Математическая биология и биоинформатика
2024. Т. 19. № 1. С. 36–51. doi: 10.17537/2024.19.36

========== **БИОИНФОРМАТИКА** ==========

GENIS – методологический подход для генотипирования *in silico* (апробация на результатах секвенирования для *Sus scrofa*)

Кипень В.Н., Снытков Е.В.

Государственное научное учреждение «Институт генетики и цитологии» Национальной академии наук Беларуси, Минск, Беларусь

Аннотация. Разработан универсальный методологический подход, который позволяет решать задачу дифференциации близкородственных видов по необработанным данным секвенирования NGS. Метод основан на использовании однонуклеотидных полиморфизмов (SNP). Данный подход универсален, его можно использовать при биоинформатическом анализе любых файлов с результатами секвенирования независимо от исследуемого биологического вида. Разработанный нами подход основан на автоматизации процесса поиска нуклеотидных последовательностей, фланкирующих искомый аллель. Поиск может проводиться на персональном компьютере исследователя, язык программирования Python v.3.10 и среда разработки программного обеспечения Jupyter Notebook бесплатны и общедоступны. Методологический подход для генотипирования *in silico* реализован в виде программы GENIS. В рамках данной работы проведена апробация программы на файлах с результатами секвенирования геномов животных рода *Sus*, выявлены полиморфизмы для дифференциации свиней породы дюрок.

Ключевые слова: *однонуклеотидный полиморфизм, генотипирование in silico, Sus scrofa, Python v.3.10, Jupyter Notebook.*

ВВЕДЕНИЕ

На данный момент в международных научных базах данных содержится значительное количество геномов животных, полученных с использованием технологии NGS (Next-Generation Sequencing). Компания Illumina (США) занимает лидирующее место – на использование ее технологии проходится более 80 % всей получаемой информации по секвенированию. Также с использованием технологии компании Pacific Biosciences (США) имеется возможность получать длинные прочтения (более 10 тыс. нуклеотидов), что позволяет нивелировать недостатки предыдущей технологии. В целом, необработанных (т.е. невыровненных) данных секвенирования, имеющих средний уровень покрытия на нуклеотид больше 5, представлено значительное количество: для рода *Sus* – более 1000 шт., для рода *Canis* – более 800, для рода *Bos* – более 2000, для рода *Hypophthalmichthys* – около 50. Необработанные данные секвенирования представлены в формате SRA (Sequence Read Archive), содержащем миллионы коротких прочтений нуклеотидных последовательностей.

Решение задачи по дифференциации близкородственных биологических видов посредством анализа молекулярных маркеров, в первую очередь, с использованием коротких нуклеотидных повторов (STR, Short Tandem Repeat) или SNP (Single Nucleotide Polymorphism), нашло свое отражение в научных работах отечественных и зарубежных исследователей [1–7]. Однако, например, ввиду прохождения популяции

дикого кабана через «бутылочное горлышко» в 2014–2016 гг. в Республике Беларусь по причине массового отстрела из-за вспышки африканской чумы свиней, использование алгоритма для дифференциации дикого кабана и домашней свиньи, основанного на расчетах частоты встречаемости аллелей в STR-локусах, должно быть перепроверено на вновь сформированной популяции. Ведь, как известно, при резкой значительной депопуляции вида происходит потеря, в первую очередь, редких аллелей STR, а частота самого распространенного, как правило, еще более возрастает. Таким образом, поиск альтернативных генетических маркеров, в частности, SNP, является весьма актуальной задачей, т.к. данный тип полиморфизма ДНК более чем в 95 % случаев находится только в двух аллельных формах.

В целом, универсального подхода, который бы позволил решать задачу дифференциации близкородственных видов с использованием SNP на основании поиска/отбора информативных SNP в fasta-файлах проектов NGS, не существует. Имеется достаточно большое количество бесплатных или условно бесплатных программ, в которые заложены алгоритмы анализа данных NGS, включая определение генотипа в SNP, локальное выравнивание, определение глубины прочтения в анализируемой нуклеотидной позиции. Еще больше программ реализовано на коммерческой основе. Однако имеющиеся в наличии и доступные нам программные продукты: например, UGENE [8], GATK [9], Annovar [10], SNPEFF [11], PLINK [12] или SNPTest [13] не могут быть использованы из-за ряда ограничений: либо необходимо загружать большие объемы данных в облачное хранилище, аренда которого платная; либо пайплайн, закодированный в программе, не способен одновременно обрабатывать несколько независимых поисковых запросов, а параллельный запуск этой же программы не предусмотрен, либо функционал программы направлен, в первую очередь, на аннотацию нуклеотидной последовательности. А тат как поиск в геноме и отбор полиморфных вариантов для дифференциации близкородственных видов предусматривает анализ сотен геномов и тысяч SNP, то распараллеливание поисковых запросов является обязательным условием для получения итогового результата в приемлемые временные сроки.

В этой связи, для автоматизации и распараллеливания процесса поиска нуклеотидных последовательностей, фланкирующих искомый аллель, т.е. для генотипирования *in silico*, нами разработана программа GENIS (**GEN**otyping **I**n **S**ilico), написанная на языке программирования Python v.3.10, в среде разработки программного обеспечения Jupyter Notebook. В рамках данной работы проведена апробация программы GENIS на файлах с результатами секвенирования геномов животных рода *Sus*, а также выявлены породоспецифичные SNP для свиней породы дюрок.

МАТЕРИАЛЫ И МЕТОДЫ

Образцы для биоинформатического анализа. Объектом для формирования перечня файлов SRA (Sequence Read Archive) стал геном *Sus scrofa*. Сформированный перечень содержал информацию об уникальном идентификационном коде (BioSample) образца, проекте (BioProject), в рамках которого была получена информация о нуклеотидных последовательностях, а также технические идентификационные номера (Experiment и Run), таблица 1.

Каждая платформа для полногеномного секвенирования характеризуется своими особенностями, для проведения данного биоинформатического анализа основными были количество прочтений (Spots) и средняя длина прочтения (Read), таблица 2. Значения данных параметров позволяют оценить общее количество нуклеотидов (Bases), определенных за один раунд работы прибора. Для анализа были отобраны пять различных платформ для NGS, с использованием которых был секвенирован геном *Sus*

scrofa – BGISEQ [17], ILLUMINA [18], LS454 [19], PACBIO SMRT [20] и OXFORD NANOPORE [21].

Таблица 1. Сводная информация об образцах *Sus scrofa*, отобранных для биоинформатического анализа

№№	BioSample[1]	BioProject[2]	Experiment[3]	Run[3]
1	SAMEA6565336	PRJEB36830	ERX3951401	ERR3943553
2	SAMEA6565336	PRJEB36830	ERX3951403	ERR3943555
3	SAMEA6565336	PRJEB36830	ERX3951405	ERR3943557
4	SAMN12612872	PRJNA530874	SRX6746775	SRR10008503
5	SAMN12633856	PRJNA530874	SRX6757578	SRR10020136
	SAMN18317865	PRJNA721459	SRX10994650	SRR14447529
6	SAMN25146071	PRJNA530874	SRX13850383	SRR17687110
7	SAMN25146078	PRJNA530874	SRX13850376	SRR17687117
8	SAMN03938573	PRJNA291011	SRX1123718	SRR2133352
9	SAMN03938573	PRJNA291011	SRX1123729	SRR2133363
10	SAMN03938573	PRJNA291011	SRX1123731	SRR2133365
11	SAMN03938573	PRJNA291011	SRX1123735	SRR2133369
12	SAMN03938573	PRJNA291011	SRX1123737	SRR2133371
13	SAMN03938573	PRJNA291011	SRX1123738	SRR2133372
14	SAMN03938573	PRJNA291011	SRX1123741	SRR2133375
15	SAMN03938573	PRJNA291011	SRX1123743	SRR2133377
16	SAMN03938573	PRJNA291011	SRX1123745	SRR2133379
17	SAMN03938573	PRJNA291011	SRX1123747	SRR2133381
18	SAMN09531794	PRJNA478804	SRX4615665	SRR7759992
19	SAMN09531794	PRJNA478804	SRX4615664	SRR7759993
20	SAMN09531794	PRJNA478804	SRX4615663	SRR7759994
21	SAMN09531794	PRJNA478804	SRX4615662	SRR7759995
22	SAMN09531794	PRJNA478804	SRX4615661	SRR7759996
23	SAMN09531794	PRJNA478804	SRX4615660	SRR7759997
24	SAMN09531794	PRJNA478804	SRX4615793	SRR7760075
25	SAMN09531794	PRJNA478804	SRX4615783	SRR7760085
26	SAMN09531794	PRJNA478804	SRX4615776	SRR7760092
27	SAMN09531794	PRJNA478804	SRX4615769	SRR7760099
28	SAMN09531794	PRJNA478804	SRX4615758	SRR7760110
29	SAMN09531794	PRJNA478804	SRX4615747	SRR7760121
30	SAMN09531794	PRJNA478804	SRX5294764	SRR8490205
31	SAMN09531794	PRJNA478804	SRX5326590	SRR8523737
32	SAMN09531794	PRJNA478804	SRX5326576	SRR8523751
33	SAMN09531794	PRJNA478804	SRX5326575	SRR8523752
34	SAMN09531794	PRJNA478804	SRX6606585	SRR9851973
35	SAMN12568965	PRJNA530874	SRX6711348	SRR9963836

Примечания: 1 – база данных BioSample [14], BioProject [15]; 3 – база данных SRA [16]

Полиморфизмы. Ранее, в ходе проведения расширенного биоинформатического анализа геномов *Sus scrofa* с целью определить специфичные SNP для дифференциации коммерческих пород свиней, было определено, что имеется шесть возможных комбинаций нуклеотидных замен, а именно: A/C, A/G, A/T, C/G, C/T, G/T [22]. С учетом данной информации был сформирован перечень SNP для анализа, включающий по 10 полиморфизмов для каждой из нуклеотидных замен. SNP располагались на всех хромосомах *Sus scrofa* на отдалении друг от друга – медианное значение составило 8 075 356 нуклеотида (минимальное – 8 195 нуклеотида, максимальное – 86 139 062 нуклеотида), таблица 3.

Таблица 2. Сводная информация о файлах с результатами секвенирования геномов *Sus scrofa*

Run[1]	Spots[2]	Read[3]	Платформа	Прибор
ERR3943553	46267664	100	BGISEQ	BGISEQ-500
ERR3943555	30074196	100	BGISEQ	BGISEQ-500
ERR3943557	41632598	100	BGISEQ	BGISEQ-500
SRR10008503	322235349	300	ILLUMINA	HiSeq X Ten
SRR10020136	296464777	300	ILLUMINA	HiSeq X Ten
SRR14447529	200000000	100	BGISEQ	BGISEQ-500
SRR17687110	400423255	300	ILLUMINA	HiSeq X Ten
SRR17687117	399043828	300	ILLUMINA	HiSeq X Ten
SRR2133352	489781	536	LS454	454 GS FLX Titanium
SRR2133363	226845	449	LS454	454 GS FLX Titanium
SRR2133365	1051317	475	LS454	454 GS FLX Titanium
SRR2133369	827873	478	LS454	454 GS FLX Titanium
SRR2133371	927165	561	LS454	454 GS FLX Titanium
SRR2133372	848669	506	LS454	454 GS FLX Titanium
SRR2133375	507164	631	LS454	454 GS FLX Titanium
SRR2133377	1154376	447	LS454	454 GS FLX Titanium
SRR2133379	1138253	453	LS454	454 GS FLX Titanium
SRR2133381	1091743	445	LS454	454 GS FLX Titanium
SRR7759992	899509	9747	PACBIO SMRT	Sequel
SRR7759993	830607	9819	PACBIO SMRT	Sequel
SRR7759994	1051269	8884	PACBIO SMRT	Sequel
SRR7759995	919029	9999	PACBIO SMRT	Sequel
SRR7759996	933714	9695	PACBIO SMRT	Sequel
SRR7759997	884016	9664	PACBIO SMRT	Sequel
SRR7760075	57853171	300	ILLUMINA	HiSeq X Ten
SRR7760085	248467649	300	ILLUMINA	HiSeq X Ten
SRR7760092	28428815	300	ILLUMINA	HiSeq X Ten
SRR7760099	22328258	300	ILLUMINA	HiSeq X Ten
SRR7760110	33148236	300	ILLUMINA	HiSeq X Ten
SRR7760121	123793033	300	ILLUMINA	HiSeq X Ten
SRR8490205	44912788	300	ILLUMINA	HiSeq X Ten
SRR8523737	103937336	300	ILLUMINA	HiSeq X Ten
SRR8523751	4520337	300	ILLUMINA	HiSeq X Ten
SRR8523752	93312636	300	ILLUMINA	HiSeq X Ten
SRR9851973	1871364	20288	OXFORD NANOPORE	PromethION
SRR9963836	315124190	300	ILLUMINA	HiSeq X Ten

Примечания: 1 – база данных SRA [16]; 2 – количество прочтений; 3 – средняя длина прочтения (количество нуклеотидов)

Таблица 3. Сводная информация о полиморфизмах, отобранных для биоинформатического анализа

Код	Affy SNP ID[1]	Хромосомная позиция[2]	Полиморфизм	*Tm*[3]	CG %[3]
SNP_01	Affx-114720993	1:31932169	C/G	70.4	37
SNP_02	Affx-114701012	1:118071231	G/T	68.7	32
SNP_03	Affx-115005763	1:118111365	A/G	68.7	32
SNP_04	Affx-114883144	1:119867557	C/T	71.0	38
SNP_05	Affx-114929237	1:185604514	A/C	66.4	27
SNP_06	Affx-115196355	2:39614412	C/G	73.9	45
SNP_07	Affx-114865670	2:39632721	A/G	75.1	48
SNP_08	Affx-114904768	2:52404337	G/T	73.3	44
SNP_09	Affx-114825778	2:120825903	A/C	74.5	46
SNP_10	Affx-114781900	2:140452243	C/T	72.8	42
SNP_11	Affx-114949970	3:58388838	C/G	70.4	37
SNP_12	Affx-115293304	3:64630653	G/T	77.4	54

Код	Affy SNP ID[1]	Хромосомная позиция[2]	Полиморфизм	*Tm*[3]	CG %[3]
SNP_13	Affx-114905741	3:70425966	A/G	64.7	23
SNP_14	Affx-114989292	3:77735111	C/T	73.3	44
SNP_15	Affx-115280725	4:7319711	C/T	75.1	48
SNP_16	Affx-114620280	4:57136761	A/G	73.9	45
SNP_17	Affx-114925363	4:75946087	C/G	67.0	28
SNP_18	Affx-114617715	4:100917120	A/T	76.8	52
SNP_19	Affx-115252158	4:105079765	A/C	74.5	46
SNP_20	Affx-115011247	5:45280392	C/T	64.7	23
SNP_21	Affx-114841044	5:45330381	G/T	68.1	31
SNP_22	Affx-114782192	5:71189934	A/G	69.3	34
SNP_23	Affx-114634296	5:76067717	A/T	70.4	37
SNP_24	Affx-115024456	6:103189180	C/T	70.4	37
SNP_25	Affx-114918769	6:112930618	G/T	71.6	39
SNP_26	Affx-115135405	6:121005974	A/G	70.4	37
SNP_27	Affx-114842388	6:135820186	A/C	72.8	42
SNP_28	Affx-114799038	6:161450197	C/G	64.7	23
SNP_29	Affx-114922841	6:161476704	A/T	68.7	32
SNP_30	Affx-114881642	7:70206542	C/T	70.4	37
SNP_31	Affx-114701447	7:83717060	G/T	73.3	44
SNP_32	Affx-115248609	7:91364288	A/G	74.5	46
SNP_33	Affx-115028950	8:22107041	C/G	72.8	42
SNP_34	Affx-115091951	8:39871989	A/G	74.5	46
SNP_35	Affx-114758152	8:44711400	C/T	68.1	31
SNP_36	Affx-114845298	8:95372504	A/T	67.6	30
SNP_37	Affx-114691953	8:101905410	G/T	69.9	35
SNP_38	Affx-115281020	8:102674459	A/C	64.1	21
SNP_39	Affx-115246344	9:47650875	A/G	78.0	55
SNP_40	Affx-114786170	9:58767893	A/C	68.7	32
SNP_41	Affx-115264145	9:64951529	C/T	74.5	46
SNP_42	Affx-115266623	9:137516833	C/G	75.4	49
SNP_43	Affx-115219849	9:138661524	G/T	73.9	45
SNP_44	Affx-114844219	10:1285200	C/T	68.1	31
SNP_45	Affx-115195072	10:51106163	A/G	76.8	52
SNP_46	Affx-115047526	11:33493174	A/C	72.2	41
SNP_47	Affx-114866577	11:33501369	A/T	69.3	34
SNP_48	Affx-115066610	12:42603997	A/T	66.4	28
SNP_49	Affx-114738752	13:50389675	C/G	76.8	52
SNP_50	Affx-114806135	13:53234428	G/T	72.8	42
SNP_51	Affx-115188264	13:126819803	A/C	70.4	37
SNP_52	Affx-114793371	13:181189056	A/T	73.3	44
SNP_53	Affx-114822549	14:93950749	C/G	70.4	37
SNP_54	Affx-115041119	15:53360154	C/G	69.9	35
SNP_55	Affx-115132762	15:53387069	A/T	73.3	44
SNP_56	Affx-114694487	15:58295974	A/C	67.0	28
SNP_57	Affx-114669233	16:68101059	A/C	71.6	39
SNP_58	Affx-115021995	17:10238699	A/T	69.3	34
SNP_59	Affx-115102059	17:10503838	G/T	72.2	41
SNP_60	Affx-115123857	X:51861764	A/T	71.6	39

Примечания: 1 – код полиморфизма в Axiom™ Porcine Genotyping Array [23]; 2 – версия сборки генома *Sus scrofa* Sscrofa11.1 (GCF_000003025.6); 3 – температура плавления ампликона, Oligo Calc: Oligonucleotide Properties Calculator [24]

Среднее значение для расчетного параметра T_m [14] для 60 SNP составило 71.3 ± 3.40 ºC, CG – 38.7 ± 8.20 %; для каждой из нуклеотидных замен: A/C – 70.2 ± 3.6 ºC (35.9 ± 8.5 %), A/G – 72.6 ± 4.2 ºC (41.8 ± 10.0 %), A/T – 70.7 ± 3.1 ºC (37.4 ± 7.4 %), C/G – 71.2 ± 3.7 ºC (38.5 ± 8.9 %), C/T – 70.8 ± 3.2 ºC (37.7 ± 7.7 %) и G/T – 72.1 ± 2.7 ºC (40.7 ± 6.7 %).

Эти два параметра не различаются в пределах нуклеотидных замен, что, потенциально, не должно было повлиять на глубину прочтения в анализируемых позициях.

Длина строки поиска для каждого полиморфизма варьировала от 15 до 25 нуклеотидов, с шагом в два нуклеотида. Пример формирования нуклеотидных последовательностей представлен в таблице 4.

Таблица 4. Нуклеотидные последовательности, используемые для генотипирования *in silico*

Код	N*	Левая фланкирующая последовательность (L)	Правая фланкирующая последовательность (R)
SNP_01	15	CAGTTATGTTTGTTT	TGTTGCTATTTAAAT
SNP_01	17	CCCAGTTATGTTTGTTT	TGTTGCTATTTAAATGT
SNP_01	19	TCCCCAGTTATGTTTGTTT	TGTTGCTATTTAAATGTGA
SNP_01	21	ACTCCCCAGTTATGTTTGTTT	TGTTGCTATTTAAATGTGAAG
SNP_01	23	TGACTCCCCAGTTATGTTTGTTT	TGTTGCTATTTAAATGTGAAGTT
SNP_01	25	TATGACTCCCCAGTTATGTTTGTTT	TGTTGCTATTTAAATGTGAAGTTTT

* количество нуклеотидов в искомой последовательности

Определение генотипа in silico. Для биоинформатического анализа были использованы геномы животных, представленные в открытом доступе в формате SRA, которые дополнительно конвертировали в формат *.fasta с использованием пакета SRA-Toolkit v.2.11. Для автоматизации процесса поиска нуклеотидных последовательностей *in silico*, фланкирующих искомый аллель, использовали программу, написанную на языке программирования Python v.3.10, в среде разработки программного обеспечения Jupyter Notebook.

Аппаратное обеспечение. Для проведения биоинформатического анализа были задействованы четыре персональные электронно-вычислительные машины (ПЭВМ), центральное процессорное устройство которых было представлено интегральными схемами (чипом) нескольких последовательных поколений от разработчика и производителя электронных устройств Intel ARK, включая последнее 13 поколение (таблица 5).

Таблица 5. Характеристики ПЭВМ (персональных компьютеров, ПК)

Код	CPU[1]	RAM[2]			Системная плата
		Тип	Объем, Гб	Частота, МГц (тайминги)	
ПЭВМ_1	Intel Core i7-7700	DDR4	16(1)*	2400 (17-17-17-40)	ASRock H110M-HDV R3.0
ПЭВМ_2	Intel Core i3-12100	DDR4	16(2)*	3200 (16-18-18-36)	MSI Pro H610M-G (MS-7D46)
ПЭВМ_3	Intel Core i7-11700	DDR4	16(2)*	3200 (16-18-18-36)	MSI MAG B560 Torpedo
ПЭВМ_4	Intel Core i7-13700K	DDR5	16(2)*	5600 (40-40-40-80)	AsRock Z690 Steel Legend/D5

Примечания: 1 – Central Processing Unit (центральное процессорное устройство); 2 – Random Access Memory (оперативная память); * 1 или 2 модуля

Анализ SRA файлов проводили с использованием твердотельного накопителя (Solid-State Drive, SSD) Patriot Viper VPN100 1TB VPN100-1TBM28H. Подключение SSD осуществляли через адаптер M.2 NVME to PCI-E 3.0x4 Expansion Card (UGREEN).

Статистический анализ данных. Для сравнения количественных переменных: время поиска нуклеотида (в секундах), глубина прочтений нуклеотида (в абсолютных

значениях), – использовали дисперсионный анализ (ANOVA – Analysis of Variation). Взаимосвязь между количественными переменными оценивали с использованием линейной регрессионной модели, рассчитывали коэффициент детерминации R^2. Статистический анализ проводили в SPSS v.20.0.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Оценка эффективности работы GENIS

В основу работы программы GENIS положен подход, основанный на регистрации вхождения искомой последовательности в последовательность в файле с результатами секвенирования, код программы GENIS представлен в файле GenisForText.pdf. В том случае, если регистрируется вхождение, программа учитывает аллель (A/T/G/C), а также общее количество подобных вхождений в файле с результатами секвенирования. Впоследствии вся полученная информация компилируется и, в зависимости от полученной информации (аллель/покрытие), принимается решение о присвоении конкретному SNP того или иного генотипа. В связи с тем, что для поиска видо- или породоспецифичных SNP, как правило, необходимо протестировать сотни файлов с результатами секвенирования для тысяч SNP, нами была также запрограммирована возможность параллельного анализа одного файла для различных SNP.

Минимальная длина строки поиска для корректного определения генотипа для платформы BGISEQ (прибор BGISEQ-500) составила 19 нуклеотидов, для платформ ILLUMINA (HiSeq X Ten), LS454 (454 GS FLX Titanium), PACBIO SMRT (Sequel), ILLUMINA (HiSeq X Ten) и OXFORD NANOPORE (PromethION) – 21 нуклеотид. Статистически значимая ассоциации между временем поиска и количеством выявленных нуклеотидных последовательностей с искомым олигонуклеотидом отсутствовала, т.е. программа затрачивала сопоставимое время на поиск одной или > 100 специфичных последовательностей. Также не было выявлено статистически значимых различий между временем поиска и длиной искомой последовательности из таблицы 4.

Прямая линейная зависимость была показана для переменных «Bases» (общее количество секвенированных нуклеотидов) и «Time_MEAN» (среднее время поиска, сек), рисунок 1. Коэффициент детерминации (R^2) для всех значений в совокупности (независимо от типа платформы для NGS) составил 0.983. Отдельно для каждой из платформ: для BGISEQ (BGISEQ-500) – 1.0; для ILLUMINA (HiSeq X Ten) – 0.999; для LS454 (454 GS FLX Titanium) – 0.987; для PACBIO SMRT (Sequel) – 0.943.

Переменная «Time_MEAN» представляла собой среднее арифметическое значений времени поиска для всех SNP при длине искомой последовательности в 21 нуклеотид, т.е. для 60 значений для каждого анализируемого файла RUN из таблицы 2. Таким образом, независимо от средней длины прочтения при использовании различных платформ NGS наблюдается выраженная линейная связь между переменной «Bases» (которая в свою очередь коррелирует со значением $R \approx 1$ с переменной «Spots») и «Time_MEAN» (среднее время поиска, сек).

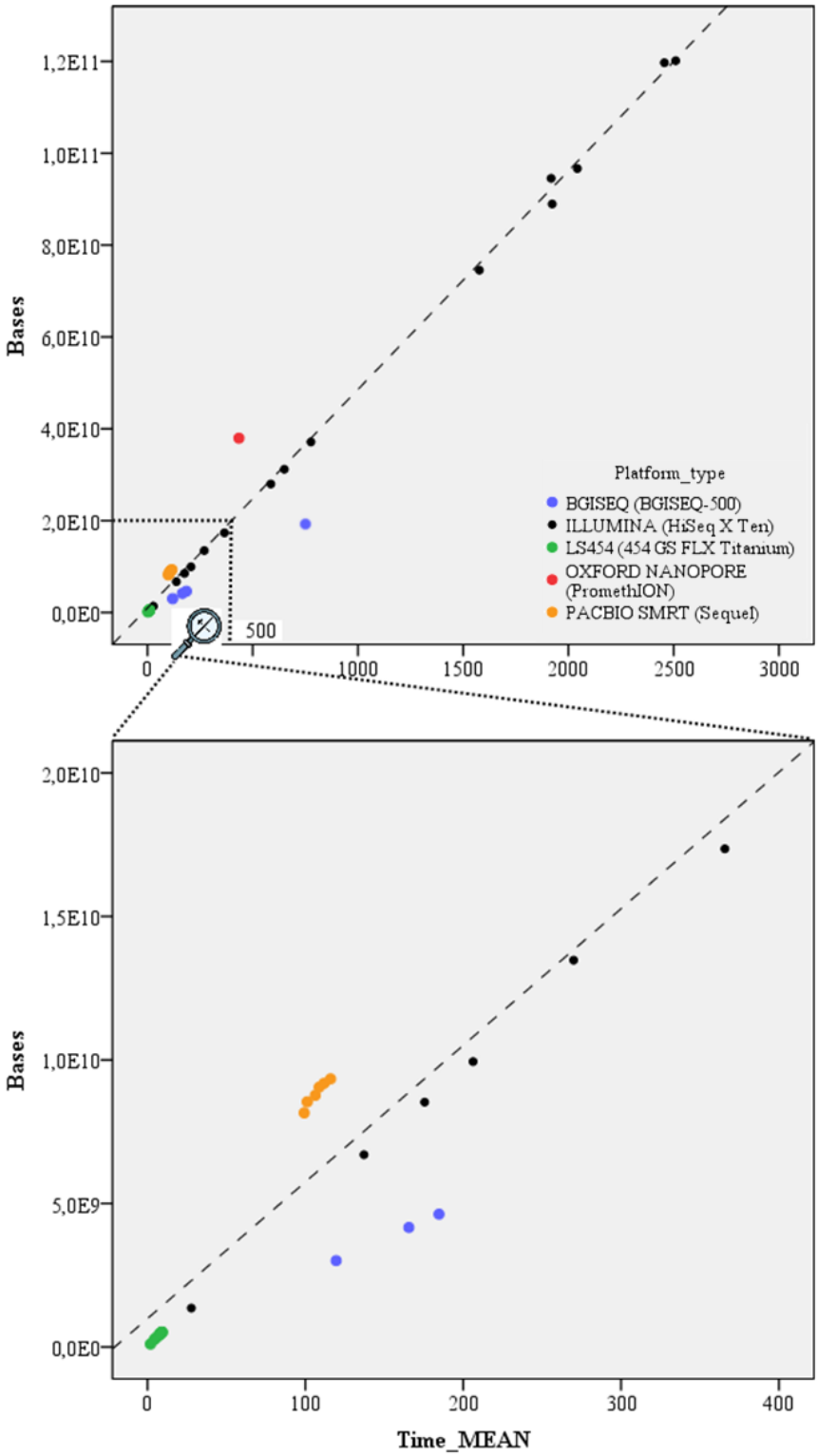

Рис. 1. Общая закономерность: зависимость между временем на определение генотипа (сек) [ось *X*] *in silico* и общим количеством нуклеотидов в рабочем файле [ось *Y*] (для ПЭВМ_1).

При использовании различных ПЭВМ (таблица 5) результаты генотипирования *in silico* полностью совпали. В то же время значения среднего времени поиска последовательности в 21 нуклеотид различались в зависимости от используемой ПЭВМ, рисунок 2.

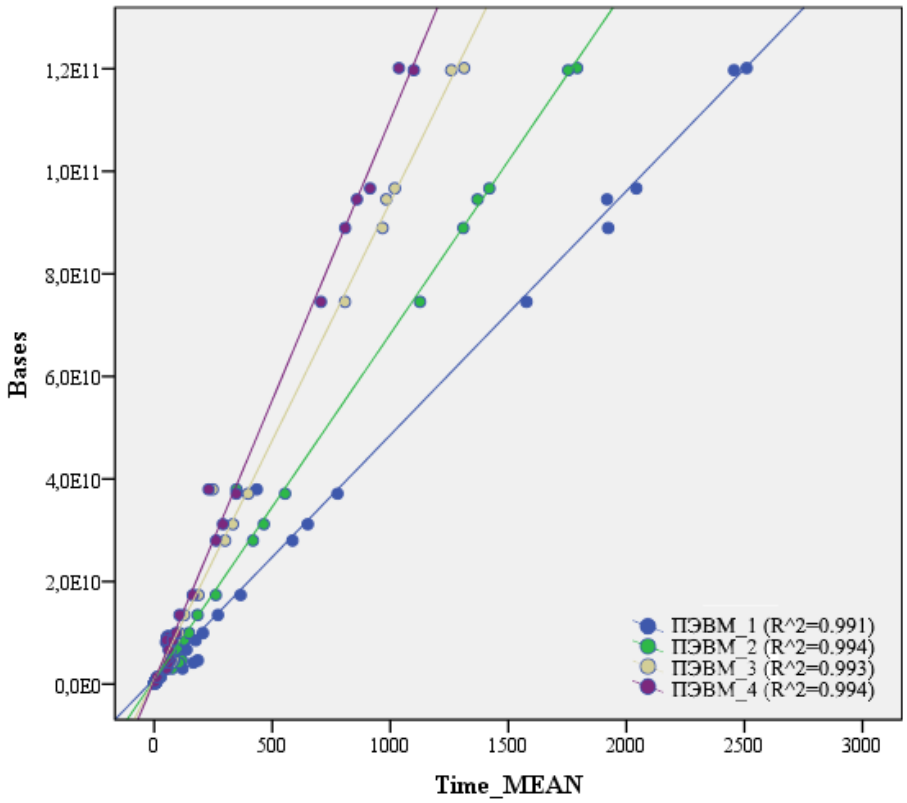

Рис. 2. Сравнительный аспект: зависимость между временем на определение генотипа (сек) *in silico* [ось *X*] и общим количеством нуклеотидов в рабочем файле [ось *Y*] для четырех ПЭВМ.

Для всех ПЭВМ использовались базовые настройки UEFI (UnifiedExtensible Firmware Interface): технология для повышения производительности – гиперпоточность (hyper-threading technology, HTT) включена; андервольтинг (Undervolting CPU) – процесс, который позволяет уменьшить энергопотребление и тепловыделение, не влияя на производительность системы, не использовался; схема управления электропитанием (Пуск\Панель управления\Все элементы панели управления\Электропитание) – сбалансированная.

Наиболее быстрым закономерно оказался поиск при использовании ПЭВМ_4, включающей процессор Intel Core i7-13700K последнего поколения по состоянию на конец 2022 г. В среднем, данная ПЭВМ_4 в 6-поточном режиме (одновременный поиск по 6 возможным нуклеотидным заменам – A/C, A/G, A/T, C/G, C/T, G/T) позволяла получать результат в 2.23 ± 0.20 раза быстрее, чем ПЭВМ_1, в 1.60 ± 0.11 раза быстрее, чем ПЭВМ_2 и в 1.15 ± 0.07 раза быстрее, чем ПЭВМ_3.

Проведенный нами анализ включал также использование жесткого диска HDD (WD Purple 4TB WD42PURZ), SSD форм-фактора 2.5” (SSD Samsung 870 Evo 250GB MZ-77E250BW) и SSD форм-фактора M.2 (Patriot Viper VPN100 1TB VPN100-1TBM28H) в унипоточном режиме (поиск только для нуклеотидной замены A/C). При сравнении значений времени поиска нуклеотидных последовательностей (длина 21 нуклеотид) статистически значимых различий в пределах использования одной ПЭВМ выявлено не было. Таким образом, даже при наличии ПЭВМ с характеристиками, не отличающимися высокой вычислительной мощностью, т.н. «офисного ПК», возможно производить генотипирование *in silico* с использованием разработанного нами алгоритма в унипоточном режиме. При использовании HDD в многопоточном режиме время на поиск возрастает пропорционально количеству потоков и при наличии одновременно работающих 4 и более потоков стремится к нерационально долгому времени.

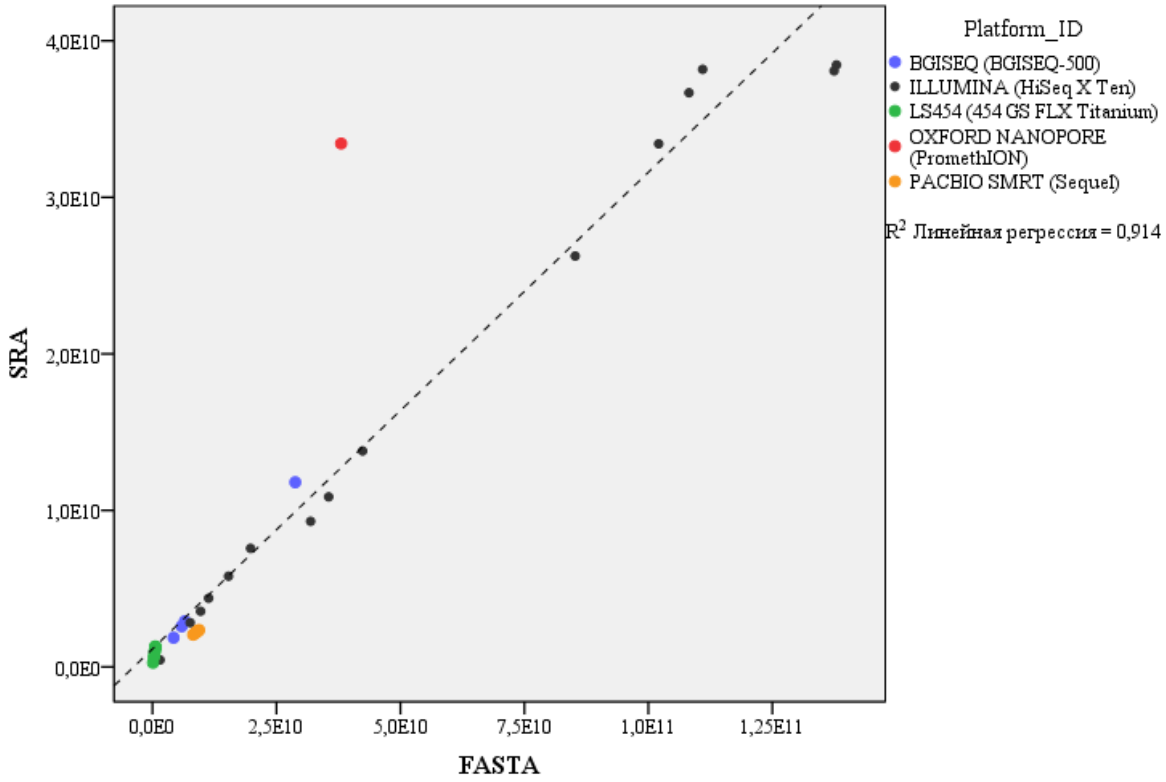

Рис. 3. Зависимость между размером файла SRA (в байтах) и файлом в формате fasta (в байтах).

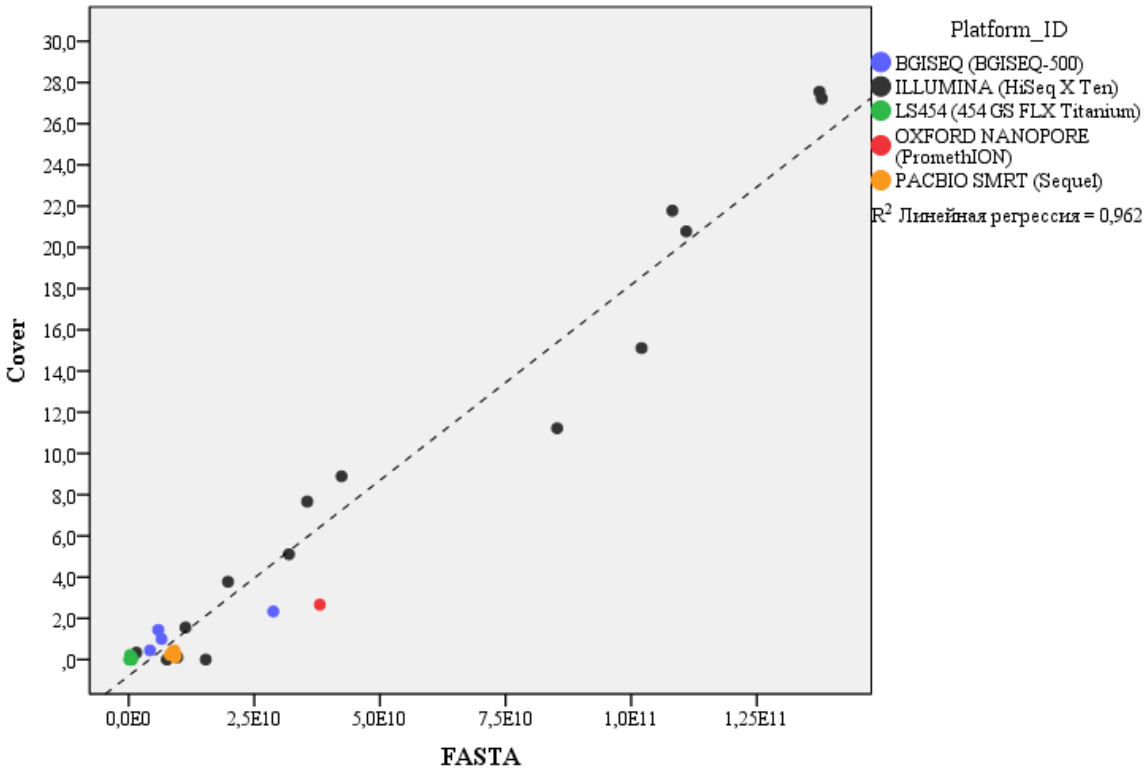

Рис. 4. Зависимость между размером файла в формате fasta (в байтах) и средней глубиной прочтения в анализируемом SNP (Cover).

Возможности ПЭВМ_3 и ПЭВМ_4 позволяют одновременно задействовать 14–20 независимых потоков для поиска нуклеотидных последовательностей, т.е. для генотипирования *in silico*, при загруженности CPU на 95–100 %. Скорость генотипирования *in silico* в многопоточном режиме на 15–20 % ниже, чем в

унипоточном режиме, однако возможность распараллеливания рабочего процесса сокращает общее время анализа в десятки раз. Подобных показателей позволяет добиться использование в качестве рабочего пространства для хранения и анализа файлов SRA твердотельных накопителей (SSD), работающих через интерфейс PCIe. Для данных дисков характерны высокая скорость последовательного чтения (например, для используемого нами Patriot Viper VPN100 1TB VPN100-1TBM28H – 3 450 МБ/с) и средняя скорость случайного чтения (600 000 IOps). Для примера, для SSD, работающих через интерфейс SATA 3.0 (например, для используемого нами SSD Samsung 870 Evo 250GB MZ-77E250BW), скорость последовательного чтения и средняя скорость случайного чтения меньше в 6.16 и 6.13 раза соответственно. Использование SSD, работающих через интерфейс PCIe, является наиболее подходящим при наличии высокопроизводительной ПЭВМ.

При конвертации файла SRA в файл с расширением fasta с использованием пакета SRA-Toolkit v.2.11 происходит изменение размера конечного файла (в байтах), обычно размер значительно увеличивается: например, для OXFORD NANOPORE (PromethION) – в среднем в 1.14 раза, для BGISEQ (BGISEQ-500) – в 2.26 раза, для ILLUMINA (HiSeq X Ten) – в 3.06 раза, для PACBIO SMRT (Sequel) – в 3.98 раза. В целом, зависимость носит линейный характер, рисунок 3.

Зависимость между размером файла с расширением fasta и средней глубиной прочтения в анализируемом SNP также носит линейный характер и представлена на рисунке 4. Исследователю не доступна точная информация о размере файла с расширением fasta на этапе отбора файлов SRA для биоинформатического анализа, однако, зная коэффициент увеличения файла при конвертации, можно с достаточной высокой точностью оценить, стоит ли отбирать данный файл в исследуемую группу. Например, для самой распространенной технологии NGS (ILLUMINA), уравнение линейной регрессии имеет следующий вид: Y[Cover, нуклеотиды] = 6×10^{-10} * X[SRA, байты] – 1.439, R^2 = 0.930.

Соответственно, средняя глубина прочтения в 10 нуклеотидов ожидается при размере файла SRA в 19.1 ГБ, средняя глубина прочтения в 5 нуклеотидов ориентировочно достижима при размере файла SRA в 10.7 ГБДля примера, разработанный нами методологический подход для генотипирования с использованием ПЭВМ_4 позволяет определить 1000 генотипов *in silico* (100 SNP для 10 образцов *Sus scrofa* со средним покрытием в искомом нуклеотиде ≈10) за 7–8 часов.

Практическое применение программы GENIS

Нами было проведено дополнительное исследование с использованием программного обеспечения GENIS, в рамках которого проведен анализ результатов секвенирования геномов *Sus scrofa domesticus* и определены SNP со значительным потенциалом для дифференциации домашних свиней породы дюрок. В перечень потенциально информативных SNP вошли как ранее описанные SNP [25], так и фланкирующие их SNP из Axiom® Porcine Genotyping Array (Axiom_PigHDv1) от Affymetrix©, а также SNP, для которых имелась доказанная ассоциация с QTL (Quantitative Trait Locus) и была представлена информация в Pig Quantitative Trait Locus Database (Pig QTLdb (https://www.animalgenome.org)). Общее количество SNP в анализе составило 261. Секверированные нуклеотидные последовательности особей *Sus scrofa domesticus*, были представлены как в формате fasta (имели статус «aligned»), так и в формате SRA. Общее количество образцов в анализе – 248 (дюрок – 85, ландрас – 46, пьетрен – 28, крупная белая – 70, йоркшир – 19).

Данные о секвенированных нуклеотидных последовательностях полных геномов *Sus scrofa domesticus* расположены в базе Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra): PRJNA41185, PRJNA176478, PRJNA186497, PRJEB1683, PRJNA239399, PRJNA260763, PRJNA255085, PRJEB9922, PRJNA309108,

PRJNA322309, PRJNA343658, PRJNA358108, PRJNA369600, PRJNA378496, PRJNA393920, PRJNA487172, PRJNA506339, PRJNA507853, PRJNA485589, PRJNA488960, PRJNA550237, PRJNA520978, PRJNA553106, PRJNA671763, PRJNA626370, PRJNA622908.

Дифференцирующий потенциал SNP определяли с использованием ROC-анализа в SPSS v.20.0. При наличии нижней границы асимптотического 95 % доверительного интервала, значение которой более 0.5, для параметра AUC (площадь под кривой) SNP позиционировался как генетический маркер со значительным дифференцирующим потенциалом, более 0.7 – как генетический маркер с высоким дифференцирующим потенциалом.

Проведенный биоинформатический анализ, направленный на определение генотипа *in silico* для животных вида *Sus scrofa domesticus*, позволил рассчитать частоты генотипов у пяти пород свиней – дюрок (выборка «DU»), ландрас («LA»), пьетрен (PI), крупная белая («LW») и йоркшир («YO»). Полученные результаты легли в основу математического анализа классификаций с применением ROC-кривых (receiver operating characteristic).

По причине наличия значительного практического материала, полученного в ходе выполнения данного исследования, считаем не возможным подробно предоставить все наши результаты. Подробный анализ для каждой породы свиней будет осуществлен в рамках отдельных публикаций. В то же время, в статье представлены данные об участках в геноме *Sus scrofa domesticus*, в которых в ходе селекции породы дюрок сформировались кластеры (гаплотипы) SNP, молекулярно-генетический анализ которых позволит оценить чистопородность конкретной особи или наличие примесей других пород в исследуемой выборке животных. Сводная информация по наличию таких наиболее информативных кластеров представлена в таблице 6.

Таблица 6. Расположение высоинформативных SNP и их кластеров в геноме *Sus scrofa* для дифференциации пород свиней дюрок

Хромосомная позиция	AUC	НГ 95 % ДИ	ВГ 95 % ДИ	p-уровень
Chr.4:9676649	0.7953	0.7434	0.8473	8.12×10^{-15}
Chr.7:106301845	0.8255	0.7710	0.8800	1.13×10^{-17}
Chr.8:45485535	0.7803	0.7255	0.8352	1.69×10^{-13}
Chr.8:47482649	0.8231	0.7718	0.8744	1.94×10^{-17}
Chr.9:25438712	0.7668	0.7122	0.8214	2.28×10^{-12}
Chr.9:49010671	0.7958	0.7405	0.8511	7.34×10^{-15}
Chr.9:49034938	0.7721	0.7149	0.8293	8.39×10^{-13}
Chr.9:49042137	0.7737	0.7176	0.8298	6.14×10^{-13}
Chr.9:49046617	0.7729	0.7167	0.8291	7.22×10^{-13}
Chr.9:138661524	0.7604	0.7047	0.8161	7.50×10^{-12}
Chr.12:11541028	0.7875	0.7305	0.8445	4.01×10^{-14}
Chr.13:182876924	0.7892	0.7324	0.8461	2.83×10^{-14}
Chr.14:73886405	0.7681	0.7109	0.8253	1.80×10^{-12}
Chr.14:99099156	0.8149	0.7585	0.8713	1.22×10^{-16}
Chr.14:107794160	0.7618	0.7065	0.8172	5.76×10^{-12}
Chr.16:32192496	0.8180	0.7671	0.8688	6.20×10^{-17}

Для проверки корректности получаемых в ходе биоинформатических исследований данных, нами была сформирована тестовая выборка животных, разводимых в Беларуси, которая включала следующие породы свиней: белорусская крупная белая – БКБ (38 шт.), белорусская мясная – БМ (12 шт.), дюрок – ДЮ (26 шт.), ландрас – ЛА (9 шт.), йоркшир – ЙО (10 шт.). Биологические образцы (выщипы ушной раковины) свиней были отобраны сотрудниками СГЦ «Заднепровский» (Витебская обл., Беларусь). Для выделения ДНК использовали набор «Нуклеосорб» (Праймтех, Беларусь).

Концентрацию ДНК и степень ее очистки определяли с использованием спектрофотометра Implen Nano Photometer N50 (Implen, Германия). Для генотипирования по Chr.7:106301845A > G (rs80967182), Chr.8:47482649G > T (rs81333725) и Chr.14:99099156T > C (rs80859281), использовали технологию KASP (Kompetitive allele specific PCR, LGC Biosearch Technologies, Великобритания). ПЦР проводили на амплификаторе Quant Studio 5 (Applied Biosystems, США) согласно протоколу производителя. Дифференцирующий потенциал SNP определяли с использованием ROC-анализа в SPSS v.20.0.

Исследуемые SNP в порядке увеличения дифференцирующего потенциала расположились в следующей последовательности: Chr.8:47482649G > T (AUC = 0.887. 95 % ДИ = [0.820 – 0.953], $p < 7.1 \times 10^{-9}$), Chr.14:99099156T > C (AUC = 0.942, 95 % ДИ = [0.870 – 1.0], $p < 3.5 \times 10^{-11}$) и Chr.7:106301845A > G (AUC = 0.957, 95 % ДИ=[0.919 – 0.995], $p < 7.9 \times 10^{-12}$).

В результате проведенного MDR-анализа определено, что при анализе SNP Chr.7:106301845A > G, Chr.8:47482649G > T и Chr.14:99099156T > C точность дифференциации свиней породы дюрок достигает 97.5 % при значении хи-квадрат 105.961 ($p < 0.0001$).

ЗАКЛЮЧЕНИЕ

Разработан универсальный методологический подход, который позволяет определять генотип *in silico* для SNP на основании данных, полученных при NGS. Данный подход универсален, его можно использовать при биоинформатическом анализе любых файлов (*.fasta) с результатами секвенирования независимо от исследуемого биологического вида.

Разработанный нами подход основан на автоматизации процесса поиска нуклеотидных последовательностей, фланкирующих искомый аллель. Поиск проводится на ПЭВМ исследователя, не требует дорогостоящего оборудования, язык программирования Python v.3.10 и среда разработки программного обеспечения Jupyter Notebook бесплатны и общедоступны. Методологический подход для генотипирования *in silico* реализован в виде программы GENIS. В рамках данной работы проведена апробация программы на файлах с результатами секвенирования геномов животных рода *Sus*.

Разработанный нами подход может применяться для широкого круга задач, основными из которых, на наш взгляд, являются: поиск видо-(родо-) и/или породоспецифичных SNP, которые способны дифференцировать (различать) исследуемые выборки между собой, например, для пар *Sus scrofa domesticus* (домашняя свинья) и *Sus scrofa scrofa* (дикий кабан), *Canis lupus familiaris* (домашняя собака) и *Canis lupus* (волк), *Bos grunniens* (домашний як) и *Bos taurus* (крупный рогатый скот), *Hypophthalmichthys nobilis* (пестрый толстолобик) и *Hypophthalmichthys molitrix* (белый толстолобик) и др. Подобные исследования уже проводятся, результаты некоторых их них представлены в научных публикациях [26, 27].

С использованием программы GENIS уже подтвержден высокий дифференцирующий потенциал ряда SNP для дифференциации свиней породы дюрок.

В перспективе, полученные результаты будут способствовать понимаю динамики и вектора процессов, происходящих при селекции в сельском хозяйстве и аквакультуре, т.к. SNP, обладающие значительным потенциалом для дифференицации близкородственных видов, также могут быть ассоциированы с локусами количественных признаков (QTL, Quantitative Trait Loci). Также оценка частоты данных SNP будет способствовать расширению научных представлений о динамике эволюционных изменений в кратко- и среднесрочной перспективе для одомашненных видов животных в сравнении с их дикими сородичами, т.к. закрепление тех или иных

полиморфных локусов в геноме ассоциировано с искусственным отбором, проводимым человеком.

Исследование выполнено в рамках НИР «Биоинформатический подход к анализу данных полногеномного секвенирования для поиска однонуклеотидных замен, способных дифференцировать близкородственные биологические виды» (БРФФИ, 2023-2025 гг., Б23-060).

СПИСОК ЛИТЕРАТУРЫ

1. Рябцева А.О., Цыбовский И.С., Котова С.А. Микросателлитные маркеры в исследовании полиморфизма дикого кабана (*Sus scrofa*) и свиньи домашней (*Sus scrofa domestica*), обитающих на территории Республики Беларусь. *Молекулярная и прикладная генетика*. 2018. Т. 25. С. 56–65.
2. Rębała K., Rabtsava A.A., Kotova S.A., Kipen V.N., Zhurina N.V., Gandzha A.I., Tsybovsky I.S. STR Profiling for Discrimination between Wild and Domestic Swine Specimens and between Main Breeds of Domestic Pigs Reared in Belarus. *PLoS One*. 2016. V. 11. № 11. P. e0166563. doi: 10.1371/journal.pone.0166563
3. Носова А.Ю., Кипень В.Н., Царь А.И., Лемеш В.А. Дифференциация гибридного потомства белого (*Hypophthalmichthys molitrix Val.*) и пестрого (*H. nobilis Rich.*) толстолобиков на основании полиморфизма микросателлитных локусов. *Генетика*. 2020. Т. 56. № 3. С. 313–320. doi: 10.31857/S0016675820030121
4. Conyers C.M., Allnutt T.R., Hird H.J., Kaye J., Chisholm J. Development of a microsatellite-based method for the differentiation of European wild boar (*Sus scrofa scrofa*) from domestic pig breeds (*Sus scrofa domestica*) in food. *Journal of Agricultural and Food Chemistry*. 2012. V. 60. № 13. P. 3341–3347. doi: 10.1021/jf205109b
5. Lorenzini R., Fanelli R., Tancredi F., Siclari A., Garofalo L. Matching STR and SNP genotyping to discriminate between wild boar, domestic pigs and their recent hybrids for forensic purposes. *Scientific Reports*. 2020. V. 10. P. 3188. doi: 10.1038/s41598-020-59644-6
6. Koseniuk A., Smołucha G., Gurgul A., Szmatoła T., Oczkowicz M., Radko A. Differentiation of the domestic pig and wild boar using genotyping-by-sequencing. *Folia Biologica (Kraków).* 2023. V. 71. № 1. P. 1–11. doi: 10.3409/fb_71-1.01
7. Koseniuk A., Smołucha G., Natonek-Wiśniewska M., Radko A., Rubiś D. Differentiating Pigs from Wild Boars Based on *NR6A1* and *MC1R* Gene Polymorphisms. *Animals* (*Basel*). 2021. V. 11. № 7. P. 2123. doi: 10.3390/ani11072123
8. *Unipro UGENE*. URL: https://ugene.net/ (accessed 16.01.2024).
9. *Genome Analysis Toolkit.* URL: https://gatk.broadinstitute.org/hc/en-us (accessed 16.01.2024).
10. *ANNOVAR.* URL: https://annovar.openbioinformatics.org/en/latest/ (accessed 16.01.2024).
11. *SnpEff & SnpSift.* URL: http://pcingola.github.io/SnpEff/ (accessed 16.01.2024).
12. *Whole genome association analysis toolset.* URL: https://zzz.bwh.harvard.edu/plink/ (accessed 16.01.2024).
13. *SNPTEST*. URL: https://www.well.ox.ac.uk/~gav/snptest/ (accessed 16.01.2024).
14. *BioSample*. URL: https://www.ncbi.nlm.nih.gov/biosample (accessed 16.01.2024).
15. *BioProject.* URL: https://www.ncbi.nlm.nih.gov/bioproject (accessed 16.01.2024).
16. *Sequence Read Archive (SRA).* URL: https://www.ncbi.nlm.nih.gov/sra (accessed 16.01.2024).
17. *BGISEQ*. URL: https://www.bgi.com/global/overview/business-overview?i=3 (accessed 16.01.2024).
18. *Illumina sequencing platforms*. URL: https://www.illumina.com/systems/sequencing-platforms.html (accessed 16.01.2024).

19. *Roche Sequencing Solutions.* URL: https://diagnostics.roche.com/global/en/about/about-roche-sequencing-solutions.html (accessed 16.01.2024).
20. *PACBIO SMRT, Sequencing systems.* URL: https://www.pacb.com/sequencing-systems/ (accessed 16.01.2024).
21. *OXFORD NANOPORE.* URL: https://nanoporetech.com/products (accessed 16.01.2024).
22. Кипень В.Н., Снытков Е.В., Михайлова М.Е., Шейко Р.И. Дифференциация пород домашних свиней с использованием расширенного биоинформатического анализа SNP. *Доклады Национальной академии наук Беларуси*. 2022. Т. 66. № 3. С. 301–309. doi: 10.29235/1561-8323-2022-66-3-301-309
23. *Axiom™ Porcine Genotyping Array.* URL: https://www.thermofisher.com/order/catalog/product/550588 (accessed 16.01.2024).
24. *Oligo Calc: Oligonucleotide Properties Calculator.* URL: http://biotools.nubic.northwestern.edu/OligoCalc.html (accessed 16.01.2024).
25. Кипень В.Н., Михайлова М.Е., Снытков Е.В., Романишко Е.Л., Иванова Е.В., Шейко Р.И. Биоинформатический анализ геномов коммерческих пород домашних свиней для идентификации породоспецифичных SNP. *Известия Национальной академии наук Беларуси. Серия аграрных наук.* 2021. Т. 59. № 4. С. 464–476. doi: 10.29235/1817-7204-2021-59-4-464-476
26. Кипень В.Н., Иванова Е.В., Снытков Е.В., Верчук А.Н. Анализ полиморфизма гена гефестина (*HEPH*) на Х-хромосоме для установления принадлежности биологических образцов к диким или домашним представителям вида *Sus scrofa*. *Генетика*. 2020. Т. 56. № 9. С. 1054–1064. doi: 10.31857/S0016675820080068
27. Кипень В.Н., Рябцева А.О., Котова С.А., Журина Н.В., Ганджа А.И., Цыбовский И.С. Оценка интрогрессии генов свиньи домашней (*Sus scrofa domesticus*) в генофонд дикого кабана (*Sus scrofa scrofa*) на основе исследования полиморфизма генов MC1R и NR6A1. *Молекулярная и прикладная генетика*. 2019. Т. 26. С. 83–95.

Рукопись поступила в редакцию 29.04.23, переработанный вариант поступил 16.01.2024.
Дата опубликования 02.03.2024.

========== **BIOINFORMATICS** ==========

GENIS – methodological approach for *in silico* genotyping (validation on *Sus scrofa* sequencing results)

Kipen V.N., Snytkov E.V.

Institute of Genetics and Cytology of the National Academy of Sciences of Belarus, Minsk, f Belarus

Abstract. A universal methodological approach has been developed that allows solving the problem of differentiating closely related species using raw NGS sequencing data. The method is based on the use of single nucleotide polymorphisms (SNPs). This approach is universal; it can be used in the bioinformatic analysis of sequencing results, regardless of the biological species under study. The approach we developed is based on automating the process of searching for nucleotide sequences flanking the desired allele. The search is carried

out on the researcher's personal computer, does not require expensive equipment, the Python v.3.10 programming language and the Jupyter Notebook software development environment are free and publicly available. The methodological approach for *in silico* genotyping is implemented in the form of the GENIS software. Within the framework of this work, the program was tested on files with the results of genome sequencing of animals of the genus *Sus*. Revealed polymorphisms for the differentiation of pigs of the Duroc breed.

Key words: *Single Nucleotide Polymorphism, SNP, in silico genotyping, Sus scrofa, Python v.3.10, Jupyter Notebook.*