

Poszukiwanie genetycznych podstaw choroby

Analiza sprzężeń i analiza asocjacji

Jakie cechy badamy

- Choroby - zmienność patologiczna
 - genetyczne - zależne od zmian w genach
 - dziedziczne - nie wszystkie choroby genetyczne są dziedziczne!
 - wrodzone - nie wszystkie choroby wrodzone są dziedziczne lub genetyczne!
 - inne - choroby, w których jest składowa genetyczna i składowa środowiskowa
- Zmienność prawidłowa
 - każdy z nas jest inny

Jak dziedziczą się cechy człowieka?

- Mendlowskie - jeden gen = jedna cecha
 - zmienność konkretnego genu decyduje o fenotypie konkretnej cechy
 - w przypadku chorób - tzw. mutacje sprawcze
- Wielogenowe - zależne od działania kilku - kilkunastu genów
 - np. kolor oczu, włosów
- Wieloczynnikowe - zależą od współdziałania wielu (setek, tysięcy) genów i środowiska

Jak dziedziczą się cechy człowieka?

- Mendlowskie - jeden gen = jedna cecha
 - dobrze potrafimy identyfikować geny, analizować dziedziczenie, wykrywać mutacje i przewidywać fenotyp
- Wielogenowe - zależne od działania kilku - kilkunastu genów
 - potrafimy analizować i przewidywać fenotyp, ale nie ze 100% dokładnością
- Wieloczynnikowe - zależą od współdziałania wielu (setek, tysięcy) genów i środowiska
 - nie potrafimy dobrze przewidywać, dopiero zaczynamy poznawać złożoność i odkrywać korelacje statystyczne

Jak dziedziczą się cechy człowieka?

- Mendlowskie - jeden gen = jedna cecha
 - znamy wiele chorób, które tak się dziedziczą, ale są to choroby rzadkie
 - tylko pojedyncze przykłady cech zmienności prawidłowej
- Wieloczynnikowe - zależą od współdziałania wielu (setek, tysięcy) genów i środowiska
 - praktycznie wszystkie aspekty zmienności prawidłowej
 - większość często występujących chorób

Badanie relacji genotyp-fenotyp u człowieka

- Analiza sprzężeń - poszukiwanie rejonów chromosomu położonych blisko genu determinującego daną cechę
 - Allele o dużej penetracji (allele sprawcze) - cechy mendlowskie
- Analiza asocjacji - poszukiwanie korelacji polimorfizmów genetycznych z występowaniem fenotypu
 - Także allele o niskiej penetracji (przy dużych zbiorach danych)
 - Cechy wieloczynnikowe

Asocjacje a sprzężenie

- Sprzężenie - wspólna segregacja alleli genów leżących blisko siebie na chromosomie
 - dotyczy loci, nie konkretnych alleli
 - proste podłoże biologiczne (chromosomy, rekombinacja)
 - badana w rodowodach i/lub parach krewnych
- Dotyczy cech mendlowskich - wysoka odziedziczalność, mutacje sprawcze pojedynczych genów

Asocjacja a sprzężenie

- Osoby z allelem a jakiegoś genu mają większe prawdopodobieństwo fenotypu X
 - Czy zawsze asocjacja oznacza zależność przyczynową?
 - Czy każda asocjacja ma wartość diagnostyczną?
 - Czy asocjacja odkrywa “gen na”?
- Odpowiedzi na te i inne pytania - wykład

Asocjacje a sprzężenie

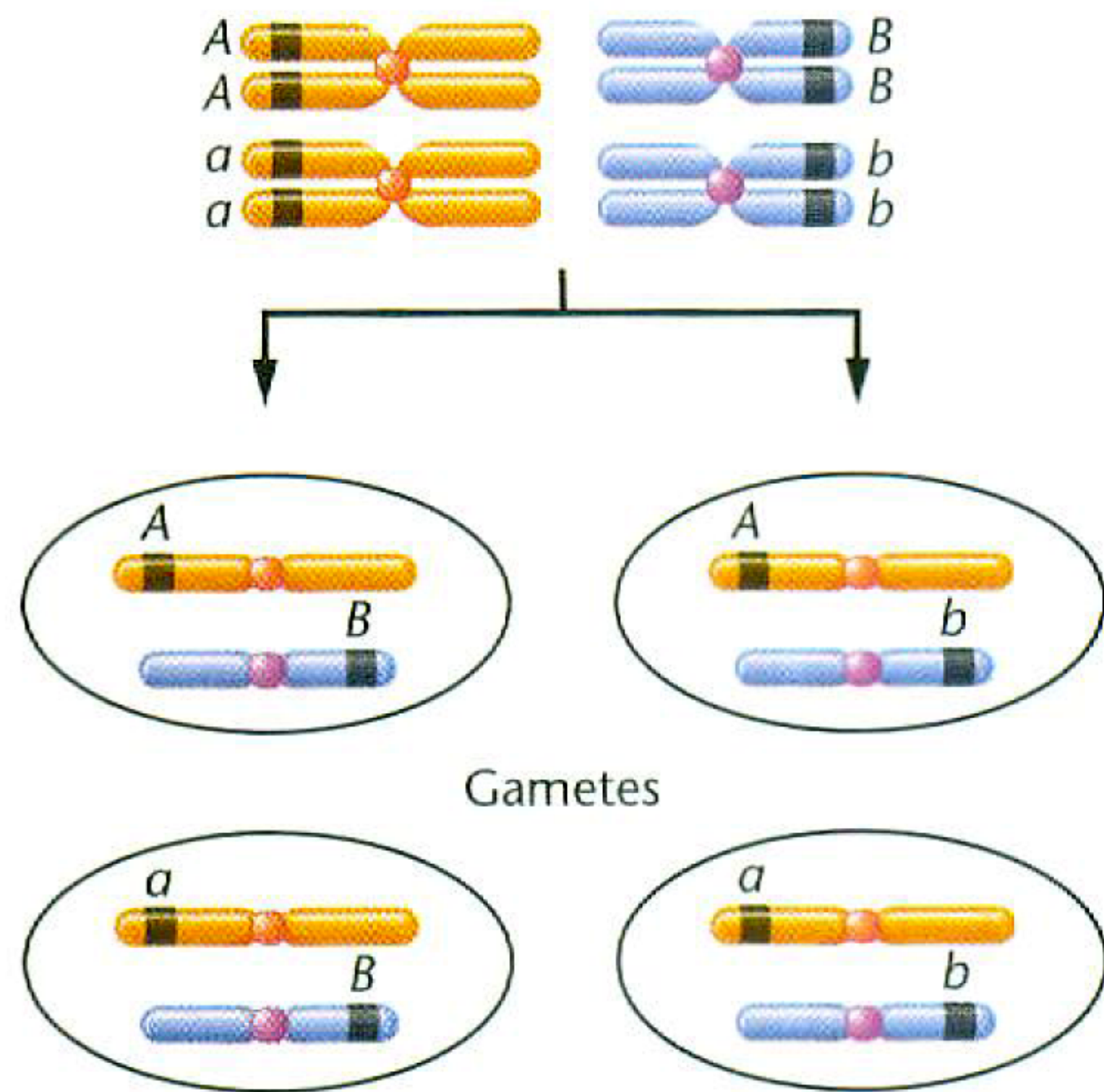
- Asocjacja - korelacja występowania konkretnych alleli genów w populacji
 - dotyczy konkretnych alleli
 - często złożone i/lub niejasne podłoże biologiczne - zjawisko statystyczne, niekiedy bez związku przyczynowego
 - dotyczy populacji lub grupy, ale bez wymogu pokrewieństwa
 - może niekiedy być związana ze sprzężeniem (nierównowaga sprzężeń)
 - dotyczy cech wieloczynnikowych

Asocjacja a sprzężenie

Asocjacja	Sprzężenie
Dotyczy alleli	Dotyczy genów (<i>loci</i>)
Na poziomie populacji	W rodzinie
Dziedziczenie wieloczynnikowe	Dziedziczenie mendlowskie - mutacje sprawcze

Geny i chromosomy

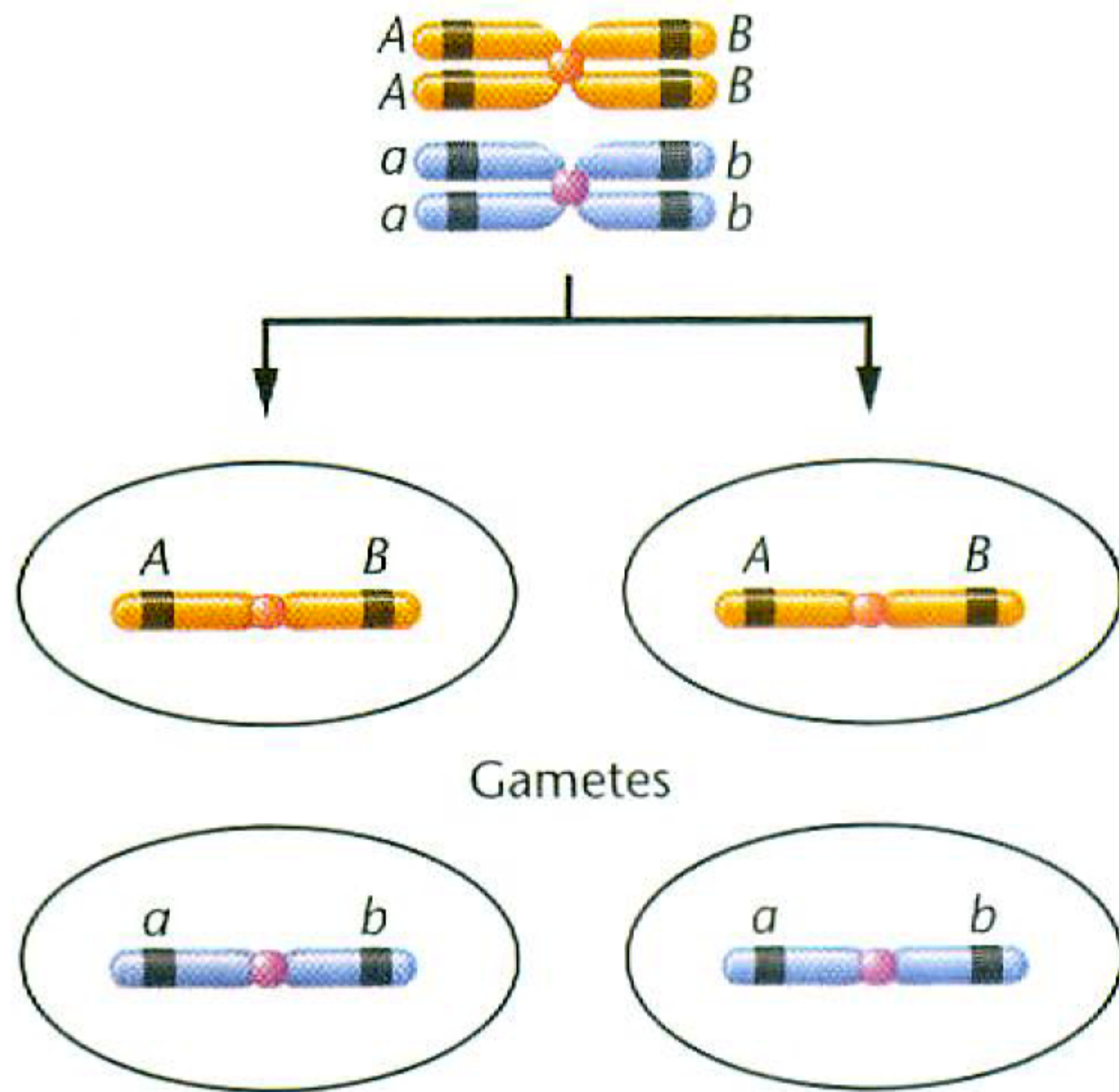
Allele genów zlokalizowanych na różnych chromosomach segregują niezależnie (II prawo Mendla)



Dla 2 genów:
4 równoliczne klasy gamet

Sprzężenie

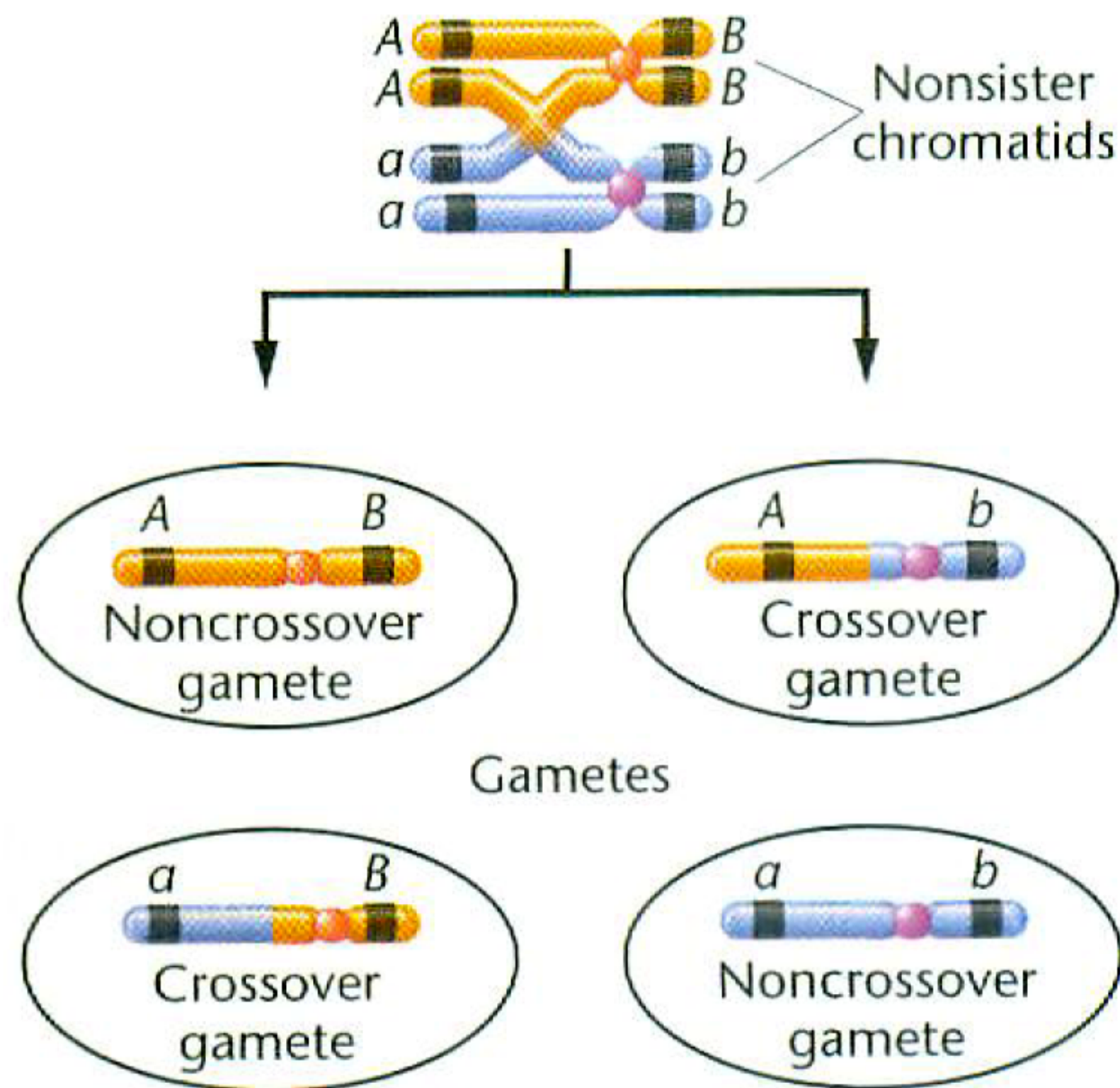
Allele genów leżących na tym samym chromosomie dziedziczą się razem - sprzężenie



Dla 2 genów i pełnego sprzężenia
2 równoliczne klasy gamet rodzicielskich

Sprzężenie

Crossing-over (rekombinacja chromatyd niesiostrzanych w mejozie)



Dla 2 genów:

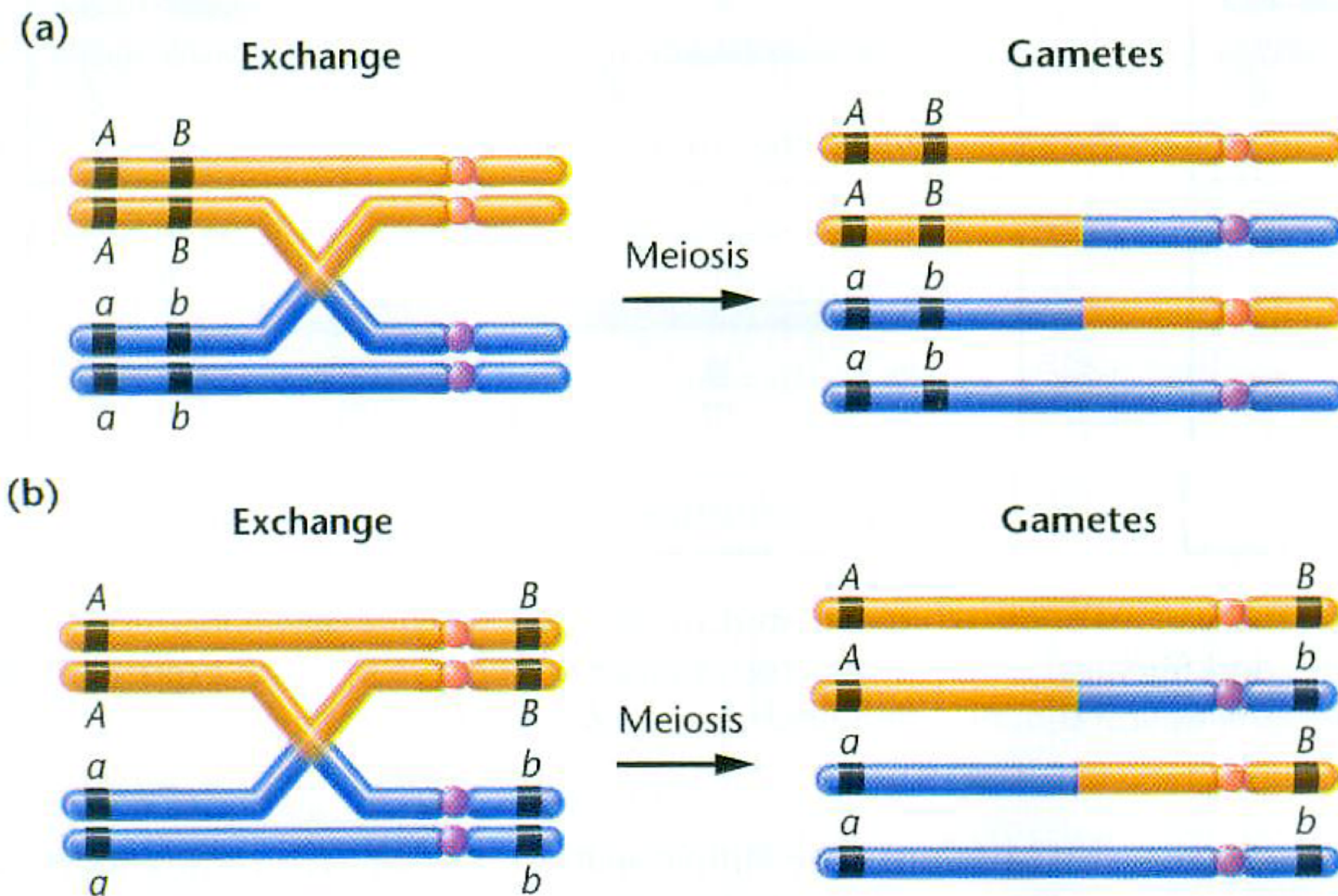
2 równoliczne klasy gamet rodzicielskich

2 równoliczne klasy gamet rekombinowanych

Klasy rekombinowane mniej liczne od rodzicielskich

Mapowanie genetyczne

Aby powstały rekombinowane gamety, crossing-over musi zająć **pomiędzy** genami (*loci*)



powstają gamety rodzicielskie

powstają gamety rekombinowane

Liczba rekombinantów jest miarą odległości genetycznej

Mapowanie genów

- Prawdopodobieństwo crossing-over pomiędzy genami jest proporcjonalne do odległości między nimi na chromosomie
- Liczebność klas zrekombinowanych w potomstwie jest miarą odległości genetycznej
- U *Drosophila* najlepiej mapować za pomocą krzyżówki heterozygotycznej samicy i samca recesywnego
- A u człowieka?

Metody

- Sprzężenie: analiza sprzężeń (mapowanie)
 - metody parametryczne - dla dużych rodowodów, nawet niewielu
 - metody nieparametryczne - dla dużej liczby małych rodowodów
- Asocjacje - badanie korelacji (testy statystyczne)

Analiza sprzężeń u człowieka

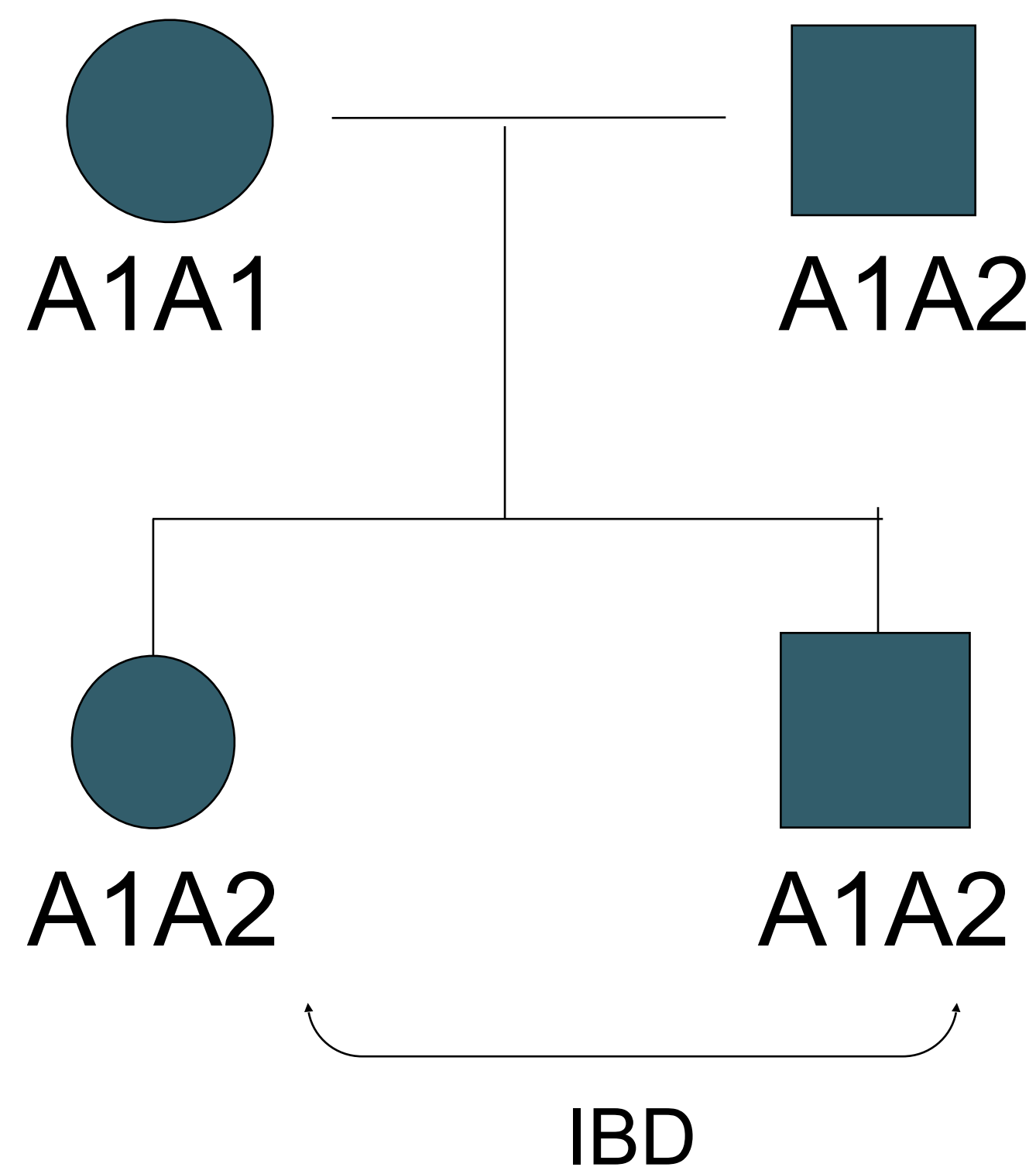
- Geny człowieka są rozdzielone długimi obszarami międzygenowymi
- Sprzężenie pomiędzy genami, których allele dają obserwowalne fenotypy jest bardzo rzadkie
- Wykorzystuje się markery molekularne (RFLP, VNTR, etc.)
 - mapy genetyczne człowieka (np. CEPH, HapMap)
 - poszukiwanie markera sprzężonego z locus choroby

Metody mapowania

- Parametryczne (oparte na modelach dziedziczenia)
 - dwupunktowa
 - wielopunktowa
- Nieparametryczna analiza sprzężeń
 - czy jest statystycznie istotne odchylenie od założenia o niezależnym przekazywaniu alleli (II prawo Mendla) dla danych *loci* w populacji
 - współwystępowanie (korelacje) alleli u spokrewnionych osób

Analiza nieparametryczna

Dwa allele są identyczne przez pochodzenie (IBD) jeżeli są kopiami tego samego allelu rodzicielskiego



Metody nieparametryczne

- Korelacja względnego podobieństwa u par mapowanej cechy z podobieństwem markera
 - Badania bliźniąt
 - Badania chorego rodzeństwa (*affected siblings method*):
 - czy w parach chorych krewnych allele markera (nieważne które) są wspólne częściej, niż w reszcie populacji?
 - Badania niewielkich rodowodów (z pojedynczymi chorymi)

Metody parametryczne

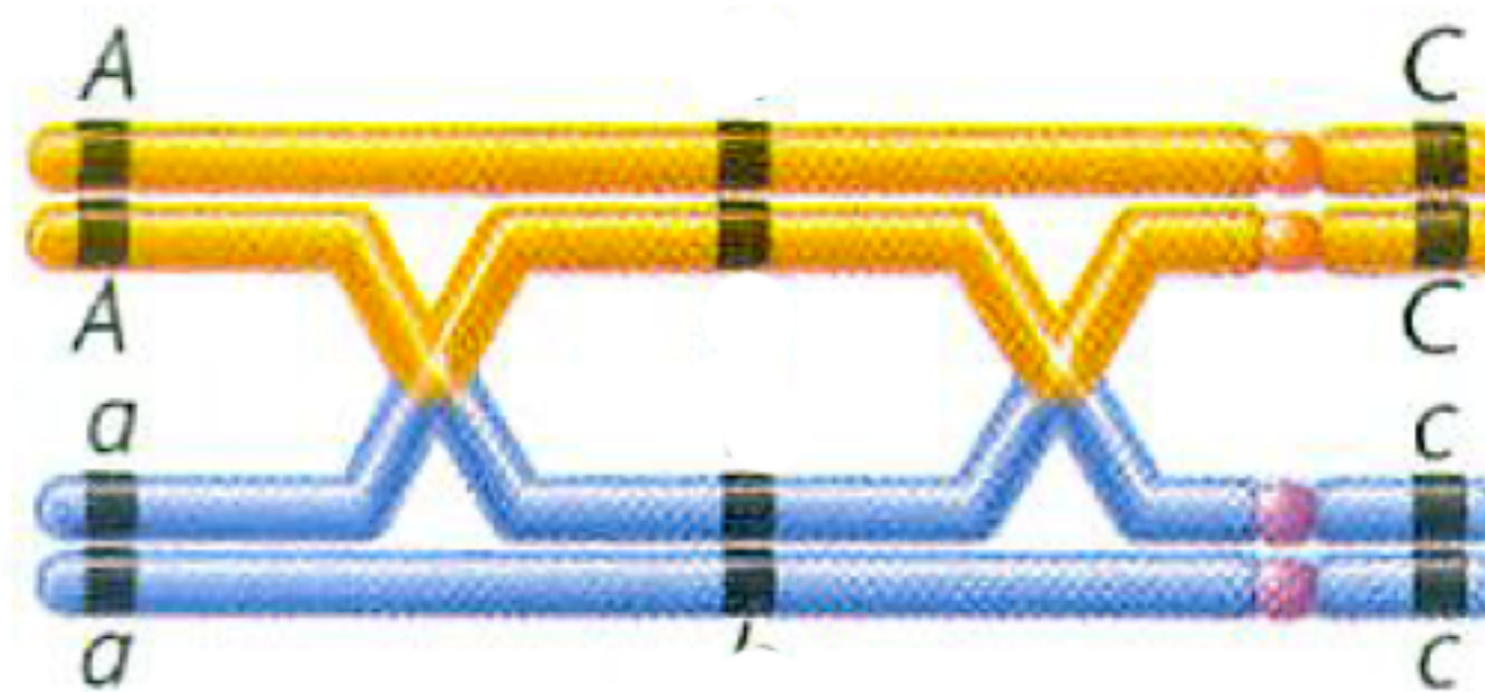
- Zakładamy, że jest znany model dziedziczenia
 - 2 lub więcej genów i/lub markerów - *loci*
 - genotypy rodziców i genotypy potomstwa
 - dane (liczba rekombinantów) pozwalają na wyznaczenie parametrów modelu
 - odległości między *loci*

Miarą odległości jest częstość rekombinacji

- Częstość rekombinacji θ = prawdopodobieństwo przekazania rekombinowanej gamety
- Loci na różnych chromosomach segregują niezależnie
 $\Rightarrow \theta = 0,5$
- Loci blisko sprzężone segregują razem
 $\Rightarrow \theta = 0$
- Zatem
 - $\theta < 0,5$ sprzężenie
 - $\theta = 0,5$ brak sprzężenia

Mapowanie

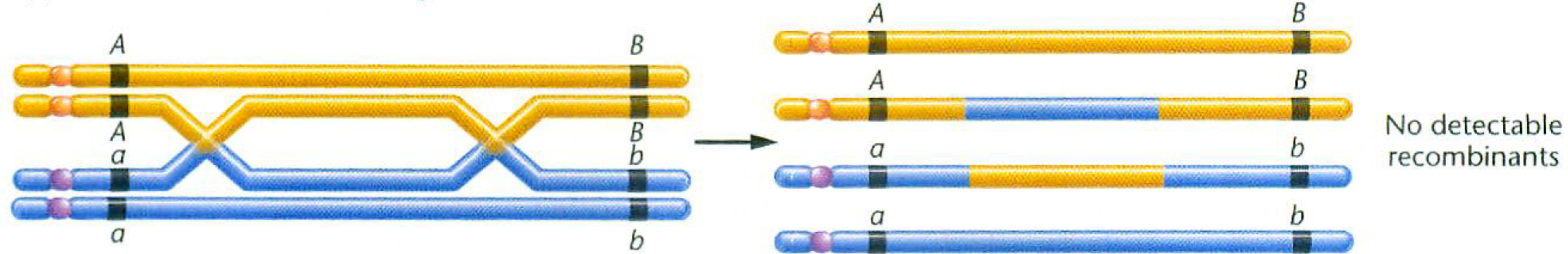
- Jednostka: cM (centymorgan) = 1% rekombinacji
- Zależność nie jest liniowa



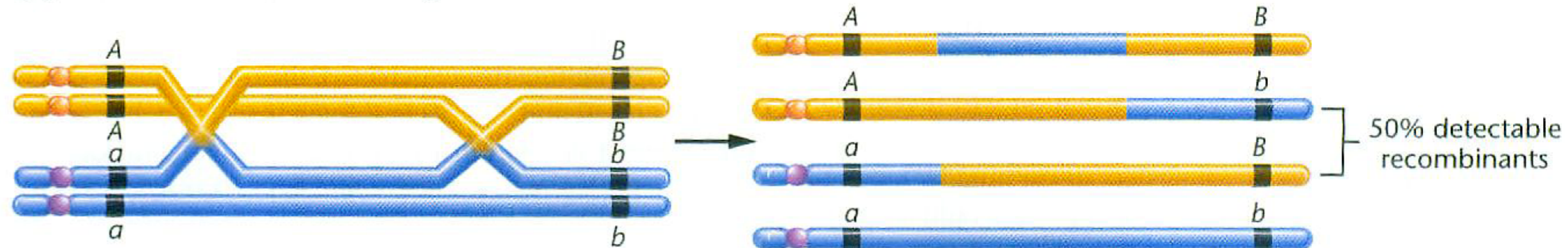
- Podwójny crossing-over – gamety typu rodzicielskiego
- Interferencja – crossing-over w danym miejscu zmienia prawdopodobieństwo zajścia kolejnego w pobliżu

Podwójny c-o – złożony obraz

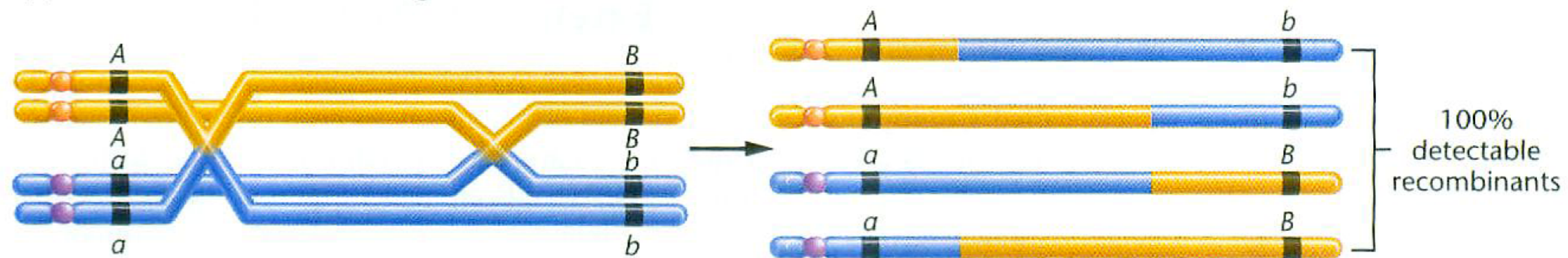
(a) Two-strand double exchange



(b) Three-strand double exchange



(c) Four-strand double exchange



Średnio 50% rekombinantów. Podobnie dla potrójnego c-o, itp.

Funkcja mapowa

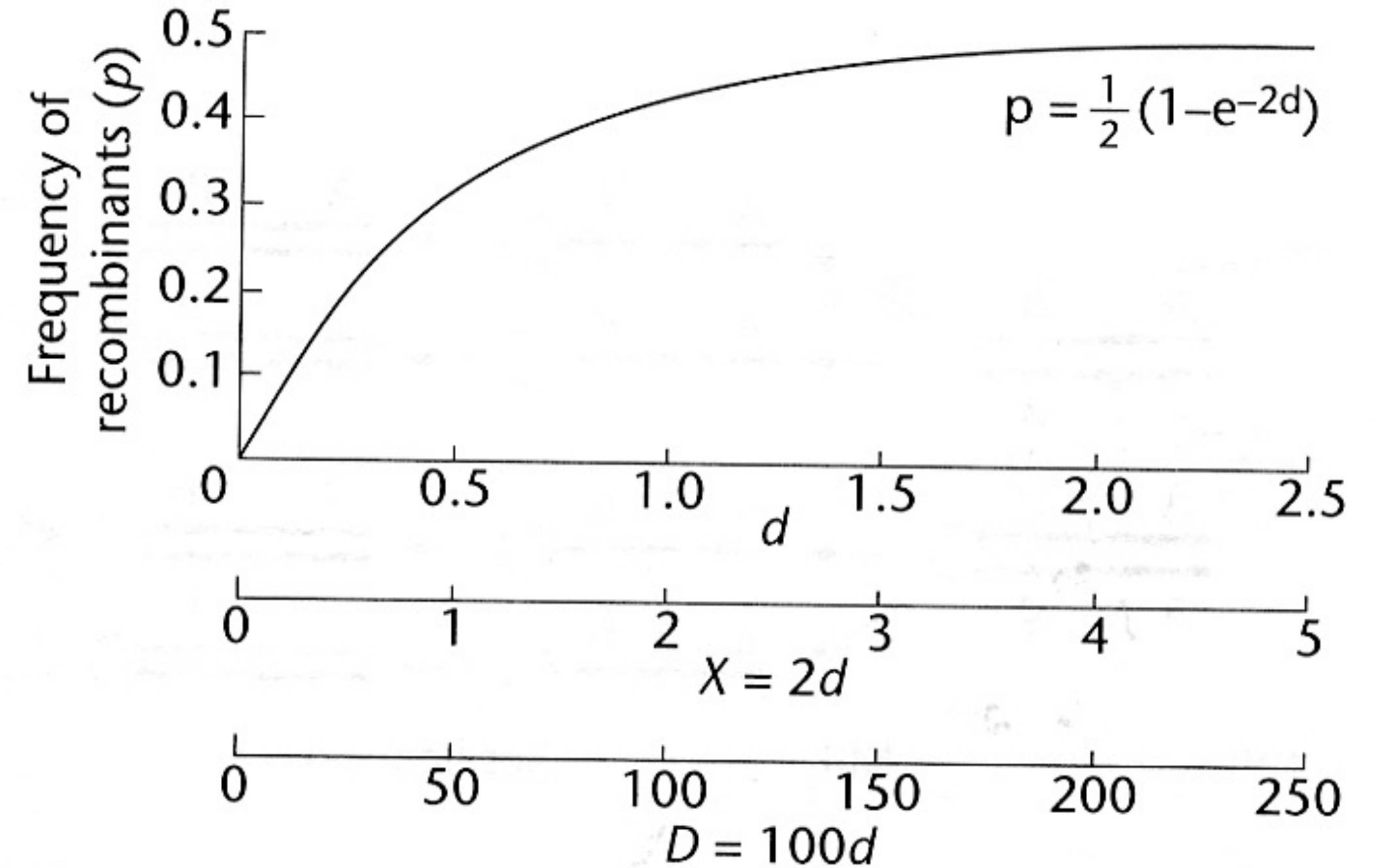
- Zależność odległości genetycznej od częstości rekombinacji
- Funkcja mapowa Haldane'a
 - wielokrotne c-o, bez interferencji
- Funkcja Kosambi'ego
 - uwzględnia też interferencję, szeroko stosowana
- Dla małych θ : $d \approx \theta$

$$d = \frac{\ln(1 - 2\theta)}{2}$$

$$d = \frac{\ln\left(\frac{1 + 2\theta}{1 - 2\theta}\right)}{4}$$

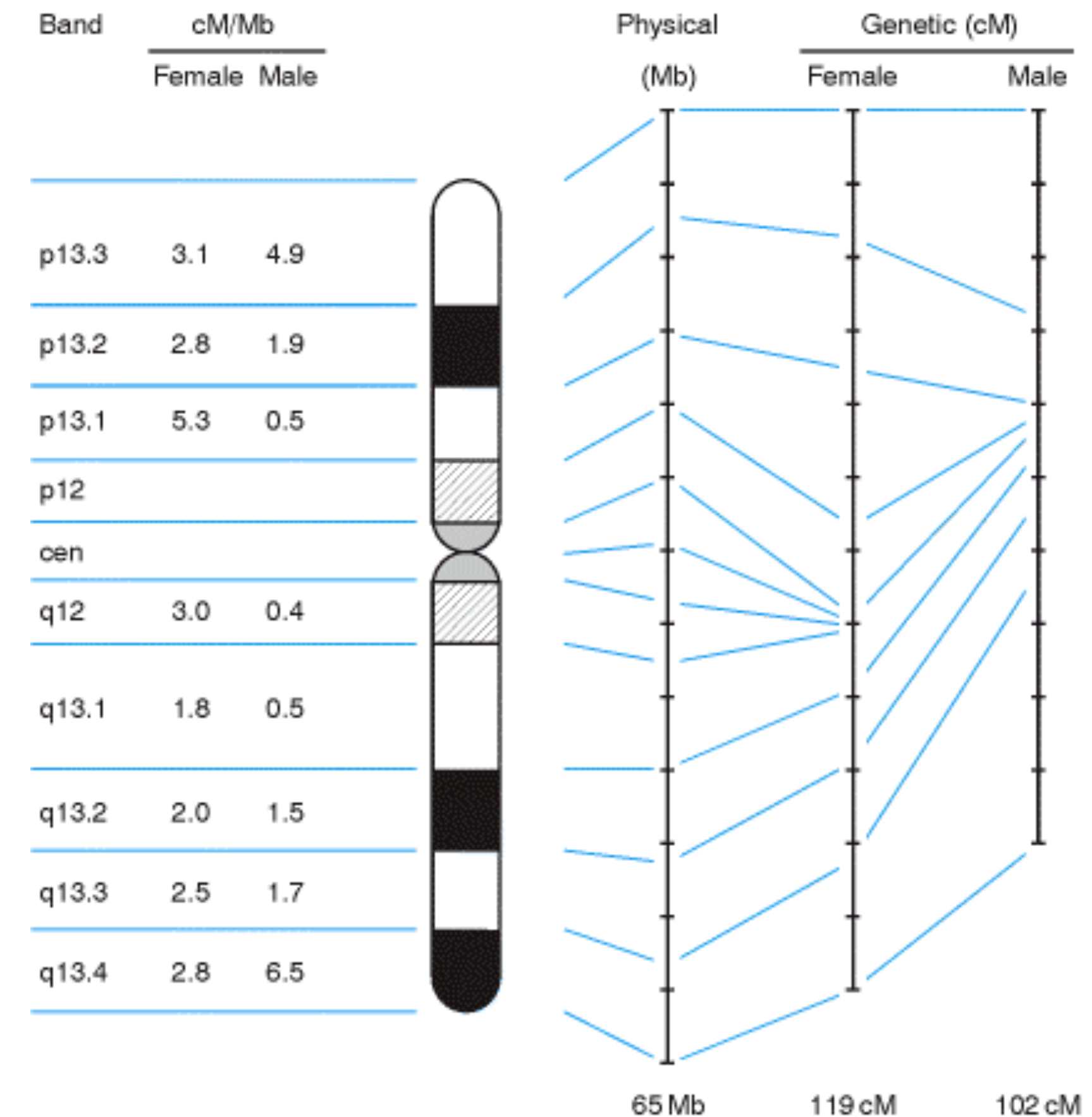
Funkcja mapowa

- Wraz ze wzrostem odległości częstość c-o dąży do 0,5
- Dla genów niesprzężonych “rekombinantów” jest 50%, podobnie jak dla genów leżących w dużej odległości



Płeć a częstość rekombinacji

- Całkowita mapa mężczyzny = 2851cM (autosomy)
- Całkowita mapa kobiety = 4296 cM (autosomy)
- Dla ~3000Mb genomu autosomów
 - 1 cM u mężczyzn $\approx 1,05$ Mb
 - 1 cM u kobiet $\approx 0,7$ Mb
 - średnia 1 cM $\approx 0,88$ Mb
- stosunek różny w różnych obszarach genomu



Analiza parametryczna

- U *Drosophila* najprościej skrzyżować samicę podwójną heterozygotę z samcem podwójnie recesywnym i policzyć klasy w potomstwie.
- A u człowieka?
 - nie zakłada się krzyżówek doświadczalnych
 - mało liczne potomstwo, długi czas generacji (25 lat/pokolenie)

Wiarygodność (*likelihood*)

- Wiarygodność: prawdopodobieństwo uzyskania zaobserwowanych danych przy założeniach modelu i jego określonych parametrach

Wiarygodność (*likelihood*)

- W rodowodzie w pełni informatywnym
 - dane: R =liczba rekombinantów; NR =liczba genotypów rodzicielskich
 - parametr: częstość (prawdopodobieństwo) rekombinacji θ
- Hipoteza zerowa – brak sprzężenia ($\theta=0,5$)
- Stosunek wiarygodności dla danej wartości θ : $L(\theta)/L(\theta=0,5)$
- lod score (Z) = *logarithm of odds* – logarytm dziesiętny stosunku wiarygodności

Proste przykłady obliczeń *lod*

Dla danego rodowodu (i), *lod* dla danego θ wynosi:

$$Z_i(\theta) = \log_{10} \frac{L(\text{rodowód} / \theta)}{L(\text{rodowód} / \theta = 0,5)}$$

Dla danej wartości θ , sumuje się *lod*-score z różnych rodowodów (F):

$$Z(\theta) = \sum_{i=1}^F Z_i(\theta)$$

Analiza dwupunktowa

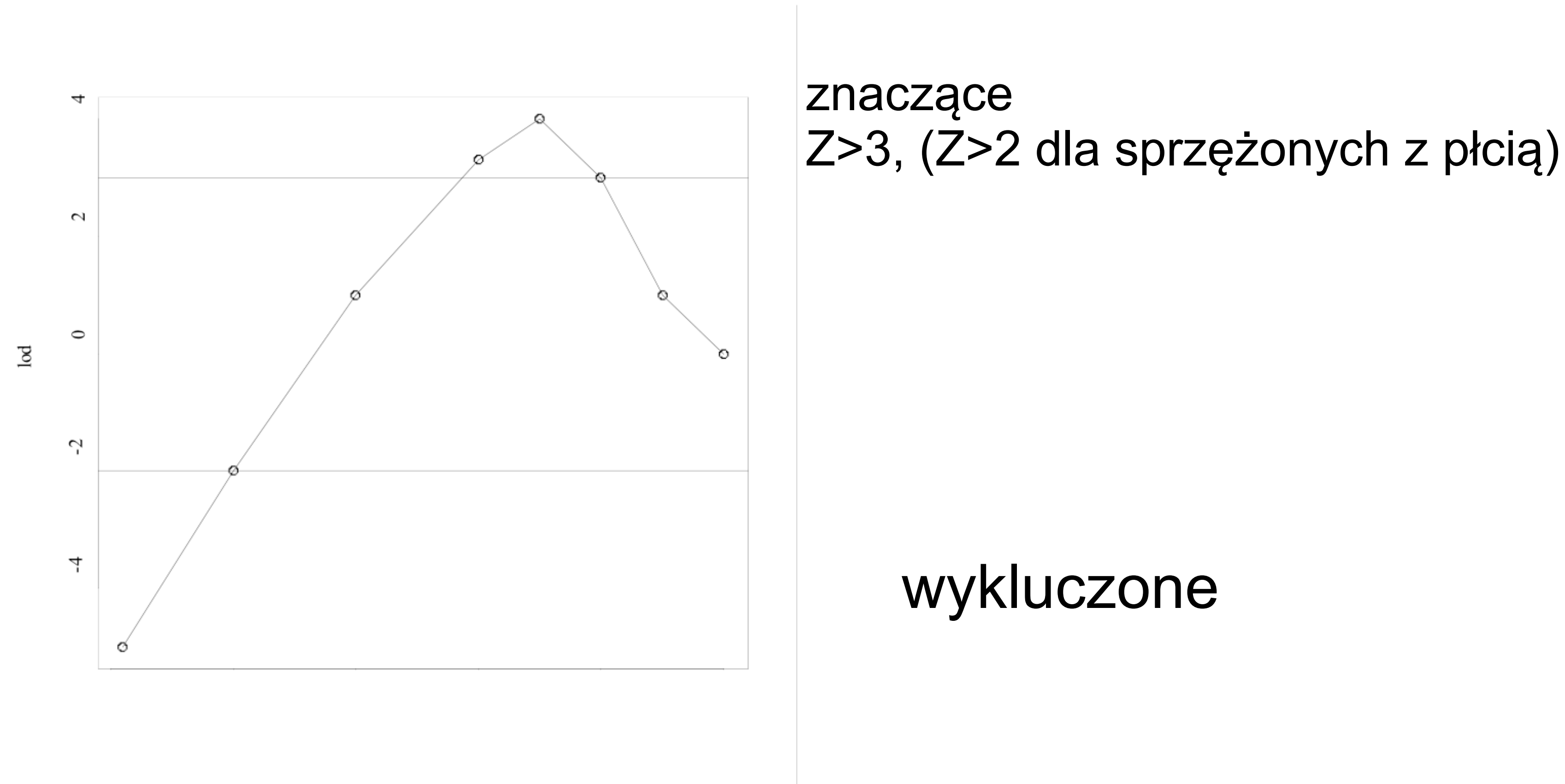
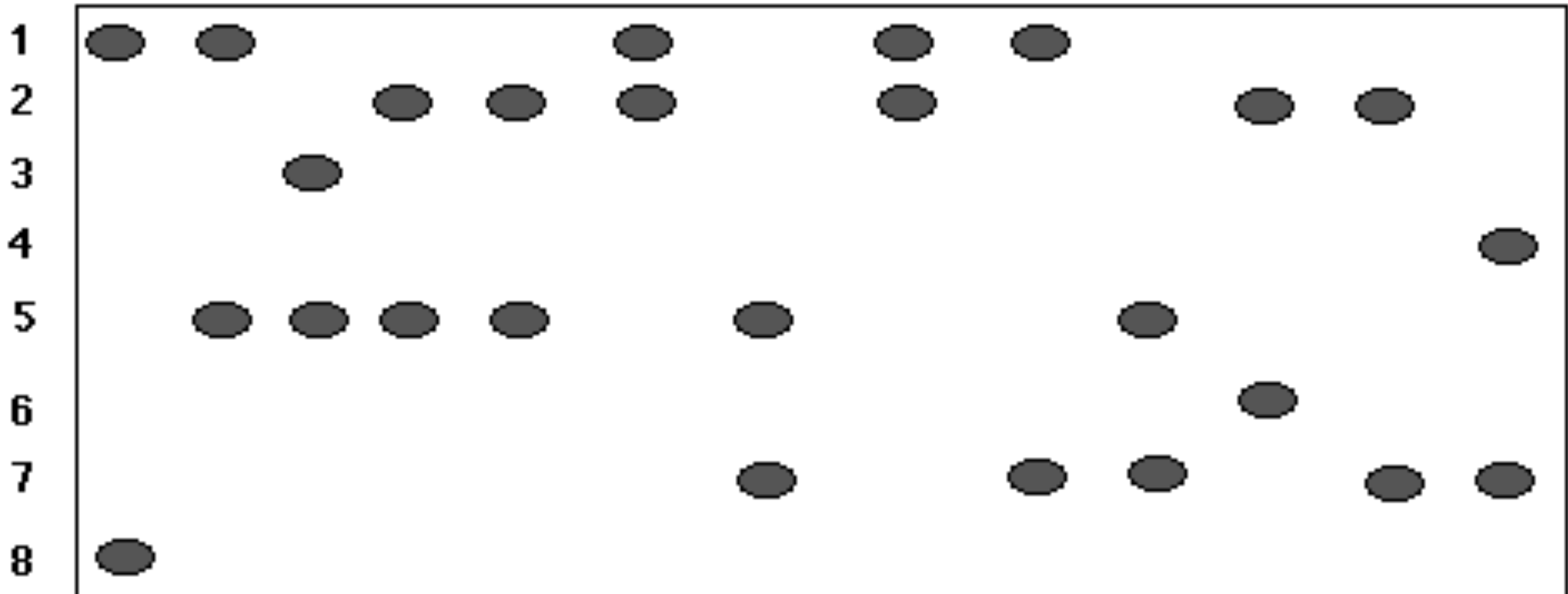


Tabela								
$\theta =$	0.01,	0.10,	0.20,	0.30,	0.35,	0.40,	0.45,	0.50
lod=	-5.0,	-2.0,	1.0,	3.3,	4.0,	3.0,	1.0,	0.0

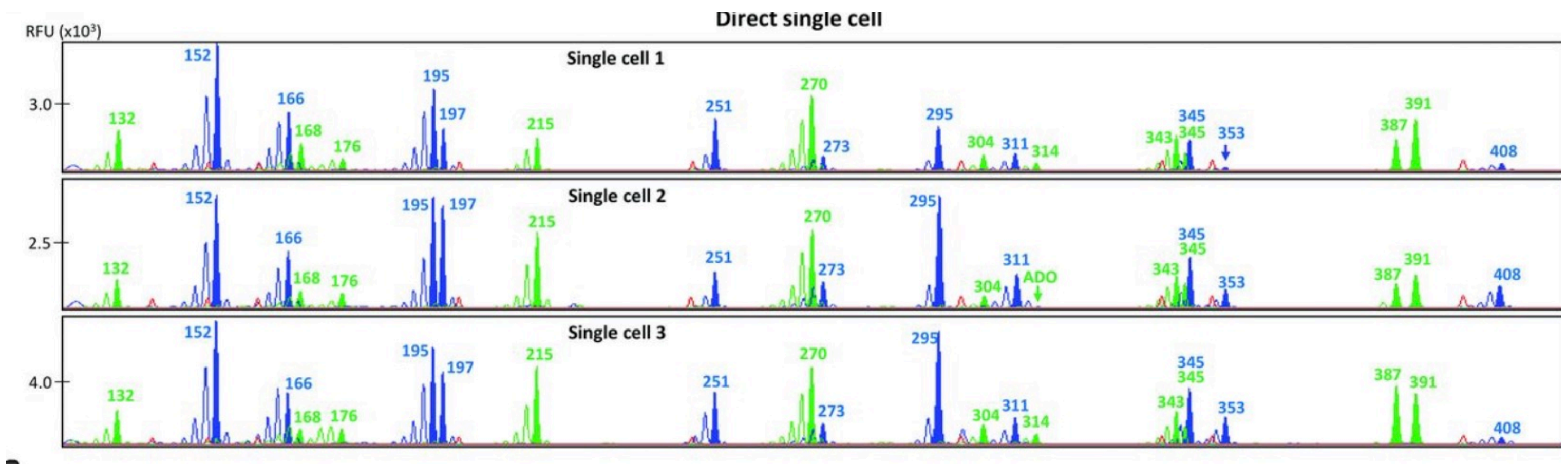
Markery w analizie sprzężeń u człowieka

- Sprzężenie dwóch genów o obserwowalnym fenotypie – praktycznie niespotykane
 - wyjątek – zespół paznokciowo-rzepkowy (NPS – *Nail Patella Syndrome*) i grupy krwi AB0
 - *Loci* w obrębie kompleksów MHC
- Markery molekularne
 - PCR, RFLP



Współczesne techniki

- Elektroforeza kapilarna i detekcja fluorescencji dla markerów PCR
- Sekwencjonowanie NGS dla markerów SNP



Problem fazy

- Większość technik genotypowania molekularnego daje wynik w postaci genotypów poszczególnych zmiennych loci, ale nie podaje haplotypu
 - układ alleli na poszczególnych chromosomach

Unphased genotypes	Possible phasing A		Possible phasing B		Possible phasing C		Possible phasing D	
A/C	A	C	A	C	A	C	A	C
G/T	G	T	G	T	T	G	T	G
A/T	A	T	T	A	A	T	T	A
Population haplotype frequency	55%	0%	15%	5%	2%	3%	0%	20%

Haplotype phasing: existing methods and new developments

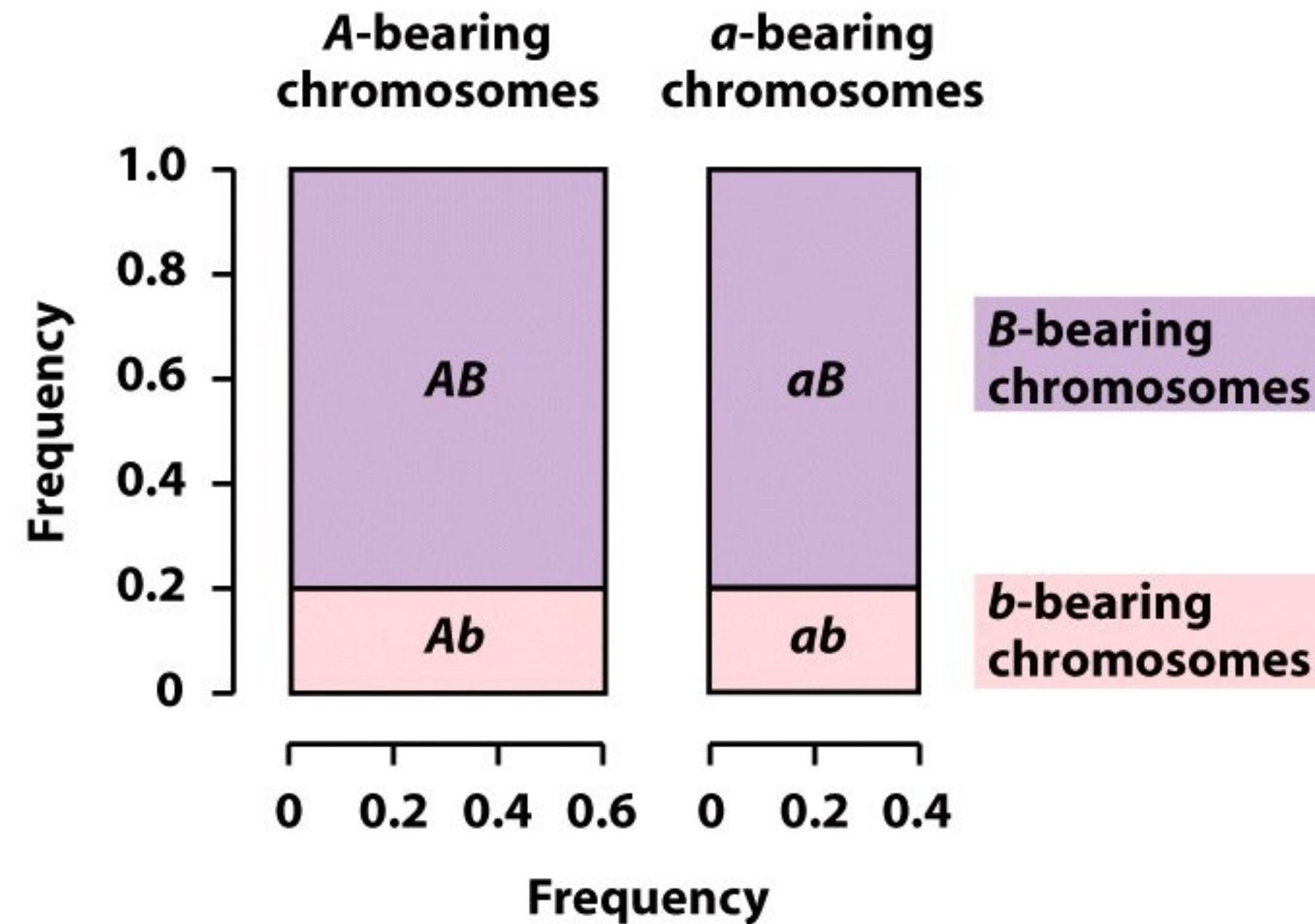
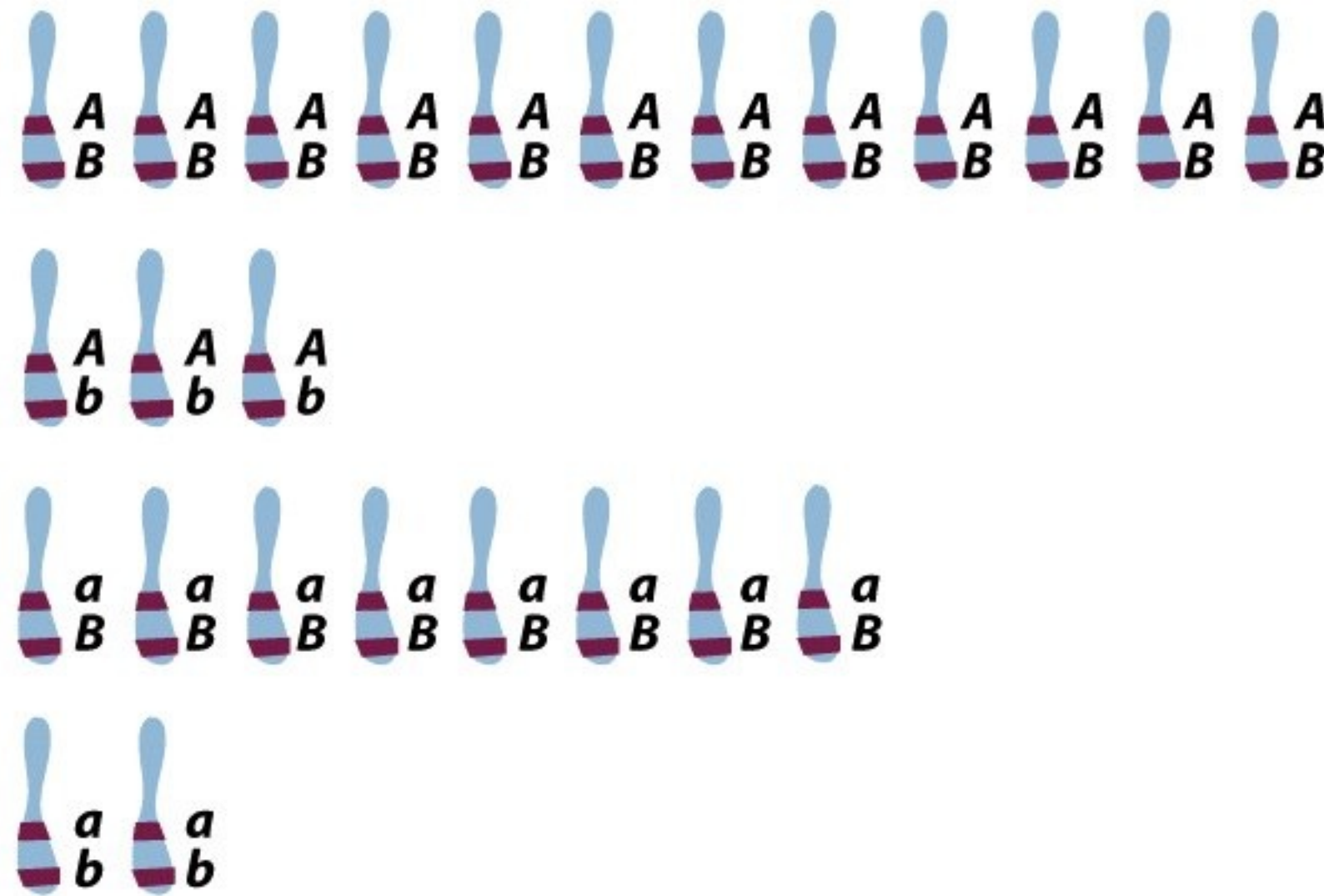
Sharon R. Browning* and Brian L. Browning†

Sprzężenia i równowaga sprzężeń

- Równowaga sprzężeń – genotyp w jednym locus jest niezależny od genotypu w drugim
- Haplotyp – genotyp (zbiór alleli) dla wielu loci danego chromosomu (lub gamety)

Równowaga sprzężeń

A population in linkage equilibrium



Frequency calculations

Allele **A**: $15 \div 25 = 0.6$

a: $10 \div 25 = 0.4$

B: $20 \div 25 = 0.8$

b: $5 \div 25 = 0.2$

Chromosome **AB**: $12 \div 25 = 0.48$

Ab: $3 \div 25 = 0.12$

aB: $8 \div 25 = 0.32$

ab: $2 \div 25 = 0.08$

Allele **B** on **A** chromosomes: $12 \div 15 = 0.8$

B on **a** chromosomes: $8 \div 10 = 0.8$

Równowaga sprzężeń

- W populacji będącej w stanie równowagi częstość haplotypu to iloczyn częstości alleli

A a B b

p q s t

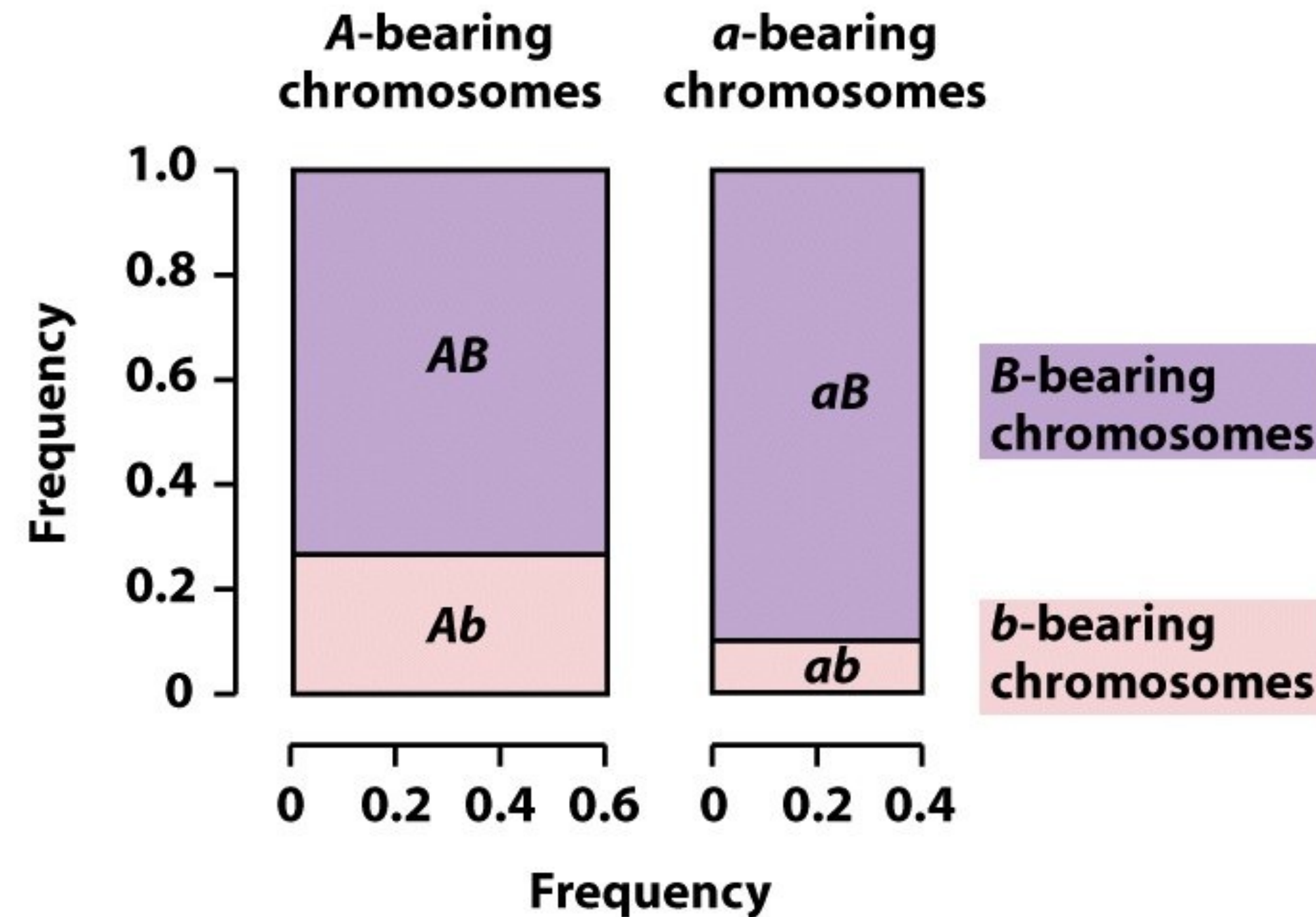
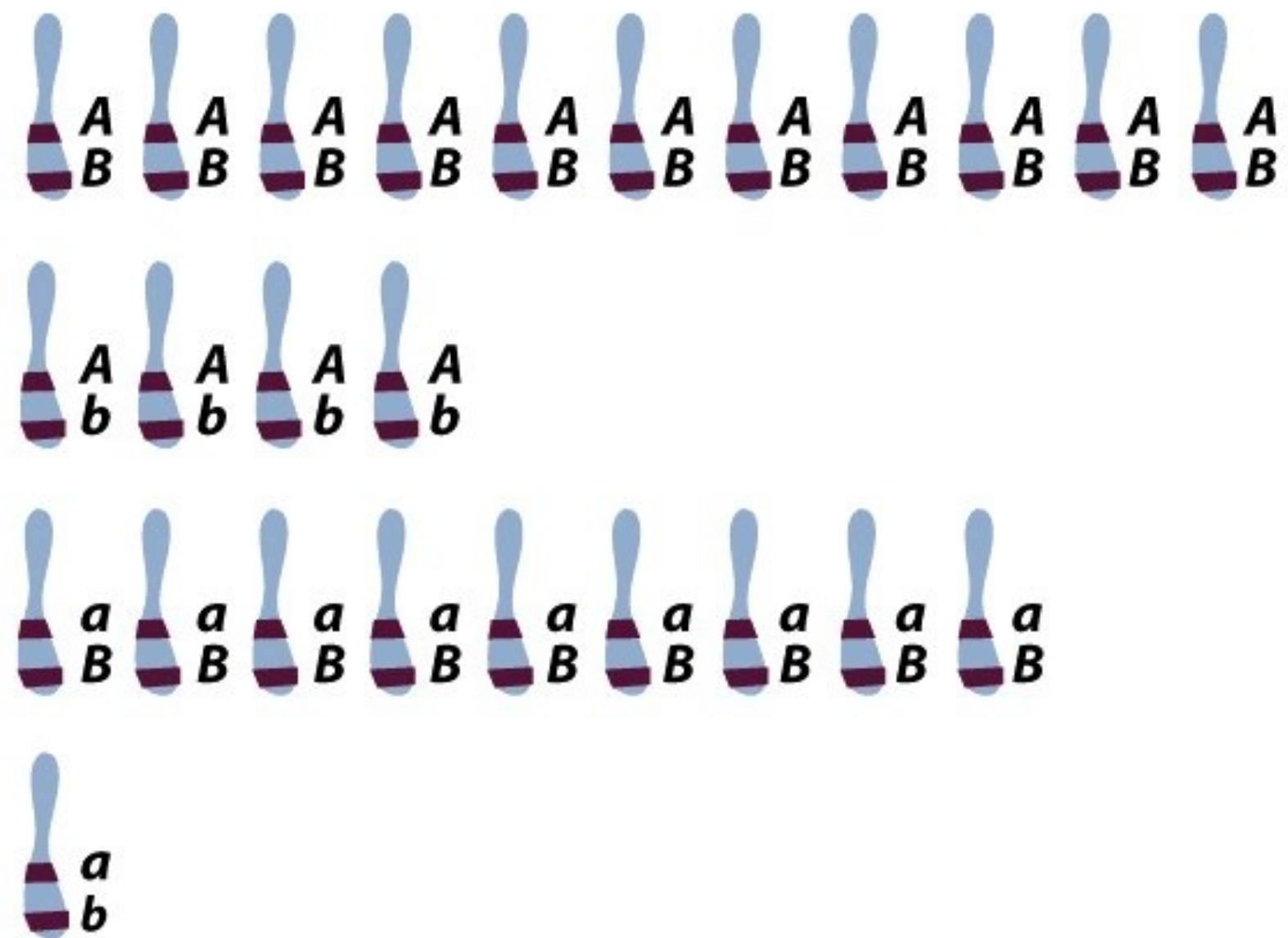
Haplotypy

AB Ab aB ab

ps pt qs qt

Nierównowaga sprzężeń

A population in linkage disequilibrium



Frequency calculations

Allele A: $15 \div 25 = 0.6$

a: $10 \div 25 = 0.4$

B: $20 \div 25 = 0.8$

b: $5 \div 25 = 0.2$

Chromosome AB: $11 \div 25 = 0.44$

Ab: $4 \div 25 = 0.16$

aB: $9 \div 25 = 0.36$

ab: $1 \div 25 = 0.04$

Allele B on A chromosomes: $11 \div 15 = 0.73$

B on a chromosomes: $9 \div 10 = 0.9$

Skąd bierze się nierównowaga sprzężeń

- Migracje
- Dobór naturalny działający na genotyp wielu loci
- Kombinacja doboru w jednym z loci i dryfu

Nierównowaga sprzężeń a sprzężenie

Allel genu d w pobliżu markera A zmutował dając związany z chorobą allel D wiele pokoleń temu - efekt założyciela



Jeżeli odległość A do d jest niewielka, wówczas większość chromosomów w populacji niosących D niesie też A1

Ale nie odwrotnie!

Nierównowaga sprzężeń (*linkage disequilibrium*) – nieprzypadkowe skojarzenie alleli w sprzężonych loci – efekt założyciela. Maleje z czasem.

Lokalizowanie genu z mutacją sprawczą

- I etap – zgrubne (markery co 8-20 cM) – ustalenie chromosomu, stwierdzenie czy we wszystkich rodzinach ten sam *locus* itp.
- II etap – dokładne (markery co 1-4 cM)

Sprzężenia w epoce genomu

- Techniki sekwencjonowania całogenomowego są obecnie coraz bardziej dostępne
- Czy badanie sprzężeń ma jeszcze sens?

Table 1 | Summary of 1000 Genomes Project phase I data

	Autosomes	Chromosome X	GENCODE regions*
Samples	1,092	1,092	1,092
Total raw bases (Gb)	19,049	804	327
Mean mapped depth (×)	5.1	3.9	80.3
SNPs			
No. sites overall	36.7 M	1.3 M	498 K
Novelty rate†	58%	77%	50%
No. synonymous/non-synonymous/nonsense	NA	4.7/6.5/0.097 K	199/293/6.3 K
Average no. SNPs per sample	3.60 M	105 K	24.0 K
Indels			
No. sites overall	1.38 M	59 K	1,867
Novelty rate†	62%	73%	54%
No. inframe/frameshift	NA	19/14	719/1,066
Average no. indels per sample	344 K	13 K	440
Genotyped large deletions			
No. sites overall	13.8 K	432	847
Novelty rate†	54%	54%	50%
Average no. variants per sample	717	26	39

NA, not applicable.
* Autosomal genes only.
† Compared with dbSNP release 135 (Oct 2011), excluding contribution from phase I 1000 Genomes Project (or equivalent data for large deletions).

*Lists of participants and their affiliations appear at the end of the paper.

An integrated map of genetic variation from 1,092 human genomes

Table 1 | Median autosomal variant sites per genome

	AFR		AMR		EAS		EUR		SAS	
Samples	661		347		504		503		489	
Mean coverage	8.2		7.6		7.7		7.4		8.0	
	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons
SNPs	4.31M	14.5k	3.64M	12.0k	3.55M	14.8k	3.53M	11.4k	3.60M	14.4k
Indels	625k	-	557k	-	546k	-	546k	-	556k	-
Large deletions	1.1k	5	949	5	940	7	939	5	947	5
CNVs	170	1	153	1	158	1	157	1	165	1
MEI (Alu)	1.03k	0	845	0	899	1	919	0	889	0
MEI (L1)	138	0	118	0	130	0	123	0	123	0
MEI (SVA)	52	0	44	0	56	0	53	0	44	0
MEI (MT)	5	0	5	0	4	0	4	0	4	0
Inversions	12	0	9	0	10	0	9	0	11	0
Nonsynon	12.2k	139	10.4k	121	10.2k	144	10.2k	116	10.3k	144
Synon	13.8k	78	11.4k	67	11.2k	79	11.2k	59	11.4k	78
Intron	2.06M	7.33k	1.72M	6.12k	1.68M	7.39k	1.68M	5.68k	1.72M	7.20k
UTR	37.2k	168	30.8k	136	30.0k	169	30.0k	129	30.7k	168
Promoter	102k	430	84.3k	332	81.6k	425	82.2k	336	84.0k	430
Insulator	70.9k	248	59.0k	199	57.7k	252	57.7k	189	59.1k	243
Enhancer	354k	1.32k	295k	1.05k	289k	1.34k	288k	1.02k	295k	1.31k
TFBSs	927	4	759	3	748	4	749	3	765	3
Filtered LoF	182	4	152	3	153	4	149	3	151	3
HGMD-DM	20	0	18	0	16	1	18	2	16	0
GWAS	2.00k	0	2.07k	0	1.99k	0	2.08k	0	2.06k	0
ClinVar	28	0	30	1	24	0	29	1	27	1

What is ClinVar?

ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. ClinVar thus facilitates access to and communication

D862–D868 *Nucleic Acids Research*, 2016, Vol. 44, Database issue
doi: 10.1093/nar/gkv1222

Published online 17 November 2015

ClinVar: public archive of interpretations of clinically relevant variants

Melissa J. Landrum^{*}, Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, Wonhee Jang, Kenneth Katz, Michael Ovetsky, George Riley, Amanjeev Sethi, Ray Tully, Ricardo Villamarin-Salomon, Wendy Rubinstein and Donna R. Maglott

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20893, USA

Received September 14, 2015; Revised October 15, 2015; Accepted October 24, 2015

Sprzężenia w epoce genomu

- Między dwiema osobami możemy spodziewać się nawet 4 milionów różnic na poziomie sekwencji
 - U bliskich krewnych mniej, ale nadal sporo
- Stwierdzenie, które z tych różnic odpowiada za fenotyp nie zawsze jest ewidentne
 - Łatwiej w obszarach kodujących
- Sekwencjonowanie stosowane w rzadkich chorobach, gdzie nie ma dostatecznie dużo rodowodów

Wiele genów
niskie penetracje

Podłoże
genetyczne

Pojedyncze geny
wysoka penetracja

Analiza asocjacji
Wielogenowe współczynniki ryzyka

Dane

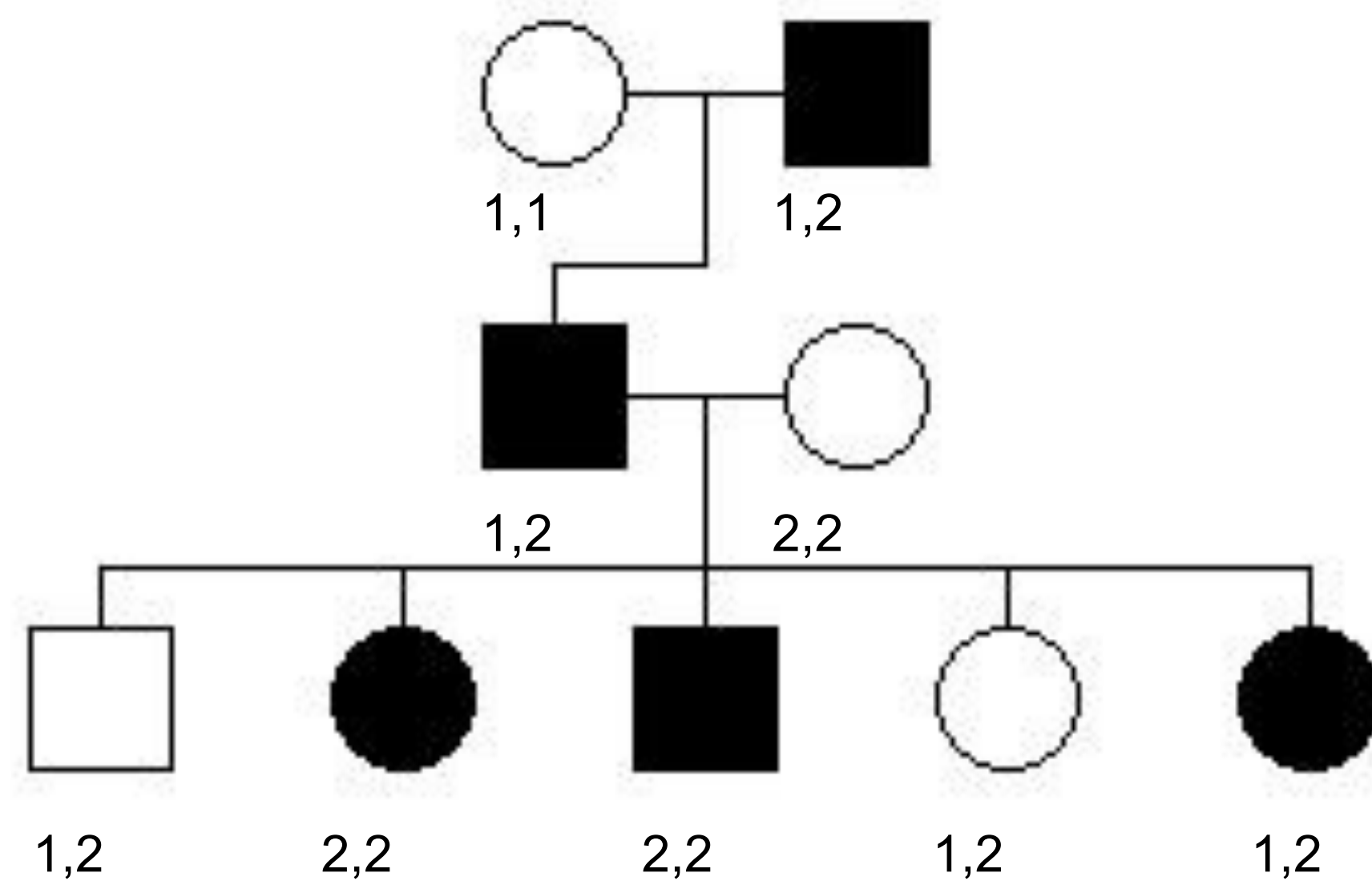
Dużo małych rodowodów

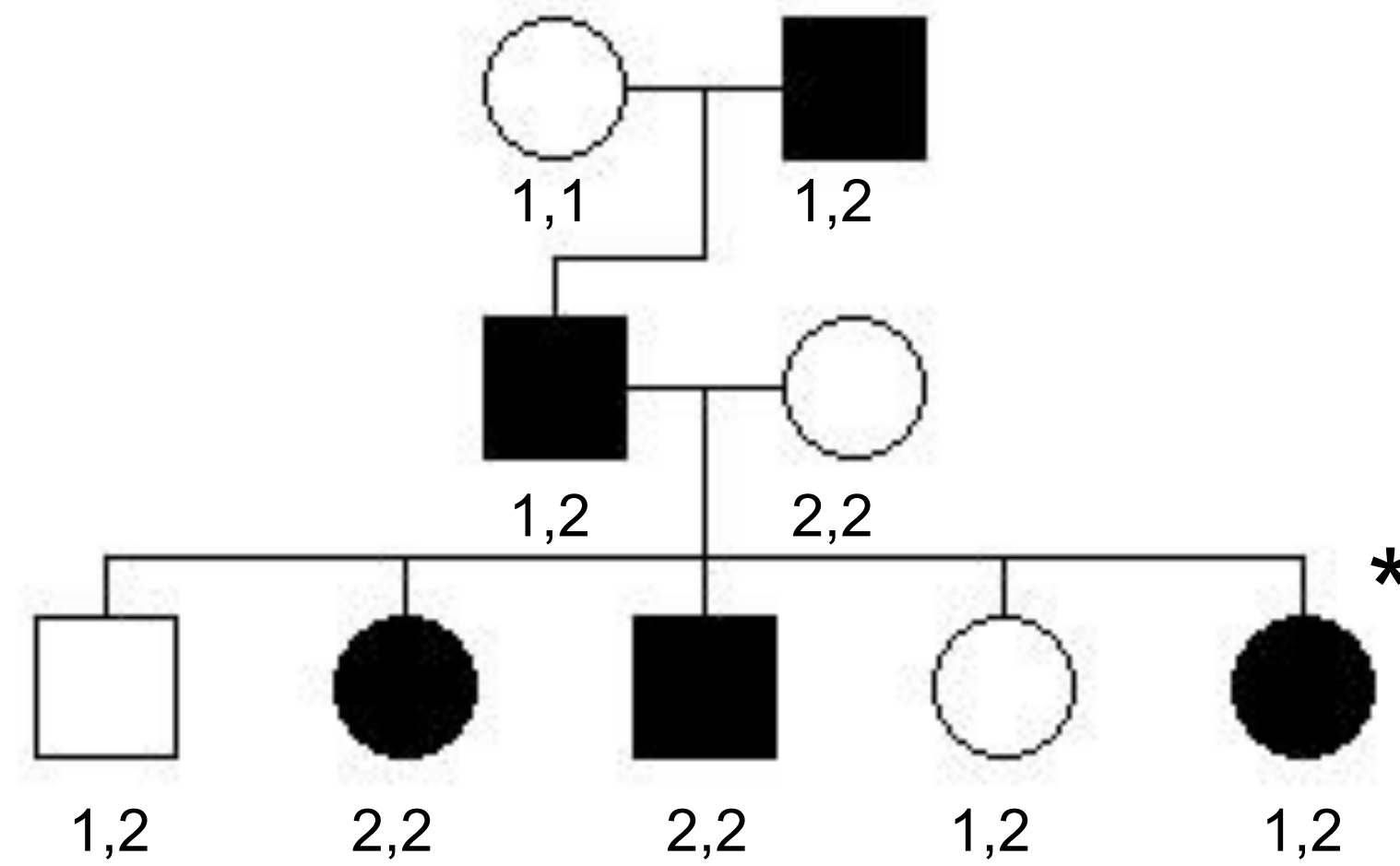
Duże rodowody

Nieparametryczna
analiza sprzężeń

Parametryczna
analiza sprzężeń

Ćwiczenie

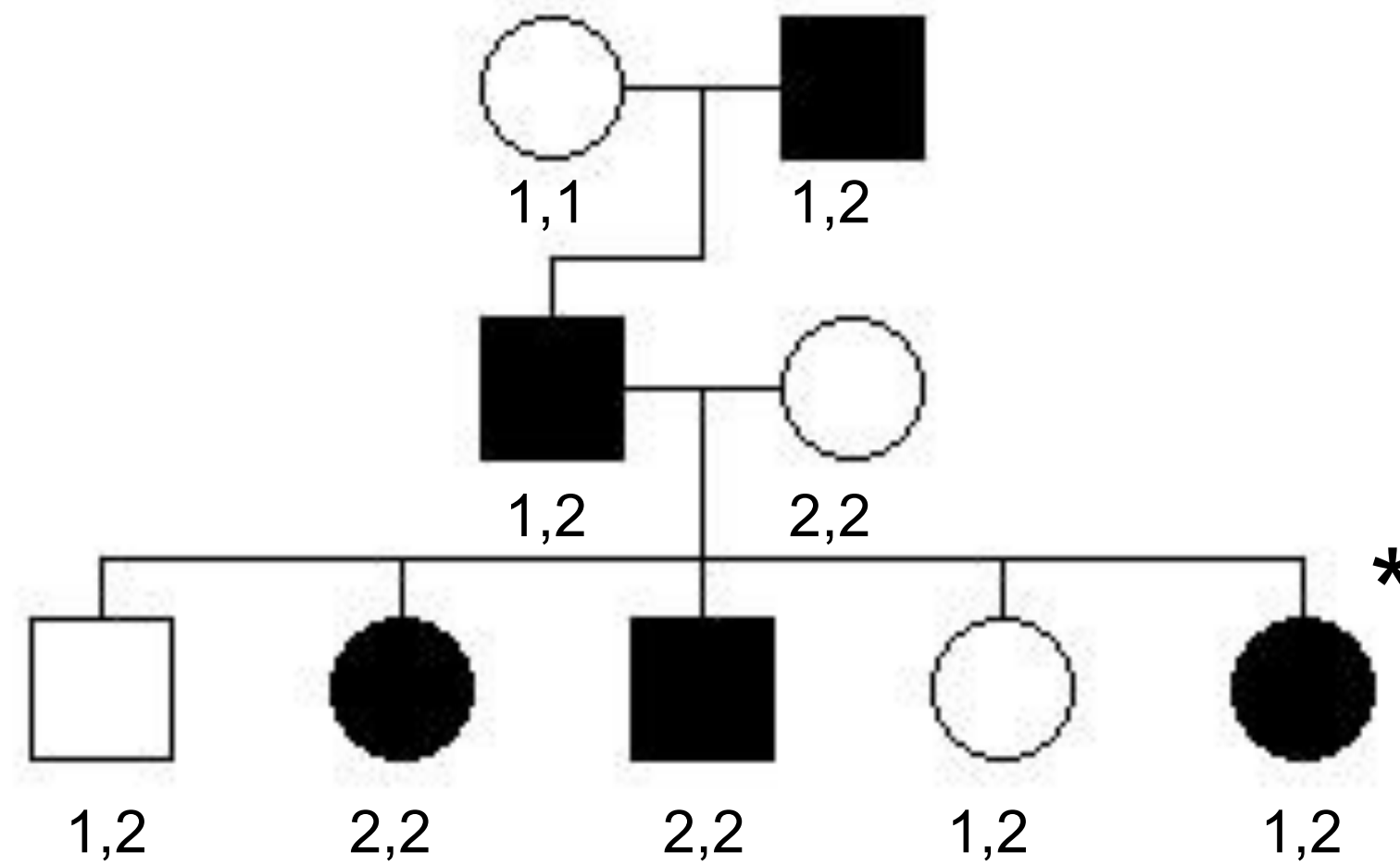




1 rekombinant (R); 4 rodzicielskie (NR)

Przy braku sprzężenia ($\theta=0,5$) prawdopodobieństwo uzyskania R i NR jest jednakowe i wynosi $\frac{1}{2}$

$$L(\theta=0,5) = \left(\frac{1}{2}\right)^5$$

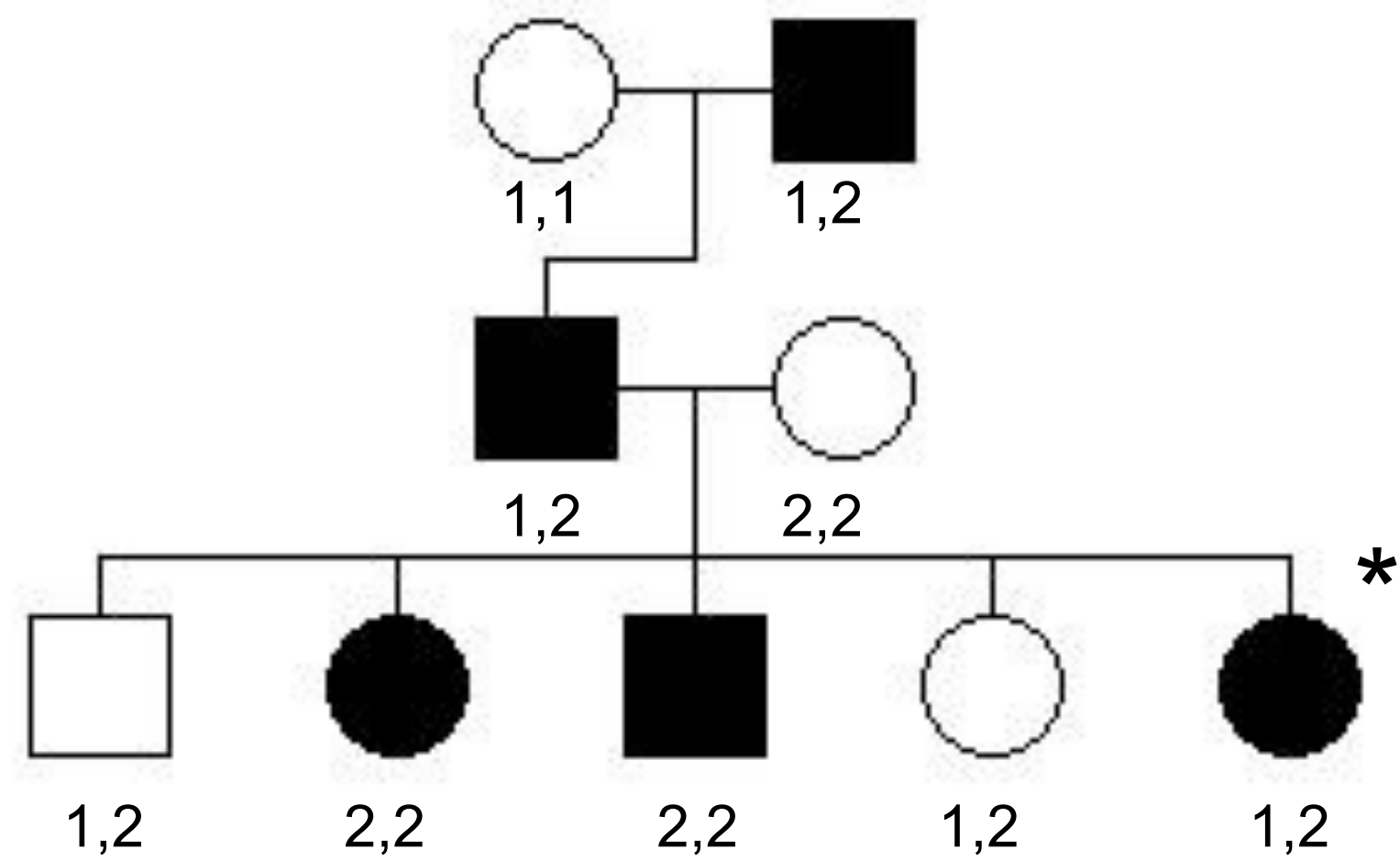


1 rekombinant (R); 4 rodzicielskie (NR)

Dla danej wartości θ prawdopodobieństwo uzyskania R wynosi θ (z definicji), prawdopodobieństwo uzyskania NR wynosi zatem $1 - \theta$

$$L(\theta) = \theta \cdot (1 - \theta)^4$$

1R 4NR



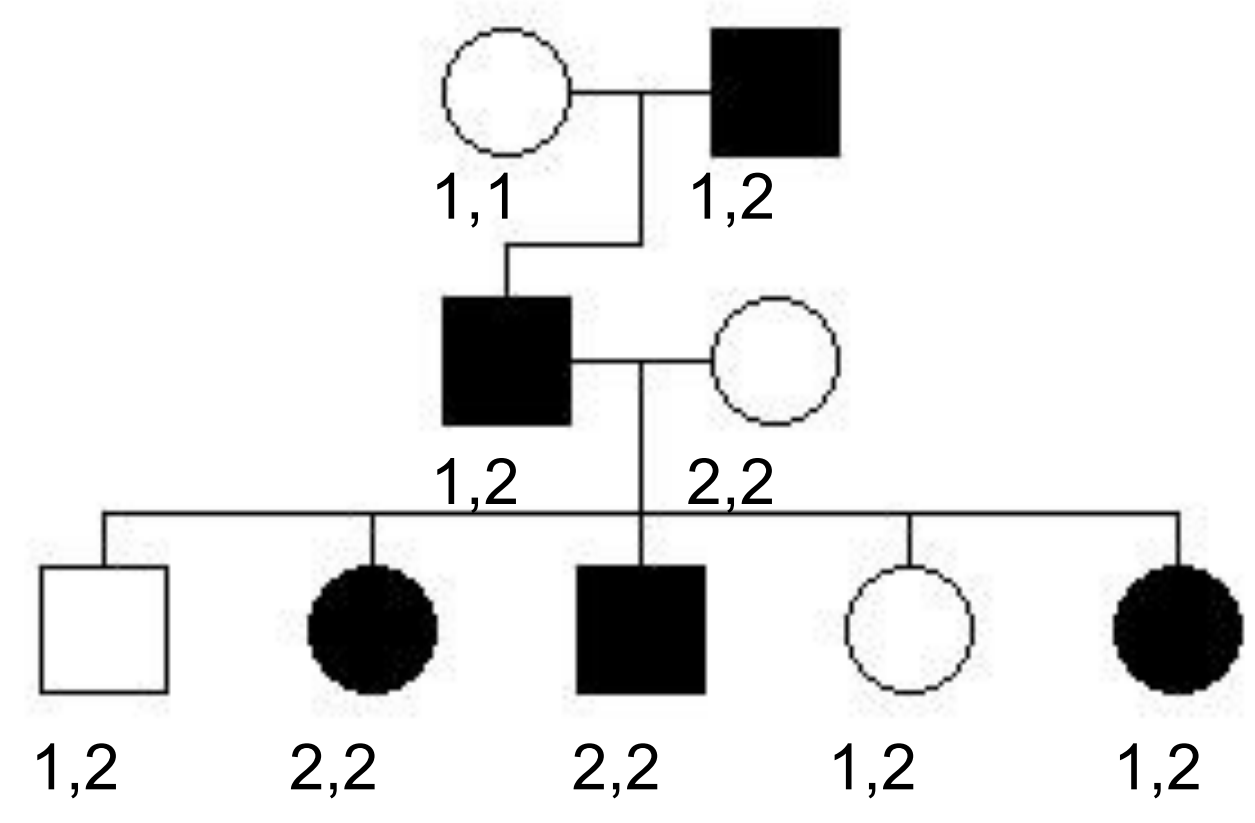
1 rekombinant (R); 4 rodzicielskie (NR)

$$L(\theta=0,5) = (1/2)^5$$

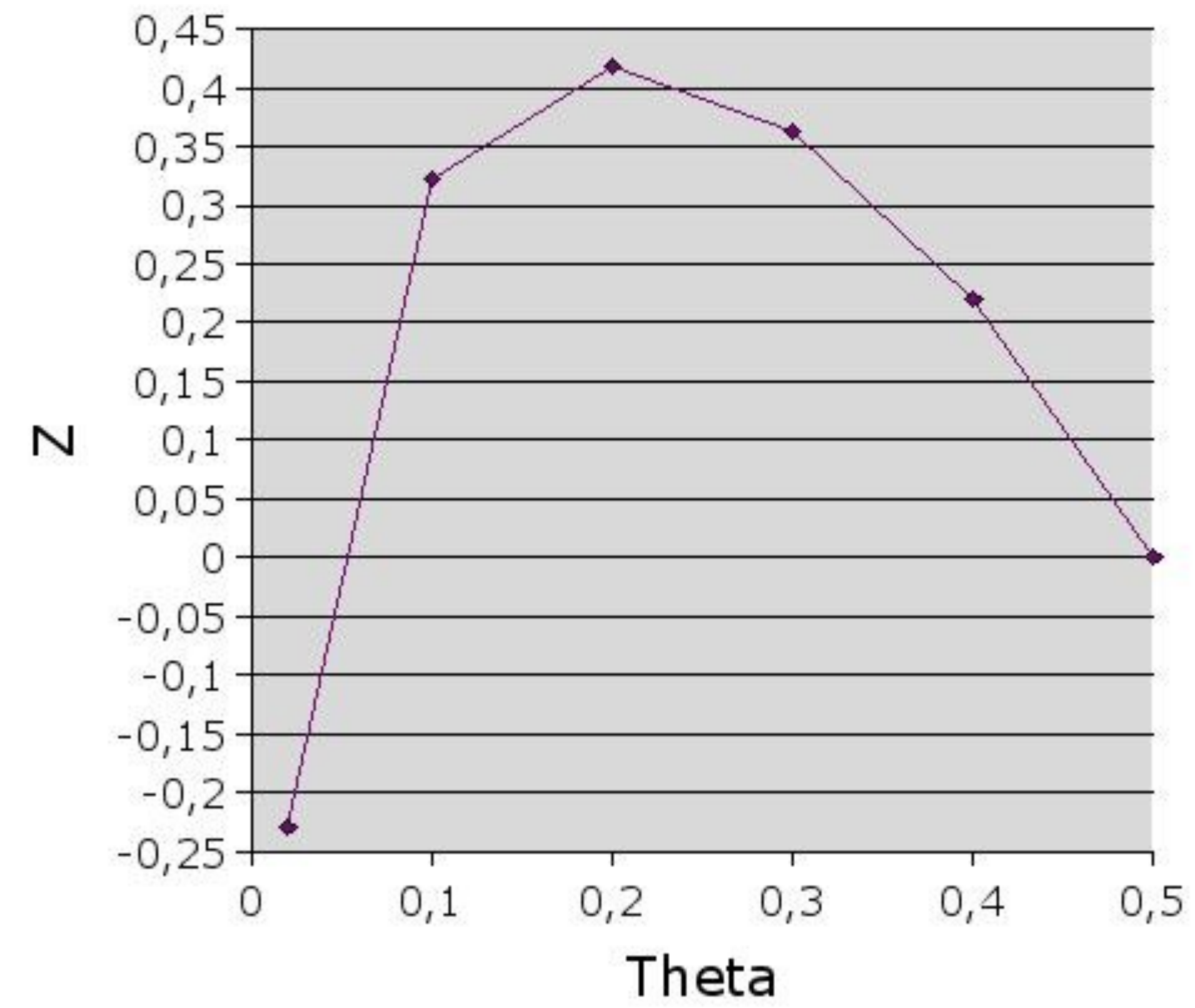
$$L(\theta) = \theta \cdot (1 - \theta)^4$$

$$\text{Dla } \theta=0,1 \quad L(\theta=0,1) = 0,1 \cdot (0,9)^4$$

$$Z(\theta = 0,1) = \log_{10} \left(\frac{0,1 \cdot 0,9^4}{0,5^5} \right) \approx 0,32$$

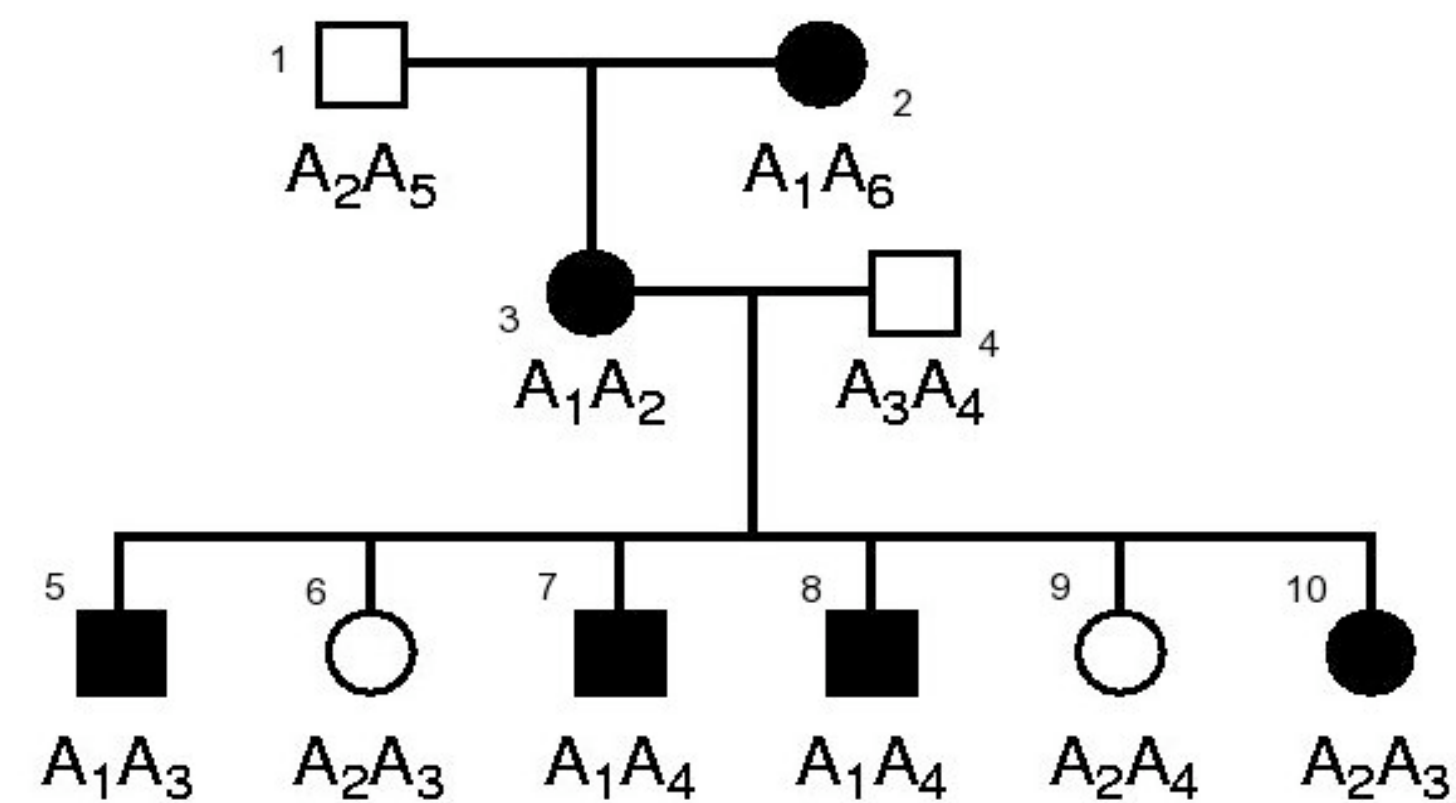


0	0,02	0,1	0,2	0,3	0,4	0,5
$-\infty$	-0,23	0,32	0,42	0,36	0,22	0



0	0,02	0,1	0,2	0,3	0,4	0,5
$-\infty$	-0,23	0,32	0,42	0,36	0,22	0

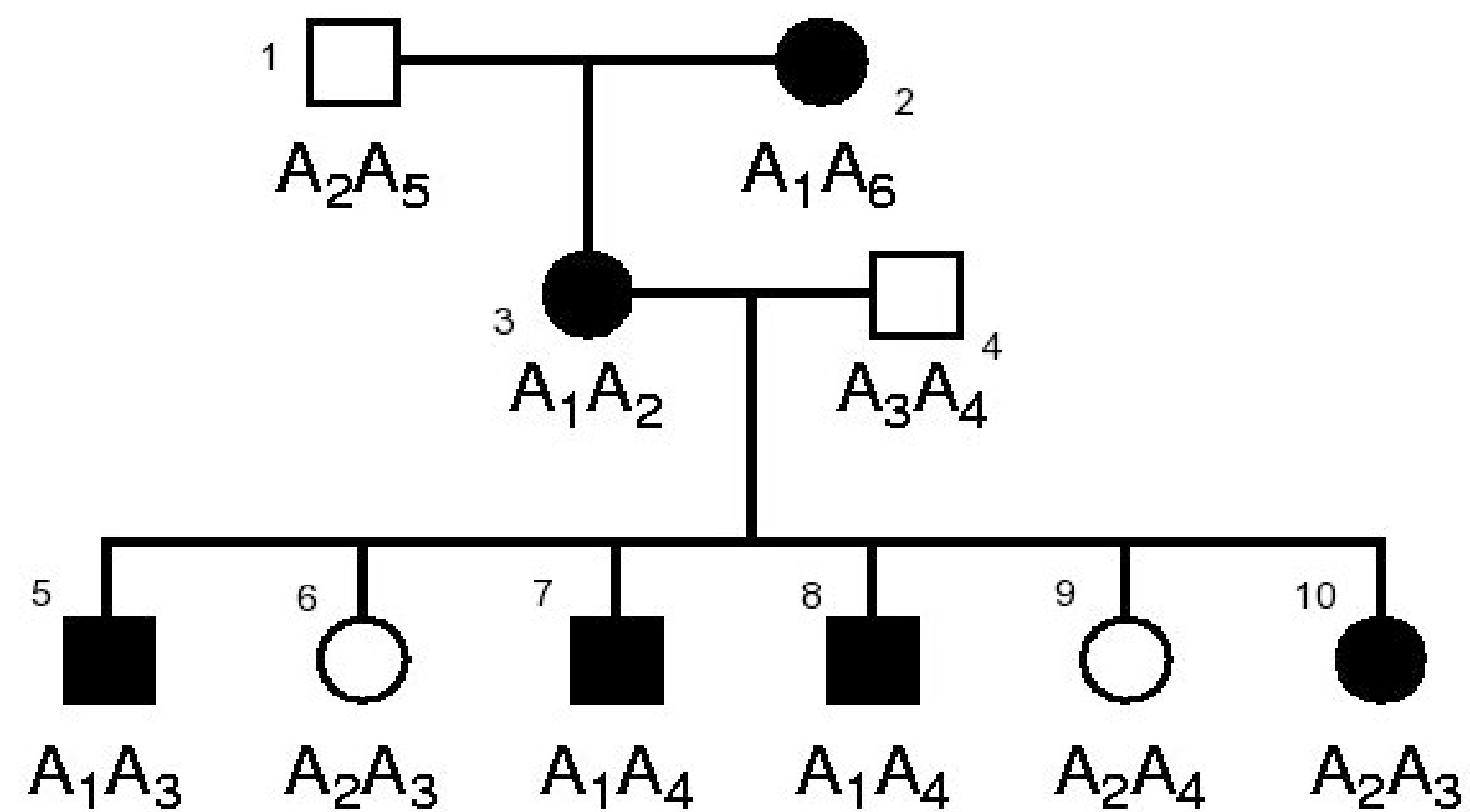
Kodowanie rodowodu



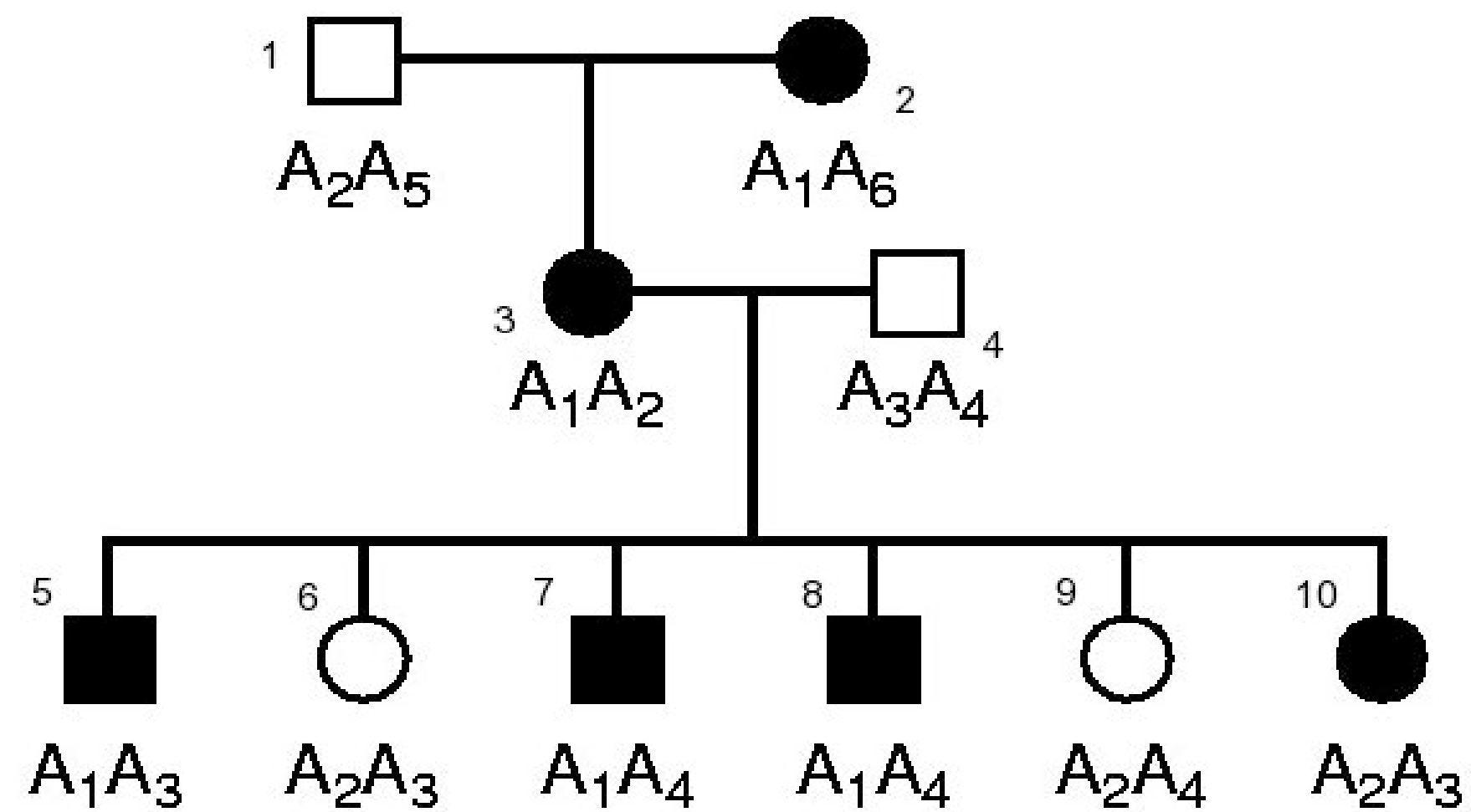
Płeć: 1 mężczyzna, 2 kobieta, 0 nieznana
Choroba: 1 zdrowa(y), 2 chora(y), 0 nieznana
0 zawsze oznacza nieznane/brak danych!!!

Rodzina	osoba	ojciec	matka	płeć	choroba	marker1a1	marker1a2
001	1	0	0	1	1	2	5
001	2	0	0	2	2	1	6
001	3	1	2	2	2	1	2
001	4	0	0	1	1	3	4
001	5	4	3	1	2	1	3

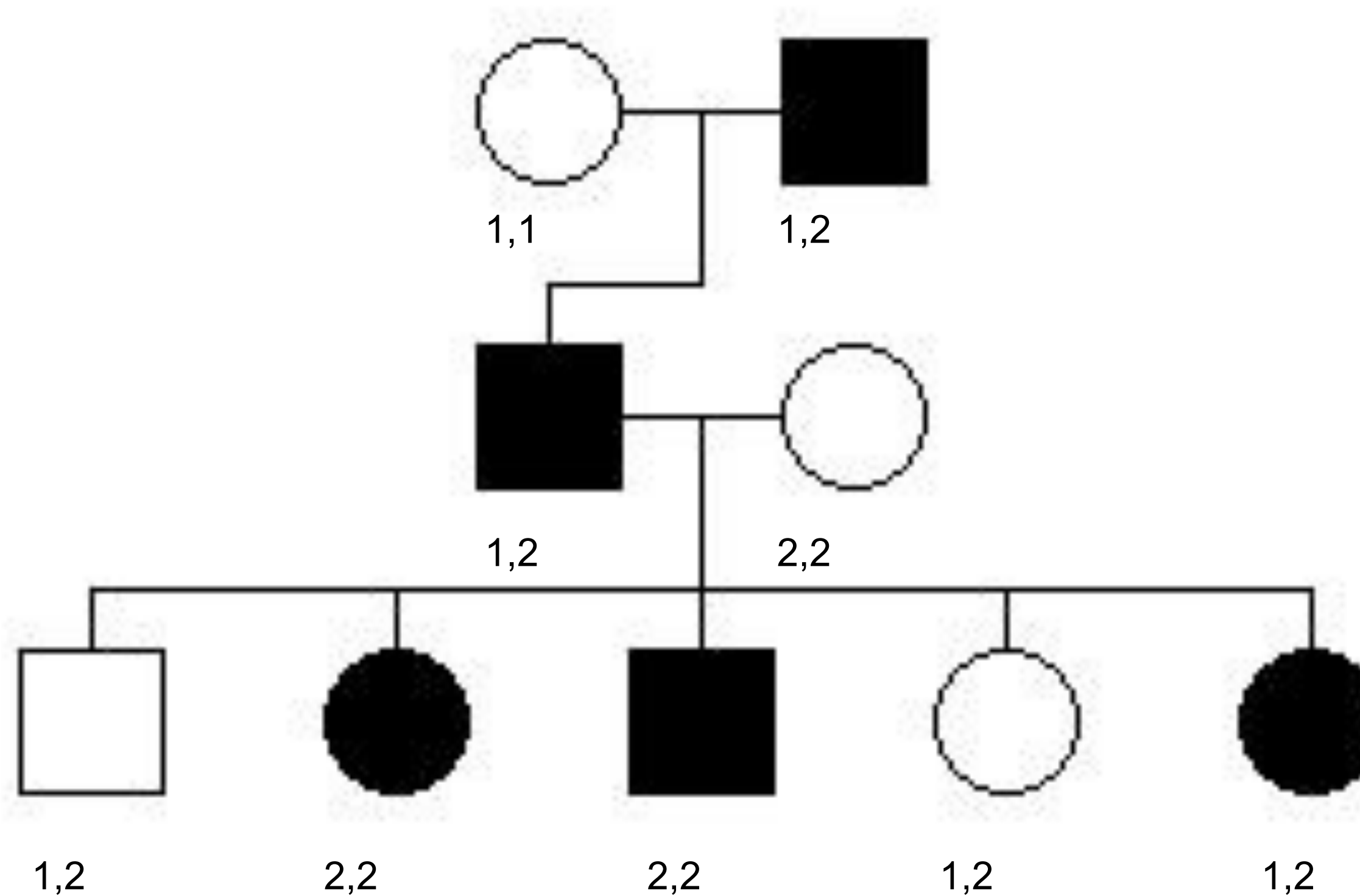
Liczba spacji nieistotna (minimum 1)



001	1	0	0	1	1	2	5
001	2	0	0	2	2	1	6
001	3	1	2	2	2	1	2
001	4	0	0	1	1	3	4
001	5	4	3	1	2	1	3
001	6	4	3	2	1	2	3
001	7	4	3	1	2	1	4
001	8	4	3	1	2	1	4
001	9	4	3	2	1	2	4
001	10	4	3	2	2	2	3



Rodz1	dziadek	0	0			1	1	2	5
Rodz1	babka	0	0			2	2	1	6
Rodz1	matka			dziadek	babka	2	2	1	2
Rodz1	ojciec	0	0			1	1	3	4
Rodz1	dz1			ojciec	matka	1	2	1	3
Rodz1	dz2			ojciec	matka	2	1	2	3
Rodz1	dz3			ojciec	matka	1	2	1	4
Rodz1	dz4			ojciec	matka	1	2	1	4
Rodz1	dz5			ojciec	matka	2	1	2	4
Rodz1	dz6			ojciec	matka	2	2	2	3



Rodzina osoba ojciec matka płeć choroba marker1a1 marker1a2

Płeć: 1 mężczyzna, 2 kobieta, 0 nieznana

Choroba: 1 zdrowa(y), 2 chora(y), 0 nieznana

0 zawsze oznacza nieznane/brak danych!!!