

Technologia sekwencjonowania pozwala masowo odczytywać sekwencję nukleotydów w krótkich fragmentach DNA. W większości eksperymentów wykorzystujących sekwencjonowanie pierwszy etap analizy polega na zrekonstruowaniu obszarów pochodzenia odczytów w genomie. W tym celu odczyty *mapuje się* na *genom referencyjny*, złożony z najbardziej typowych dla danego gatunku wariantów sekwencji. Genomy referencyjne są obecnie dostępne dla tysięcy gatunków, a główny ciężar badań przenosi się w kierunku analizy różnorodności genomowej w obrębie poszczególnych gatunków. Jest to szczególnie widoczne w przypadku genomiki człowieka, gdzie rozwój jest napędzany perspektywą zastosowań w medycynie spersonalizowanej. Jednakże, pomimo ogromnej i wciąż rosnącej ilości dostępnych danych, obecnie stosowane procedury analizy ograniczają się do wykorzystania genomów referencyjnych, co niepotrzebnie zaburza wyniki badań.

Temu właśnie problemowi poświęcony jest niniejszy projekt. Wprowadzimy w nim pojęcie multi-genomu referencyjnego, który pozwoli modelować różnorodne warianty poszczególnych obszarów genomu. Opracujemy narzędzia, które włączą multi-genomy w obecnie stosowane procedury analizy danych z sekwencjonowania. Skoncentrujemy się na etapie mapowania odczytów, co pozwoli zachować do dalszych etapów odczyty nie pasujące do genomu referencyjnego. Zalety naszego podejścia zostaną zilustrowane analizą danych z sekwencjonowania w badaniach powstawania pęknięć i mechanizmów naprawy DNA.

Podsumowując, w wyniku naszego projektu powstanie kompletny zestaw narzędzi do przetwarzania danych z sekwencjonowania DNA, który (1) wykorzysta dostępne informacje o różnorodności genomowej, (2) będzie gotowy do zastosowania przez nie-informatyków w ramach różnorodnych procedur analizy danych oraz (3) w praktyce wykaże swe przewagi nad tradycyjnymi metodami. Nasze podejście będzie użyteczne dla szerokiej gamy badań naukowych korzystających z technologii sekwencjonowania DNA, w tym dla genomiki nowotworów i medycyny spersonalizowanej.