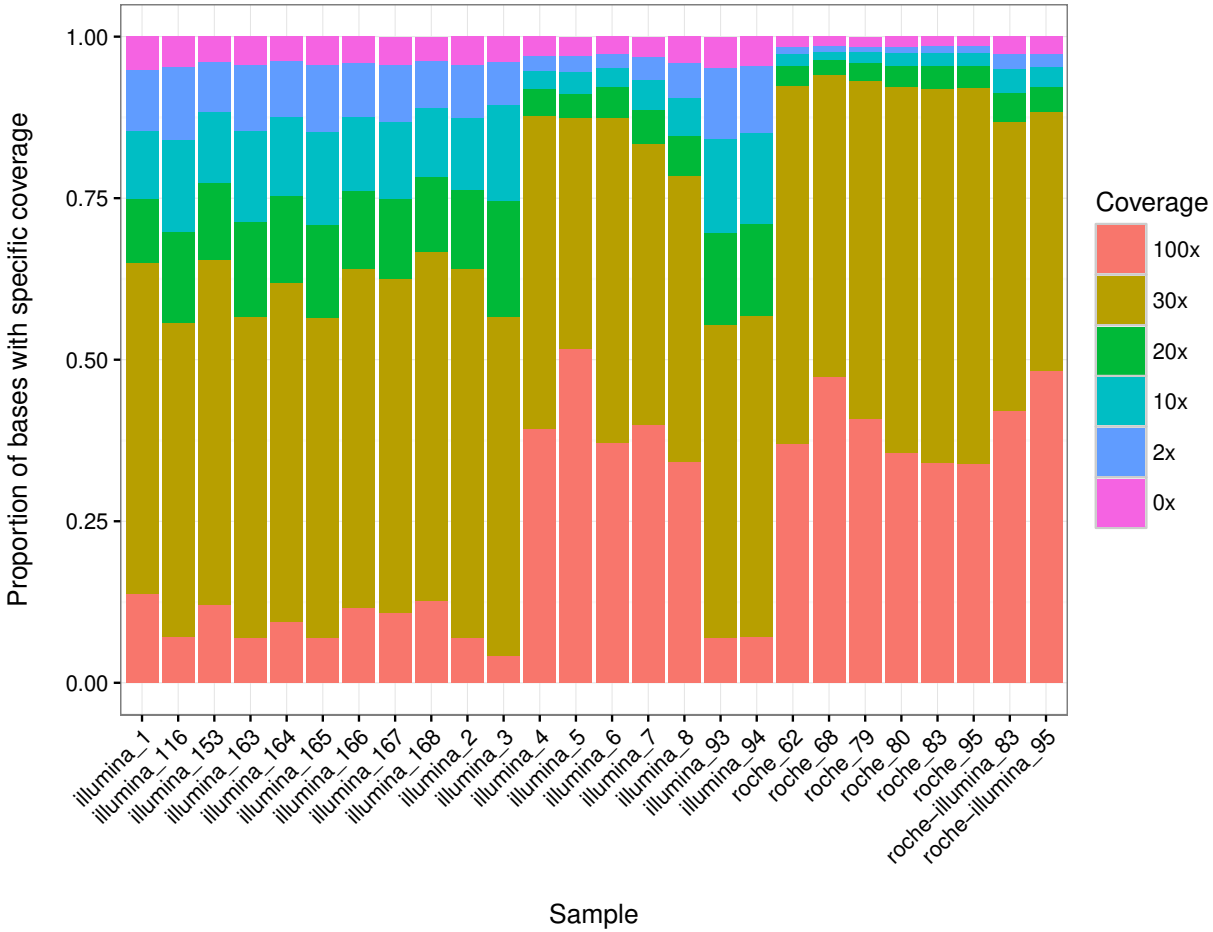


NOS2

Покрытие экзома - в цифрах

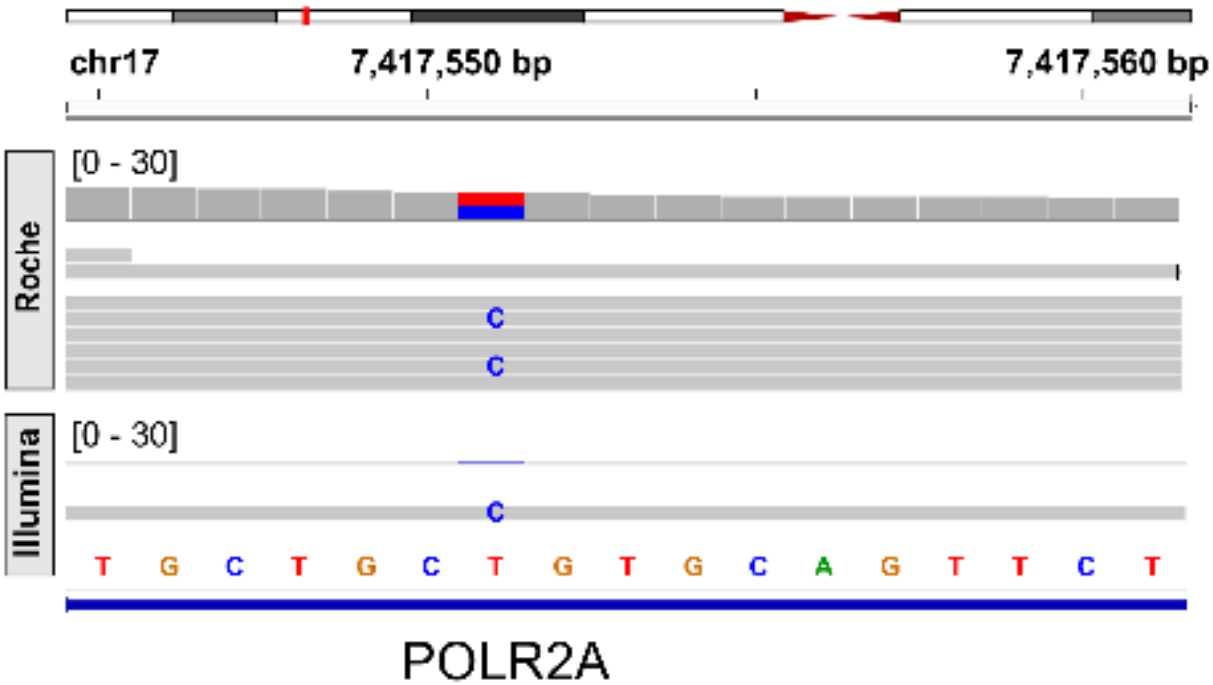


Как это выглядит в числах:



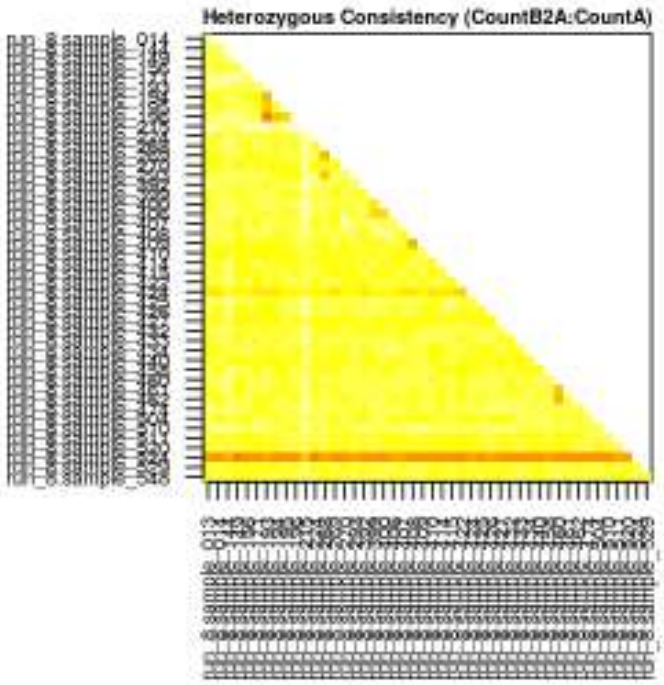
Покрытие экзома - еще пример

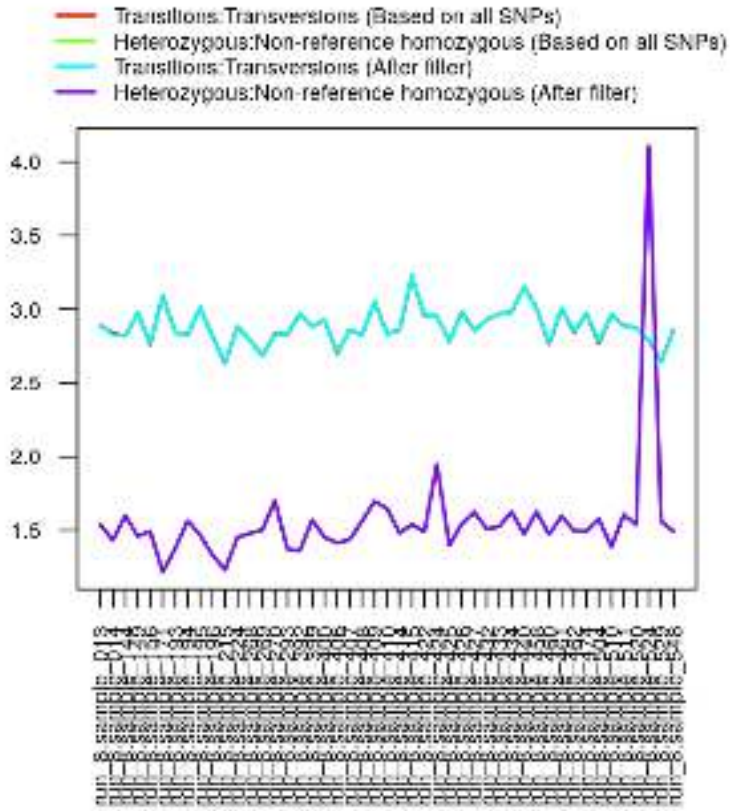
И как это выглядит в клиническом смысле:



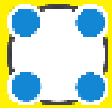
QC of variant calls

Как проконтролировать загрязнение образцов?





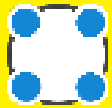
Клиническая интерпретация



Для клинической интепретации существуют различные сложные рекомендации. Мы воплотили собственную метрику, *IVS*, для приоритизации важных вариантов. Она складывается из:

- Эффекта варианта. Чем выше вероятность loss-of-function - тем лучше.
- Частоты варианта в 1000 геномов, ExAC-e, ESP6500.
- Частоты варианта в нашей когорте.
- Предсказаний патогенности - PROVEAN, SIFT, PolyPhen2 (HVAR), fathmm-MKL.

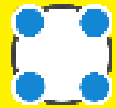
Клиническая интерпретация



Пример подтвержденной по Сэнгеру патогенной замены у одного из пациентов, легко обнаруживаемой при помощи нашей метрики.

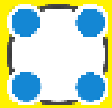
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Gene	Chr	Position	rsID	REF	ALT	EffType	IVS	OMIM	ClinVar	PROVEAN	SIFT	Polyphen2	fathmm-MI	1000G-AF	ExAc-AF	ESP6500	OUR_AF	S27	S59
2	GCK	7	44187340	-	C	A	PROTEIN_INTERACTION	12	-	-	Deleterious	Damaging	Damaging	Pathogenic	0	0	0	0.006757	0/0:29/0	0/1:11/27
3	GCK	7	44198764	rs150560724	C	T	UTR_5_PRIME	3.83335	-	-	-	-	-	Benign	0.00599042	0.008475	0.007535	0.016	0/0:13/0	0/1:9/10
4	ABCC8	11	17426996	rs200276273	G	T	INTRON	3.78456	-	-	-	-	-	Benign	0.00259585	0.011	0.011012	0.048	0/0:7/0	0/1:2/4
5	GCGR	17	79768834	rs140065949	C	T	INTRON	3.54084	-	-	-	-	-	Benign	0.0183706	0.023	0.024091	0.007042	0/0:40/0	0/1:25/24
6	MOG	6	29643875	-	A	ATG	DOWNSTREAM	1.868	-	-	-	-	-	-	0	0	0	0.066	0/0:22/0	0/1:15/5/0
7	SLC16A1	1	113460676	-	CA	CA	INTRON	1.606	-	-	-	-	-	-	0	0	0	0.197	0/0:9/0/2	0/1:0/2/21/0
8	SGCG	13	23808732	-	CT	C	INTRON	1.088	-	-	-	-	-	-	0	0	0	0.456	0/1:42/9	0/1:47/10
9	HNF1A	12	121439588	rs11065390	G	A	UTR_3_PRIME	0.95304	-	-	-	-	-	Benign	0.120008	0.042	0	0.097	0/0:36/0	0/1:8/16
10	SCEL	13	78178550	rs2274016	G	A	NON_SYNONYMOUS_CO	0.87695	-	-	Neutral	Damaging	Benign	Pathogenic	0.235224	0.145	0.098431	0.162	0/1:26/9	0/1:7/13
11	LIN51	15	101109683	rs1047320	T	C	SYNONYMOUS_CODING	0.85718	-	-	-	-	-	-	0.0888562	0.016	0.018069	0.061	0/0:35/0	0/1:42/71
12	HADH	4	108911051	rs17550794	T	C	UTR_5_PRIME	0.81232	-	-	-	-	-	Benign	0.13099	0.08	0.144243	0.115	0/0:36/0	0/1:6/16
13	LOC10193	2	88861757	rs1800980	T	G	INTRON	0.76042	-	-	-	-	-	Benign	0.132388	0.08	0.15774	0.103	0/0:56/0	0/1:26/16
14	PAX4	7	127253898	rs77039439	G	A	SYNONYMOUS_CODING	0.71847	-	Benign	-	-	-	-	0.0177716	0.045	0.05213	0.061	0/0:59/0	0/1:38/26
15	LIN51	15	101114482	rs34231390:r	AAC	A	INTRON	0.61863	-	-	-	-	-	-	0.127596	0.171	0.154137	0.21	0/1:37/4	0/1:26/18
16	PTF1A	10	23482850	rs10828415	G	A	UTR_3_PRIME	0.57172	-	-	-	-	-	Benign	0.212859	0.127	0.058588	0.054	0/0:57/0	0/1:50/41
17	INSR	19	7184651	-	GG	GG	INTRON	0.45191	-	-	-	-	-	-	0	0.215	0	0.242	0/1:4/7/	0/1:26/4/0/0
18	EIF2AK3	2	88926729	-	CC	CC	CODON_DELETION	0.45	-	-	-	-	-	-	0	0	0	0.527	1/1:0/2	0/1:5/4/0

Результаты работы и выводы



1. Нами построен и отлажен пайплайн анализа данных WES.
2. На основе проведенного QC библиотек нами сделан выбор набора для экзомного обогащения.
3. Разработана метрика для клинического ранжирования вариантов.

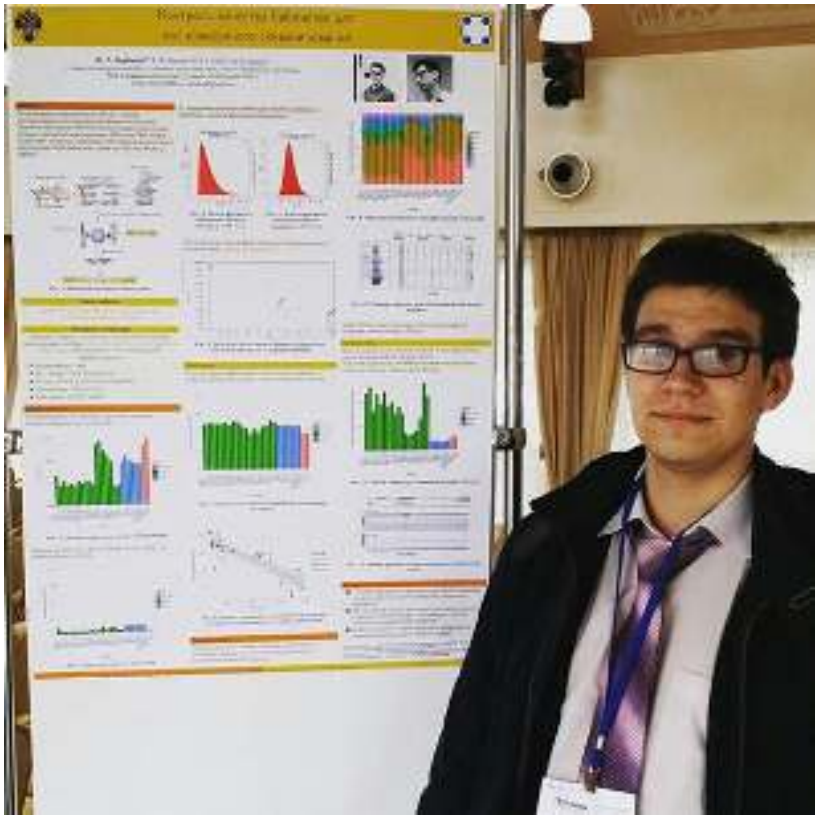
To be continued



Планы на будущее

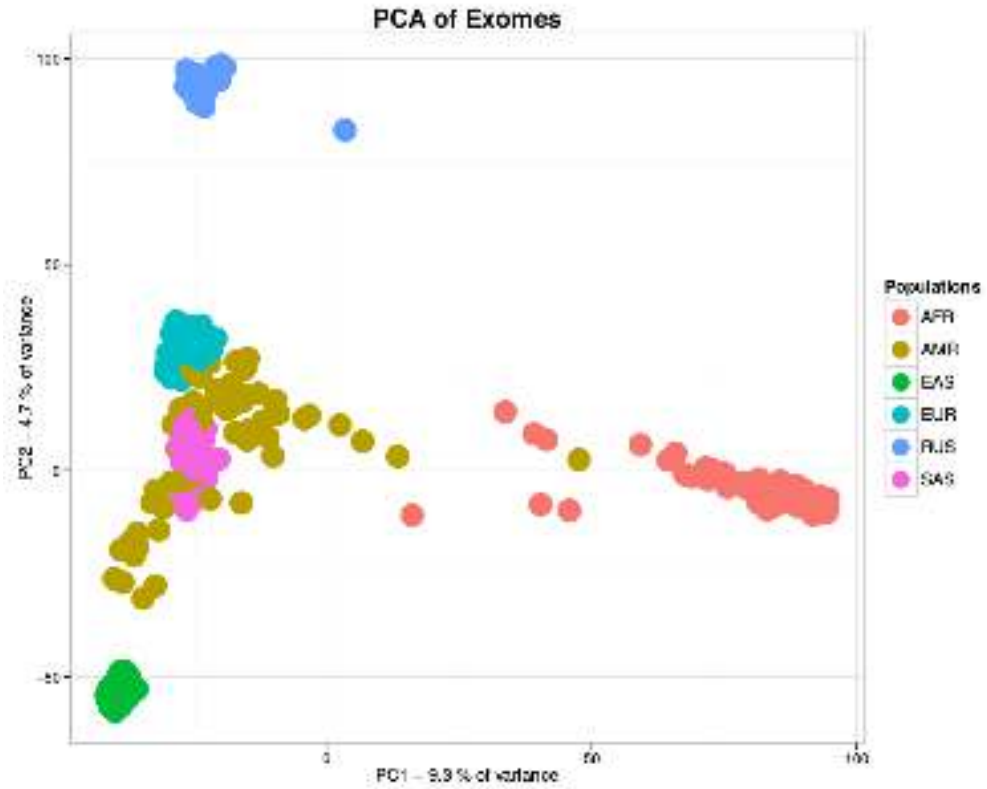
- Произвести контроль качества работы наборов Agilent.
- Написать удобный браузер вариантов.
- Написать несколько статей. :)

Спасибо за внимание!

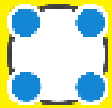


Бонус - особая русская духовность

Анализ главных компонент по данным генотипирования наших образцов и образцов из тысячи геномов.



Благодарности



- Ресурсному центру "БИОБАНК" СПбГУ и НИИ Акушерства и Гинекологии им. Д.О. Отта, а также лично Глотову А.С., Жуковой Е.А. и Полеву Д.Е. за предоставленные данные и обсуждение результатов
- Центру "Лаборатория алгоритмической биотехнологии" СПбГУ за доступ к вычислительным мощностям