Кафедра математической теории интеллектуальных систем

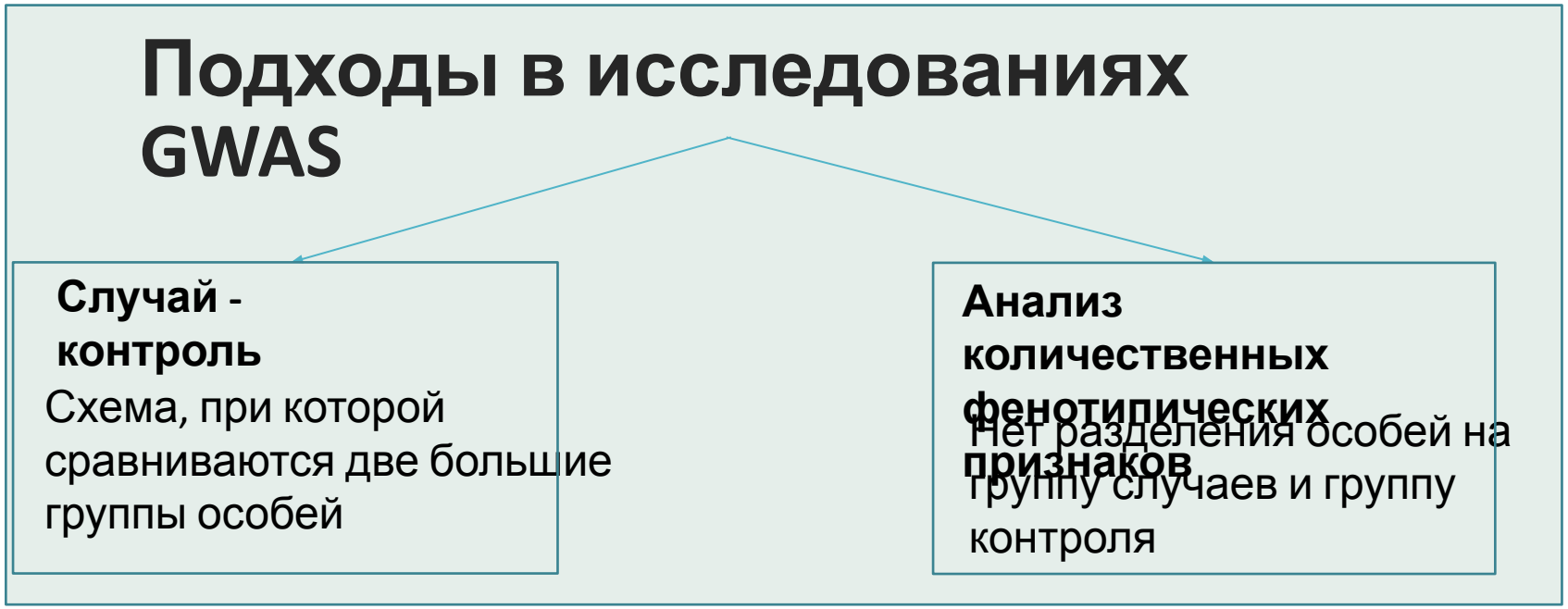
Полногеномный поиск ассоциаций для множества фенотипических признаков

Выполнила студентка
607 группы
Петрова Александра Алексеевна

Научный руководитель
Старший научный сотрудник
Галатенко Алексей Владимирович

Полногеномный поиск ассоциаций

Полногеномный поиск ассоциаций (GWAS, Genome-Wide Association Studies) – направление биологических исследований, связанных с изучением ассоциаций между геномными вариантами и фенотипическими признаками.

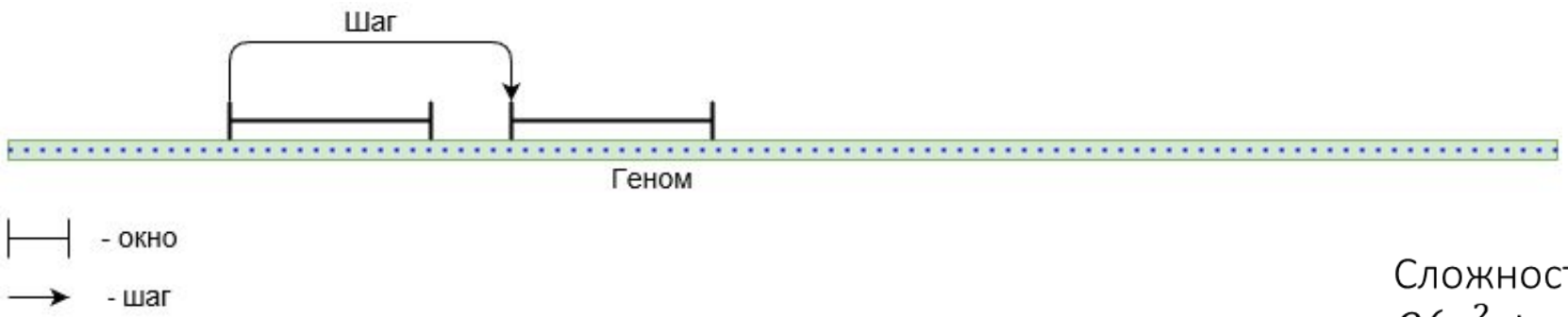


SNP - отличия последовательности ДНК размером в один нуклеотид (A, T, G или C) в геноме представителей одного вида или между гомологичными участками гомологичных хромосом.

Алгоритм mtSet программного комплекса Limix

Используется следующая статистическая модель:

$$Y = FB + U_r + U_g + \Psi$$



Сложность для одного окна составляет

$$O(n^2 + nl^2m^2 + nlm^4) + O(n^3)$$

Y – матрица фенотипа $N \times P$,
N – количество индивидов в выборке,
P – количество фенотипических признаков

F – матрица фиксированных эффектов

B – матрица весов

U_r – матрица SNP

U_g – матрица полигенных мутаций

Ψ – матрица остаточного шума

m – количество фенотипических признаков

n – количество особей

l – количество аллелей

Цель работы

Цель данной работы заключается в том, чтобы разработать алгоритм, который, возможно, с небольшой потерей точности, мог бы производить полногеномный поиск ассоциаций множественных фенотипических признаков быстрее, чем алгоритм mtSet.

Для достижения поставленной цели в работе решаются следующие задачи:

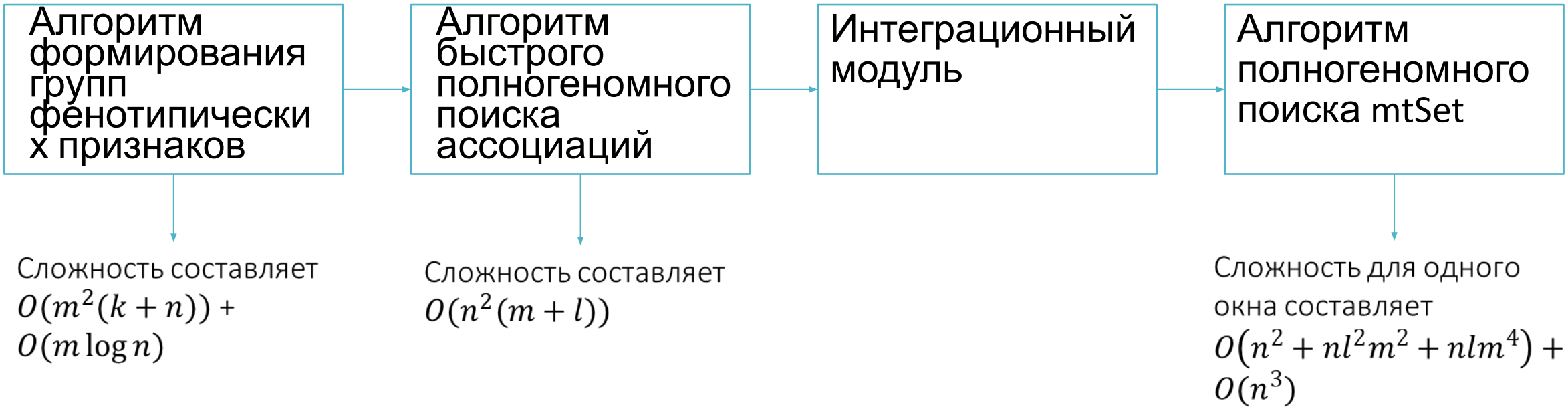
1. Разработка алгоритма формирования групп фенотипических признаков.

2. Разработка алгоритма быстрого полногеномного поиска ассоциаций для группы фенотипических признаков.

3. Реализация интеграционного модуля между алгоритмом быстрого полногеномного поиска ассоциаций и алгоритмом

4. Выполнение тестирования на данных фенотипа и генотипа растения Arabidopsis thaliana

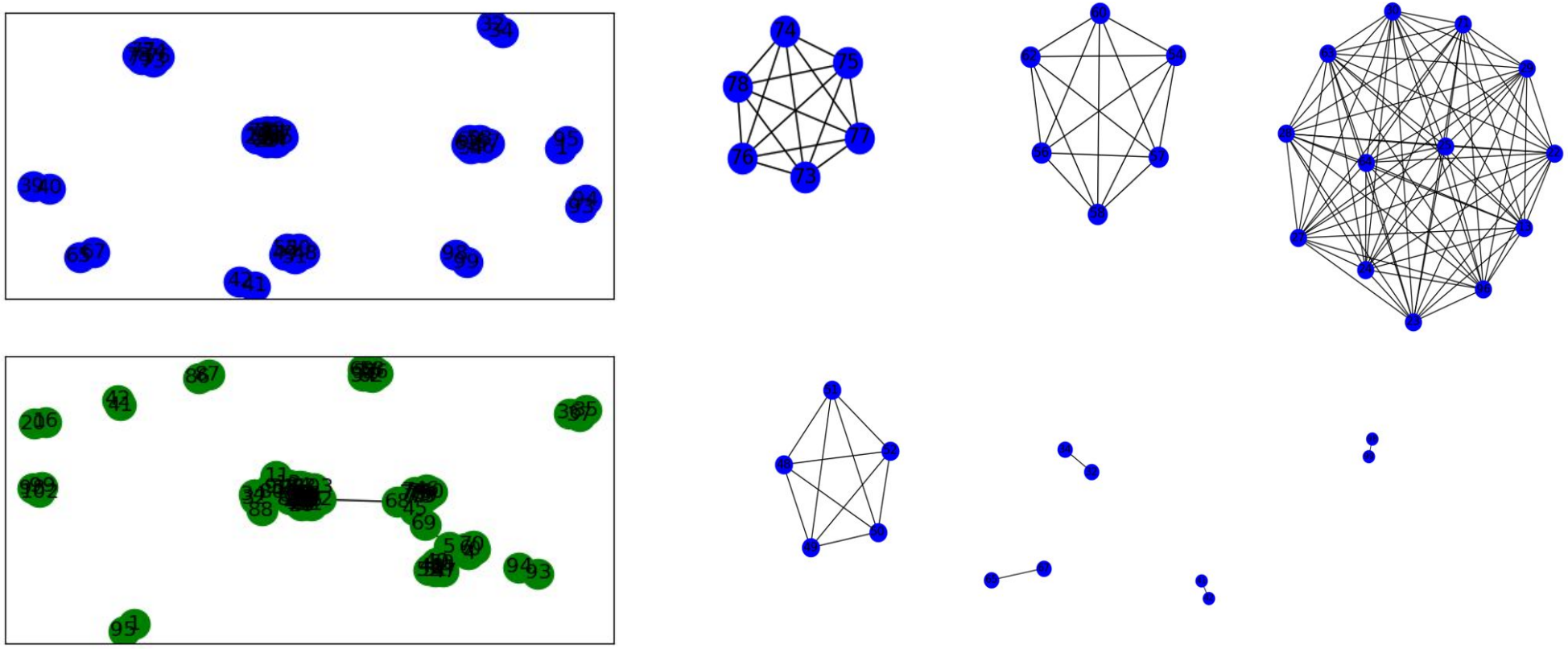
Разработанный алгоритм



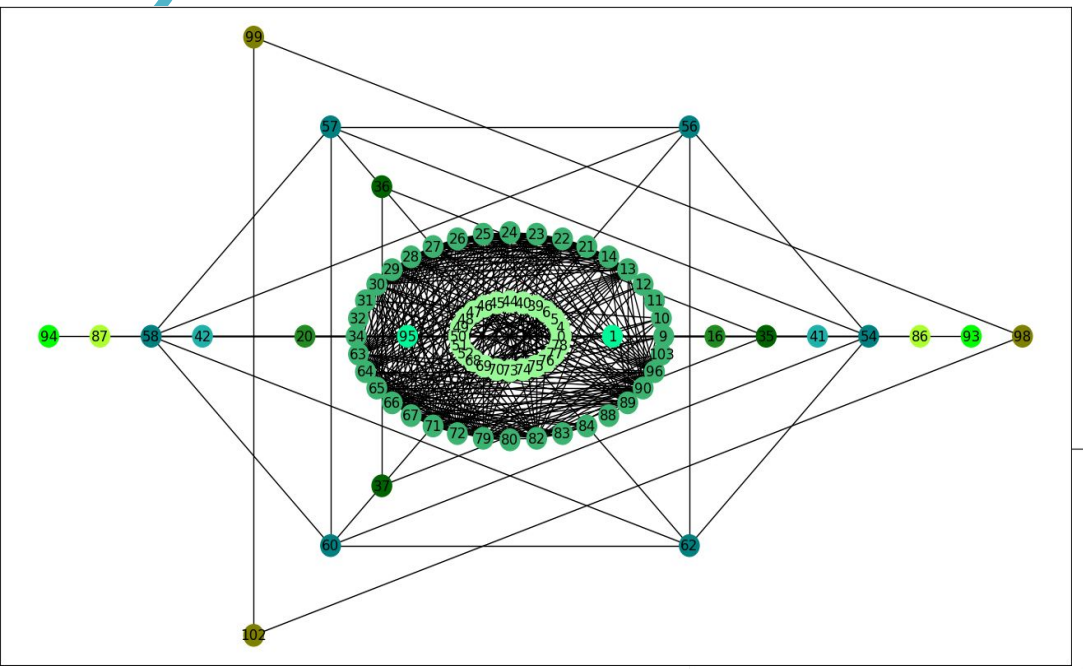
k – количество связей между признаками
m – количество фенотипических признаков
n – количество особей
l – количество нуклеотидов

Алгоритм формирования групп фенотипических признаков

Алгоритм формирования групп фенотипических признаков часть 1

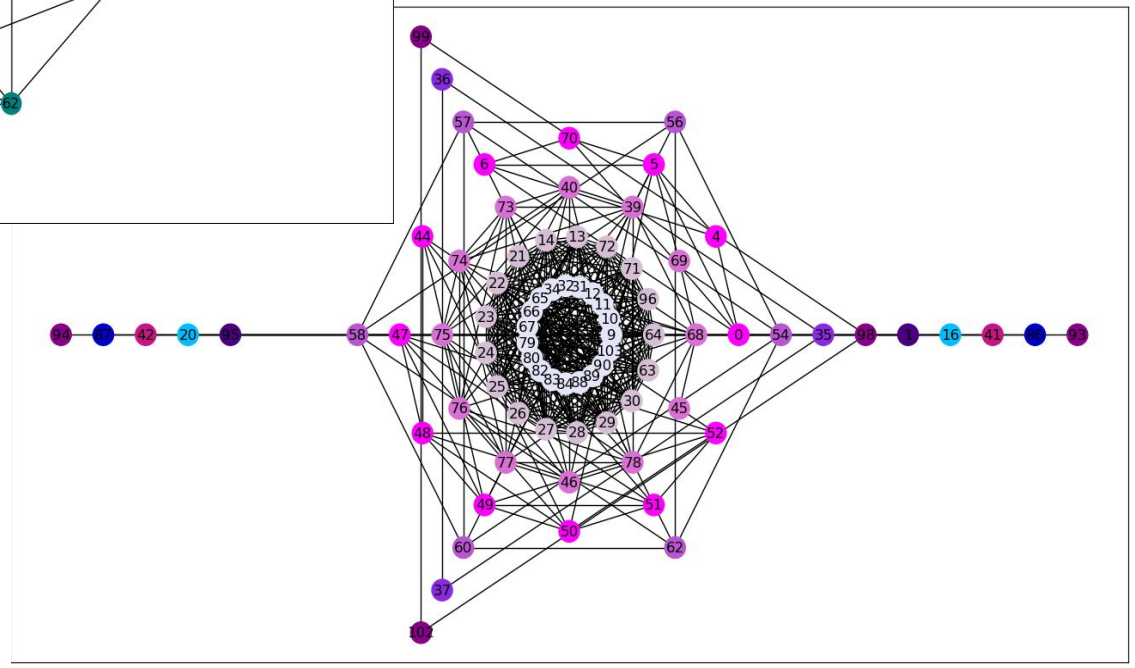


Алгоритм формирования групп фенотипических признаков часть 2



Результат применения алгоритма Гирвен-Ньюмен к зеленому графу

Было выделено 10 групп признаков



Результат применения алгоритма Гирвен-Ньюмен к зеленому графу

Было выделено 12 групп признаков

Алгоритм быстрого полногеномного поиска ассоциаций часть 1

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{m1} \\ y_{12} & & y_{m2} \\ \vdots & \ddots & \vdots \\ y_{1n} & \cdots & y_{mn} \end{pmatrix}$$

Y – матрица фенотипа

$$d_{ij} = \sqrt{\sum_{m} (y_i^m - y_j^m)^2}$$

Расстояние между фенотипами особи i и особи j

m – количество признаков

n – количество особей

$$W = \begin{pmatrix} w_1 \\ w_2 \\ \cdots \\ w_n \end{pmatrix}$$

W – вектор весов

$$w_i = \left(\prod_{j=1}^{n} \frac{d_{ij}}{p_{ij}}\right)^{1/n}$$

$$\begin{cases} p_{ij} = 0, r_{ij} = d_{ij} \\ d_{ij} = 0, r_{ij} = 1 \end{cases}$$

Показатель вклада i-го SNP в данную группу фенотипических признаков

$$X = \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ x_{12} & & x_{n2} \\ \vdots & \ddots & \vdots \\ x_{1l} & \cdots & x_{nl} \end{pmatrix}$$

X – матрица SNP

$$\rho_{ij} = \sum_{k=1}^{l} |x_{ik} - x_{jk}|$$

Расстояние Хэмминга для SNP особи i и особи j

l – количество аллелей

Алгоритм быстрого полногеномного поиска ассоциаций часть 2

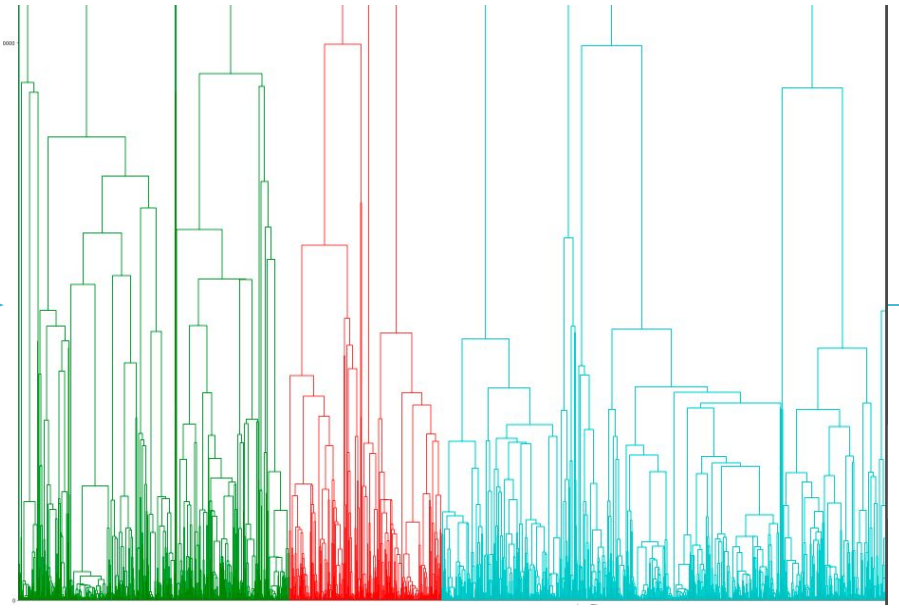
X_{max} – вектор аллелей с максимальным весом

X_{max-1} –вектор аллелей с ближайшим к максимальному весом

X_{min} – вектор аллелей с минимальным весом

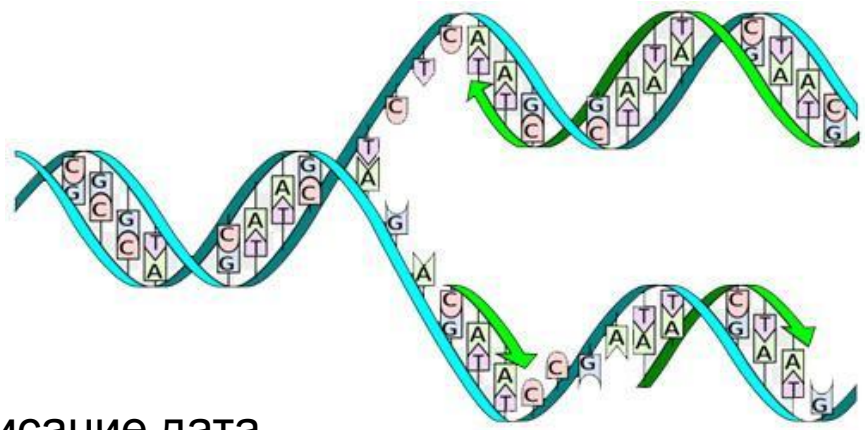
$S = S_1 \cap S_2 = (X_{min} \Delta X_{max}) \cap (X_{min} \Delta X_{max-1})$ - набор координат аллелей

Алгоритм иерархической кластеризации методом ближайшего соседа



Получили набор отрезков генома, которые затем направляются на вход алгоритма mtSet

Тестирование

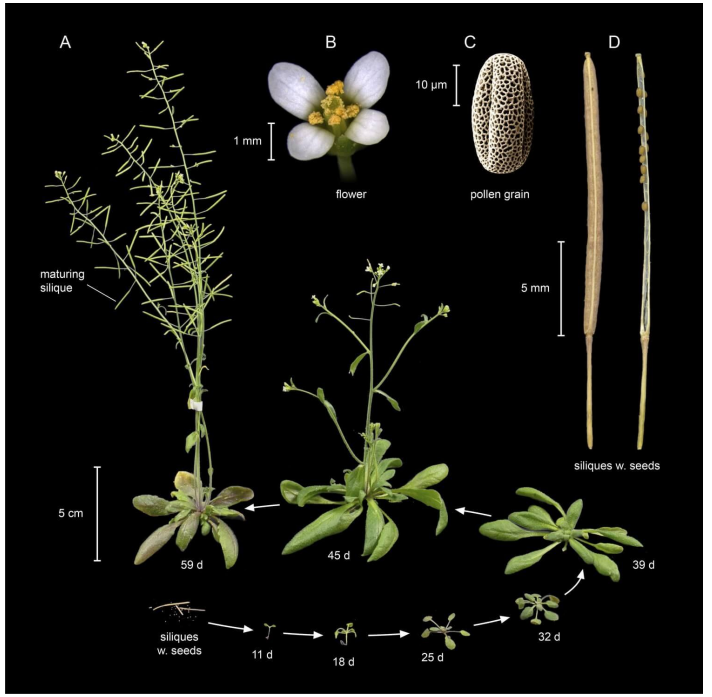


Описание дата-

```
[ 2. 2. 2. 0. 0. 2. 2. 0. 0. 0. 0. 2. 0. 2. 0. 2. 0. 0. 2. 2. 0. 2. 2. 0.
 2. 0. 0. 0. 2. 2. 2. 0. 0. 2. 0. 0. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.
 2. 2. 2. 2. 0. 2. 0. 0. 2. 2. 2. 2. 2. 2. 0. 2. 2. 2. 0. 2. 0. 2. 2. 2.
 2. 2. 2. 0. 2. 2. 2. 2. 2. 2. 0. 0. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.
 2. 2. 0. 2. 2. 2. 2. 0. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 0. 0. 0. 2. 2. 2.
 2. 2. 2. 2. 2. 2. 2. 0. 2. 2. 0. 2. 2. 2. 2. 2. 2. 0. 2. 2. 2. 2. 2. 2. 2. 2.
 2. 0. 0. 2. 2. 0. 2. 2. 2. 2. 2. 2. 2. 0. 2. 2. 2. 0. 0. 2. 2. 2. 2. 2.
 2. 2. 0. 0. 2. 2. 2. 2. 0. 2. 2. 0. 2. 2. 2. 2. 2. 0. 2. 2. 2. 2. 2. 2. 2. 2.
 0. 2. 2. 0. 2. 2. 0.]
```

1	snp1	0	657	T	C
1	snp2	0	3102	G	A
1	snp3	0	4648	A	C
1	snp4	0	4880	T	C
1	snp5	0	5975	G	T
1	snp6	0	6063	T	C
1	snp7	0	6449	C	T

00 Homozygous for first allele in .bim file
01 Missing genotype
10 Heterozygous
11 Homozygous for second allele in .bim file



Была взята группа из следующих Фенотипических признаков:

0W
2W
4W

Различные фенотипические признаки, связанные с количеством дней до цветения

*Atwell, S., Huang, Y., Vilhjálmsson, B. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. // *Nature* **465,** 627–631 (2010). https://doi.org/10.1038/nature08800

Результаты тестирования

mtSet

chrom	start	end	nsnps	pvalue
1	23027657	23029657	4	1.407491340158984e-05
1	23026657	23028657	4	1.4074913378844082e-05
2	-	-	-	-
3	4840394	4842394	5	6.018123991825176e-05
3	6699394	6701394	8	5.379841202833628e-05
3	3087394	3089394	5	4.469149961092417e-06
3	22927394	22929394	4	3.952481315105589e-05
5	17631168	17633168	5	9.442412753814328e-05
5	17632168	17634168	10	8.844718885135648e-05
5	6260168	6262168	12	8.089592572482654e-05
5	6252168	6254168	9	4.9363925309055216e-05
5	6261168	6263168	14	3.128437670747888e-05
5	6251168	6253168	6	2.9857721790735625e-05
5	6259168	6261168	6	2.7911367522644893e-05
5	19533168	19535168	10	1.4342158951067126e-05

Разработанный алгоритм

chrom	start	end	nsnps	pvalue
1	21600657	21602657	7	2.882718539223049e-05
1	23026657	23028657	4	2.062649559325109e-06
1	23027657	23029657	4	2.062649514997124e-06
2	2187651	2189651	4	3.2434067167313365e-07
3	4840394	4842394	5	1.0549827625388847e-05
3	6699394	6701394	8	9.129301954821966e-06
3	3087394	3089394	5	3.639910828428796e-07
3	4839394	4841394	5	7.534514265526442e-05
3	5687394	5689394	9	4.121299191161856e-05
5	6260168	6262168	12	5.524710499344688e-05
5	6252168	6254168	9	3.2285581621374096e-05
5	6251168	6253168	6	1.8679042454198182e-05
5	6261168	6263168	14	1.8039599838108703e-05
5	6259168	6261168	6	1.735714867806214e-05
5	14700168	14702168	4	9.144756095729539e-05

Дальнейшие шаги

1. Скорректировать кластеризацию координат аллелей для получения более точных отрезков генома для их дальнейшей подачи в mtSet
2. Выполнить сравнение полученных участков с результатами, опубликованными в Nature
3. Произвести анализ полученных участков генома на наличие в них каких-то генов

Список литературы

1. Casale, F., Rakitsch, B., Lippert, C. et al. Efficient set tests for the genetic analysis of correlated traits. // Nat Methods 12, 755–758 (2015). https://doi.org/10.1038/nmeth.3439
2. Уткин Л.В., Жук Ю.А. Полногеномный поиск ассоциаций с использованием матриц парных сравнений // Труды СПИИРАН. 2016. Вып. 47. С. 225-240.
3. Atwell, S., Huang, Y., Vilhjálmsson, B. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. // *Nature* **465,** 627–631 (2010). https://doi.org/10.1038/nature08800

Спасибо за внимание