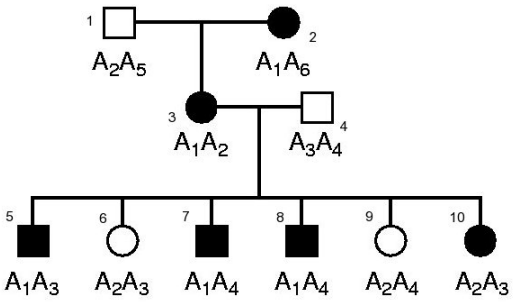
Formaty danych w analizie genetycznej człowieka i analiza sprzężeń

Pliki danych. Dane do analizy zawarte są w plikach tekstowych. Najważniejsze z nich, niezbędne w każdej analizie, to: plik **rodowodu** (zawierający też genotypy) i plik **mapy** markerów, w analizach sprzężeń dodatkowo potrzebny jest plik **opisu *loci*,** i (w analizach parametrycznych) plik **modelu** dziedziczenia.

Rodowód. Rodowód zapisywany jest w tzw. formacie LINKAGE (od nazwy jednego z najstarszych programów do analizy sprzężeń). Jeden plik może zawierać wiele rodowodów (rodzin). Każda linijka opisuje jedną osobę, a dane podane są w kolumnach oddzielonych spacjami lub tabulatorami (dowolna liczba). Zwyczajowo plik z rodowodem ma rozszerzenie .ped, ale nie jest to wymagane. W analizach GWAS często każda osoba jest *de facto* odrębną rodziną (unika się krewnych), ale format musi być zachowany.

Dla uproszczenia warto zacząć od ponumerowania osób w rodowodzie (lub nadania im innych unikatowych identyfikatorów). Format pliku rodowodu objaśnia prosty przykład poniżej:



Plik .ped dla tego rodowodu wygląda następująco

```
1      2  3  4  5  6  7  8
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
```

Te numery kolumn podano dla ułatwienia, nie występują w pliku!

Pierwsze pięć kolumn jest stałe i są to:

1 – identyfikator rodziny, w tym pliku jest tylko jedna rodzina, ale identyfikator musi zawsze występować
2 – identyfikator kolejnej osoby
3 – identyfikator ojca, 0 oznacza, że nie występuje w tym rodowodzie
4 – identyfikator matki, 0 oznacza, że nie występuje w tym rodowodzie. Jeżeli w rodowodzie brak jednego z rodziców, to drugi też musi być oznaczony jako 0!
5 – płeć: 1 – mężczyzna, 2 – kobieta

Kolejne kolumny kodują fenotypy i genotypy, zwykle zaczynając od genotypu choroby. To, jakie cechy tu zakodowano jest opisane w pliku opisu *loci* (.dat). W tym rodowodzie mamy jedną cechę choroby, (kolumna *6*, 1 – zdrowy, 2 – chory, 0 – brak danych) i jeden marker (kolumny *7* i *8* dla obu alleli u każdego osobnika). Jeżeli kodujemy markery leżące na chromosomie X, to u mężczyzn podajemy allel dwukrotnie (tak, jakby byli homozygotami).

Taki format zapisu rodowodu jest wykorzystywany przez wiele programów do analiz genetycznych.

Niekiedy dane te rozdziela się na dwa pliki: właściwy rodowód (kolumny od *1* do *5*) i genotypy (pozostałe kolumny). Rozdział na plik rodowodu i plik genotypów jest standardem w programach do analizy asocjacji. Do konwersji między różnymi formatami plików w analizach genetycznych można stosować np. program Mega2[1] albo proste skrypty.

Mapa. W prowadzonych współcześnie analizach dysponujemy mapą markerów molekularnych, np. miejsc polimorfizmu sekwencji (SNP). Mapę tę zawiera plik standardowo z rozszerzeniem .map. Jest on wymagany nawet wtedy, gdy (jak w naszym prostym przykładzie) mamy tylko jeden marker. W pliku są trzy kolumny: chromosom, nazwa markera i pozycja na chromosomie (w cM). Umieśćmy nasz przykładowy marker na chromosomie 1 w pozycji 0. Plik będzie wtedy wyglądał tak (pierwsza linijka z nagłówkiem nie jest obowiązkowa, ale może dla ułatwienia zostać w pliku):

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

Nazwy markerów muszą oczywiście odpowiadać zadeklarowanym w pliku .dat! Można też uwzględniać różnice w częstości rekombinacji u kobiet i u mężczyzn, wtedy dodajemy dwie dodatkowe kolumny, odpowiednio FEMALE_POSITION i MALE_POSITION.

W rzeczywistych analizach występuje oczywiście o wiele więcej markerów, a ich realne pozycje muszą być znane. Dla wielu stosowanych w analizach sprzężeń zestawów markerów dostępne są gotowe pliki z mapami[2].

Kolejne dwa typy plików stosowane są w analizie sprzężeń, nie są konieczne w analizach typu GWAS

Opis *loci*. Ten plik (standardowo z rozszerzeniem .dat) opisuje rodzaj informacji zawartych w pliku rodowodu w kolumnie *6* i kolejnych. W naszym przypadku pierwszą cechą jest choroba (skrót A od ang. *affection*). Markery oznaczamy kodem M.
Poza tymi typowymi cechami możemy jeszcze mieć cechy ilościowe (T od ang. *trait*) opisywane liczbą rzeczywistą (np wzrost, poziom enzymu itp.) oraz kowarianty (C od ang. *covariate*), czyli wielkości zależne dla określania klas ryzyka (np. wiek, czynniki środowiskowe).

W drugiej kolumnie podajemy nazwę cechy - może być dowolna (byle nie zawierała spacji lub przecinków), ale tę samą nazwę musimy wykorzystywać we wszystkich plikach w analizie. Dla naszego przykładowego rodowodu plik .dat wyglądać może tak:

```
A     Choroba
M     Marker1
```

Opis modelu. W tym pliku, typowo z rozszerzeniem .model (lub .mod) podajemy model dziedziczenia. W pliku są cztery kolumny: nazwa choroby (taka sama, jak w pliku .dat), częstość allelu sprawczego, penetracje (prawdopodobieństwo zachorowania odpowiednio dla 0, 1 i 2 kopii allelu sprawczego w genotypie) i nazwa modelu (dowolna). Wiersz nagłówka nie jest obowiązkowy, ale może dla czytelności występować w pliku. Dla naszego prostego przykładu będzie to:

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

Penetracje oddzielamy przecinkami. Pamiętajmy, że w standardzie anglosaskim separatorem dziesiętnym jest kropka!

---

[1] https://watson.hgen.pitt.edu/docs/mega2_html/

[2] Np. http://compgen.rutgers.edu/download_maps.shtml

Plik modelu może opisywać dużo bardziej złożone modele, np. uwzględniając różne prawdopodobieństwa zachorowania zależnie od płci. Oto opis choroby sprzężonej z płcią, gdzie kobiety - nosicielki mają 50% ryzyko zachorowania.

```
DISEASE              ALLELE_FREQ     PENETRANCES          LABEL
Choroba2             0.05            *                    Kodominacja
   SEX = FEMALE                      0.00,0.50,1.00
   OTHERWISE                         0.00,0.00,1.00
```

A oto przykład[3] cechy, gdzie występują różne grupy ryzyka zależnie od płci i kowariantu, jakim jest wiek.

```
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
```

Spróbuj opisać słowami model z powyższego pliku.

Powyższe cztery pliki są niezbędne w każdej analizie parametrycznej (w analizach nieparametrycznych nie jest oczywiście potrzebny plik .model). Dodatkowo można w osobnym pliku podać dane o częstości występowania alleli każdego markera w populacji.

Analiza sprzężeń. Gdy mamy już wszystkie pliki z danymi, sama analiza jest prosta. W przypadku, gdy mamy tylko jeden marker musimy podać, dla jakich odległości od markera (w cM) mamy obliczyć *LOD*.

Odpowiednia komenda to:

```
merlin -d przyklad.dat -p przyklad.ped -m przyklad.map --model przyklad.model --positions:0,1,5,10,20,30,40
```

Pierwsze cztery parametry są oczywiste: plik opisu, rodowód i mapa (odpowiednio -d, -p i -m), następnie specyfikacja modelu parametrycznego (--model, zwróć uwagę na podwójną kreskę!) i wreszcie lista odległości od markera w cM, dla których chcemy mieć obliczone *LOD*.

Ćwiczenie.

1. Dla przykładowego rodowodu opisanego powyżej oblicz *LOD* dla 0, 0.05, 0.1, 0.2, 0.3 i 0.4 częstości rekombinacji prostą metodą omówioną wcześniej. Następnie przygotuj pliki i przeprowadź analizę programem MERLIN (pamiętaj, że 0.1 częstość rekombinacji to 10 cM). Czy wyniki są dokładnie takie same? Czy maksimum *LOD* występuje dla tej samej wartości? Czy odchylenia są większe dla małych, czy dla dużych odległości? Co pamiętasz ze wstępu o zależności częstości rekombinantów od odległości genetycznej?
2. We wskazanym przez prowadzących katalogu znajdziesz cztery pliki do analizy mającej na celu zlokalizowanie mutacji sprawczej dla pewnej choroby (dziedziczna niepełnosprawność intelektualna) dziedziczonej jako cecha sprzężona z płcią. Ile markerów wykorzystano w badaniu? Dla analiz z wieloma markerami nie podaje się odległości od markera (jak w pierwszym przykładzie), tylko w ilu punktach (jak gęsto) pomiędzy każdą parą markerów należy obliczyć *LOD*. Dla cech sprzężonych z płcią komenda to nie `merlin` tylko `minx`. W tym przypadku:

   ```
   minx -p FamD.ped -d FamD.dat -m FamD.map --model FamD.model --steps 0
   ```

   W której pozycji uzyskaliśmy najwyższą wartość *LOD* i któremu markerowi (patrz plik .map) ona odpowiada? Wyniki można też przedstawić w formie graficznej dodając do powyższej komendy opcję `--pdf`

[3] Ze strony http://csg.sph.umich.edu/abecasis/merlin/reference/parametric.html