Poszukiwanie genetycznych podstaw fenotypu

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

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

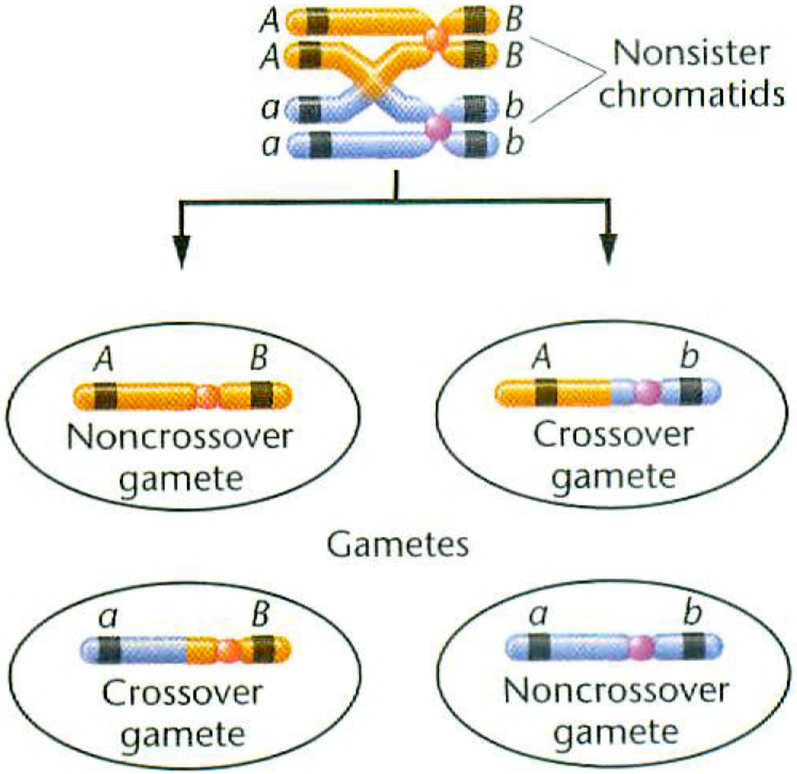
- Osoby z allelem a jakiegoś genu mają większe prawdopodobieństwo fenotypu X
    - Czy zawsze asocjacja oznacza zależność przyczynową?
    - Czy każda asocjacja ma wartość diagnostyczną?
    - Czy asocjacja odkrywa “gen na ....”?
- Odpowiedzi na te i inne pytania - wykład

Asocjacja a sprzężenie

Asocjacja	Sprzężenie
Dotyczy alleli	Dotyczy genów (*loci*)
Na poziomie populacji	W rodzinie
Dziedziczenie wieloczynnikowe	Dziedziczenie mendlowskie - mutacje sprawcze

Sprzężenie

Crossing-over (rekombinacja chromatyd niesiostrzanych w mejozie)

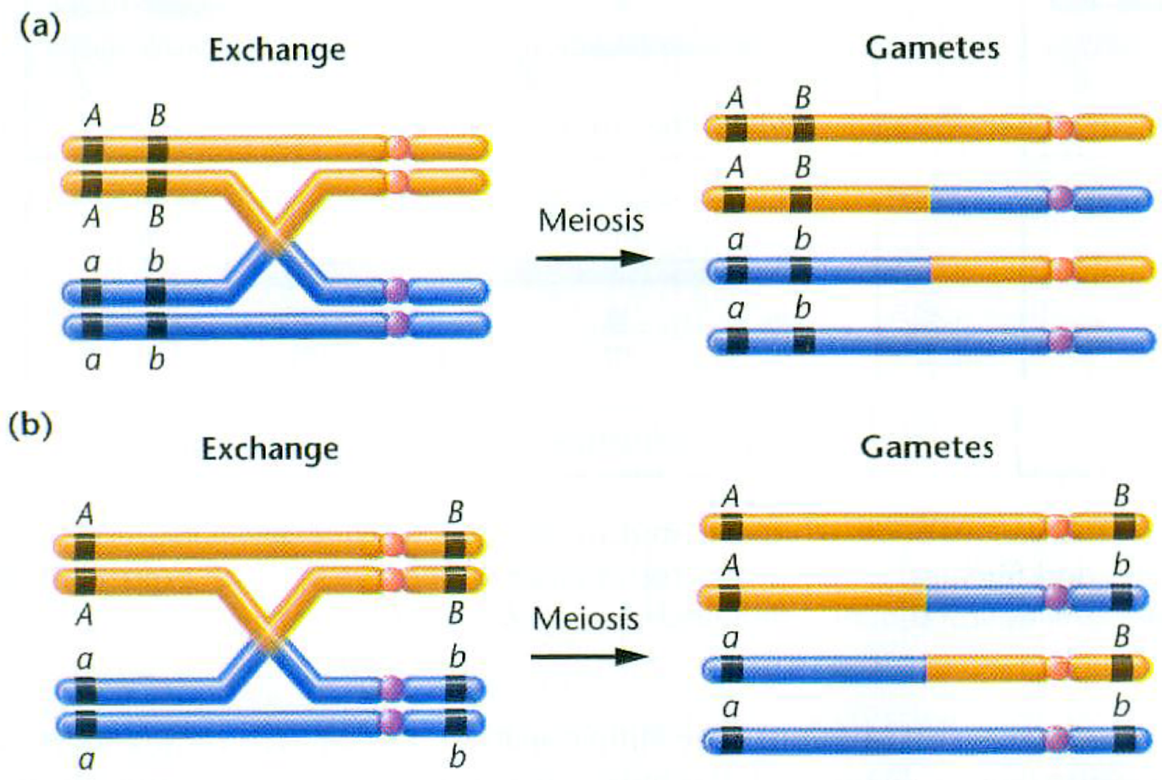


Dla 2 genów:
2 równoliczne klasy gamet rodzicielskich
2 równoliczne klasy gamet rekombinowanych
Klasy rekombinowane mniej liczne od rodzicielskich

W. S Klug, M.R Cummings “Concepts of Genetics” 8th edition, Prentice Hall, 2005

Mapowanie genetyczne

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



powstają gamety rodzicielskie

powstają gamety rekombinowane

Liczba rekombinantów jest miarą odległości genetycznej

W. S Klug, M.R Cummings “Concepts of Genetics” 8th edition, Prentice Hall, 2005

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 nieparametyczne - 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, SNP.)
  - 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

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

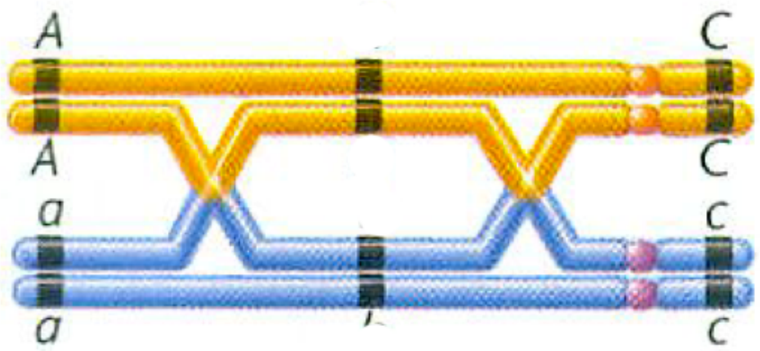
- Zakładamy, że jest znany model dziedziczenia
  - 2 lub więcej genów i/lub markerów - *loci*
  - genotypy rodziców i genotypy potomstwa
  - dane (liczba rekombinantów) pozwalają na wyznaczenie parametrów modelu - odległości między *loci*

Miarą odległości jest częstość rekombinacji

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

Mapowanie

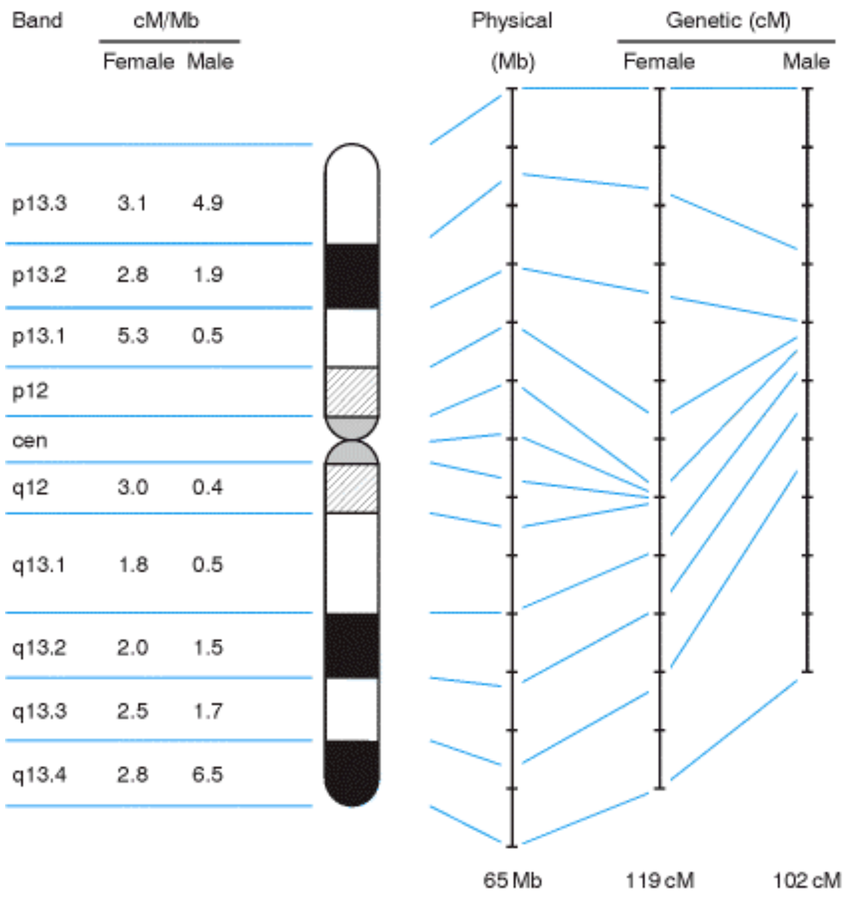
- Jednostka: cM (centymorgan) = 1% rekombinacji
- Zależność nie jest liniowa



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

Płeć a częstość rekombinacji

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



Wykorzystanie analizy sprzężeń

- Uwzględnienie wielokrotnych crossing-over i interferencji: funkcja mapowa
- Obecnie analizę sprzężeń wykorzystuje się jeszcze tylko do ustalania, w pobliżu jakiego znanego markera/polimorfizmu należy szukać genu powiązanego z chorobą
- Nie musimy więc przejmować się dokładną zależnością odległości i częstości obserwowanych rekombinantów

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 (θ=0,5)
- Stosunek wiarygodności dla danej wartości θ: L(θ)/L(θ=0,5)
- lod score (Z) = *logarithm of odds* – logarytm dziesiętny stosunku wiarygodności

Proste przykłady obliczeń *lod*

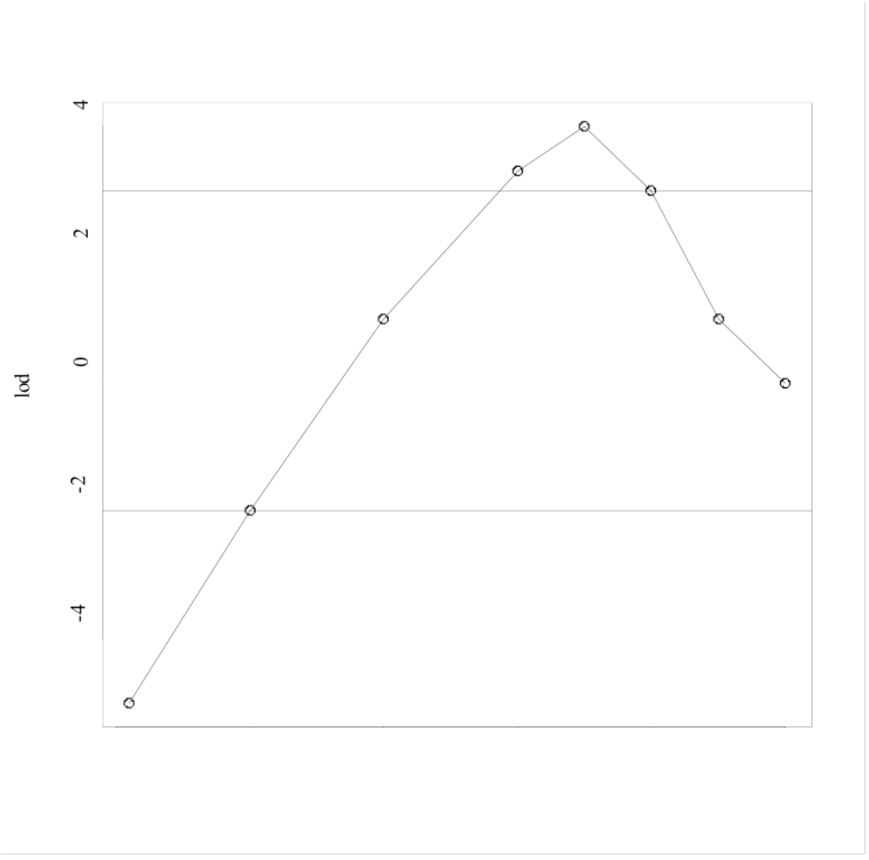
Dla danego rodowodu (i), *lod* dla danego θ wynosi:

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

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

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

Analiza dwupunktowa



znaczące
Z>3, (Z>2 dla sprzężonych z płcią)

wykluczone

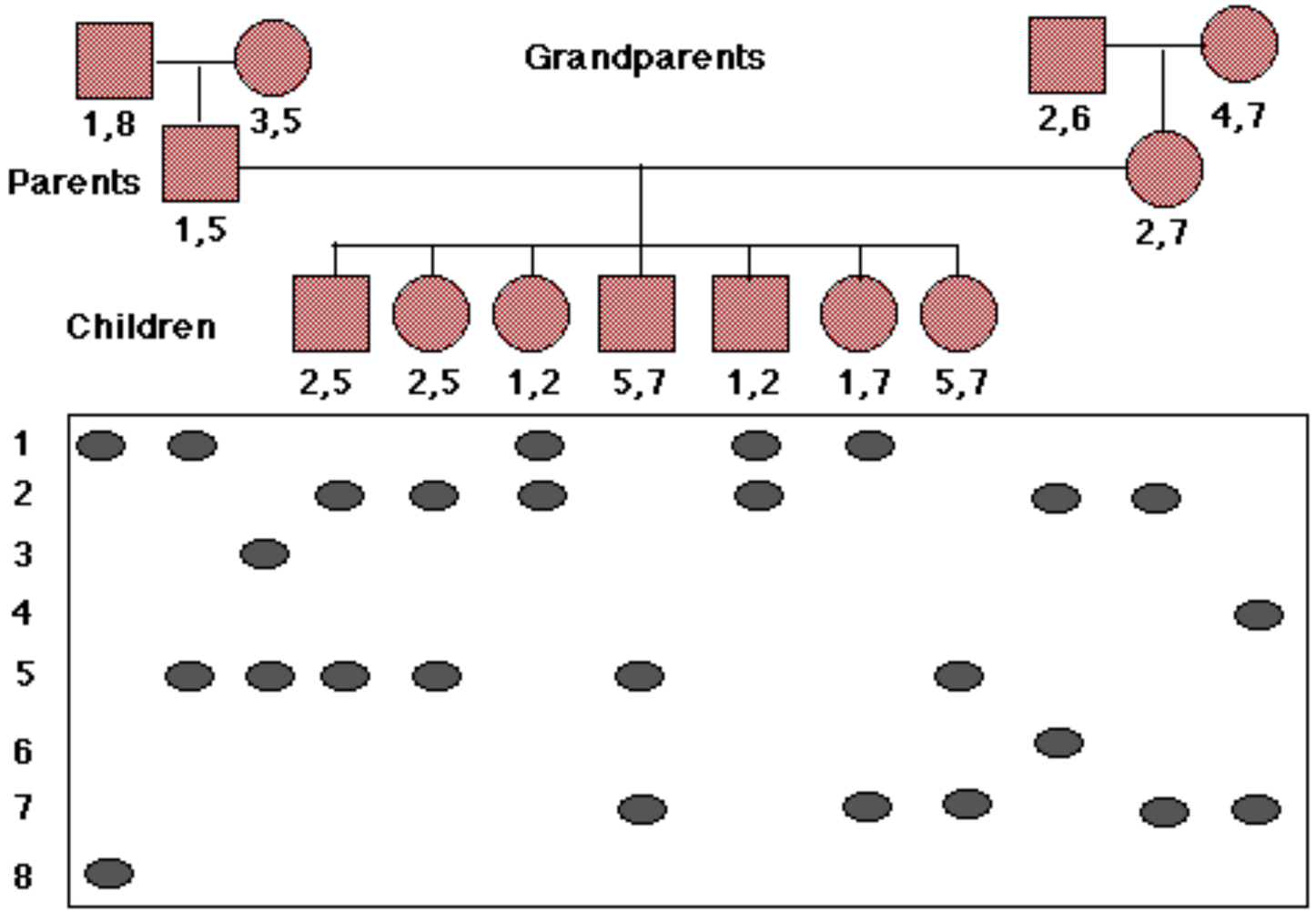
Tabela

θ =	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

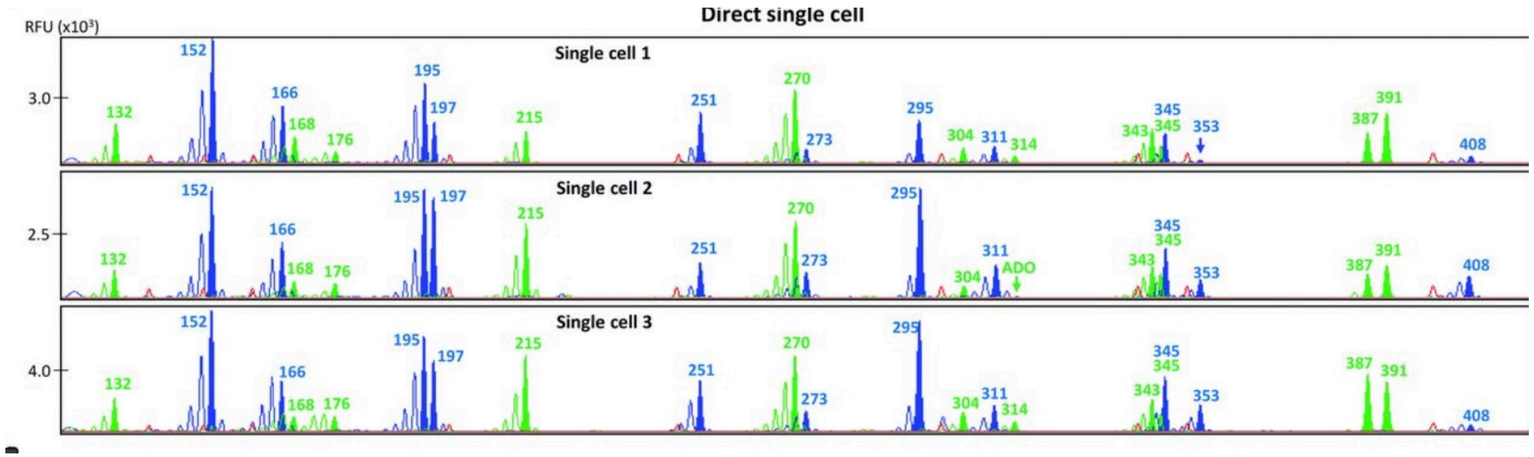
- Sprzężenie dwóch genów o obserwowalnym fenotypie – praktycznie niespotykane
  - wyjątek – zespół paznokciowo-rzepkowy (NPS – *Nail Patella Syndrome*) i grupy krwi AB0
  - *Loci* w obrębie kompleksów MHC
- Markery molekularne
  - obecnie głównie SNP

Markery


Grandparents
1,8
3,5
2,6
4,7
Parents
1,5
2,7
Children
2,5
2,5
1,2
5,7
1,2
1,7
5,7
1
2
3
4
5
6
7
8

Współczesne techniki

- Elektroforeza kapilarna i detekcja fluorescencji dla markerów PCR
- Mikromacierze i 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‡*

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

The 1000 Genomes Project Consortium*

Projekt 1000 genomów

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

NATURE | VOL 526 | 1 OCTOBER 2015

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

D862–D868 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 17 November 2015*
doi: 10.1093/nar/gkv1222

ClinVar: public archive of interpretations of clinically relevant variants

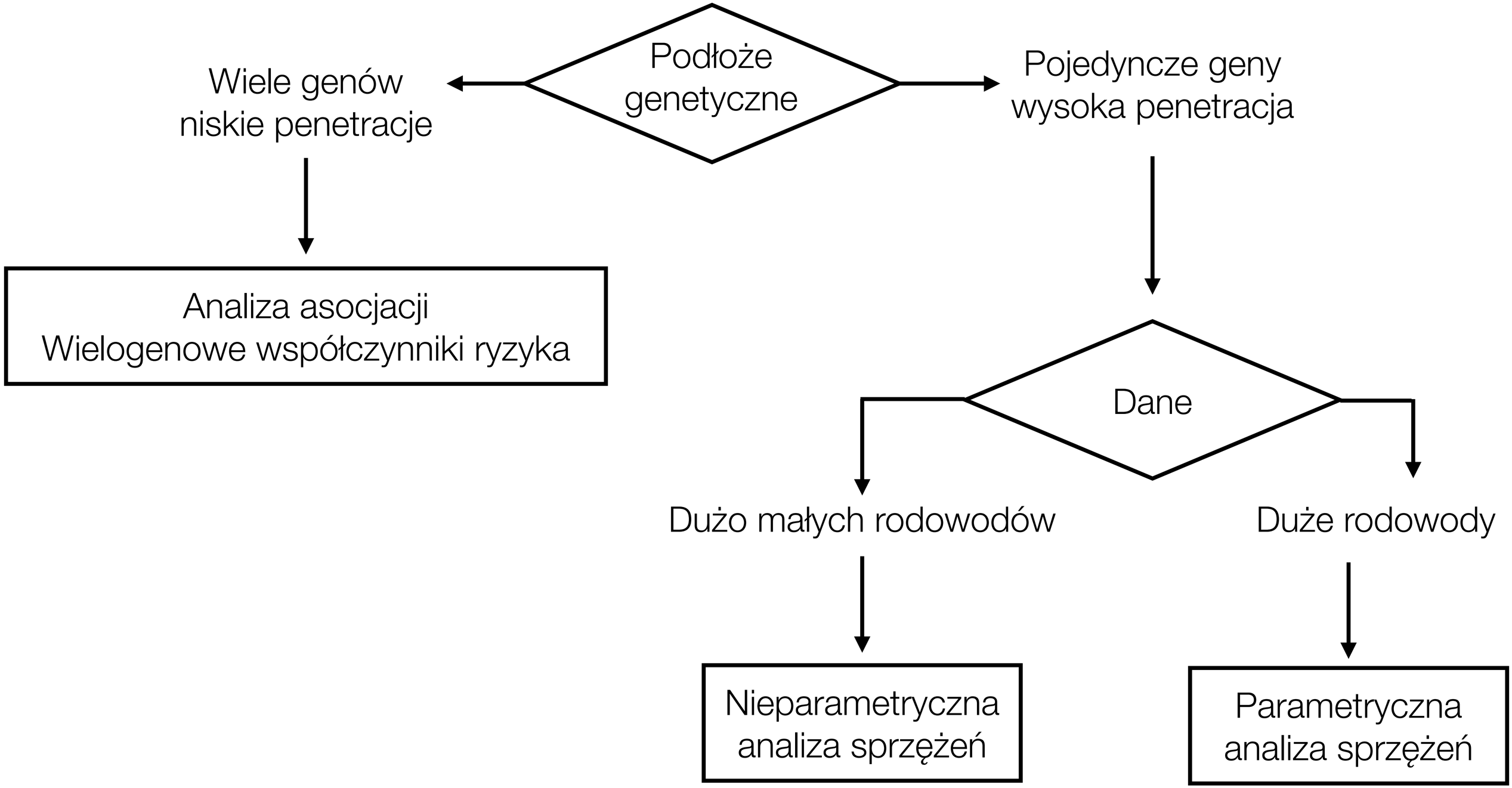
Melissa J. Landrum*, Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, Wonhee Jang, Kenneth Katz, Michael Ovetsky, George Riley, Amanjeev Sethi, Ray Tully, Ricardo Villamarin-Salomon, Wendy Rubinstein and Donna R.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

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

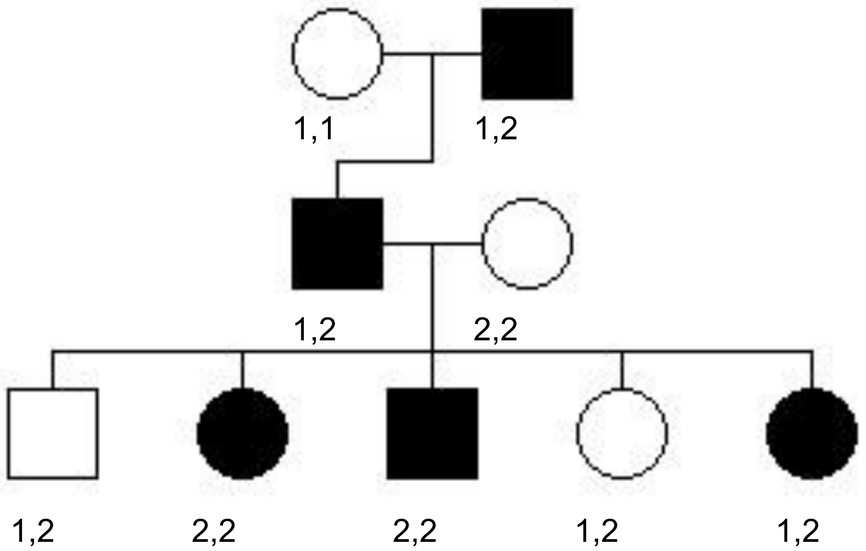
Received September 14, 2015; Revised October 15, 2015; Accepted October 24, 2015

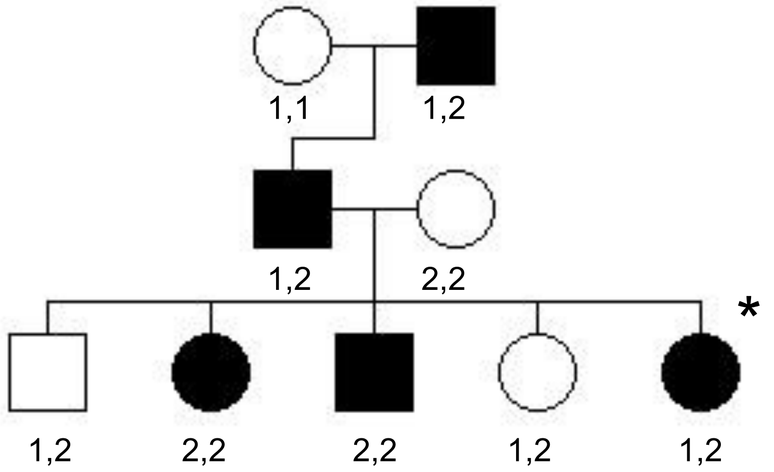
Sprzężenia w epoce genomu

- 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 odpowiadają za fenotyp nie zawsze jest ewidentne
  - Łatwiej w obszarach kodujących
- Sekwencjonowanie stosowane w rzadkich chorobach, gdzie nie ma dostatecznie dużo rodowodów


Podłoże genetyczne
Wiele genów niskie penetracje
Pojedyncze geny wysoka penetracja
Analiza asocjacji Wielogenowe współczynniki ryzyka
Dane
Dużo małych rodowodów
Duże rodowody
Nieparametryczna analiza sprzężeń
Parametryczna analiza sprzężeń

Ćwiczenie

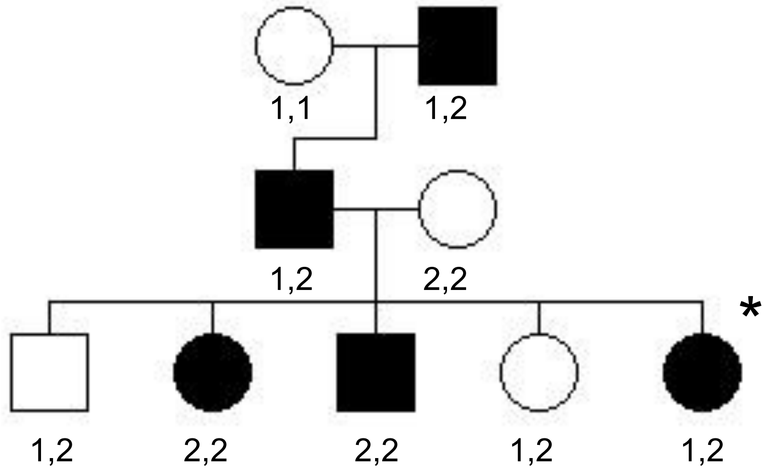

1,1
1,2
1,2
2,2
1,2
2,2
2,2
1,2
1,2



1 rekombinant (R); 4 rodzicielskie (NR)

Przy braku sprzężenia (θ=0,5) prawdopodobieństwo uzyskania R i NR jest jednakowe i wynosi ½

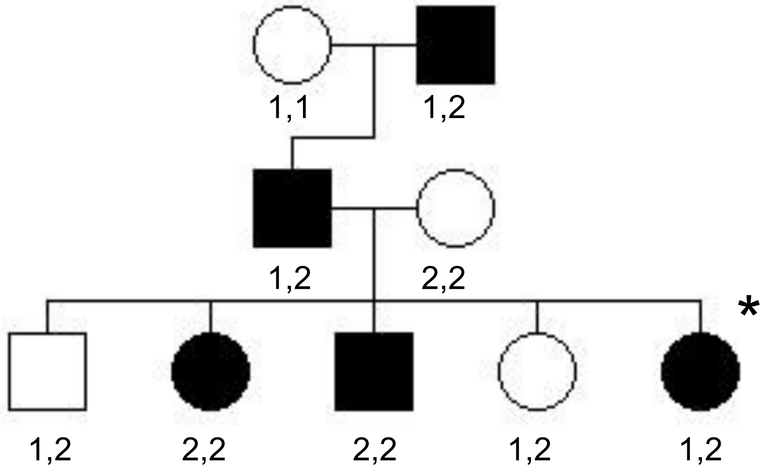
$L(\theta=0,5)= (½)^5$



1 rekombinant (R); 4 rodzicielskie (NR)

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

$$L(\theta) = \theta^{1R} \cdot (1-\theta)^{4NR}$$

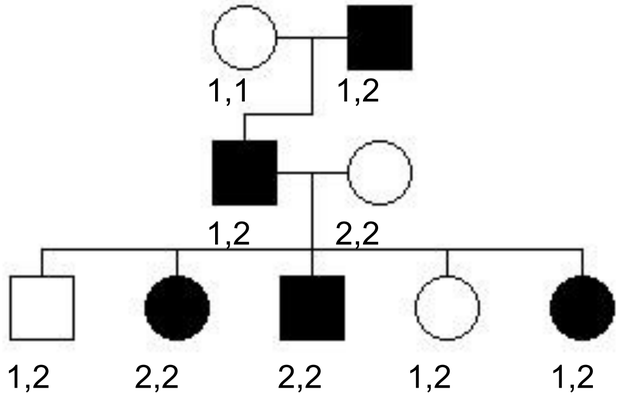


1 rekombinant (R); 4 rodzicielskie (NR)

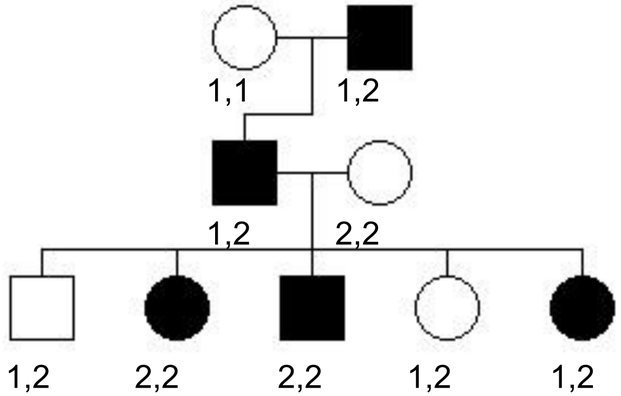
$L(\theta=0,5)= (½)^5$ $L(\theta)= \theta\cdot(1 - \theta)^4$

Dla $\theta=0,1$ $L(\theta=0,1) = 0,1\cdot(0,9)^4$

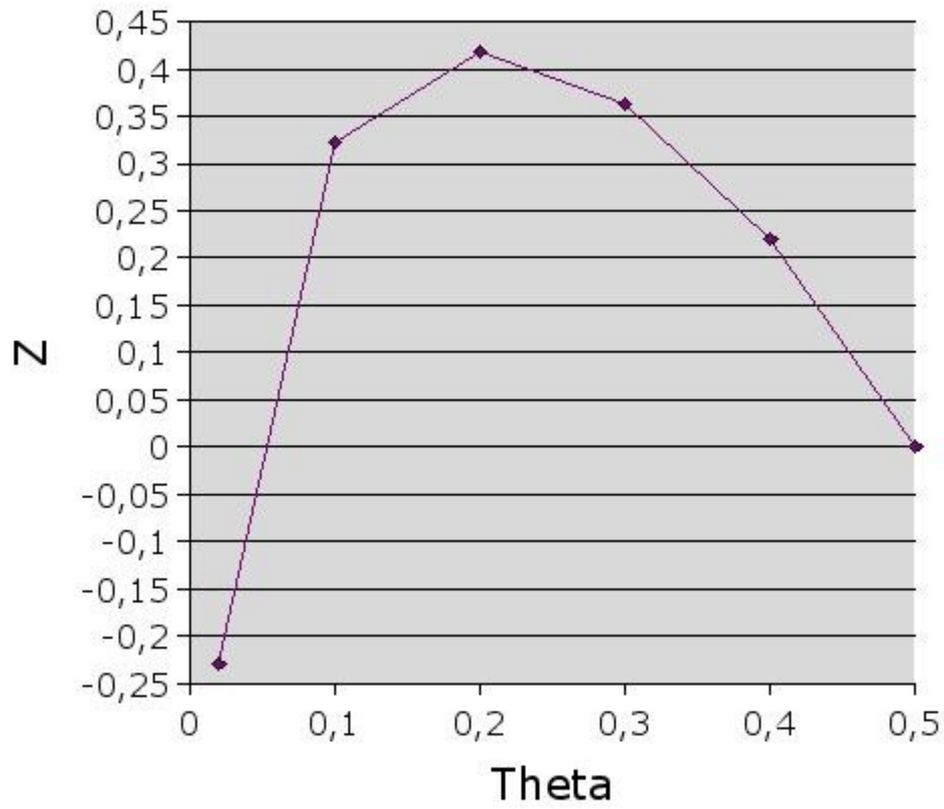
$$Z(\theta = 0,1) = \log_{10}\left(\frac{0,1\cdot 0,9^4}{0,5^5}\right) \approx 0,32$$



0	0,02	0,1	0,2	0,3	0,4	0,5

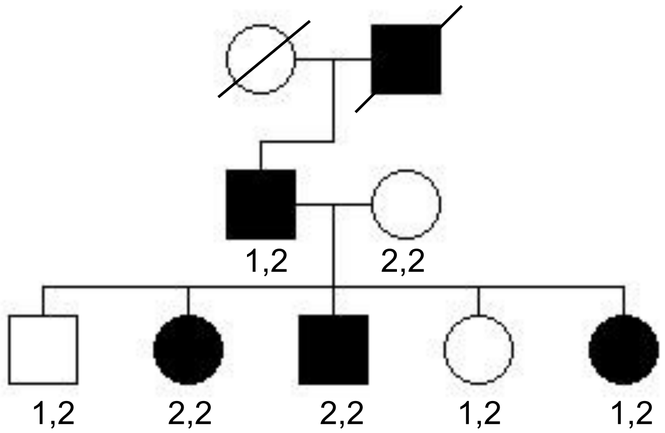


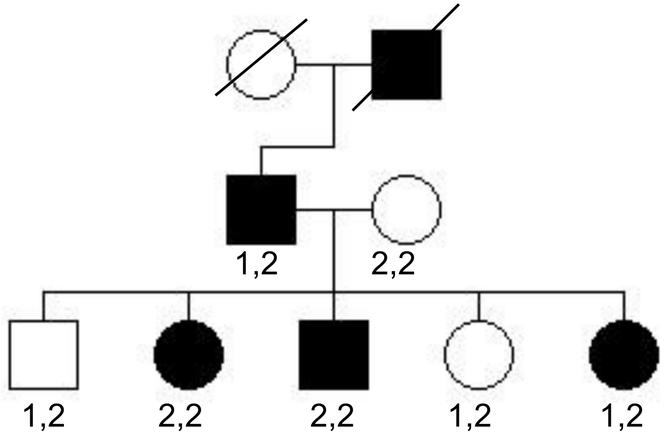
0	0,02	0,1	0,2	0,3	0,4	0,5
-∞	-0,23	0,32	0,42	0,36	0,22	0

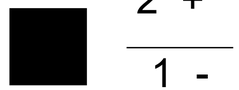


0	0,02	0,1	0,2	0,3	0,4	0,5
-∞	-0,23	0,32	0,42	0,36	0,22	0

Nieznana faza markera u ojca





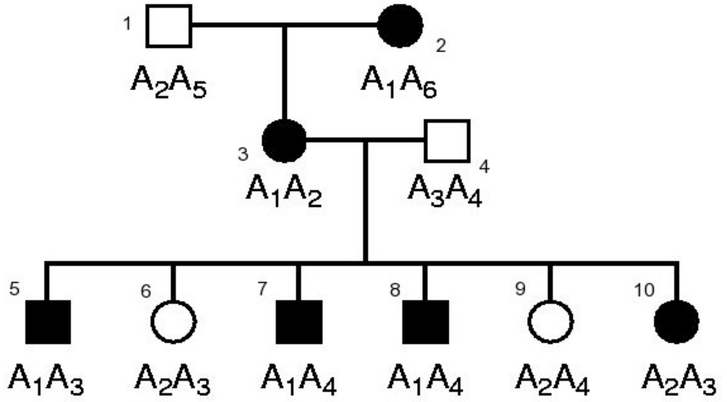
$\frac{1\ +}{2\ -}$ albo  $\frac{2\ +}{1\ -}$

1R 4NR 1NR 4R

$$L(\theta = 0{,}2) = \left(\frac{0{,}2 \cdot 0{,}8^4}{2} \right) + L(\theta = 0{,}2) = \left(\frac{0{,}2^4 \cdot 0{,}8}{2} \right)$$

$$Z(\theta = 0{,}2) = \log_{10} \left(\frac{\frac{0{,}2 \cdot 0{,}8^4}{2} + \frac{0{,}2^4 \cdot 0{,}8}{2}}{0{,}5^5} \right) \approx 0{,}12$$

Kodowanie rodowodu

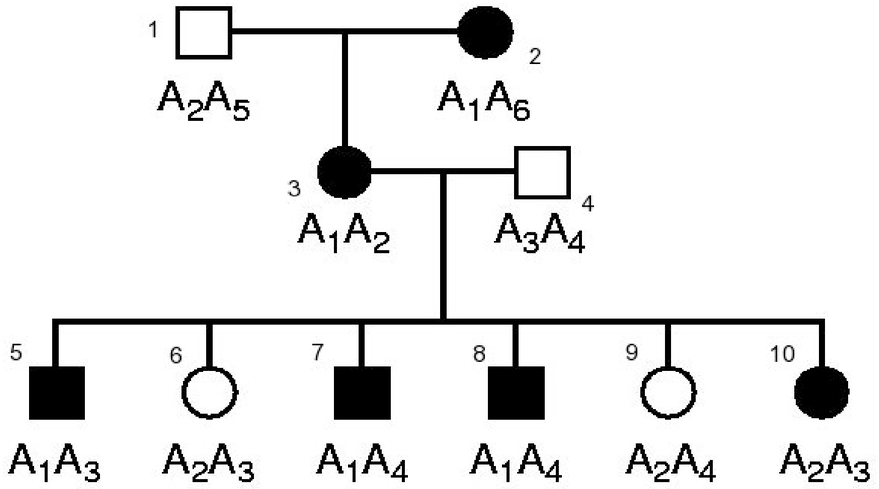


Płeć: 1 mężczyzna, 2 kobieta, 0 nieznana
Choroba: 1 zdrowa(y), 2 chora(y), 0 nieznana
0 zawsze oznacza nieznane/brak danych!!!

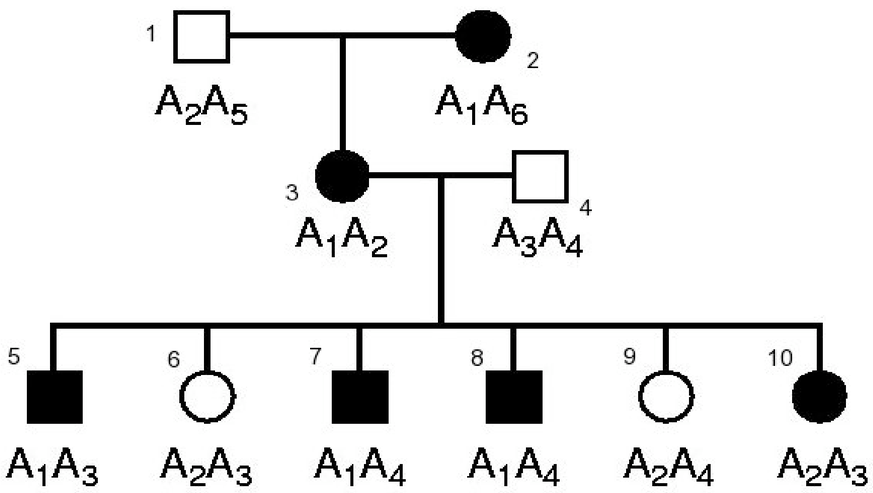
Rodzina	osoba	ojciec	matka	płeć	choroba	marker1a1	marker1a2
001	1	0	0	1	1	2	5
001	2	0	0	2	2	1	6
001	3	1	2	2	2	1	2
001	4	0	0	1	1	3	4
001	5	4	3	1	2	1	3

Liczba spacji nieistotna (minimum 1)

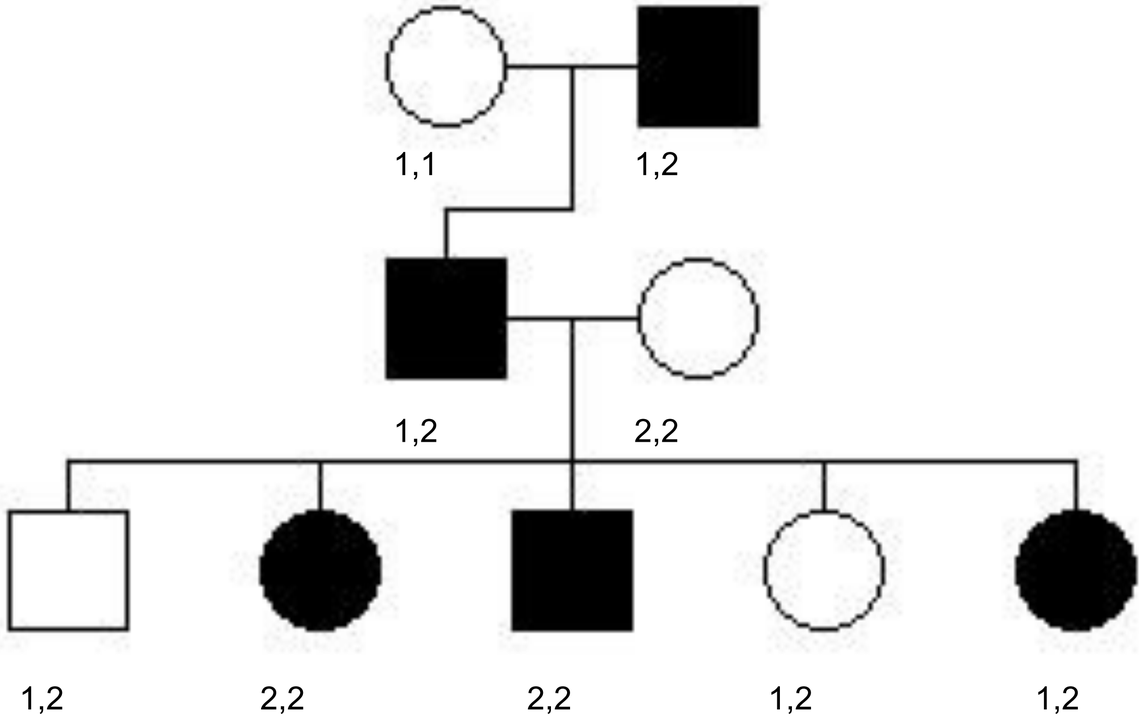
Plik .ped




```
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!!!

Pliki danych w analizach sprzężeń i GWAS

- rodowód + genotypy i fenotypy (niekiedy w osobnych plikach)
- mapa markerów

dodatkowo w analizach sprzężeń:

- opis *loci*
- opis modelu dziedziczenia

Pliki mapy

```
CHROMOSOME MARKER      POSITION
1          Marker1     0
```

format .map

```
1 rs12073590 0.029735 1205155 C A
1 rs6685064  0.029785 1211292 T C
1 rs61559999 0.030045 1235792 T C
1 rs62623580 0.030111 1254255 A G
```

format .bim

Plik opisu *loci* (.dat)

Oznaczenia:

- A (Affecction) : chory/zdrowy
- M (Marker)
- T (Trait): cecha ilościowa
- C (Covariate): dodatkowy parametr (klasa ryzyka, wiek, itp.)

- Nazwy markerów muszą być takei same, jak w pliku .map

```
A       Choroba
M       Marker1
```

Plik modelu (.model)

- Nazwa choroby musi być taka sama, jak w pliku .dat

```
DISEASE          ALLELE_FREQ      PENETRANCES      LABEL
Choroba          0.001            0.0,1.0,1.0      Dominujaca
```

```
DISEASE          ALLELE_FREQ      PENETRANCES              LABEL
PROSTATE_CANCER  0.001            *                        Complex
   SEX = FEMALE                   0.000,0.000,0.000
   AGE < 50                       0.001,0.050,0.100
   AGE < 70                       0.002,0.200,0.400
   OTHERWISE                      0.004,0.500,0.800
```

Analiza sprzężeń

- komenda merlin lub minx (dla cech sprzężonych z płcią)
- parametry: pliki z danymi i odległości
- podawanie odległości:
  - bezpośrednio w cM: --positions:0,1,5,10,20,30,40
  - równomierne kroki --steps 5
    - przedział między każdą parą markerów podzielony na 5 części