# Primenjena bioinformatika

## Uvodni pojmovi

**Bioinformatika** je multisciplinarna oblast koja kombinuje biologiju, informatiku i statistiku za razvijanje metoda i alata za predstavljanje i analiziranje bioloških podataka. **DNK (dezoksiribonukleinska kiselina)** je nukleinska kiselina koja čuva informacije za pravilno funkcionisanje živih bića. Sadrži genetičke informacije. Sastoji se od nukleotida **adenina**, **citozina**, **timina** i **guanina**. Komplementarne baze su A = T i G = C. Nukleotidi su raspoređeni duž DNK lanca i formiraju složene sekvence. DNK se nalazi u hromozomima unutar jezgra ćelija. DNK je dvolančan. **Genom** predstavlja sekvencu nukleotida koja predstavlja celokupan genetski materijal. Sadrži sve informacije potrebne za razvoj, rast, funkcionisanje i održavanje organizma, a odgovoran je i za različite karakteristike jedinke i bolesti. **Geni** su delovi genoma, tj. segmenti DNK koji sadrže uputstva za sintezu proteina i regulaciju različitih bioloških procesa. Različite varijacije vode ka različitim karakteristikama. **Aleli** su različite varijante istog gena. **Genomika** je disciplina koja proučava ceo genom.

## Centralna dogma molekularne biologije

**Centralna dogma molekularne biologije** objašnjava tok genetičke informacije od njenog skladištenja u DNK molekulu, preko prenosa na RNK, do konačne sinteze proteina. **DNK replikacija** je proces u kom DNK pravi svoju identičnu kopiju pomoću enzima **DNK polimeraze**. **Transkripcija** je proces u kojem se genetska informacija iz DNK prepisuje u molekul RNK pomoću enzima **RNK polimeraze**. **Translacija** je proces u kojem se informacija iz RNK koristi za sintezu odgovarajućeg proteina pomoću **ribozoma**. **Kodon** je skup od 3 nukleotida koji određuje aminokiselinu. Redosled kodona u RNK određuje redosled aminokiselina u proteinskom lancu. **RNK (ribonukleinska kiselina)** sadrži **uracil** umesto timina. RNK je jednolančan.

DNK replikacija počinje kada enzim **helikaza** razdvoji dva lanca dvostruke spirale. Na svakoj novootvorenoj niti, **RNK primaza** postavlja kratak **RNK prajmer** - nukleinsko-kiselinski lanac koji služi kao početna tačka za rad **DNK polimeraze**. Na **vodećem lancu**, polimeraza dodaje nukleotide kontinuirano u smeru otvaranja spirale, dok se na **zaostajućem lancu** sinteza odvija isprekidano, u vidu kratkih delova zvanih **Okazakijevi fragmenti**, jer polimeraza može da radi samo u 5' $\rightarrow$ 3' smeru. Kasnije se ti fragmenti spajaju pomoću enzima **DNK ligaze**. Na kraju nastaju dva nova dvostruka lanca DNK, svaki sa jednim starim i jednim novim lancem.

Glavni tipovi RNK:

- **ribozomalna RNK (rRNK)** - formira bitan deo ribozoma.

- **informaciona (messenger) RNK (mRNK)** - sadrži instrukcije za polipeptidnu sintezu.
- **prenosna (transfer) RNK (tRNK)** - nosi aminokiseline do ribozoma i uklapa ih u mRNK kodiranu poruku.

**Prekursorska informaciona RNK (pre-mRNK)** je prvi proizvod procesa transkripcije i sadrži egzone i introne. Od lanca ove RNK nastaje informaciona RNK tako što se pojedini egzoni zadržavaju, a svi introni izbacuju. Svi tipovi RNK su stvoreni transkripcijom. **Alternativno isecanje (splicing)** je mogućnost da se od jednog segmenta gena dobije više različitih mRNK u zavisnosti od toga koje egzome pokupimo. RNK sekvenca koja nastaje kao rezultat transkripcije jednog gena naziva se **transkript** tog gena. Kada su geni proizveli nešto kažemo da se **ekspresovani**. Proučavanje RNK može biti kvalitativno (transkriptivna rekonstrukcija) ili kvantitativno (evaluacija različitih genskih ekspresija). **Epigenetika** je oblast koja se bavi pitanjem kako od istog genetskog materijala dobijamo različite RNK. **Euhromatin** je labavo upakovan oblik hromatina, bogat aktivnim genima jer je dostupan za transkripciju. **Heterohromatin** je gusto upakovan, uglavnom neaktivan oblik hromatina koji nije dostupan za transkripciju.

## Sekvenciranje

**Sekvenciranje** je proces određivanja tačnog redosleda nukleotida u molekulu DNK ili RNK. Pomaže u razumevanju genetičkog sastava organizma, identifikaciju gena za specifične osobine i bolesti, otkrivanje mutacija i istraživanje genetičke varijabilnosti unutar populacija. **Sekvenceri** su uređaji koji iz uzorka pljuvačke ili krvi analizom daju sekvence genoma. Izlaz nije savršen, već zahteva rekonstrukciju, tj. manje sekvence treba složiti da bi se dobio genom od početka do kraja.

**PCR (Polymerase Chain Reaction)** je lančana polimerazna reakcija kojom se jedan uzorak DNK umnožava u veliki broj kopija za kratko vreme. Zasniva se na hemijskim procesima koj se odvijaju tokom cikličnih promena temperature radi odvajanja i spajanja komplementarnih lanaca DNK. Ova metoda je jeftina. Osnovne faze PCR-a:

- **denaturacija** - razdvajanje lanca DNK. Postiže se zagrevanjem.
- **hibridizacija (aniling)** - dodavanje prajmera na 3' kraj.
- **produženje (elongacija)** - nastavljanje niza nukleotida koristeći **TAQ polimerazu** - slična kao DNK polimeraza, ali može da funkcioniše na visokoj temperaturi.

**Sanger sekvenciranje** se zasniva na tome da DNK polimeraza gradi novi lanac, ali se povremeno ubaci posebna baza (dideoksinukleotid) koja zaustavlja sintezu. Tako nastaju fragmenti različite dužine, pri čemu je svaki kraj obeležen bojom koja pokazuje na kojoj bazi se stalo. Kada se fragmenti poređaju po dužini i očitaju boje, dobija se ceo redosled DNK. **TruSeq PCR Free** je metoda pripreme DNK za sekvenciranje kod koje se izbegava umnožavanje DNK pomoću PCR-a. Time se smanjuju greške i pristrasnosti, pa se dobija precizniji prikaz stvarne sekvence. **Flow Cell** je staklena pločica u uređaju za sekvenciranje na kojoj se DNK fragmenti vežu i kopiraju. Na njoj se odvija stvarno očitavanje sekvenci, obično uz fluorescentno označavanje, što omogućava istovremeno sekvenciranje milionima

fragmenata. **Base Call Accuracy** je procenat očitanih baza koje su ispravno identifikovane u sekvenci DNK. **Phred kvalitet ($Q$)** je numerički skor koji kvantifikuje tačnost očitane baze. Što je $Q$ veći, verovatnoća greške je manja. Verovatnoća da je baza pogrešno očitana jednaka je:

$$P = 10^{-\frac{Q}{10}}$$

Odavde je base call accuracy jednak $1 - P$. Postoje i druge tehnologije sekvenciranja kao što su SMRT/ZMV Sequencing i Oxford NANOPORE Technologies.

**Genomsko sekvenciranje** je proces određivanja kompletnog redosleda nukleotida u celom genomu nekog organizma, čime se dobija potpuna genetska informacija tog organizma. Proces:

1. Izolacija DNK iz uzorka.
2. Fragmentacija na male fragmente, obično veličine od 200 do 300 baznih parova.
3. Povezivanje adaptera na krajeve.
4. Priprema sekvenijske biblioteke. Sadrži DNK fragmente spremne za sekvenciranje.
5. Primena na flowcell.
6. Generisanje klastera. Vrši se **Solid-phase PCR** koji formira klastere koji sadrže više kopija istog fragmenta. Faze:
   - **prikrečenje** - DNK fragmenti se vezuju za čvrstu podlogu (flow cell) pomoću adaptera.
   - **denaturacija** - dvostruki lanci DNK se razdvajaju u pojedinačne niti.
   - **amplifikacija** - vezane niti služe kao šabloni za PCR umnožavanje i stvaranje novih kopija.
   - **sinteza klastera** - formiraju se lokalni „klasteri“ velikog broja identičnih kopija istog fragmenta, spremni za sekvenciranje.

**Paired-end sekvenciranje** podrazumeva da se DNK fragmenti sekvenciraju sa oba kraja, stvarajući parove sekvenci. Omogućava bolje razumevanje udaljenosti između fragmenata i poboljšava tačnost rekonstrukcije.

## Asembliranje (assemby)

**Asembliranje** je povezivanje read-ova u duže sekvence kako bismo sakupljanjem i poravnanjem dobili celokupan genom. **Read** je sekvenca nukleotida određene dužine. Ako je read dužine $k$ naziva se $k$**-mer**. Pohlepan algoritam radi tako što se za svaki par read-ova računa dužina preklapanja, a zatim se dva read-a sa najvećim preklapanjem spajaju u jedan. Ovo se ponavlja sve dok ne ostane samo jedan fragment.

**Referentni genom** je reprezentativan primer genoma vrste. Često je izgrađen od genoma više individua, koji se razlikuju u nekom delovima. **Varijacije genoma** su različitosti od referentnog genoma. Referentni model pruža koordinatni sistem za komunikaciju genetskih podataka. Ljudski genom se sastoji od 23 para hromozoma sa oko 3 milijarde nukleotida.

Muškarci imaju 22 autozoma, jedan X i jedan Y hromozom, a žene 22 autozoma i dva X hromozoma. **Human genome project** je međunarodni naučni istraživački projekat tokom koga je generisana prva sekvenca ljudskog genoma. Prvobitno je ispisan u 130 knjiga, fontom 4. Trenutno se koristi **HG38**. Genom se skladišti kao niska u binarnom ili tekstualnom formatu. **Egzom** je deo genoma koji kodira proteine. Čini oko 2% genoma.

**FASTA** je tekstualni format koji skladišti sekvence molekula DNK i proteina. Koristi jednu liniju po sekvenci. Linija može biti opisna kada, sadrži ime sekvence i njen opis, ili sekventna kada sadrži samo sekvencu nukleotida. Koristi se za duže sekvence. Koriste se **IUPAC kodovi**:

- $R$ - A ili G
- $Y$ - C, T ili U
- $N$ - bilo koja baza

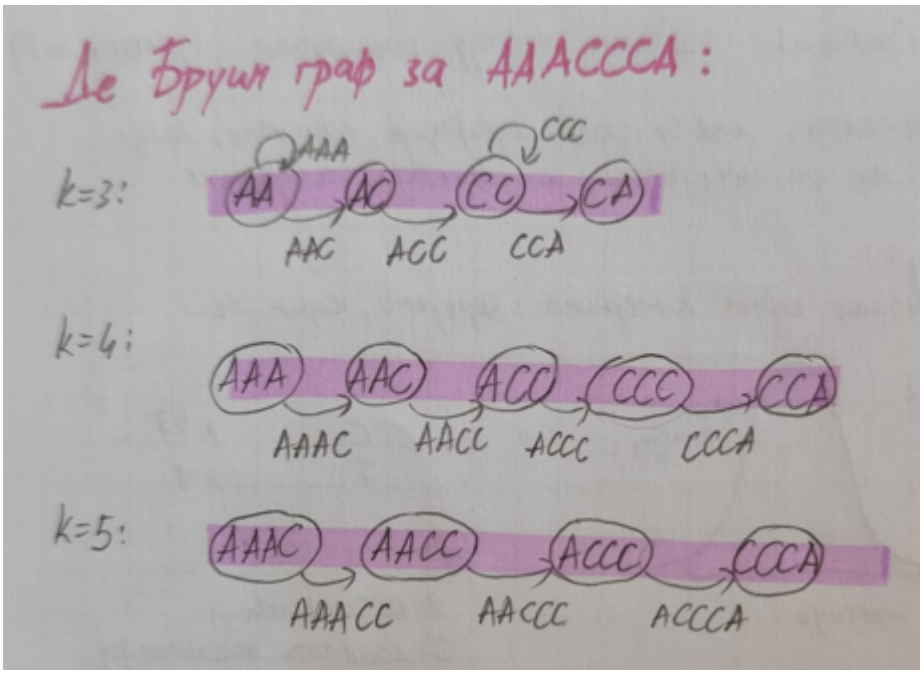
Hromozomi mogu biti **imenovani** (npr. chr1, chr2, chr3) ili **neimenovani** (npr. 1, 2, 3). **FASTA index (fai)** je fajl koji sadrži informacije o FASTA fajlu sa istim imenom. Sadrži ime sekvence, dužinu sekvence, bajt ofset prve baze, broj baza u svakoj liniji i broj bitova u svakoj liniji. **FASTQ** je tekstualni format koji dodatno skladišti i kvalitet ($Q$) sekvence. Za svaku sekvencu se koriste 4 linije. Koristi se za kraće sekvence.

**De novo whole-genome shotgun assembly** je metoda za rekonstruisanje celog genoma iz kratkih, nasumično dobijenih read-ova bez referentnog genoma. Tipičan put asembliranja:

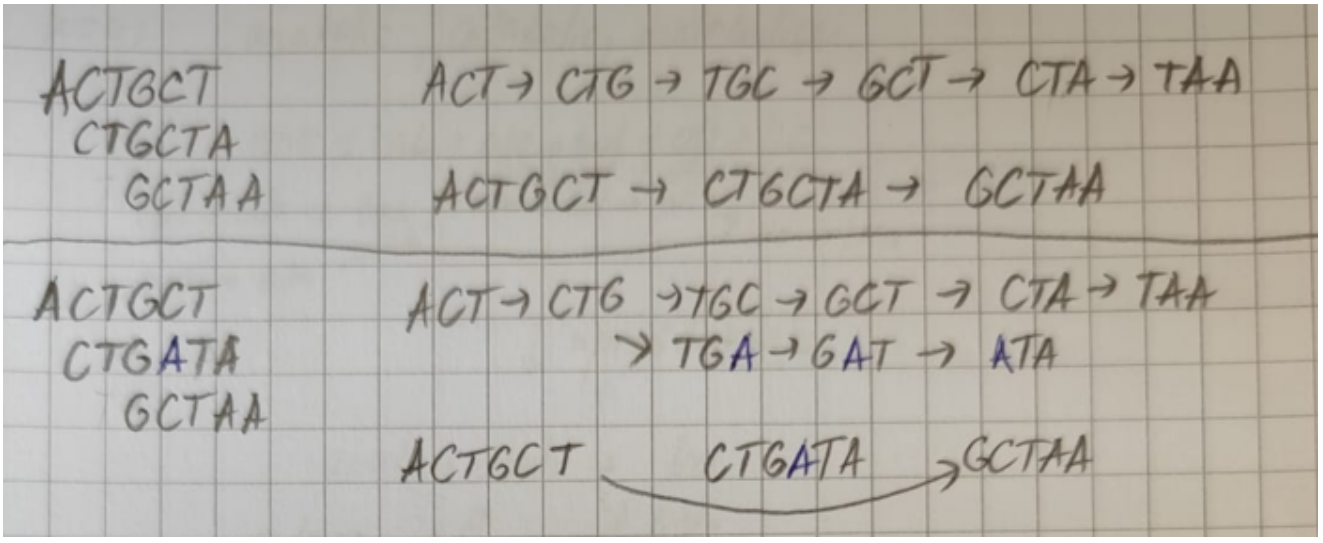
- **Popravka grešaka na sirovim read-ovima** - uklanjaju se sekvence sa greškama ili niskim kvalitetom kako bi se poboljšala preciznost asembliranja. Greške ometaju preklapanja. **Korekcija grešaka** podrazumeva odlučivanje kojim $k$-merima verujemo, a kojima ne.
- **Indeksiranje i asembl samostalnih read-ova** - read-ovi se organizuju i spajaju u kraće kontinualne sekvence (**kontige**) koristeći preklapanja.
- **Skafoldiranje uparenih read-ova** - kontigi se povezuju pomoću informacija iz paired-end read-ova ili mate-pair read-ova, formirajući duže strukture sa poznatom udaljenošću između kontiga (**skafolde**).
- **Zatvaranje rupa skafolda** - pokušavaju se popuniti prazni prostori između kontiga unutar skafolda korišćenjem dodatnih read-ova ili lokalnog asemblija, tako da se dobije što potpuniji genom.

**Grafovi** služe za predstavljanje prekapanja i sličnosti između read-ova. Mogu biti **stringovni** i **De Bruijn grafovi**. Za fiksiran prirodni broj $k$, struktura De Bruijn grafa je sledeća:

- čvorovi: svi $k-1$-meri prisutni u read-ovima.
- grane: za svaki $k$-mer $X$ prisutan u read-ovima: $k$-mer je na grani između $k-1$ prefiksa od $X$ i $k-1$ sufiksa od $X$.




Na osnovu sekvence i veličine $k$-mera, možemo jedinstveno da kreiramo De Bruijnov graf. Ne može svaki De Bruijnov graf da rekonstruiše jedinstvenu sekvencu. Stringovni graf se dobija od grafa preklapanja uklanjanjem redudantnosti. Read-ovi mogu biti redudantni i tranzitivno redudantni.




Prednosti De Bruijn grafova su brzina i jednostavnost, a ograničenja gubitak koherencije i nekonzistentnost. Stringovni grafovi hvataju celu informaciju, ali su read-ovi duži. Za odabir $k$ potrebno je balansirati. Veće $k$ dovodi do manje povezanosti, a manje $k$ do dvosmislenih puteva.

**Konstrukcija kontiga** uključuje sekvenciranje, heširanje, uprošćavanje i uklanjanje grešaka. Analizira se topologija grafa da bi se uočile potencijalne greške. Primeri:

- isecanje mrtvih krajeva (krajevi koji nisu povezani sa ostatkom grafa)
- pucanje mehura (alternativni put iz jednog čvora koji sa kasnije spaja na glavnu putanju)
- isecanje himeričkih read-ova (spajaju inače nepovezane delove)

**Konstrukcija skafolda** koristi uparene informacije i biblioteke koje sadrže višestruke veličine ubacivanja. Zatvaraju se i rupe između kontiga.

Ne postoje trivijalne **metrike** za asembli. **Metrike bez referenci** su broj kontiga/skafolda, dužina asembla, dužina najdužeg kontiga/skafolda, procenat rupa u skafoldima, $Nx$ i $NGx$ kontiga/skafolda. **Metrike sa referencom** su pokrivenost i greške asembla. $Nx$ je najveća dužina kontiga koja pokriva $x\%$ totalne dužine asembla. Izražava se u procentima. $N50$ je medijana. $NGx$ je najveća dužina kontiga koji pokriva $x\%$ totalne dužine genoma. Ove metrike računamo tako što sortiramo kontige i tražimo odgovarajući.

## Poravnanje (alignment)

**Alignment (poravnanje)** je proces usklađivanja dve ili više bioloških sekvenci (DNK, RNK ili proteina) kako bi se identifikovale sličnosti i razlike. **Naivni algoritam (gruba sila)** radi tako što na svakom mogućem offset-u u referenci proveravamo da li se svi karakteri poklapaju. Ovo je preskupo jer za referencu dužine $n$ i read dužine $m$ imamo $(n - m + 1) \cdot m$ = $O(nm)$ provera.

**Niz sufiksa** je tehnika koja podrazumeva da se referenca rasparčna u $n$ sufiksa koje sortiramo azbučno (alfabetno). Na pronalaženje onda možemo da koristimo binarnu pretragu. Ovaj pristup je komplikovaniji i zauzima više memorije, ali je brži jer se izvršava u $O(m \log_2 n)$ provera.

**Heš mapa** je tehnika koja za ideju ima da napravimo tabelu sa svakim $k$-merom u referenci. Broj mogućih sekvenci dužine $k$ je $4^k$. Možemo koristiti **heš funkciju** da smanjimo domen: $[0,\ R] \to [0,\ H]$. Na primer:

$$hash(n) = H\frac{n}{R} \text{ ili } hash(n) = n\ \%\ H, \text{ gde je } n \text{ broj sekvence}$$

Konstrukcija sada zahteva vreme $O(n)$, a traženje $O(1)$.

**Burrows-Wheeler Transformation (BWT)** je tehnika koja permutuje ulazni string u novi string, pogodniji za kompresiju i pretragu. Na nisku dodajemo karakter $, pravimo sve moguće rotacije i sortiramo ih. BWT string dobijemo tako što uzmemo poslednji karakter svake od rotacija. Ima osobinu čuvanja ranka, što je potrebno za LF mapiranje. Ovaj algoritam je reversibilan - krenemo od prvog reda i primenimo **LF (last-first) mapiranje**: LF $(i)$ znači "koji karakter u prvom redu odgovara karakteru na poziciji $i$ u poslednjem redu?"

**FM Index** je tehnika koja pronalazi sva ponavljanja šeme $P$ u tekstu $T$ koristeći BWT($T$), odnosno traži redove u BWT($T$) sa $P$ kao prefiksom. Uradimo to za $P$-ov najveći sufiks i onda ga povećavamo. LF($(i,\ q_c)$) određuje rank karaktera $c$ u redu $i$. Pozicije u originalnom tekstu se određuju korišćenjem niza sufiksa.

**Alajnovanje bazirano na skoru (oceni)** podrazumeva da se read-ovi postavljaju tako da skor bude maksimalan. Veliki broj read-ova se neće savršeno alajnovati sa referencom usled genetskih varijacija ili grešaka sekvenciranja. Računamo ocenu baziranu na udaljenosti. Izvršava se u dva koraka:

- **seed** - pronalaze se kratka podudaranja između read-a i reference.

- **extend** - podudaranja se proširuju kako bi se dobilo celokupno poravnanje.

**Binary Alignment Map (BAM)** je standardizovan format za čuvanje alajnovanih read-ova. Predstavlja binarnu verziju **Sequence Alignment Map (SAM)** fajla koji sadrži plain text. Sadrži red sekvence i kvalitet, poziciju (hromozom i prvo poklapanje), CIGAR nisku, flagove, poziciju para read-ova i opcione tagove. Pozicije u SAM kreću od 1, a u BAM od 0. Read-ovi mogu biti sortirani po poziciji read-a ili po imenu read-a. Koordinate mogu biti indeksirane, za brz pristup (fajlovi sa ekstenzijom **.bai**).

**Smith-Waterman aligner** predstavlja algoritam dinamičkog programiranja za lokalno alajnovanje. Matrica se popunjava prema formuli:

$$\mathrm{Score}_{i,\,j} = \max \begin{cases} 0 \\ \mathrm{Score}_{i-1,\,j-1} + M \\ \mathrm{Score}_{i-1,j} - G \\ \mathrm{Score}_{i,j-1} - G \end{cases},$$

gde je $M$ match score ako se poklapaju, a inače mismatch penalty ($\pm 1$) i $G$ gap penalty. Smerovi u matrici prate putanju optimalnog poravnanja: dijagonalno je pogodak (match ili mismatch), na desno delecija (gap u read-u), a na dole insercija (gap u referenci).

**CIGAR niske** su kompaktan zapis poravnanja, koristeći run-length encoding. Kodovi:

- **M** - match/mismatch
- **I** - insertion
- **D** - deletion
- **S** - soft clip (deo koji nije alajnovan)

## Varijante

**Genomske varijante** su razlike u DNK sekvenci između genoma različitih jedinki iste vrste. Najčešće varijante:

- **Single Nucleotide Variant (SNV)** - promena jednog nukleotida. Jednostavna promena ali može da dovede do značajnih promena (npr. daltonizam). Mogu biti **bezopasne (silent)** koje nemaju efekat na jedinku ili **opasne (missense i nonsense)** koje dovode do promena. Missense menja kodon tako da on kodira drugu amino kiselinu. Nonsense pretvara kodon u neki stop kodon čime se prevremeno prekida translacija.
- **Indel (insercije i delecije)** - umetanje ili brisanje baze ili do 50 baza.
- **Strukturne varijante** - veće promene poput **duplikacije** (ponavljanje dela), delecije i insercije većeg segmenta, **inverzije** (deo DNK je obrnut) i **translokacije** (deo DNK premešten na drugo mesto). **Balansirane** SV ne menjaju dužinu genoma. To su inverzije i translokacije. **Nebelansirane** SV menjaju dužinu i to su insercije, delecije i duplikacije. Ukoliko za SV detekciju koristimo veće read-ove imaćemo veće prostiranje, tj. možemo da vidimo celu varijantu. Sa druge strane, greške su veće.

**Variant calling** je proces pronalaženja razlika između reference genoma i posmatrane sekvence. Potrebni su nam alajnovani read-ovi po referentnom genomu da bismo mogli da nađemo (pozovemo) varijante. **Pileup** je tekstualni format koristi u bioinformatici da bi se opisalo usklađivanje sekvenci sa referentnim genomom. Svaka linija u fajlu odgovara jednoj bazi u referentnom genomu. Sadrži informaciju o nazivu hromozoma, poziciju u genomu, bazu u referentnom genomu, broj alajnovanih read-ova, same baze iz tih read-ova i njihov kvalitet. Očitavanja koja se poklapaju sa referencom su **REF**. Očitavanja koja podržavaju varijantu (alt sekvencu) su **ALT**. **Dubina (depth)** je broj read-ova koji pokrivaju tu poziciju, tj. REF+ALT. **Učestalost varijantnog alela (Variant Allele Frequency)** je **VAF** = $\frac{\text{ALT}}{\text{REF+ALT}}$.

Ljudski genom je **diploidni** - dva alela za svaki gen. Varijante:

- **0/0 homozigot** - oba alela se poklapaju sa referencom.
- **0/1 heterozigot** - jedan alel se poklapa, a drugi ne.
- **1/1 homozigot** - oba alela se ne poklapaju i isti su.
- **1/2 heterozigot** - oba alela se ne poklapaju, ali nisu isti.

Uvek pretpostavljamo diploidnost. Ako imamo više od 2 različita slova, uzimamo 2 najčešća. **VCP** je format za čuvanje pozvanih varijanti (nakon variant calling-a). Sadrži informaciju o hromozomu, poziciji varijante, ID varijante, referentnu i alternativnu bazu, kvalitet i ostale informacije.

**Normalizacija** pretvara brojeve očitanih read-ova u vrednosti koje omogućavaju pravično poređenje ekspresije gena unutar uzorka i između različitih uzoraka. Metode unutar jednog uzorka su **RPKM (Reads per kilobase million)**, **FPKM (Fragments per kilobase million)** i **TPM (Transcripts per million)**. Metode za poređenje između više uzoraka su **TMM** i **DESeq**. **Diferencijalna ekspresija** je proces gde su različiti geni aktivirani u ćeliji, dajući ćeliji specifičnu funkciju.

**Mere statističkog značaja** su pokazatelji koji nam govore da li je posmatrana razlika stvarna ili je nastala slučajno. Neke od mera su:

- **null hypothesis** - pretpostavka da nema stvarne razlike između posmatranih grupa ili uslova.
- $p$**-value** - verovatnoća da bi se dobio rezultat jednak ili ekstremniji od posmatranog, pod pretpostavkom da je nulta hipoteza tačna.
- **alternative hypothesis** - pretpostavka da postoji stvarna razlika između posmatranih grupa ili uslova.

## Pitanja

**Gde se nalazi DNK u ćeliji kod eukariota?**

Nalazi se u jedru.

**Objasniti razliku između transkripcije i translacije.**

Procesom transkripcije se od DNK sintetiše RNK i to se vrši u jedru. Procesom translacije se od RNK sintetišu proteini i to se vrši u citoplazmi.

**Objasniti pojam alel.**

Aleli su različiti oblici jednog istog gena. Mogu biti dominantni ili recesivni. Proizvode varijacije u određenim naslednim karakteristikama kao što je boja očiju. Na primer, jedan alel je plave boje, a drugi smeđe.

**Šta je prajmer i koju ulogu ima u procesu sekvenciranja?**

Prajmer je niz nukleinskih kiselina koji služi kao početna tačka za sintezu DNK. Prajmeri su neophodni jer enzimi koji vrše replikaciju (DNK polimeraza) mogu samo da dodaju nove nukleotide na već postojeći lanac.

**PCR ciklus se sastoji od sledeća 3 procesa: ?, ?, ?.**

1. denaturacija - razdvajanje lanaca DNK
2. hibridizacija (aniling) - dodavanje prajmera na 3' krajeve lanca
3. produženje (elongacija) - nastavljanje niza nukleotida pomoću DNK polimeraze

**Ako sekvencirana baza ima Phred kvalitet 10, kolika je verovatnoća da je ta baza pogrešno očitana?**

$$P = 10^{-\frac{Q}{10}} = 10^{-\frac{10}{10}} = 10^{-1} = \frac{1}{10} = 10\%$$

**Kako je nastao referentni ljudski genom?**

Referentni genom je izgrađen od više različitih individua u okviru projekta "Human Genome Project", gde je ispisan u 130 knjiga.

**Ljudski genom se sastoji od oko ? baznih parova.**

3 milijarde

**Koja je razlika između FASTA i FASTQ formata?**

FASTA je tekstualni format koji skladišti sekvence molekula DNK i proteine. Koristi jednu liniju po sekvenci. Koristi se za duže sekvence. FASTQ je tekstualni format koji dodatno skladišti i kvalitet sekvence. Za svaku sekvencu se koriste 4 linije. Koristi se za kraće sekvence.

**Šta je asembly? Šta je alignment?**

Assembly predstavljanja sklapanje manjih sekvenci pomoću preklapanja delova sa ciljem sklapanja celog genoma ili neke korisne informacije. Alignment predstavlja sklapanje sekvenci uz pomoć referentnog ljudskog genoma.

**Koja je razlika između De Bruijn i string grafa?**

De Bruijn grafovi se koriste za kratke sekvence. Čvor predstavlja $(k-1)$-mere koji su prisutni u read-u. Za svaki $k$-mer $X$ prisutan u read-ovima: $k$-mer je na grani između $k-1$ prefiksa od $X$ i $k-1$ sufiksa od $X$. String grafovi se koriste za duže sekvence. Čvorovi su read-ovi koji ne moraju biti iste dužine, a grane između čvorova postoje ukoliko postoji preklapanje u tim read-ovima.

**Nacrtati De Bruijn graf za sekvencu "AAACCCA" za $k = 4$.**

$$AAA \xrightarrow{AAAC} AAC \xrightarrow{AACC} ACC \xrightarrow{ACCC} CCC \xrightarrow{CCCA} CCA$$

**Koliko se poređenja izvrši u najgorem slučaju naivnog algoritma za pretragu niski gde je referenca dužine $n$, a read dužine $m$.**

Imamo $(n - m + 1) \cdot m$ poređenja, pa je složenost $O(nm)$.

**Pored naivnog, koji algoritam za pretragu niski se još često koristi?**

Smith-Waterman aligner.

**Navesti nekoliko (bar tri) tipa struktura podataka koje se često koriste kod modernog pristupa mapiranja genoma.**

- Niz sufiksa
- Stabla sufiksa
- Heš mape
- FM index

**Navesti četiri tipa varijanti.**

- delecije
- insercije
- inverzije
- duplikacije
- translokacije

**Koji od navedenih tipova uticaja varijanti je uglavnom neškodljiv: a. Silent c. Nonsense b. Missense d. Frameshift**

a. Silent

**Izračunati VAF (Variant Allele Frequency) ako je pokrivenost (Depth of Coverage) 30 i postoji 10 varijanti koje podržavaju REF i 20 varijanti koje podržavaju ALT.**

$$\mathrm{VAF} = \frac{\mathrm{ALT}}{\mathrm{REF} + \mathrm{ALT}} = \frac{\mathrm{ALT}}{\mathrm{depth}} = \frac{20}{30} = 0.67$$

**Odrediti genotipe (0/0, 0/1, 1/1, 1/2) u sledećim primerima:**

**a) REF: A, A A A A A A**

**b) REF: A, C C C G G G**

**c) REF: A, T T T T T T**

**d) REF: A, A A A T T T**

a) 0/0 - oba se poklapaju sa REF
b) 1/2 - oba se ne poklapaju i različiti su
c) 1/1 - oba se ne poklapaju i isti su
d) 0/1 - jeda se poklapa, a drugi ne

**Zaokružiti nebalansirane (unbalanced) strukturne varijante (balansirane strukturne varijante ne menjaju dužinu genoma): a. insercije b. alele c. translokacije d. pirimidini e. delecije f. izoformi g. inverzije h. duplikacije**

a) insercije
e) delecije
h) duplikacije

**Dat je referentni genom (REF) i sekvencirani genom (OBS). Sekvenciranje je urađeno tehnikom uparenih read-ova (paired-end sequencing) gde je dužina svakog read-a 150 baznih parova, a prosečna veličina fragmenta je 600 baznih parova. Urađeno je mapiranje sekvenciranih read-ova na referentni genom. Grafički prikazati izgled mapiranja read-ova na referencu u oblasti gde se u sekvenciranom genomu nalazi delecija veličine 200 baznih parova. Prikazati read pair, read depth i split read signale. Da li je iz svakog od ovih signala ponaosob moguće odrediti koja je strukturna varijanta prisutna u sekvenciranom genomu?**

?????

**Razlika izmedju DNK i RNK molekula?**

DNK (dezoksiribonukleinska kiselina) sastoji se od dva duga polinukleotidna lanca građena od deoksinukleotida. Sadrži baze adenin, timin, citozin i guanin. Sadrži sećer dezoksiribozu. RNK (ribonukleinska kiselina) sastoji se od jednog dugog polinukleitidnog lanca. Sadrži baze adenin, uracil, citozin i guanin. Sadrži sećer ribozu.

**Prekursorska informaciona RNK (eng. pre-mRNA) je prvi proizvod procesa transkripcije i sadrži ? i ?. Od lanca ove RNK nastaje informaciona RNK (eng. mRNA) tako što se pojedini ? zadržavaju, a svi ? izbacuju.**

Sadrži egzone i introne. Zadržavaju se pojedini egzoni, a izbacuju svi introni.

**RNK sekvenca koja nastaje kao rezultat transkripcije jednog gena naziva se ? tog gena.**

Transkript.

**Šta je alternativno isecanje (eng. alternative splicing)?**

Alternativno isecanje (splicing) je mogućnost da se od jednog segmenta gena dobije više različitih mRNK u zavisnosti od toga koje egzome pokupimo.

**Navesti dve metode normalizacije koje se koriste u RNA-Seq eksperimentima (normalizuju u odnosu na dužinu transkripta i ukupnog broja sekvenciranih ridova).**

- RPKM
- FPKM
- TPM
- TMM
- DESeq

**Prilikom statističkog testiranja hipoteza, verovatnoća da se pri uslovu nulte hipoteze dobiju još ekstremniji rezultati naziva se ?.**

$p$-vrednost.

**Metod normalizacije koji prvo normalizuje u odnosu na broj sekvenciranih ridova a zatim u odnosu na dužinu transkripta naziva se ?.**

TPM.

**Šta je rak?**

Rak je bolest u kojoj se neke telesne ćelije nekontrolisano dele i nekontrolisano rastu i šire na druge delove tela.

**Šta su tumor-supresorski geni?**

Tumor-supresorski geni su geni za supresiju tumora i oni regulišu ćeliju tokom deobe i replikacije. Paze da se ćelija samouništi nakon određenog broja deoba.

**Zašto čistoća tumora (eng. tumor purity) predstavlja problem kod pronalaženja varijanti?**

U uzorku tumora imamo i zdrave ćelije, pa nam je bitan odnos tumorskih i zdravih ćelija.