Bioinformatyczna obróbka danych

Wykład 7

Michał Szczesniak, 2025

Ogólny przebieg analizy

- Kontrola jakości odczytów (QC)
- Przycinanie odczytów
- Mapowanie do genomu referencyjnego
- Usuwanie/oznaczanie duplikatów PCR
- Wywoływanie wariantów (variant calling)
- Filtrowanie i adnotacja wariantów

QC

Podczas kontroli jakości zwracamy uwagę m.in. na:

- Jakość nukleotydów na końcach odczytów
- Obecność adapterów
- Rozkład GC
- Duplication levels

QC

- Narzędzia takie jak **Trim Galore**, **Cutadapt** czy **Trimmomatic** umożliwiają automatyczne znalezienie i usunięcie takich sekwencji
- Trim Galore jest często używanym rozwiązaniem typu *all-in-one* – łączy funkcjonalność wyszukiwania adapterów (opartego o **Cutadapt**) z przycinaniem złej jakości nt na końcach. Działanie tych narzędzi polega m.in. na:
  - Usuwaniu adapterów Illumina
  - Przycinaniu ze względu na jakość
  - Usuwaniu zbyt krótkich odczytów

Mapowanie

- **BWA-MEM** (Burrows-Wheeler Aligner, algorytm MEM) jest obecnie uważany za standardowy aligner dla sekwencji genomowych o długościach odczytów ~100-150 bp. Jest on zoptymalizowany pod kątem szybkości i dokładności mapowania.
- BWA-MEM wykonuje dopasowanie lokalne

Usuwanie duplikatów PCR

- Najpopularniejszym narzędziem jest **Picard MarkDuplicates** (obecnie rozwijany w ramach Picard/GATK). Działa ono w ten sposób, że przegląda posortowany plik BAM i grupuje odczyty mające taką samą pozycję początkową zmapowania na referencji oraz taką samą mapującą parę (dla danych sparowanych). Takie odczyty uważa się za duplikaty pochodzące z jednego fragmentu DNA. Picard oznacza wszystkie poza jednym z nich jako duplikaty (dodając flagę “duplicate” w rekordzie BAM). Można następnie takie odczyty usunąć lub po prostu pozostawić oznaczone - nowoczesne pipeline’y zwykle nie wyrzucają ich fizycznie z pliku, lecz oznaczają (*mark*) i pozostawiają.
- Definicja duplikatu w tym kontekście obejmuje zarówno duplikaty PCR, jak i tzw. duplikaty optyczne (błędy sekwenatora, gdzie intensywne klastry na flowcellu są policzone wielokrotnie). MarkDuplicates domyślnie wykrywa jedne i drugie.
- W eksperymentach typu **PCR-free** (biblioteki bez amplifikacji) oczekujemy bardzo niskiego odsetka duplikatów, natomiast w targetowanych panelach czy WES, gdzie DNA wejściowego jest mniej, odsetek duplikatów może być wyższy (czasem 10-20%).

Wywoływanie wariantów dziedzicznych

- Standardowo wykorzystywanym narzędziem jest **GATK HaplotypeCaller** będący częścią pakietu GATK (Genome Analysis Toolkit). HaplotypeCaller działa w trybie *assembly-based*: dla każdej podejrzanej pozycji tworzy lokalny graf haplotypów i próbuje zrekonstruować najbardziej prawdopodobne haplotypy, a następnie zmapować odczyty do tych haplotypów w celu wygenerowania wiarygodnych wywołań wariantów. Jest to podejście bardziej zaawansowane niż proste zliczanie odchyleń od referencji, dzięki czemu HaplotypeCaller radzi sobie lepiej w regionach z wieloma blisko położonymi zmianami oraz z indelami.
- Innym popularnym narzędziem jest **bcftools (samtools) mpileup + call**, które wykorzystuje klasyczny model samych stosów odczytów (pileup) i statystyki bayesowskie do wywoływania SNP/indeli. Bcftools działa szybciej i bywa używane np. w analizach populacyjnych lub wtedy, gdy zależy nam na prostej i szybkiej analizie.
- GATK z HaplotypeCaller oferuje zwykle wyższą czułość i specyficzność, zwłaszcza w trudnych rejonach genomu, kosztem większych wymagań obliczeniowych.

Plik VCF

Głównym wynikiem analizy wariantów genomowych jest plik w formacie VCF (ang. *Variant Call Format*). Plik VCF zawiera informacje o wykrytych wariantach sekwencyjnych w genomie, przedstawione w polach oddzielonych znakiem tabulacji. Te pola to:

1. Nazwa chromosomu, np. **chr1**.
2. Pozycja wariantu na chromosomie, np. **3319632**.
3. Identyfikator wariantu w bazach danych, np. **rs2282198** w bazie danych dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/). W przypadku gdy wariant nie jest zaadnotowany w bazach danych w dbSNP lub ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), w tym miejscu pojawia się kropka.
4. Nukleotyd lub nukleotydy (w przypadku indeli) znajdujące się we wcześniej określonej pozycji w genomie referencyjnym, np. **ATTATT**.
5. Lista alternatywnych alleli w tej samej pozycji co w pkt 4, np. **ATTATTTTTATT**.
6. Jakość przypisana wykrytemu wariantowi, np. **228.0**.
7. Pole opisujące czy wykryty wariant spełnia założone kryteria filtrowania. Zazwyczaj pole to ma wartość **PASS**.
8. Opis i adnotacja wariantu; różnego rodzaju informacje nt. wariantu podawane są w formacie klucz=wartość i są oddzielone średnikiem. Szczegółowy opis można znaleźć w dokumentacji formatu VCF, na stronie https://samtools.github.io/hts-specs/VCFv4.2.pdf.
9. Pole 9 i dalsze pola służą do opisu próbek użytych w analizie danych.

Oprócz tego, plik VCF posiada nagłówek, z liniami rozpoczynającymi się znakiem „#”. Zawiera on tzw. metadane, między innymi opis uzyskanych wyników i użytej metodologii, np. jaki program i z jakimi parametrami był wykorzystany w analizie. Szczegółowe informacje można znaleźć w dokumentacji formatu VCF: https://samtools.github.io/hts-specs/VCFv4.2.pdf.

Filtrowanie i adnotacja wariantów

- **Filtrowanie proste (hard filters):** Dla każdej klasy wariantów ustalamy progi dla wybranych metryk – np. odrzucamy SNP o jakości QUAL < 30, o wartości DP (głębia) zbyt niskiej lub zbyt wysokiej (wskazującej na błąd lub duplikację), o wartości wskaźnika Phred-scaled Fisher Strand (FS) powyżej pewnego poziomu (co sugeruje stronniczość w rozkładzie odczytów forward/reverse). GATK udostępnia zalecane progi, np. **FS > 60** i **QD < 2** są często używane do odfiltrowania kiepskich SNP. Warianty niespełniające kryteriów otrzymują flagę FILTER w VCF (np. “FILTER=LowQual”).
- **VQSR (Variant Quality Score Recalibration):** Jest to zaawansowana metoda używana w GATK, gdzie model machine learning uczy się na podstawie zbioru wysokiej jakości wariantów (np. znanych z baz danych) oraz oczywistych negatywnych przykładów, a następnie nadaje każdemu wariantowi score VQSLOD. Następnie wybiera się próg odcięcia tego score, aby osiągnąć pożądaną równowagę czułości i precyzji. VQSR potrafi wykryć warianty podejrzane o bycie artefaktami, biorąc pod uwagę wiele cech jednocześnie. Wymaga jednak dość dużej liczby wariantów i obecności sprawdzonych zestawów treningowych – stąd bywa trudny do zastosowania w małych zbiorach (np. pojedynczy egzom), ale świetnie działa w dużych projektach populacyjnych.

Filtrowanie i adnotacja wariantów

Narzędzia takie jak **ANNOVAR**, **Ensembl VEP (Variant Effect Predictor)** czy **SnpEff** automatycznie określają dla każdego wariantu:

- w jakim znajduje się genie (lub regionie międzygenowym),
- czy zmienia sekwencję kodującą białko (np. nonsens, zmiana sensu, zmiana ramki odczytu), czy może jest intronowy, UTR, itp.,
- czy jest znany w bazach populacyjnych (np. czy występuje w gnomAD i z jaką częstością),
- czy jest notowany w bazach chorobowych (np. ClinVar – czy został zgłoszony jako patogenny/łagodny),
- jakie przewidywane skutki dla funkcji białka ma dana zmiana (np. wyniki programów predykcji typu **SIFT**, **PolyPhen**, **CADD**).

Warianty somatyczne

- Wykrywanie mutacji somatycznych jest bardziej złożone głównie z dwóch powodów: **mieszana populacja komórek** oraz **konieczność odróżnienia zmian germinalnych od somatycznych**. W próbce guza DNA pochodzi z wielu komórek nowotworowych (które mogą być zróżnicowane genetycznie) oraz często z domieszki komórek prawidłowych. Zatem wariant somatyczny obecny tylko w części komórek nowotworu może pojawiać się z niską częstością allelu (np. 5%, 10% odczytów), co jest sygnałem dużo słabszym niż heterozygotyczny wariant dziedziczny ~50%.
- Z danych sekwencjonowania nie da się od razu stwierdzić, czy dane odchylenie od referencji w guzie to unikalna mutacja somatyczna, czy po prostu pacjent miał ten wariant odziedziczony. Dlatego standardem jest posiadanie dla porównania materiału **normalnego** od tego samego pacjenta (np. krew, zdrowa tkanka) i wykonywanie analizy pary **tumor/normal**.
- W pipeline somatycznym najważniejszy etap to *somatic variant calling*, w którym narzędzie porównuje ułożenie odczytów z guza i z materiału referencyjnego (normalnego). Wariant somatyczny powinien występować w guzie, a nie występować w DNA normalnym.

Warianty somatyczne

- **Mutect2** (GATK) jest aktualnie jednym z najczęściej używanych callerów somatycznych.
- **Strelka2** to inne popularne narzędzie, które działa bardzo wydajnie i dokładnie, ale wymaga, by podać parę *tumor + normal* (Strelka2 nie działa w trybie tumor-only).
- **VarScan2** jest nieco starszym programem, który operuje prostszym sposobem (zlicza odczyty w tumor i normal i wykonuje test statystyczny na różnicę częstości allelu). VarScan2 bywa czuły, ale generuje więcej false positives i zazwyczaj wymaga intensywnego filtrowania (np. za pomocą skryptu *somaticFilter* i dodatkowych metod).

Warianty somatyczne

- Mutect2 modeluje tło błędów: analizuje każdą pozycję, patrzy ile odczytów w normalnej próbce ma dany alternatywny allel (idealnie 0 lub minimalnie), a ile w tumor, i na tej podstawie liczy prawdopodobieństwo, że obserwowany sygnał w guzie jest istotny. Dodatkowo stosuje mechanizmy usuwania artefaktów: np. wykorzystanie **Panel of Normal (PoN)** – listy wariantów często pojawiających się w kontrolach (np. błędy technologiczne), które są automatycznie odrzucane. Wykorzystuje też bazy populacyjne (gnomAD) do odrzucenia częstych wariantów (bo raczej dziedziczne) oraz specjalny etap filtrowania **Mutect2 Filter** do oznaczania podejrzanych pozycji.
- Przy niskich częstościach alleli rozróżnienie prawdziwego wariantu od szumu bywa trudne – dlatego często stosuje się dodatkowe potwierdzenie (np. sekwencjonowanie głębokie ukierunkowane lub techniki niezależne). Mimo to, nowoczesne callery jak Mutect2 i Strelka2 radzą sobie imponująco – w testach osiągają dobre wyniki w zakresie czułości i precyzji przy wykrywaniu mutacji nawet o częstości <5%, choć kosztem spadku pewności.
- W praktyce analizy nowotworów, pipeline wygląda następująco: mapujemy (tworzymy BAM tumor i BAM normal, wykonujemy mark duplicates, etc. dla obu), następnie uruchamiamy np. Mutect2: program przyjmuje jednocześnie BAM z guza i próbki normalnej i generuje VCF z wariantami somatycznymi. Potem typowo stosuje się skrypt **FilterMutectCalls** (GATK) do wstępnego filtrowania oraz ewentualnie dodatkowe filtry jak **FilterByOrientationBias**. Dalsze etapy to adnotacja (czy dana mutacja jest w onkogenie, czy zmienia funkcję białka, czy jest „actionable” klinicznie, itd.).

Wykrywanie wariantów strukturalnych

Podstawowe sygnały SV w danych short-read:

- **Discordant read pairs:** W sekwencjonowaniu paired-end oczekujemy, że oba końce pary mapują w poprawnej orientacji i w odległości zgodnej z wielkością fragmentu. Jeśli fragment zawiera dużą delecję, to dwa odczyty mogą mapować dużo dalej od siebie niż zwykle (para o zbyt dużym insert size) – to tzw. para niezgodna, sygnalizująca delecję. Jeśli jest insercja w próbce, jeden z odczytów może nie mapować, lub mapować gdzie indziej. Np. w przypadku translokacji – jedna połowa pary mapuje na chromosomie 1, druga na chromosomie 8.
- **Split reads:** Pojedynczy odczyt, który mapuje częściowo do dwóch różnych miejsc referencji. Np. odczyt przecinający punkt pęknięcia (breakpoint) dużej insercji – jedna jego część mapuje przed insercją, druga za insercją, z przerwą. Alignery często zgłaszają taki odczyt jako tzw. *soft-clipped*, gdzie część sekwencji nie dopasowała się w jednym ciągu. Analiza tych miękko odciętych odczytów pozwala dokładnie zlokalizować miejsce rearanżacji genomowej.
- **Zmiany w pokryciu:** Duplikacja dużego fragmentu objawi się wzrostem liczby odczytów mapujących do tego regionu (lokalny skok pokrycia), a delecja – spadkiem pokrycia.

Wykrywanie wariantów strukturalnych

Narzędzia do wykrywania SV często łączą powyższe sygnały, by zwiększyć wiarygodność detekcji. Przykładowo:

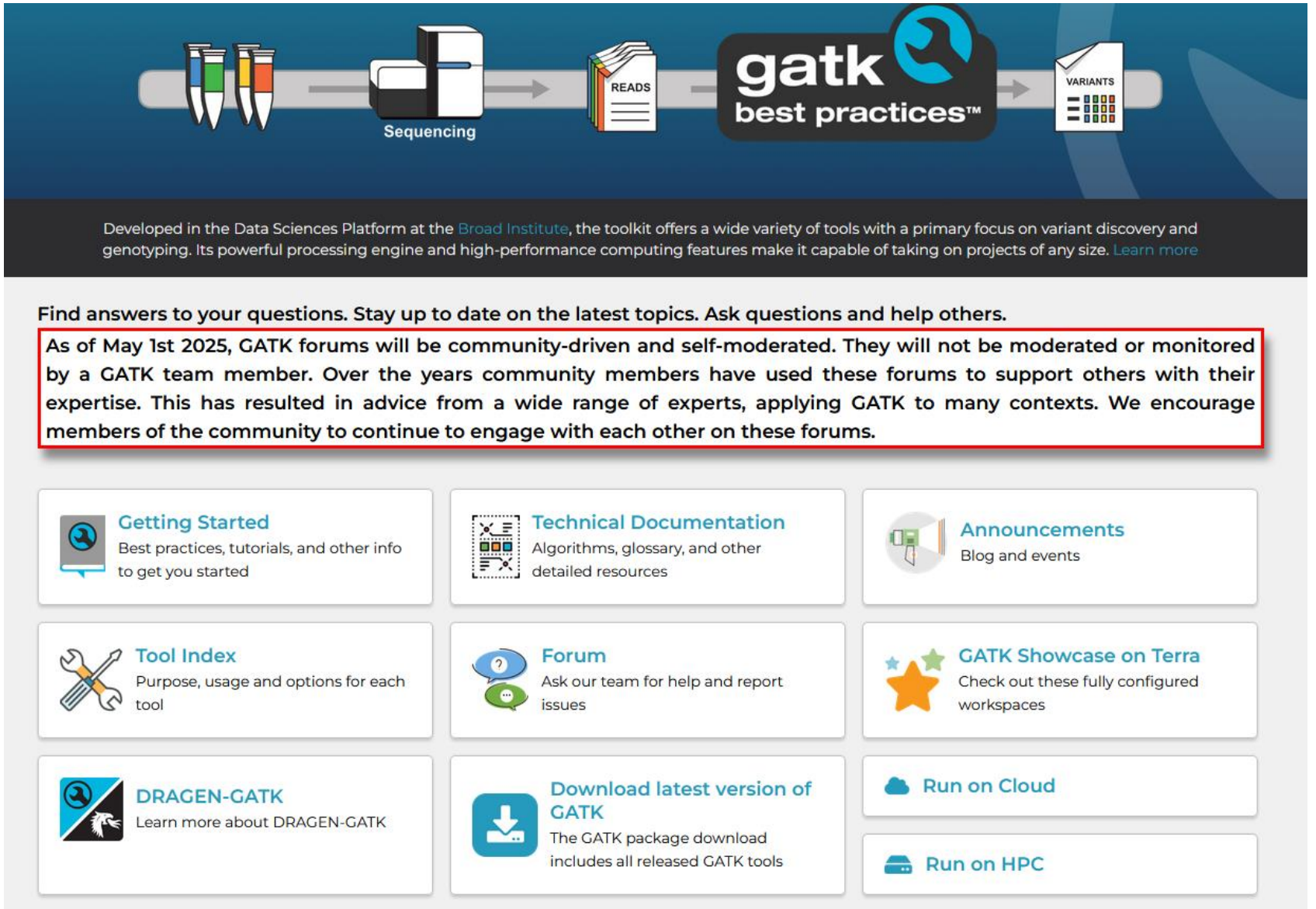
- **Manta** (Illumina) – popularny SV caller, który wykrywa kandydatów SV w oparciu o pary discordant i split reads. Manta potrafi identyfikować różne klasy SV (delecje, insercje, inwersje, translokacje) i jest stosunkowo szybka. Wymaga danych z dość głębokim pokryciem (gł. WGS), na WES też bywa używana, choć czułość spada dla zmian, które akurat są na granicy egzonów.
- **LUMPY** – narzędzie o podejściu probabilistycznym, integrujące wiele sygnałów jednocześnie (paired-end, split-read, read-depth) i mogące korzystać z danych wielu próbek jednocześnie. LUMPY generuje listę breakpointów
- **GRIDSS** – nowszy framework, który idzie krok dalej: zawiera własny assembler do lokalnego składania sekwencji na końcach potencjalnych przerw (break-ends) i łączy dowody z assembly, odczytów dzielonych i par niezgodnych z modelem statystycznym. GRIDSS osiąga wysoką czułość i specyficzność, ale jest intensywny obliczeniowo (składanie wielu kandydatów). Często używany jest w trybie somatycznym (GRIDSS2 dla par tumor/normal).

Inne narzędzia warte uwagi to np. **Delly**, **SvABA**, **MELT** (dedykowany insercjom transpozonów), **CNVnator** (typowo do zmian kopii na bazie pokrycia). Każde z nich ma swój algorytm, a w praktyce w projektach WGS często uruchamia się kilka SV-callerów równolegle i potem łączy ich wyniki (tzw. konsensus, bo wykrywanie SV bywa trudniejsze).

Coraz częściej do wykrywania SV wykorzystuje się również **długie odczyty** (PacBio, Oxford Nanopore), gdzie obserwowanie dużych insercji czy translokacji jest prostsze.

Wybrane pipeline do analizy wariantów

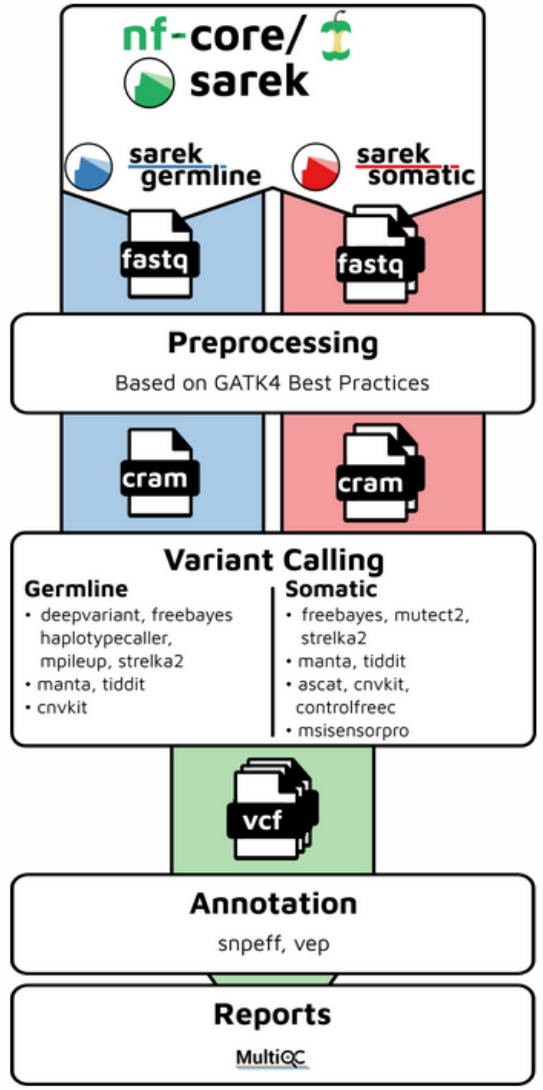
- **GATK Best Practices** – Jest to zestaw oficjalnych zaleceń od Broad Institute dla wykrywania wariantów dziedzicznych i somatycznych. Obejmuje konkretne narzędzia (np. BWA-MEM, Picard MarkDuplicates, BaseRecalibrator dla BQSR, HaplotypeCaller, VariantRecalibrator dla VQSR).
- GATK Best Practices jest swego rodzaju “złotym standardem” i punktem odniesienia, potwierdzonym przez tysiące eksperymentów.
- Zapewnia wysoką dokładność wykrywania wariantów, jednak za cenę czasu obliczeniowego. Dlatego do projektów wielkoskalowych z użyciem GATK często wykorzystuje się klastry HPC lub chmury. Best Practices zakładają też wykorzystanie dodatkowych danych (np. znane warianty do BQSR, bazy kontrolne do VQSR).



https://gatk.broadinstitute.org/hc/en-us

Wybrane pipeline do analizy wariantów

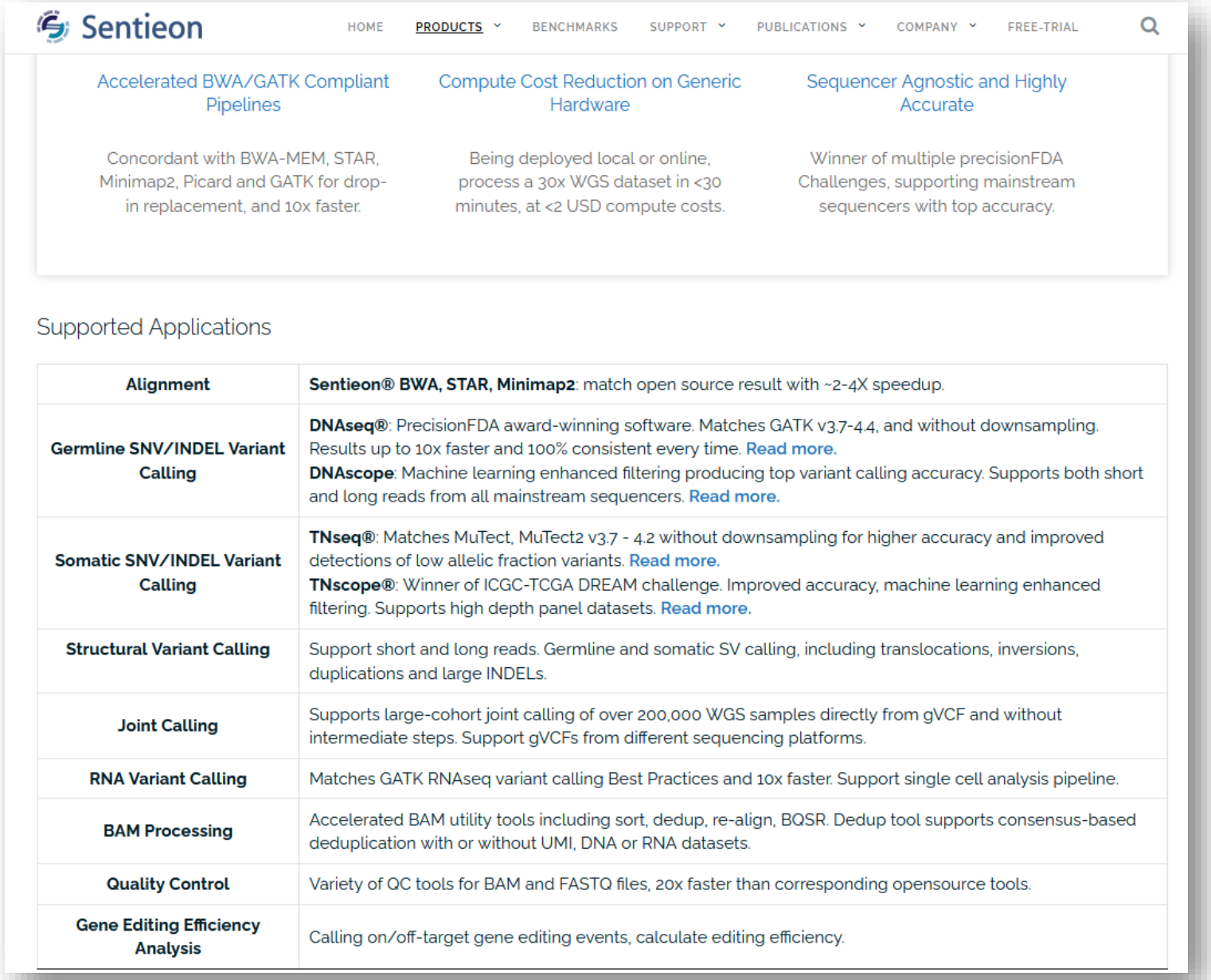
- **nf-core/sarek** – gotowy pipeline oparty o framework Nextflow.
- Sarek jest przeznaczony do wykrywania zarówno wariantów dziedzicznych, jak i somatycznych dla WGS/WES. Pipeline ten integruje narzędzia: od mapowania (BWA), poprzez GATK (lub alternatywy) po wywoływanie wariantów kilkoma metodami i adnotację. Jego dużą zaletą jest automatyzacja – użytkownik dostarcza pliki FASTQ i parametry, a Sarek wykonuje całość na infrastrukturze klastrowej lub w chmurze.
- nf-core/sarek kładzie nacisk na powtarzalność i best practices, więc domyślnie używa kontenerów z zainstalowanymi właściwymi wersjami narzędzi, standardów nazw plików itp.
- nf-core/sarek stał się w wielu jednostkach domyślnym narzędziem do analizy danych klinicznych, bo zapewnia kompleksowość (od raw data do wariantów i raportów) oraz aktualizuje się wraz z rozwojem narzędzi.



https://nf-co.re/sarek/3.2.3/

Wybrane pipeline do analizy wariantów

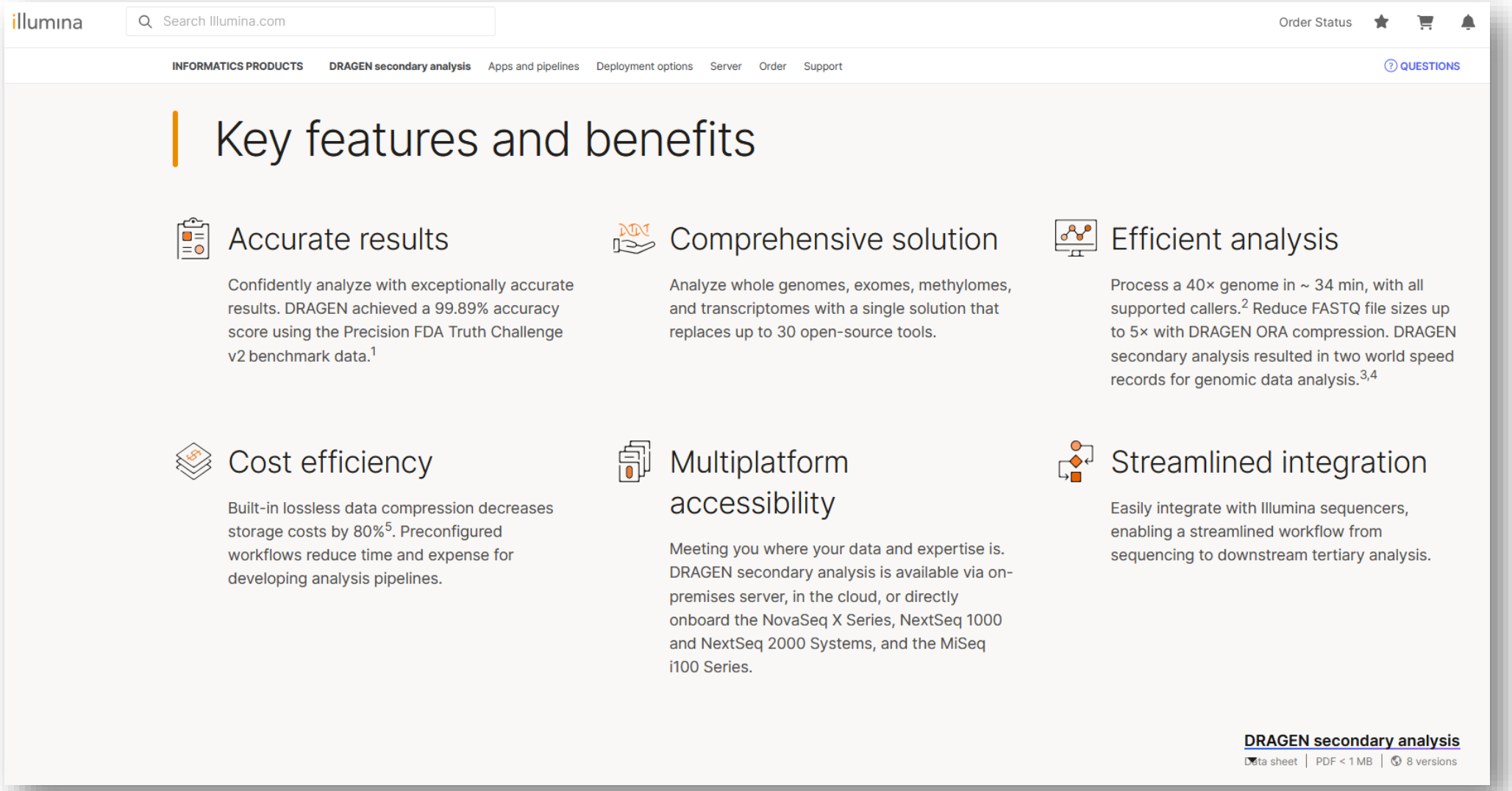
- **Sentieon DNAseq** – komercyjny pakiet będący zoptymalizowaną reimplementacją pipeline'u GATK. Firma Sentieon przepisała algorytmy GATK (takie jak realignment, BQSR, HaplotypeCaller) w wydajniejszy sposób (niski poziom C/C++), dzięki czemu osiąga znaczne przyspieszenie bez utraty dokładności.
- W publikacjach i testach wykazano, że Sentieon generuje (praktycznie) identyczne wyniki co GATK Best Practices, tylko że wielokrotnie szybciej – np. ponad **20 razy szybciej** na tym samym sprzęcie.
- Dla dużych instytucji sekwencjonujących setki genomów Sentieon bywa opłacalny, bo oszczędza czas/koszty obliczeń.



https://www.sentieon.com/products/#who_we_serve

Wybrane pipeline do analizy wariantów

- **DRAGEN** (Dynamic Read Analysis for GENomics) został zaprojektowany, by wykonywać **ekstremalnie szybkie** analizy WGS – potrafi przetworzyć pełny genom 30x w kilkanaście minut. Osiąga to poprzez przeniesienie kluczowych zadań (indeksowanie, mapowanie, sortowanie, call SNP/indel) do sprzętu FPGA.
- DRAGEN nie obniża przy tym jakości – w konkursach (np. PrecisionFDA) uzyskiwał najwyższe oceny dokładności zarówno dla SNP, jak i indeli.
- W praktyce, sekwenatory Illumina NovaSeq mogą być wyposażone w moduł DRAGEN, co oznacza, że po zakończeniu sekwencjonowania wyniki analizy wariantów są od razu dostępne, bez potrzeby przesyłania terabajtów danych do serwerów.
- Wady*:* koszt (dedykowany sprzęt), mniejsza elastyczność (pipeline jest zoptymalizowany pod konkretny cel - głównie WGS; choć obsługuje też egzomy i panele). Pod względem wymagań sprzętowych – wymaga albo posiadania karty FPGA Illumina, albo wynajmu mocy w chmurze. Nie jest to narzędzie “open-source”, więc użytkownik nie zajrzy do środka algorytmów.
- DRAGEN stał się popularny w dużych inicjatywach (np. sekwencjonowanie całych populacji), gdzie przepustowość ma krytyczne znaczenie.



https://emea.illumina.com/products/by-type/informatics-products/dragen-secondary-analysis.html