Sekwencjonowanie genomów ma bezprecedensowy wpływ na nasze zrozumienie organizmów żyjących na Ziemi. Od 2007 roku, kiedy firma Illumina zaczęła dominować na rynku, koszt sekwencjonowania spadł drastycznie. Przykładowo, dla genomu ludzkiego była to zmiana z ok. 10 milionów USD do 1 tysiąca USD. Pozwoliło to na uruchomienie wielu ambitnych projektów sekwencjonowania, takich jak: The 1000 Genomes Project, The Genomics England 100,000 Genomes Project, The French Plan for Genomic Medicine 2025, Earth BioGenome Project, The Human Cell Atlas, The 100K Pathogen Genome Project. Niedawne szacunki wskazują, że w skali globalnej już w 2025 roku rocznie sekwencjonowanych będzie rzędu 1 zeta par zasad z czego 2–40 EB będzie musiało zostać przechowywane przez lata.

Współcześnie zdecydowana większość zsekwencjonowanych danych została otrzymana za pomocą urządzeń firmy Illumina. Technologia ta ma wiele zalet, włączając w to m.in. wysoką przepustowość i bardzo dobrą jakość wyników. Niestety, cechuje się ona też pewnymi wadami. Najbardziej istotną z nich jest to, że pozwala na poznanie tylko krótkich fragmentów genomu (odczytów), o długości ok. 250 par zasad, co, w porównaniu do rozmiaru ludzkiego genomu wynoszącego ok. 3 miliardy par zasad, jest wartością niewielką. W związku z tym, poznanie genomu przypomina układanie puzzli. Odczyty pochodzą z losowych miejsc w genomie, a ich sumaryczna długość jest zwykle dziesiątki razy większa niż jego długość. Naturalne jest więc, że pomiędzy odczytami występują spore nakładki. Oczywiście problemem może być rozmiar danych, ponieważ z pojedynczego eksperymentu możemy uzyskać nawet miliard odczytów. Co więcej, w genomach występują fragmenty powtarzalne o długości znacznie przekraczającej długość odczytu. Z tego też powodu, używając tylko tej technologii nie jest możliwe poznanie całego genomu. Dla przykładu ciągle ok. 5% ludzkiego genomu pozostaje dla nas zagadką.

Ograniczenia sekwenatorów firmy Illuminy dały sposobność do rozwoju urządzeń sekwencjonowania trzeciej generacji. Najbardziej znanymi przykładami są technologie Single Molecule Real Time (SMRT) firm Oxford Nanopore Technology (ONT) oraz Pacific Biosciences (PacBio). Szczegóły techniczne są różne, ale urządzenia obu firm pozwalają na otrzymanie znacznie dłuższych odczytów, osiągających długość nawet miliona par zasad. Niestety poważnym problemem jest tutaj stopa błędu wynosząca ok. 10% dla pojedynczej pary zasad; dla porównania analogiczna stopa błędu dla urządzeń Illuminy wynosi 0.1%. Tym niemniej, znaczącym symptomem, że w najbliższym czasie doświadczymy poważnych zmian na rynku jest fakt, że firma Pacific Biosciences została ostatnio przejęta przez lidera rynku, firmę Illumina.

Wysoka stopa błędu i znacznie większa długość odczytów powoduje konieczność przeprojektowania istniejących, bądź stworzenia nowych narzędzi do analizy i kompresji danych z sekwencjonowania. W ostatnich latach poczynionych zostało już sporo prób w tym kierunku. Tym niemniej, istniejące narzędzia trudno nazwać doskonałymi. W związku z tym, głównym celem niniejszego projektu jest opracowanie nowych algorytmów i struktur danych dla kilku problemów, które można napotkać w trakcie przetwarzania odczytów z urządzeń trzeciej generacji. Niektóre z tych narzędzi będą dotyczyły problemu kompresji odczytów oraz wyników analiz. Celem innych będzie znajdowanie nakładek pomiędzy długimi i zaszumionymi odczytami. Jednym z celów będzie też opracowanie narzędzia, które może zostać wykorzystane w diagnostyce medycznej (prawie) czasu rzeczywistego.