Od Sangera do sekwencjonowania genomów – przegląd technologii sekwencjonowania DNA

STRESZCZENIE

Nie ma techniki, która wniosłaby większy wkład w rozwój genetyki, biologii molekularnej i medycyny niż sekwencjonowanie DNA. Przez wiele lat złoty standard w tym zakresie stanowiła metoda oparta na enzymatycznej syntezie DNA opracowana przez Frederica Sangera, a jej modyfikacje stosowane są do dziś. Pod koniec XX wieku nastąpił dynamiczny rozwój technologii sekwencjonowania nowej generacji (NGS), które zakończyły erę analizy pojedynczych genów i zapoczątkowały erę sekwencjonowania genomów. Pomimo ostrej konkurencji, jedna z technologii NGS praktycznie całkowicie zdominowała światowy rynek. W artykule przedstawiamy autorski przegląd metod sekwencjonowania DNA, począwszy od metody Sangera po wysokoprzepustowe technologie sekwencjonowania drugiej i trzeciej generacji, ze szczególnym uwzględnieniem tych, które odniosły komercyjny sukces. Prezentujemy ich krótką historię, zasady działania, możliwości techniczne, zastosowania i ograniczenia. W podsumowaniu komentujemy koszty sekwencjonowania genomu człowieka na obecnym etapie genomicznej rewolucji i nakreślamy perspektywy dalszego rozwoju genomiki.

WPROWADZENIE

Historia sekwencjonowania DNA rozpoczęła się niedługo po tym jak poznano jego strukturę [1,2] i mechanizm replikacji [3], lecz dopiero 100 lat po tym jak Friedrich Miescher po raz pierwszy zidentyfikował DNA w komórkach (1869). Wyizolował on z jąder leukocytów substancję, którą nazwał „nukleiną" (ang. *nuclein*). Kiedy okazało się, że nie jest to białko, nazwę zmieniono na „kwas nukleinowy", a po dokładniejszym poznaniu jego składu chemicznego - na „kwas deoksyrybonukleinowy" (DNA). Pierwsze próby odczytu sekwencji, czyli kolejności nukleotydów w łańcuchu DNA, pochodzą z początku lat 60. XX wieku. Prawdziwy przełom nastąpił w drugiej połowie lat 70., kiedy to brytyjski badacz, Frederic Sanger, opracował metodę sekwencjonowania DNA na drodze enzymatycznej syntezy *in vitro* [4]. Imituje ona naturalny proces replikacji DNA, w którym potomna cząsteczka DNA powstaje na zasadzie komplementarności do cząsteczki matrycowej.

Co ciekawe, DNA wcale nie był pierwszym biologicznym polimerem, którego sekwencję udało się odczytać. Sposoby sekwencjonowania RNA opracowano 10 lat wcześniej [5,6], a metody oznaczania sekwencji aminokwasów w białkach na przełomie lat 50. [7-9]. Już w roku 1958 przyznano Nagrodę Nobla w dziedzinie chemii za zasługi w badaniu struktury białek i ustalenie sekwencji aminokwasowej insuliny. Otrzymał ją właśnie Frederic Sanger. Druga Nagroda Nobla w dziedzinie chemii trafiła do rąk Frederica Sangera w roku 1980, za wkład w rozwój metod oznaczania sekwencji nukleotydów w kwasach nukleinowych. Tym razem Frederic Sanger współdzielił ją z Walterem Gilbertem, współtwórcą metody sekwencjonowania DNA przez degradację chemiczną (tzw. metody Maxama-Gilberta [10]) oraz Paulem Bergiem, prowadzącym badania biochemiczne kwasów nukleinowych z wykorzystaniem rekombinowanego DNA.

Przez wiele lat metoda Sangera stanowiła złoty standard w biologii molekularnej, a jej modyfikacje stosowane są powszechnie do dziś. Określa się je wspólnym mianem metod sekwencjonowania przez syntezę (ang. *sequencing by synthesis*, SBS). Postęp technologiczny, który dokonał się w latach 90. umożliwił odczytywanie sekwencji setek tysięcy cząsteczek DNA jednocześnie. W dobie masowego równoległego sekwencjonowania, która rozpoczęła się pod koniec XX wieku i trwa do dziś, odczytywanie kolejnych genomów stało się rutyną. Niebagatelne znaczenie miał fakt, że wraz ze spadkiem czasu potrzebnego na sekwencjonowanie genomu drastycznie zmalały jego koszty. O tym, jak ważna jest technologia sekwencjonowania DNA świadczy liczba publikacji w bazie PubMed zawierających termin „DNA sequencing", przekraczająca 550 tys. (Ryc. 1).

dr inż. Małgorzata Marcinkowska-Swojak[1],

mgr Magdalena Rakoczy[1],

dr Jan Podkowiński[1],

Jurand Handschuh[2],

dr inż. Paweł Wojciechowski[1,3],

dr hab. Luiza Handschuh, prof. ICHB PAN[1]✉

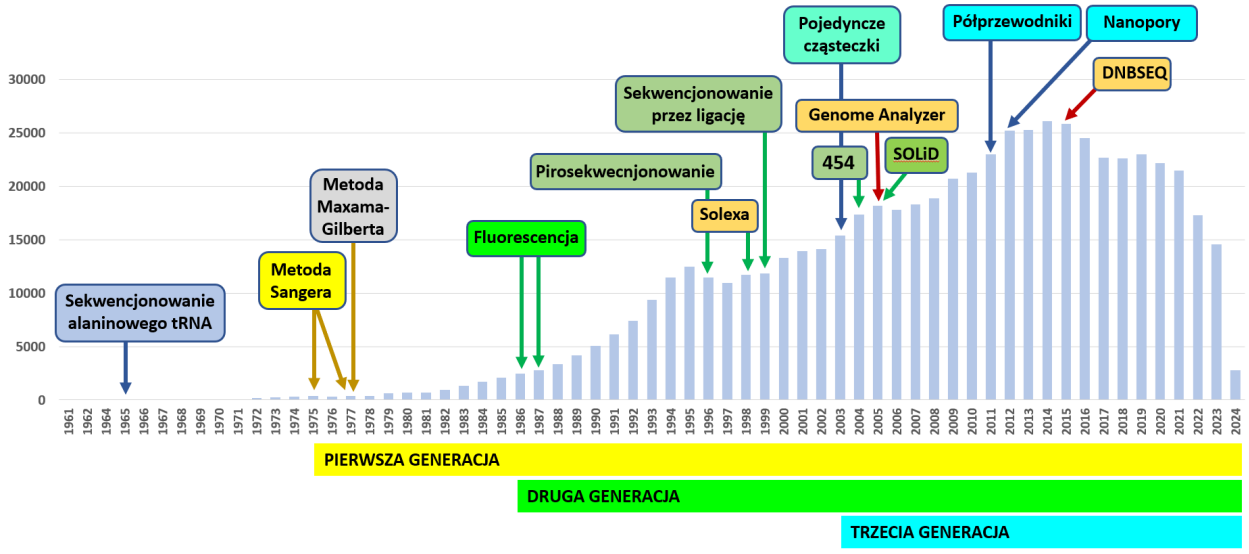
[1]Pracownia Genomiki, Instytut Chemii Bioorganicznej Polskiej Akademii Nauk, Poznań
[2]Politechnika Poznańska (student bioinformatyki)
[3]Instytut Informatyki, Politechnika Poznańska

https://doi.org/10.18388/pb.2021_534

✉autor korespondujący: luizahan@ibch.poznan.pl

Słowa kluczowe: sekwencjonowanie DNA metodą Sangera, sekwencjonowanie nowej generacji, Illumina, DNBSEQ, PacBio, Nanopore

Wykaz stosowanych skrótów: CLR – pełne długie odczyty (ang. *complete long read*); DNBSEQ – sekwencjonowanie nanokulek DNA (ang. *DNA Nanoballs Sequencing*); FC – płytka przepływowa/sekwencyjna (ang. *flow cell*); NGS – sekwencjonowanie nowej generacji (ang. *next generation sequencing*); ONT – technologia sekwencjonowania nanoporowego (ang. *Oxford Nanopore Technology*); RCR – replikacja wg modelu toczącego się koła (ang. *rolling circle replication*); SBB – sekwencjonowanie przez wiązanie (ang. *sequencing by binding*); SBS – sekwencjonowanie przez syntezę (ang. *sequencing by synthesis*); SMRT – sekwencjonowanie pojedynczej cząsteczki w czasie rzeczywistym (ang. *Single-Molecule Real-Time*); stLFR – sekwencjonowanie fragmentów długich odczytów w jednej probówce (ang. *single tube long fragment read*); ZMW – optyczne studzienki reakcyjne (ang. *Zero-Mode Waveguides*)

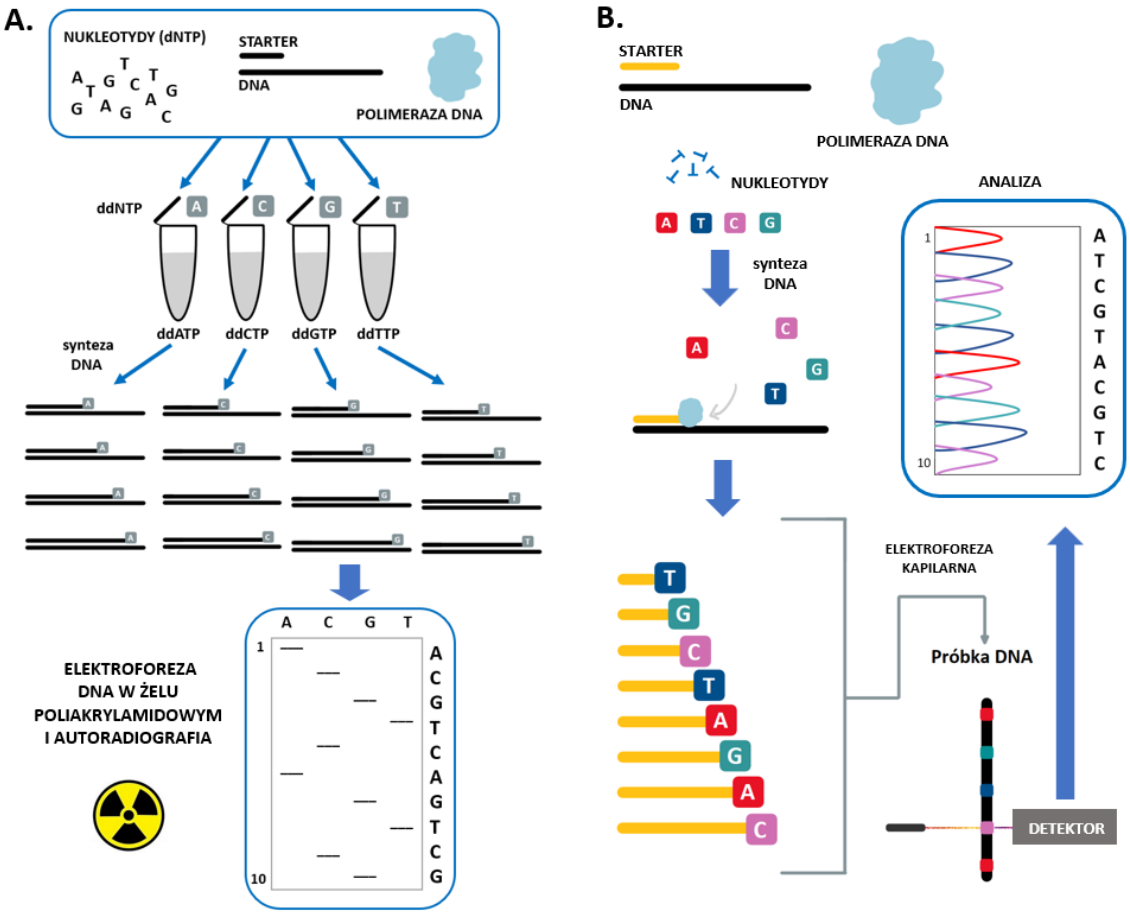


Rycina 1. Kamienie milowe w rozwoju technologii sekwencjonowania DNA, przedstawione na tle liczby publikacji zawierających wyrażenie kluczowe „DNA sequencing", zdeponowanych w bazie PubMed do dnia 02.03.2024 (łącznie 553 614 publikacji).

PIERWSZA GENERACJA: METODA SANGERA I METODA MAXAMA-GILBERTA

Frederic Sanger pierwszą wersję swojej metody sekwencjonowania DNA nazwał metodą „plus i minus" [4]. Do syntezy wykorzystał jednoniciowy DNA, krótki starter, polimerazę DNA I oraz znakowane radioizotopowo trifosforany deoksyrybonukleozydów (dNTP). Podobnie jak to się dzieje w komórkach, polimeraza wydłużała odcinek starterowy DNA dodając do niego nukleotydy i katalizując tworzenie wiązania fosfodiestrowego pomiędzy wolną grupą 3'-OH ostatniego nukleotydu w rosnącej nici DNA a 5'-fosforanem kolejnego nukleotydu. Kolejność nukleotydów była determinowana przez sekwencję matrycowej nici DNA. Obecność w mieszaninie reakcyjnej radioaktywnie znakowanego nukleotydu pozwalała na późniejszą wizualizację produktów reakcji. Ponieważ znakowany radioizotopowo nukleotyd, np. α[^{32}P]-dATP, był dodawany do reakcji w ilości znacząco mniejszej niż pozostałe trzy nieznakowane nukleotydy, szybciej ulegał wyczerpaniu. Powodowało to terminację nowo syntetyzowanych łańcuchów DNA w miejscach poprzedzających ten nukleotyd. Następnie uzyskane produkty syntezy Frederic Sanger rozdzielał w żelach poliakrylamidowych techniką jonoforezy. Dwa lata później udoskonalił metodę wprowadzając do reakcji tzw. terminatory, czyli trifosforany nukleozydów w formie dideoksy (ddNTP) [11]. Synteza odbywała się w 4 osobnych reakcjach, przy czym w każdej z nich oprócz tej samej matrycy i standardowych składników, w tym znakowanego radioizotopowo (^{32}P) dowolnego trifosforanu nukleozydu, znajdował się jeden rodzaj ddNTP (ddATP, ddCTP, ddGTP lub ddTTG). Włączenie go do powstającej cząsteczki DNA kończyło jej syntezę, gdyż ze względu na brak grupy -OH w pozycji 3' deoksyrybozy, do łańcucha nie mógł już zostać przyłączony kolejny nukleotyd. Stężenia ddNTP były znacznie niższe od stężeń dNTP, więc w każdej z czterech reakcji powstawały różnej długości produkty, kończące się określoną resztą terminatora. Rozdział znakowanych radioaktywnie produktów (każda mieszanina reakcyjna w osobnej linii) w żelach akrylamidowych o dużej rozdzielczości, a następnie autoradiografia żelu, umożliwiały odczyt sekwencji badanej cząsteczki DNA z powstałego radiogramu. Ta właśnie wersja jest znana do dziś pod nazwą „metody Sangera" (Ryc. 2A). Warto pamiętać, że do rozwoju metod analizy sekwencji DNA przyczyniły się także prace innych badaczy, np. Wu i Kaisera [12], którzy jako pierwsi zastosowali metodę wydłużania startera do określenia krótkiego, 12-nukleotydowego fragmentu sekwencji bakteriofaga λ. Z kolei Padmanabhan i Wu [13] wykazali możliwość wykorzystania radioaktywnie znakowanego startera i podobnie znakowanych nukleotydów w syntezie DNA *in vitro*.

Alternatywna wobec metody Sangera metoda Maxama-Gilberta bazowała na chemicznej degradacji DNA, znakowanego na jednym końcu przez przyłączenie radioaktywnego izotopu fosforu (na końcu 5' przy użyciu kinazy polinukleotydowej, na końcu 3' przy użyciu terminalnej transferazy) [10]. Kluczowym etapem była modyfikacja zasad azotowych z wykorzystaniem różnych zestawów odczynników chemicznych, prowadząca do oderwania zmodyfikowanej zasady od reszty cukrowej, a następnie pęknięcia łańcucha DNA w miejscu pozbawionym zasady. Równolegle prowadzone były 4 oddzielne reakcje: (1) modyfikacja G (w mniejszym stopniu również A) – metylacja za pomocą siarczanu dimetylu (zachodząca szybciej dla G niż dla A), a następnie odłączenie puryny od deoksyrybozy poprzez ogrzewanie; (2) modyfikacja A (w mniejszym stopniu również G) – metylacja za pomocą siarczanu dimetylu, a następnie rozerwanie wiązania glikozydowego (słabszego w przypadku metylowanej A niż G) pod wpływem rozcieńczonego kwasu; (3) jednoczesna modyfikacja T i C za pomocą hydrazyny (odłączenie pirymidyny od deoksyrybozy), a następnie piperydyny (rozerwanie łańcucha DNA); (4) modyfikacja wyłącznie C pod wpływem hydrazyny w obecności 2M NaCl, a następnie pęknięcie łańcucha DNA po dodaniu piperydyny. Produkty reakcji, podobnie jak w przypadku metody Sangera, rozdzielano w żelach akrylamidowych. Sekwencję określano na podstawie autoradiogramu.



Rycina 2. Schematyczne przedstawienie procesu sekwencjonowania DNA oryginalną (**A**) i zmodyfikowaną (**B**) metodą Sangera, na drodze enzymatycznej syntezy *in vitro* z wykorzystaniem radioizotopowo (**A**) lub fluorescencyjnie (**B**) znakowanych nukleotydów (opis w tekście). Sekwencja jest odczytywana na podstawie autoradiogramu (**A**) lub sygnałów fluorescencyjnych emitowanych przez znaczniki przyłączone do poszczególnych ddNTP (**B**).

Początkowo możliwe było odczytanie stosunkowo krótkiej sekwencji DNA, ok. 100 nt w przypadku metody Maxama-Gilberta [10] i ok. 100–200 nt w przypadku metody Sangera [11]. Głównym ograniczeniem była rozdzielczość żelu poliakrylamidowego. W późniejszym okresie zwiększono długość odczytywanego DNA dla obu metod, ale tylko jedna z nich, metoda Sangera, weszła na stałe do praktyki laboratoryjnej. Jej przewaga polegała na prostocie sekwencjonowania przez syntezę, stosowaniu mniej toksycznych odczynników i niższej częstości błędów. Co więcej, w metodzie Sangera wystarczył tylko jeden enzym – fragment Klenowa polimerazy DNA I z *Escherichia coli* który już wówczas był dostępny komercyjnie. W kolejnych latach rozwijano i unowocześniano metodę Sangera, wprowadzając różnego rodzaju modyfikacje, co pozwoliło m.in. na odczytywanie coraz dłuższych sekwencji DNA.

MODYFIKACJE METODY SANGERA

Jedną z ważniejszych modyfikacji metody Sangera było zastąpienie znakowania radioaktywnego znakowaniem fluorescencyjnym. Początkowo znakowano startery [14], następnie wprowadzono znakowane fluorescencyjnie terminatory, dzięki czemu można było połączyć wszystkie reakcje w jednej probówce. Odczyt wyników nie wymagał już radiografii żelu, lecz skanowania laserem, a światło emitowane przez fluorofory wykrywane było przez detektor (Ryc. 2B). Zastosowanie fluorescencji wyeliminowało niedogodności związane z używaniem znaczników radioaktywnych i pozwoliło zautomatyzować proces zbierania danych. Przyłączenie do ddNTP barwników fluorescencyjnych miało jednak wadę – spowodowało na tyle duże zmiany w strukturze nukleotydów, że nie mogły one być wydajnie wbudowywane w syntezowaną cząsteczkę DNA przez enzym Klenowa. Rozwiązanie tego problemu przyniosły odkrycia nowych polimeraz DNA z wirusów, bakterii i archeonów, w tym z mikroorganizmów ekstremofilnych, a także postęp w klonowaniu DNA i inżynierii białek. Istotne dla rozwoju metod sekwencjonowania DNA było wprowadzenie zmodyfikowanej polimerazy DNA bakteriofaga T7, tzw. sekwenazy [15].

Inną istotną modyfikacją metody Sangera było połączenie sekwencjonowania DNA z metodą PCR, czyli tzw. „sekwencjonowanie PCR-em" (ang. *PCR sequencing*, *cycling sequencing*). Metoda ta posługuje się technologią znaną z amplifikacji DNA przy pomocy PCR z tą różnicą, iż w reakcji bierze udział tylko jeden starter [16]. Matrycą może być zarówno dwuniciowy, jak i jednoniciowy DNA, a jako enzym stosuje się termostabilną polimerazę DNA włączającą do syntetyzowanego łańcucha DNA analogi nukleotydów używane jako terminatory. W odróżnieniu od standardowej reakcji PCR, nie dochodzi tu do wykładniczej amplifikacji

DNA. Materiał jest amplifikowany liniowo, a więc po *n* cyklach powstaje *n* razy więcej cząsteczek DNA niż było ich w chwili startu. Służy to wzmocnieniu sygnału fluorescencyjnego i pozwala obniżyć stężenie matrycy w reakcji. Dzięki temu, a także dzięki zastosowaniu poprzedzającej syntezę denaturacji DNA, metoda nadaje się do sekwencjonowania śladowych ilości materiału genetycznego oraz tzw. trudnych matryc, przybierających złożone struktury drugorzędowe. W ostatnich latach XX wieku metodę tę stosowano także do sekwencjonowania genomów, od prostych genomów mikroorganizmów po genom człowieka. Pierwszy genom, zsekwencjonowany w roku 1995, należał do bakterii *Haemophilus influenzae* [17]. Rok później ukończono i udostępniono sekwencję genomu pierwszego organizmu eukariotycznego, drożdży *Saccharomyces cerevisiae*. Był to efekt współpracy 633 uczonych z ponad 100 ośrodków [18,19]. Publikacje opisujące ten sukces ukazały się w kolejnym roku w specjalnym suplemencie magazynu „Nature" [20,21]. W roku 1999 poznano sekwencję pierwszego genomu roślinnego - rośliny modelowej *Arabidopsis thaliana* [22,23], a w roku 2001 międzynarodowe konsorcjum (*International Human Genome Sequencing Consortium)* opublikowało na łamach „Nature" wstępną wersję genomu człowieka [24].

Kolejną innowacją, która usprawniła sekwencjonowanie metodą Sangera było zastosowanie elektroforezy kapilarnej do separacji cząsteczek DNA. Migracja DNA jest wymuszona przez różnicę potencjału elektrycznego, podobnie jak podczas elektroforezy w żelu poliakrylamidowym znajdującym się między dwoma szklanymi płytami, ale odbywa się w żelu umieszczonym w kapilarze o submilimetrowej średnicy i długości ok. 30-50 cm (Ryc. 2). Główne zalety tej techniki wynikają z bardzo skutecznego chłodzenia kapilary, co pozwala na stosowanie wysokiego napięcia, przy jednoczesnym zapewnieniu optymalnych dla rozdziału DNA warunków termicznych (podwyższona temperatura). Metoda ta umożliwia separację cząsteczek DNA o długości do ok. 700 nt z rozdzielczością do jednego nukleotydu.

Opisane powyżej modyfikacje metody Sangera znacząco usprawniły proces sekwencjonowania DNA i przyspieszyły jego automatyzację. W 1986 r. pojawił się na rynku aparat firmy Applied Biosystems, 370A DNA Sequencig System, uważany za pierwszy automatyczny sekwenator. Pomimo wysokich cen automatyczne sekwenatory stosunkowo szybko przyjęły się na rynku i były intensywnie rozwijane. Przykładowo ABI Prism 3100 Genetic Analyzer z 1998 r., wspólny produkt firm Applied Biosystems i Hitachi, który pozwalał na równoczesną analizę 16 próbek na kapilarach z żelem krzemionkowym i mógł sekwencjonować do 348 próbek w jednym przebiegu urządzenia, kosztował ok. 300.000 USD. Ten model sekwenatora wywarł znaczący wpływ na rozwój nauk biologicznych i był na dużą skalę wykorzystywany w projekcie sekwencjonowania genomu człowieka (ang. *Human Genome Project*, https://www.genome.gov/human-genome-project).

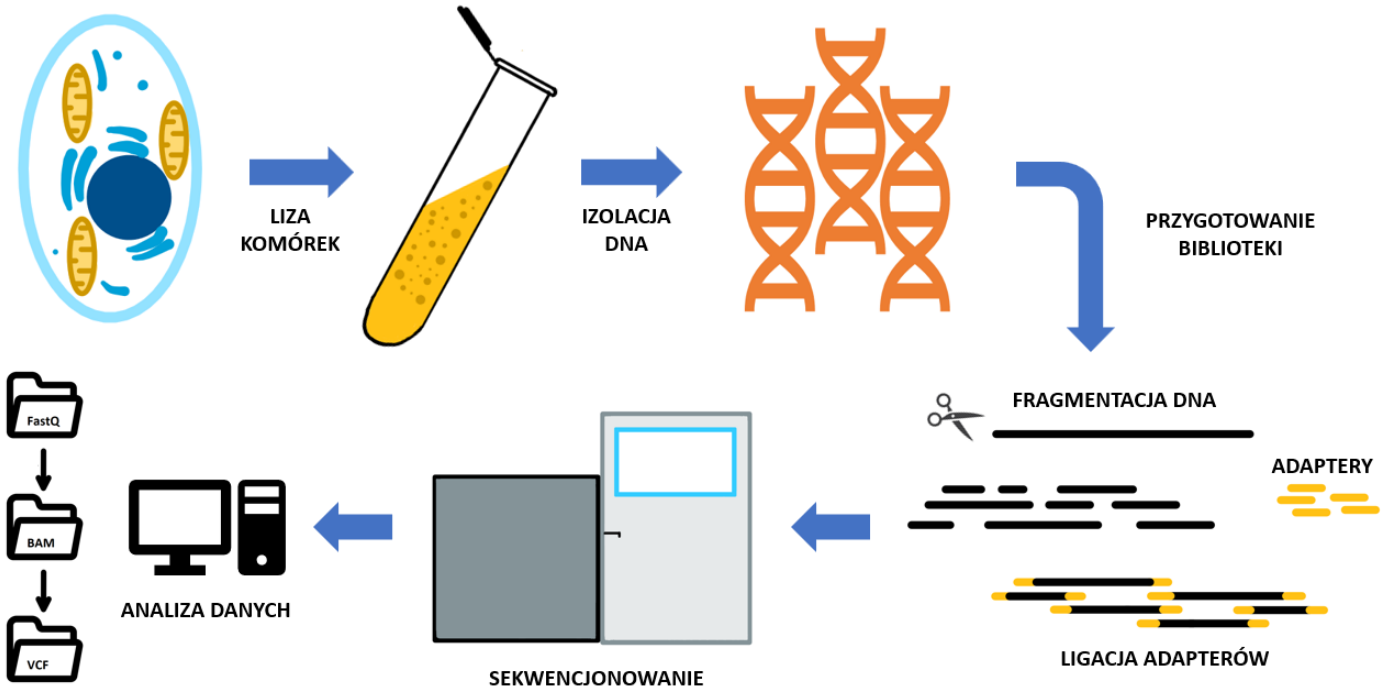
Sekwenatory bazujące na metodzie Sangera, np. aparat Spectrum Compact CE System (Promega) są wciąż obecne na rynku i powszechnie stosowane w badaniach naukowych i laboratoriach diagnostycznych. Nie zawsze jest bowiem konieczne sekwencjonowanie całego genomu. Poza tym do dobrej praktyki laboratoryjnej należy weryfikacja wyników sekwencjonowania wysokoprzepustowego za pomocą metody alternatywnej. Jest to szczególnie wskazane w przypadku identyfikacji nowych lub patogennych mutacji. Jednak główną zaletą sekwencjonowania Sangera jest bardzo niska częstość błędów, <0,01% (mniej niż 1 błąd na 10.000 nukleotydów). Metoda ta jest niezawodna do takich zastosowań jak wykrywanie mutacji i SNP (ang. *Single Nucleotide Polymorphism*, polimorfizm pojedynczego nukleotydu), analiza heterozygotyczności, sekwencjonowanie trudnych matryc, weryfikacja poprawności sekwencji konstruktów genetycznych, syntetycznych DNA czy produktów PCR.

NOWA GENERACJA: SEKWENCJONOWANIE WYSOKOPRZEPUSTOWE, CZYLI WIĘCEJ, SZYBCIEJ I TANIEJ

Sekwencjonowanie nowej generacji (ang. *next generation sequencing*, NGS), zwane też masowym sekwencjonowaniem równoległym (ang. *massive parallel sequencing*, MPS) to szeroki termin, obejmujący różnego rodzaju technologie sekwencjonowania wysokoprzepustowego, w tym zarówno te zaliczane do drugiej jak i trzeciej generacji. Ich uniwersalną cechą jest możliwość sekwencjonowania setek tysięcy cząsteczek DNA jednocześnie.

Kluczowym krokiem poprzedzającym sekwencjonowanie jest przygotowanie tzw. biblioteki (ang. *library*), czyli zestawu fragmentów DNA, które mogą być analizowane równolegle, przy pomocy automatycznego sekwenatora. Sposób przygotowania biblioteki zależy od rodzaju technologii sekwencjonowania, ale cały proces składa się z kilku podstawowych etapów, które są wspólne dla wszystkich technologii NGS (Ryc. 3). Pierwszym z nich jest fragmentacja. Cząsteczki DNA, które tworzą genomy, są bowiem zbyt długie, by mogły być sekwencjonowane w całości. Przykładowo, genom człowieka składa się z 23 par chromosomów o długości od kilkudziesięciu do kilkuset milionów nukleotydów każdy (https://www.ncbi.nlm.nih.gov/grc/human/data). Łącznie to ponad 3 miliardy nt, podczas gdy zakres długości sekwencji odczytywanych przez sekwenatory wynosi średnio od kilkuset nt do kilkudziesięciu tysięcy nt.

Następnie do obu końców fragmentów DNA przyłączane (ligowane) są dodatkowe, krótkie odcinki DNA o znanej sekwencji, tzw. adaptery. Umożliwiają one związanie biblioteki z podłożem, na którym odbywa się sekwencjonowanie. Adapter zawiera również sekwencję komplementarną do sekwencji startera (jeśli sekwencjonowanie odbywa się przez syntezę) oraz unikatową sekwencję stanowiącą znacznik, tzw. indeks (ang. *index, barcode*), inny dla każdej analizowanej próbki. Jest to konieczne wtedy, gdy sekwencjonujemy wiele próbek na raz. Ostatnim (opcjonalnym) etapem przygotowania biblioteki jest jej namnożenie (amplifikacja metodą PCR) celem zwiększenia ilości DNA służącego do sekwencjonowania. Wyjściowym materiałem do sekwencjonowania może być także RNA, ale póki co tylko jedna z dostępnych technologii wysokoprzepustowych (Oxford Nanopore, opisana poniżej) pozwala na bezpośrednią analizę cząsteczek RNA. W pozostałych przypadkach RNA musi zostać najpierw „przepisany" na DNA w proce-



Rycina 3. Ogólny schemat sekwencjonowania nowej generacji (NGS). Niezależnie od stosowanej technologii proces tworzenia biblioteki przebiega podobnie: izolowany z komórek DNA poddawany jest fragmentacji (np. mechanicznej, termicznej czy enzymatycznej), a do powstałych fragmentów przyłączane są uniwersalne adaptery umożliwiające związanie biblioteki z podłożem, np. płytką przepływową. Sekwencjonowanie odbywa się w sposób automatyczny. Dane generowane przez sekwenator są następnie przetwarzane za pomocą programów komputerowych służących do analizy wyników NGS.

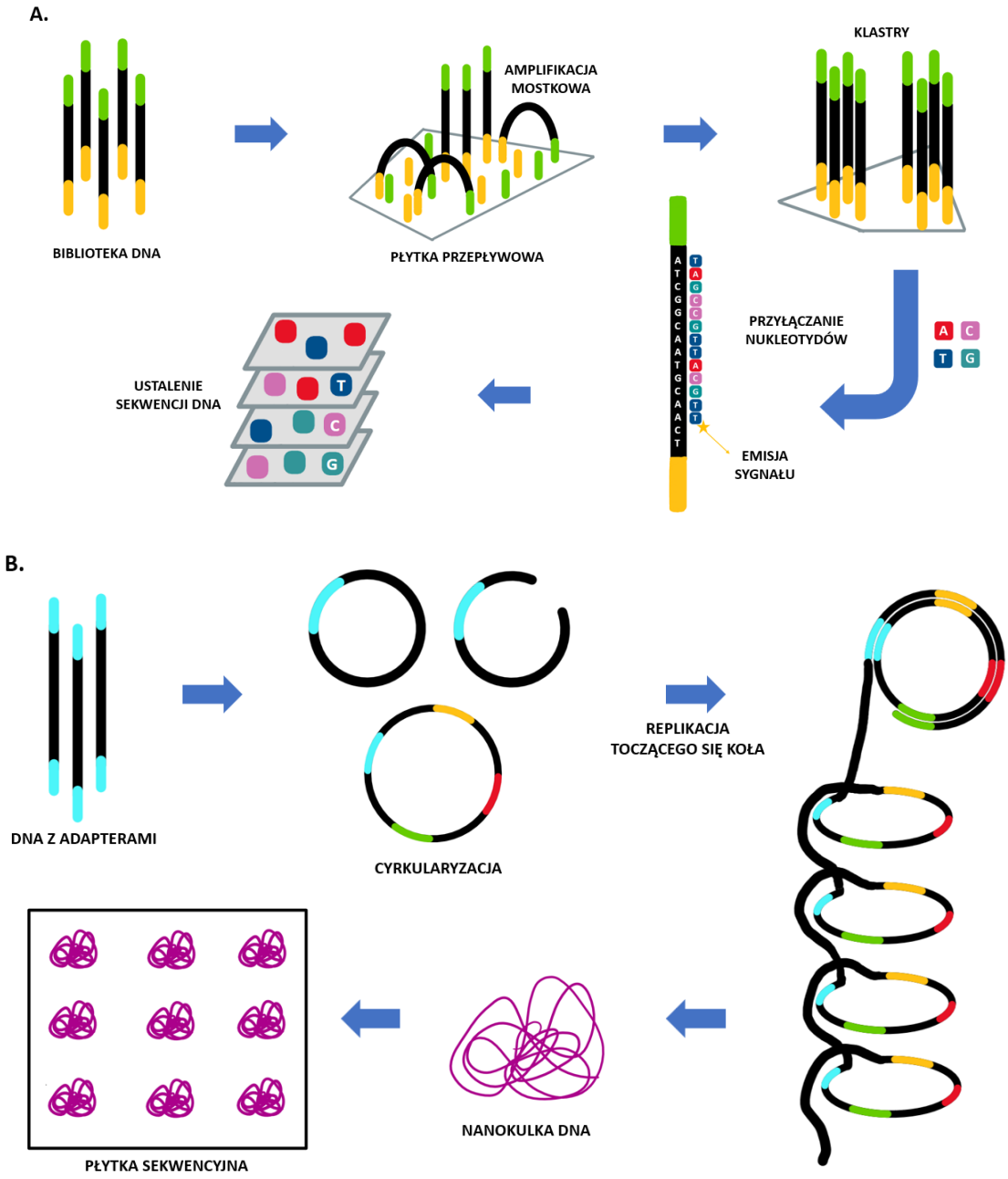
sie odwrotnej transkrypcji. Sekwencjonowanie transkryptomów (tzw. RNA-seq), obok sekwencjonowania genomów, należy do głównych zastosowań technologii NGS.

DRUGA GENERACJA: KRÓTKIE ODCZYTY, DUŻA DOKŁADNOŚĆ

Wąskim gardłem metody Sangera, nawet w jej zautomatyzowanej wersji, był rozdział elektroforetyczny cząsteczek DNA po zakończeniu etapu syntezy. Z tego względu w latach 1980-1990 wiele wysiłku włożono w poszukiwanie alternatywnych metod detekcji, które pozwalałyby na odczyt kolejnych nukleotydów już w trakcie syntezy [25]. Trzy z nich zostały wdrożone do produkcji i pojawiły się na rynku pod nazwami 454 (Life Sciences, później Roche), Solexa (później Illumina) i SOLiD (Applied Biosystems, później Life Technologies). Pierwszą z nich, opublikowaną w 2005 r., była metoda 454, zwana też pirosekwencjonowaniem. Bazowała ona na detekcji pirofosforanu (PP_i), który jest uwalniany każdorazowo po przyłączeniu nukleotydu do rosnącego łańcucha DNA [26,27]. PP_i był przekształcany w ATP przez enzym sulfurylazę. Powstały ATP napędzał kolejną reakcję enzymatyczną, katalizowaną przez lucyferazę. W wyniku tej reakcji, polegającej na konwersji lucyferyny do oksolucyferyny, generowany był sygnał świetlny rejestrowany za pomocą fotodiody. Niewłączone do łańcucha DNA dNTP degradowane były za pomocą apyrazy.

Z metodą 454 wiązano początkowo duże nadzieje. Rozgłos przyniosło jej sekwencjonowanie genomu neandertalczyka [28] i Jamesa Watsona, odkrywcy struktury DNA [29]. W tym właśnie okresie (2007 r.) firma 454 Life Sciences została nabyta przez duży koncern farmaceutyczny Roche, za niebagatelną kwotę prawie 155 milionów dolarów. Wbrew oczekiwaniom zainteresowanie tą technologią stopniowo spadało. W roku 2013 Roche ogłosił decyzję o zamknięciu 454, a trzy lata później zaprzestał wspierania tej platformy. Podobny los spotkał technologię SOLiD (ang. *Sequencing by Oligonucleotide Ligation and Detection*), która pojawiła się na rynku w roku 2006 i choć zapowiadała się obiecująco, nie odniosła sukcesu. Była to oryginalna strategia sekwencjonowania przez ligację [27,30]. W metodzie tej, do startera przyłączonego do matrycowej nici DNA, dołączane (ligowane) były na zasadzie komplementarności względem matrycy krótkie, znakowane fluorescencyjnie oligonukleotydy. Zastosowanie znacznika specyficznego dla kombinacji dwóch pierwszych nukleotydów od miejsca ligacji, wymagało skomplikowanego systemu detekcji i analizy danych. Być może właśnie to zdecydowało o niepowodzeniu tej platformy, mimo stosunkowo niskich kosztów sekwencjonowania i wysokiej precyzji odczytu, sięgającej 99,94%. Problem stanowiły też sekwencje powtarzające się i palindromowe [31]. Zarówno metoda 454 jak i SOLiD wykorzystywały do przygotowania biblioteki tzw. emulsyjny PCR, w którym amplifikacja zachodziła w roztworze, na kulkach (ang. *beads*) opłaszczonych DNA [32].

Żadna z tych dwóch metod nie wytrzymała jednak konkurencji z trzecią, która pod marką Illumina całkowicie zdominowała rynek sekwencjonowania DNA na świecie w drugiej dekadzie XXI w. Miała ona swoje początki w niewielkiej spółce Solexa, założonej w 1998 r. przez profesorów Uni-



Rycina 4. Schematyczne przedstawienie procesu sekwencjonowania drugiej generacji za pomocą technologii Illumina (**A**) i DNBSEQ (**B**). **A.** DNA (czarna linia) wchodzące w skład biblioteki Illuminy przyłączają się końcami (adapterami (żółte i zielone linie)) do komplementarnych jednoniciowych oligonukleotydów (oznaczonych również kolorem żółtym i zielonym) znajdujących się na powierzchni płytki przepływowej , przyjmując strukturę mostka. W wyniku „mostkowej" amplifikacji tworzą się klastry, czyli skupiska identycznych cząsteczek DNA. W trakcie sekwencjonowania przez syntezę rejestrowane są sygnały pochodzące od kolejnych nukleotydów przyłączanych do wydłużanej nici DNA w każdym klastrze, a specjalne oprogramowanie przekształca je następnie w sekwencje DNA. **B.** Fragmenty DNA biblioteki DNBSEQ (DNA – kolor czarny, adaptery – kolor jasnoniebieski, żółty, zielony i czerwony) łączone są ze sobą w kolistą matrycę, którą poddaje się procesowi replikacji wg modelu toczącego się koła. Powstała w efekcie długa jednoniciowa cząsteczka spontanicznie zwija się w nanokulkę DNA. Nanokulki przyłączają się do płytki sekwencyjnej, na której odbywa się sekwencjonowanie.

wersytetu w Cambridge w Wielkiej Brytanii, Davida Klenermana i Shankara Balasubramaniana, który badali przemieszczanie się polimerazy w trakcie syntezy pojedynczej cząsteczki DNA unieruchomionej na podłożu stałym [33]. W 2006 r. pojawił się pierwszy sekwenator Solexy, Genome Analyzer, który generował 1 Gb (ang. *gigabase*) danych w jednym przebiegu urządzenia [34]. W 2007 roku amerykańska firma Illumina nabyła Solexę za 600 milionów dolarów i była to z pewnością najlepsza decyzja zarządu tego przedsiębiorstwa. Roczne przychody Illuminy w roku 2021 przekroczyły 4,5 miliarda dolarów, a wartość wyemitowanych akcji spółki w roku 2023 wynosiła blisko 30 miliardów dolarów (https://beststocks.com/illumina-a-leading-player-in-the-life-science/). Szacuje się, że dziś na sekwenatorach Illuminy odczytuje się milion genomów rocznie [35], co czyni Illuminę niewątpliwym liderem w branży. Znajduje to swoje potwierdzenie także w liczbie prac naukowych – w bazie PubMed ponad 32 tys. publikacji zawiera termin „Illumina" (stan na marzec 2024).

Podobnie jak metoda Sangera, technologia Illuminy opiera się na sekwencjonowaniu przez syntezę. Sekwencjono-

wanie odbywa się na płytce przepływowej (ang. *flow cell*, FC), dawniej wykonywanej ze szkła, a obecnie z tworzywa sztucznego o dobrych parametrach optycznych. W płytce znajdują się mikroskopijne kanały, przez które przepływają reagenty chemiczne, niezbędne na kolejnych etapach procesu sekwencjonowania, takich jak hybrydyzacja, amplifikacja i odczyt sygnałów fluorescencyjnych [36]. Wewnętrzne powierzchnie kanałów FC pokryte są jednoniciowymi oligonukleotydami, komplementarnymi do adapterów wykorzystywanych w procesie przygotowania biblioteki do sekwencjonowania. Dzięki temu możliwe jest związanie biblioteki z płytką. Przyłączone jednym końcem do FC fragmenty DNA przyjmują strukturę przypominającą mostek (ang. *bridge*), gdyż ich wolny koniec łączy się z innym komplementarnym oligonukleotydem znajdującym się w sąsiadującym miejscu na płytce (Ryc. 4A). Następnie pojedyncze nici DNA są poddawane cyklicznej reakcji „mostkowej" amplifikacji (ang. *bridge amplification*). Dochodzi w niej na przemian do syntezy nowych nici DNA i ich denaturacji, podobnie jak w standardowej reakcji PCR, tyle że DNA przez cały czas pozostaje związany z podłożem. Prowadzi to do utworzenia tzw. klastrów (ang. *clusters*), czyli skupisk cząsteczek DNA o identycznej sekwencji (klonów). Każda z nich staje się następnie matrycą w reakcji SBS. Przez płytkę przepływa roztwór stanowiący mieszaninę znakowanych fluorescencyjnie nukleotydów, polimerazy i innych składników niezbędnych do syntezy komplementarnych nici DNA. W każdym cyklu reakcji sekwencjonowania przyłączany jest wyłącznie jeden nukleotyd, komplementarny do matrycy DNA w danym klastrze. Po zakończonym cyklu sygnały emitowane przez fluorofory rejestrowane są przez system detektorów i kamer. Następnie znaczniki fluorescencyjne zostają usunięte, tak, by kolejne nukleotydy mogły zostać przyłączone do wydłużanej cząsteczki DNA. Zebrane w kolejnych cyklach sygnały fluorescencyjne, uwalniane od przyłączanych nukleotydów, są przekształcane w sekwencje DNA przez odpowiednie oprogramowanie.

Pierwotnie w sekwenatorach Illuminy stosowano czterokanałowy system detekcji - każdy z czterech nukleotydów oznaczony był unikalnym barwnikiem fluorescencyjnym. Stosunkowo niedawno, wraz z premierą urządzenia HiSeq w roku 2014, Illumina wprowadziła dwukanałowy system detekcji fluorescencji [37]. Uproszczenie to pozwoliło obniżyć koszty produkcji urządzeń przy zachowaniu wysokiej dokładności odczytu. Obecnie cztery sekwenatory Illuminy dostępne na rynku wykorzystują system dwukanałowy: MiniSeq, NextSeq 500/550, NovaSeq 6000 i NovaSeq X/X Plus. W systemie tym nukleotydy znakowane są w następujący sposób: tyminy fluoroforem zielonym, cytozyny czerwonym, adeniny oboma na raz, natomiast guaniny nie są wyznakowane żadnym barwnikiem. Zastosowanie powyższej kombinacji wymaga użycia odpowiednich algorytmów bioinformatycznych do dekodowania sekwencji [38].

Inne ulepszenie technologiczne wprowadzone przez Illuminę dotyczyło sposobu organizacji i lokalizacji klastrów DNA na FC [39]. Na tradycyjnych płytkach oligonukleotydy przyłączone są w sposób losowy, nieuporządkowany (ang. *random FC*). Prowadzi to do powstawania klastrów o różnej wielkości i gęstości, a w efekcie może wpłynąć negatywnie na jakość i ilość danych z sekwencjonowania. Większe i gęsto rozmieszczone klastry mogą zlewać się ze sobą, co utrudnia ich rozróżnienie i prowadzi do błędów w odczycie sekwencji. Z kolei mniejsza gęstość klastrów ogranicza przepustowość sekwencjonowania, przez co generuje większe koszty w przeliczeniu na próbkę. Z tego względu, w roku 2014 wprowadzono do użycia FC pokryte miliardami mikroskopijnych dołków rozmieszczonych regularnie na całej powierzchni według ustalonego wzoru (ang. *patterned FC*). Komplementarne do adapterów oligonukleotydy DNA są immobilizowane wyłącznie w nanodołkach, co porządkuje proces tworzenia klastrów. Jednorodność i optymalna gęstość klastrów zapewnia z kolei uzyskanie wysokiej jakości danych z sekwencjonowania [40].

Obecnie w ofercie Illuminy znajduje się 7 platform, które różnią się przepustowością, kosztami analiz, czasem sekwencjonowania, a także zastosowaniami. Wybór konkretnego sekwenatora zależy więc zarówno od rodzaju planowanej analizy jak i budżetu projektu (Tab. 1).

Konkurencyjną dla Solexy technologię sekwencjonowania DNA na nanokulkach samoorganizujących się w nanomacierze (ang. *self-assembled DNA nanoarrays*) opracowała amerykańska firma Complete Genomics, założona w roku 2005. Cztery lata później firma miała na swoim koncie 50 zsekwencjonowanych ludzkich genomów i publikację w Science [41]. W roku 2013 Complete Genomics została przejęta przez chińską firmę BGI (Beijing Genomics Institute), a w 2018 stała się częścią spółki od niej zależnej, MGI. Oferowana przez MGI technologia sekwencjonowania DNA znana jest pod nazwą DNBSEQ (od ang. *DNA Nanoballs Sequencing*). Cały proces rozpoczyna się od fragmentacji DNA do odcinków o długości 100–350 pz i przyłączenia do nich odpowiednich adapterów na obu końcach. Następnie oligonukleotyd komplementarny do obu adapterów, tzw. *splint oligo*, hybrydyzuje z nimi, dzięki czemu dochodzi do ligacji końców i cyrkularyzacji jednoniciowego DNA. Pozostałe liniowe fragmenty usuwane są za pomocą egzonukleazy III [42]. Kilkukrotne powtórzenie procesu ligacji i cyrkularyzacji DNA sprawia, że powstaje matryca zawierająca kilka sekwencji adapterowych (Ryc. 4B). W kolejnym kroku kolistą matrycę DNA poddaje się amplifikacji w procesie replikacji wg modelu toczącego się koła (ang. *rolling circle replication*, RCR). Powstająca w wyniku replikacji długa cząsteczka spontanicznie zwija się w ciasną kulkę o średnicy ok. 300 nm. Nanokulki pochodzące z różnych matryc DNA pozostają od siebie oddzielone, gdyż z uwagi na ujemny ładunek DNA odpychają się wzajemnie [41].

W pierwszym etapie sekwencjonowania nanokulki przyłączają się do płytki przepływowej, która pokryta jest dwutlenkiem krzemu, tytanem, heksametylodisilazanem (HMDS) i odpowiednim materiałem światłoczułym. Dodatni ładunek HMDS przyciąga nanokulki ujemnie naładowanego DNA, które wiążą się z płytką w wysoce uporządkowany sposób (ang. *patterned array flow cell*). Każde włączenie nukleotydu w rosnący łańcuch DNA jest monitorowane podobnie jak w technologii Illuminy - fluorofor wzbudzany przy pomocy lasera emituje światło o określonej długości fali, rejestrowane przez wysokorozdzielczą kamerę. Następnie za pomocą odpowiedniego oprogramowania intensywność sygnału przetwarzana jest na sekwencję DNA.

Tabela 1. Porównanie podstawowych parametrów poszczególnych sekwenatorów wiodących platform sekwencjonowania drugiej generacji - Illumina i MGI (na podstawie danych prezentowanych na stronach producentów (https://www.illumina.com/systems/sequencing-platforms.html; https://en.mgi-tech.com/products, stan na dzień 04.05.2024).

Platforma	Sekwenator	Max. liczba FC na jeden cykl pracy	Czas pracy urządzenia [godz]	Max. wydajność urządzenia	Max. liczba odczytów w trakcie jednego cyklu pracy*	Min długość odczytu [nt]**	Max długość odczytu [nt]**	Max. liczba genomów***, które można odczytać w jednym cyklu pracy urządzenia
Illumina	iSeq 100	1	9,5–19	1,2 Gb	8 mln	36	150	0,01
	MiniSeq	1	5–24	7,5 Gb	50 mln	75	150	0,06
	MiSeq	1	5,5–56	15 Gb	50 mln	25	300	0,12
	NextSeq 550	1	11–29	120 Gb	800 mln	75	150	1
	NextSeq 1000	1	8-42	240 Gb	800 mln	50	300	2
	NextSeq 2000	1	8-44	540 Gb	3,6 mld	50	300	4,5
	NovaSeq 6000	2	13-44	6 Tb	40 mld	35	250	48
	NovaSeq X	2	17-48	16 Tb	104 mld	50	150	128
MGI	DNBSEQ-E25	1	5-20	7,5 Gb	25 mln	100	150	0,03
	DNBSEQ-G99	2	5-30	96 Gb	160 mln	100	300	0,2
	DNBSEQ-G50	1	9-40	150 Gb	500 mln	50	150	0,6
	DNBSEQ-G400	2	13-109	1,44 Tb	3,6 mld	50	400	6
	DNBSEQ-T7	4	16-24	7 Tb	23,2 mld	100	150	29
	DNBSEQ-T20x2	6	60-80	72 Tb	210 mld	100	150	262
	DNBSEQ-T10x4RS	8	96-106	76,8 Tb	256 mld	100	150	320

*dotyczy odczytów sparowanych (ang. *paired-end reads*), dla pojedynczych odczytów (ang. *single-end reads*) liczba ta będzie dwukrotnie mniejsza;
**długość pojedynczego odczytu, kombinacja długości odczytów jest zależna od typu FC;
***genom człowieka lub porównywalny, ze średnim pokryciem 30 x

Obecnie MGI oferuje 7 sekwenatorów o zróżnicowanych parametrach technicznych (Tab. 1). Technologia DNBSEQ zapewnia podobną przepustowość, długość odczytów, tempo zbierania danych i jakość wyników co technologia Illuminy [43]. Zastosowania obu technologii są praktycznie takie same. Niewątpliwą przewagą DNBSEQ jest jednak niższy koszt. Wejście na światowy rynek BGI/MGI miało pozytywny skutek także dla użytkowników Illuminy, bo sprawiło, że Illumina znacząco obniżyła koszty sekwencjonowania. Rywalizacja pomiędzy dwoma dużymi graczami na rynku sekwencjonowania genomowego stała się na tyle zacięta, że doszło do batalii sądowych. Firmy złożyły przeciwko sobie w różnych krajach pozwy antymonopolowe oraz oskarżenia o naruszenie patentów. Illumina wygrała pozew o naruszenie patentu przeciwko BGI w Wielkiej Brytanii [44], ale przegrała inny pozew w USA [45]. W roku 2022 doszło do ugody na rynku amerykańskim, lecz w innych krajach spór nadal trwa.

TRZECIA GENERACJA: DŁUGIE ODCZYTY, POJEDYNCZE CZĄSTECZKI

Jednym z największych przełomów w dziedzinie sekwencjonowania DNA są technologie, które zaliczamy do tzw. trzeciej generacji. Oferują one zupełnie nowe możliwości analizy genomów i transkryptomów, sięgają bowiem do poziomu długich, pojedynczych cząsteczek, odczytywanych w czasie rzeczywistym [46]. Co więcej, do sekwencjonowania trzeciej generacji nie jest wymagana amplifikacja matrycy, co pozwala uniknąć błędów wprowadzanych na etapie PCR, które zdarzają się mimo ograniczania liczby cykli i stosowania polimeraz o wysokiej precyzji. Sekwenatory trzeciej generacji potrafią odczytać sekwencje o długości od kilku - kilkunastu tysięcy do ponad stu tysięcy nukleotydów, a więc odcinki DNA zawierające całe geny. W genomie możemy badać w ten sposób nie tylko pojedyncze mutacje oraz krótkie insercje i delecje (ang. *indels*), ale też większe rearanżacje i regiony składające się z sekwencji wielokrotnie powtórzonych. W przypadku RNA daje to możliwość analizy pełnej długości transkryptów, genów fuzyjnych czy

produktów alternatywnego splicingu. Początkowo ograniczeniem technicznym platform trzeciej generacji był stosunkowo wysoki odsetek błędów, co sprawiało, że stosowano je raczej jako podejścia suplementarne względem technologii drugiej generacji. Z czasem jednak udoskonalono techniki sekwencjonowania długich odczytów do tego stopnia, że odsetek błędów znacząco zmalał [47].

Pierwszy sekwenator trzeciej generacji wprowadziła w 2008 r. firma Helicos (Cambridge, USA). Wykorzystywał on technologię tSMS (ang. *true Single Molecule Sequencing*), umożliwiającą bezpośredni odczyt sekwencji pojedynczych cząsteczek DNA, co eliminowało potrzebę amplifikacji i fragmentacji materiału genetycznego [48]. Zasada działania była podobna jak w technologii Illuminy i polegała na sekwencjonowaniu przez syntezę. Sygnał pochodzący z fluoroforu usuwanego po przyłączeniu nukleotydu do rosnącej nici DNA, wykrywany był za pomocą detektora o wysokiej czułości, co zapewniało dużą rozdzielczość odczytu. Technologia Helicos nie osiągnęła jednak komercyjnego sukcesu i w 2012 zaprzestano produkcji tego typu urządzeń.

Nukleotydy z przyłączonymi fluoroforami wykorzystuje również jedna z dwóch najpopularniejszych dziś technologii sekwencjonowania trzeciej generacji – technologia sekwencjonowania pojedynczej cząsteczki w czasie rzeczywistym (ang. *Single-Molecule Real-Time,* SMRT) [49]. Ta innowacyjna metoda opracowana przez firmę Pacific Biosciences (PacBio, Menlo Park, USA) umożliwia otrzymywanie odczytów o średniej długości kilkunastu tysięcy nt przy jednocześnie wysokiej precyzji odczytu, sięgającej nawet 99,9%. Aktualnie PacBio oferuje dwa systemy umożliwiające generowanie długich odczytów: Sequel oraz Revio. Oba oferują wysoką jakość danych z sekwencjonowania SMRT i choć różnią się przepustowością ich zastosowania są podobne (Tab. 2). Niezwykle praktyczne w tych systemach jest trwałe zintegrowanie aparatów z aplikacjami umożliwiającymi analizę danych i interpretację wyników. Dla porównania, Illumina wprowadziła takie ulepszenie dopiero w najnowszych maszynach - NextSeq 1000 i NextSeq 2000 oraz NovaSeq X Plus. Co ciekawe, od niedawna w ofercie PacBio znajduje się również sekwenator Onso generujący krótkie odczyty. Zastąpienie w nim technologii SBS technologią SBB, czyli sekwencjonowaniem przez wiązanie (ang. *sequencing by binding*) pozwoliło na uzyskiwanie danych o bardzo wysokiej jakości (Q40) [50].

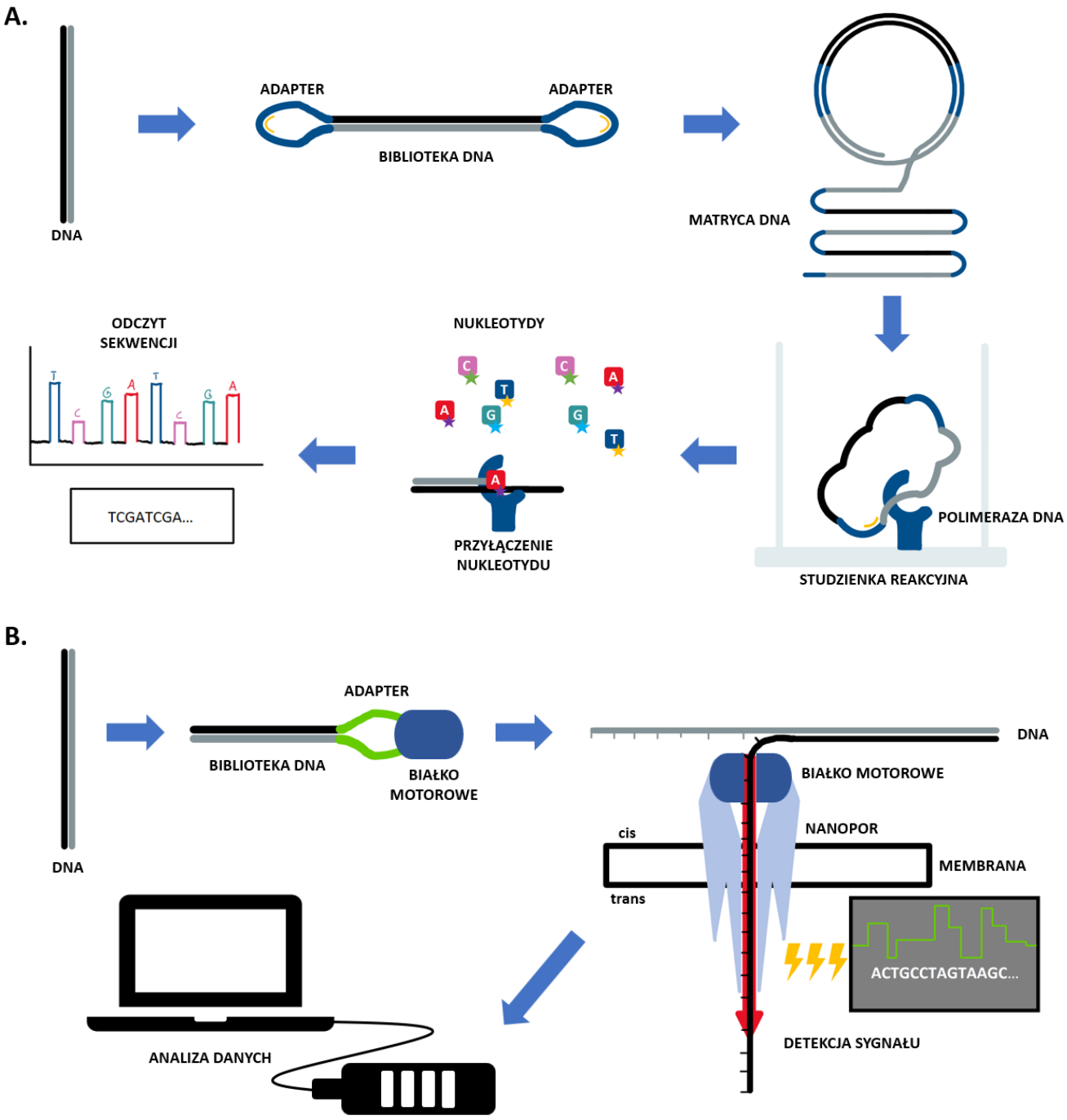
W technologii SMRT przyłączona do polimerazy DNA matryca umieszczana jest na dnie optycznych studzienek reakcyjnych, tzw. ZMW (ang. *Zero-Mode Waveguides*) [51] znajdujących się na płytce do sekwencjonowania (SMRT Cell). Na dnie studzienek znajdują się miniaturowe otwory, przez które wpada światło. Średnica otworu jest zbyt mała, aby umożliwić propagację światła w zakresie długości fal używanym do detekcji. Gdy polimeraza przyłącza kolejny nukleotyd, emitowany jest sygnał fluorescencyjny, rejestrowany w czasie rzeczywistym. Maksymalne ograniczenie pola detekcji sprawia, że sygnały fluorescencyjne pochodzące z tła nie zakłócają procesu rejestracji. Ponieważ znaczniki fluorescencyjne generują różne widma emisyjne, detektor odczytuje impulsy świetlne i identyfikuje konkretne nukleotydy. W wersji HiFi (ang. *High Fidelity*) matryca DNA dzięki odpowiednim adapterom, które przyłączają się do jej obu końców, tworzy kolistą strukturę. W efekcie polimeraza wielokrotnie powiela ten sam długi fragment, zwiększając liczbę uzyskanych z niego odczytów i minimalizując w ten sposób ryzyko błędu [52,53] (Ryc. 5A).

Wysoka precyzja długich odczytów technologii PacBio pozwala na analizę złożonych genomów oraz ich części trudnych do zsekwencjonowania za pomocą alternatywnych metod, np. regionów zawierających powtórzenia tandemowe [54]. Użyteczność technologii SMRT wykazano m.in. w badaniach genomów człowieka (*Homo sapiens*) [55], roślin, w tym ryżu (*Oryza* sp.) [56] i pszenicy (*Triticum turgidum*) [57], czy mikroorganizmów chorobotwórczych, np. wirusa zapalenia wątroby typu C (ang. *Hepatitis C Virus,* HCV) [58]. Sekwencjonowanie SMRT można również wykorzystać do mapowania modyfikacji chemicznych nukleotydów, np. metylacji. Ponieważ detekcja sygnału odbywa się w sposób bardzo precyzyjny, każda zmiana jego kinetyki, wynikająca z konkretnych modyfikacji, jest wykrywalna [59]. Innym ciekawym zastosowaniem technologii SMRT, wykraczającym poza genomikę, jest badanie procesu syntezy białek w czasie rzeczywistym [60]. Pojedyncze rybosomy, wraz z $tRNA^{fMet}$ i biotynylowanym mRNA, unieruchamia się w ZMW jako kompleksy inicjujące proces translacji. Następnie do mieszaniny reakcyjnej dodaje się fluorescencyjnie znakowane tRNA. Dzięki wyznakowaniu różnych cząsteczek tRNA odrębnymi fluoroforami możliwa jest ich detekcja na etapie związania z rybosomem w trakcie syntezy białka, a tym samym określenie sekwencji powstającego białka.

Drugą obok PacBio technologią sekwencjonowania trzeciej generacji jest tzw. sekwencjonowanie nanoporowe firmy Oxford Nanopore Technologies (ONT). Ta innowacyjna technologia opiera się na stymulowanej napięciem elektrycznym migracji pojedynczej nici DNA przez kanały (nanopory) w membranie. W czasie przejścia DNA przez nanopor mierzone są zmiany potencjału elektrycznego, wywoływane pojawieniem się danego nukleotydu w miejscu detekcji. Pozwala to na odczyt sekwencji nici DNA z dużą szybkością, czułością i precyzją [61].

Technologia nanoporowa wykorzystuje występujące w przyrodzie mechanizmy transportu cząsteczek przez błony za pośrednictwem kanałów jonowych. Syntetyczne membrany, wykonane najczęściej z polimeru, silikonu czy związków krzemu, mają mikroskopijne otwory o średnicy 1-100 nm (nanopory), które powstają w wyniku trawienia chemicznego czy bombardowania powierzchni matrycy jonami [62]. Grubość membrany jest tak dobrana by umożliwić odczyt tylko jednego nukleotydu w danym momencie, choć same nanopory mogą mieć różne właściwości.

Przygotowanie biblioteki DNA rozpoczyna się od fragmentacji DNA i selekcji fragmentów odpowiedniej długości. Następnie do jednego z końców dwuniciowego DNA ligowany jest adapter o strukturze spinki do włosów, który pozwala na przyłączenie DNA do membrany z nanoporami (Ryc. 5B). Do DNA przyłącza się też specjalne białko motorowe, które po zetknięciu się z nanoporem rozplata podwójną helisę DNA i unieruchamia jedną z nici, zapewniając przy okazji optymalny czas przejścia DNA przez



Rycina 5. Schematyczne przedstawienie procesu sekwencjonowania trzeciej generacji za pomocą technologii PacBio (**A**) i Oxford Nanopore (**B**). **A.** W technologii PacBio do pofragmentowanego DNA na obu końcach przyłączane są adaptery o strukturze spinki do włosów (kolor niebieski), co powoduje, że matryca staje się kolista. Matryca wraz z polimerazą umieszczana jest na dnie studzienek reakcyjnych, w których przebiega reakcja sekwencjonowania. Fluorescencyjnie znakowane nukleotydy emitują sygnały w momencie wbudowania się do syntetyzowanej nici DNA. Polimeraza wielokrotnie powiela ten sam długi fragment DNA, zwiększając liczbę uzyskanych odczytów i minimalizując ryzyko wystąpienia błędu. **B.** W technologii nanoporowej do długich fragmentów DNA na jednym z końców przyłączany jest adapter o strukturze spinki do włosów (kolor zielony). Związane z adapterem białko motorowe umożliwia przyłączenie DNA do nanopora znajdującego się w membranie. Następnie białko motorowe rozplata dwuniciowy DNA umożliwiając przejście przez nanopor pojedynczej nici DNA. W czasie, gdy przechodzi ona przez nanopor, następuje zmiana potencjału elektrycznego rejestrowana przez czujniki znajdujące się na płytce do sekwencjonowania.

nanopor [63]. Gotowa biblioteka jest wprowadzana do urządzenia, w którym zachodzi odczyt sekwencji. W urządzeniu membrana z nanoporami zanurzana jest w roztworze, a dzięki gradientowi stężeń kompleksy DNA-białko są kierowane do nanopora. Przepływ jonów przez otwór jest kontrolowany przez napięcie elektryczne. DNA z uwagi na swój ujemny ładunek przesuwa się w roztworze w kierunku elektrody o ładunku dodatnim. Kiedy cząsteczka DNA przechodzi przez nanopor, jonowy przepływ jest częściowo zablokowany, co powoduje zmniejszenie natężenia prądu elektrycznego. Odczyt sekwencji DNA odbywa się na podstawie pomiarów potencjału elektrycznego, który zmienia się w zależności od tego, jaki nukleotyd aktualnie przechodzi przez nanopor [64].

Tabela 2. Porównanie parametrów sekwenatorów wiodących platform sekwencjonowania trzeciej generacji - PacBio i ONT (na podstawie danych prezentowanych na stronach producentów (https://www.pacb.com/sequencing-systems; https://nanoporetech.com/products/specifications, stan na dzień 04.05.2024).

Parametry		Max. liczba FC na jeden cykl pracy	Średnia długość odczytu [kb]	Max. czas pracy sekwencjonowania na 1 FC [godz]	Max. wydajność urządzenia
PacBio	Sequel	12	10–15	20	10 Gb
	Sequel II / IIe	8	10–20	30	500 Gb
	Revio	8	15–20	<4	15 Tb
Oxford Nanopore	MinION	1	25–50*	72	15-35 Gb
	GridION	5	25–50*	72	75-175 Gb
	PromethION 2 Solo / Integrated	2	25–50*	72	555 Gb
	PromethION 24 / 48	24/48	25–50*	72	6,6 Tb / 13,3 Tb

*średnia długość przy opcji „ultra-long" (maksymalna >4Mb)

Sekwenatory ONT, w przeciwieństwie do aparatów np. Illuminy czy PacBio, to niewielkie i poręczne urządzenia, w najmniejszej wersji przypominające duży pendrive, który można podłączyć do laptopa. Nie bez przyczyny więc technologię tę określa się mianem „kieszonkowej". Sekwenator ONT zawiera chipy z czujnikami mierzącymi potencjał elektryczny, elektrodami i jedną lub wiele membran z nanoporami. Odczyty generowane w tych aparatach mogą osiągać długość nawet kilku milionów nt. Aktualnie w ofercie Oxford Nanopore dostępne są trzy podstawowe platformy: MinION, GridION i PromethION, które różnią się między sobą wielkością, przepustowością i ilością generowanych danych (Tab. 2).

Sekwencjonowanie nanoporowe to bardzo wszechstronna technologia, która znajduje zastosowanie w wielu dziedzinach nauki i medycyny. Umożliwia szybkie i coraz bardziej dokładne sekwencjonowanie całych genomów [61,65], monitorowanie składu mikroorganizmów w środowisku (ze względu na małe rozmiary urządzenia MinION badania takie można przeprowadzić w terenie) [66], analizę zmian epigenetycznych, takich jak metylacja DNA i inne modyfikacje chromatyny [67]. Co jednak najważniejsze, sekwencjonowanie nanoporowe jest jedyną technologią, która pozwala na bezpośrednie sekwencjonowanie cząsteczek RNA, podczas gdy inne metody wymagają przepisania RNA na cDNA. Badanie RNA bez konieczności stosowania odwrotnej transkrypcji, oferowane przez ONT, otwiera zupełnie nowy rozdział w obszarze transkryptomiki [61,68]. Przykładem są badania wirusa SARS-CoV-2, w których osiągnięto najdłuższy (~26 kb) ciągły odczyt sekwencji pochodzącej z genomu tego koronawirusa oraz wykryto nowe subgenomy i miejsca metylacji wirusowego RNA [69]. Opracowano także wysokiej rozdzielczości mapy transkryptomu i epitranskryptomu wirusa SARS-CoV-2 [70].

Oprócz wyżej wymienionych technologii NGS, w roku 2010 pojawiła się jeszcze jedna, znana jako Ion Torrent, czyli „potok jonów" [71], która może być zaliczana zarówno do technologii sekwencjonowania drugiej, jak i trzeciej generacji. Mimo, iż sekwenatory produkcji Ion Torrent Systems Inc. z serii Ion GeneStudio S5 nadal widnieją w ofercie Thermo Fisher Scientific, obecnie technologia ta nie jest zbyt szeroko stosowana. Warto jednak o niej wspomnieć ze względu na jej innowacyjność. Jak większość metod NGS opiera się na sekwencjonowaniu przez syntezę, ale jest jedyną technologią, która pozwala wykrywać naturalne, niezmodyfikowane nukleotydy. Nie wymaga też stosowania optyki. Sekwencjonowanie odbywa się na płytce półprzewodnikowej z mikrostudzienkami, z których każda zawiera polimerazę, klonalne kopie jednej cząsteczki DNA oraz jonoczuły detektor (ang. *ion-sensitive field-effect transistor*, ISFET). Płytka jest kolejno przemywana roztworami zawierającymi jeden typ nukleotydu. Włączenie nukleotydu do rosnącej nici DNA uwalnia PPi oraz proton wodoru, który powoduje lokalną zmianę pH, wykrywaną przez detektor. Gdy przyłącza się kilka nukleotydów, sygnał jest odpowiednio większy. Serie impulsów elektrycznych w czasie rzeczywistym przesyłane są do komputera, który przekształca je w zapis sekwencji DNA. Użycie elektronicznych chipów i rezygnacja z fluorescencji oraz złożonych systemów optycznych pozwoliło znacząco obniżyć koszty sekwencjonowania [72]. Ograniczeniem technologii Ion Torrent są za to problemy z interpretacją sygnałów przy dłuższych sekwencjach składających się z powtórzeń tego samego nukleotydu (homopolimerach) oraz długość odczytu, zbliżona do tej oferowanej przez technologie drugiej generacji (200–600 nt).

ROZWIĄZANIA POŚREDNIE: KRÓTKIE ODCZYTY Z DŁUGICH SEKWENCJI

Mimo szeregu zalet technologii sekwencjonowania drugiej generacji, ich największym ograniczeniem jest długość odczytu, która wynosi zwykle 150 nt, a w najlepszym przypadku sięga 300 nt (Tabela 1). Składanie pełnego genomu z tak krótkich odcinków jest bardzo trudne lub wręcz niemożliwe. Jednym z rozwiązań tego problemu było opracowanie metod przygotowania biblioteki w taki sposób, by w pierwszym etapie znakować długie cząsteczki genomowego DNA przez dodanie do nich specjalnego znacznika – fragmentu DNA zawierającego unikatową sekwencję.

Pierwsza tego typu strategia Illuminy nosiła nazwę „*mate pair sequencing*" [73] i polegała na generowaniu odcinków DNA o długości kilku tys. nt, dodaniu na ich końcach nukleotydów znakowanych biotyną i łączeniu w kolistą cząsteczkę. Kolisty DNA poddawano fragmentacji, a region zawierający biotynę wyłapywano za pomocą streptawidyny i poddawano tradycyjnemu sekwencjonowaniu za pomocą krótkich odczytów. Przy składaniu odczytów brano pod uwagę dodatkową informację o długości fragmentu DNA dzielącego oba odczyty. Podobne strategie oferowały w pewnym momencie także platformy 454 i SOLiD [74].

Aktualnie Illumina testuje nowe rozwiązania, łącząc innowacyjną technologię długiego odczytu o nazwie CLR (ang. *Complete Long Read*) z bardzo dokładnymi krótkimi odczytami. Technologia ta pozwala zsekwencjonować przy pomocy krótkich odczytów cząsteczki o długości 5-7 tys., a nawet 10 tys. nt, zużywając do tego mniej materiału wejściowego niż wymaga tego np. PacBio [75].

Z kolei wprowadzona przez BGI metoda stLFR (ang. *single tube long fragment read*) wykorzystuje do przygotowania biblioteki kulki z dołączonymi do nich znacznikami. W jednej reakcji znajduje się 10-50 milionów kulek, z których każda posiada inny znacznik. Następnie do kulek przyłączane są długie cząsteczki DNA, do których wcześniej, za pomocą enzymu transpozazy, wklejono co 200-1000 pz uniwersalne sekwencje umożliwiające hybrydyzację z kulką i ligację z sekwencją znacznika. Cząsteczki DNA zawierające znacznik podlegają w kolejnym etapie amplifikacji i cyrkularyzacji, po czym tworzą nanokulki, które są sekwencjonowane przy użyciu technologii DNBSEQ.

ANALIZA DANYCH Z SEKWENCJONOWANIA DNA

Analiza wyników sekwencjonowania DNA oryginalną metodą Sangera czy metodą Maxama-Gilberta była stosunkowo prosta – polegała na odczytaniu kolejności nukleotydów z obrazu prążków utrwalonych na kliszy rentgenowskiej (Ryc. 2A). Znacznie więcej umiejętności technicznych i czasu pracy wymagało przeprowadzenie samego eksperymentu. Dopiero automatyzacja metody Sangera sprawiła, że pojawiła się potrzeba stworzenia oprogramowania do analizy wyników. Dziś pliki z wynikami generowane przez sekwenatory kapilarne analizuje się za pomocą takich programów jak np. Sequence Scanner (Thermo Fisher Scientific), Finch TV (Digital World Biology), Mutation Surveyor (Softgenetics), czy Geneious (Biomatters). Część z nich jest dostępna bez opłat, niektóre zapewniają producenci aparatów, inne wymagają zakupu osobnej licencji. Użytkownik zaznajomiony z techniką sekwencjonowania kapilarnego bez najmniejszych problemów poradzi sobie z analizą wykresu przedstawiającego sygnały fluorescencyjne (Ryc. 2B).

Zupełnie inaczej wygląda bioinformatyczna analiza danych pochodzących z sekwencjonowania DNA drugiej i trzeciej generacji. Jest to bardzo obszerny temat, którego omówienie wykracza poza ramy niniejszego artykułu. Poniżej jedynie krótko przedstawiamy podstawowe etapy analizy, które są wspólne dla różnych technologii NGS oraz wyzwania związane z przetwarzaniem dużych zbiorów danych.

Pierwszym etapem jest tzw. *basecalling*, czyli przetworzenie sygnału odczytanego z sekwenatora i zapisanie go w postaci sekwencji nukleotydów. Proces ten jest zwykle przeprowadzany automatycznie z wykorzystaniem układów obliczeniowych zainstalowanych w samym sekwenatorze (obecność takiego układu oraz jego moc obliczeniowa zależy od modelu sekwenatora). Często oprogramowanie służące do basecallingu wykorzystuje procesory kart graficznych (ang. *Graphics Processing Unit*, GPU).

Kolejnym etapem jest tzw. *demultiplexing*, który polega na przydzieleniu odczytanych sekwencji do poszczególnych próbek, na podstawie sekwencji indeksu (ang. *barcode*). Etap ten jest konieczny tylko w przypadku sekwencjonowania wielu próbek jednocześnie.

W większości przypadków następnym etapem jest analiza jakości i filtrowanie danych. Ma to na celu usunięcie odczytów o niskiej jakości lub niewystarczającej długości, co jest często stosowane w przypadku sekwencjonowania trzeciej generacji. Kolejne etapy analiz bioinformatycznych zależą ściśle od rodzaju eksperymentu i celu badawczego. W przypadku sekwencjonowania DNA organizmu, którego genom został już wcześniej poznany, stosuje się tzw. mapowanie do genomu referencyjnego. W oparciu o referencyjne bazy danych przeprowadza się również adnotację (ang. *annotation*). Dzięki temu można określić położenie w genomie i sekwencję konkretnych genów, zidentyfikować znajdujące się w nich mutacje i ocenić ich potencjalne znaczenie biologiczne.

Rozmiar danych pochodzących z sekwenatorów NGS w zasadzie wyklucza możliwość ich analizy na komputerach domowych. Dla przykładu, rozmiar pliku z danymi z sekwencjonowania pełnego genomu człowieka ze średnim pokryciem 30 x w zależności od technologii wynosi:

- dla Illuminy ok. 100 GB – pliki fastq (zawierające sekwencje odczytów i dane na temat ich jakości) dla odczytów sparowanych o długości 150 nt (2 x 150 pz);
- dla PacBio HiFi – pliki z matrycowym DNA, bez sekwencji adapterów (tzw. *subreads*) w binarnym formacie bam zajmują ok. 2,9 TB, natomiast plik bam z sekwencją uzgodnionych odczytów HiFi (tzw. ccs) ~86 GB. Należy pamiętać, że pliki z matrycowym DNA zawierają dodatkowe informacje umożliwiające badanie metylacji genomu;
- dla ONT – pliki w formacie fast5 (które umożliwiają badanie metylacji genomu) zajmują około 1 TB, podczas gdy plik fastq po basecallingu ~86 GB. W roku 2023 firma Oxford Nanopore wprowadziła nowy format plików pod5, który wymaga ok. 30-50% mniej miejsca niż pliki w formacie fast5.

Na powyższym przykładzie widać, że każda z technologii wykorzystuje inne formaty plików, co więcej, w zależności od etapu ich przetwarzania formaty te są różne. Rozmiary plików dla danych z sekwencjonowania mogą się różnić w zależności od stopnia ich kompresji. Natomiast nie należy oczekiwać, że będą one znacząco mniejsze przy wyższym stopniu kompresji. Warto również wspomnieć, że wciąż trwają prace nad wydajniejszymi algorytmami kompresji danych genomicznych i samymi formatami plików. Pomimo osiągnięć na tym polu [76], problematyczna staje się dalsza analiza skompresowanych danych. Wyspecjalizowane algorytmy stosują bowiem formaty plików nieobsługiwane przez praktycznie żadne istniejące potoki przetwarzania danych, a przy znaczących rozmiarach plików konwersja

między formatami jest procesem czasochłonnym i wymagającym znacznych zasobów obliczeniowych.

Kolejnym problemem pośrednio powiązanym z rozmiarami danych pochodzących z sekwencjonowania jest czas obliczeń. Dla przykładu, wykrywanie germinalnych wariantów genetycznych dla genomu człowieka zsekwencjonowanego z 30-krotnym pokryciem trwa ok. 30 godzin przy założeniu wykorzystania pakietu GATK4 [77]. Czas ten podano jednak dla serwera obliczeniowego z 64 procesorami firmy AMD wyposażonego w 512 GB pamięci RAM (przy czym z naszego doświadczenia wynika, że pełne wykorzystanie wszystkich rdzeni tych procesorów przez cały czas trwania obliczeń jest mało prawdopodobne). Wniosek z powyższych obserwacji jest taki, że analiza danych z sekwencjonowania wymaga znaczącej infrastruktury obliczeniowej oraz przestrzeni dyskowych i jest praktycznie niemożliwa do przeprowadzenia na standardowych komputerach stacjonarnych czy laptopach. Istnieją natomiast rozwiązania sprzętowe, które znacząco skracają czas wykonywania pewnych analiz. Na przykład, dla wspomnianego wykrywania wariantów genetycznych dostępne są na rynku trzy wiodące rozwiązania: DRAGEN firmy Illumina [78] i seria rozwiązań BOLT firmy MGI [79] (dokładniej MegaBOLT, ZBOLT i ZBOLT Pro) korzystające z procesorów FCPGA oraz NVIDIA Clara for Genomics wykorzystujących procesory kart graficznych GPU. Każde z tych rozwiązań umożliwia skrócenie czasu analiz dla pojedynczej próbki z 30 godzin do 20-30 minut.

Wybór właściwych narzędzi do analizy danych z sekwencjonowania nie jest prosty. Każda z platform sekwencjonowania generuje dane o specyficznych właściwościach (takich jak np. długość odczytów, ich jakość oraz charakterystyczne błędy sekwencjonowania), a co za tym idzie mają one odmienne potoki przetwarzania danych. Decyzja odnośnie wyboru platformy sekwencjonowania wynika z obranego celu badawczego. Opracowano bazę danych narzędzi służących analizie danych dla sekwencjonowania trzeciej generacji, która obecnie zawiera ponad 870 różnych metod [80]. Co więcej, często wykorzystuje się równolegle kilka technologii sekwencjonowania do osiągnięcia jednego celu. Jako przykład może posłużyć tutaj asemblacja (składanie) *de novo* kompletnego genomu człowieka [81], w której wykorzystano zarówno krótkie odczyty Illuminy, długie odczyty ultra-long ONT i PacBio HiFi oraz dodatkowo inne technologie uzupełniające. Przetwarzanie danych z sekwencjonowania jest dziedziną, która bardzo szybko się rozwija, a dodatkowo niektóre sekwenatory są obecne na rynku zaledwie od kilku lat. Wiąże się to często z brakiem standardów analizy danych, a w związku z tym badacze zajmujący się analizą danych muszą na bieżąco śledzić najnowsze prace naukowe w tej dziedzinie.

PODSUMOWANIE

Kiedy po raz pierwszy sekwencjonowano genom człowieka, zajęło to 13 lat i pochłonęło blisko 3 miliardy dolarów [82]. Dziś na najbardziej wysokoprzepustowych sekwenatorach równoległe sekwencjonowanie ponad 100 genomów człowieka zajmuje 2 doby. Jeszcze w trakcie trwania Projektu Poznania Genomu Człowieka (ang. *Human Genome Project*, HGP) rozpoczął się dynamiczny rozwój technik sekwencjonowania nowej generacji, które umożliwiły zwiększenie przepustowości i zmniejszenie kosztów sekwencjonowania przy jednoczesnym zachowaniu dokładności uzyskiwanych odczytów [83]. Mimo, iż nie wszystkie technologie osiągnęły komercyjny sukces, zaproponowano wiele ciekawych, innowacyjnych rozwiązań.

Druga generacja technologii NGS weszła do powszechnego użytku głównie za sprawą amerykańskiej firmy Illumina. Monopol na światowym rynku sekwencjonowania DNA przełamało dopiero pojawienie się chińskiego koncernu BGI. Już w roku 2018 BGI oferował sekwencjonowanie pełnego genomu człowieka za cenę 600 USD podczas gdy to samo na platformie Illuminy wymagało wówczas 1000 USD. W marcu 2023 MGI przedstawiło swój nowy aparat, DNBSEQ-T20×2, na którym genom człowieka można zsekwencjonować za rekordowo niską cenę 100 USD [83]. Najnowszy sekwenator Illuminy, NovaSeq X/X Plus, pozwala zredukować koszty do ok. 200 USD. Czy to już granice, których nie da się przekroczyć? Czy tak samo stanie się w przypadku technologii trzeciej generacji, które póki co są nieporównywalnie droższe? Kiedy każdy człowiek na świecie będzie miał zsekwencjonowany genom? Na te pytania nie ma jednoznacznych odpowiedzi. Jedno jest pewne – jesteśmy świadkami prawdziwej rewolucji genomicznej. W wielu krajach prowadzone są duże projekty sekwencjonowania genomów, których celem jest poznanie zmienności genetycznej mieszkańców, wykrywanie rzadkich chorób genetycznych czy np. określenie podatności na choroby nowotworowe. W Wielkiej Brytanii, USA i Chinach istnieją już bazy, w których znajduje się po kilkaset tysięcy genomów. W Polsce pierwszym tego typu projektem jest ECBiG (Europejskie Centrum Bioinformatyki i Genomiki), w ramach którego powstała Genomiczna Mapa Polski (https://www.genompolski.pl/). Do jej sporządzenia wykorzystano sekwencje ok. 6 tys. genomów.

Rozwój technologiczny sprawił, że samo sekwencjonowanie przestało być wąskim gardłem – teraz jest nim analiza i archiwizacja danych, których z każdym rokiem przybywa w oszałamiającym tempie. Mamy do czynienia wręcz z eksplozją danych NGS. A w natłoku informacji ciągle daleko nam do tego, by zrozumieć jak działa genom. Nawet jeśli znamy pełną sekwencję DNA, funkcja wielu regionów genomu nadal pozostaje dla nas tajemnicą. Dlatego właśnie w tym kierunku powinny zmierzać dalsze badania genomiczne.

PIŚMIENNICTWO

1. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171: 737-738
2. Franklin RE, Gosling RG (1953) Molecular configuration in sodium thymonucleate. Nature 171: 740-741
3. Meselson M, Stahl FW (1958) The Replication of DNA in Escherichia Coli. Proceedings of the National Academy of Sciences of the United States of America 44: 671-682
4. Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of molecular biology 94: 441-448
5. Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A (1965) Structure of a Ribonucleic Acid. Science 147: 1462

6. Brownlee GG, Sanger F, Barrell BG (1967) Nucleotide Sequence of 5s--Ribosomal Rna from Escherichia Coli. Nature 215: 735
7. Edman P, Agren G (1947) The Amino Acid Composition of Secretin. Arch Biochem 13: 283-286
8. Sanger F, Tuppy H (1951) The Amino-Acid Sequence in the Phenylalanyl Chain of Insulin .2. The Investigation of Peptides from Enzymic Hydrolysates. Biochem J 49: 481-490
9. Sanger F, Thompson EOP (1953) The Amino-Acid Sequence in the Glycyl Chain of Insulin .2. The Investigation of Peptides from Enzymic Hydrolysates. Biochem J 53: 366-374
10. Maxam AM, Gilbert W (1977) New Method for Sequencing DNA. P Natl Acad Sci USA 74: 560-564
11. Sanger F, Nicklen S, Coulson AR (1977) DNA Sequencing with Chain--Terminating Inhibitors. P Natl Acad Sci USA 74: 5463-5467
12. Wu R, Kaiser AD (1968) Structure and Base Sequence in Cohesive Ends of Bacteriophage Lambda DNA. J Mol Biol 35: 523
13. Padmanabhan R, Wu R (1972) Nucleotide Sequence Analysis of DNA .9. Use of Oligonucleotides of Defined Sequence as Primers in DNA Sequence Analysis. Biochem Bioph Res Co 48: 1295
14. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH, Hood LE (1986) Fluorescence Detection in Automated DNA-Sequence Analysis. Nature 321: 674-679
15. Zhu B (2014) Bacteriophage T7 DNA polymerase – sequenase. Frontiers in Microbiology: Evolutionary and Genomic Microbiology 5: 1-5
16. Kretz K, Callen W, Hedden V (1994) Cycle Sequencing. Pcr Meth Appl 3: S107-S112
17. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269: 496-512
18. Yeast Genome *https://yeastgenome.org*
19. Yeast Genome Downloads *http://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/S288C_reference_genome_R1-1-1_19960731.tgz*
20. Goffeau A (1997) The yeast genome directory. Nature 387: 5
21. Mewes HW, Albermann K, Bahr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F, Zollner A (1997) Overview of the yeast genome. Nature 387: 7-8
22. Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin XY, Town CD, Venter JC, Fraser CM, Tabata S, Nakamura Y, Kaneko T, Sato S, Asamizu E, Kato T, Kotani H, Sasamoto S, Ecker JR, Theologis A, Federspiel NA, Palm CJ, Osborne BI, Shinn P, Conway AB, Vysotskaia VS, Dewar K, Conn L, Lenz CA, Kim CJ, Hansen NF, Liu SX, Buehler E, Altafi H, Sakano H, Dunn P, Lam B, Pham PK, Chao Q, Nguyen M, Yu GX, Chen HM, Southwick A, Lee JM, Miranda M, Toriumi MJ, Davis RW, Wambutt R, Murphy G, Dusterhoft A, Stiekema W, Pohl T, Entian KD, Terryn N, Volckaert G, Salanoubat M, Choisne N, Rieger M, Ansorge W, Unseld M, Fartmann B, Valle G, Artiguenave F, Weissenbach J, Quetier F, Wilson RK, de la Bastide M, Sekhon M, Huang E, Spiegel L, Gnoj L, Pepin K, Murray J, Johnson D, Habermann K, Dedhia N, Parnell L, Preston R, Hillier L, Chen E, Marra M, Martienssen R, McCombie WR, Mayer K, White O, Bevan M, Lemcke K, Creasy TH, Bielke C, Haas B, Haase D, Maiti R, Rudd S, Peterson J, Schoof H, Frishman D, Morgenstern B, Zaccaria P, Ermolaeva M, Pertea M, Quackenbush J, Volfovsky N, Wu DY, Lowe TM, Salzberg SL, Mewes HW, Rounsley S, Bush D, Subramaniam S, Levin I, Norris S, Schmidt R, Acarkan A, Bancroft I, Quetier F, Brennicke A, Eisen JA, Bureau T, Legault BA, Le QH, Agrawal N, Yu Z, Martienssen R, Copenhaver GP, Luo S, Pikaard CS, Preuss D, Paulsen IT, Sussman M, Britt AB, Selinger DA, Pandey R, Mount DW, Chandler VL, Jorgensen RA, Pikaard C, Juergens G, Meyerowitz EM, Theologis A, Dangl J, Jones JDG, Chen M, Chory J, Somerville MC, In AG (2000) Analysis of the genome sequence of the flowering plant. Nature 408: 796-815
23. Bevan M, Walsh S (2005) The Arabidopsis genome: a foundation for plant research. Genome Res 15: 1632-1642
24. Lander ES, Consortium IHGS et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921
25. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH (2017) DNA sequencing at 40: past, present and future. Nature 550: 345-353
26. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242: 84-89
27. Orlowska M, Sobczyk M (2017) Metody sekwencjonowania nowej generacji oraz ich wykorzystanie w genetyce, hodowli i biotechnologii roślin. Aparatura badawcza i dydaktyczna 1: 54-61
28. Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Pääbo S (2006) Analysis of one million base pairs of Neanderthal DNA. Nature 444: 330-336
29. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872-875
30. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu YT, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu HN, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, De la Vega FM, Blanchard AP (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19: 1527-1541
31. Huang YF, Chen SC, Chiang YS, Chen TH, Chiu KP (2012) Palindromic sequence impedes sequencing-by-ligation mechanism. BMC Syst Biol 6: S10
32. Kotowska M, Zakrzewska-Czerwińska J (2010) Kurs szybkiego czytania DNA - nowoczesne techniki sekwencjonowania. Biotechnologia 4: 24-38
33. Bennett ST, Barnes C, Cox A, Davies L, Brown C (2005) Toward the $1000 human genome. Pharmacogenomics 6: 373-382
34. EMEA Illumina History *https://emea.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history.html*
35. Department of Chemistry University of Cambridge UK *https://www.ch.cam.ac.uk/collaboration-and-impact/solexa-sequencing*
36. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P (2014) Library construction for next--generation sequencing: Overviews and challenges. Biotechniques 56: 61
37. van Pelt-Verkuil E, van Leeuwen WB, te Witt R (2019) Molecular Diagnostics Part 1: Technical Backgrounds and Quality Aspects. Springer Nature, Singapore
38. Andrews S (2016) Illumina 2 colour chemistry can overcall high confidence G bases. QC Fail
39. Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J (2016) High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian J Microbiol 56: 394-404
40. EMEA Illumina Patterned FC *https://emea.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/patterned-flow-cells.html*
41. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, Dahl F, Fernandez A, Staker B, Pant KP, Baccash J, Borcherding AP, Brownley A, Cedeno R, Chen LS, Chernikoff D, Cheung A, Chirita R, Curson B, Ebert JC, Hacker CR, Hartlage R, Hauser B, Huang S, Jiang Y, Karpinchyk V, Koenig M, Kong C, Landers T, Le C, Liu J, McBride CE, Morenzoni M, Morey RE, Mutch K, Perazich H, Perry K, Peters BA, Peterson J, Pethiyagoda CL, Pothuraju K, Richter C, Rosenbaum AM, Roy S, Shafto J, Sharanhovich U, Shannon KW, Sheppy CG, Sun M, Thakuria JV, Tran A, Vu D, Zaranek AW, Wu XD, Drmanac S, Oliphant AR, Banyai WC, Martin B, Ballinger DG, Church GM, Reid CA (2010) Human Ge-

nome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science 327: 78-81

42. Chen Y, Wang H, Yang J, Yang HM, Zhang WW, Drmanac R, Xu CJ (2021) Reusable and sensitive exonuclease III activity detection on DNB nanoarrays based on cPAS sequencing technology. Enzyme Microb Tech 150: 109878
43. Jeon SA, Park JL, Park SJ, Kim JH, Goh SH, Han JY, Kim SY (2021) Comparison between MGI and Illumina sequencing platforms for whole genome sequencing. Genes Genom 43: 713-724
44. Illumina vs BGI bussineswire *https://www.businesswire.com/news/home/20210120005644/en*
45. Illumina vs BGI reuters *https://www.reuters.com/legal/litigation/bgi-group-units-illumina-settle-us-lawsuits-over-dna-sequencing-2022-07-14*
46. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong XX, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma CC, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323: 133-138
47. Marx V (2021) Long road to long-read assembly. Nat Methods 18: 125-129
48. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z (2008) Single-molecule DNA sequencing of a viral genome. Science 320: 106-109
49. McCarthy A (2010) Third Generation DNA Sequencing: Pacific Biosciences' Single Molecule Real Time Technology. Chem Biol 17: 675-676
50. PacBio SBB *https://www.pacb.com/technology/sequencing-by-binding*
51. Iizuka R, Yamazaki H, Uemura S (2022) Zero-mode waveguides and nanopore-based sequencing technologies accelerate single-molecule studies. Biophys Physicobiol 19: e190032
52. Rhoads A, Au KF (2015) PacBio Sequencing and Its Applications. Genom Proteom Bioinf 13: 278-289
53. Hon T, Mars K, Young G, Tsai YC, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC, Knapp SJ, Ware D, Shapiro B, Peluso P, Rank DR (2020) Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data 7: 399
54. Jiao WB, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol 36: 64-70
55. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian YF, Chin CS, Phillippy AM, Schate MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR, Hunkapiller MW (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37: 1155-1162
56. Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang CJ, Chougule K, Gao DY, Iwata A, Goicoechea JL, Wei SR, Wang J, Liao Y, Wang MH, Jacquemin J, Becker C, Kudrna D, Zhang JW, Londono CEM, Song X, Lee S, Sanchez P, Zuccolo A, Ammiraju JSS, Talag J, Danowitz A, Rivera LF, Gschwend AR, Noutsos C, Wu CC, Kao SM, Zeng JW, Wei FJ, Zhao Q, Feng Q, El Baidouri M, Carpentier MC, Lasserre E, Cooke R, Farias DD, da Maia LC, dos Santos RS, Nyberg KG, McNally KL, Mauleon R, Alexandrov N, Schmutz J, Flowers D, Fan CZ, Weigel D, Jena KK, Wicker T, Chen MS, Han B, Henry R, Hsing YIC, Kurata N, de Oliveira AC, Panaud O, Jackson SA, Machado CA, Sanderson MJ, Long MY, Ware D, Wing RA (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus. Nat Genet 50: 1618-1618
57. Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K, Jordan KW, Golan G, Deek J, Ben-Zvi B, Ben-Zvi G, Himmelbach A, MacLachlan RP, Sharpe AG, Fritz A, Ben-David R, Budak H, Fahima T, Korol A, Faris JD, Hernandez A, Mikel MA, Levy AA, Steffenson B, Maccaferri M, Tuberosa R, Cattivelli L, Faccioli P, Ceriotti A, Kashkush K, Pourkheirandish M, Komatsuda T, Eilam T, Sela H, Sharon A, Ohad N, Chamovitz DA, Mayer KFX, Stein N, Ronen G, Peleg Z, Pozniak CJ, Akhunov ED, Distelfeld A (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357: 93-96
58. Takeda H, Yamashita T, Ueda Y, Sekine A (2019) Exploring the hepatitis C virus genome using single molecule real-time sequencing. World J Gastroentero 25: 4661-4672
59. Flusberg B, Webster D, Travers K, Olivares E, Korlach J, Turner S (2010) Direct detection of DNA methylation and mutagenic damage through single-molecule, real-time (SMRT™) DNA sequencing. Cancer Res 70: 461-465
60. Uemura S, Aitken CE, Korlach J, Flusberg BA, Turner SW, Puglisi JD (2010) Real-time tRNA transit on single translating ribosomes at codon resolution. Nature 464: 1012-1073
61. Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17: 239
62. Zmienko A, Satyr A (2020) Sekwencjonowanie nanoporowe i jego zastosowanie w biologii. Postepy Biochemii 66: 193-204
63. Loose M, Malla S, Stout M (2016) Real-time selective sequencing using nanopore technology. Nat Methods 13: 751-754
64. Feng YX, Zhang YC, Ying CF, Wang DQ, Du CL (2015) Nanopore-based Fourth-generation DNA Sequencing Technology. Genom Proteom Bioinf 13: 4-16
65. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF (2021) Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol 39: 1348-1365
66. Zhu XJ, Yan SS, Yuan FH, Wan SG (2020) The Applications of Nanopore Sequencing Technology in Pathogenic Microorganism Detection. Can J Infect Dis Med 2020: 6675206
67. Laver T, Harrison J, 'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. . Biomol Detect Quantif 3: 1-8
68. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, Sadowski N, Holmes N, de Jesus JG, Jones KL, Soulette CM, Snutch TP, Loman N, Paten B, Loose M, Simpson JT, Olsen HE, Brooks AN, Akeson M, Timp W (2019) Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods 16: 1297-1305
69. Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Holzer M, Marz M (2019) Direct RNA nanopore sequencing of full--length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome Res 29: 1545-1554
70. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H (2020) The Architecture of SARS-CoV-2 Transcriptome. Cell 181: 914-921
71. Liu L, Li YH, Li SL, Hu N, He YM, Pong R, Lin DN, Lu LH, Law M (2012) Comparison of Next-Generation Sequencing Systems. J Biomed Biotechnol 2012: 251364
72. Merriman B, Rothberg JM, Team ITRD (2012) Progress in Ion Torrent semiconductor chip based sequencing. Electrophoresis 33: 3397-3417
73. EMEA Illumina Mate-pair seq *https://www.illumina.com/science/technology/next-generation-sequencing/mate-pair-sequencing.html*
74. Berglund EC, Kiialainen A, Syvänen AC (2011) Next-generation sequencing technologies and applications for human genetic history and forensics. Investig Genet 2: 23
75. EMEA Illumina Long-read seq *https://emea.illumina.com/science/technology/next-generation-sequencing/long-read-sequencing.html*
76. Deorowicz S (2020) FQSqueezer: k-mer-based compression of sequencing data. Scientific reports 10: 1-9
77. Zhao S, Agafonov O, Azab A, Stokowy T, Hovig E (2020) Accuracy and efficiency of germline variant calling pipelines for human genome data. Scientific reports 10: 20222

78. Illumina DRAGEN *https://www.illumina.com/products/by-type/informatics-products/dragen-secondary-analysis/order.html*

79. Li Z, Xie Y, Zeng W, Huang Y, Gu S, Gao Y, Huang W, Lu L, Wang X, Wu J, Yin X, Zhu R, Huang G, Lu L, Tang J, Zheng Y, Liu Q, Zhou X, Shan R, Wang B, Fang M, Jin X (2023) An efficient large-scale whole--genome sequencing analyses practice with an average daily analysis of 100Tbp: ZBOLT. Clinical and Translational Discovery 3: e252

80. Amarasinghe SL, Ritchie ME, Gouil Q (2021) long-read-tools.org: an interactive catalogue of analysis methods for long-read sequencing data. GigaScience 10: giab003

81. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S, Hoyt SJ, Diekhans M, Logsdon GA, Alonge M, Antonarakis SE, Borchers M, Bouffard GG, Brooks SY, Caldas GV, Chen NC, Cheng H, Chin CS, Chow W, de Lima LG, Dishuck PC, Durbin R, Dvorkina T, Fiddes IT, Formenti G, Fulton RS, Fungtammasan A, Garrison E, Grady PGS, Graves-Lindsay TA, Hall IM, Hansen NF, Hartley GA, Haukness M, Howe K, Hunkapiller MW, Jain C, Jain M, Jarvis ED, Kerpedjiev P, Kirsche M, Kolmogorov M, Korlach J, Kremitzki M, Li H, Maduro VV, Marschall T, McCartney AM, McDaniel J, Miller DE, Mullikin JC, Myers EW, Olson ND, Paten B, Peluso P, Pevzner PA, Porubsky D, Potapova T, Rogaev EI, Rosenfeld JA, Salzberg SL, Schneider VA, Sedlazeck FJ, Shafin K, Shew CJ, Shumate A, Sims Y, Smit AFA, Soto DC, Sovic I, Storer JM, Streets A, Sullivan BA, Thibaud-Nissen F, Torrance J, Wagner J, Walenz BP, Wenger A, Wood JMD, Xiao C, Yan SM, Young AC, Zarate S, Surti U, McCoy RC, Dennis MY, Alexandrov IA, Gerton JL, O'Neill RJ, Timp W, Zook JM, Schatz MC, Eichler EE, Miga KH, Phillippy AM (2022) The complete sequence of a human genome. Science 376: 44-53

82. Human Genome *https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project*

83. Reuter JA, Spacek DV, Snyder MP (2015) High-Throughput Sequencing Technologies. Mol Cell 58: 586-597

From Sanger to genome sequencing – an overview of DNA sequencing technologies

Małgorzata Marcinkowska-Swojak[1], Magdalena Rakoczy[1], Jan Podkowiński[1], Jurand Handschuh[2], Paweł Wojciechowski[1,3], Luiza Handschuh[1]✉

[1] Laboratory of Genomics, Institute of Bioorganic Chemistry PAS
[2] Poznań University of Technology (bioinformatics student)
[3] Poznan University of Technology, Institute of Computing Science

✉corresponding author: luizahan@ibch.poznan.pl

Keywords: Sanger method of DNA sequencing, next generation sequencing, Illumina, DNBSEQ, PacBio, Nanopore

ABSTRACT

There is no technique that would make a greater contribution to the development of genetics, molecular biology and medicine than DNA sequencing. For many years, the method based on enzymatic DNA synthesis developed by Frederic Sanger was the gold standard in this area and its modifications are still used today. At the end of the 20th century, there was a dynamic development of next-generation sequencing (NGS) technologies, which ended the era of single gene analysis and initiated the era of genome sequencing. Despite fierce competition, one NGS technology has practically completely dominated the global market. In the article, we present our own review of DNA sequencing methods, starting from the Sanger method to high-throughput second- and third-generation sequencing technologies, with particular emphasis on those that have achieved commercial success. We present their short history, principles of operation, technical possibilities, applications and limitations. In the summary, we comment the human genome sequencing costs at the current stage of the genomic revolution and outline the prospects for further development of genomics.