Identyfikacja biomarkerów chorób rozszczepu i braków zębowych przy użyciu danych z paneli genetycznych.

Genomika jest istotnym narzędziem w nowoczesnej medycynie, które pozwala na lepsze zrozumienie, diagnostykę i leczenie wielu chorób. Nasza analiza opierać się będzie na danych pochodzących z paneli genetycznych (Comprehensive Genomic Profiling - CGP) dotyczących 444 genów.
W tym zadaniu skupimy się na analizie danych genetycznych, której celem jest identyfikacja genów, które z istotnością statystyczną rozróżniają pacjentów z rozszczepem wargi i podniebienia oraz pacjentów z brakami zębowymi od grupy kontrolnej.
Następnie zbudujemy model ML, który będzie trenowany na danych genomicznych, w celu klasyfikacji nowych próbek do kategorii: kontrola, rozszczep lub braki zębowe.

Zadania:

1. Przygotowanie środowiska pracy:
    a. Przygotowanie repozytorium w gitlabie (branch system main -> feature)
        i. Każdy pracuje na własnym feature branch
        ii. Wszyscy robią rebase/merge main do feature zanim otworzą merge request
        iii. Merge request wraz z code review przez pozostałych członków zespołu
        iv. Merge do main
        v. Pull z najnowszego brancha main przed rozpoczęciem pracy na feature.
2. Czyszczenie i przygotowanie danych:
    a. Pobranie danych (link od p. Liliany)
    b. Łączenie plików i przypisywanie kategorii (1)
    c. Czyszczenie danych: zamiana “.” na NaN, zamiana spacji na “_” (1)
    d. Zliczenia:
        i. Przypisz do zmiennej liczbę pacjentów w każdej z kohorcie (1)
        ii. Filtrowanie: usuń, zachowując kolejność (5):
            1. synonimiczne zmiany na poziomie białka,
            2. Warianty, które mają mniej niż 3 metryki wskazujące na negatywny efekt wariantu na białko (SIFT, LRT, MutationAssessor, FATHMM, PROVEAN, MetaSVM),
            3. warianty łagodne i potencjalnie łagodne,
            4. Warianty które występują w populacji europejskiej (1000gp_EUR_freq) z częstością większą niż 1%
            5. Wyodrębnij dwa wyniki, jeden gdzie są tylko warianty homozygotyczne i drugi gdzie są warianty homo i heterozygotyczne
        iii. Binaryzacja (4)
            1. Na danych z punktu II, dla każdego pacjenta, dla każdego genu przypisz wartość 1 jeśli występuje jakakolwiek mutacja (z potencjałem patogenności) i 0 jeśli jej brak.

2. Zapisz w macierzy której kolumny to geny a rzędy to pacjenci (jedna kolumna powinna zawierać typ), podczas transformacji zamień NaN na 0

iv. Przygotowanie macierzy do testu Chi2/fisher exact. Wewnątrz kohorty, ilu jest pacjentów z co najmniej jedną mutacją wskazującą na patogenność per gen (po odfiltrowaniu wariantów i binaryzacji). Wewnątrz kohorty, ilu jest pacjentów bez mutacji wskazujących na patogenność per gen (4)

	cleft_with_pathogenic	cleft_without_pathogenic	control_with_pathogenic	control_without_pathogenic
gene1				
gene2				
gene3				

3. Analiza eksploracyjna i testy statystyczne
    a. Chi2: (4)
        i. per kategoria (braki zębowe - kontrola, rozszczep - kontrola) zrób analizę statystyczną używając testu Chi2
        ii. Dla każdej pary (braki zębowe - kontrola, rozszczep - kontrola) wskaż, które geny są istotnie różne między grupami (kontrola - choroba)
    b. Fisher exact test: (4)
        i. per kategoria (braki zębowe - kontrola, rozszczep - kontrola) zrób analizę statystyczną używając testu fisher exact test
        ii. Dla każdej pary (braki zębowe - kontrola, rozszczep - kontrola) wskaż, które geny są istotnie różne między grupami (kontrola - choroba)
    c. Heatmapa z wyników z binaryzacji po odfiltrowaniu nieistotnych genów z Chi2/fisher (2)
        i. razem dla kontroli, braków zębowych i rozszczepów
        ii. Oddzielnie dla kontroli + rozszczep i kontroli + braki zębowe
    d. Do wyboru po jednym genie z wynikiem istotnym statystycznie (3):
        i. Przeszukaj literaturę (np. Pubmed, OMIM) pod kątem udziału genu w patologii powiązanej choroby (rozszczep lub braki zębowe) z uwzględnieniem ścieżki w której uczestniczy gen i opracuj slajd do prezentacji. Zaznacz wszystkie zidentyfikowane geny na ścieżce.
4. Modelowanie predykcyjne
    a. Przeanalizuj wszystkie uzyskane transformacje danych (wyniki punktów 2d i 3) i zaproponuj feature'y na zagregowanych danych do modelowania
    b. Pytania pomocnicze:
        i. Jakie cechy musi mieć wariant, żeby był istotny dla choroby?
        ii. Jakie warianty/geny sumarycznie tłumaczą fenotyp choroby?
    c. Wybierz model i zrób klasyfikację w oparciu o jednostkę chorobową