Streszczenie

Najnowsze odkrycia w sekwencjonowaniu wysokoprzepustowym (HTS) przyczyniły się do bezprecedensowego wzrostu ilości generowanych danych multiomicznych. Z jednej strony, przekroczenie w połowie pierwszej dekady XXI w. historycznego progu, jakim był koszt kompletnej sekwencji genomu człowieka (sekwencjonowanie całego genomu, WGS) poniżej tysiąca dolarów, otworzyło drzwi do wielu narodowych inicjatyw genomowych, takich jak 100,000 Genomes Project w Wielkiej Brytanii, czy 1000 Polskich Genomów.

Z drugiej strony, wiele popularnych metod bioinformatycznych do analizy drugo- i trzeciorzędowej wykazuje dużą złożoność obliczeniową. Większość istniejących narzędzi i algorytmów analizy genomicznej jest z natury sekwencyjna i nie jest w stanie w pełni wykorzystać możliwości rozproszonego modelu obliczeń, co czyni sytuację jeszcze trudniejszą. W szczególności, jak dotąd nie zaproponowano prawdziwie skalowalnych metod dla typowych operacji genomicznych, takich jak obliczanie głębokości pokrycia, podsumowywanie krótkich odczytów (ang. *pileup*) i łączenie zbiorów danych za pomocą przecięć przedziałowych.

Ponadto, znikoma liczba badań podejmuje temat wyzwań związanych z projektowaniem genomicznych platform chmurowych do rozproszonego przetwarzania i analizy danych pochodzących z HTS. Podobnie mało uwagi poświęcono idei wykorzystania zunifikowanego podejścia, realizującego deklaratywny paradygmat programowania do wyrażania operacji genomicznych przy użyciu języka Structured Query Language (SQL).

Niniejsza rozprawa ma na celu wypełnienie tych luk poprzez przedstawienie koncepcji Genomicznej Platformy Danych typu Lakehouse oraz zaprezentowanie projektu SeQuiLa, implementującego nowatorskie skalowalne metody dla wyżej wymienionych, obliczeniowo wymagających, operacji genomicznych. Na poniższą pracę składa się seria sześciu publikacji poprzedzonych wstępem, w którym Autor opisuje wyzwania i najnowsze osiągnięcia w dziedzinie analizy danych genomicznych.

Słowa kluczowe: Big Data, obliczenia rozproszone, obliczenia chmurowe, sekwencjonowanie wysokoprzepustowe, architektura Data Lakehouse