ПОИСК НОВЫХ РЕГУЛЯТОРНЫХ SNPs В ГЕНОМЕ ЧЕЛОВЕКА И ОПРЕДЕЛЕНИЕ ИХ ФЕНОТИПИЧЕСКИХ ПРОЯВЛЕНИЙ

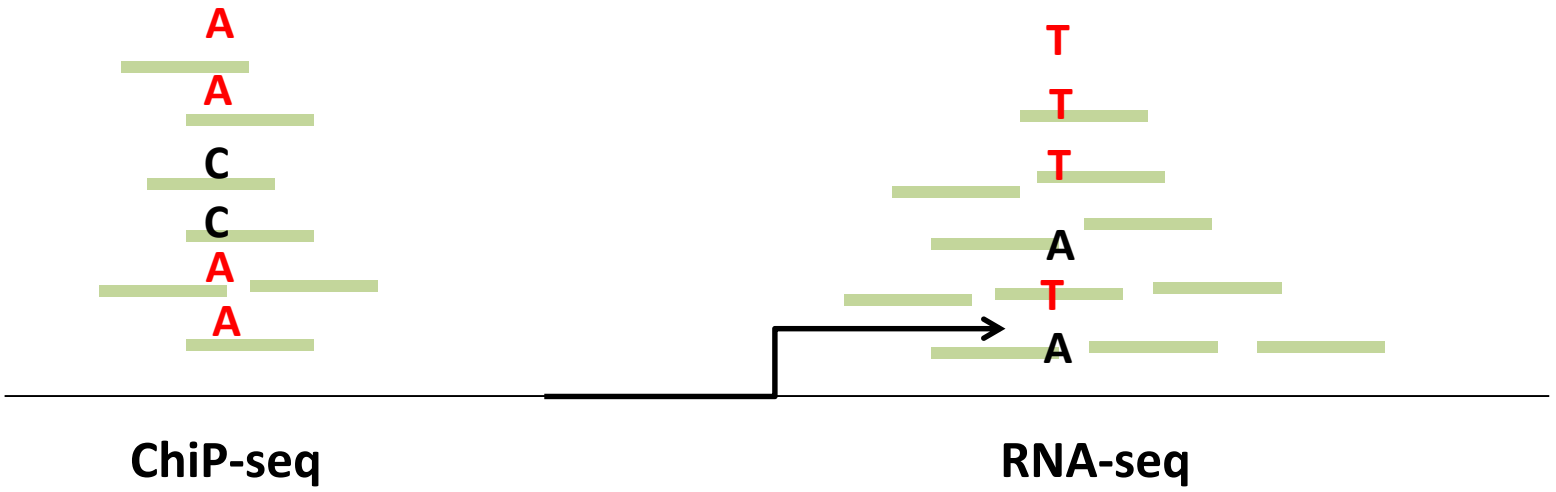
Выполнила: Устроханова Д. З., 2 курс магистратуры
Научный руководитель: к.б.н. Брызгалов Л. О.
Институт цитологии и генетики СО РАН
Лаборатория регуляции экспрессии генов

Новосибирск 2022

Анализ аллель-специфичных событий в NGS данных в гетерозиготных локусах (асимметрии представленности аллелей в данных ChIP-seq и RNA-seq)

Аллель-специфичное связывание (АСС)

Аллель-специфичная экспрессия (АСЭ)



rSNP - SNP, оказывающий влияние на

- функционирование регуляторного района;
- экспрессию гена с данным регуляторным районом

Цели и задачи

Цель работы: реализовать биоинформатический алгоритм для полногеномного поиска регуляторных SNPs на основании анализа аллель-специфичных событий.

Задачи:

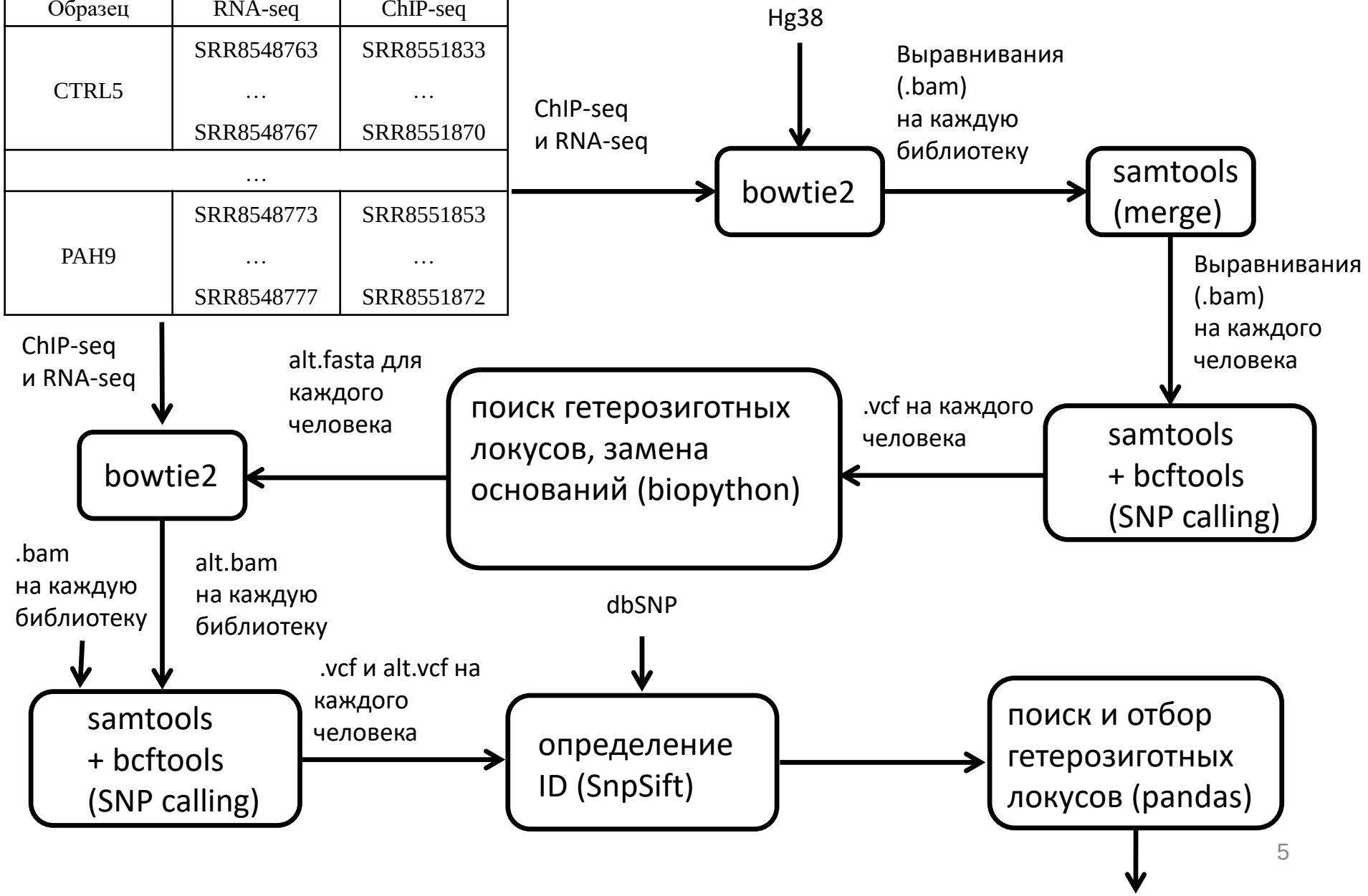
1. Осуществить поиск данных экспериментов ChIP-seq и RNA-seq, полученных для одних и тех же образцов и пригодных для анализа аллель-специфичных событий;
2. Определить события аллель-специфичного связывания и аллель-специфичной экспрессии в наборах данных ChIP-seq и RNA-seq;
3. Выбрать алгоритм машинного обучения и обучить модель для задачи классификации SNP (регуляторные/нерегуляторные);
4. С использованием обученной модели отобрать потенциально регуляторные SNPs;
5. Реализовать программу, автоматизирующую предыдущие этапы;
6. Проанализировать полученные в лаборатории данные экспериментов ChIP-seq и RNA-seq;
7. Для полученных rSNPs найти локусы количественного признака экспрессии (eQTLs), проанализировав независимый набор транскриптомных данных.

Данные

- NCBI GEO SRA (GSE126325), эндотелиальные клетки легочной артерии человека;
- семь человек;
- три библиотеки ChIP-seq на каждого человека (H3K27ac, H3K4me1, H3K4me3);
- пять библиотек RNA-seq на каждого человека;

Биоинформатический конвейер

Образец	RNA-seq	ChIP-seq
CTRL5	SRR8548763 … SRR8548767	SRR8551833 … SRR8551870
…		
PAH9	SRR8548773 … SRR8548777	SRR8551853 … SRR8551872



Биоинформатический конвейер

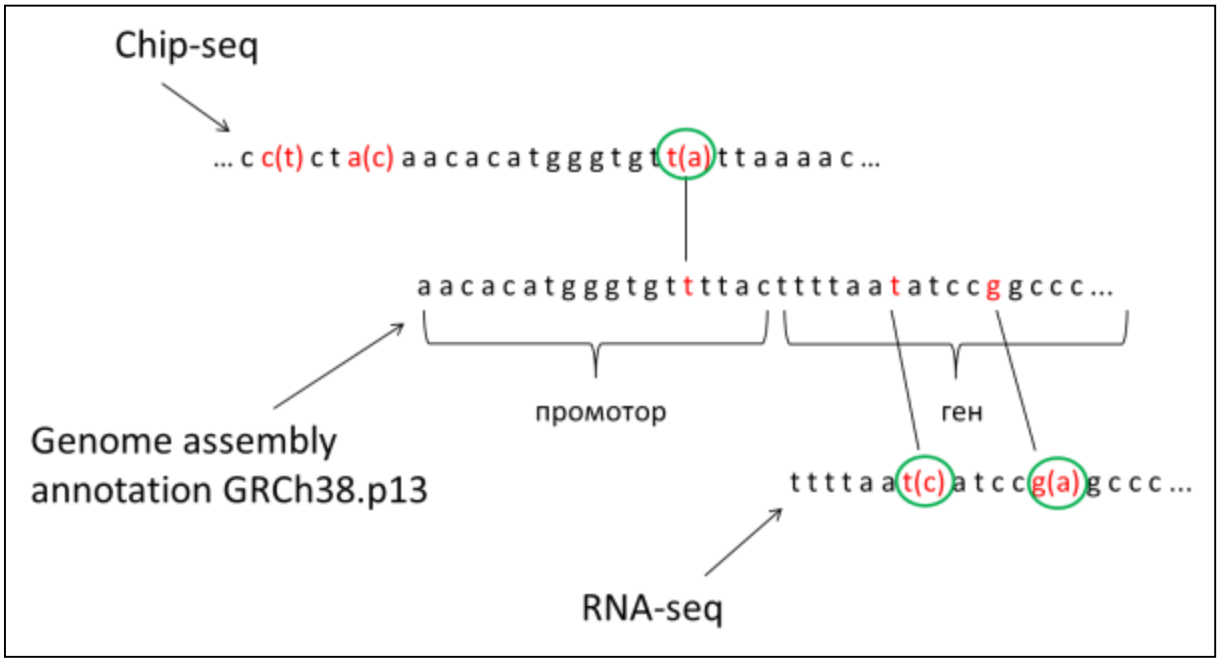
Genome assembly
annotation GRCh38.p13

пересечение гетерозиготных SNPs в ChIP-seq с регул. районами, отбор (pybedtools)

потенциальные rSNPs

пересечение гетерозиготных SNPs в RNA-seq с генами, отбор (pybedtools)

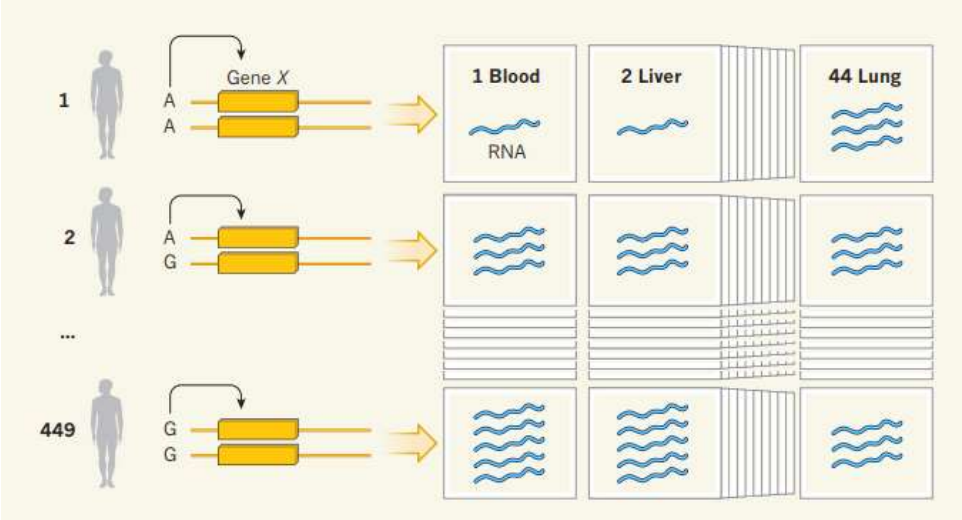
GTEx (v.08)
1,2 млн eQTL



SNPs	Признак 1	Признак 2	…	GTEx
rs1				0
rs2				1
…				…

Размеры выборки: 2375 SNPs

Genotype-Tissue Expression (GTEx)



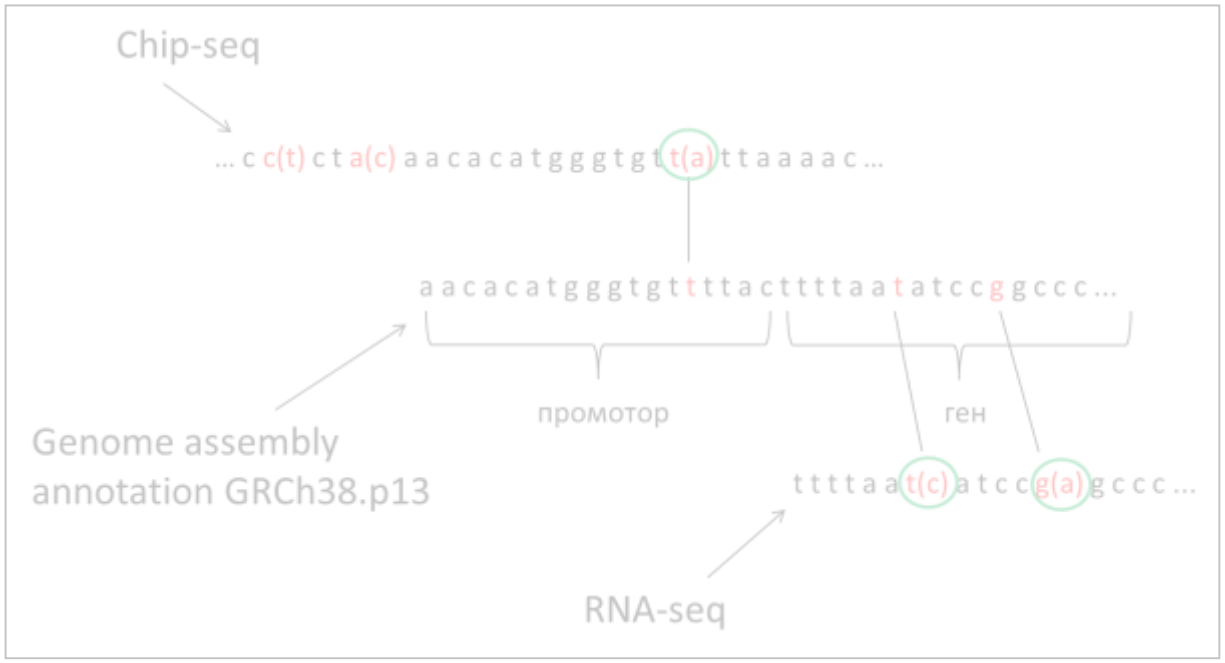
Genotype-Tissue Expression (GTEx) – это консорциум, предоставляющий общедоступный ресурс для изучения тканеспецифической экспрессии и регуляции генов.

Основная задача: идентификация ассоциации между уровнями экспрессии всех экспрессируемых генов (**eGenes**) и генетическими вариантами (**eVariants**).

eVariants принимаются за **eQTLs**

Локус количественного признака экспрессии (eQTL) - это геномный локус, от генотипа которого зависит уровень экспрессии гена-мишени.

Биоинформатический конвейер



Genome assembly
annotation GRCh38.p13

пересечение гет-ных SNPs в ChIP-seq с регул. районами, отбор (pybedtools, pandas)

потенциальные rSNPs

пересечение гет-ных SNPs в RNA-seq с генами, отбор (pybedtools, pandas)

GTEx (v.08)
1,2 млн eQTL

SNPs	Признак 1	Признак 2	…	GTEx
rs1				0
rs2				1
…				…

Размеры выборки: 2375 SNPs

Обучающие признаки

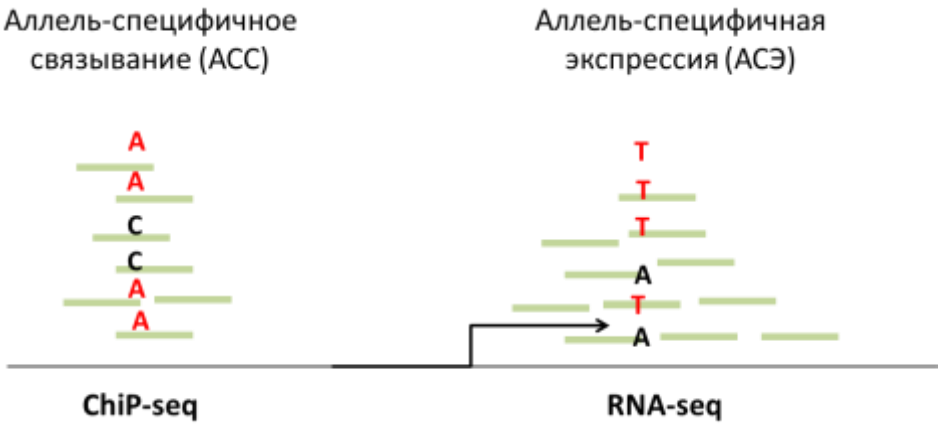
АСС (аллель-специфичное связывание)	$\left\lvert \log_{10} \frac{DPref(ChiP-seq)}{DPal(ChiP-seq)} \right\rvert$
АСЭ (аллель-специфичная экспрессия)	$\left\lvert \log_{10} \frac{DPref(RNA-seq)}{DPal(RNA-seq)} \right\rvert$
Соотношение АСС и АСЭ	$\left\lvert \log_{10} \frac{\left\lvert \log_{10} \frac{DPref(ChiP-seq)}{DPal(ChiP-seq)} \right\rvert}{\left\lvert \log_{10} \frac{DPref(RNA-seq)}{DPal(RNA-seq)} \right\rvert} \right\rvert$
Плотность rSNPs	$\frac{\text{Кол} - \text{во } SNPs \text{ на } POS \pm 400 \text{ п. о.}}{400 \text{ п. о.}}$
Возможные комбинации реф. и альт. аллелей*	Бинарный признак

** Для случайного леса – алгоритма, применяемого в работе далее*

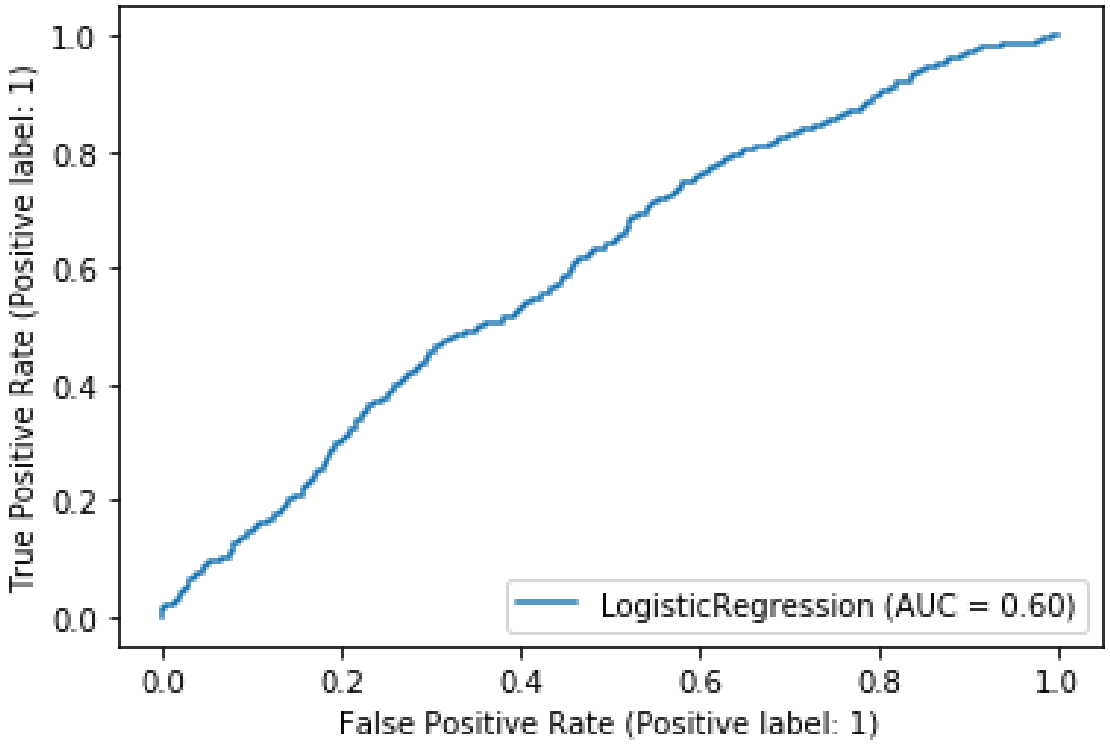
DP – покрытие аллеля в библиотеке

Логистический регрессионный анализ для прогнозирования вероятности того является ли rSNP eQTL по данным GTEx

	коэффициент	P>|z|
intercept	0.9363	
Плотность SNP	-0.0264	0.323
ACC (H3K4me3)	**0.2346**	**0.000**
ACC (H3K27ac)	-0.0516	0.346
ACC (H3K4me1)	0.1249	0.279
АСЭ	0.4603	0.194
Соотношение ACC и АСЭ (H3K4me3)	**-0.1923**	**0.000**
Соотношение ACC и АСЭ (H3K27ac)	-0.0031	0.916
Соотношение ACC и АСЭ (H3K4me1)	-0.0306	0.279

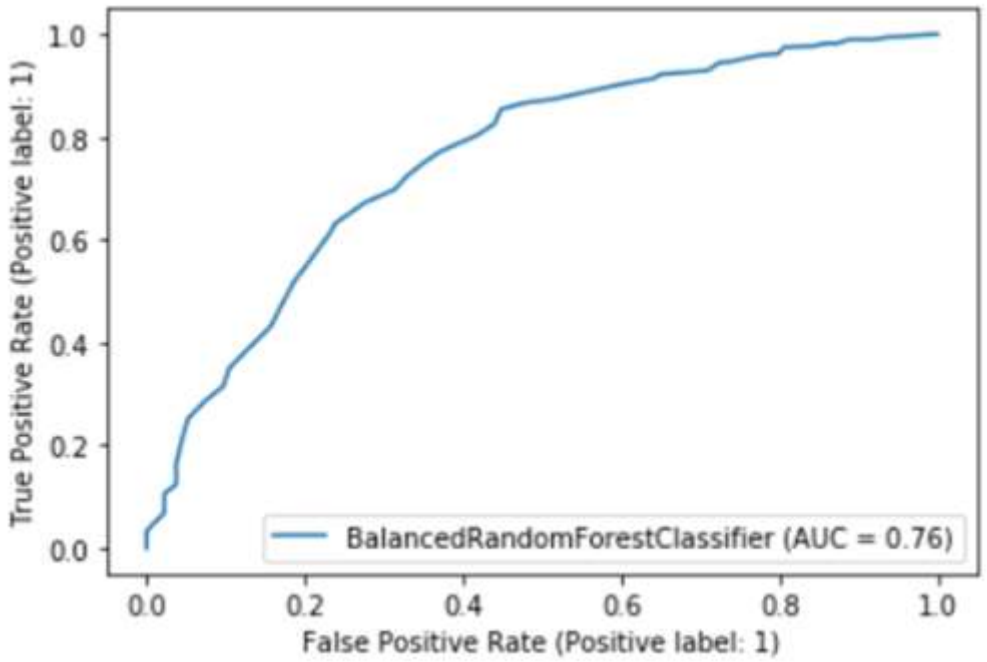


ROC кривая классификации методом логистической регрессии



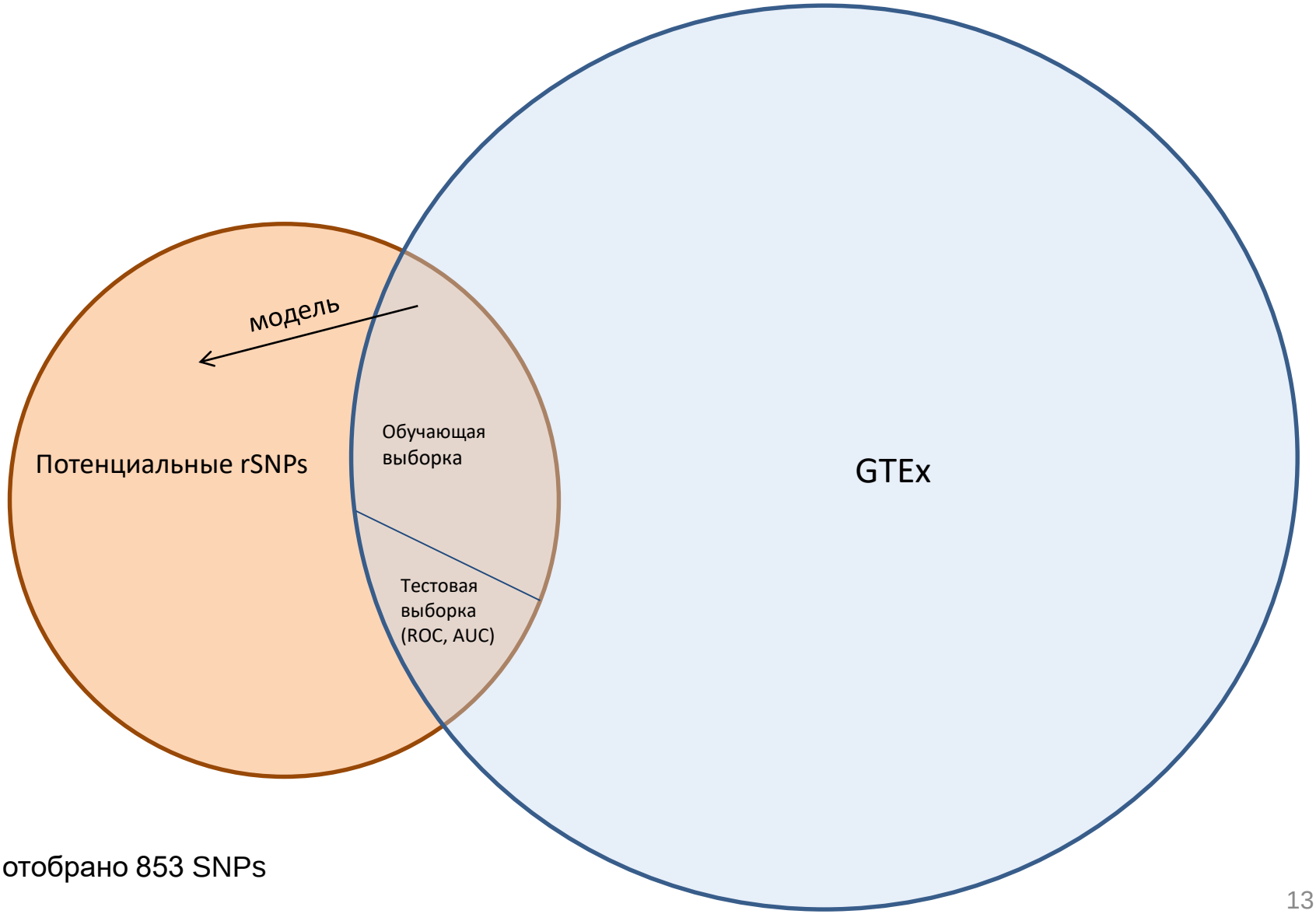
Логистическая регрессия является не лучшей моделью для данной классификации

ROC кривая классификации методом случайного леса



Параметры модели подобраны на скользящем контроле;
AUC 0,76 на тестовой выборке (на данных, которые модель «не видела при обучении»),

Обучение модели и применение ко всем отобранным rSNPs



отобрано 853 SNPs

Применение конвейера к данным полученным в лаборатории регуляции экспрессии генов ИЦиГ СО РАН

- Данные:
  - два человека;
  - моноциты крови;
  - три библиотеки ChIP-seq на каждого человека;
  - две библиотеки RNA-seq на каждого человека;

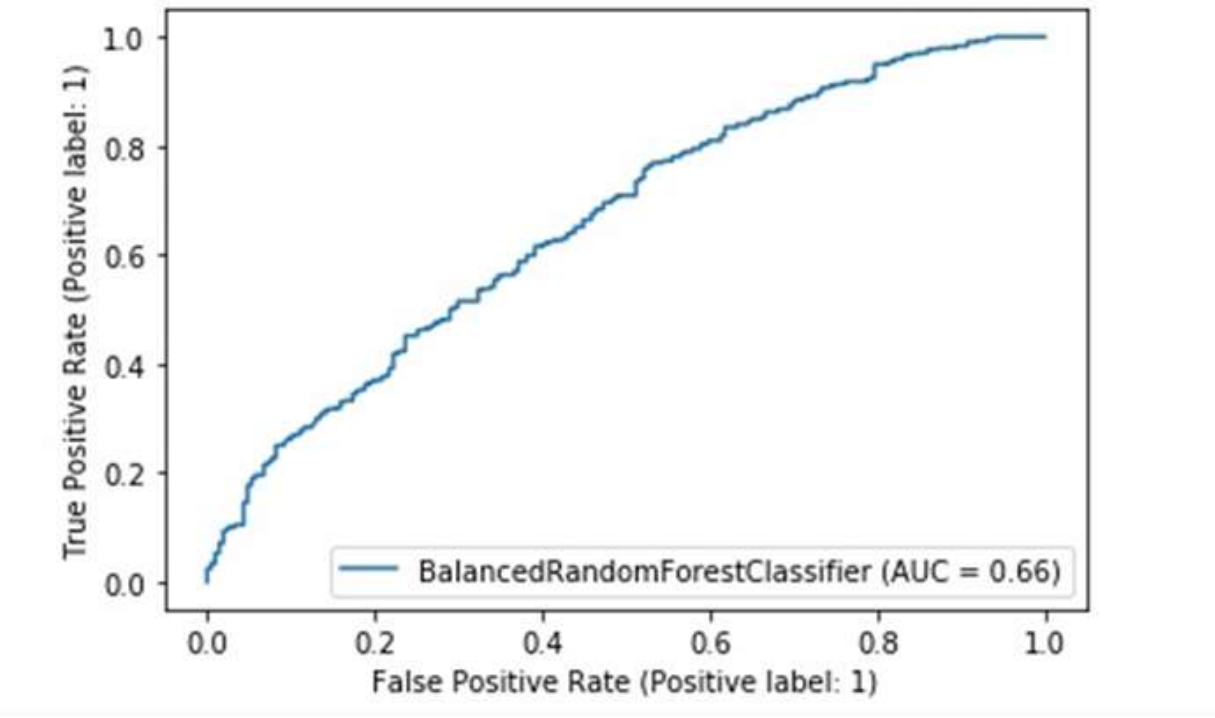
- Всего 2356 SNPs;

- Обучающая выборка была сформирована путем пересечения с данными GTEx;

- Размер обучающей выборки составил 1612 SNPs;

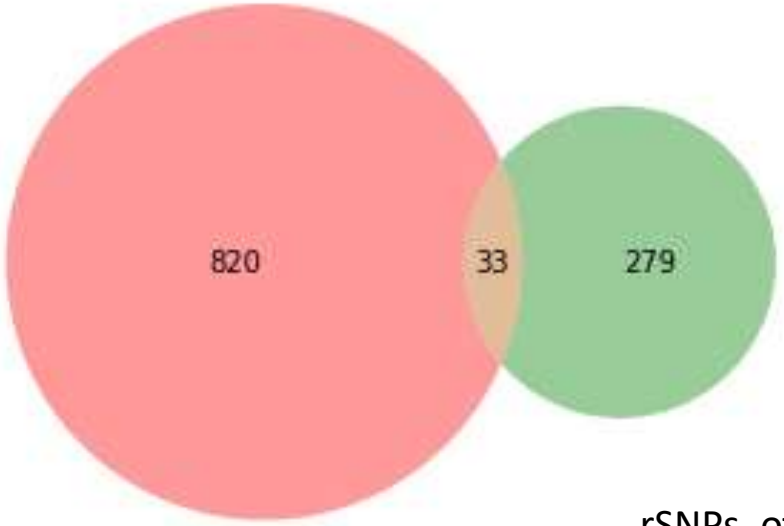
- Обучающие признаки формировались аналогично.

ROC кривая классификации методом случайного леса



Параметры модели подобраны на скользящем контроле;
AUC 0,66 на тестовой выборке,
отобрано 312 SNPs

Диаграмма Венна для отобранных rSNPs, по анализам данных нашей лаборатории и данных, полученных из архива NCBI



rSNPs, отобранные по анализам данных нашей лаборатории

rSNPs, отобранные по анализам данных, полученных из архива NCBI

Анализ независимого набора транскриптомных данных. Аннотация rSNPs

- Данные:
  - список из 312 rSNPs, отобранных на предыдущем этапе;
  - транскриптомные данных для тканей мозга 281 человека из NCBI;
  - данные секвенирования генома 2504 отдельных геномов, отобранных из пяти суперпопуляций из проекта 1000 Genomes Project

- Хотим:
  - разделить набор транскриптомных данных по аллельной комбинации каждого из отобранных 312 rSNPs на группы. Найти дифференциально экспрессирующиеся гены (ДЭГ) в этих группах.

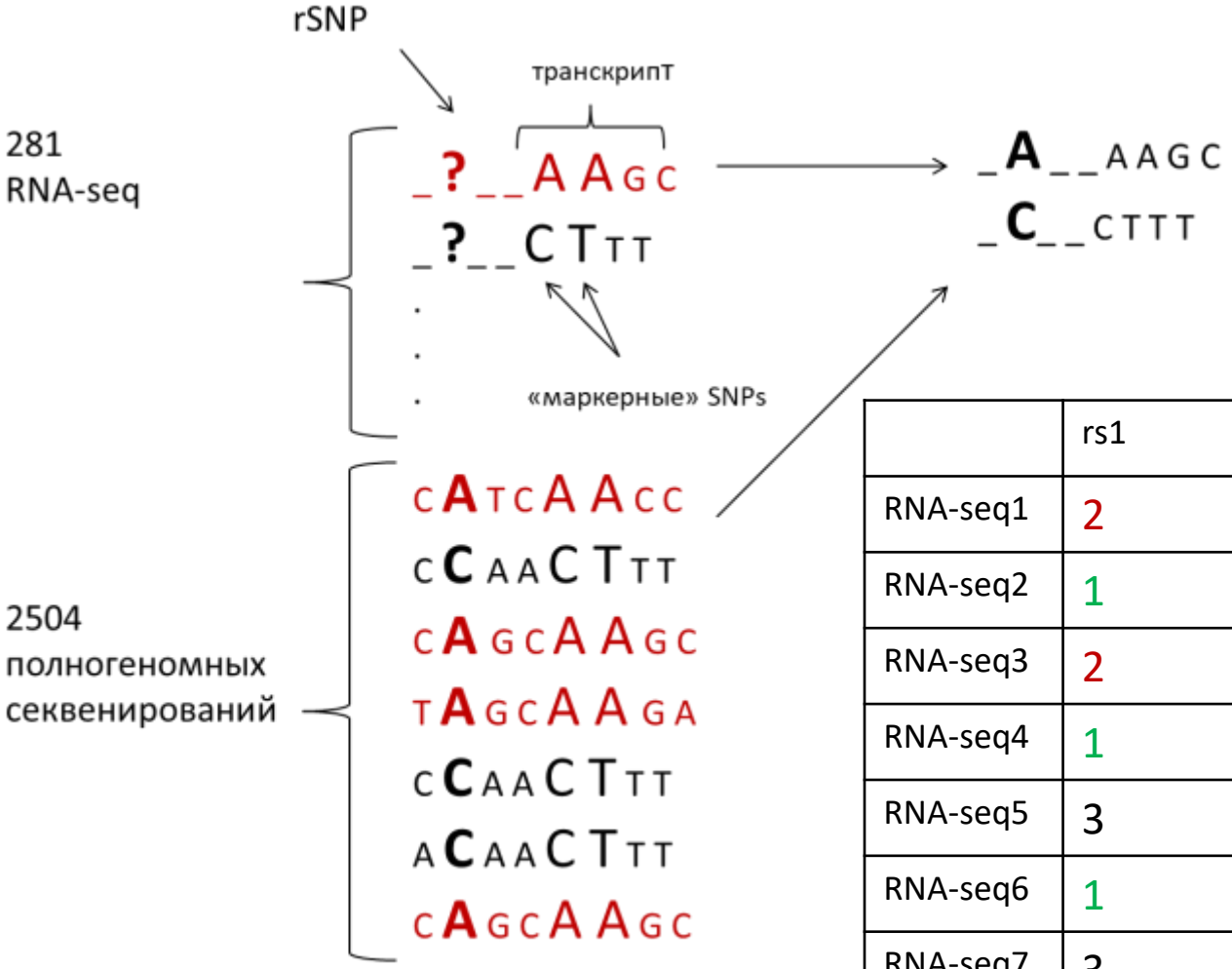
Анализ независимого набора транскриптомных данных. Аннотация rSNPs

- Проблема:
  - большинство идентифицированных rSNPs расположены в некодирующих геномных областях. Следовательно, нет возможности предсказать аллель по транскриптомным данным.

- Решение:
  - находим SNPs в кодирующих областях, связанные с отобранными rSNPs;
  - для определения аллельного варианта каждого rSNPs по данным RNA-seq используем оценку условной вероятности и концепцию расстояния Хэмминга

Postovalov S. et al. On the Relationship between Regulatory and Exomic DNA Markers // Proc. - 2020 Ural Symp. Biomed. Eng. Radioelectron. Inf. Technol. USBEREIT 2020. 2020. P. 97–100

Анализ независимого набора транскриптомных данных. Аннотация rSNPs



«1» – гомозигота по одному аллелю
«2» – гомозигота по другому аллелю
«3» – гетерозигота

	rs1	rs2	…	rs312
RNA-seq1	2	2		1
RNA-seq2	1	3		2
RNA-seq3	2	1		3
RNA-seq4	1	3		2
RNA-seq5	3	2		1
RNA-seq6	1	2		3
RNA-seq7	3	1		3
…				

группа ДЭГ

↓

Функциональная аннотация (KEGG, GO, DO)

Функциональная аннотация дифференциально экспрессирующихся генов (на примере rs16910241 и rs56119169)

Существуют общие ДЭГи для rs16910241 и rs56119169;

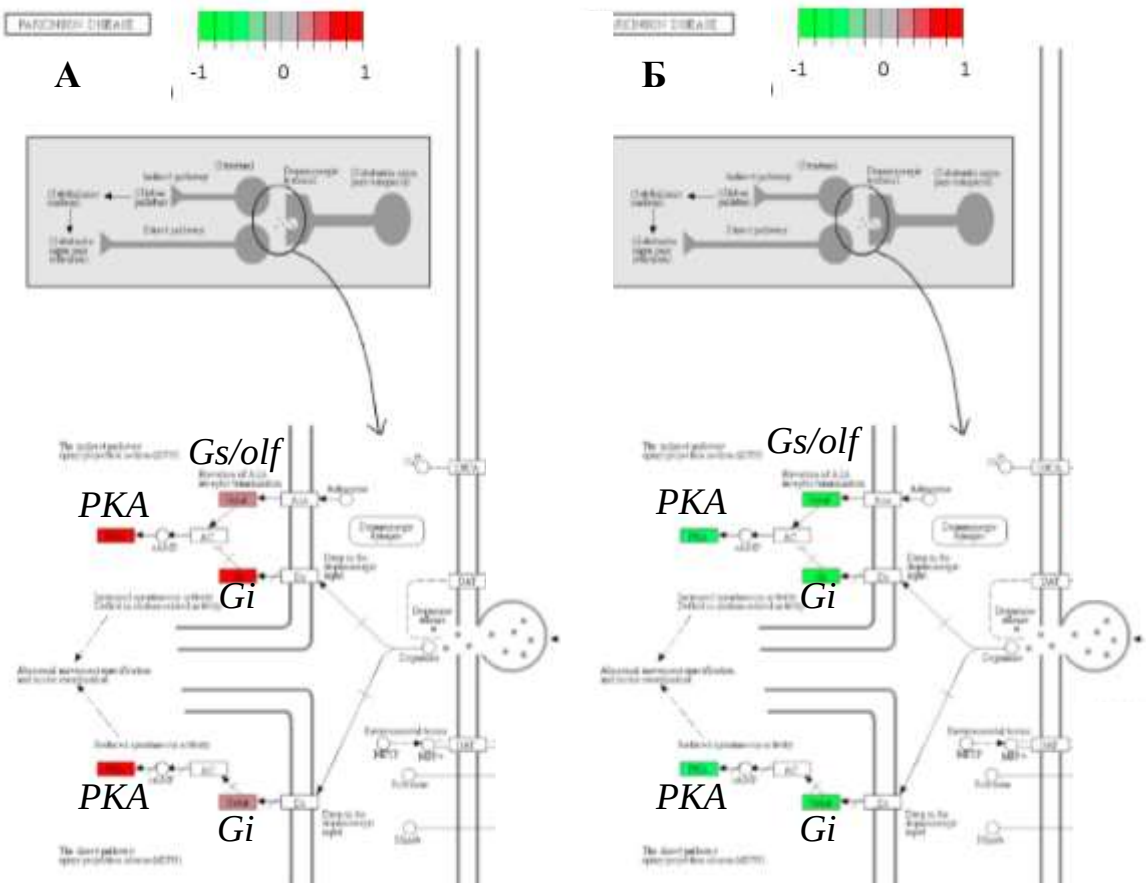
В терминах DO (Disease Ontology), KEGG:

- ассоциация с болезнью Паркинсона для обеих групп генов;

Продукты ДЭГов для rs16910241 и rs56119169, включенные в KEGG Pathway (Parkinson disease) в области пресинаптического окончания:

- субъединица альфа G-белка G(olf);
- субъединица альфа-1 G-белка G(i);
- каталитическая субъединица альфа цАМФ-зависимой протеинкиназы.

Направление изменения экспрессии генов в путях болезни Паркинсона



Пути болезни Паркинсона (Kyoto Encyclopedia of Genes and Genomes, KEGG)

Показана область пресинаптического окончания. Направление изменения экспрессии указано цветом: rs16910241 повышает экспрессию всех трех генов, тогда как rs56119169 снижает ее

Выводы

1. Реализован новый биоинформатический алгоритм для поиска функциональных вариантов rSNPs, в основе которого лежит анализ аллель-специфичных событий. В результате анализа данных (ChIP-seq и RNA-seq) было отобрано 1132 потенциальных rSNPs, в том числе 312 rSNPs из данных полученных в нашей лаборатории.

2. Показано, что одновременный анализ аллель-специфичных событий в данных ChIP-seq и RNA-seq позволяет эффективно выявлять регуляторные SNPs и значительно увеличивает эффективность обнаружения регуляторных вариантов.

3. По результатам анализа транскриптомных данных, среди 1132 потенциальных rSNPs, 721 rSNPs проявили себя как eQTLs (76 rSNPs из данных полученных в нашей лаборатории) — для них были найдены гены, дифференциально экспрессирующиеся у групп людей с разными генотипами по этим rSNPs.

4. Осуществлена функциональная аннотация обогащения групп дифференциально экспрессирующихся генов терминами GO, DO и KEGG.

5. Показано, что варианты, демонстрирующие регуляторный потенциал (влияние на изменение экспрессии генов) в одной ткани, могут демонстрировать его и в других тканях..

По результатам работы:

Korbolina, E. E., Bryzgalov, L. O., Ustrokhanova, D. Z., Postovalov, S. N., Poverin, D. V., Damarov, I. S., Merkulova, T. I. (2021). A panel of rsnps demonstrating allelic asymmetry in both chip-seq and rna-seq data and the search for their phenotypic outcomes through analysis of degs. International Journal of Molecular Sciences, 22(14).

Международная научная студенческая конференция - 2022, диплом 3 ст. (подсекция биоинформатика)

Данные NCBI GEO SRA (GSE126325), эндотелиальные клетки легочной артерии человека

Образец	RNA-seq	ChiP-seq		
		H3K27ac	H3K4me1	H3K4me3
CTRL5	SRR8548763 SRR8548764 SRR8548765 SRR8548766 SRR8548767	SRR8551833	SRR8551851	SRR8551870
CTRL8	SRR8548778 SRR8548779 SRR8548780 SRR8548781 SRR8548782	SRR8551836	SRR8551854	SRR8551873
PAH2	SRR8548793 SRR8548794 SRR8548795 SRR8548796 SRR8548797	SRR8551839	SRR8551858	SRR8551877
PAH3	SRR8548798 SRR8548799 SRR8548800 SRR8548801 SRR8548802	SRR8551840	SRR8551859	SRR8551878
PAH6	SRR8548813 SRR8548814 SRR8548815 SRR8548816 SRR8548817	SRR8551843	SRR8551862	SRR8551881
PAH8	SRR8548818 SRR8548819 SRR8548820 SRR8548821 SRR8548822	SRR8551845	SRR8551864	SRR8551883

$$\frac{\sum_{i=1}^{N} \left| R_{i,rSNP} - R_{i,eSNP} \right|}{\sum_{i=1}^{N} \left(R_{i,rSNP} + R_{i,eSNP} \right)}$$

где Rij = 0, 1 или 2, и это количество минорных аллелей в j-й позиции i-го человека из данных 1000Genomes (N = 2504).

Чем меньше расстояние Хэмминга, тем более сильная связь наблюдается между двумя маркерами SNPs. NHD равен нулю, если во всех случаях (у всех людей) минорные аллели в одной позиции соответствуют редким аллелям в другой позиции.

Оценка условной вероятности по таблице сопряженности

Предположим, что в RNA-Seq получена аллель "A" для полиморфизма rs738904. Какова вероятность, что у этого же человека хотя бы в одной копии хромосомы 22 полиморфизм rs7289432 имеет аллель "G"?

Обозначим для полиморфизма rs738904 через A* генотип человека, в котором один аллель равен A.

Аналогично обозначим для полиморфизма rs7289432 через G* генотип человека, в котором один аллель равен G.

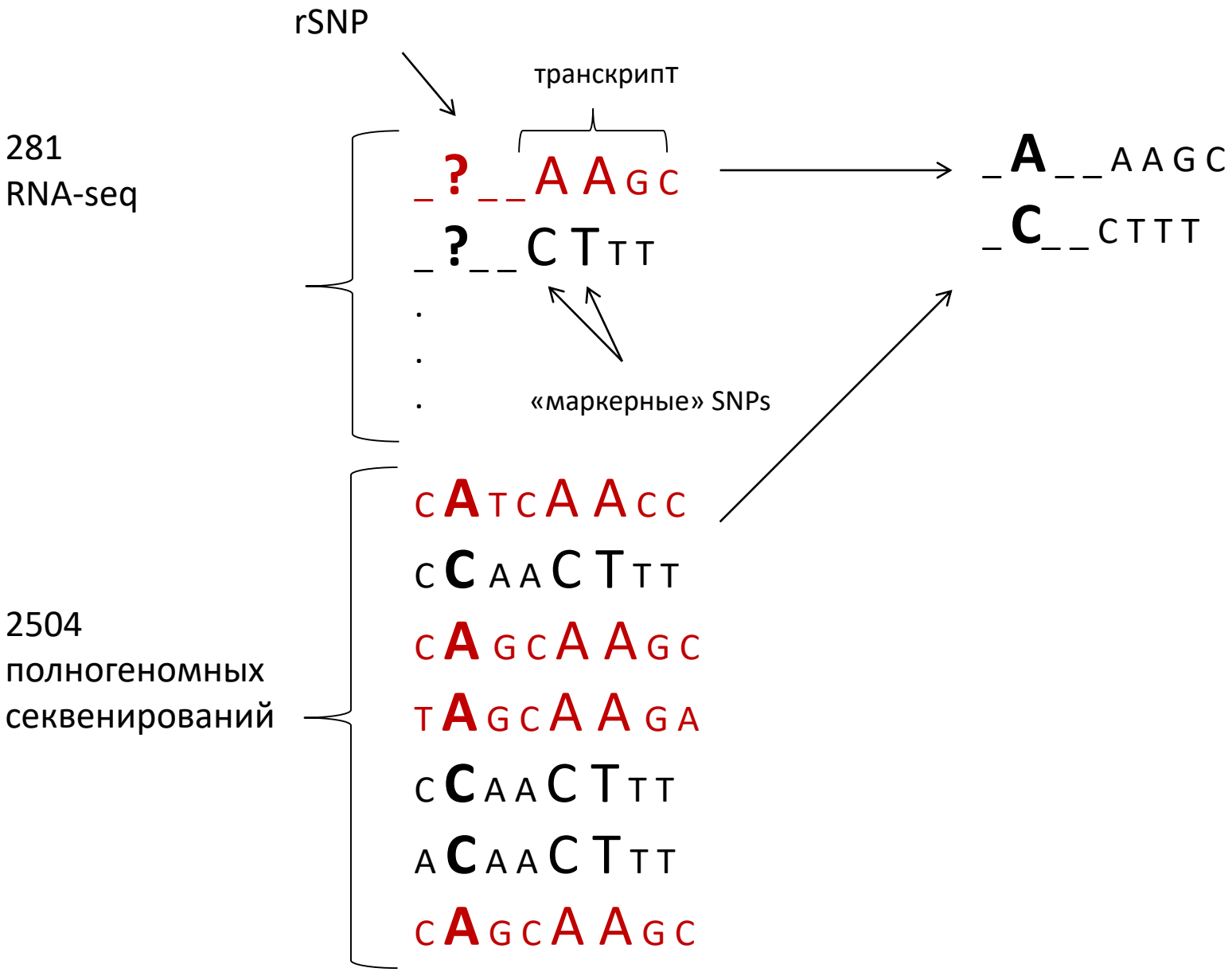
Тогда по формуле условной вероятности можно найти
P{"G*"|"A*"}= P{"G*","A*"}/P{A*}.

P{"G*","A*"} = (1042+5+7+352)/2504=1406/2504

P{"A*"} = (1054+357)/2504=1411/2504
Тогда оценка P{"G*"|"A*"} = 1406/1411 = 0,996.

rSNP (rs7289432)	cSNP (rs738904)			Total
	CC	AC	AA	
AA	1084	5	0	1089
AG	9	1042	5	1056
GG	0	7	352	359
Total	1093	1054	357	2504

Для предсказания присутствия регуляторных полиморфизмов берем условную вероятность не менее 0.9. Из предсказанных регуляторных полиморфизмов берем тот, у которого HPX меньше.


rSNP
транскрипт
281
RNA-seq
_ ? _ _ A A G C
_ ? _ _ C T T T
.
.
.
«маркерные» SNPs
_ A _ _ A A G C
_ C _ _ C T T T
2504
полногеномных
секвенирований
C A T C A A C C
C C A A C T T T
C A G C A A G C
T A G C A A G A
C C A A C T T T
A C A A C T T T
C A G C A A G C