Algoritamski alati u bioinformatici

Uvodni pojmovi

Bioinformatika je interdisciplinarna istraživačka i industrijska oblast, koja povezuje biologiju i računarske nauke. Značaj bioinformatike:

- personalizovana medicina - prilagođavanje terapije genetskom profilu pacijenta.
- farmaceutska industrija - dizajn novih lekova i predviđanje njihovih efekata biomarkera i dijagnostika bolesti.
- biološko inženjerstvo - selekcija i unapređivanje biljnih i životinjskih vrsta, selekcija GMO, analiza poljoprivrednih sistema.
- ekologija - praćenje i analiza ekosistema, vrsta i klimatskih promena.
- razumevanje života - ključ za otkrivanje, razumevanje i integraciju velikih količina bioloških podataka.
- evo-devo istraživanje - proučavanje evolucije i razvoja tkiva, organa i organizama.

Glavni izazovi u bioinformatici:

- obim, kompleksnost, integracija i obrada podataka
- zastarevanje podataka i istraživanja
- standardi (nomenklatura i protokoli)
- interpretacija (rezultati, izveštaji, vizualizacija)
- rapidan tehnološki napredak
- nedostatak kvalifikovanog kadra

Sekvenciranje je proces određivanja tačnog redosleda nukleotida u molekulu DNK ili RNK, odnosno čitanje genetske informacije u obliku niza baza. Glavni koraci:

1. izolacija pojedinačnih ćelija
2. amplifikacija uzorka za sekvenciranje
3. prečišćavanje magnetnim kuglicama
4. pripremanje biblioteke
5. sekvenciranje metodama sledeće generacije
6. komputaciona obrada i vizualizacija

Bioinformatičke platforme su softverski sistemi koji integrišu alate, baze podataka i algoritme za obradu, analizu i interpretaciju bioloških podataka. Njihova uloga je da omoguće istraživačima da:

- bezbedno skladište i organizuju biološke podatke

- vrše sekvenciranje, poravnanja, anotacije i analize
- vizualizuju i interpretiraju rezultate i povezuju ih sa postojećim biološkim znanjem

Bioinformatički tokovi (pipelines) su serije komputacionih koraka, gde ulaz u jedan korak predstavlja izlaz iz prethodnog koraka. Komunikacija se vrši preko rezultujućih fajlova. **Nextflow** je okvir koji služi za automatizaciju pipeline-ova. Svaki alat je upakovan u docker kontejner i pokreće se u okviru njega, a nextflow orkestrira pokretanje kontejnera i dodeljuje resurse. Na ovaj način dobijamo veću modularnost i ponovnu iskoristivost.

Bioinformatičke analize su računarski zasnovane metode za obradu i tumačenje bioloških podataka. One obuhvataju niz postupaka kojima se iz velikih i složenih skupova podataka (npr. DNK/RNK sekvence) izvlače biološki smisleni zaključci. Koraci analize su organizovani u pipeline-ove.

Ključne veštine jednog bioinformatičara su:

- **informatika**: Python/R, Nextflow, Unix, Cloud, algoritmi, ...
- **biohemija/biologija**: genetika, centralna dogma, PCR, sekvenciranje, ...
- **data science**: statistika, modeliranje, transformacija podataka, predviđanje, istraživanje, ...

Mikrobiologija

Ljudsko telo je sistem **organa**, a organi su sistemi **tkiva**. Tkiva mogu biti vezivna, mišićna, epitelna i nervna. Tkiva su sistemi **ćelija**, a ćelije su jedinice života. Veliki deo onoga što ćelije rade čine **proteini**. "Blueprints" za proteine nalaze se u jezgru ćelije. Molekul **DNK** je niz nukleotida (baza). Sadrži **adenin**, **timin**, **citozin** i **guanin**. DNK je sastavljen od dva uvijena lanca. U ljudskom organizmu imamo 3 milijarde baznih parova. Parovi su $A = T$ i $C \equiv G$. **Hromatin** je kompleks DNK i proteina u jedru ćelije. Organizuje DNK tako da stane u jedro i reguliše ekspresiju gena. **Histoni** su osnovni proteini u hromatinu oko kojih se DNK namotava. Pomažu u pakovanju DNK i učestvuju u regulaciji gena. **Nukleozom** je osnovna jedinica hromatina. Ljudski organizam ima 23 para **hromozoma**, od toga 22 **autozomna** para i 1 par **polnih hromozoma** (žene XX, muškarci XY).

Replikacija je proces udvostručavanja DNK, kojim nastaju dve identične kopije genetskog materijala. **Helikaza** je enzim koji razdvaja lance DNK. DNK lanci imaju smer. **DNK polimeraza** je enzim koji vrši replikaciju i on može da radi samo u smeru 5' $\rightarrow$ 3'. **Transkripcija** je proces prepisivanja DNK u RNK (najčešće **mRNK - informaciona RNK**). **Egzoni** su delovi DNK koji se pretvaraju u protein, a **introni** su nekodirajući delovi koji se uklanjaju tokom transkripcije. **Translacija** je proces prevođenja informacija iz mRNK u protein, pomoću ribozoma i **tRNK (transportna RNK)**. **Kodoni** predstavljaju tri uzastopne baze u mRNK koje kodiraju određenu aminokiselinu ili signal za početak/kraj translacije. **Aminokiseline** su osnovni gradivni elementi proteina. Postoji 20 standardnih aminokiselina u genetskom kodu, a njihovi redosledi određuju strukturu i funkciju proteina. **Centralna dogma molekularne biologije** opisuje osnovni tok genetičke informacije u ćeliji: genetička

informacija se čuva u DNK, prepisuje u RNK i na kraju prevodi u proteine, koji obavljaju funkcije u ćeliji.

Geni su delovi DNK sa informacijom o pravljenju proteina. Jednu karakteristiku mogu da određuju više gena koji međusobno interaguju. **Aleli** su različiti oblici jednog istog gena. Jedan gen se sastoji od dva alela, jedan koji se nalazi na hromozomu nasleđenom od majke, a drugi na homolognom hromozomu nasleđenom od oca. Za ćeliju se kaže da je **homozigotna** za određeni gen kada su identični aleli gena prisutni na oba hromozoma. Za ćeliju se kaže da je **heterozigotna** za određeni gen kada su različiti aleli gena prisutni na hromozomima. **Dominantni alel** je alel čija osobina može da se izrazi i kada je prisutan samo jedan primerak tog alela (heterozigot). **Recesivni alel** je alel čija se osobina izražava samo kada su prisutna dva primerka tog alela (homozigot). **Genotip** je skup svih alela. **Fenotip** je skup svih osobina jednog organizma koje su nastale zajedničkim delovanjem genotipa. **Genska mutacija** je promena u sekvenci nukleotida jednog gena koja može da utiče na funkciju ili ekspresiju proteina koji taj gen kodira. Vrste mutacija:

- **tačkaste mutacije (SNP - Single Nucleotide Polymorphism)** - zamena jedne baze za drugu bazu. Ovo su veoma učestale varijacije.
- **delecije** - obrisana jedna ili više baza.
- **insercije** - dodata jedna ili više baza.
- **InDel** - na istom mestu obrisano i dodato više baza.
- **varijacija broja kopija (CNV - Copy Number Variations)** - promena u broju kopija nekog dela.
- **translokacije** - jedan deo izmešten na drugo mesto.
- **inverzije** - obrnut redosled nekog dela.

Sekvenciranje

Sekvenciranje je proces određivanja tačnog redosleda nukleotida u molekulu DNK ili RNK, odnosno čitanje genetske informacije u obliku niza baza. Sekvenciranje ima veliku ulogu u istraživanju ekspresije gena i istraživanju kancera. **Sanger sekvenciranje** je prvi algoritam sekvenciranja. Koraci:

1. **PCR (Polymerase Chain Reaction)** i **razdvajanje lanca**: PCR je laboratorijska metoda koja omogućava brzo i ciljano umnožavanje specifičnih sekvenci DNK. Za razdvajanje lanaca se koristi mešavina normalnih i fluorescentnih di-deoksi-nukleotida (**ddNTP**). Oni nemaju OH grupu na kraju, pa prekidaju lanac. Kao rezultat ovog koraka dobijamo milijarde oligonukleotida različitih dužina.
2. **Southern Blotting** i **gel za elektroforezu**: U gelu DNK se na jedan kraj ubacuju gel-matrice i pušta se struja. DNK je naelektrisana negativno, pa se oligonukleotidi kreću ka drugoj strani gela. Svi DNK fragmenti imaju istu masu pa oligonukleotidi koji su najmanji prvi stižu na drugu stranu.
3. **Čitanje lanca**: ddNTP su fluorescnentni pa emituju određenu broju kada se tretiraju laserom.

Sekvenciranje sledeće generacije (NGS) je moderna tehnologija koja omogućava istovremeno sekvenciranje miliona fragmenata DNK ili RNK. Glavne tehnologije:

- **CRT/TIRF (Illumina) Imaging**
- **SMRT/ZMW Imaging**
- **Nanopore Sequencing**

Zasnovano je na **sekvenciranju pomoću sinteze (Sequencing by Synthesis - SBS)**, a čvrsto se oslanja i na PCR. Koraci PCR:

1. **denaturacija** - razdvajanje lanaca na visokoj temperaturi.
2. **aniling** - dodavanje prajmera na 3' kraj.
3. **elongacija** - produžavanje lanca pomoću **TAQ polimeraze** koja je rezistentna na više temperature.

CRT/TIRF (Illumina) sekvenciranje počinje pripremom uzorka, gde se DNK ili RNK fragmentuje na manje delove i dodaju se adapteri - kratki sekvencijalni segmenti koji omogućavaju vezivanje na površinu sekvencijskog instrumenta. Ovi adapteri takođe sadrže bar kodove koji omogućavaju identifikaciju različitih uzoraka u istom eksperimentu, što je poznato kao multiplexing. Nakon pripreme, fragmenti se učitavaju na **flow cell**, tanku staklenu ploču prekrivenu oligonukleotidima koji omogućavaju vezivanje i fiksaciju fragmenta. Nakon vezivanja, fragmenti DNK se amplifikuju pomoću **bridge amplification** procesa. Tokom ovog procesa, svaki molekul DNK formira most sa susednim oligonukleotidom, stvarajući klastere identičnih molekula. Ovi klasteri omogućavaju detekciju signala sa svakog fragmenta pojedinačno, čime se povećava osetljivost i preciznost sekvenciranja. Samo sekvenciranje se vrši dodavanjem fluorescentno obeleženih nukleotida (dNTP). Svaki dNTP je označen specifičnom bojom koja odgovara jednoj od baza. Kada se nukleotid inkorporuje u rastući lanac DNK, emituje se fluorescentni signal koji se detektuje kamerom. Nakon svakog ciklusa, signal se snima, a boja se uklanja kako bi se omogućila sledeća inkorporacija. Ovaj proces se ponavlja u ciklusima, omogućavajući određivanje redosleda nukleotida u svakom fragmentu. Jedna od ključnih prednosti Illumina sekvenciranja je njegova sposobnost da generiše milione čitanja (reads) u jednom eksperimentu, što omogućava duboko pokrivanje genoma i precizno mapiranje sekvenci. Takođe, tehnologija omogućava **paired-end sekvenciranje**, gde se oba kraja fragmenta sekvenciraju, pružajući dodatne informacije o strukturi i organizaciji genoma. **Kvalitet baze** je verovatnoća kojom vidimo boju vezanu za bazu. Za skladištenje sekvenci koristi se **FASTQ** fajl format koji za svaku sekvencu sadrži labelu, samu sekvencu i kvalitete za svaku od baza u sekvenci.

Sekundardna bioinformatička analiza obuhvata analizu sekvenciranog DNK gde obično tražimo mutacije na DNK, kao i analizu sekvenciranog RNK gde obično posmatramo diferencijalnu ekspresiju, tj. kako će geni biti ekspresovani. Sekvenciranje DNK može biti:

- **sekvenciranje celog genoma (Whole Genome Sequencing - WGS)**: traže se nove varijante u svim delovima genoma. Pokrivenost read-ovima je oko 30.

- **sekvenciranje celog egzoma (Whole Exome Sequencing - WES)**: sekvencira se samo kodirajući deo genoma. Pokrivenost read-ovima je oko 100.
- **targetirano sekvenciranje**: sekvencira se jedan ili mali broj gena. Pokrivenost read-ovima je oko 1000.

Asembliranje genoma je proces kojim se kratki fragmenti DNK (reads) dobijeni sekvenciranjem ponovo spajaju u ceo genom. **Greedy algoritam**:

1. Izračunati poravnanje u pravima svih fragmenata.
2. Izabrati dva fragmenta sa najvećim preklapanjem.
3. Spojiti ta dva fragmenta.
4. Ponoviti korake 2 i 3 dok ne ostane samo jedan fragment.

Ovo nije optimalno rešenje, a čak i najpraktičnija rešenja mogu biti problematična jer zahtevaju mnogo memorije i imaju visoku kompjutacionu cenu. **Poravnanje** predstavlja mapiranje read-ova na poznati referentni genom. Ovo je mnogo brže i memorijski efikasnije od asembliranja, gde nemamo nikakvu referencu. **Referentni genom** je reprezentativni primer genoma jedna biološke vrste. Obično je napravljen od genoma nekoliko individualnih organizama te vrste. Svaki organizam jedne vrste se razlikuje od referentnog genoma na nekim mestima i ta razlika se naziva **genska varijacija**. **Projekat ljudskog genoma** je međunarodna naučna inicijativa da se napravi referentni genom čoveka. Trenutna verzija hromozoma je **HG38**, iako se koristi i HG37.

FASTA je jednostavan fajl format za čuvanje sekvenci. Postoje dva tipa linija:

- **opisna linija** - počinje sa '>' i sadrži ime sekvence i opcionu deskripciju.
- **linija sa sekvencom** - u novom redu nakon opisne linije. Sadrži samu nukleotidnu sekvencu. Obično imamo 70 ili manje baza u jednom redu.

Koriste se **IUPAC kodovi** za baze: **A**, **C**, **T**, **G**, **U**. Postoje i dodatni kodovi za nepoznate baze: **R** (A ili G), **Y** (C, T ili U) i **N** (bilo koja baza). Hromozomi mogu biti imenovani sa 1, 2, 3, ... ili sa chr1, chr2, chr3, ... **FASTA indeks (fai)** je fajl format koji opsiuje FASTA fajl sa istim imenom. Sadrži:

- ime sekvence
- dužinu sekvence
- ofset prve baze u fajlu
- broj baza u svakoj liniji FASTA fajla
- broj bajtova u svakoj liniji FASTA fajla

Anotacije su informacije koje se dodaju na genom ili sekvence DNK/RNK da bi se objasnilo značenje ili funkcija određenih regiona. Čuvaju se u nekoliko sličnih tipova fajlova kao što su BED, GTF, GFF. **BED** format ima sledeću strukturu:

CHROM START END NAME score ticks blocks

Poravnanje

Poravnanje predstavlja mapiranje read-ova na poznati referentni genom. Neki read-ovi (očitavanja) biće savršeno poravnati sa referencom, ali mnogi neće usled genomskih varijacija i grešaka prilikom sekvenciranja. Alati za poravnanje obično računaju poene ili skor zasnovan na distanci između očitavanja i lokalne sekvence reference. Očitavanje se onda postavlja na mestu gde je skor maksimalan, što se naziva **poravnanje bazirano na poentiranju pogotka/promašaja**. Postoje razni algoritmi u odnosu na izbor između brzine i preciznosti. Neki od njih su BBMap, BigBWA, Bowtie, BWA, Novoalign. Većina alata koristi pristup iz dva koraka:

1. **seed** - biramo "gruba" poravnanja. Ima mnogo lažno pozitivnih poravnanja, ali je ovaj korak obično veoma brz. Uobičajeni pristupi su k-mer heširanje, koren-drvo ili FM indeks (Burrows-Wheeler).
2. **extend** - tražimo "fina" poravnanja. Biramo poravnanja sa najvećim skorom. Ovaj korak je sporiji od prvog. Uglavnom je baziran na dinamičkom programiranju.

CIGAR string je niz simbola koji kodira kako se read poravnava sa referentnim genomom. Kodovi:

- **M** - pogodak ili promašaj
- **I** - insercija
- **D** - delecija
- **S** - soft clip (deo koji nije poravnat)

SAM (Sequence Align Map) je standardizovani format za čuvanje poravnanih očitavanja. Sadrži sekvence poravnanja i kvalitet, poziciju (hromozom i prva baza koja se poklapa), CIGAR string, poziciju para, druge flagovi i tagove. **BAM (Binary SAM)** je binarna verzija SAM fajla. Indeksiranje kreće od 0, dok u SAM kreće od 1. BAM fajlovi mogu biti sortirani po koordinatama ili po imenu. **BAM Index (bai)** je fajl koji omogućava indeksni pristup BAM fajlu.

Burrows-Wheeler transformacija (BWT) je algoritam koji prebacuje jedan string u drugi string koji je lakši za analizu.

BW Transformacija				
1. Ulaz	2. Sve rotacije	3. Sortiranje - leksikografski	4. Poslednja kolona	5. Izlaz
GATTACA$	GATTACA$ ATTACA$G TTACA$GA TACA$GAT ACA$GATT CA$GATTA A$GATTAC $GATTACA	$GATTACA A$GATTAC ACA$GATT ATTACA$G CA$GATTA GATTACA$ TACA$GAT TTACA$GA	$GATTACA A$GATTAC ACA$GATT ATTACA$G CA$GATTA GATTACA$ TACA$GAT TTACA$GA	ACTGA$TA

Niz sufiksa (Suffix Array) podrazumeva da označimo sve rotacije stringa, a zatim da ih sortiramo kao u BWT.

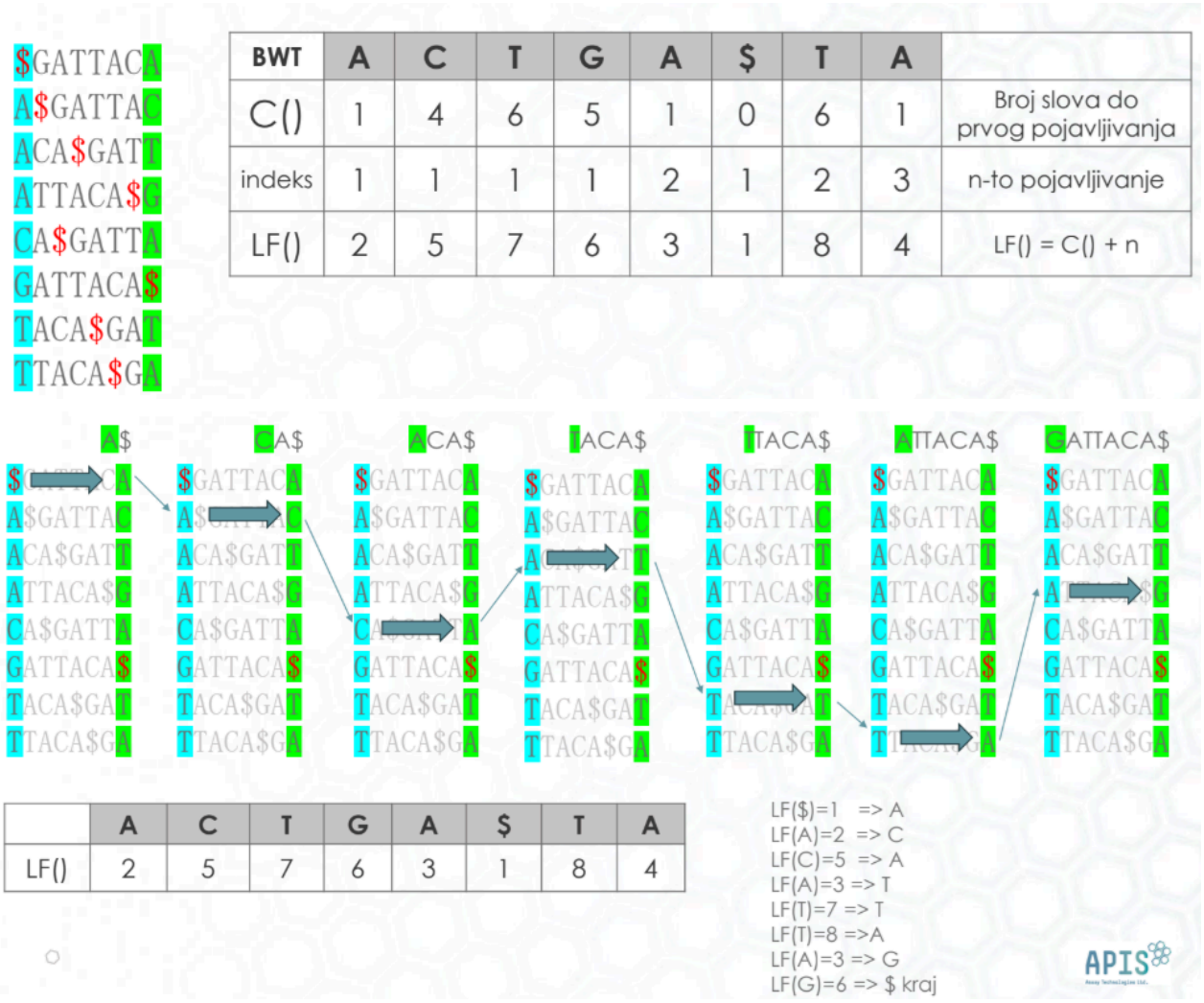
GATTACA$ - 1	$GATTACA - 8
ATTACA$G - 2	A$GATTAC - 7
TTACA$GA - 3	ACA$GATT - 5
TACA$GAT - 4	ATTACA$G - 2
ACA$GATT - 5	CA$GATTA - 6
CA$GATTA - 6	GATTACA$ - 1
A$GATTAC - 7	TACA$GAT - 4
$GATTACA - 8	TTACA$GA - 3

Veza između BWT i SA:

$$\text{BWT}(i) = \begin{cases} T(\text{SA}(i)-1), & \text{SA}(i) > 1 \\ \$, & \text{SA}(i) = 1 \end{cases}, \text{ gde je } T \text{ originalni string}$$

Naivni algoritam za BWT bi imao složenost $O(n^2 \log n)$. Možemo da zanemarimo sve iza znaka $ pa je ukupna složenost $O(n \log n)$. Dodatno, ako sortiramo sufikse do prvog, drugog, ..., 2^k karaktera redom, onda sortiranje sufiksa do 2^{k+1} možemo uraditi sa složenošću $O(k \log k)$. Sada smo složenost sveli na $O(\log n \log n)$, a uz radix sortiranje i na $O(\log n)$. Osobine BWT:

1. **pogodnija za kompresiju** - zbog sortiranja u BWT stringu ćemo imati duže nizove slova.
2. **reverzibilna**
3. **čuva redosled slova** - n-to pojavljivanje slova x u poslednjoj koloni je jednako n-tom pojavljivanju slova x u prvoj koloni. **LF (Last-First) mapiranje** mapira n-to pojavljivanje slova u poslednjoj koloni na prvu kolonu. Ako slovo x ima n-to pojavljivanje u BWT onda je $\text{LF}(x) = C(x) + n$, gde je $C(x)$ pozicija prvog pojavljivanja u prvoj koloni.



BWT	A	C	T	G	A	$	T	A	
C()	1	4	6	5	1	0	6	1	Broj slova do prvog pojavljivanja
indeks	1	1	1	1	2	1	2	3	n-to pojavljivanje
LF()	2	5	7	6	3	1	8	4	LF() = C() + n

	A	C	T	G	A	$	T	A
LF()	2	5	7	6	3	1	8	4

FM (Full-text Minute-space) index kombinuje BWT sa malim, dodatnim strukturama podataka. Sastoji se od prve i poslednje kolone iz BW matrice. Ovom tehnikom možemo pretragu da spustimo sa $O(n)$ na $O(1)$, ali nam treba matrica pojavljivanja u poslednjoj koloni. Možemo da računamo samo neke redove (npr. svaki i-ti) koje nazivamo **tačke provere (checkpoints)**.

	A	T	G	C
$GACTATA	1	0	0	0
A$GACTAT	1	1	0	0
ACTATA$G	1	1	1	0
ATA$GACT	1	2	1	0
CTATA$GA	2	2	1	0
GACTATA$	2	2	1	0
TA$GACTA	3	2	1	0
TATA$GAC	3	2	1	1

A	T	G	C
1	0	0	0
1	2	1	0
3	2	1	0

FM indeks ne zahteva mnogo memorije.

Burrows-Wheeler Aligner (BWA) je poravnanje koje se bazira na BWT i FM indeksu. U okviru BWA postoji nekoliko varijanta poravnanja kao što su BWA-backtrack, BWA-SW i

BWA-MEM. **BWA-MEM** koristi **seed-and-extend** pristup. Prvo se traže semena sa **supermaksimalnim tačnim poklapanjem (SMEM)**, pri čemu se koristi minimalna dužina semena. Da bi se smanjio broj promašenih očitavanja, primenjuje se postupak **re-seed**: ako SMEM ima dužinu L i pojavljuje se N puta u referentnom genomu, a L je preveliko, pravi se re-seed koristeći najduže poklapanje koje pokriva srednju bazu SMEM-a i pojavljuje se $N+1$ puta. **Lanci (chains)** su grupe semena koja su kolinearna i međusobno bliska. Tokom seed faze, greedy algoritam spaja lance i uklanja kratka semena koja su sadržana u lancu. Ovo filtriranje pomaže kasnijoj fazi ekstenzije. Semena se rangiraju prema dužini i dužini lanca kojem pripadaju. Zatim, za svako seme od najlošije do najbolje rangiranog, algoritam ili uklanja semena koja su već poravnata, ili ih proširuje. Proširenje se vrši pomoću **dinamičkog programiranja**, koristeći **Smith-Waterman algoritam**. Tokom ekstenzije algoritam prati najbolje skorove kako se približava kraju upitne sekvence. Ako je razlika između najboljeg skora do kraja sekvence i lokalnog skora ispod određenog praga, lokalno poravnanje se ignoriše. Postoji više metoda za određivanje **sličnosti sekvenci**:

- **Hamming distanca** - broj promena iz jednog stringa u drugi.
- **Edit distanca** - minimalan broj promena, insercija i delecija prilikom transformacije jednog stringa u drugi.

Pozivanje varijanti

Genske varijacije su razlike sekvenci DNK među pojedincima ili populacijama. Varijante u kodirajućem delu sekvence nazivaju se **mutacije**. Prema nasledivosti, mutacije mogu biti:

- **somatske** - zahvataju sve ćelije osim gameta.
- **gametske (germinativne)** - nastaju u gametima, pa se prenose na potomstvo.

Prema uzorku, mutacije mogu biti:

- **spontane** - nastaju slučajno, bez uočljivog delovanja nekog mutagenog faktora.
- **indukovane** - posledice dejstva poznatih ili nepoznatih fizičkih, hemijskih ili bioloških agenasa na eksponirani genetički materijal.

Prema dejstvu na fenotip, mutacije mogu biti:

- **misens** - promena jednog nukleotida dovodi do promene kodona tako da se ugradi drugačija aminokiselina u protein. To može promeniti strukturu ili funkciju proteina.
- **sinonimne** - promena nukleotida menja kodon, ali aminokiselina ostaje ista. Fenotipski obično nemaju efekta, iako ponekad mogu uticati na ekspresiju ili stabilnost RNK.
- **besmislene** - promena nukleotida stvara **stop kodon** umesto kodona za aminokiselinu, pa se prevod proteina prekida prerano. Rezultat je skraćen, obično nefunkcionalan protein.

Hromozomske mutacije su mutacije na nivou hromozoma. Genetske varijante je teško proučavati zbog složenosti i diverziteta genoma, ograničenja sekvenciranja DNK i lažno pozitivnih/negativnih rezultata.

GATK (The Genome Analysis Toolkit) je strukturirani programski okvir dizajniran da omogući brz razvoj efikasnih i robusnih alata za analizu DNK sekvencera sledeće generacije. **GATK najbolje prakse (best practices)** pružaju korak po korak preporuke za izvođenje analize otkrivanja varijanti. Postoji nekoliko različitih tokova rada GATK best practices, prilagođenih određenim aplikacijama u zavisnosti od vrste varijacije i primenjene tehnologije. Koraci u GATK tokovima:

1. **Data pre-processing** - pretprocesiranje podataka pre analize. Uključuje čišćenje i pripremu podataka (npr. uklanjanje loših sekvenci).
2. **Coverage** - procena koliko puta je svaka pozicija u genomu „pokrivena“ čitanjima. Koristi se da se proveri kvalitet i pouzdanost podataka.
3. **Pozivanje varijanti** - koriste se metode zasnovane na verovatnoći (FreeBayes, Strelka2), mašinskom učenju (DeepVariant), graficima (Dragen), kao i heurističke metode.
4. **Dodatni koraci** - filtriranje i anotacija varijanti (npr. uklanjanje lažnih pozitivnih i dodavanje informacija o genima, bolestima ili učestalosti u populaciji).

VCF (Variant Call Format) je fajl format za čuvanje pozvanih varijanti. **Frekvencija alela** računa se kao količnik broja kopija nekog alela u populaciji i ukupnog broja svih alela tog gena u populaciji. **Frekvencija alela populacije** označava učestalost određenog alela u celoj populaciji, što uključuje i referentni i alternativni alel. **Frekvencija alela varijante** označava koliko je konkretna mutacija/varijanta (npr. alternativni alel u odnosu na referentni) zastupljena u populaciji.

HaplotypeCaller je jedan od glavnih alata u okviru GATK koji se koristi za pozivanje varijanti. Radi tako što lokalno rekonstruiše moguće haplotipove i na osnovu njih odlučuje gde postoje varijante. **Haplotip** je niz varijanti koje se nalaze zajedno. Koraci:

1. **Identifikacija aktivnih regiona** - klizni prozor se pomera duž reference i meri broj neusklađenosti (SNP-ova, indela, soft clip-ova). Samo regioni sa dovoljno nepoklapanja (iznad praga) ulaze u dalju analizu.
2. **Asembliranje/sklapanje verovatnih haplotipova** - unutar aktivnih regiona pravi se graf sekvenci i pronalaze se najverovatniji haplotipovi. Zatim se ti haplotipovi ponovo poravnavaju na referentni genom (npr. Smith-Waterman algoritmom). Ovo uklanja lažne varijante i bolje opisuje indele.
    - **Haplotype calling** - dodeljivanje očitavanja odgovarajućim haplotipovima.
    - **PL vrednosti (Phred-scaled Likelihoods)** - numeričke vrednosti koje predstavljaju verovatnoću da određeni genotip objašnjava očitavanja. Što je niža PL, to je genotip verovatniji.
3. **Utvrđivanje verovatnoća haplotipova** u odnosu na očitavanja - koristi se i kvalitet baza (BQSR korigovani podaci). Izračunava se verovatnoća da očitavanje potiče od određenog haplotipa, a to se zatim prevodi u verovatnoće za svaki mogući alel ("per allele" verovatnoća).

4. **Određivanje genotipa** - na osnovu dobijenih verovatnoća određuje se najverovatniji genotip za svaki uzorak na svakoj poziciji varijante.

GVCF (Genomic VCF) je poseban format izlaza HaplotypeCallera u kojem se beleže i varijante i ne-varijantne pozicije sa verovatnoćama. **Joint calling** podrazumeva kombinovanje više GVCF fajlova (od različitih uzoraka) kako bi se zajedno odredili genotipovi. Ovo daje konzistentnije i tačnije rezultate u studijama sa više uzoraka.

Somatske varijante i tumor

Rak (tumor) predstavlja grupu bolesti koje uključuju abnormalni rast ćelija sa potencijalom invazije i širenja na druge delove tela. Rak mogu da prouzorkuju EM zračenje, hemijski agensi i slobodni radikali, genetski faktori i infekcije i virusi. Tipovi tumora:

- **karcinomi** - maligni tumori koji nastaju iz epitelnog tkiva (koža, sluzokoža, ...).
- **sarkomi** - maligni tumori koji nastaju iz vezivnog, mišićnog, koštanog ili masnog tkiva.
- **adenomi** - benigni tumori žlezdanog epitela.
- **leukemija** - maligni poremećaji krvotvornih ćelija koštane srži.

Hallmarks of cancer (obeleživači raka) su skup osnovnih bioloških osobina koje ćelije raka stiču tokom razvoja tumora. Glavne osobine:

- **Mitoza** - učestala i nekontrolisana deoba. **Antionkogeni (tumor supresorski geni)** su geni čija je normalna uloga da koče deobu ćelija, popravljaju oštećenja DNK ili podstiču ćelijsku smrt (**apoptozu**) kada je ćelija previše oštećena.
- **Mobilnost** - ćelije raka mogu da se odvoje, kreću kroz tkivo, krv i limfu, što omogućava invaziju. **Metastaza** je proces širenja ćelija raka iz primarnog tumora u udaljene organe ili tkiva.
- **Besmrtnost** - izbegavaju normalno ograničenje broja deoba. Aktiviraju **telomerazu** - enzim koji obnavlja **telomere** (krajeve hromozoma). Normalne ćelije gube telomere pri svakoj deobi i na kraju umiru, dok ćelije raka stalnim produžavanjem telomera mogu da se dele beskonačno.
- **Angiogeneza** - stimulišu stvaranje novih krvnih sudova da bi se obezbedio kiseonik i hranljive materije potrebne za dalji rast.

Heterogenost tumora opisuje razlike između tumora istog tipa kod različitih pacijenata, razlike između ćelija raka unutar jednog tumora ili razlike između primarnog tumora i sekundarnog tumora. **Tumor-only sequencing** podrazumeva da se sekvencira samo DNK tumora, bez uzorka normalnog tkiva istog pacijenta. Dobija se genetski profil tumorskih ćelija. **Tumor-only variant calling** podrazumeva da nakon sekvenciranja, softver identifikuje varijante (mutacije) u tumoru u odnosu na referentni genom. Pristup je brži i jeftiniji, ali može dati više lažnih varijanti jer ne razlikuje somatske od germinalnih mutacija. **Tumor-only pipeline**:

1. **Demultiplexing** - ako je sekvenciranje rađeno na više uzoraka, read-ovi se razdvajaju po uzorku. Kao rezultat dobijaju se FASTQ fajlovi koji sadrže read-ove za konkretne uzorke tumora.
2. **Trimming** - uklanjaju se adapter sekvence i niske kvalitativne baze sa krajeva čitanja u cilju smanjivanja greške.
3. **Alignment** - read-ovi se ravnaju na referentni genom. U ovom koraku se uklanjaju i označavaju duplikati nastali kao rezultat PCR. Takođe se vrši i **Base Quality Score Recalibration (BQSR)** koji podrazumeva da se koriguje procena kvaliteta baza da bi pozivanje varijanti bilo preciznije. Kao rezultat dobijaju se BAM fajlovi.
4. **Variant Calling** - pre samog pozivanja varijanti uklanjaju se varijante koje su verovatno nastale greškom sekvenciranja ili poravnanja (**Variant Filtering**). Nakon pozivanja varijanti dodaju se i dodatne informacije o njima (**Annotation**). Kao rezultat dobijaju se VCF fajlovi.

Tumor purity označava procenat ćelija u uzorku koje su stvarno tumorske, u odnosu na normalne (nezaražene) ćelije. Ako je uzorak "prljav" (nizak purity), mutacije tumora će biti razblažene normalnim DNK, pa ih softver može propustiti ili označiti slabije.

Mutect2 je alat koji se koristi za identifikaciju somatskih mutacija u tumorima iz NGS podataka. Može da detektuje SNV i Indel varijante. Upoređuje read-ove tumora sa referentnim genom. Kod tumor-only podataka, koristi i **baze poznatih germinalnih varijanti** (npr. gnomAD) da filtrira nasledne varijante. Procenjuje verovatnoću da je svaka varijanta somatska (nastala u tumoru) ili greška sekvenciranja. **PON (Panel of Normals)** je zbirka sekvenciranih uzoraka (VCF format) normalnog tkiva od više zdravih osoba, koja se koristi u analizi tumora. Neki regioni genoma imaju sistemske greške u sekvenciranju ili poravnanju, koje mogu izgledati kao mutacije. PON pomaže da se ove greške prepoznaju i filtriraju, jer se pojavljuju i u normalnim uzorcima, a nisu stvarne somatske mutacije.