Projekt HapMap

prezentacja danych oraz przykładowe wyniki analizy sprzężeń pomiędzy loci

Przemysław Biecek

Plan referatu

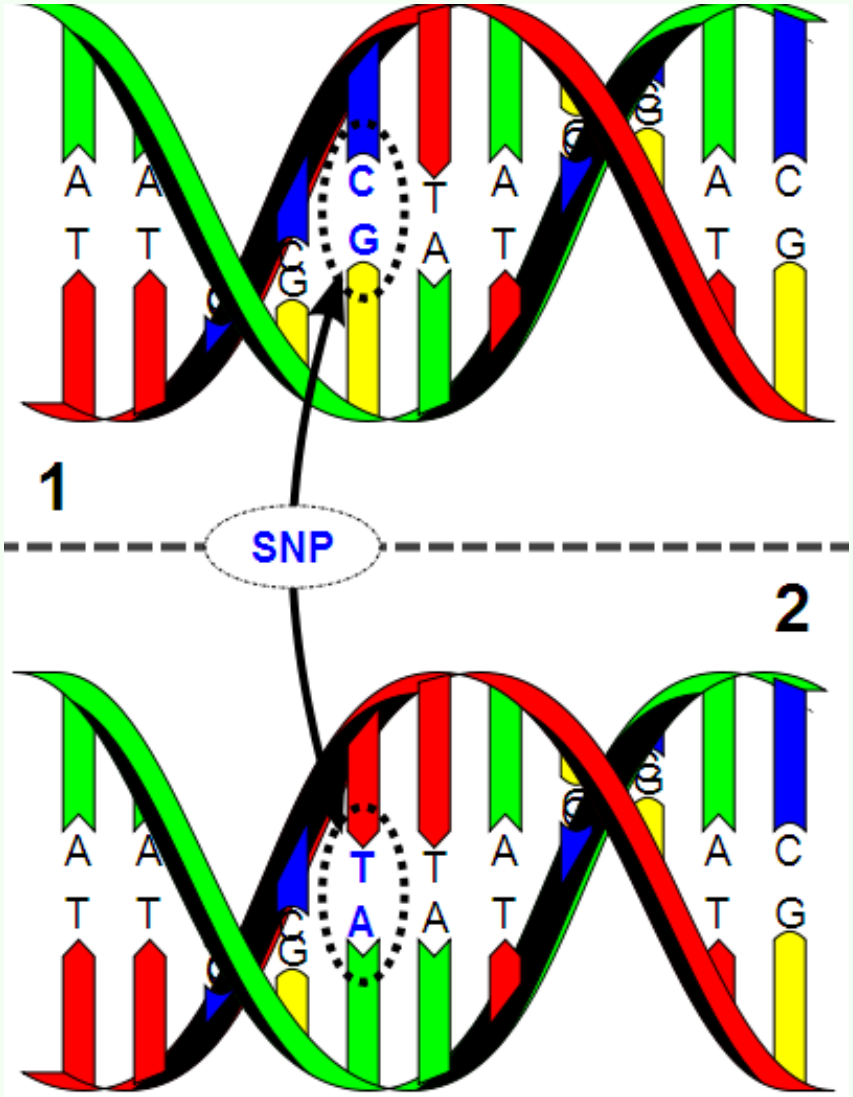
1. Przedstawienie projektu HapMap.
2. Przykłady analiz wykorzystujących dane z projektu HapMap
   - pozytywna presja selekcyjna,
   - QTL mapping,
   - tagSNP.
3. Analiza SNP zlokalizowanych w genach OR
   - motywacje płynące z symulacji,
   - wyniki dla danych z HapMap.

O projekcie HapMap

- W roku 19990 rozpoczął się projekt „Human Genome". Jednym z celów projektu było określenie sekwencji genomu człowieka. Po 13 latach, w kwietniu 2003 roku, na ramach Science i Nature ogłoszono zakończenie tego projektu. Projekt ukończono mniejszym nakładem środków niż planowano 2 lata przed czasem.
- W październiku 2002 rozpoczął się projekt „HapMap". Planowany budżet wynosił 12 milionów USD. Celem tego projektu było zbadanie zmienności genetycznej przedstawicieli czterech populacji. Drugi etap projektu zakończył się w październiku 2007.

Analizie poddano 270 osobników z czterech, separowanych geograficznie, populacji:

- 90 osobników (30 rodzin) z populacji Yoruba z Ibadan (Nigeria). Każdy uczestnik oświadczył, że wszyscy jego dziadkowie należeli do populacji Yoruba.
- 90 osobników (30 rodzin) zamieszkałych w Utah (UAS), których przodkowie byli imigrantami z zachodniej i północnej Europy. Dane były zebrane jeszcze w 1980 roku i pochodzą z the Centre d'Etude du Polymorphisme Humain CEPH.
- 45 osobników (niespokrewnionych) z Pekinu (Chiny) studentów jednej z uczelni. Każdy uczestnik oświadczył, że przynajmniej 3 z 4 jego dziadków należało do narodu Han.
- 45 osobników (niespokrewnionych) z Tokyo (Japonia). Każdy uczestnik oświadczył, że wszyscy jego dziadkowie pochodzili z Japonii.



SNP to polimorfizm pojedynczego nukleotydu (ang. Single Nucleotide Polymorphism).

Realizacje projektu podzielono na trzy etapy.

- Etap 1 (zakończono w październiku 2005).
  Polegał na zebraniu informacji o 1.000.000 SNP (średnio 1 SNP na 5kb), dla wszystkich 270 osobników.
- Etap 2 (zakończono w październiku 2007).
  Polegał na zebraniu informacji o ponad 3.500.000 SNP (średnio 1 SNP na 1kb), dla wszystkich 270 osobników.
- Etap 3.
  Polegał na zgenotypowaniu 10 regionów o długości 500kb u 48 osobników (16 YRI, 16 CEU, 8 JAP, 8 CHB).

W projekcie uczestniczyło wiele ośrodków badawczych.

Country	Research Group		Role	Percent Genome		Chromosomes
Japan	Yusuke Nakamura RIKEN and University of Tokyo		Genotyping -Third Wave	24.3%		5, 11, 14, 15, 16, 17, 19
United Kingdom	David Bentley Wellcome Trust Sanger Institute		Genotyping -Illumina BeadArray	23.7%		1, 6, 10, 13, 20
Canada	Thomas Hudson McGill University and Génome Québec		Genotyping -Illumina BeadArray	10.1%		2, 4p
China	Huanming Yang	Huanming Yang	Genotyping - Sequenom MassExtend, Illumina BeadArray	5.9%		3q, 8p,
		Wei Huang Chinese National at Shanghai	Genotyping -Illumina BeadArray	3.6%	9.5%	21
United States	Arnold Oliphant Illumina		Genotyping -Illumina BeadArray	16.1%		8q, 9, 18q, 22, X
	David Altshuler Broad Institute of Harvard and MIT		Genotyping -Sequenom MassExtend, Illumina	9.7%	32.4%	4q, 7q, 18p, Y

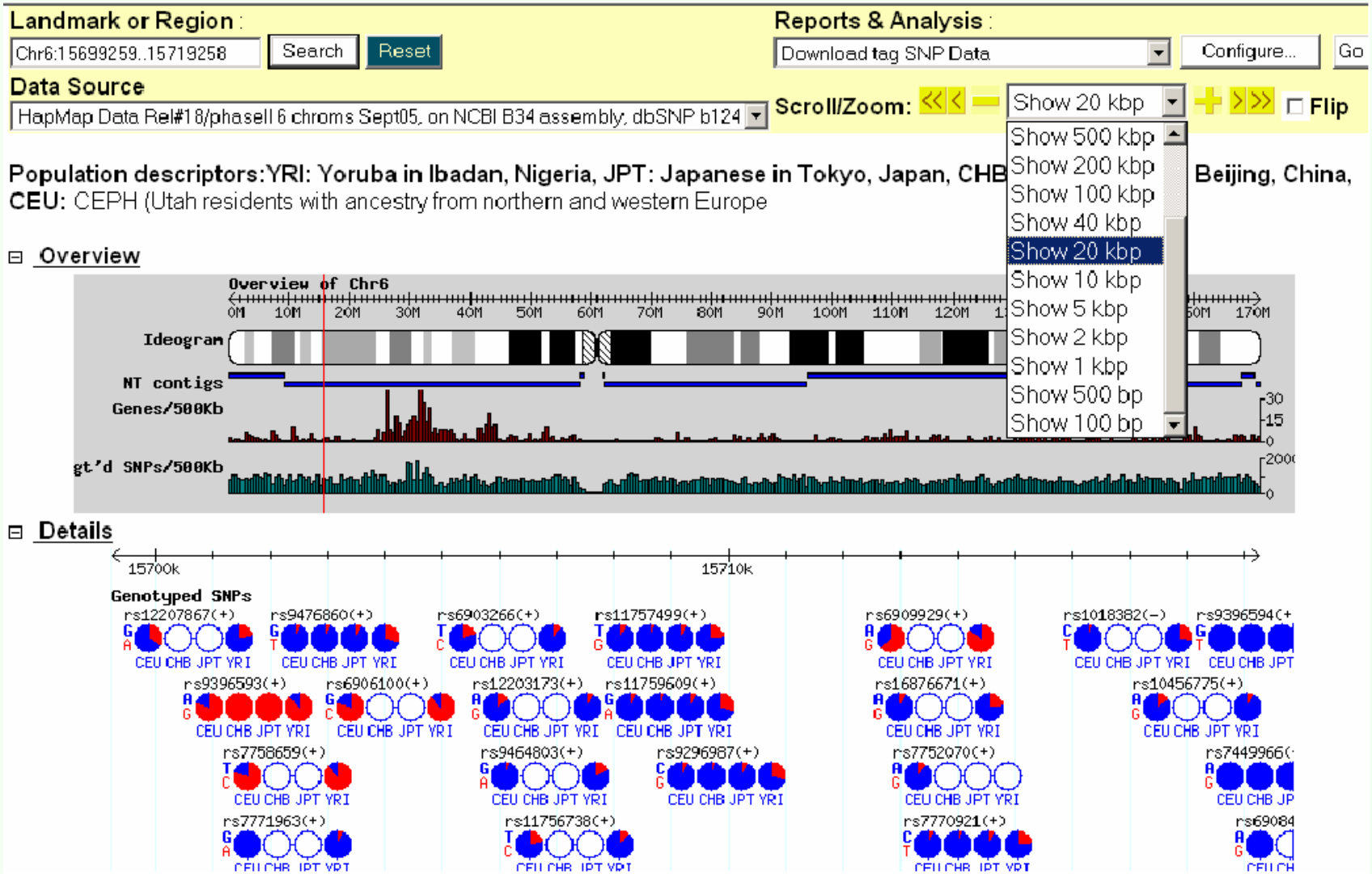
Proces Quality Control przeszło ponad 3,5 miliona SNP (liczba SNP różni się pomiędzy populacjami!!!).

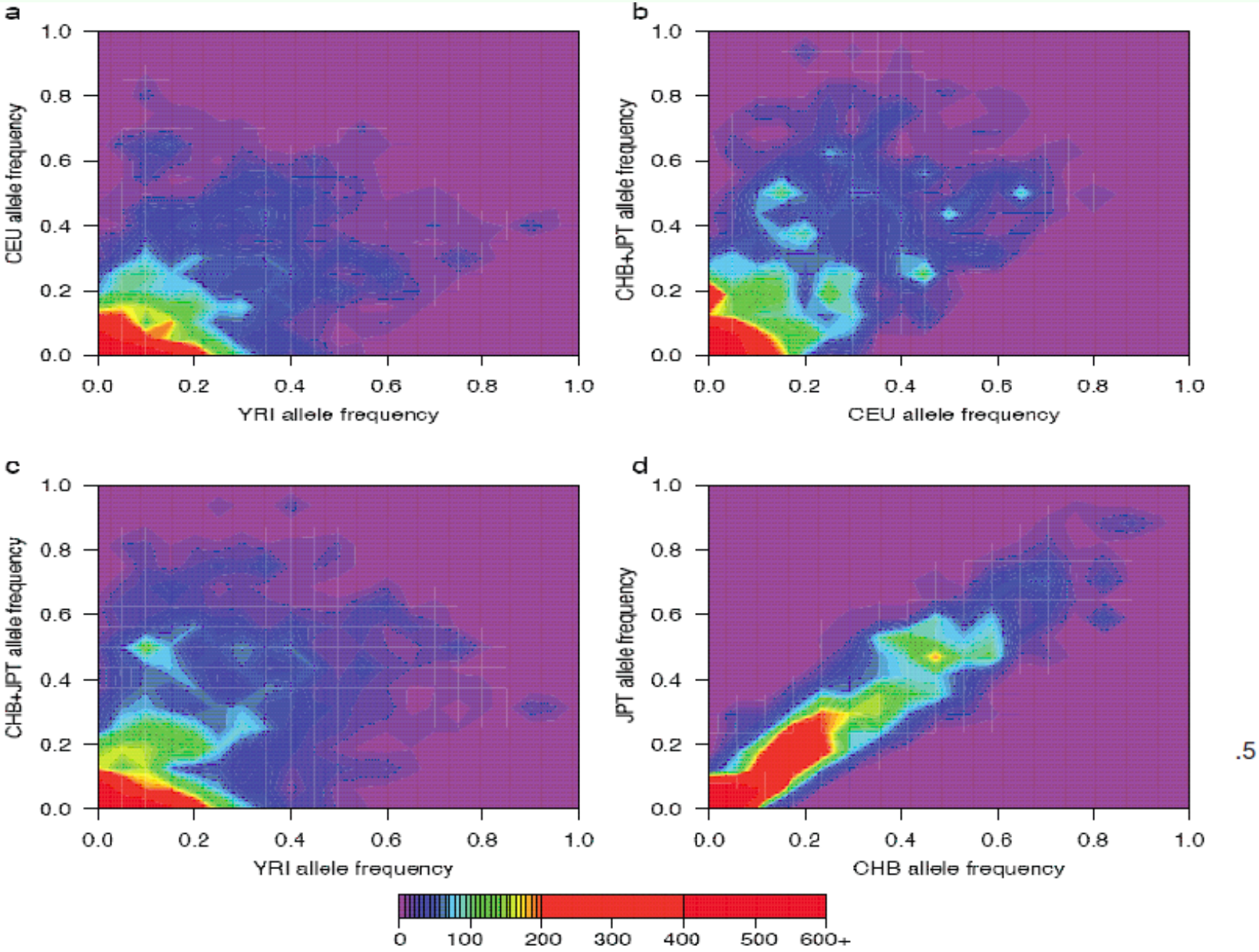
	CEU	JPT+CHB	YRI
chr1	296740	300842	295042
chr2	317542	318502	310060
chr3	246273	246742	241274
chr4	235794	236037	229979
chr5	240113	240886	234787
chr6	261164	265372	258665
chr7	206003	206785	201730
chr8	207474	211166	206488
chr9	176336	178339	175111
chr10	203991	206390	202537
chr11	198312	199751	193165
chr12	186292	187515	185020

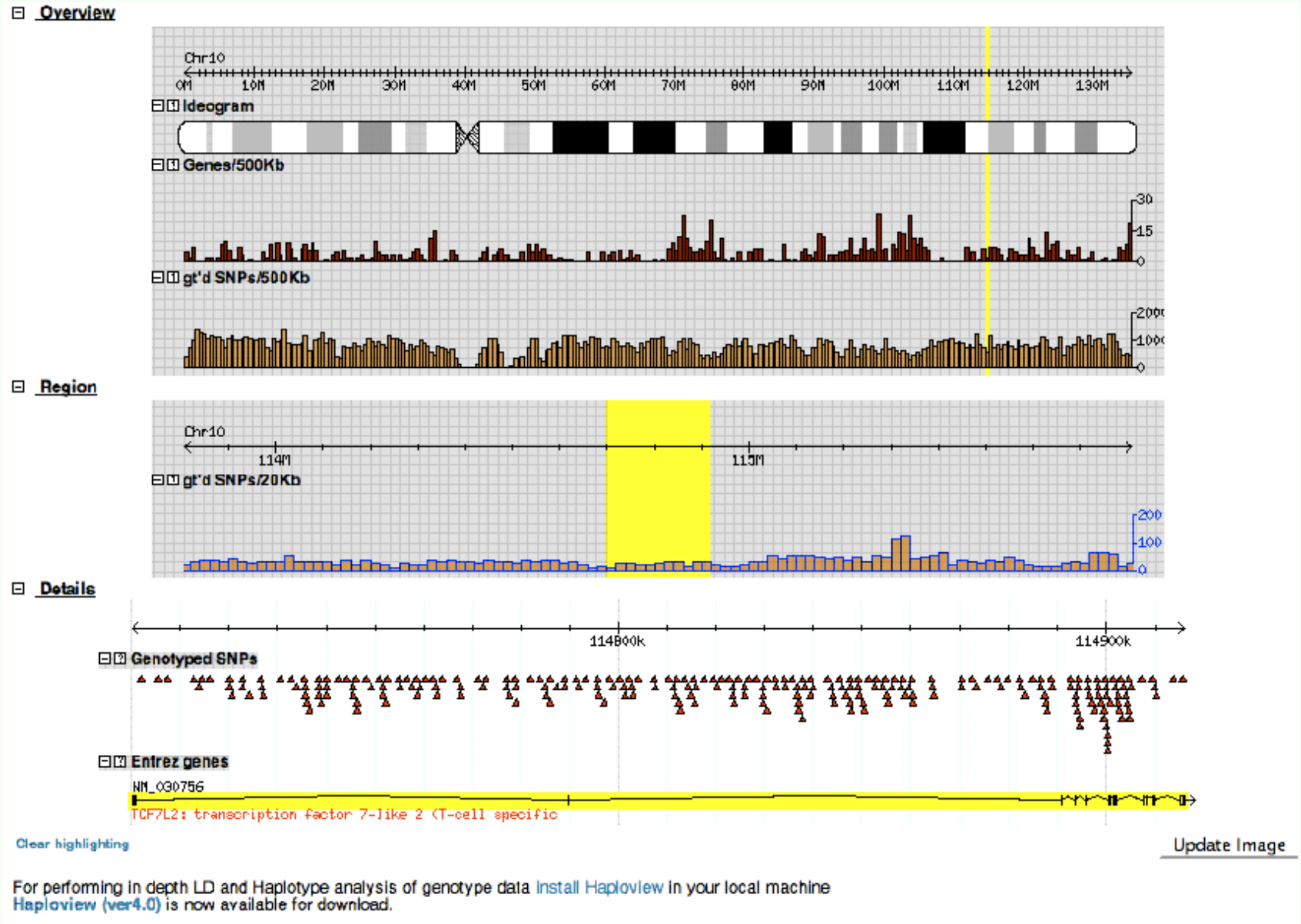
	CEU	JPT+CHB	YRI
chr13	151413	153859	150960
chr14	119915	120682	117364
chr15	104116	104725	101682
chr16	106660	106754	103910
chr17	86953	86898	84919
chr18	115675	116564	114373
chr19	54237	54387	53003
chr20	116276	116309	114079
chr21	48217	50053	48541
chr22	52986	54881	54008
chrX	106779	107709	106025
chrY	69	67	63
Total	3839330	3871213	3782785

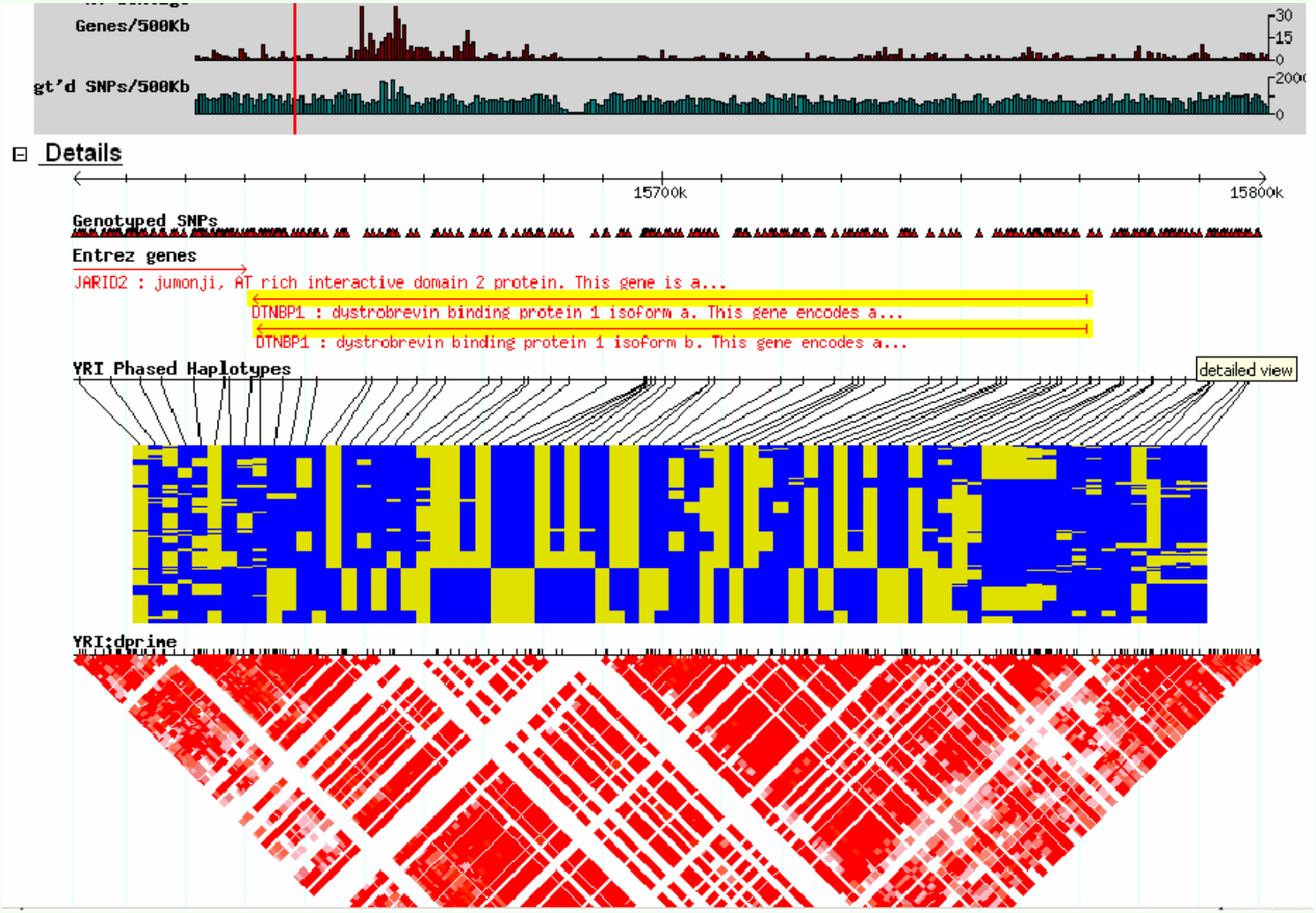
Dane są dostępne na stronie projektu **http://www.hapmap.org** (release 22 to ponad 6GB przetworzonych danych).
Na tej stronie dostępnych jest również wiele publikacji i tutoriali poświęconych projektowi.

Na stronie projektu dostępne są programy Genome Browser i MartView pozwalający na zarządzanie danymi oraz wykonanie różnorakich analiz.




a
CEU allele frequency
YRI allele frequency
b
CHB+JPT allele frequency
CEU allele frequency
c
CHB+JPT allele frequency
YRI allele frequency
d
JPT allele frequency
CHB allele frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
600+


Overview
Chr10
0M 10M 20M 30M 40M 50M 60M 70M 80M 90M 100M 110M 120M 130M
Ideogram
Genes/500Kb
30
15
0
gt'd SNPs/500Kb
Region
Chr10
114M
115M
gt'd SNPs/20Kb
200
100
0
Details
114800k
114900k
Genotyped SNPs
Entrez genes
NM_030756
TCF7L2: transcription factor 7-like 2 (T-cell specific
Clear highlighting
Update Image
For performing in depth LD and Haplotype analysis of genotype data Install Haploview in your local machine
Haploview (ver4.0) is now available for download.


Genes/500Kb
30
15
0
gt'd SNPs/500Kb
200
0
Details
15700k
15800k
Genotyped SNPs
Entrez genes
JARID2 : jumonji, AT rich interactive domain 2 protein. This gene is a...
DTNBP1 : dystrobrevin binding protein 1 isoform a. This gene encodes a...
DTNBP1 : dystrobrevin binding protein 1 isoform b. This gene encodes a...
YRI Phased Haplotypes
detailed view
YRI:dprime

Przykładowe kierunki badań

Taka ilość danych pozwala na przeprowadzanie różnorodnych analiz. Przedstawimy trzy przykłady:

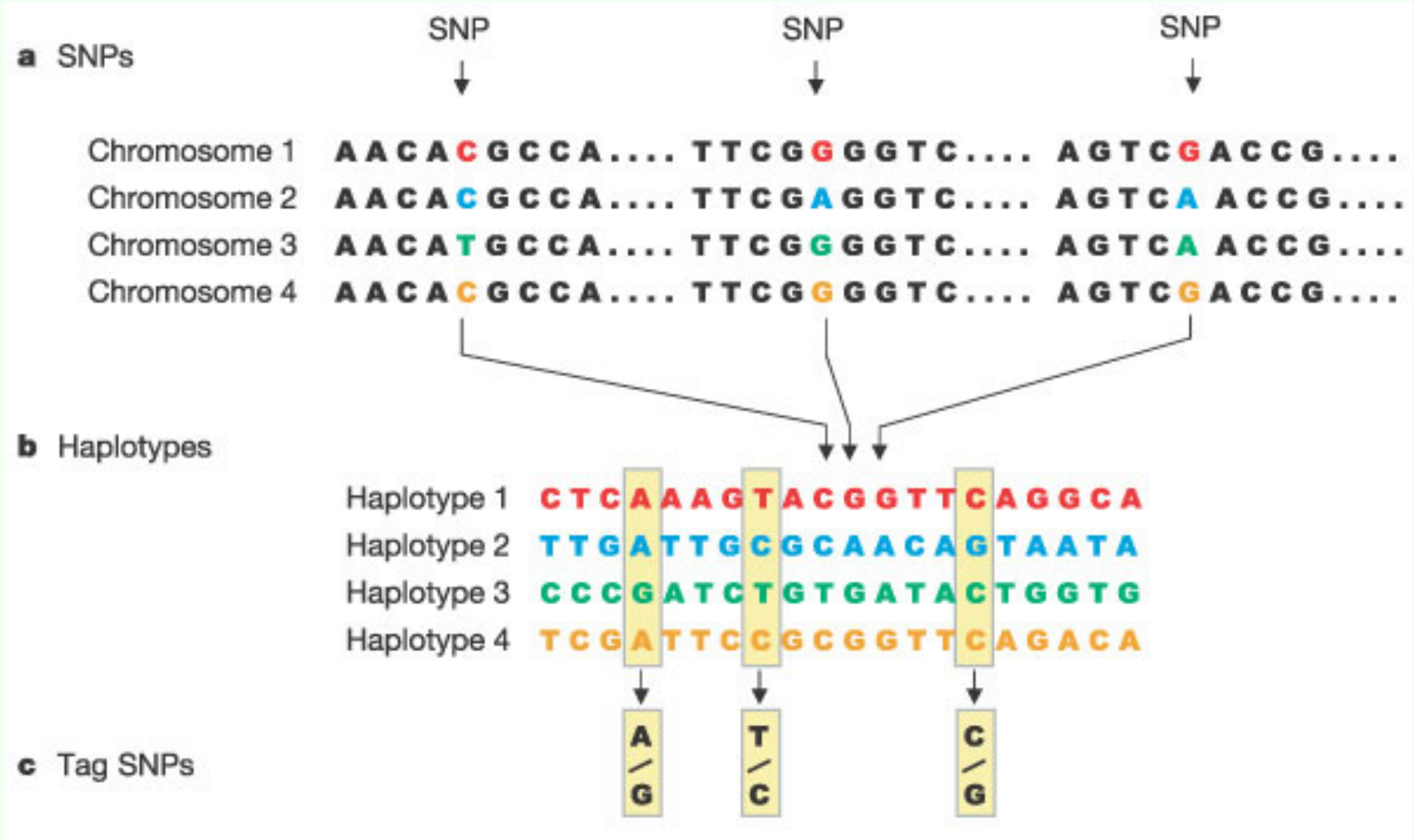
- wyznaczanie tagSNP,
- lokalizacja genów o geograficznie uwarunkowanej presji selekcjnej,
- lokalizacja QTLi związanych z ryzykiem zapadnięcia na wybraną chorobę.

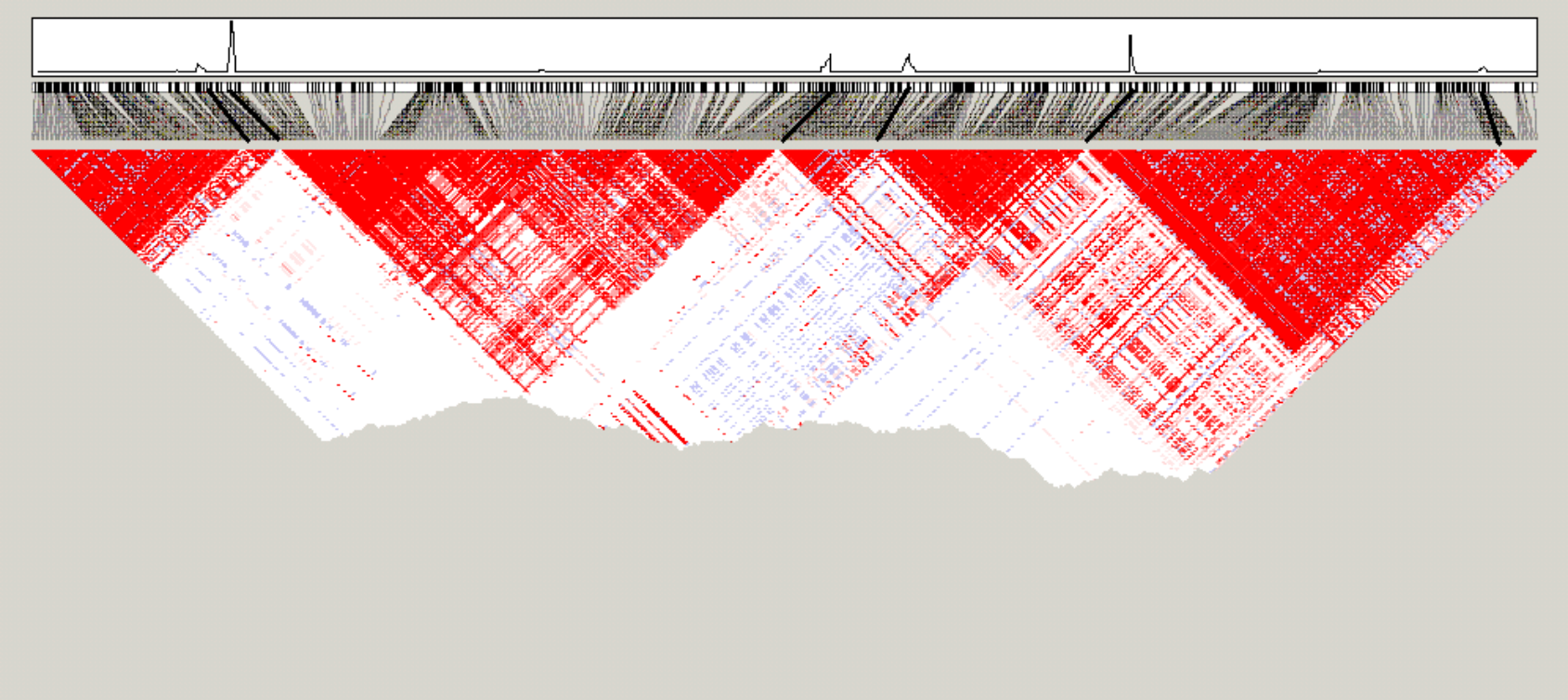
Wyznaczanie tagSNP

Tag SNPs

Minimalny zbiór SNP wymagany do identyfikacji haplotypu.
Współczynnik $r^2 = 1$ pomiędzy parą SNP oznacza, że jeden ze nicj jest nadmiarowy, ponieważ drugi „mówi wszystko" o nim.

Ocenia się, że w danej populacji 55% osób ma jedną (najczęstszą) wersję haplotypu, 30% ma inną, 8% ma kolejną a reszta zmienności jest związana z mniej częstymi haplotypami.
Wniosek: aby określić haplotyp osobnika, nie trzeba genotypować całego jego genomu!!!
W miarę dokładne przybliżenie można uzyskać dzięki mniejszemu zbiorowi SNP.


a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G



Geograficznie uwarunkowana presja selekcyjna

Zakłada się, że SNP to miejsca cichych mutacji. Gdyby tak nie było, to presja selekcyjna spowodowałaby, że dany allel rozpowszechni się w populacji.
Może się jednak zdarzyć, że

- dany allel jest bardzo młody i jeszcze nie zdominował całej populacji,
- dany allel jest korzystny w jednej populacji ale niekorzystny w innej (anemia sierpowata i malaria).

W obu przypadkach w jednej populacji będziemy obserwować wysoką częstość występowania danego allelu, a w innej bardzo niską.

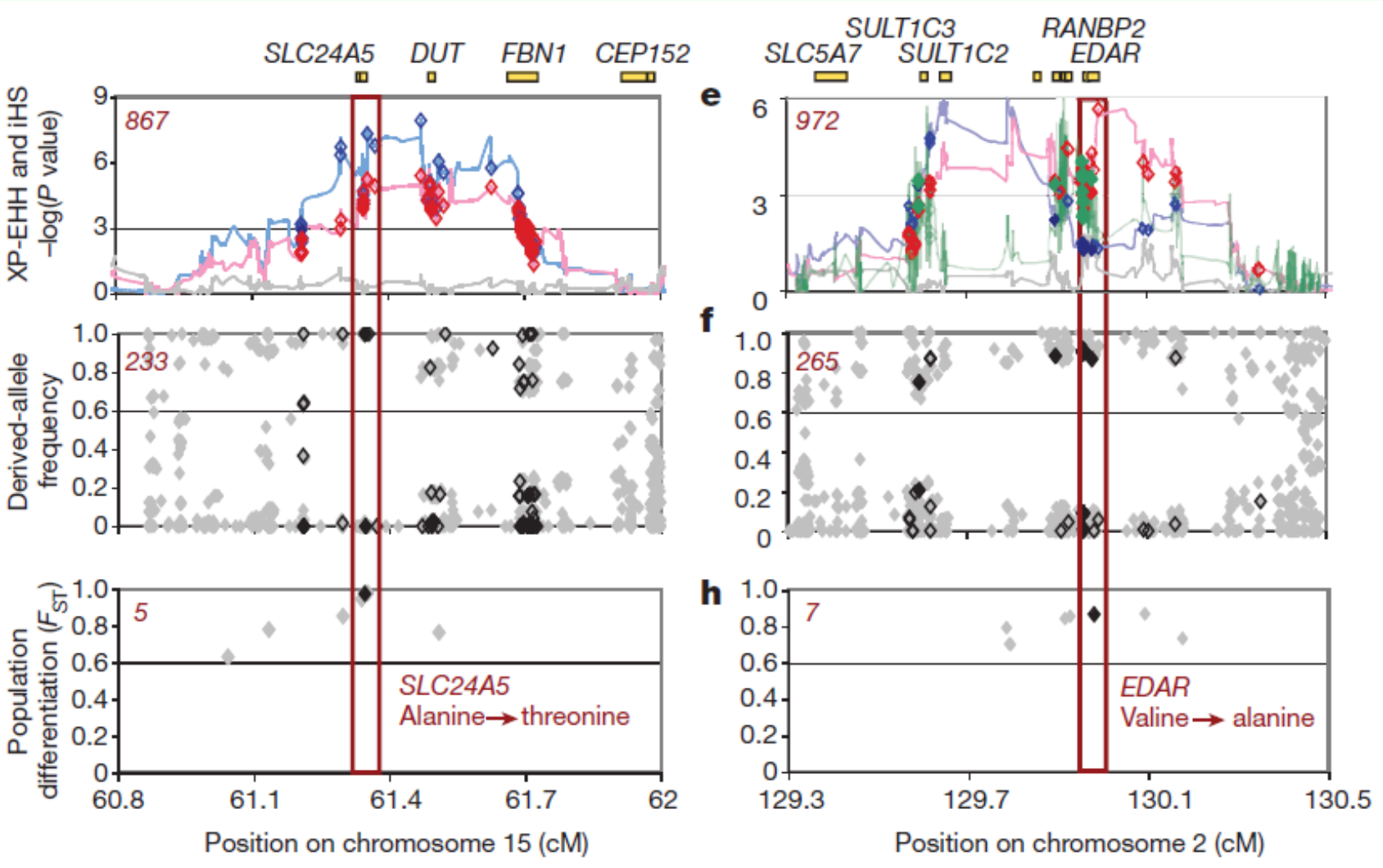
Sabeti, Varilly, Ben Fry, et. al. „Genome-wide detection and characterization of positive selection in human populations". Vol 449 Nature Letters 18 October 2007.

Pozytywna presje selekcyjną można więc poznać po częstości występowania allelu bliskiej fixacji (100%) oraz po dalekozasiegowm związku „linkage association" (association jest eliminowane przez rekombinacje, jego obecność oznacza, że częstość występowania danego allalu zmieniła się niedawno).

Table 1 | The twenty-two strongest candidates for natural selection

Region	Chr:position (MB, HG17)	Selected population	Long Haplotype Test	Size (Mb)	Total SNPs with Long Haplotype Signal	Subset of SNPs that fulfil criteria 1	Subset of SNPs that fulfil criteria 1 and 2	Subset of SNPs that fulfil criteria 1, 2 and 3	Genes at or near SNPs that fulfil all three criteria
1	chr1:166	CHB + JPT	LRH, iHS	0.4	92	39	30	2	*BLZF1, SLC19A2*
2	chr2:72.6	CHB + JPT	XP-EHH	0.8	732	250	0	0	
3	chr2:108.7	CHB + JPT	LRH, iHS, XP-EHH	1.0	972	265	7	1	*EDAR*
4	chr2:136.1	CEU	LRH, iHS, XP-EHH	2.4	1,213	282	24	3	*RAB3GAP1, R3HDM1, LCT*
5	chr2:177.9	CEU,CHB + JPT	LRH, iHS, XP-EHH	1.2	1,388	399	79	9	*PDE11A*
6	chr4:33.9	CEU,YRI, CHB + JPT	LRH, iHS	1.7	413	161	33	0	
7	chr4:42	CHB + JPT	LRH, iHS, XP-EHH	0.3	249	94	65	6	*SLC30A9*
8	chr4:159	CHB + JPT	LRH, iHS, XP-EHH	0.3	233	67	34	1	
9	chr10:3	CEU	LRH, iHS, XP-EHH	0.3	179	63	16	1	
10	chr10:22.7	CEU, CHB + JPT	XP-EHH	0.3	254	93	0	0	
11	chr10:55.7	CHB+JPT	LRH, iHS, XP-EHH	0.4	735	221	5	2	*PCDH15*
12	chr12:78.3	YRI	LRH, iHS	0.8	151	91	25	0	
13	chr15:46.4	CEU	XP-EHH	0.6	867	233	5	1	*SLC24A5*
14	chr15:61.8	CHB + JPT	XP-EHH	0.2	252	73	40	6	*HERC1*
15	chr16:64.3	CHB + JPT	XP-EHH	0.4	484	137	2	0	
16	chr16:74.3	CHB + JPT, YRI	LRH, iHS	0.6	55	35	28	3	*CHST5, ADAT1, KARS*
17	chr17:53.3	CHB + JPT	XP-EHH	0.2	143	41	0	0	
18	chr17:56.4	CEU	XP-EHH	0.4	290	98	26	3	*BCAS3*
19	chr19:43.5	YRI	LRH, iHS, XP-EHH	0.3	83	30	0	0	
20	chr22:32.5	YRI	LRH	0.4	318	188	35	3	*LARGE*
21	chr23:35.1	YRI	LRH, iHS	0.6	50	35	25	0	
22	chr23:63.5	YRI	LRH, iHS	3.5	13	3	1	0	
		Total SNPs		16.74	9,166	2,898	480	41	

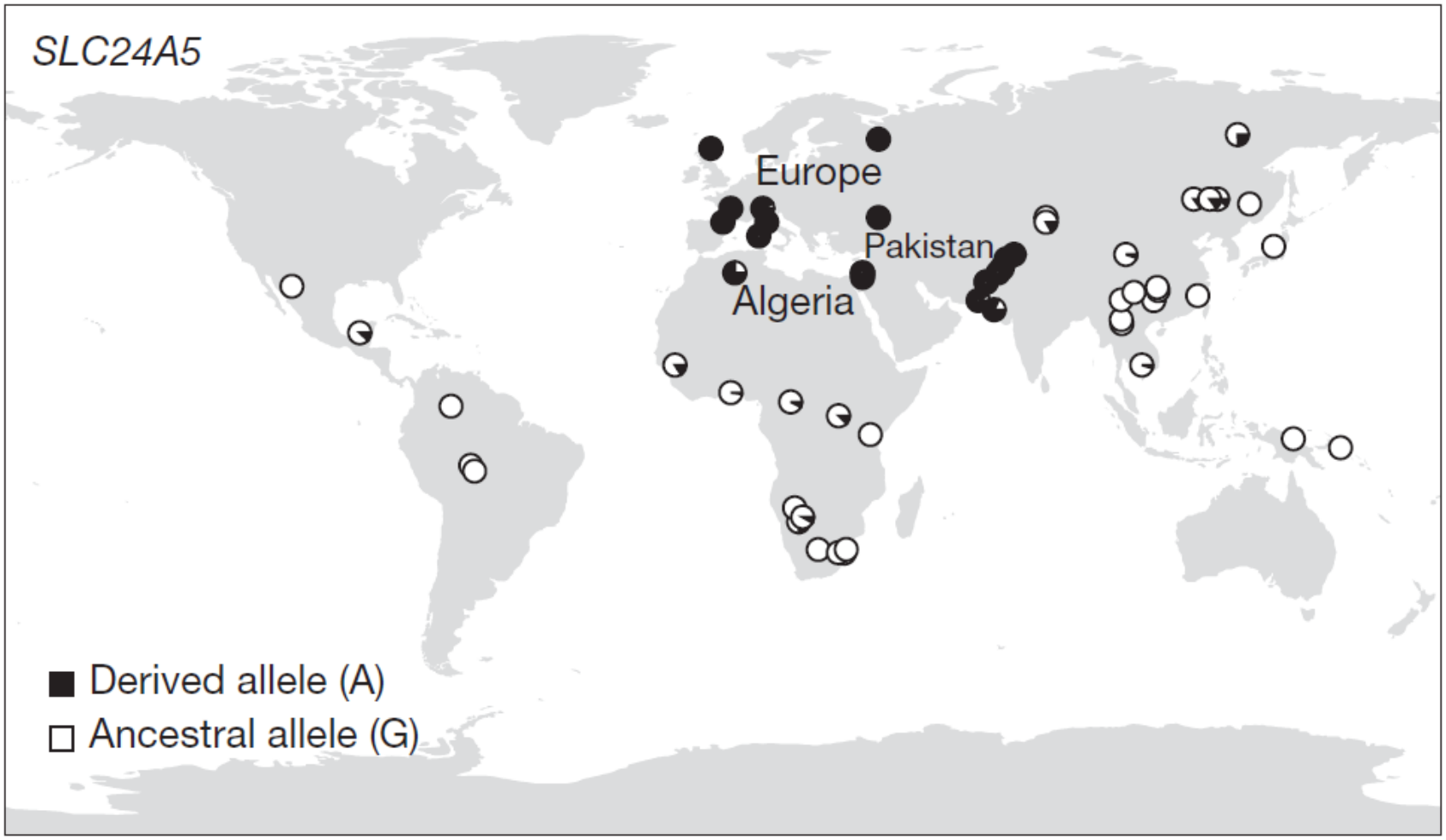
Sabeti, Varilly, Ben Fry, et. al. „Genome-wide detection and characterization of positive selection in human populations". Vol 449 Nature Letters 18 October 2007.

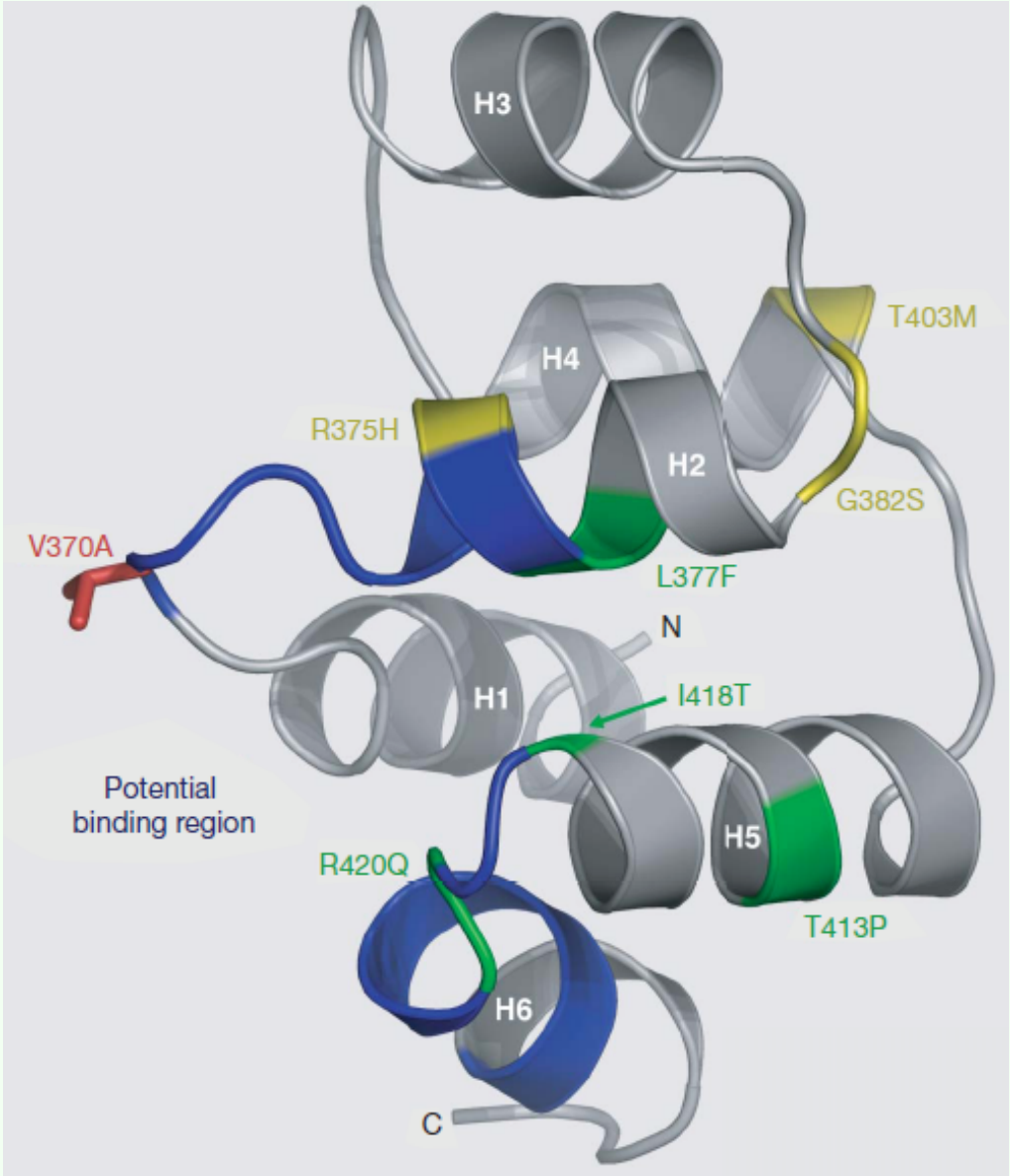

SLC24A5
DUT
FBN1
CEP152
SULT1C3
RANBP2
SLC5A7
SULT1C2
EDAR
XP-EHH and iHS
–log(P value)
867
972
e
Derived-allele
frequency
233
265
f
Population
differentiation (F_{ST})
5
7
h
SLC24A5
Alanine→threonine
EDAR
Valine→alanine
60.8
61.1
61.4
61.7
62
129.3
129.7
130.1
130.5
Position on chromosome 15 (cM)
Position on chromosome 2 (cM)

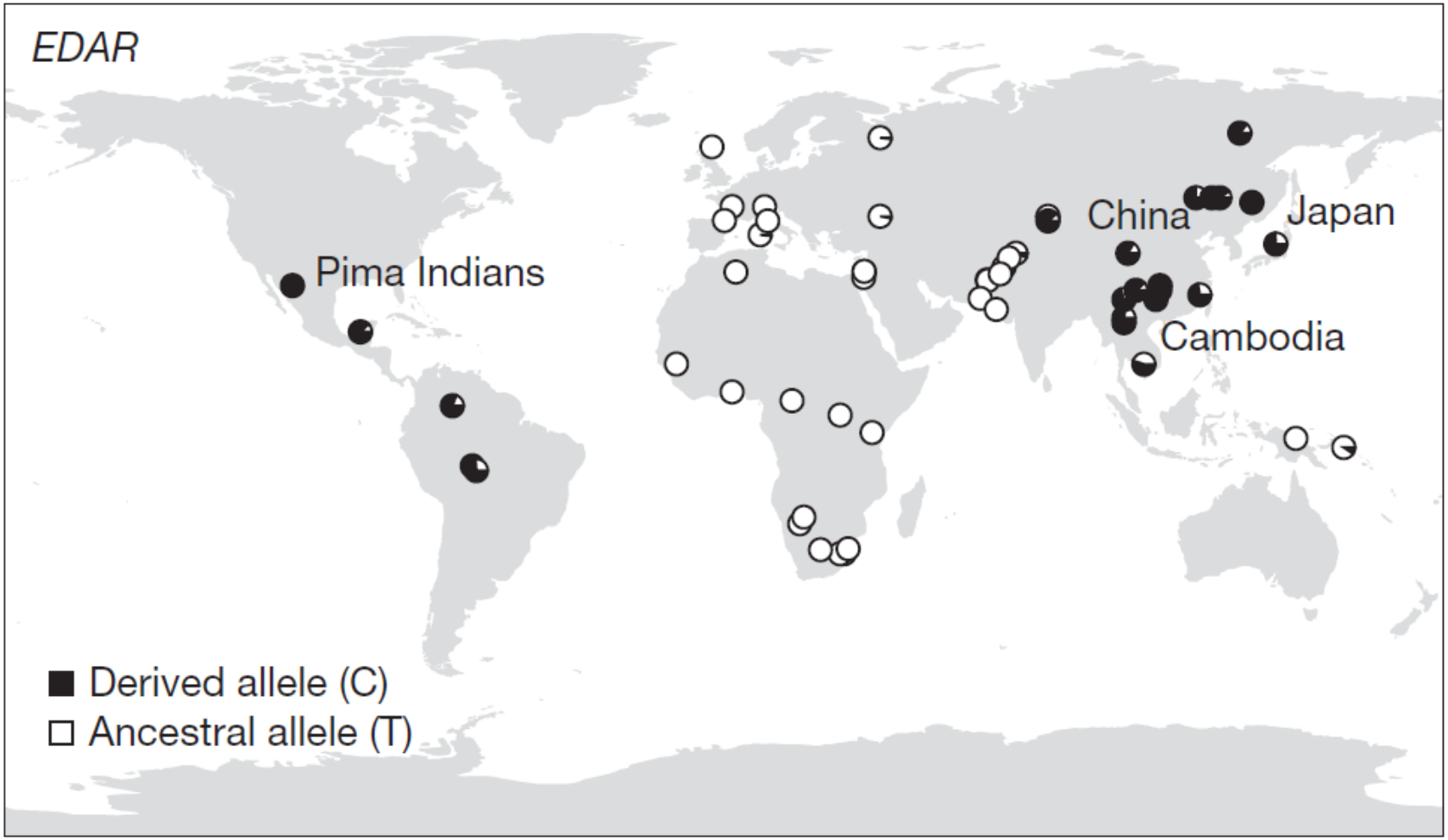
Wybrane polimorfizmy genów o różnej presji selekcyjnej.

- polimorfizm A111 genu SLC24A5 związany z różnicami w pigmencie, rozprzestrzeniony w Europie,
- polimorfizm genu LARGE, mutacja ta jest krytyczna dla wiązania się wirusa Lassa (powodującego gorączkę Lassa), rozprzestrzeniony w Afryce,
- polimorfizm genu EDAR wpływający na proces rozwoju torebek włosowych, rozprzestrzeniony w Azji i Ameryce (Indianie Pima).

Autory sami przyznają, że jedynie zbliżyli się do tematu presji selekcyjnej w historii człowieka.


SLC24A5
Europe
Pakistan
Algeria
Derived allele (A)
Ancestral allele (G)


H3
T403M
H4
R375H
H2
G382S
V370A
L377F
N
H1
I418T
Potential
binding region
H5
R420Q
T413P
H6
C


EDAR
Pima Indians
China
Japan
Cambodia
Derived allele (C)
Ancestral allele (T)

Lokalizacja QTLi

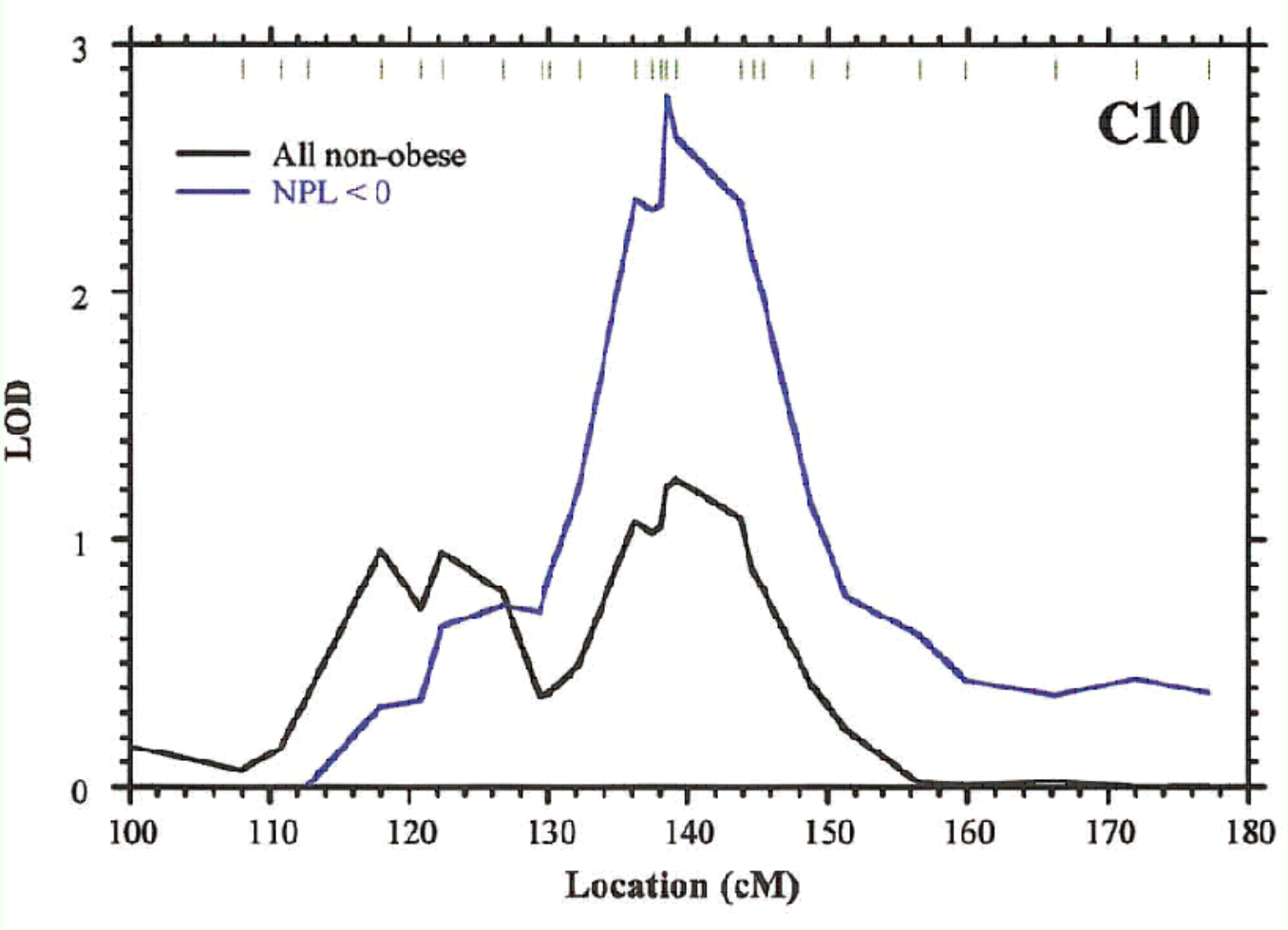
Dane SNP można zastosować do lokalizacji genów sprzężonych z cechą ilościową lub jakościową. Do oceny siły oraz istotności sprzężenia wykorzystuje się modele liniowe.
Związek pomiędzy genotypem a fenotypem opisuje się następującym modelem

$$Y_i = \mu + \sum \beta_j X_{i,j} + \varepsilon,$$

gdzie

$$X_{i,j} = \begin{cases} 1 & dlagenotypuAA \\ 0 & dlagenotypuAT \\ -1 & dlagenotypuTT \end{cases}.$$

Oceny istotności efektów testuje się testem ilorazu wiarygodności LOD.


C10
All non-obese
NPL < 0
LOD
Location (cM)
0
1
2
3
100
110
120
130
140
150
160
170
180

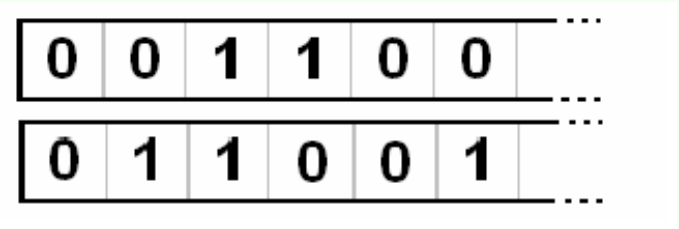
Analiza SNP w genach OR

Wyniki otrzymane we współpracy z

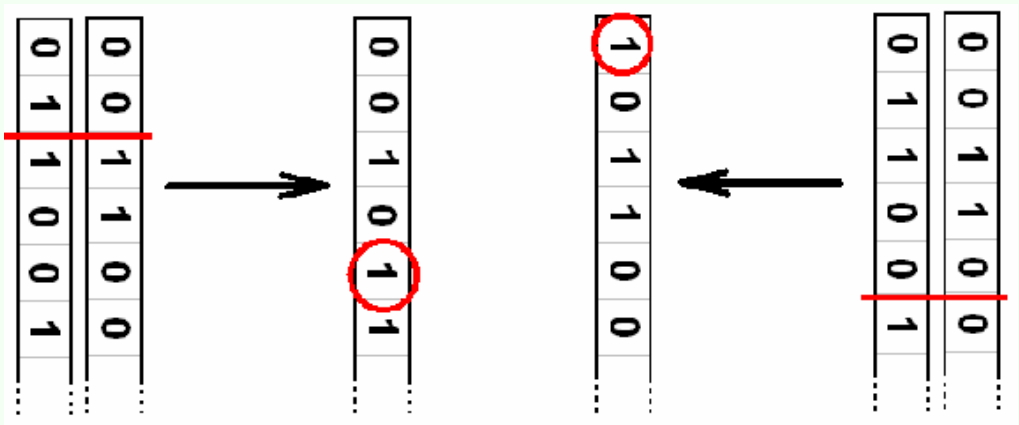
- Stanisław Cebrat,
- Paweł Mackiewicz,
- Dorota Mackiewicz,
- Marta Zawierta,
- Wojciech Waga.

Model symulacyjny

Obserwujemy ewolucje populacji N osobników. Każdy osobnik jest reprezentowany przez diploidalny M bitowy chromosom.



Dopuszczamy mutacje (wyłącznie defektywne) i rekombinacje (z częstością $p_{mut} = 1$, p_{rek} na chromosom). Genotyp 11 uznajemy za letalny i takie osobniki są usuwane z populacji.

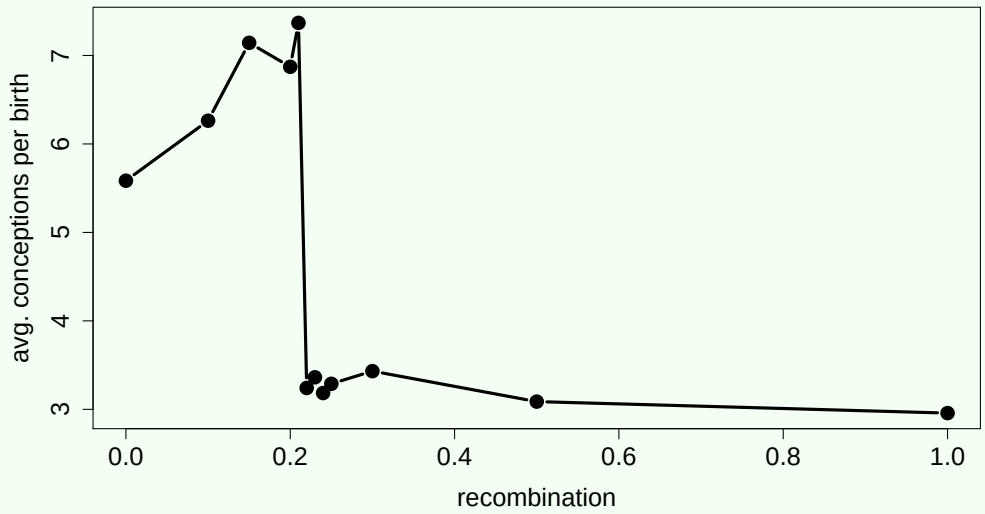


W ewolucji populacji można zaobserwować dwie strategie

- komplementującą - występuje dla niskiej częstości rekombinacji,
- oczyszczającą - występuje dla wysokiej częstości rekombinacji.

Strategia komplementująca oznacza, że w populacji zmniejsza się różnorodność, tworzą się dwa komplementarne haplotypy. W procesie produkcji gamet nowy osobnik może powstać tylko, jeżeli otrzyma komplementarne genotypy.
Strategia oczyszczająca oznacza, że osobniki z defektem są usuwane, ogólna frakcja defektów w populacji jest więc wysoka.

Śmiertelność w populacji ze strategią komplementarną jest bardzo wysoka, aby ją zmniejszyć można (w symulacjach) wprowadzić odcinek chromosomu odpowiedzialny za preselekcje gamet.

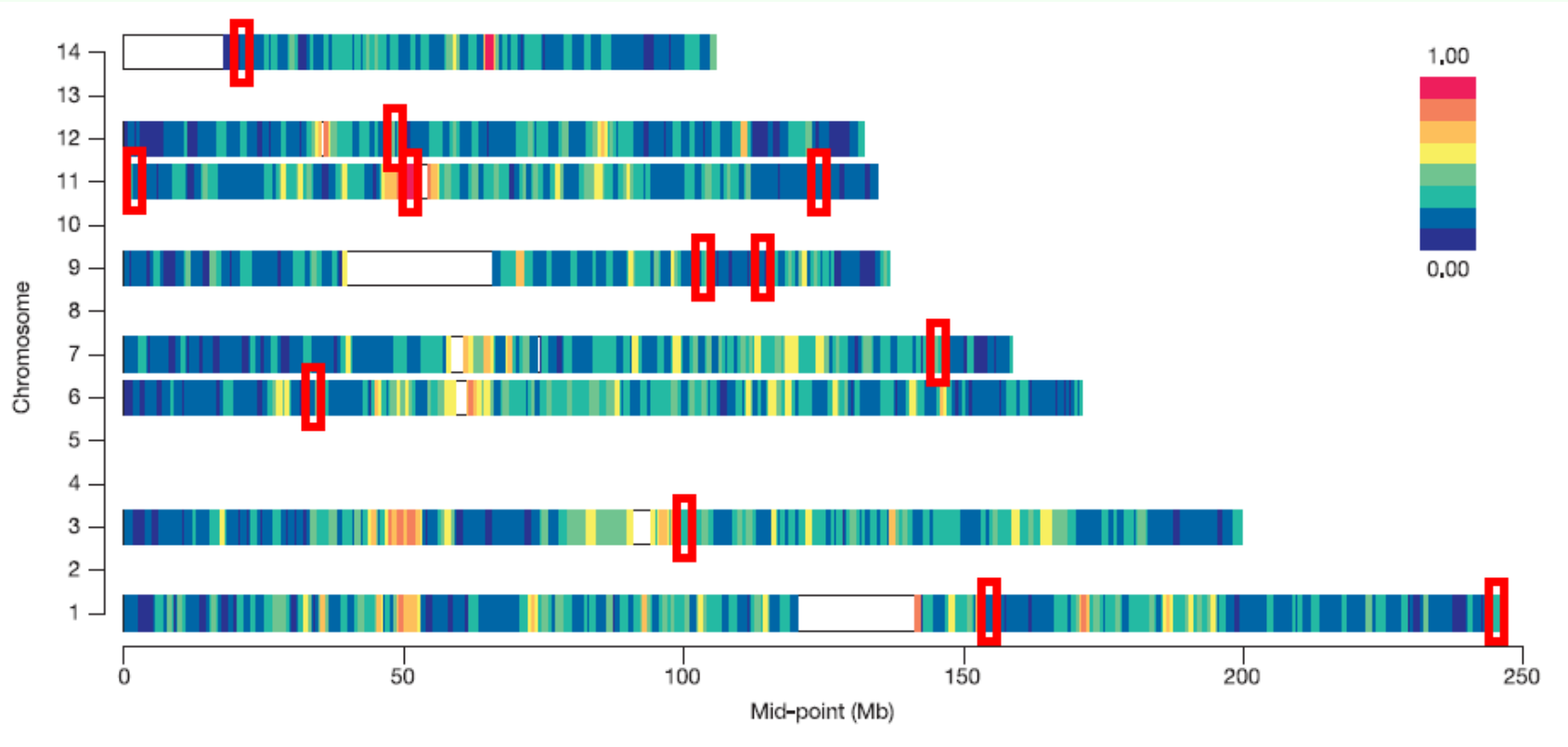


Biologiczny odpowiednik takiego mechanizmu odpowiadałby sytuacji, gdy komórka jajowa otoczona przez setki plemników mogła wybierać tego „najbardziej jej pasującego”. Taki mechanizm preselekcji jest prawdopodobny!

W genomie ludzkim znajduje się duża rodzina genów OR (Orphan Receptor). To geny o strukturze podobnej do zidentyfikowanych receptorów, ale których ligand nie został zidentyfikowany.
Geny OR rozmieszczone są na całym genomie, ale na kilku chromosomach występują w licznych grupach nazywanych CLIC (konserwatywnych z innymi organizmami).

Chr1	31, 56 genów OR,
Chr3	18 genów OR,
Chr6	34 genów OR,
Chr7	21 genów OR,
Chr11	103, 146, 42, 44 genów OR,
Chr12	28 genów OR,
Chr14	46 genów OR.

Interesująca rodzina genów



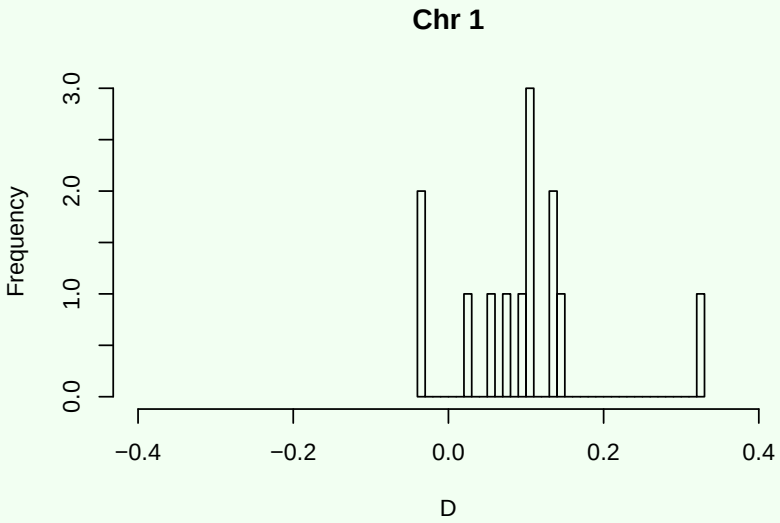
LD (linkage disequilibrium)

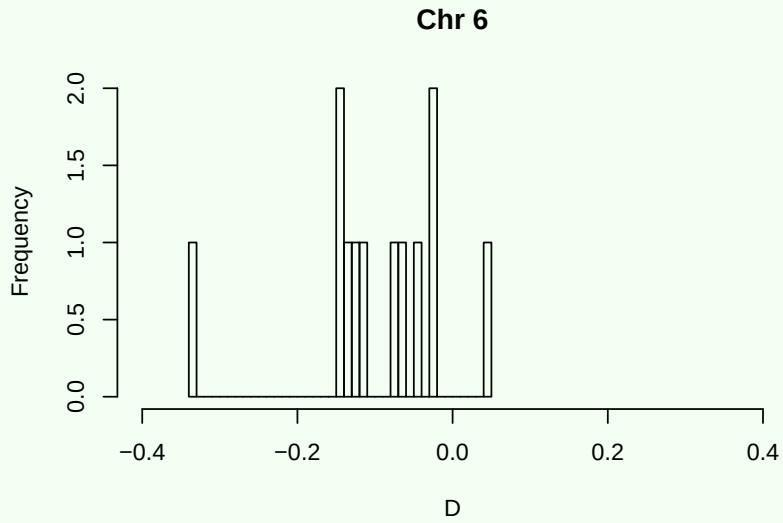
Dla pary alleli SNP współczynnik LD (oznaczane D) jest miarą odstępstwa od częstości genotypów wynikającej z losowego kojarzenia gamet. Inne miry tego odstępstwa to D', r^2, LOD.

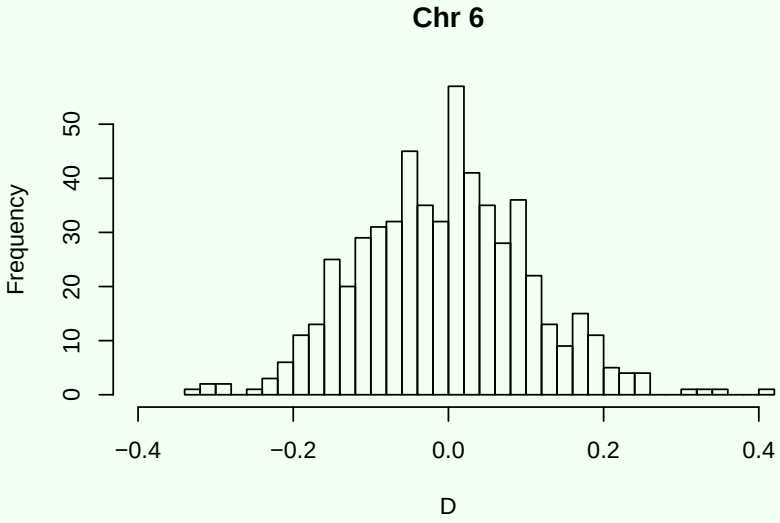
$$D = p_{AB} + p_{BA} - 2p_A p_B,$$

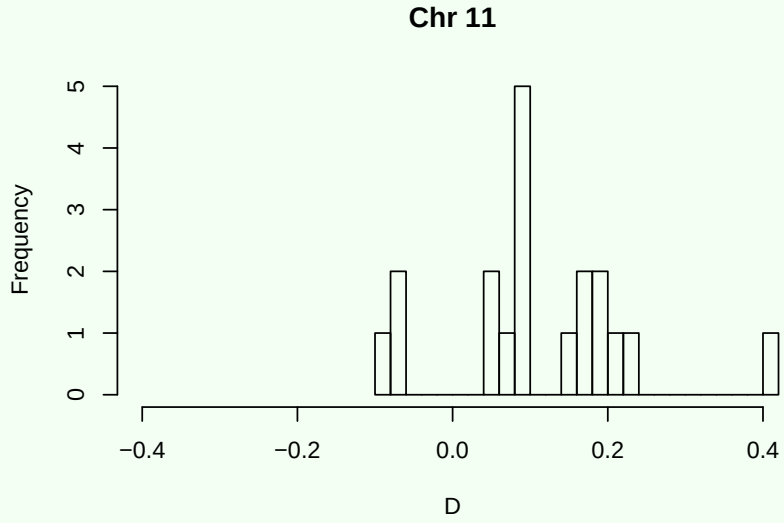
gdzie

	A	B	$\sum$
A	p_{AA}	p_{AB}	p_A
B	p_{BA}	p_{BB}	p_B
$\sum$	p_A	p_B	1


Chr 1
Frequency
0.0
1.0
2.0
3.0
−0.4
−0.2
0.0
0.2
0.4
D


Chr 6
Frequency
0.0
0.5
1.0
1.5
2.0
−0.4
−0.2
0.0
0.2
0.4
D


Chr 6
Frequency
0
10
20
30
40
50
−0.4
−0.2
0.0
0.2
0.4
D


Chr 11
Frequency
0
1
2
3
4
5
−0.4
−0.2
0.0
0.2
0.4
D

Miara LD jest czuła na założenia o homogeniczności populacji i losowm kojarzeniu. Dlatego wyniki zostaną zweryfikowane zmodyfikowanym testem TDT (transmission disequilibrium test).

	AA	AB	BB
AA	—	Pr(dziecko = AB)	—
AB	Pr(dziecko = AB)	Pr(dziecko=AB)	Pr(dziecko=AB)
BB	—	Pr(dziecko = AB)	—

Będziemy obserwować jedynie infromatywne przekazywanie gamet.

Zmodyfikowany test TDT (transmission disequilibrium test)

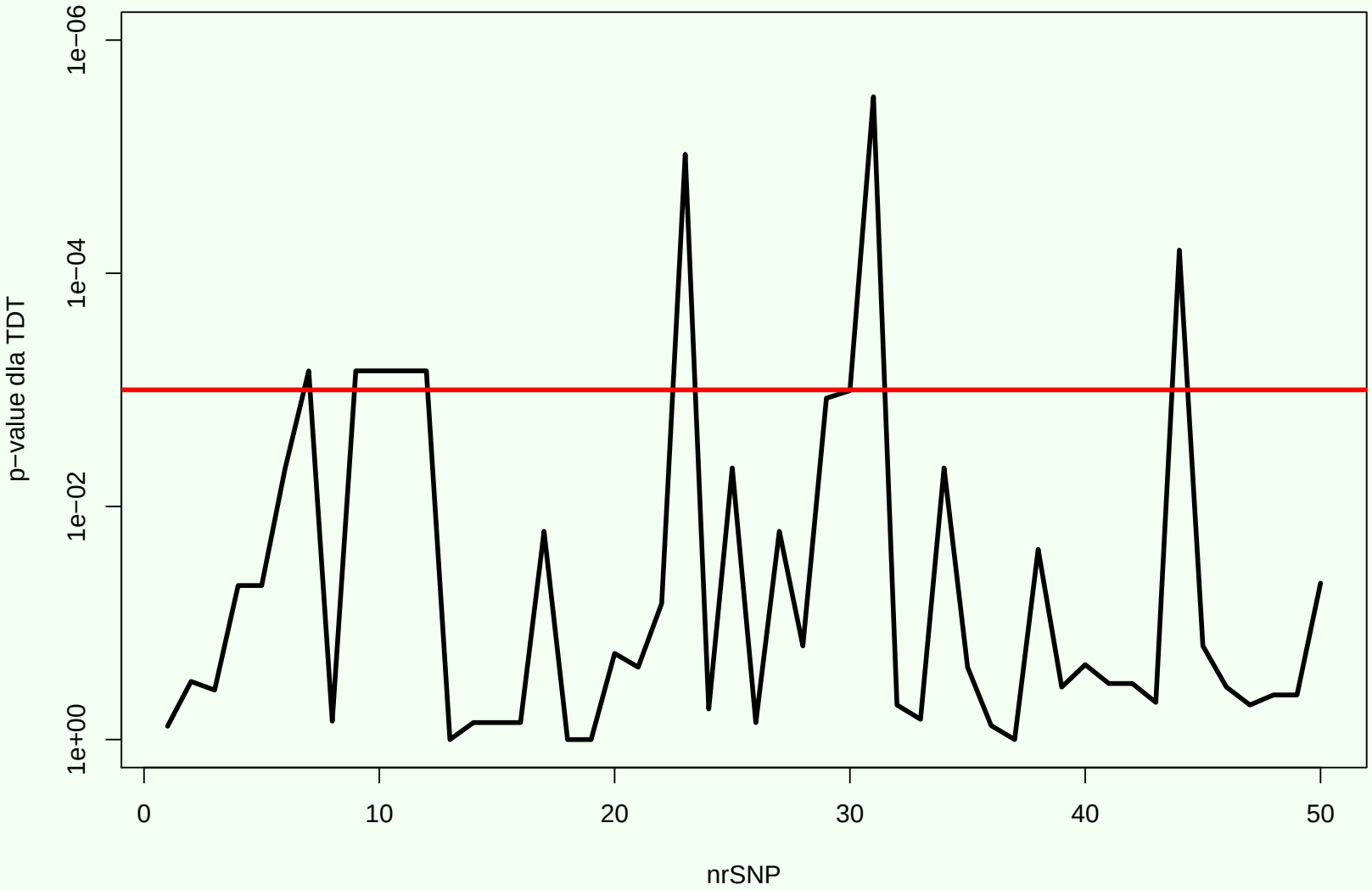
Określmy statystykę

$$S = \frac{(2L_2 - L_1)^2}{L_1}$$

gdzie

L_1= #{trójek w których przynajmniej jeden rodzic jest heteroalleliczny}
L_2= #trójek w których dziecko jest heteroalleliczne .

Można pokazać, że w przypadku losowej transmisji gamet asymptotyczny rozkład statystyki S jest rozkładem χ^2_1.


p−value dla TDT
1e−06
1e−04
1e−02
1e+00
0
10
20
30
40
50
nrSNP

Podsumowanie

- Symulacje komputerowe sugerują, że preselekcja gamet na etapie zapłodnienia komórki jest możliwa i niesie korzyści związane z tempem ewolucji (przynajmniej w komputerowym modelu).
- Analizy danych z projektu HapMap są zgodne z wynikami modelu symulacyjnego.
- Projekt HapMap to olbrzymia baza danych pozwalających na wykonywanie ciekawych analiz.