Dokładne wykrycie klinicznie istotnych wariantów liczby kopii (CNV) jest niezbędne w diagnozowaniu chorób genetycznych, ponieważ CNV są odpowiedzialne za znaczną część chorób. Podczas gdy bioinformatyczne potoki przetwarzania specjalizujące się w wykrywaniu wariantów pojedynczych nukleotydów i krótkich indeli przy użyciu danych z całego sekwencjonowania egzomu (WES) zapewniają zadowalającą wydajność (https://precision.fda.gov/challenges/consistency), identyfikacja większych delecji i duplikacji nadal stanowi wyzwanie. Chociaż opracowano mnóstwo narzędzi do wywoływania CNV z danych WES, większość z tych algorytmów charakteryzuje się ograniczoną rozdzielczością, niewystarczającą wydajnością i niezadowalającymi miernikami klasyfikacji. Chociaż dostrajanie parametrów wywoływania CNV może znacznie poprawić ogólną wydajność algorytmu, wciąż nie ma dobrze ustalonych wytycznych, które pomogłyby zoptymalizować wykrywanie CNV.

Obecnie analizy obliczeniowe, w szczególności wykrywanie CNV, jest głównym wąskim gardłem w wielu projektach sekwencjonowania organizmów. Istnieje ponad 25 narzędzi do wykrywania zmiany liczby kopii DNA (CNV) na podstawie sekwencjonowania pełnoeksomowego przy użyciu analizy głębokości pokrycia. Istniejące narzędzia składają się z kilku etapów, w tym: (i) obliczenia głębokości pokrycia dla każdego regionu sekwencjonowania, (ii) normalizacji, (iii) segmentacji i (iv) wykrywania CNV. Zasadniczym aspektem całego procesu jest etap normalizacji, w którym usuwane są systematyczne błędy i tendencyjności, a zestaw próbek odniesienia służy do zwiększenia stosunku sygnału do szumu.

Pomimo tego, że niektóre narzędzia do wykrywania CNV używają dedykowanych algorytmów do otrzymania zbioru próbek referencyjnych, większość zaawansowanych narzędzi nie zawiera tego kroku. Zbiór próbek referencyjnych jest wyznaczany na podstawie korelacji pomiędzy próbkami liczonej w regionach sekwencjonowania podzielonych na grupy wg ich przynależności do chromosomu.

W projekcie chcemy podzielić regiony na grupy nie wg przynależności do chromosomu (jak to było do tej pory), ale wg cech targetów, np. mediana głębokości pokrycia liczona wzdłuż wszystkich próbek, zawartość par GC itp.

Według naszej wiedzy będzie to pierwsza próba stworzenia algorytmu do dzielenia algorytmu na grupy. Oczekujemy, że nowy podział na grupy (nie wg przynależności do chromosomów) zmieni znacząco grupy próbek referencyjnych dla danej próbki, co znacząco poprawi jakość wykrytych CNV.