

Projekt zaliczeniowy

Analiza danych

Daria Plewa

Michał Humiński

Dominik Lisiecki

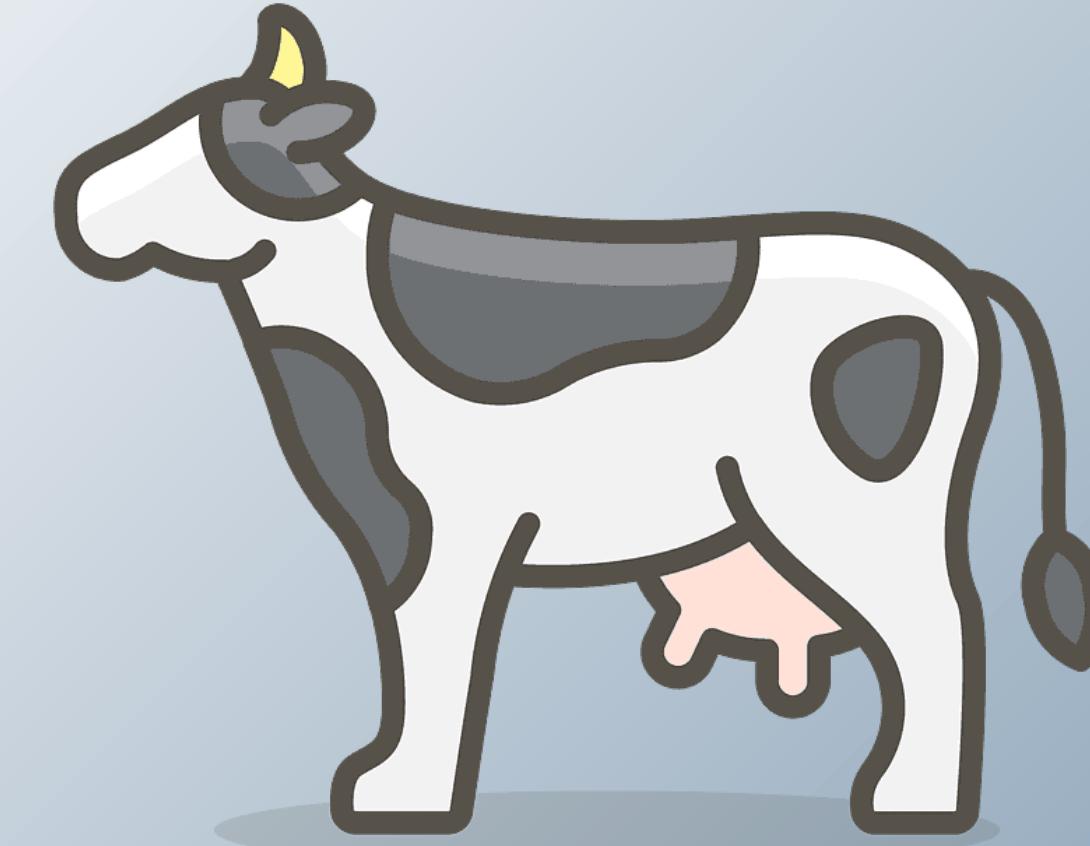
Bioinformatyka

studia licencjackie rok 3 2021/2022



Plan prezentacji

- analizowane dane
- cel badań
- hipotezy badawcze
- wykorzystane metody i programy
- wyniki i wnioski - analiza 1
- wyniki, wnioski, parallelizacja - analiza 2
- podsumowanie



creazilla.com

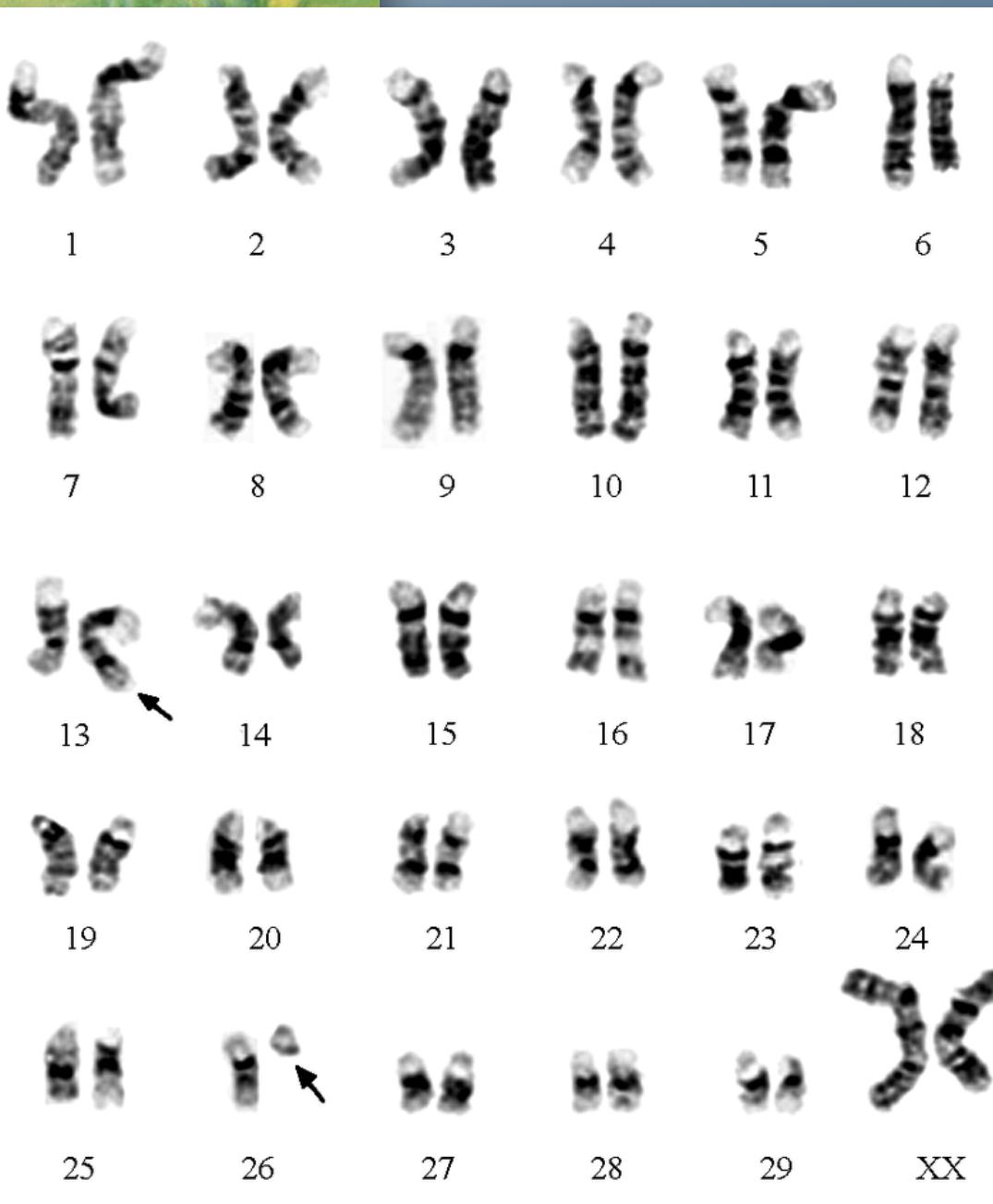
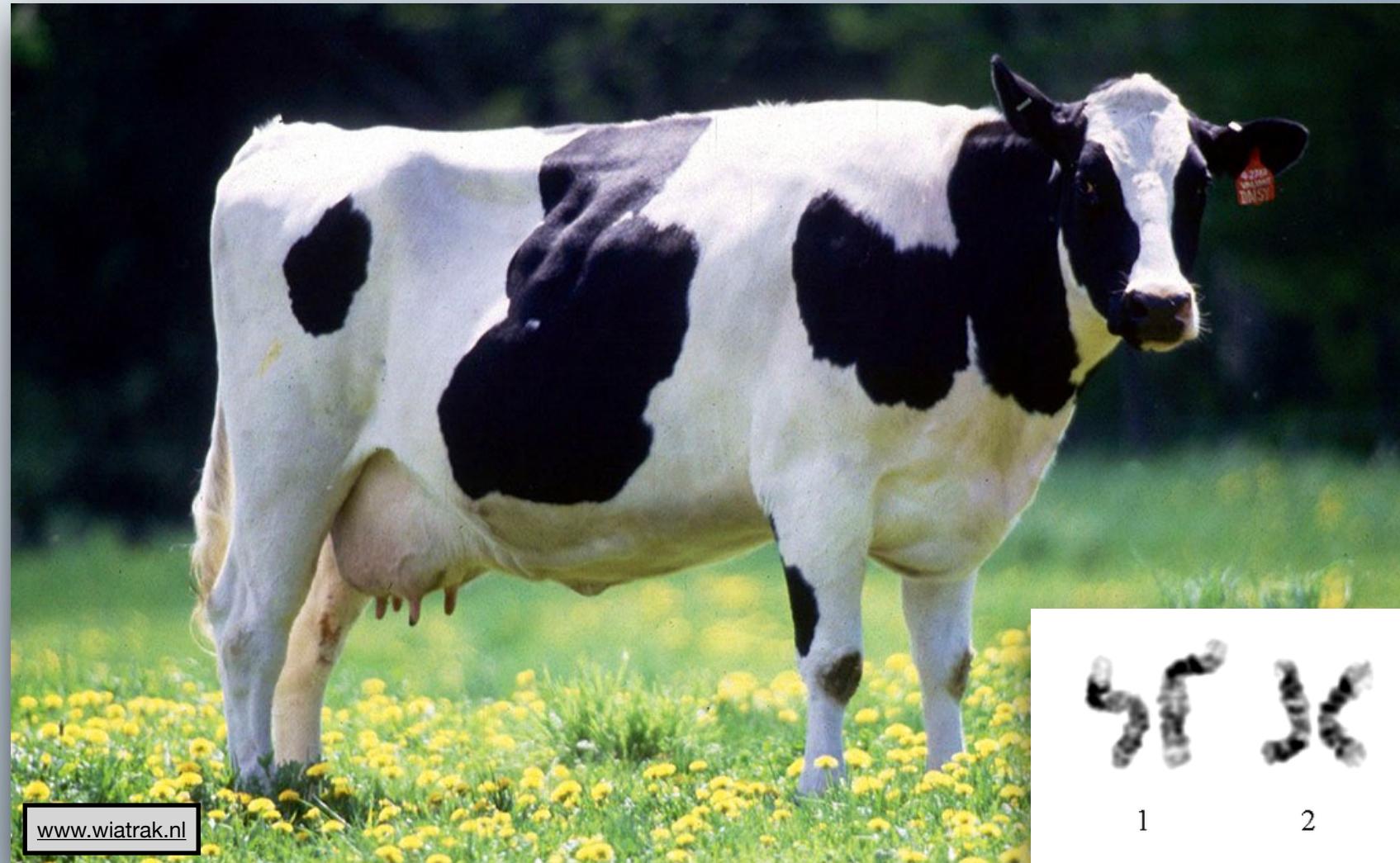


www.kindpng.com



www.kindpng.com

Analizowane dane

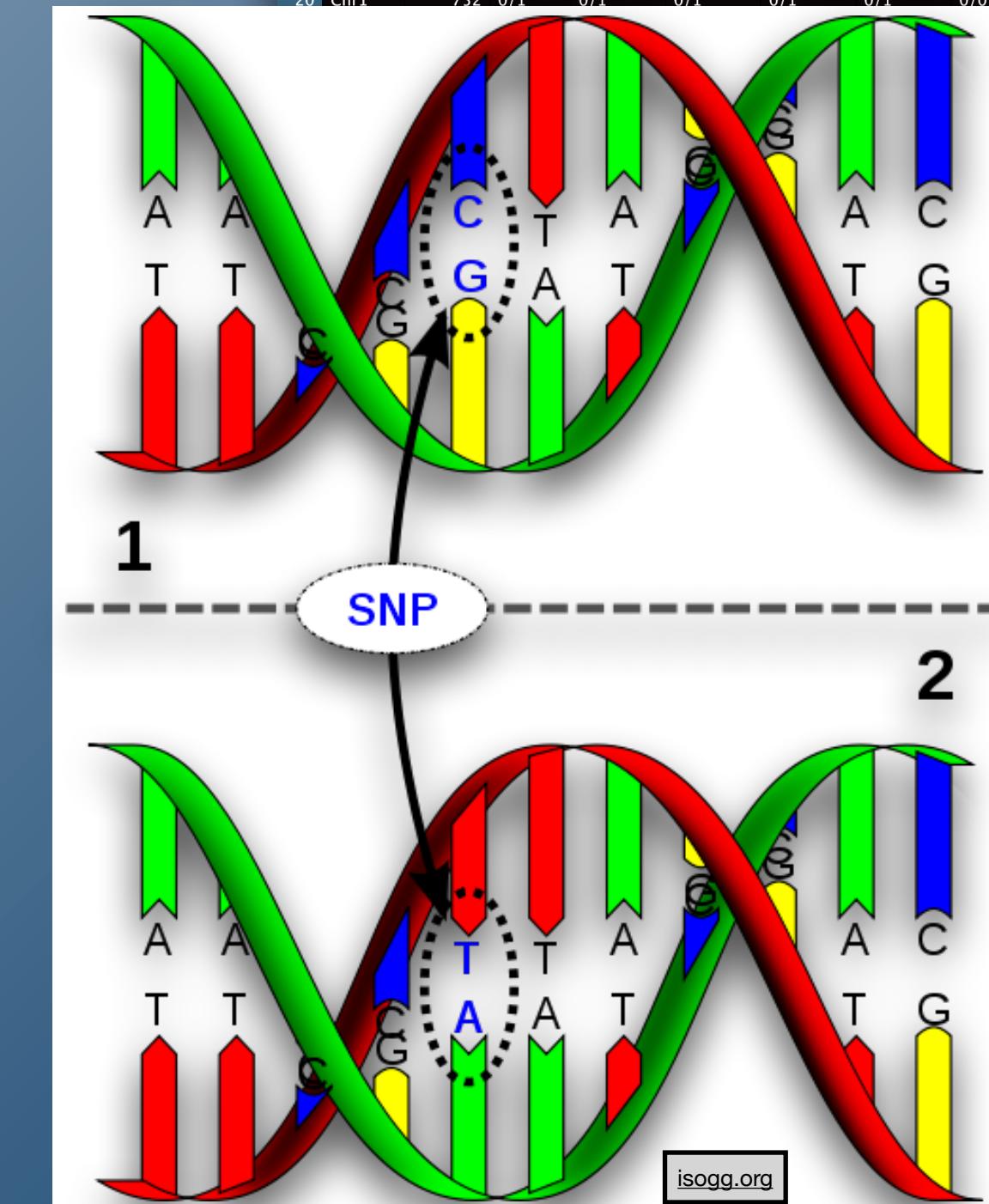


- ▶ Analizujemy dane dotyczące **krów rasy holsztyńsko-fryzyjskiej**.
- ▶ Do dyspozycji dostaliśmy **dane genetyczne krów zdrowych oraz krów chorych**.
- ▶ **Pełny kariotyp krowy składa się z **60 chromosomów**.**

Analizowane dane

- Do naszej dyspozycji otrzymaliśmy **4 pliki danych** w formacie .vcf.
- Analizowane przez nas pliki miały **ponad 14,3 mln rekordów** dla osobników chorych oraz **ponad 13,8 mln rekordów** dla osobników zdrowych.
- Otrzymane dane dotyczą **chromosomów** jednak zawierają także sekwencje kontigowe.

Chr1	x238	X0.1	X0.1.1	X0.1.2	X0.1.3	X0.1.4	X0.1.5	X0.1.6	X0.1.7	X0.1.8	X0.0	X0.1.9	X0.1.10	X1.1	X0.0.1	X0.1.11	X0.1.12
1 Chr1	300	0/1	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/0	0/0	0/1	0/1	0/1	0/1	0/1
2 Chr1	324	1/1	0/1	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
3 Chr1	340	0/1	0/1	0/0	0/1	0/0	0/0	0/1	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
4 Chr1	353	1/1	1/1	0/1	0/1	0/1	0/1	0/1	0/1	1/1	0/1	0/1	0/1	1/1	1/1	1/1	0/1
5 Chr1	355	1/1	1/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	1/1	1/1	0/1
6 Chr1	357	0/1	0/1	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	1/1	0/1	0/0
7 Chr1	380	0/1	0/1	0/0	0/1	0/1	0/0	0/1	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
8 Chr1	420	1/1	1/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
9 Chr1	435	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
10 Chr1	512	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
11 Chr1	571	1/1	1/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	1/1	0/1	0/1
12 Chr1	628	0/0	0/0	0/0	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/1	0/1	0/0
13 Chr1	637	0/1	0/1	0/0	0/1	0/1	0/0	0/1	0/0	0/1	0/1	0/1	0/0	0/1	0/1	0/1	0/1
14 Chr1	638	0/0	0/0	0/0	0/1	0/0	0/0	0/1	0/0	0/1	0/0	0/1	0/1	0/0	0/0	0/1	0/1
15 Chr1	695	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
16 Chr1	696	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
17 Chr1	734	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
18 Chr1	738	0/1	0/1	0/1	0/1	0/1	0/1	0/1	1/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
19 Chr1	751	0/1	0/1	0/1	0/1	0/1	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1
20 Chr1	752	0/1	0/1	0/1	0/1	0/1	0/1	0/0	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1



Źródło własne

Cel badań

Analiza zawartości mutacji typu SNP w chromosomach osobników krów rasy holsztyńsko-fryzyjskiej



Znalezienie mutacji typu SNP mogących mieć podłożę biologiczne dla rozwoju choroby i wyznaczenie ich zależności z wybranymi parametrami dla każdego chromosomu

Hipotezy badawcze

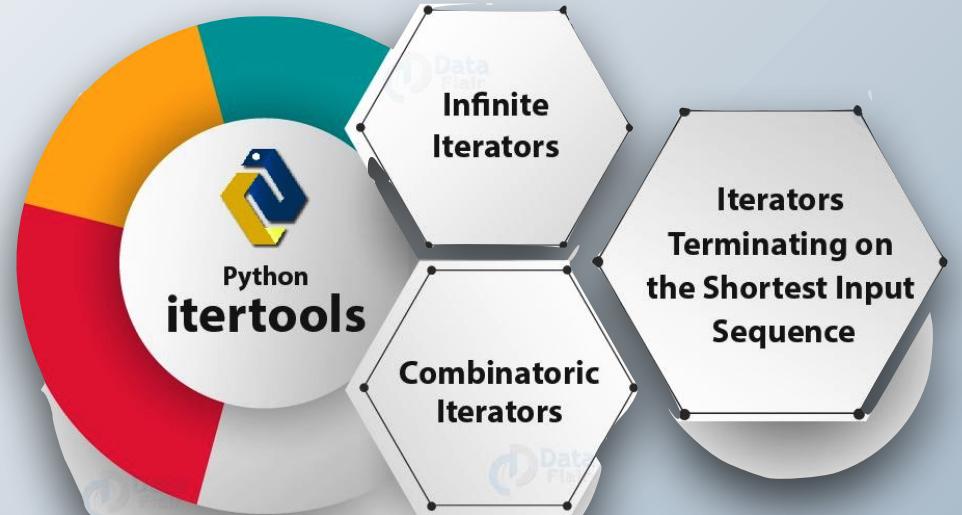
- Zawartość mutacji typu SNP zależy od długości chromosomu



creazilla.com

- Liczba SNP istotnych biologicznie zależy od długości chromosomu
- Liczba SNP istotnych biologicznie zależy od liczby znalezionych SNP dla każdego chromosomu

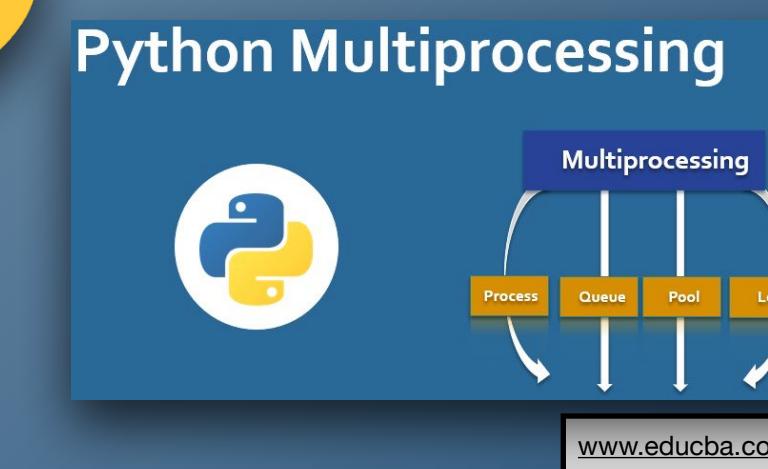
Wykorzystane metody i programy



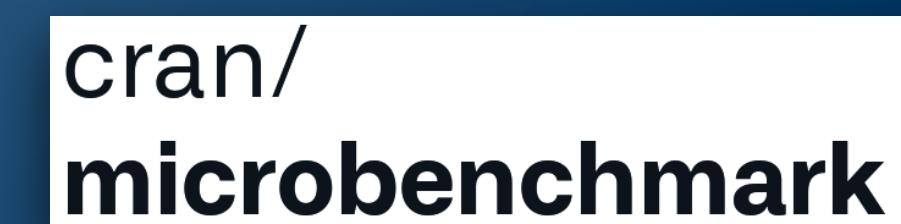
data-flair.training



en.wikipedia.org



commons.wikimedia.org



englelab.gatech.edu

Wykorzystane metody i programy

- Podczas naszych analiz użyliśmy wielu ciekawych funkcji **zaimplementowanych** do języków programowania lub też zawartych w **dodatkowych paczkach**.

- ❖ *itertools.zip_longest*
- ❖ *multiprocessing.Pool()*
- ❖ *map()*
- ❖ *pd.groupby()*
- ❖ *pd.merge()*
- ❖ *chi2_contingency()*



- ❖ *filter()*
- ❖ *inner_join()*
- ❖ *rowSums()*
- ❖ *rbind()*
- ❖ *chisq.test()*

Przygotowanie danych

Organism Overview ; Genome Assembly and Annotation report [6] ; Organelle Annotation Report [3] ID: 82

Bos taurus (cattle)
COW

Lineage: Eukaryota[8778]; Metazoa[4257]; Chordata[2107]; Craniata[2084]; Vertebrata[2084]; Euteleostomi[2067]; Mammalia[641]; Eutheria[467]; Laurasiatheria[263]; Artiodactyla[127]; Ruminantia[89]; Pecora[87]; Bovidae[63]; Bovinae[19]; Bos[8]; Bos taurus[1]

Bos taurus (cow) is an agriculturally important animal; beef and milk production are the largest manufacturing industries in the United States. The cow is an important model organism for health research in obesity, female health, and infectious diseases. Cow is also used in studies of endocrinology, physiology and reproductive techniques. The [More...](#)

Reference genome:
Bos taurus ARS-UCD1.2
Submitter: USDA ARS

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Proteins
Chr	1	NC_037328.1	CM008168.2	158.53	40.1	2,511	
Chr	2	NC_037329.1	CM008169.2	136.23	40.7	2,711	
Chr	3	NC_037330.1	CM008170.2	121.01	41.8	3,611	
Chr	4	NC_037331.1	CM008171.2	120	40.5	2,011	
Chr	5	NC_037332.1	CM008172.2	120.09	41.7	3,511	
Chr	6	NC_037333.1	CM008173.2	117.81	39.9	1,811	
Chr	7	NC_037334.1	CM008174.2	110.68	41.8	3,511	
Chr	8	NC_037335.1	CM008175.2	113.32	41.2	2,011	
Chr	9	NC_037336.1	CM008176.2	105.45	40.0	1,485	1 41 454 883 150
Chr	10	NC_037337.1	CM008177.2	103.31	41.4	2,509	- 65 533 1,445 161
Chr	11	NC_037338.1	CM008178.2	106.98	42.9	3,025	- 44 563 1,423 136
Chr	12	NC_037339.1	CM008179.2	87.22	40.4	1,014	- 50 361 674 81
Chr	13	NC_037340.1	CM008180.2	83.47	43.7	2,342	- 44 497 1,185 105
Chr	14	NC_037341.1	CM008181.2	82.4	41.3	1,387	- 37 351 784 92
Chr	15	NC_037342.1	CM008182.2	85.01	41.7	2,309	- 48 332 1,568 304
Chr	16	NC_037343.1	CM008183.2	81.01	42.6	1,903	- 32 475 1,018 120
Chr	17	NC_037344.1	CM008184.2	73.17	42.3	1,902	- 45 358 897 97
Chr	18	NC_037345.1	CM008185.2	65.82	45.4	3,438	- 51 557 1,727 173
Chr	19	NC_037346.1	CM008186.2	63.45	46.0	3,561	1 70 705 1,744 124
Chr	20	NC_037347.1	CM008187.2	71.97	41.0	776	1 31 262 576 97
Chr	21	NC_037348.1	CM008188.2	69.86	43.0	1,520	2 26 549 990 126
Chr	22	NC_037349.1	CM008189.2	60.77	43.4	1,966	1 25 370 810 72
Chr	23	NC_037350.1	CM008190.2	52.5	43.3	1,867	- 201 390 1,230 103
Chr	24	NC_037351.1	CM008191.2	62.32	41.8	974	- 19 250 518 72
Chr	25	NC_037352.1	CM008192.2	42.35	47.1	2,003	- 67 322 1,006 58
Chr	26	NC_037353.1	CM008193.2	51.99	42.9	1,362	1 14 302 606 80
Chr	27	NC_037354.1	CM008194.2	45.61	41.7	657	3 36 187 429 42
Chr	28	NC_037355.1	CM008195.2	45.94	42.1	983	- 24 170 481 69
Chr	29	NC_037356.1	CM008196.2	51.1	44.2	1,732	- 24 308 962 117
Chr	X	NC_037357.1	CM008197.2	139.01	40.4	2,471	- 75 491 1,555 333
MT		NC_006853.1	CM008198.1	0.02	39.4	13	2 22 - 37 -
Un	-	.	-	87.44	45.8	522	4 39 171 709 205

Chromosomes

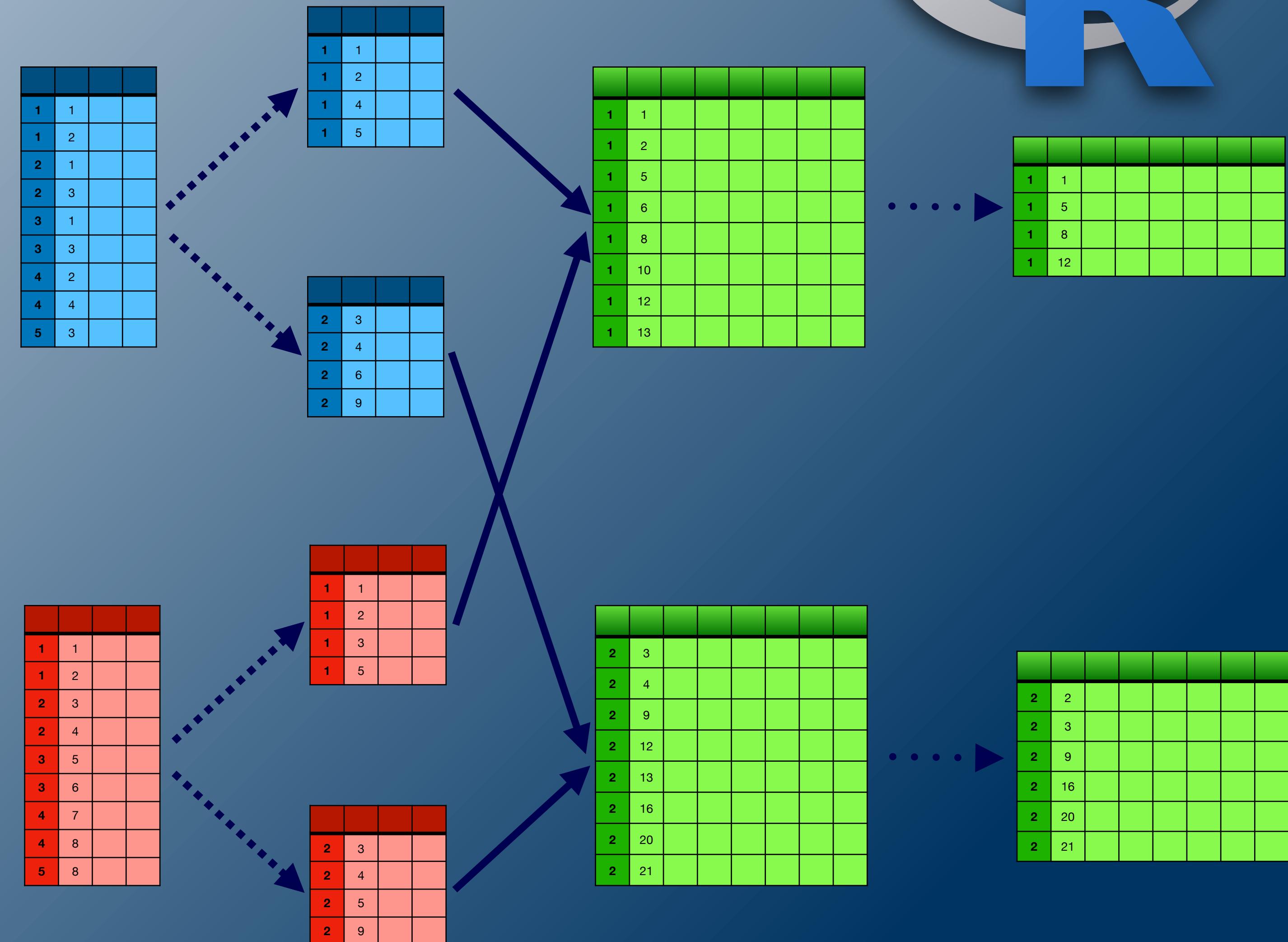
Žródło własne

- W pierwszej analizie wykorzystaliśmy **dane** pobrane z **bazy NCBI**.
- Napisany przez nas **skrypt** operował na podstawowych operacjach matematycznych.

Przygotowanie danych

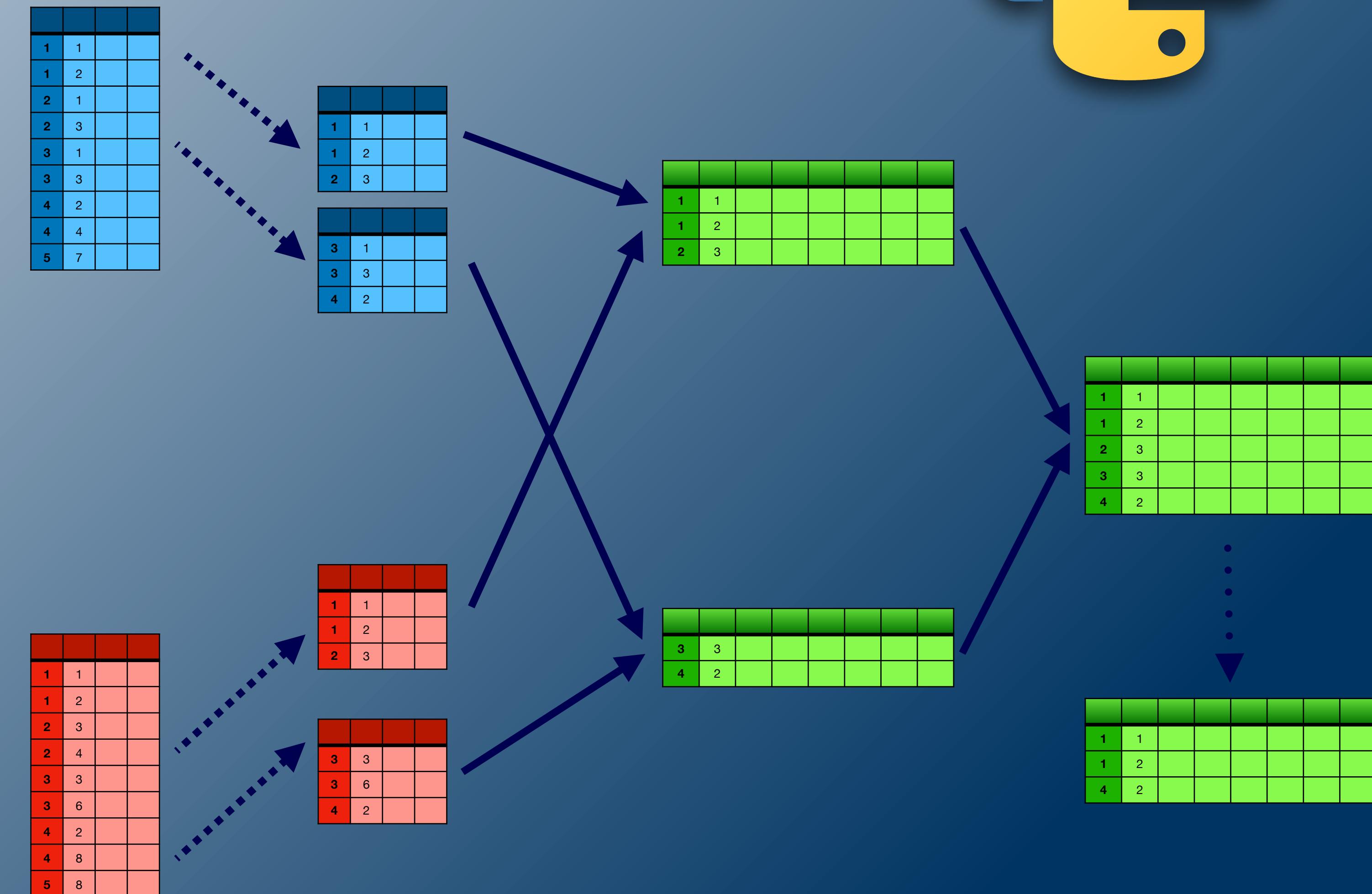
commons.wikimedia.org

- ▶ Zdecydowaliśmy się **podzielić nasze wyniki na tabele z informacjami dotyczącymi poszczególnych chromosomów dla badanych grup organizmów.**
- ▶ Następnie musielismy **porównać tabele odpowiadających sobie chromosomów względem badanej mutacji.**
- ▶ Ostatnim punktem przygotowań była **filtracja niepotrzebnych danych.**



Przygotowanie danych

- ▶ Zdecydowaliśmy się **podzielić nasze wyniki na tabele z informacjami dotyczącymi poszczególnych chromosomów dla badanych grup organizmów.**
- ▶ Następnie musielismy **porównać tabele odpowiadających sobie chromosomów względem badanej mutacji.**
- ▶ Ostatnim punktem przygotowań była **filtracja niepotrzebnych danych.**

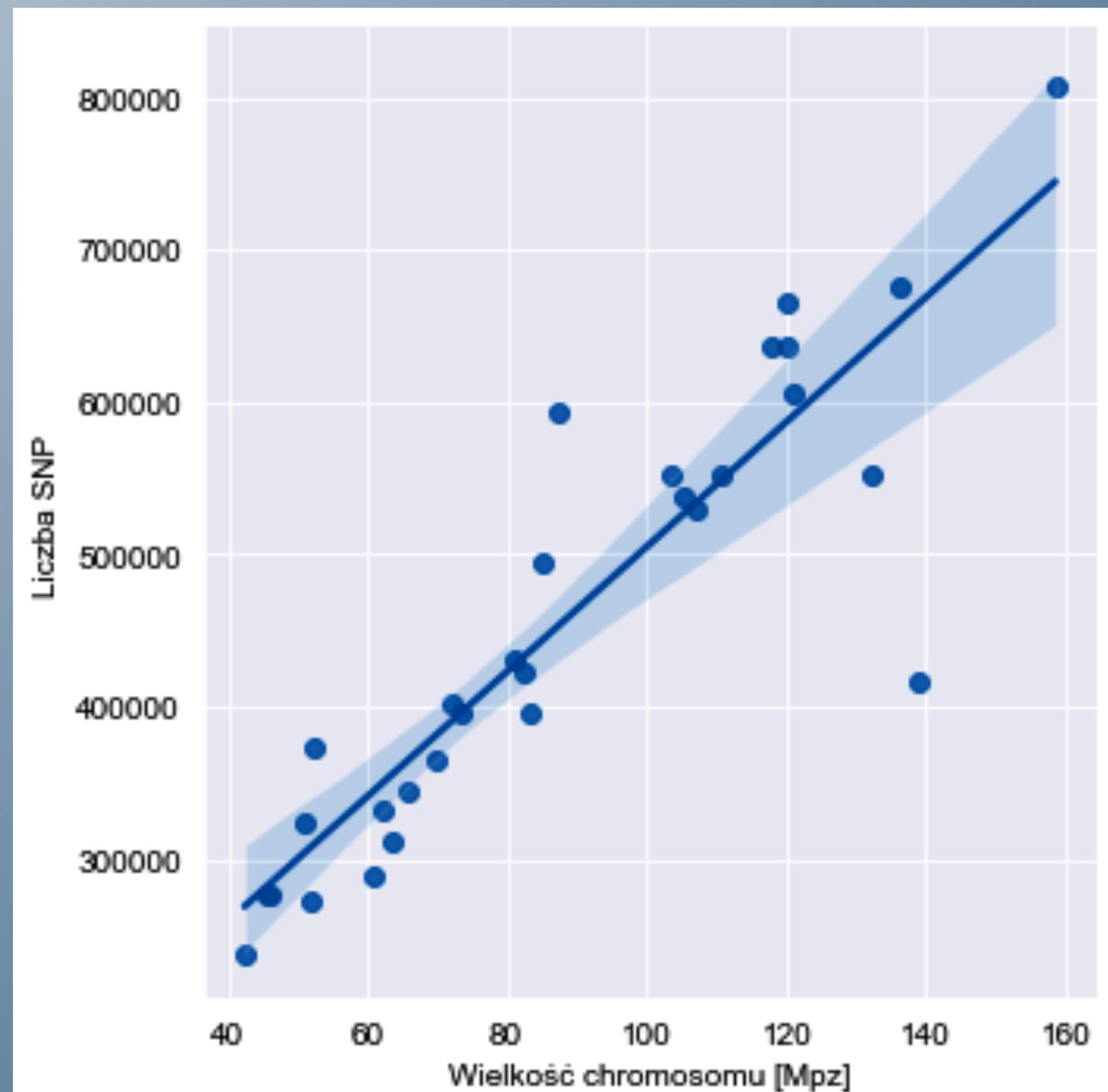


Wyniki i whioski

Wyniki - analiza 1

Zależność rozmiaru chromosomów od ilości mutacji SNP

Chromosom	Długość (Mpz)	Liczba SNP	Stosunek liczby mutacji do rozmiaru
1	158,5	806755	0,005089
2	136,2	676315	0,004965
3	121,0	605775	0,005006
4	120,0	664896	0,005541
5	120,1	635713	0,005294
6	117,8	637329	0,005410
7	110,7	551832	0,004986
8	132,3	551846	0,004171
9	105,5	537541	0,005098
10	103,3	551753	0,005341
11	107,0	529621	0,004951
12	87,2	593968	0,006810
13	83,5	395424	0,004737
14	82,4	423310	0,005137
15	85,0	494062	0,005812
16	81,0	430224	0,005311
17	73,2	396718	0,005422
18	65,8	343897	0,005225
19	63,5	311487	0,004909
20	72,0	402238	0,005589
21	69,9	365278	0,005229
22	60,8	289773	0,004768
23	52,5	373201	0,007109
24	62,3	331286	0,005316
25	42,4	237860	0,005617
26	52,0	271690	0,005226
27	45,6	275527	0,006041
28	45,9	275446	0,005996
29	51,1	323352	0,006328
X	139,0	415994	0,002993



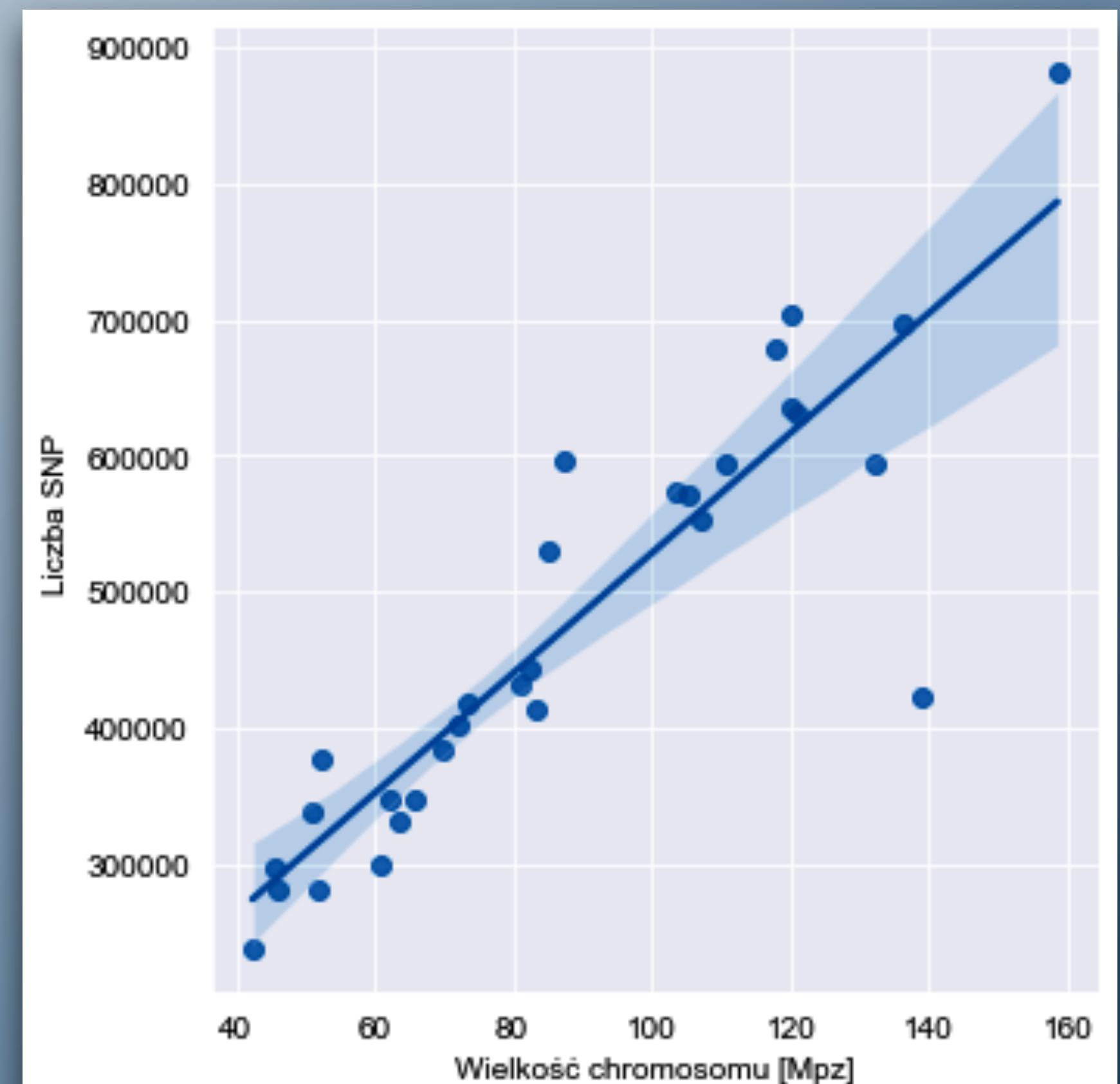
- ▶ Wyniki ukazują **silną korelację** pomiędzy liczbą znalezionych SNP oraz wielkością chromosomu u **zdrowych osobników**.
- ▶ Wartość współczynnika korelacji r-Pearsona wynosi **0,8964**.



Wyniki - analiza 1

Zależność rozmiaru chromosomów od ilości mutacji SNP

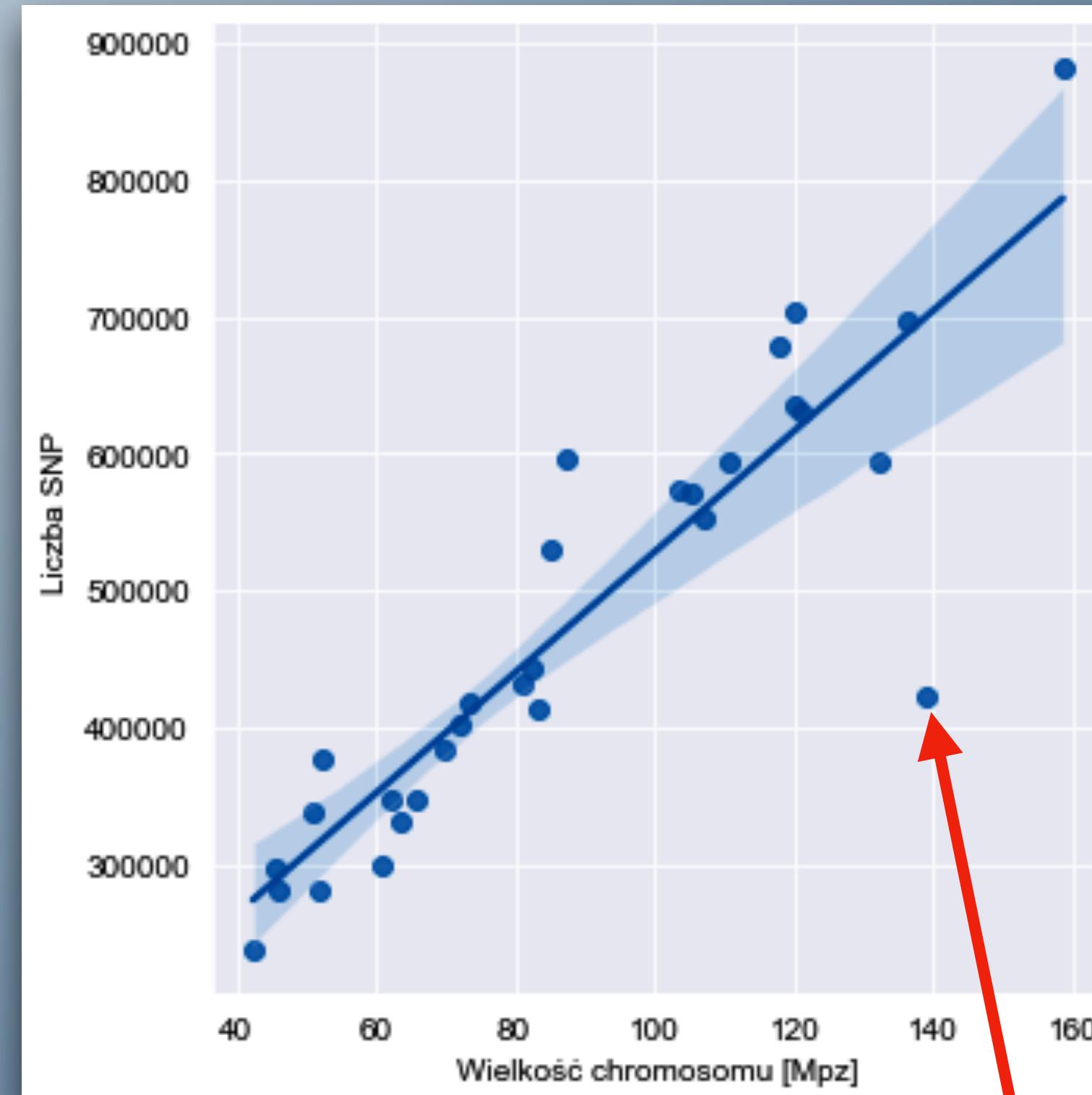
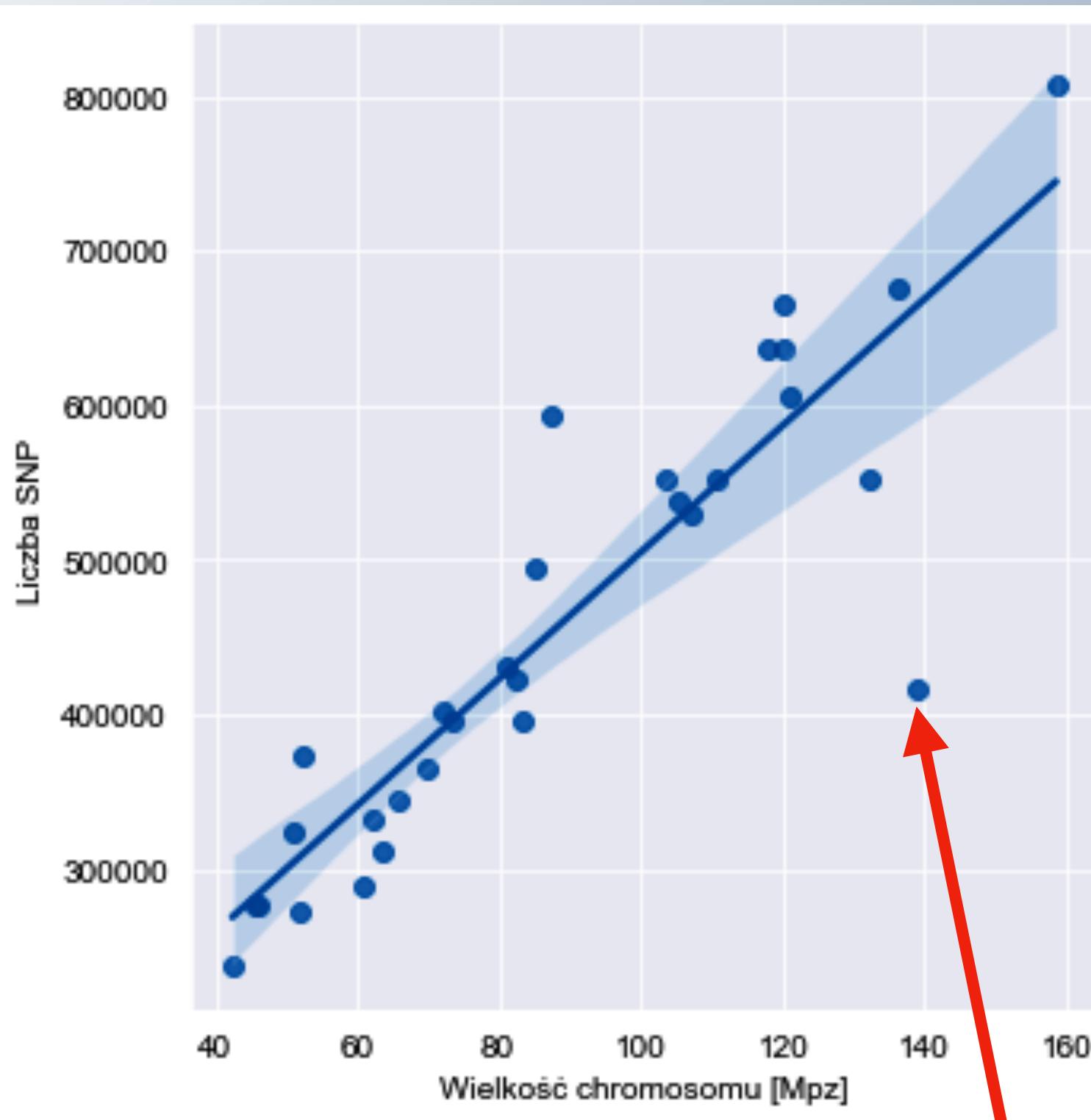
Chromosom	Długość (Mpz)	Liczba SNP	Stosunek liczby mutacji do rozmiaru
1	158,5	880814	0,005556
2	136,2	697319	0,005119
3	121,0	630235	0,005208
4	120,0	704037	0,005867
5	120,1	636185	0,005298
6	117,8	679293	0,005766
7	110,7	593746	0,005365
8	132,3	593242	0,004483
9	105,5	571778	0,005422
10	103,3	572849	0,005545
11	107,0	551783	0,005158
12	87,2	595146	0,006824
13	83,5	413776	0,004957
14	82,4	442273	0,005367
15	85,0	530278	0,006238
16	81,0	431407	0,005325
17	73,2	418412	0,005718
18	65,8	346083	0,005258
19	63,5	331040	0,005217
20	72,0	401376	0,005577
21	69,9	384387	0,005502
22	60,8	299261	0,004924
23	52,5	376643	0,007174
24	62,3	347851	0,005582
25	42,4	238131	0,005623
26	52,0	281703	0,005418
27	45,6	297330	0,006519
28	45,9	281125	0,006119
29	51,1	337635	0,006607
X	139,0	423234	0,003045



- ▶ Wyniki ukazują **silną korelację** pomiędzy liczbą znalezionych SNP oraz wielkością chromosomu u chorych osobników.
- ▶ Wartość współczynnika korelacji r-Pearsona wynosi **0,8983**.



Wnioski - analiza 1



- Należy jednak zauważyc interesujący przypadek. To wartość, która w znacznym stopniu odstaje od naszej prostej regresji - **chromosom płci**.

- Dla **obydwu grup** dostaliśmy bardzo podobne wyniki.
- Mozemy stwierdzić, że im dłuzszy chromosom analizujemy, tym więcej odnajdziemy SNP.

Wnioski - analiza 1

- ▶ Zawartość mutacji typu SNP zależy od długości chromosomu



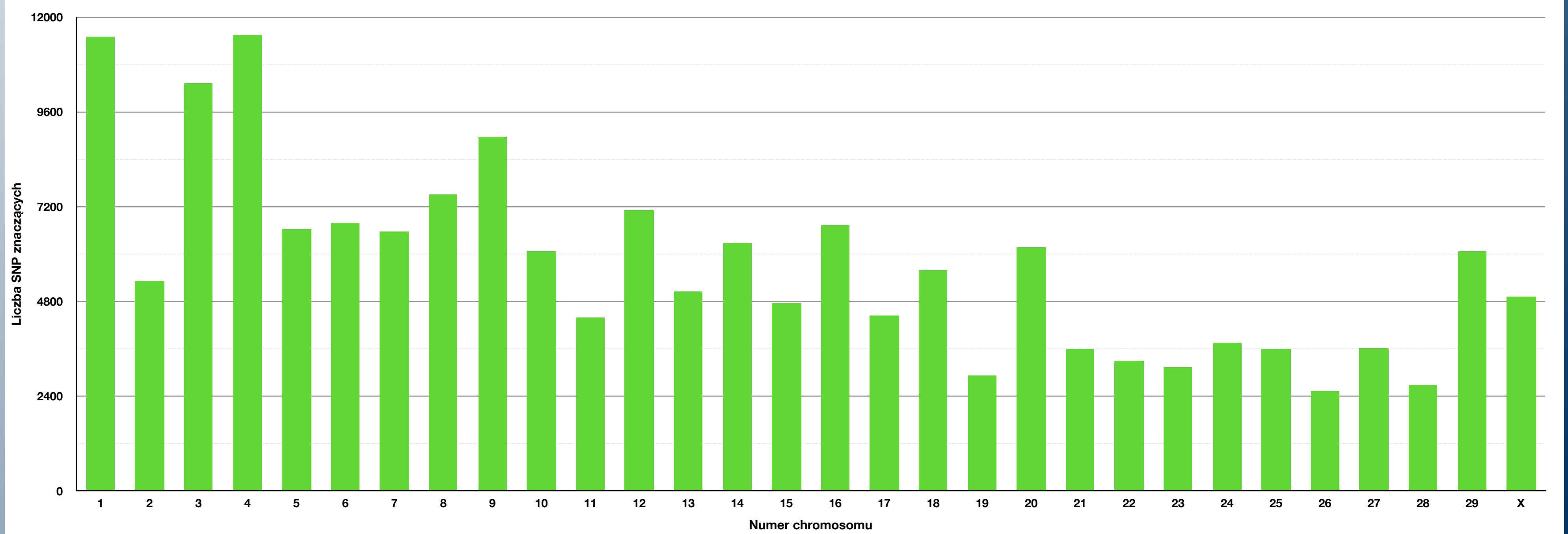
- ▶ Na podstawie długości chromosomu **можемy oszacować** zawartą w jego nici DNA liczbę mutacji typu SNP.

Wyniki - analiza 2



Analiza SNP istotnych statystycznie

Chromosom	Liczba mutacji znaczących
1	11509
2	5321
3	10325
4	11564
5	6624
6	6798
7	6569
8	7495
9	8959
10	6058
11	4397
12	7101
13	5062
14	6275
15	4749
16	6726
17	4435
18	5600
19	2928
20	6176
21	3590
22	3307
23	3134
24	3761
25	3602
26	2529
27	3622
28	2697
29	6078
X	4928

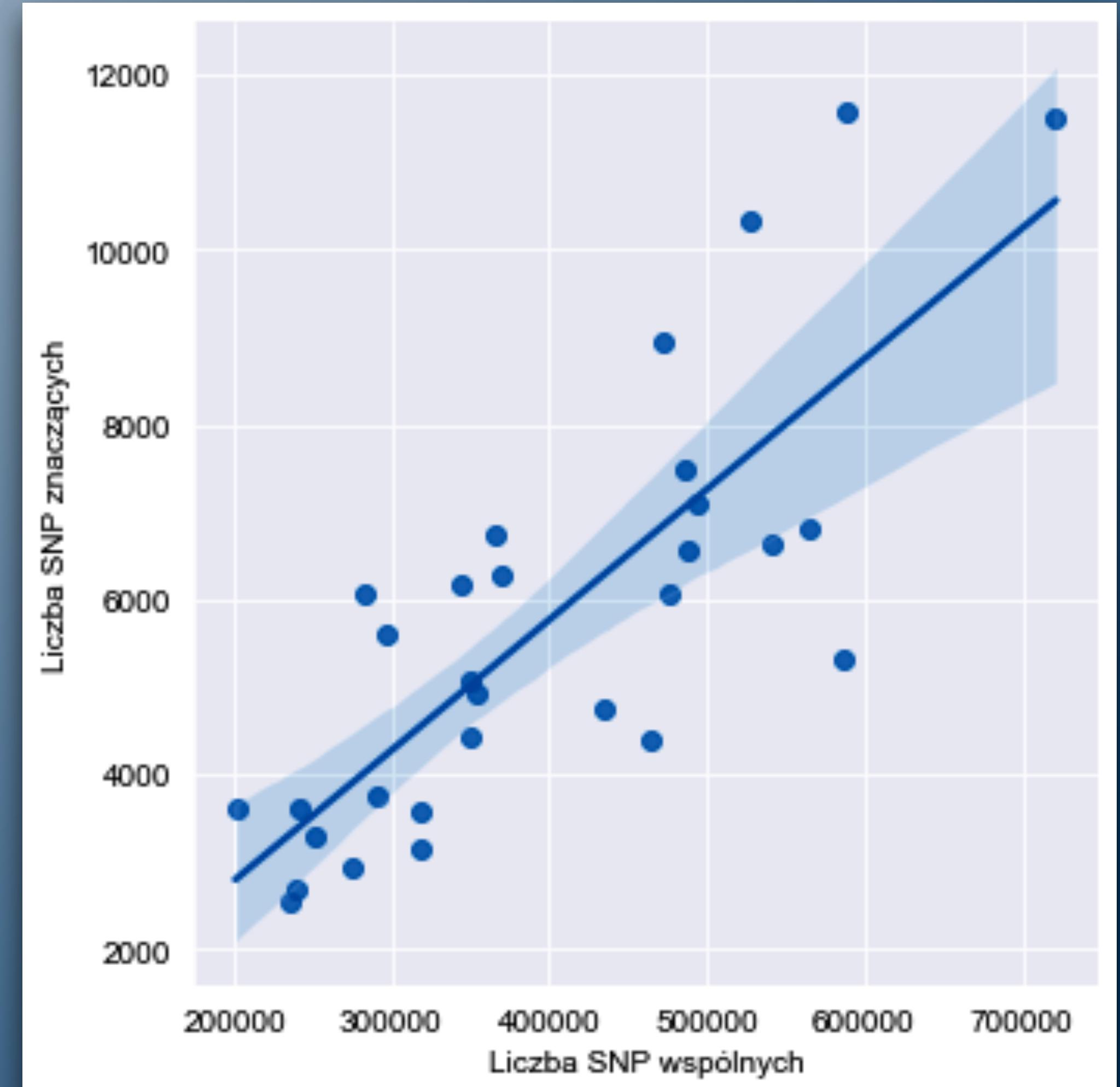


- ▶ Dla każdej mutacji wykonywaliśmy **test chi-kwadrat z poprawką Yatesa**.
- ▶ **Najwięcej miejsc SNP**, które odznaczają się innym stosunkiem genotypów, posiada **chromosom 4**, a **najmniej chromosom 26**.

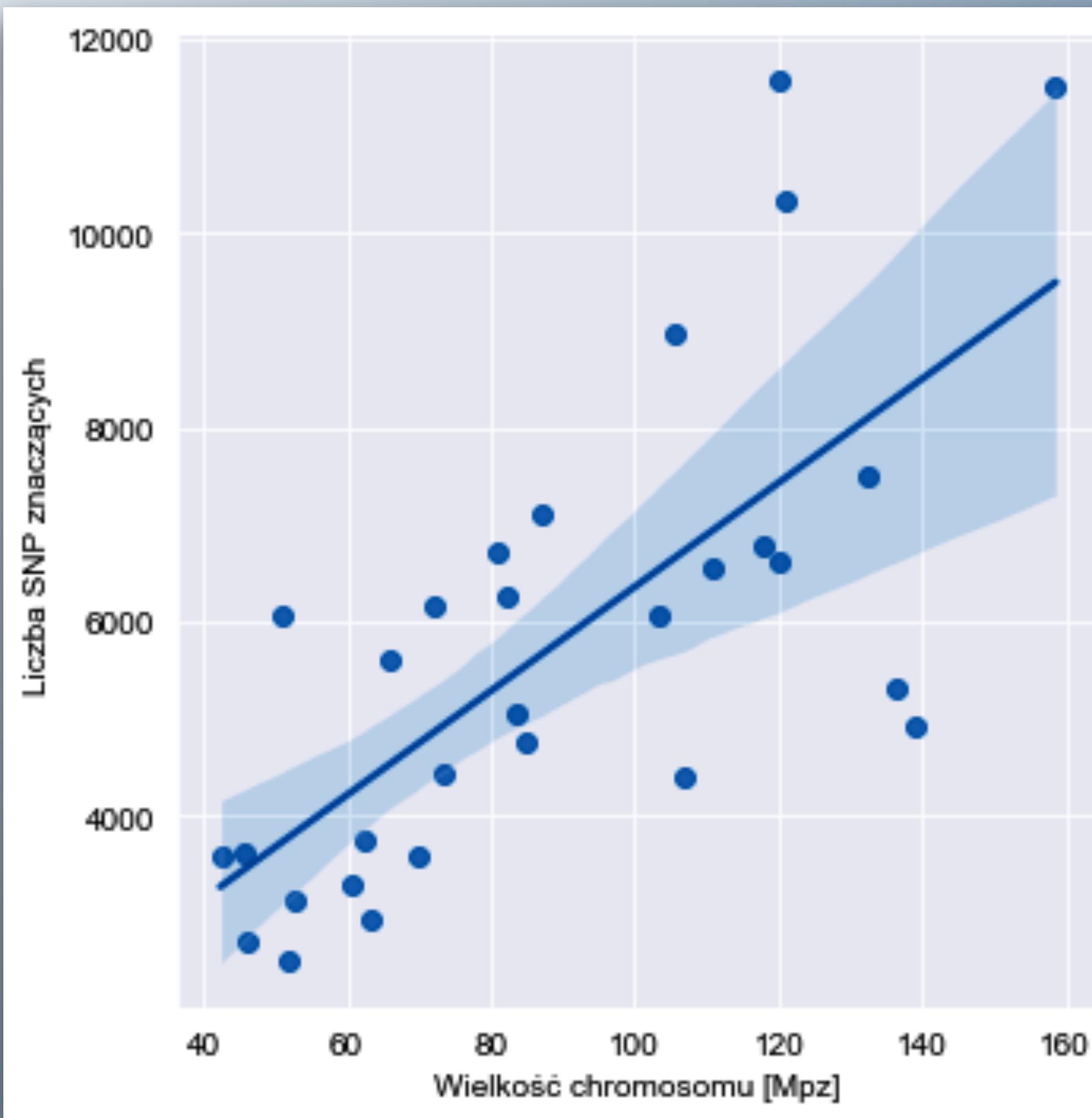
Wyniki - analiza 2



- ▶ Najpierw chcieliśmy przeanalizować czy **odsetek znalezionych mutacji typu SNP jest identyczny** względem ich **początkowej badanej liczby** dla każdego chromosomu.
- ▶ Otrzymany wynik testu korelacji r-Pearsona wynosi **0,8015**.



Wyniki - analiza 2



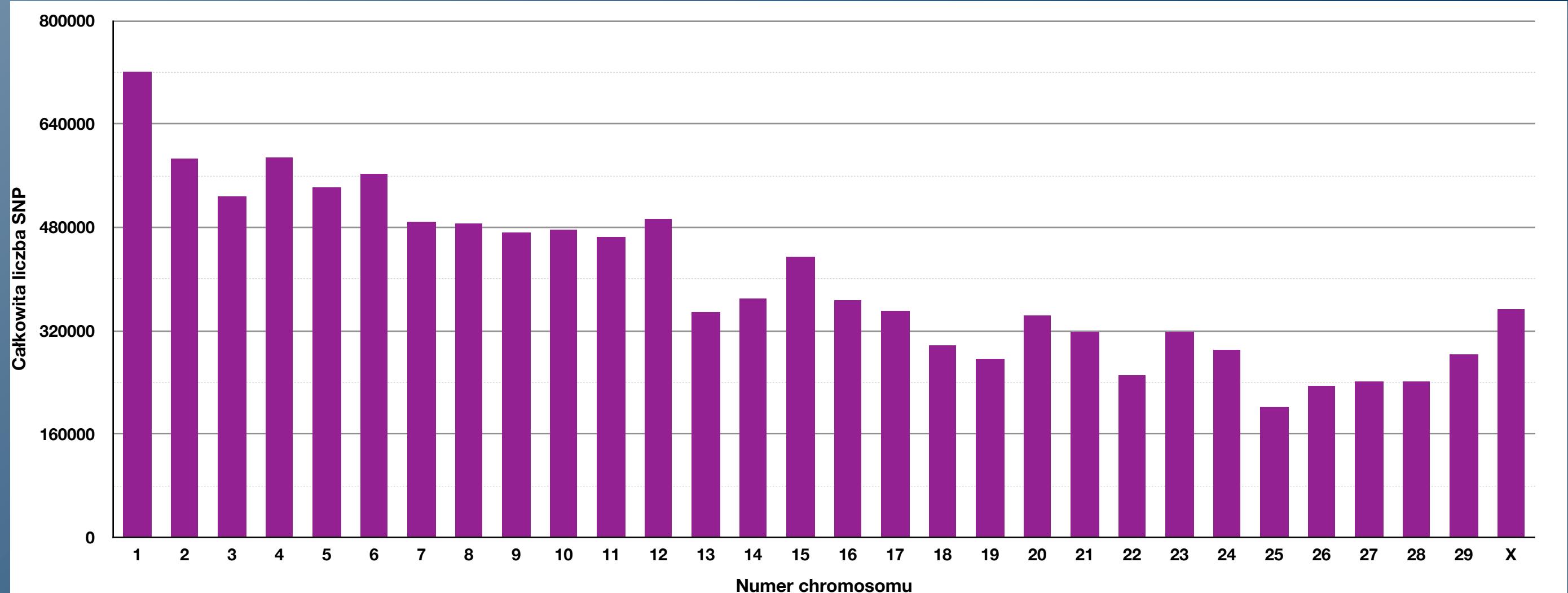
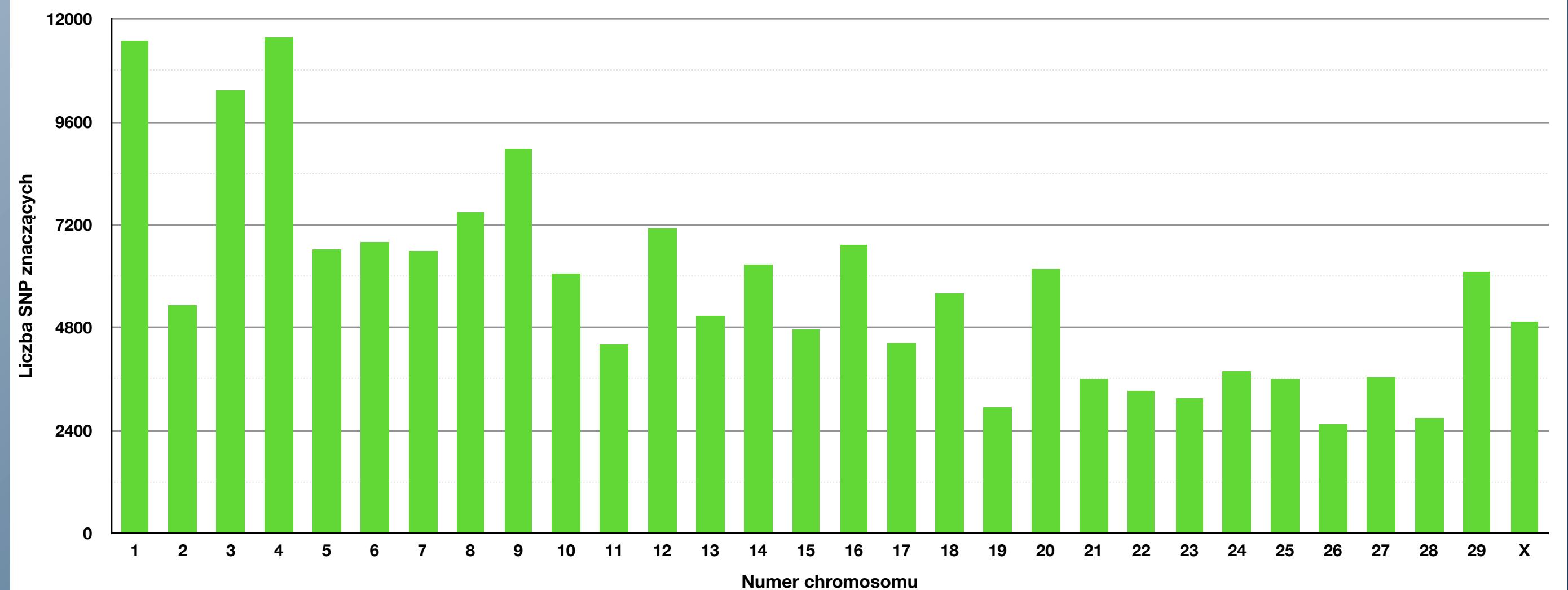
- ▶ Następna analiza dotyczyła zależności między **liczbą znalezionych mutacji typu SNP a długością chromosomu.**
- ▶ Otrzymany wynik testu korelacji r-Pearsona wynosi **0,7119**.

Wnioski - analiza 2

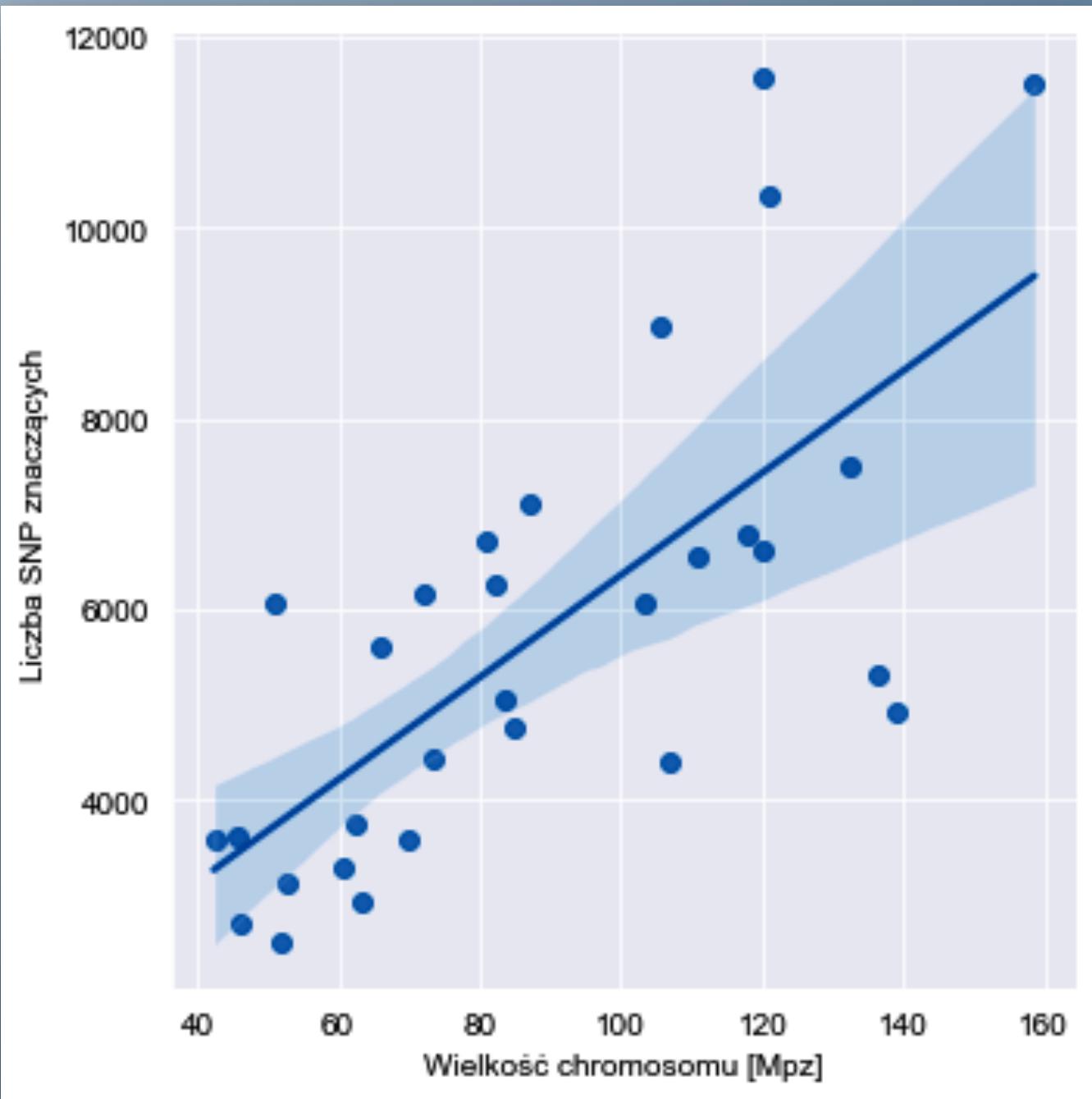
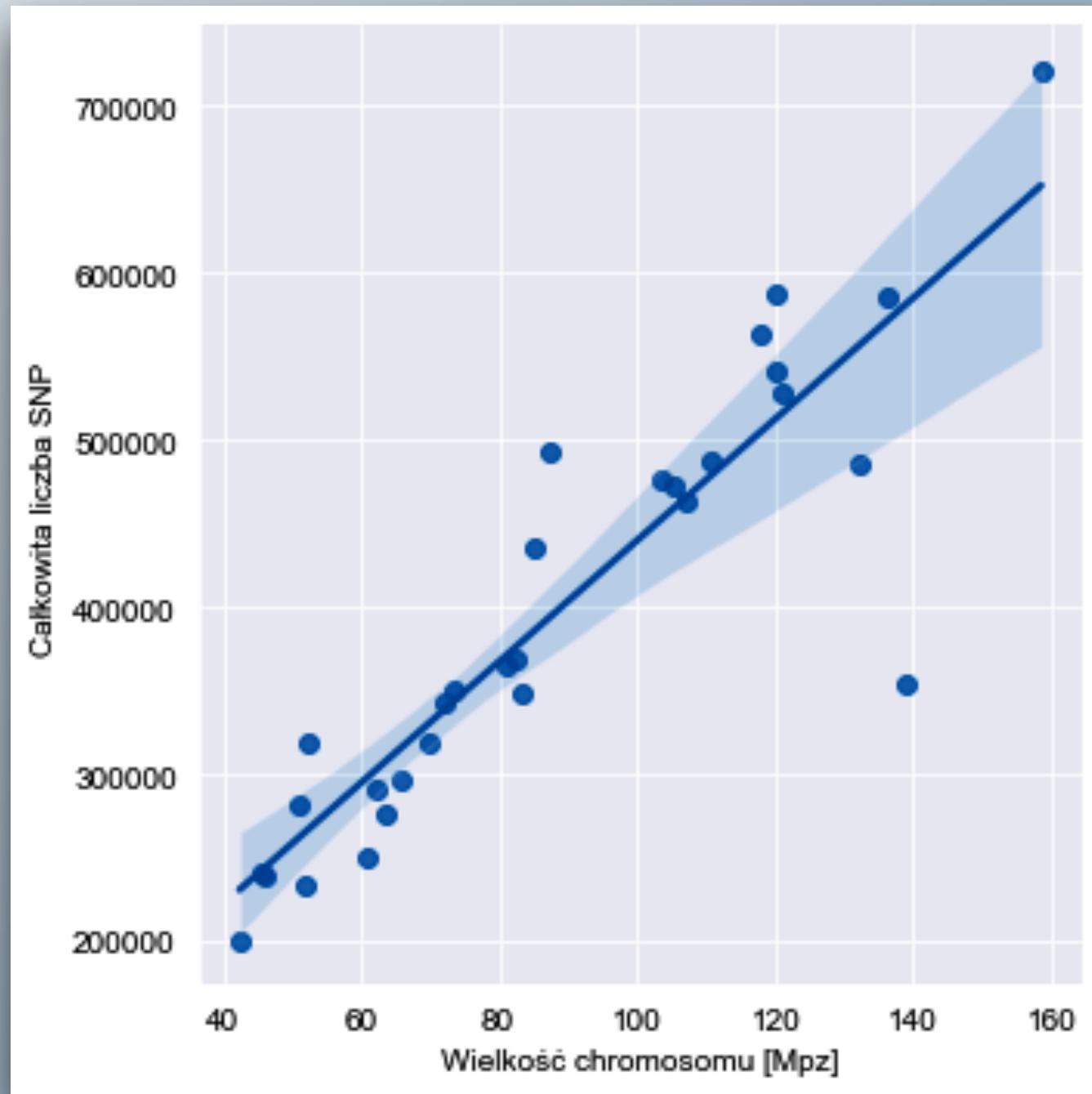


- ▶ Możemy stwierdzić **niewielkie różnice w zawartości ogólnej liczby SNP i SNP znaczących biologicznie na naszych chromosomach.**
- ▶ Widzimy **inny rozkład** naszych danych w obydwu przypadkach.

Analiza SNP	
Chromosom	Odsetek mutacji znaczących
1	0,0160
2	0,0091
3	0,0196
4	0,0197
5	0,0122
6	0,0121
7	0,0135
8	0,0154
9	0,0190
10	0,0127
11	0,0095
12	0,0144
13	0,0145
14	0,0170
15	0,0109
16	0,0184
17	0,0127
18	0,0189
19	0,0106
20	0,0180
21	0,0113
22	0,0132
23	0,0098
24	0,0129
25	0,0179
26	0,0108
27	0,0150
28	0,0112
29	0,0215
X	0,0139



Wnioski - analiza 2



- ▶ Dla obu zależności otrzymaliśmy **silne korelacje dodatnie**.
- ▶ Możemy wnioskować, że wielkość chromosomu jest **lepszym predyktorem ogólnej liczby SNP** niż liczby SNP, w których obserwujemy różnice genotypowe.

Wnioski - analiza 2



- *Liczba SNP istotnych biologicznie zależy od długości chromosomu*



- *Liczba SNP istotnych biologicznie zależy od liczby znalezionych SNP dla każdego chromosomu*



▶ Znając liczbę istotnych biologicznie SNP **możemy oszacować** długość chromosomu.

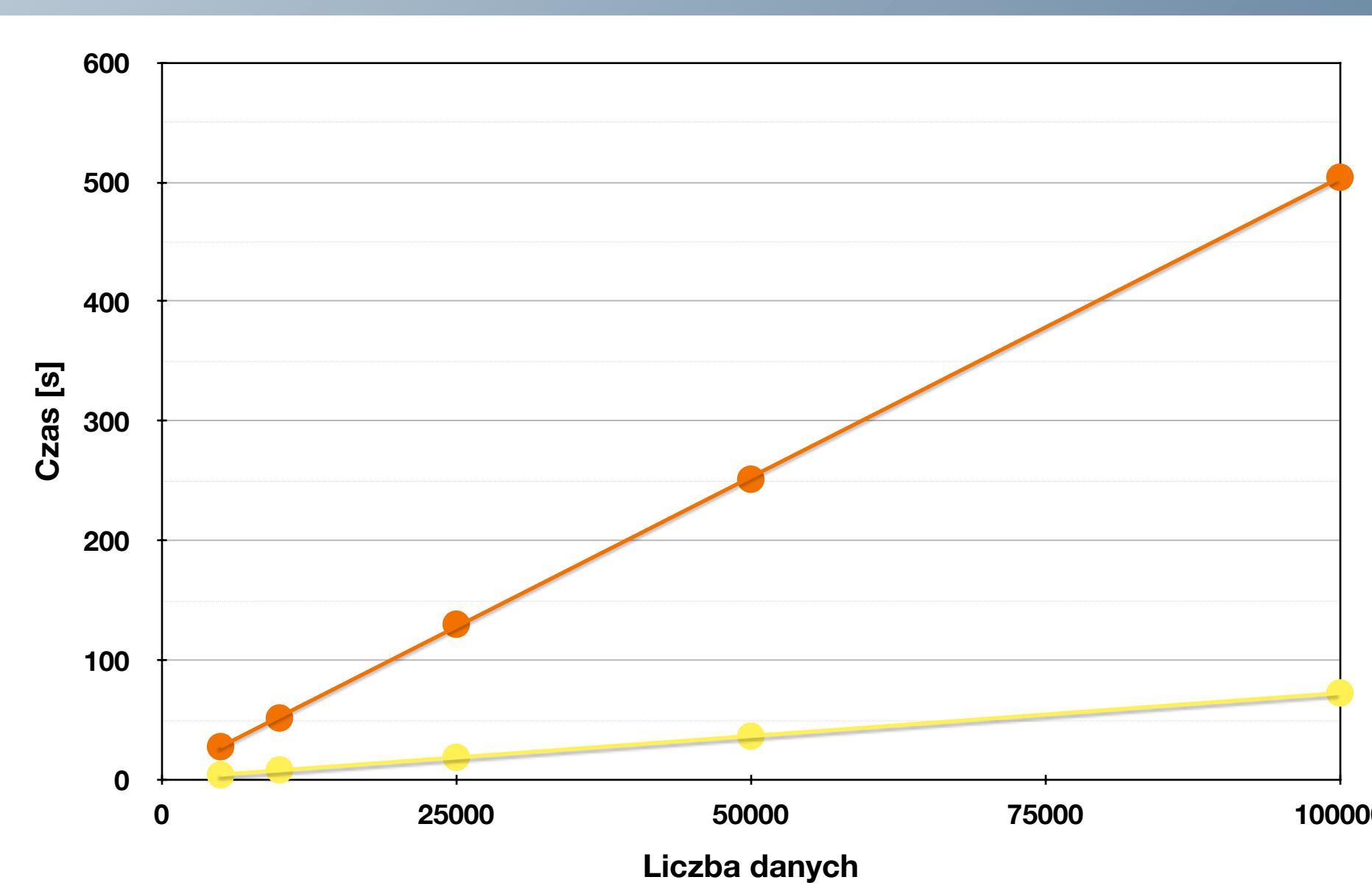
▶ Znając liczbę istotnych biologicznie SNP **możemy oszacować** całkowitą liczbę SNP dla chromosomu.

Paralelizacja - analiza 2



Wyniki czasu w zależności od ilości danych

Ilość danych	Praca na wielu rdzeniach	Praca sekwencyjna
5000	4,284	27,909
10000	8,425	51,844
25000	18,989	130,259
50000	36,619	251,479
100000	72,787	503,942



- ▶ Paralelizacja w Pythonie dotyczy całej naszej funkcji.
- ▶ Za pomocą metody ***multiprocessing.Pool()*** do zmiennej ***pool*** przypisujemy „pracowników”.

```
pool = multiprocessing.Pool()
pool = multiprocessing.Pool(processes = cpu_count())

out_data = pool.map(chunk_all,in_data)

pool.close()
```

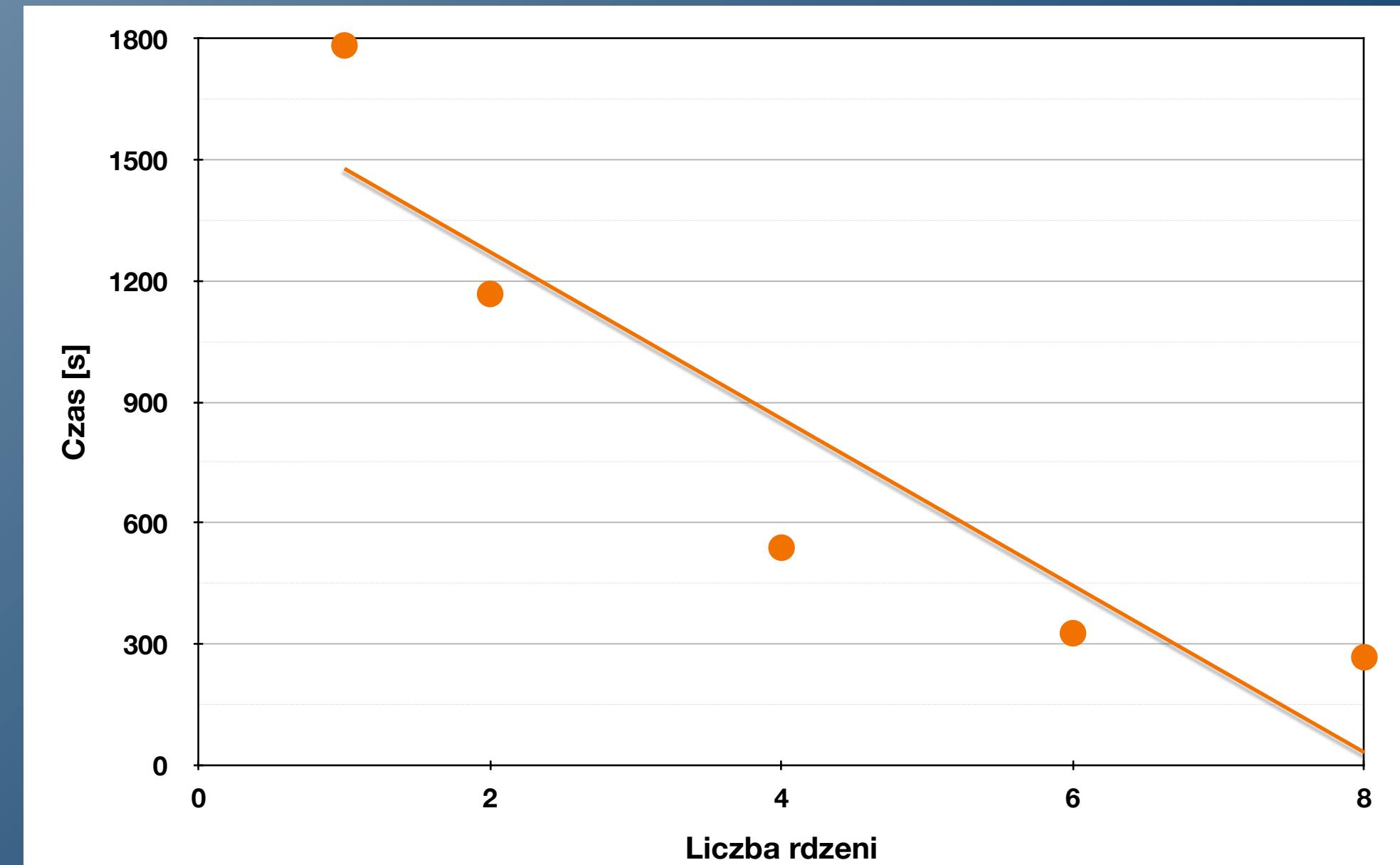
Paralelizacja - analiza 2



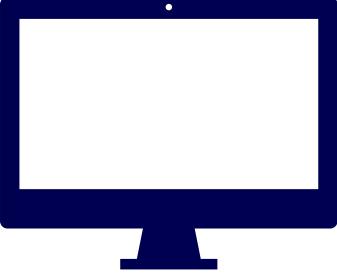
- ▶ **Obliczenia równoległe w R** dotyczyły **pętli**, która jest iterowana po wszystkich chromosomach.
- ▶ Używaliśmy do tego **klastra obliczeniowego** za pomocą funkcji *makeCluster* oraz *clusterApply*.

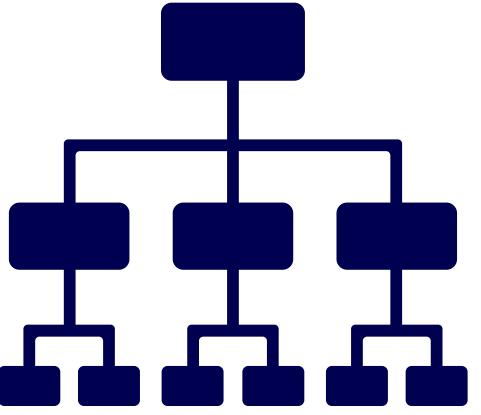
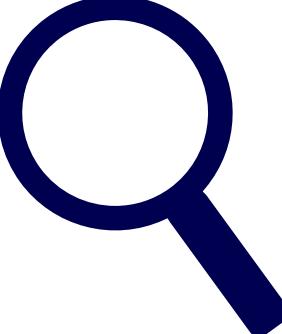
Wyniki czasu w zależności od liczby użytych rdzeni

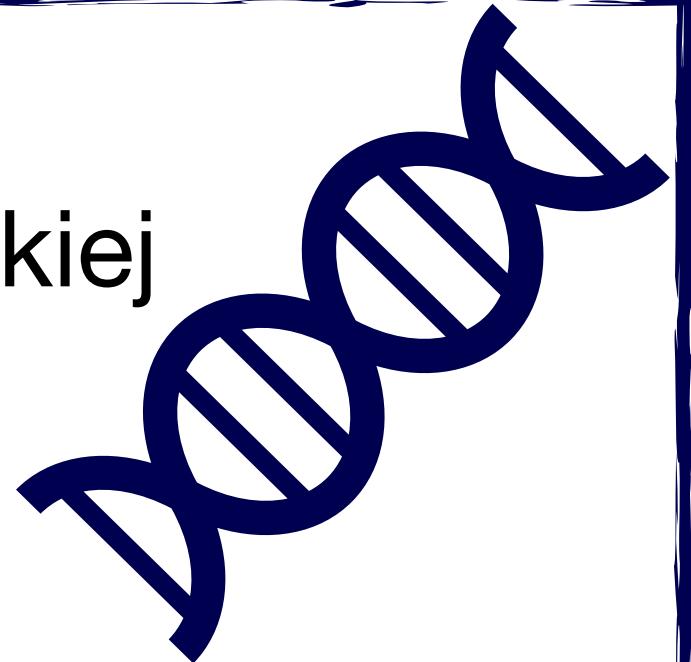
Liczba rdzeni	Czas
1	1783,651
2	1167,799
4	538,864
6	326,873
8	267,284



Podsumowanie

- ▶ Badaliśmy **zawartość mutacji typu SNP w genomach krów** z rasy holsztyńsko-fryzyjskiej otrzymanych z sekwencjonowania WGS.

- ▶ Nasze analizy prowadziliśmy w dwóch języka programowania Python oraz R.
- ▶ **Potwierdziliśmy** istnienie wszystkich badanych zależności - wszystkie nasze hipotezy badawcze okazały się poprawne.

- ▶ **Obliczenia równoległe** bazujące na pracy wielowątkowej gwarantują otrzymanie identycznych wyników w znacznie krótszym okresie czasu.

- ▶ **Praca w pakiecie R** okazała się szybsza niż w **języku Python** - być może powinniśmy spróbować przeprowadzić nasze analizy wyłącznie w oparciu o pakiet NumPy.




Projekt przygotowali:



Analiza danych

**Daria Plewa
Michał Humiński
Dominik Lisiecki**

**Bioinformatyka
studia licencjackie rok 3 2021/2022**

