Od 2008 roku możliwy jest odczyt sekwencji DNA (genomu) oraz RNA (transkryptomu) dzięki technologii sekwencjonowania w ogólności nazywanej Sekwencjonowaniem Nowej Generacji (NGS – Next Generation Sequencing). Wcześniej odbywało się to na mniejszą skalę, metodą Sangera używaną od 1977 roku. Metoda ta była używana do rozszyfrowania kompletnego ludzkiego genomu (Human Genome Project) w roku 2000. W przeciwieństwie do metody Sangera, rozwiązania NGS generują miliony krótkich odczytów (obecnie około 150 nukleotydowych par zasad) sekwencji zwanych odczytami (ang. reads). Następnie krótkie odczyty poddawane zostają analizie, do której używa się specyficznych metod bioinformatycznych. Przykładem jest składanie w większe sekwencje lub dopasowanie do istniejącego genomu.

Obecnie, pojedyncza maszyna dostępna na rynku, najbardziej zaawansowana technologicznie (Illumina 4000) może wygenerować około 0.5T nukleotydowych par zasad w ciągu jednego dnia. Co powoduje, że odczyt ok. 12 całych ludzkich genomów może być uzyskany w 3.5 dnia pracy tej maszyny [1].

W analizie sekwencji istnieje wiele, różnych typów/podejść analizy statystycznej, które przetwarzają informacje o sekwencjach DNA i RNA do postaci użytecznej wiedzy dla nauk biologicznych bądź medycyny. NGS we współpracy z technikami analizy, włączając przetwarzanie danych typu Big Data staje się podstawą dla nowego obszaru genomowej medycyny personalizowanej. Obszar ten został nazwany również „medycyną precyzyjną" w przemówieniu prezydenta Stanów Zjednoczonych Baracka Obamy z 30 stycznia 2015 roku, w którym ogłosił Inicjatywę Medycyny Precyzyjnej (ang. Precision Medicine Initiative). „Inicjatywa Medycyny Precyzyjnej ma wykorzystywać postępy w dziedzinie genomiki, pojawiających się metod zarządzania i analizy dużych zbiorów danych przy jednoczesnej ochronie prywatności, wykorzystaniu technologii informacji na temat zdrowia w celu przyspieszenia odkryć biomedycznych" [2].
Znaczenie tej inicjatywy oraz opartych na genomie rozwiązaniach medycznych w ogólności przedstawiono w pozostałej części oświadczenia Białego Domu:
"Medycyna Precyzyjna jest innowacyjnym podejściem do zapobiegania chorobom oraz ich leczenia, która uwzględnia indywidualne różnice w ludzkich genach, środowisku oraz stylu życia. Precyzyjne medycyna daje lekarzom klinicznym lepiej zrozumieć złożone mechanizmy leżące u podstaw zdrowia pacjenta, choroby lub jej etapu. Pozwoli lepiej przewidzieć, które zabiegi będą najbardziej skuteczne. Postępy w medycynie precyzyjnej doprowadziły do nowych odkryć oraz wprowadzenia nowych metod leczenia, które są dostosowane do szczególnych cech fizycznych, takich jak profil genetyczny danej osoby czy indywidualne cechy nowotworu. To prowadzi do zupełnej zmiany sposobu leczenia chorób, takich jak rak. Przykładowo, pacjenci z dowolnym typem nowotworu będą poddawani rutynowym testom molekularnym w ramach opieki nad pacjentem. Testy te umożliwią lekarzom wybrać odpowiednie leki, które zwiększą szansę na przeżycie i zmniejszą ryzyko narażenia na niekorzystne skutki. "

Obecnie cena sekwencjonowania spada: 1000 dolarów za genom od końca 2014 roku, podczas gdy jakość rośnie ze względu na bardziej stabilne i sprawdzone protokoły laboratoryjne. Wąskim gardłem NGS jest obecnie proces uzyskania użytecznej wiedzy biologicznej z dużych zbiorów danych sekwencjonowania. Odpowiedzią na potrzeby precyzyjnej analizy dużych danych genomicznych jest dalszy rozwój w dziedzinie określanej od niedawna jako "data science" (nauka o danych) i w dziedzinie obliczeń chmurowych. W ciągu ostatnich lat, paradygmaty rozwoju oprogramowania i przetwarzania danych zwane map-reduce framework, umożliwiają dystrybucję przetwarzania na wirtualnym klastrze maszyn i zapewnić skalowalność wyników w stosunku do rozmiaru danych. W ostatnich 2 latach także rozwiązania maszynowego uczenia stały się skalowalne, co pozwala im działać na setki i tysiące komputerów równoległych w razie potrzeby. To otwiera drogę do rozwoju nowych, specjalistycznych metod uczenia maszynowego, które będą w stanie rozwiązać nowe klasy problemów analizy danych NGS typu Big Data.

Dotychczas stosowane techniki analizy danych NGS w większości pochodzą z obszaru statystyki. W moim projekcie gEMiLia, chciałabym połączyć moje doświadczenie w analizie danych genomowych, które zdobyłam podczas przygotowywania pracy doktorskiej
z nowoczesnymi rozwiązaniami maszynowego uczenia, które są jednym z nurtów badań Zakładu Systemów Informatycznych, w którym pracuję od 01.01.2004.

NGS oraz obliczenia chmurowe znalazły się w 12 „przełomowych technologii" w raporcie McKinsey'a, opublikowanym w 2013 roku [3]. Wskazuje on na 12 technologii, które mogą wpłynąć na poważne zmiany gospodarcze w najbliższych latach. W moim projekcie spodziewam się połączenia ich w całość oraz dodania dodatkowego ważnego czynnika, którym jest użycie metod z obszaru maszynowego uczenia. Obszar ten jest powszechnie znanym, ale ciągle żywym i użytecznym. Takie połączenie jest nowatorskie i dotychczas niewiele na ten temat zrobiono w światowej społeczności naukowej. Połączenie genomowej wiedzy bioinformatycznej, obliczeń chmurowych oraz maszynowego uczenia jest rzadkością. Wierzę, że jest wiele do osiągnięcia poprzez uzyskanie efektu synergii z nich, z udziałem mojej codziennej pracy badawczej.
Mój plan skupia się na trzech ścieżkach analizy danych: jakości danych NGS, analizy wariantów DNA oraz analizy RNA, które technologicznie mają wiele wspólnego. Wyniki będą miały głównie formę nowych rozwiązań algorytmicznych, które można uruchomić równolegle na wielu komputerach w środowisku chmurowym. To w konsekwencji pozwala rozwiązać nowe klasy projektów analizy danych w dziedzinie biologii molekularnej i medycyny personalizowanej.

1. Illumina 4000 specifications. http://www.illumina.com/systems/hiseq-3000-4000/specifications.html
2. FACT SHEET: President Obama's Precision Medicine Initiative. Illumina 4000 specifications. http://www.illumina.com/systems/hiseq-3000-4000/specifications.html
3. Disruptive technologies: Advances that will transform life, business, and the global economy. http://www.mckinsey.com/insights/business_technology/disruptive_technologies