

analiza danych

1. Wprowadzenie
2. Regresja liniowa – przypomnienie i najprostszy klasyfikator - regresja logistyczna
3. Metody wyboru zmiennych oparte na równaniu regresji
- 4. Metody preselekcji zmiennych**
5. Transformacje – embedding ①
6. Transformacje – embedding ②
7. Prezentacja projektów

1. Przykładowe zbiory danych
2. Literatura
3. Metody wyboru zmiennych
 - Oparte na danych
 - Podobieństwo między zmiennymi
 - Korelacja zmiennej zależnej i niezależnej
4. Projekt

- 1. Przykładowe zbiory danych**
2. Literatura
3. Metody wyboru zmiennych
 - Oparte na danych
 - Podobieństwo między zmiennymi
 - Korelacja zmiennej zależnej i niezależnej
4. Projekt

Przykładowe dane ①

Dog protein **abundance** in tumor and tumor-margin tissue

sample	type	(E2-indepe	(E3-indepe	(E3-indepe	(Lyso)-N-ac	[F-actin]-m	[F-actin]-m	[F-actin]-m	[Protein AC	[Protein AC	[Pyruvate c	[Pyruvate c	[Pyruvate c	[Pyruvate c	[Pyruvate c	[Pyruvate c	1,25-dihyd	1,4-alpha-g
S01G	tumour	NA	24974	NA	NA	NA	NA	359118	NA	129502	NA	NA	NA	NA	NA	NA	NA	NA
S03G	tumour	163338	NA	1424104	NA	NA	NA	1090950	NA	NA	NA	NA	NA	NA	NA	NA	134016	141816
S03M	margin	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
S04M	margin	526914	NA	27456	NA	NA	NA	306816	NA	283098	NA	NA	NA	NA	NA	NA	NA	NA
S05G	tumour	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
S05M	margin	NA	NA	NA	7754	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
S13G	tumour	NA	NA	NA	NA	67136	505120	NA	NA	693274	NA	NA	NA	NA	NA	NA	NA	NA
S28M	margin	NA	NA	NA	NA	NA	NA	NA	NA	176943	NA	NA	NA	NA	NA	NA	NA	NA
S31M	margin	NA	NA	NA	NA	NA	NA	837804	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
S32G	tumour	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
S34G	tumour	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
S34M	margin	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
S35G	tumour	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
S40G	tumour	NA	NA	NA	NA	103730	NA	NA	NA	NA	NA	NA	NA	NA	76096	NA	NA	NA

Przykładowe dane ②

Human gut microbiome abundance

#OTUID	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri	k__Bacteri
ID_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.06939	0	0	0	0
ID_2	0	0	0	0	0.08773	0	0	0.11273	0	0	0	0	0	0.0108	0	0.01557	0	0
ID_3	0	0	0	0	0.05996	0	0	0.05692	0	0.15201	0	0	0	0	0	0.00917	0	0
ID_4	0	0	0	0	0	0	0	0.15253	0	0	0	0	0	0.02423	0	0	0	0
ID_5	0	0	0	0	0	0	0	0.29028	0	0.02485	0	0	0	0.00638	0	0	0	0
ID_6	0	0	0	0	0.02209	0	0	0.00448	0	0.05072	0	0	0	0	0	0	0	0
ID_7	0	0	0	0	0.08967	0	0	0.00366	0	0	0.00501	0	0	0.03247	0	0.01645	0	0
ID_8	0	0	0	0	0	0	0	5.57473	0	0	0	0	0	0.35336	0	0.00489	0	0
ID_9	0	0	0.03917	0	0.06967	0	0	0.07892	0	0.37835	0	0	0	0	0	0	0	0
ID_10	0	0	0	0	0	0	0	0.06806	0	0	0	0	0	0.01665	0	0	0	0
ID_11	0	0	0	0	0	0	0	0.02955	0	0.10734	0.00804	0	0	0.07372	0	0	0	0
ID_12	0	0	0	0	0	0	0	0.14006	0	0.21112	0	0	0	0	0	0	0	0
ID_13	0	0	0	0	0	0	0	0.00584	0	0.05466	0	0	0	0.02156	0	0	0	0
ID_14	0	0	0	0	0.05216	0	0	0.00377	0	0	0.00143	0	0	0.02301	0	0.00742	0	0
ID_15	0	0	0	0	0	0	0	0.008	0	0.23046	0	0	0	0	0	0.00129	0	0
ID_16	0	0	0	0	0	0	0	0.03929	0	0.1214	0	0	0	0.02315	0	0.00247	0	0
ID_17	0	0	0	0	0	0	0	0.24682	0	0	0	0	0	0	0	0	0	0
ID_18	0	0	0	0	0.03772	0	0	0.95512	0	0	0.00099	0	0	0.00258	0	0.01063	0	0
ID_19	0	0	0	0	0.08596	0	0	0	0	0.0161	0	0	0	0	0	0.00517	0	0
ID_20	0	0	0	0	0	0	0	0	0	0.08814	0	0	0	0.04788	0	0.00517	0	0
ID_21	0	0	0	0.01118	0.06618	0	0	0.16579	0	0.24536	0	0	0	0.00871	0	0	0	0
ID_22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Przykładowe dane ③

Post-natal depression of fathers - questionnaire

EPDS6. Wydarzenia przerastały i przytłaczały mnie: 3 Tak, przez większość czasu i nigdy nie umiałem sobie z nimi poradzić 2 Tak, czasami nie potrafiłem radzić sobie tak dobrze jak zwykle 1 Przez większość czasu radziłem sobie zupełnie nieźle 0 Radziłem sobie tak samo dobrze jak dawniej	EPDS7. Czułem się tak nieszczęśliwy, że nie mogłem spać w nocy: 3 Tak, przez większość czasu 2 Tak, czasami 1 Rzadko 0 Nie, wcale nie	EPDS8. Byłem smutny i czułem się nieszczęśliwy: 3 Tak, przez większość czasu 2 Tak, dość często 1 Niezbyt często 0 Nie, nigdy	EPDS9. Czułem się tak nieszczęśliwy, że wybuchałem płaczem: 3 Tak, przez większość czasu 2 Tak, dość często 1 Tylko czasami 0 Nie, nigdy	EPDS10. Przychodziło mi do głowy, by zrobić sobie krzywdę: 3 Tak, dość często 2 Czasami 1 Praktycznie nigdy 0 Nigdy	GOT1. Byłem bardziej zestresowany niż zwykle, miałem niższy próg odporności na stres 0 Zupełnie nieprawdziwe 1 nieprawdziwe 2 prawdziwe 3 całkowicie prawdziwe
2	1	1	0	0	1
1	1	1	1	0	1
1	1	0	0	0	1
0	0	0	0	0	1
0	0	0	0	0	0
1	0	0	0	0	1
0	1	1	0	0	2
0	0	0	0	0	3
0	0	0	0	0	0
1	0	0	0	0	3
0	0	0	0	0	2
0	0	0	0	0	1
0	0	0	0	0	0
0	0	0	0	0	2
0	0	0	0	0	2
0	0	0	0	0	0

Przykładowe dane ④

Stres

		skala wynagrodzenia r		poziom int	nasilenie sy	nasilenie sy	nasilenie sy	nasilenie a	nasilenie al	nasilenie d	suma nasile	nasilenie p	wskaźnik k	wskaźnik n
ID_kod		SWM	Wiek	Raven_TM	SCL_social	SCL_vegeta	SCL_pain	SCL_agora	SCL_depr_	SCL_depr_	SCL_all	AUDIT_sun	Cum_Stres	Mismatch
518600296		20.20	18.00	29	11.00	6.00	11.00	5.00	3.00	1.00	37.00	3	-0.50	-1.23
388600126		22.70	20.00	15	1.00	4.00	3.00	0.00	3.00	0.00	11.00	5	-2.27	0.03
388610215		18.30	20.00	27	3.00	7.00	11.00	1.00	3.00	0.00	25.00	5	0.11	0.71
388600164		18.30	21.00	26	7.00	8.00	10.00	1.00	8.00	0.00	34.00	9	-1.89	-0.35
518600080		22.70	19.00	13	3.00	3.00	6.00	1.00	3.00	0.00	16.00	8	0.21	0.61
518600016		23.20	22.00	21	5.00	5.00	6.00	0.00	5.00	0.00	21.00	7	-2.01	0.29
518600197		21.10	21.00	25	4.00	1.00	12.00	5.00	7.00	0.00	29.00	6	0.10	-0.81
388600182		20.00	19.00	19	9.00	7.00	11.00	5.00	4.00	4.00	40.00	16	4.16	-1.30
518600112		44.10	20.00	11	1.00	0.00	7.00	0.00	0.00	0.00	8.00	9	-2.36	0.13
518600099		20.30	21.00	6	16.00	12.00	12.00	11.00	20.00	5.00	76.00	15	2.33	1.04
518600104	#NULL!		22.00	16	6.00	3.00	11.00	0.00	3.00	1.00	24.00	9	-2.36	0.13
388600176	#NULL!		20.00	18	3.00	0.00	5.00	0.00	5.00	0.00	13.00	2	-0.81	1.12
518600106		22.40	21.00	25	3.00	4.00	8.00	1.00	0.00	0.00	16.00	4	-2.36	0.13
518600226		31.90	24.00	12	0.00	0.00	3.00	0.00	1.00	0.00	4.00	2	-2.17	-0.06
388600163		21.00	20.00	23	17.00	12.00	7.00	10.00	10.00	5.00	61.00	12	-0.06	-1.16
518600127		18.30	23.00	13	4.00	0.00	4.00	0.00	2.00	0.00	10.00	9	-2.27	0.03
518600274		19.50	20.00	21	8.00	15.00	20.00	10.00	12.00	4.00	69.00	6	-1.92	0.19
388600076		68.10	23.00	13	1.00	0.00	6.00	0.00	0.00	0.00	7.00	4	-0.43	0.74
518600108		22.70	21.00	28	3.00	5.00	7.00	0.00	0.00	0.00	15.00	2	-2.17	-0.06
518610326		20.30	23.00	18	0.00	4.00	2.00	0.00	2.00	0.00	8.00	5	-2.27	0.03
388600044	#NULL!		19.00	14	5.00	6.00	10.00	3.00	2.00	2.00	28.00	7	-2.17	-0.06
518600201		23.50	22.00	21	13.00	1.00	9.00	5.00	6.00	1.00	35.00	1	3.88	0.50
518600260		21.00	19.00	11	2.00	2.00	7.00	3.00	3.00	0.00	17.00	6	-2.36	0.13

1. Przykładowe zbiory danych
- 2. Literatura**
3. Metody wyboru zmiennych
 - Oparte na danych
 - Podobieństwo między zmiennymi
 - Korelacja zmiennej zależnej i niezależnej
4. Projekt

Original Article | Published: 23 January 2009

Genome-wide association study of panic disorder in the Japanese population

[Takeshi Otowa](#), [Eiji Yoshida](#), [Nagisa Sugaya](#), [Shin Yasuda](#), [Yukika Nishimura](#), [Ken Inoue](#), [Mamoru Tochigi](#),
[Tadashi Umekage](#), [Taku Miyagawa](#), [Nao Nishida](#), [Katsushi Tokunaga](#), [Hisashi Tanii](#), [Tsukasa Sasaki](#) ,
[Hisanobu Kaiya](#) & [Yuji Okazaki](#)

[Journal of Human Genetics](#) **54**, 122–126 (2009) | [Cite this article](#)

Genotyping

The genomic DNA was extracted from peripheral leukocytes by standard procedures.

Genotyping was conducted using the GeneChip Human Mapping 500 K Array Set (Affymetrix, CA, USA). A total of 500 568 SNPs on the chip were genotyped. This array set consisted of two chips (*Styl* and *Nspl*) with approximately 250 000 SNPs on each chip.

Original Article | Published: 23 January 2009

Genome-wide association study of panic disorder in the Japanese population

[Takeshi Otowa](#), [Eiji Yoshida](#), [Nagisa Sugaya](#), [Shin Yasuda](#), [Yukika Nishimura](#), [Ken Inoue](#), [Mamoru Tochigi](#),
[Tadashi Umekage](#), [Taku Miyagawa](#), [Nao Nishida](#), [Katsushi Tokunaga](#), [Hisashi Tanii](#), [Tsukasa Sasaki](#) ,
[Hisanobu Kaiya](#) & [Yuji Okazaki](#)

[Journal of Human Genetics](#) **54**, 122–126 (2009) | [Cite this article](#)

Statistical analysis

After the genotyping was completed, data cleaning was conducted as follows: (1) SNPs with a call rate of <95% in patients or controls were excluded, and (2) SNPs with a P -value of <0.1% for Hardy–Weinberg's equilibrium in the controls were excluded. (3) Then, the criterion regarding minor allele frequency (MAF) was determined as follows: the ratios of the observed and expected (by chance) numbers of SNPs with several levels of allele P -values (1.0×10^{-2} , 10^{-3} and 10^{-4}) were evaluated among SNPs with MAF of $\geq 1, 5, 10\%$ in all samples according to Miyagawa *et al.*¹⁵ When the observed numbers of SNPs with P -values were considered to be increased too much at a certain MAF compared with the expected numbers, SNPs with that (or a lower) level of MAF may be excluded.

1. Przykładowe zbiory danych
2. Literatura
- 3. Metody wyboru zmiennych**
 - Oparte na danych
 - Podobieństwo między zmiennymi
 - Korelacja zmiennej zależnej i niezależnej
4. Projekt

Metody wyboru zmiennych oparte na strukturze danych

- % brakujących danych
- Cechy biologiczne np. MAF
- Informatywność – wariancja / odchylenie standardowe, współczynnik zmienności

$$\sigma^2 = \frac{\sum_{i=1}^N (\bar{x} - x_i)^2}{N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (\bar{x} - x_i)^2}{N}}$$

$$CV = \frac{\sigma}{\bar{x}}$$

$$GCV = e^{\sigma_{ln}} - 1$$

Metody wyboru zmiennych oparte na podobieństwie między dwoma zmiennymi

- Korelacja Pearsona

$$\sigma_{x1,x2} = \frac{\sum_{i=1}^N (\overline{x1} - x1_i)(\overline{x2} - x2_i)}{\sqrt{(\overline{x1} - x1_i)^2 (\overline{x2} - x2_i)^2}}$$

- Korelacja Spearmana

$$r_{x1,x2} = 1 - \frac{6 \sum_{i=1}^N (r1_i - r2_i)^2}{N(N^2 - 1)}$$

Metody wyboru zmiennych oparte na podobieństwie między dwoma zmiennymi

Abbre.	Method	Symbol	Description	References
Pearson	Pearson's	r	Linear, widely-used, no parameter, coeff. $\in [-1, 1]$	Pearson, 1895
Spearman	Spearman's	ρ	Monotonic, rank-based, no parameter, coeff. $\in [-1, 1]$	Spearman, 1904
Kendall	Kendall's	τ	Monotonic, rank-based, no parameter, coeff. $\in [-1, 1]$	Kendall, 1938
Hoeffding	Hoeffding's	D	Non-linear, rank-based, no parameter, coeff. $\in [0, 1]$	Hoeffding, 1948
Blomqvist	Blomqvist's	β	Monotonic, rank-based, no parameter, coeff. $\in [-1, 1]$	Blomqvist, 1950
Goodman	Goodman and Kruskal's	γ	Monotonic, cross classifications, rank-based, no parameter, coeff. $\in [-1, 1]$	Goodman and Kruskal, 1954
WWH	Wang, Waterman, Huang's	wwh	Monotonic, rank-based, no parameter, coeff. $\in [0, +\infty]$	Wang et al., 2014
MI	Mutual information	I	Non-linear, entropy-based, no parameter, coeff. $\in [0, +\infty]$	Shannon, 1948
MIC	Maximum information correlation	mic	Non-linear, entropy-based, 1 parameter, coeff. $\in [0, 1]$	Reshef et al., 2011
Wilks	Wilks'	W	Linear, covariance-based, no parameter, coeff. $\in [0, 1]$	Wilks, 1935
KCCA	Kernel canonical correlation analysis	$kcca$	Non-linear, covariance-based, 1 parameter, coeff. $\in [0, 1]$	Bach and Jordan, 2002
dCor	Distance correlation	$dCor$	Non-linear, covariance-based, 1 parameter, coeff. $\in [0, 1]$	Szekely and Rizzo, 2009
CMMD	copula-based maximum mean discrepancy	$cmmd$	Non-linear, copulas-based, 1 parameter, coeff. $\in [0, 1]$	Poczos et al., 2012
RDC	Randomized dependence coefficient	rdc	Non-linear, copulas-based, 2 parameters, coeff. $\in [0, 1]$	Lopez-Paz et al., 2013

Metody wyboru zmiennych oparte na podobieństwie

między grupami zmiennych

- Korelacja wieloraka

$$R_{x_1 x_n} = \begin{bmatrix} r_{x_1 x_2} & r_{x_1 x_3} & r_{x_1 x_4} \end{bmatrix} \begin{bmatrix} 1 & r_{x_2 x_3} & r_{x_2 x_4} \\ & 1 & r_{x_3 x_4} \\ \text{symm} & & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_{x_1 x_2} \\ r_{x_1 x_3} \\ r_{x_1 x_4} \end{bmatrix}$$

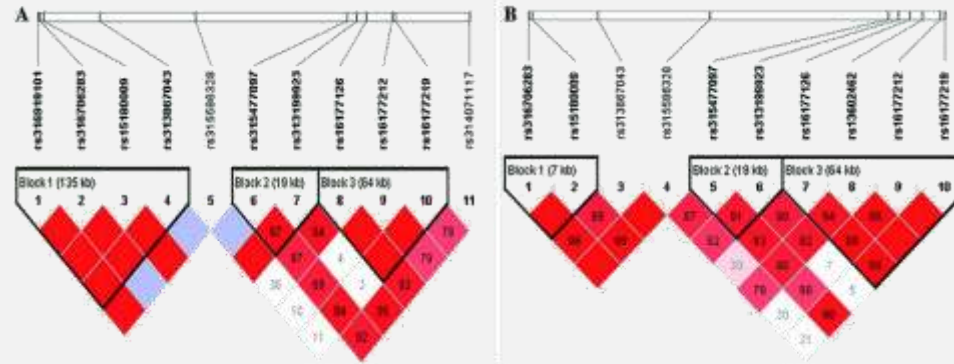
Metody wyboru zmiennych oparte na podobieństwie między grupami zmiennych

- heatmap



Metody wyboru zmiennych oparte na podobieństwie między SNP

- metoda redukcji liczby SNP → linkage disequilibrium LD
- HaploBlock



- PLINK

plink2-users	GitHub	File formats	PLINK 1.9 index	PLINK 2.0
------------------------------	------------------------	------------------------------	---------------------------------	---------------------------

PLINK 1.9 beta

PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

```
plink --tag example.vcf --tag-r2 0.7 --tag-window 50kb
```

Metody wyboru zmiennych oparte na podobieństwie

między zmienną zależną i niezależną

- Korelacja Pearsona
- Korelacje rangowe (np. Spearmana)
- Korelacja wieloraka

Metody wyboru zmiennych oparte na podobieństwie między zmienną zależną i niezależną

- CFS → Correlation based Feature Selection
- Wybór zbioru zmiennych niezależnych najbardziej skorelowanych ze zmienną niezależną, a najmniej skorelowanych ze sobą

$$CFS = \frac{k\bar{r}_{yx_i}}{\sqrt{k + k(k-1)\bar{r}_{x_ix_j}}}$$

- Modyfikacje CFS:
 - Fast Correlation-Based Filter FCBF
 - INTERACT
 - mRMR → minimum Redundancy Maximum Relevance → korelacja zastąpiona miarą *mutual information*
 - Md → korelacja zastąpiona miarą *monotone independence*

Metody wyboru zmiennych oparte na podobieństwie

między zmienną zależną i niezależną

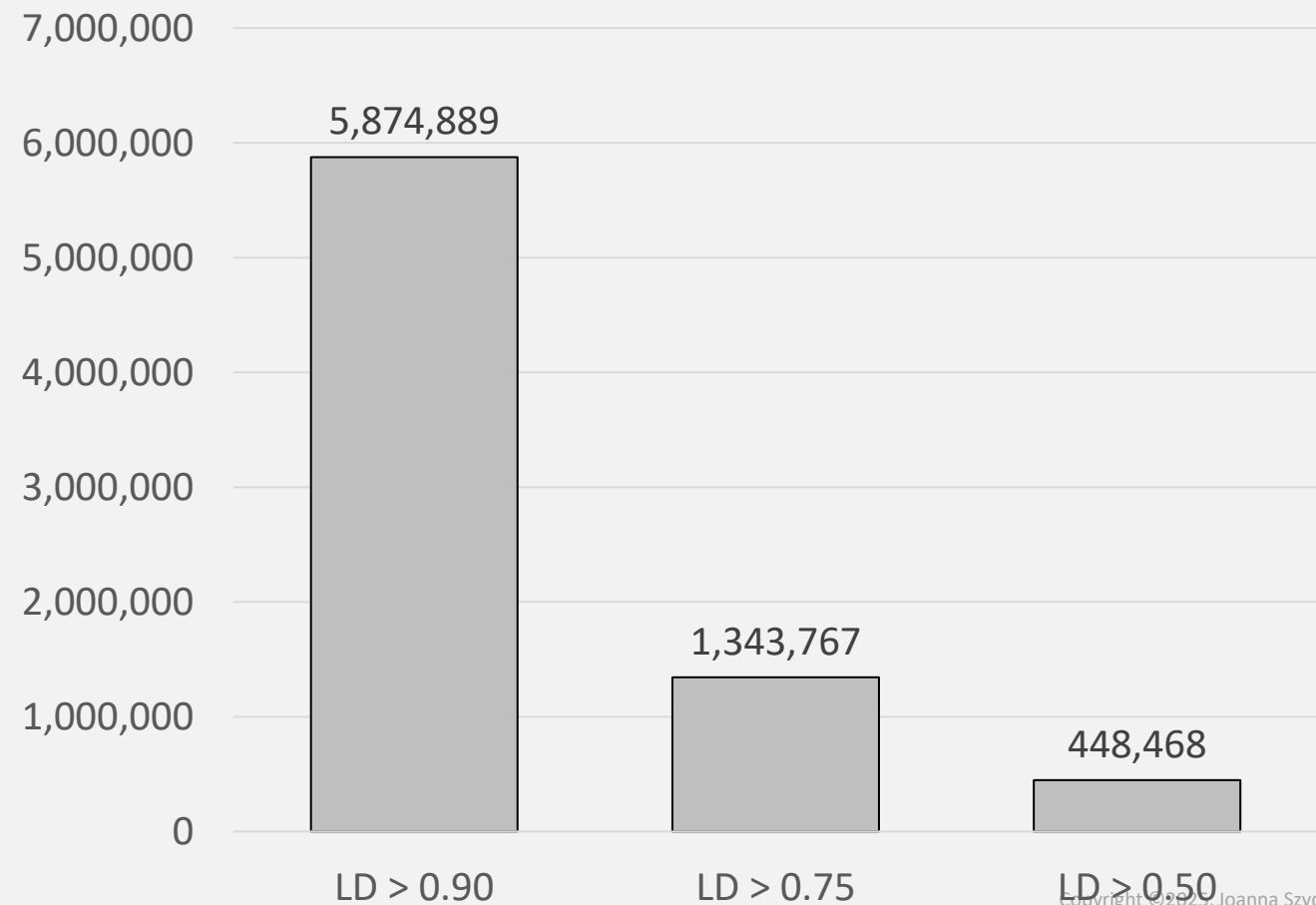
- Algorytmy oparte o metodę k-najbliższych sąsiadów
 - ReliefF

1. Przykładowe zbiory danych
2. Literatura
3. Metody wyboru zmiennych
 - Oparte na danych
 - Podobieństwo między zmiennymi
 - Korelacja zmiennej zależnej i niezależnej
- 4. Projekt**

Projekt

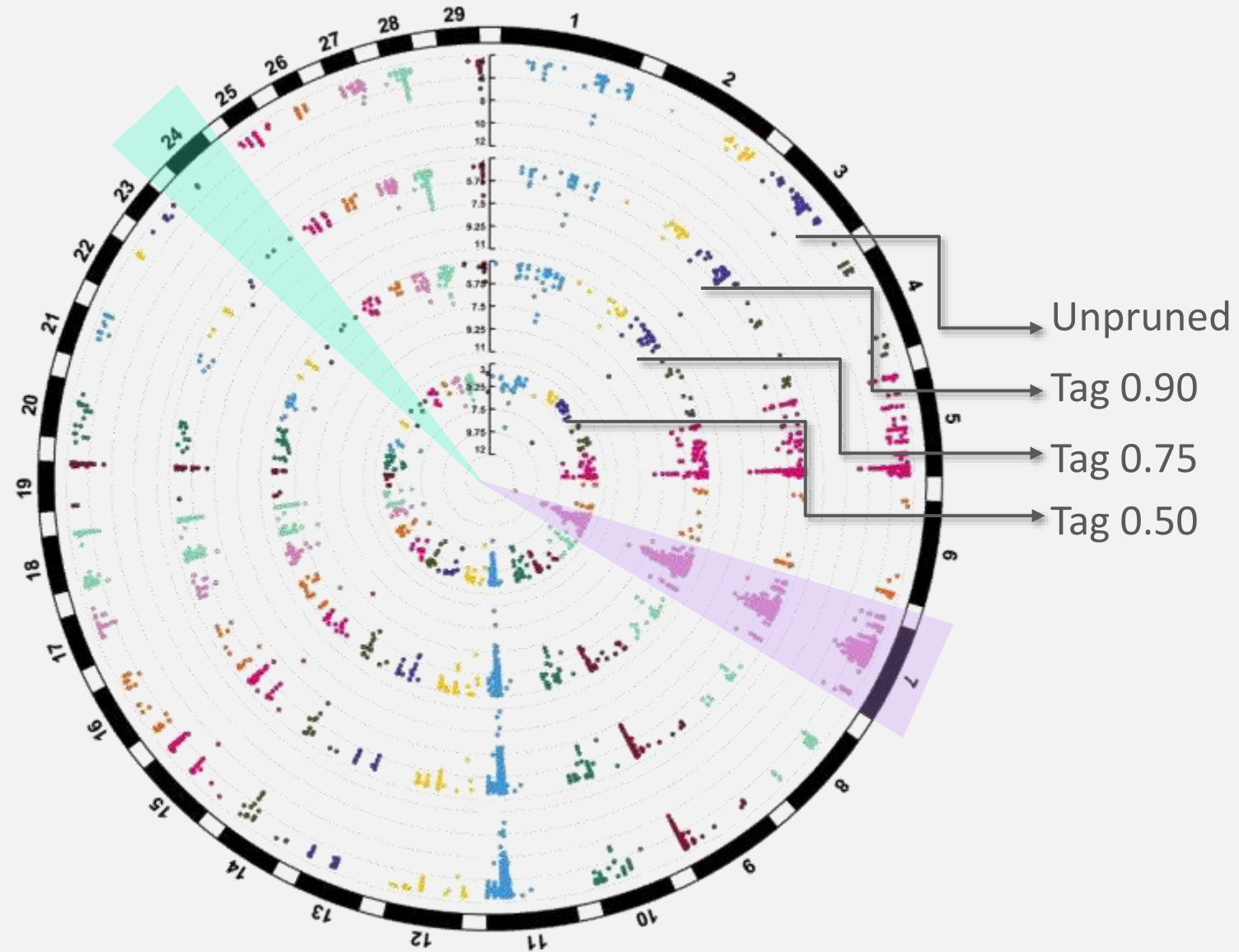
Materiał

- 1,000 Bull Genomes Project
- 528 HF buhajów
- Pseudofenotypy (wartości hodowlane) → wysokość w krzyżu
- Pełny zbiór – 10,771,020 SNP
- Podzbiory tagSNP → PLNK



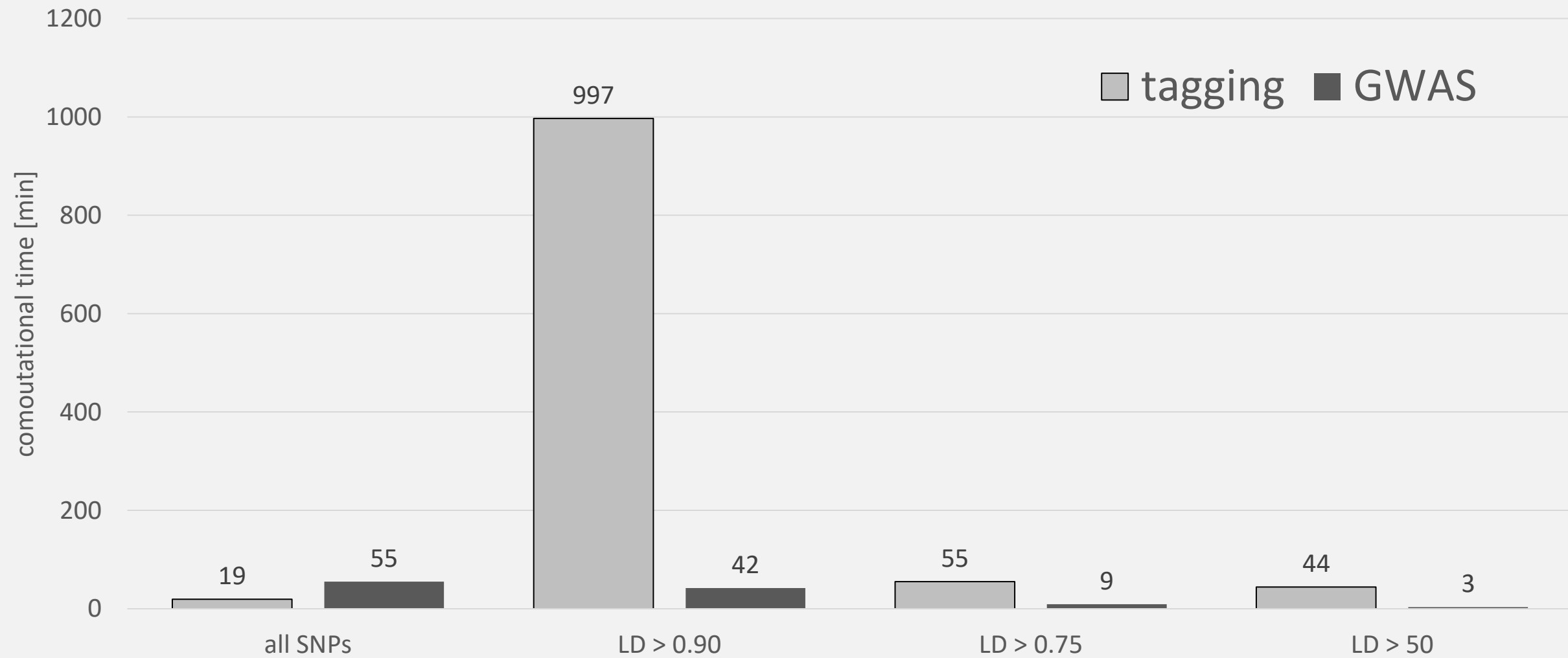
Projekt

Wyniki



Projekt

Czas obliczeń



1. Przykładowe zbiory danych
2. Literatura
3. Metody wyboru zmiennych
 - Oparte na danych
 - Podobieństwo między zmiennymi
 - Korelacja zmiennej zależnej i niezależnej
4. Projekt

1. Po co jest potrzebny wybór zmiennych w modelu ?
2. Proszę podać kryteria wyboru zmiennych
3. Które kryteria są lepsze – wybór w oparciu o model, czy wybór bez uwzględniania modelu ?