Sekwencjonowanie NGS krok po kroku: przewodnik biologa

Od izolacji DNA do wynikow bioinformatycznych - kompletny przewodnik po technologii nowej generacji

Marta Kowalczyk · 26.10.2026

Pierwsza probka, ktora wyslalam do sekwencjonowania NGS, wrocila z wynikiem, ktory kompletnie mnie zdezorientowal. Dane FASTQ zajmowaly kilkadziesiat gigabajtow, raport jakosci wygladal niepokojaco, a ja - swiezo po studiach, z pipeta w reku i teoria w glowie - nie wiedzialam, od czego zaczac interpretacje. Gdybym wtedy miala taki przewodnik, zaoszczedzilabym sobie kilku tygodni bladzenia po forach bioinformatycznych.

Sekwencjonowanie nowej generacji (NGS, Next-Generation Sequencing) to dzis standardowe narzedzie w diagnostyce molekularnej, badaniach genomicznych i medycynie personalizowanej. Tylko w 2025 roku globalny rynek NGS przekroczyl wartosc 12 miliardow dolarow, a liczba opublikowanych genomow w bazie NCBI wzrosla o kolejne 40%. Mimo to wiele osob wchodzacych w te technologie traktuje ja jak czarna skrzynke: wsadzasz DNA, wychodzi plik z wynikami. Rzeczywistosc jest znacznie bardziej zlozona - i znacznie bardziej interesujaca.

Ponizej znajdziesz opis kazdego etapu procesu NGS, ze wskazaniem konkretnych decyzji, ktore trzeba podjac, typowych bledow oraz roznic miedzy glownymi platformami. Zaczynamy od DNA, konczymy na interpretacji wariantow.

1. Izolacja i ocena jakosci DNA - fundament, ktory decyduje o wszystkim

Kazda analiza NGS zaczyna sie jeszcze przed laboratoriami sekwencjonowania - zaczyna sie przy probowce z materialem biologicznym. Jakosc wyjsciowego DNA determinuje jakosc koncowych wynikow w stopniu, ktorego nie da sie nadrobic zadnym algorytmem bioinformatycznym.

Do izolacji genomowego DNA z tkanek stalych najczesciej stosuje sie zestawy kolumnowe (np. DNeasy Blood & Tissue Kit firmy Qiagen) lub metody fenol-chloroformowe. Ta druga daje zwykle wyzsze stezenia, ale wymaga staranniejszego postepowania z toksycznymi odczynnikami. Z krwi pelnej standardem jest izolacja z leukocytow po lizie erytrocytow.

Po izolacji konieczna jest ocena jakosci w trzech wymiarach:

- Stezenie - mierzone fluorometrycznie (Qubit, PicoGreen), nie spektrofotometrycznie. Nanodrop zawyza wyniki, bo mierzy absorpcje przy 260 nm lacznie z RNA i wolnymi nukleotydami.
- Czystosc - stosunek A260/A280 powinien wynosic 1,8-2,0 dla DNA. Wartosci nizsze sugeruja zanieczyszczenie bialkami lub fenolami.
- Integralnosc - oceniana elektroforeza zelowa lub systemem Agilent Bioanalyzer/TapeStation. Fragment DNA powinien byc mozliwie dlugi i jednorodny.

Typowy blad na tym etapie: uzycie DNA z tkanki utrwalonej formaldehydem (FFPE).

Formalin powoduje sieciowanie bialek z DNA i fragmentacje kwasu nukleinowego. Biblioteki z materialu FFPE wymagaja specjalnych protokolow naprawy DNA i maja znacznie wyzszy odsetek artefaktow sekwencjonowania.

2. Przygotowanie biblioteki genomowej - serce calego procesu

Biblioteka genomowa (sequencing library) to zbior fragmentow DNA przygotowanych w taki sposob, by mogly byc odczytane przez sekwencer. Przygotowanie biblioteki to etap, ktory pochlania najwiecej czasu i generuje najwiecej bledow.

Fragmentacja DNA

Genomowy DNA musi zostac pociety do odpowiedniej wielkosci - zazwyczaj 150-500 bp dla Illumina, 5-30 kb dla PacBio lub Oxford Nanopore. Fragmentacje przeprowadza sie mechanicznie (sonikacja, np. Covaris M220) lub enzymatycznie (endonukleazy). Sonikacja jest bardziej powtarzalna, ale wymaga kalibracji parametrow dla kazdego rodzaju tkanki.

Naprawa koncow i ligacja adapterow

Po fragmentacji konce czasteczek DNA sa nierowne - wymagaja naprawy (end-repair) do postaci tepych koncow, a nastepnie dodania adeniny na koncu 3' (A-tailing). Dopiero wtedy liguje sie adaptery - krotkie sekwencje jednoniciowe, ktore umozliwiaja przylaczenie fragmentow do powierzchni flow cell i amplifikacje.

Adaptery zawieraja tez indeksy (barcody) - unikalne sekwencje identyfikujace probke. Dzieki nim mozna sekwencjonowac wiele probek jednoczesnie (multiplexing). Przy projektowaniu eksperymentu nalezy zadbac o odpowiednia roznorodnosc indeksow - uzycie zbyt podobnych sekwencji prowadzi do bledow demultipleksacji.

Amplifikacja PCR i jej pulapki

Wzbogacona biblioteka jest amplifikowana PCR. Tu pojawia sie jeden z najpowazniejszych problemow: bledy PCR i duplikaty. Polimeraza wprowadza bledy (ok. 1 na 10?-10? bp), a preferencyjna amplifikacja pewnych fragmentow zaburza rownomiernosc pokrycia. Dlatego w analizach wymagajacych wysokiej czulosci (np. wykrywanie mutacji somatycznych w onkologii) stosuje sie unikalne identyfikatory molekularne (UMI, Unique Molecular Identifiers), ktore pozwalaja odroznic prawdziwe warianty od artefaktow PCR.

3. Kontrola jakosci biblioteki przed sekwencjonowaniem

Przed zaladowaniem biblioteki na sekwencer nalezy potwierdzic jej jakosc. Ten etap jest czesto pomijany przez osoby pracujace pod presja czasu - i czesto za to placa zmarnowanym biegiem sekwencjonowania.

Standardowe metody kontroli jakosci biblioteki:

Optymalny rozklad wielkosci fragmentow dla biblioteki Illumina to waski pik w okolicach 300-400 bp (lacznie z adapterami ~120 bp). Szeroki lub wielomodalny rozklad swiadczy o problemach z fragmentacja lub ligacja.

4. Sekwencjonowanie - jak dzialaja poszczegolne platformy

Tu zaczyna sie magia - i tu tez pojawiaja sie fundamentalne roznice miedzy technologiami. Wybor platformy powinien wynikac z pytania biologicznego, nie z tego, co akurat jest dostepne w laboratorium.

Illumina - sekwencjonowanie przez synteze

Illumina dominuje rynek NGS z udzialem szacowanym na ponad 70% w 2026 roku. Technologia opiera sie na sequencing by synthesis (SBS): fragmenty DNA przylaczaja sie do powierzchni flow cell i tworza klastry przez amplifikacje mostkowa. Nastepnie polimeraza dodaje kolejno znakowane fluorescencyjnie nukleotydy, a kamera rejestruje sygnal po kazdym cyklu.

Zalety Illumina: bardzo wysoka dokladnosc (Q30 >85%, czyli blad ponizej 1 na 1000 bp), niski koszt na baze (~0,006 USD/Mb na NovaSeq X), ogromna przepustowosc. Wady: krotkie odczyty (2×150 bp lub 2×250 bp), trudnosci z regionami repetytywnym i strukturalnymi wariantami genomu.

Oxford Nanopore - sekwencjonowanie przez nanopory

Oxford Nanopore Technologies (ONT) oferuje zupelnie inna filozofie: czasteczka DNA przechodzi przez biologiczny nanopor w blonie, a zmiany pradu jonowego sa rejestrowane w czasie rzeczywistym i dekodowane na sekwencje. Odczyty moga miec dlugosc od kilku do ponad 2 milionow par zasad (ultra-long reads).

To rewolucja dla badan strukturalnych wariantow genomu, montazu de novo i wykrywania modyfikacji epigenetycznych (np. metylacji 5mC bezposrednio z odczytow, bez bisulfitowania). Cena urzadzenia MinION to ok. 1000 dolarow - nieosiagalna wczesniej dostepnosc sekwencjonowania. Wada jest wyzsza stopa bledow surowych odczytow (~5% dla najnowszego modelu R10.4.1), choc algorytmy bazecallingu (Dorado od ONT) stale ja obnizaja.

PacBio - dlugie odczyty z wysoka dokladnoscia

Pacific Biosciences stosuje technologie SMRT (Single Molecule Real-Time): polimeraza DNA odczytuje pojedyncza czasteczke w czasie rzeczywistym w zeromodowych falowodach. Tryb HiFi (CCS, Circular Consensus Sequencing) pozwala uzyskac odczyty 15-25 kb z dokladnoscia powyzej 99,9% - laczac zalety dlugich odczytow z precyzja Illumina.

Wybor miedzy Illumina, ONT a PacBio to nie kwestia prestizu platformy, lecz dopasowania do pytania badawczego. Wariantowanie SNP w duzej kohorcie? Illumina. Montaz kompletnego genomu? PacBio lub ONT. Szybka diagnostyka w terenie? MinION.

5. Bioinformatyczna analiza danych - od surowych odczytow do wariantow

Dane z sekwencera to surowe pliki FASTQ - kazdy odczyt wraz z ocena jakosci kazdej bazy (wartosc Phred Q). To dopiero punkt startowy analizy bioinformatycznej, ktora dla genomu ludzkiego (30x pokrycie) moze zajac kilka godzin na klastrze obliczeniowym.

Krok 1: Kontrola jakosci odczytow (FastQC, MultiQC)

Pierwszym narzedziem jest FastQC - generuje raport jakosci odczytow, pokazujac m.in. rozklad wartosci Phred wzdluz odczytu, zawartosc GC, obecnosc adapterow i duplikaty. MultiQC agreguje raporty z wielu probek. Typowe problemy to spadek jakosci na koncach 3' odczytow (normalny dla Illumina) i zanieczyszczenie sekwencjami adapterow.

Krok 2: Przycinanie adapterow i filtrowanie (Trimmomatic, fastp)

Narzedzia takie jak Trimmomatic lub fastp usuwaja sekwencje adapterow i odczyty niskiej jakosci. Parametry przycinania maja znaczacy wplyw na wyniki - zbyt agresywne filtrowanie usuwa prawdziwe dane, zbyt lagodne zostawia artefakty. Fastp jest szybszy i ma lepszy interfejs, ale Trimmomatic jest wciaz szeroko stosowany w starszych pipeline'ach.

Krok 3: Mapowanie do genomu referencyjnego (BWA-MEM2, Minimap2)

Odczyty sa mapowane do genomu referencyjnego (dla czlowieka: GRCh38/hg38 lub T2T-CHM13 z 2022 roku). BWA-MEM2 jest standardem dla krotkich odczytow Illumina; Minimap2 obsluguje zarowno krotkie, jak i dlugie odczyty ONT i PacBio. Wynikiem jest plik BAM - skompresowany, posortowany plik z informacja o pozycji kazdego odczytu w genomie.

Krok 4: Przetwarzanie BAM i usuwanie duplikatow (GATK, samtools)

Duplikaty PCR sa oznaczane przez narzedzia takie jak Picard MarkDuplicates lub GATK MarkDuplicatesSpark. Jesli uzyto UMI, ten krok jest bardziej precyzyjny. Nastepuje tez rekalibracja jakosci baz (BQSR, Base Quality Score Recalibration), ktora koryguje systematyczne bledy polimerazy sekwencera.

Krok 5: Wykrywanie wariantow (variant calling)

To etap, w ktorym bioinformatyka spotyka sie z biologia. GATK HaplotypeCaller jest zlotym standardem dla wariantow germinalnych (SNP i indele). Dla wariantow somatycznych (np. w nowotworach) stosuje sie Mutect2. Wynikiem jest plik VCF (Variant Call Format) zawierajacy liste roznic miedzy probka a genomem referencyjnym.

Plik VCF z analizy calego genomu moze zawierac 4-5 milionow wariantow. Kluczowe pytanie nie brzmi "ile wariantow mamy