DNA aflezen met NGS

Antwoorden oefenvragen en opdrachten

H2

Vraag 2.1

Voorbeelden van belangrijke overwegingen zijn:

- read lengte
- prijs
- throughput (aantal reads dat wordt geproduceerd)
- accuraatheid (aantal fouten)
- hoeveelheid input materiaal dat nodig is
- andere antwoorden zijn ook mogelijk

Vraag 2.2

- Illumina: verschillende inputmoleculen worden geïmmobiliseerd op het glas van een flowcell. Deze worden vermeerderd met behulp van bridge-PCR en afgelezen.
- Ion Torrent: verschillende inputmoleculen worden geïmmobiliseerd op het oppervlak van magnetische beads. Deze worden vermeerderd met behulp van PCR en afgelezen.
- PacBio: verschillende inputmoleculen worden samen met een DNA-polymerase geïmmobiliseerd op de bodem van een gaatje in een zero-mode waveguide. Het polymerase maakt een nieuwe streng en dat proces wordt gefilmd.
- Oxford Nanopore: verschillende inputmoleculen worden elk verbonden met een motoreiwit. Dit complex bindt aan het transmembraaneiwit dat door de membraan steekt (de nanopore). Het DNA-molecuul wordt door de nanopore geduwd en de verstoringen die dat geeft in het elektrische veld dat over de membraan staat, wordt gemeten.

H3

Vraag 3.1

De gebruikte tekens in de Ascii tabel beginnen bij 33 (dat is de '!'). Om de Phred score die bij een teken hoort te bepalen, moet je dus 33 van de waarde in de Ascii tabel aftrekken. Omgekeerd moet je 33 optellen bij de Phred score om het juiste teken te achterhalen. In dit geval 43 + 33 = 76. Dit staat gelijk aan het teken ‘L’. Het gaat dus om positie 5.

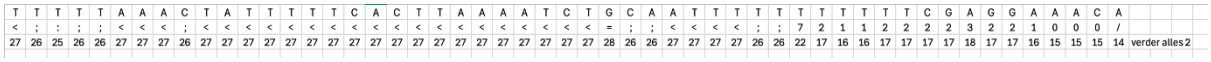
Vraag 3.2

a. De eerste grafiek laat zien dat de kwaliteit van de reads op alle posities erg laag is. Na positie 10 stijgt de kwaliteit wel enigszins, maar blijft onder de 20. Goede kwaliteit Illumina data zitten in het groene gebied, boven de 28.
   In de tweede grafiek is te zien dat de verdeling van de verschillende basen in de eerste tien posities van de reads sterk afwijkt van die bij de rest van de reads.
b. De kwaliteit van de reads is niet te verbeteren. Het selecteren van hoge kwaliteit reads is ook geen optie, want alles zit onder de 20. Bij analyse moet je dus hopen dat er genoeg coverage is om de fouten te verbeteren. Wel kan het nuttig zijn om de eerste tien basen van alle reads te verwijderen.

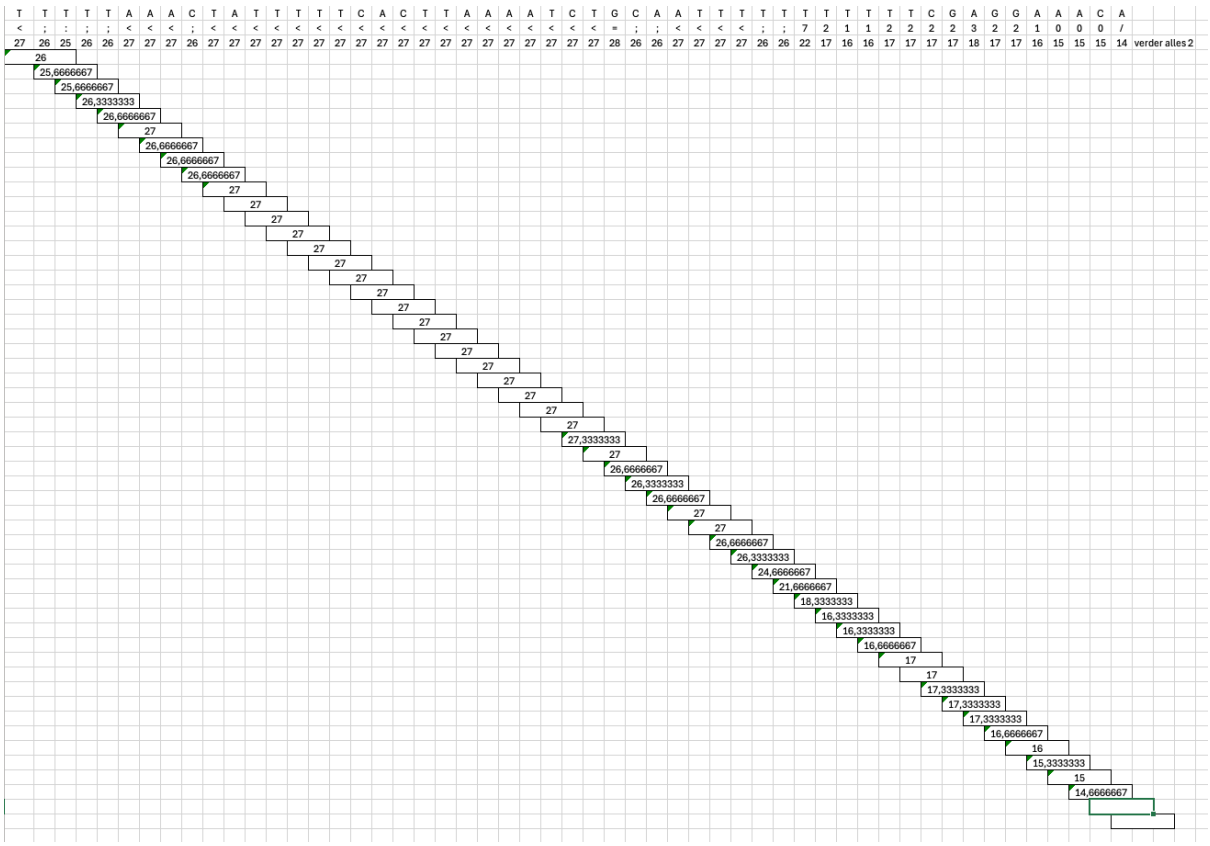
Opdracht 3.3

Ga bij deze vraag uit van Phred +33.

Stap 1: Reken de tekens om naar Phred scores met behulp van de Ascii tabel.



Stap 2: Bereken de gemiddelden van alle windows van 3.



Stap 3: Zoek waar de waarde onder de 15 komt en trim vanaf daar. In dit geval dus geen trimming aan de voorkant. Trim aan de achterkant alles vanaf positie 49 (de `A` na `GAGGA`).

H4

Opdracht 4.1

Stap 1: Bepaal alle rotaties van de sequentie `CGATGGCTAA$`

```
CGATGGCTAA$
$CGATGGCTAA
A$CGATGGCTA
AA$CGATGGCT
TAA$CGATGGC
CTAA$CGATGG
GCTAA$CGATG
GGCTAA$CGAT
TGGCTAA$CGA
ATGGCTAA$CG
GATGGCTAA$C
```

Stap 2. Sorteer op alfabetische volgorde. Dit is de LF-matrix.

```
$CGATGGCTAA
A$CGATGGCTA
AA$CGATGGCT
ATGGCTAA$CG
CGATGGCTAA$
CTAA$CGATGG
GATGGCTAA$C
GCTAA$CGATG
GGCTAA$CGAT
TAA$CGATGGC
TGGCTAA$CGA
```

De BWT is de laatste kolom: `AATG$GCGTCA`

Om `GGC` te alignen:
Zoek `C` op in de linkerkolom => regels 5 en 6
Kijk welke van deze twee regels eindigt op `G` => regel 6
Dit is enige mogelijkheid, dus positie is gevonden.

Opdracht 4.2

a. Positienummers:

```
CGATGGCTAATGGTGGTAA
00000000011111111111 (tientallen)
12345678901234567890
```

Hash tabel:

```
AAT  9
ATG  3,10
CGA  1
CTA  7
GAT  2
GCT  6
GGC  5
GGT  12,15
GTA  16
GTG  13
TAA  8,17
TGG  4,11,14
```

b. Goede voorbeelden: alle k-meren die maar één keer in de tabel voorkomen. Die beperken de zoekruimte het meest.

Opdracht 4.3

Stap 1: Zet de sequenties in een matrix.

	-	T	A	G	C	T	G	T	G	C	T	A	G
-													
G													
C													
T													
G													
A													
G													

Stap 2: Vul de velden in met de alignment scores.

	–	T	A	G	C	T	G	T	G	C	T	A	G
–	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	5	0	0	5	0	5	0	0	0	5
C	0	0	0	0	10	5	0	5	0	10	5	0	0
T	0	5	0	0	5	15	0	5	0	5	15	0	0
G	0	0	5	5	0	10	20	15	10	5	10	15	5
A	0	0	5	5	1	5	15	16	15	10	5	15	15
G	0	0	0	10	5	0	10	11	21	16	11	10	20

Stap 3: Bepaal de optimale alignment via de traceback.

	–	T	A	G	C	T	G	T	G	C	T	A	G
–	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	5	0	0	5	0	5	0	0	0	5
C	0	0	0	0	10	5	0	5	0	10	5	0	0
T	0	5	0	0	5	15	0	5	0	5	15	0	0
G	0	0	5	5	0	10	20	15	10	5	10	15	5
A	0	0	5	5	1	5	15	16	15	10	5	15	15
G	0	0	0	10	5	0	10	11	21	16	11	10	20

Stap 4: Bepaal de lokale alignment.

```
GCTGTG
|||| |
GCTGAG
```

H5

Vraag 5.1

a. Illumina. Deze genereert veel en hoge kwaliteit. Daarmee zul je dus SNPs kunnen detecteren met hoge zekerheid.
b. Illumina. Deze genereert veel en hoge kwaliteit. De reads zijn lang genoeg om korte indels te detecteren. Daarmee zul je dus korte indels kunnen detecteren met hoge zekerheid.
c. PacBio en Oxford Nanopore. De lange reads zullen vaker zowel de sequentie voor de inversie als de sequentie erachter bevatten, zodat de inversie precies in kaart te brengen is. Sequence foutjes zijn hierbij minder van belang.
d. Illumina, want daarmee kun je nauwkeurig naar verschillen in coverage kijken.

Vraag 5.2

a. Insertie. De alternatieve allelen (`AACTC,ACTC`) zijn langer dan de referentiesequentie (`A`).
b. Twee (`AACTC,ACTC`).
c. 26 (DP=26).
d. Heterozygoot (AF=0.5)

H6

Opdracht 6.1

1. Alignment

```
ATGTCT
 TGTCAGAAT
  GTCTGAATCACAC
        ATCACACA
            CACATCCAG
                 CCAGTAAG
                     TAAGTCACA
                      AAGTCACACA
            CACAGGGCAC
              CAGGGCACCATT
```

2. Graaf

```
ATGTCT -> TGTCAGAAT -> GTCTGAATCACAC -> ATCACACA => split
tak 1 => CACATCCAG -> CCAGTAAG -> TAAGTCACA -> AAGTCACACA
tak 2 => CACAGGGCAC -> CAGGGCACCATT
```

3. Twee mogelijkheden: eerste stuk + tak 1 en eerste stuk + tak 2. Let op dat tak 2 ook aan het einde van tak 1 kan alignen (met `CACA`). Zoals bij 1. weergegeven, zijn de onderste twee reads gealignd met reads 1-4 en *niet* met reads 5-8.

4. Contigs

```
1.ATGTCTGAATCACACA
2.CACACATCCAGTAAGTCACACA
3.CACAGGGCACCATT
```

Opdracht 6.2

a. Totale lengte van de reads = 95. De helft daarvan is 95/2 = 47,5.
   Je moet de langste vijf reads bij elkaar optellen om tot 47,5 te komen.
   De kortste van die vijf is 10 basen lang.
   De n50 is dus 10.
b. De drie contigs zijn samen 51 lang.
   De helft daarvan is 25,5.
   Je moet de twee langste reads optellen om tot 25,5 te komen.
   De kortste van die twee reads is 17.
   De n50 = 17.

Opdracht 6.3

De volgorde is 1-2-3

```
ATGTCTGAATCACACATCCAGTAAGTCACACAGGGCACCATT
----------********--*********************--**--
```

H7

Vraag 7.1

a. Omdat je probes of primers moet ontwerpen.
b. Omdat je vaak geïnteresseerd bent in kleine variaties, zoals SNPs. Ook met dergelijke kleine verschillen zullen de probes nog steeds binden. Bij PCR binden de primers aan weerszijden van het stuk waar je de variatie verwacht.

Vraag 7.2

De transcriptiefactor bindt aan de regio aan het begin van het gen of er net voor. Daar ligt de promotor regio. Dit betekent dat het gen hoog tot expressie zal komen. De regio aan het begin van het gen is geacetyleerd, wat betekent dat het DNA toegankelijk is op die plek. Deze regio zal dus hoog tot expressie komen onder de condities waaronder dit experiment is uitgevoerd.

H8

Vraag 8.1

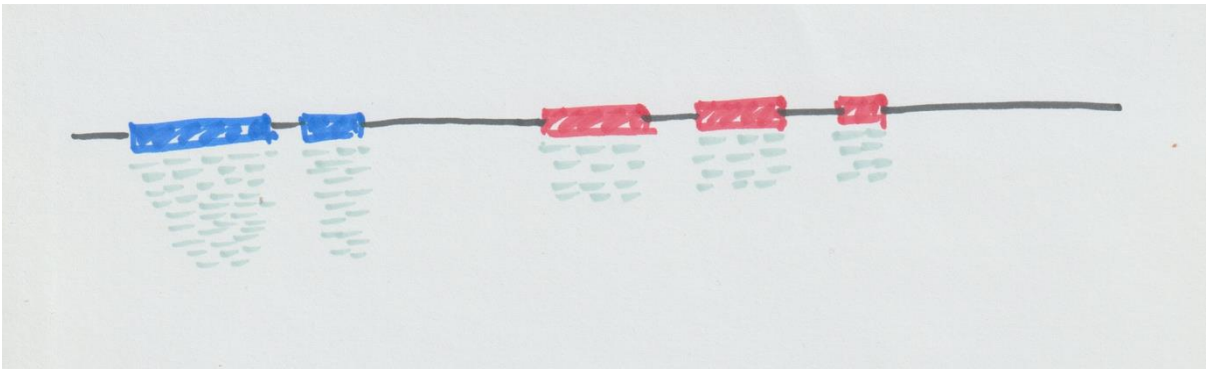
Korte reads kunnen over de grens tussen twee exonen gaan en dus laten zien welke twee exonen naast elkaar liggen op het mRNA. Ook paired-end en mate-pair informatie kunnen helpen te bepalen welke exonen naast elkaar op het transcript liggen.

Vraag 8.2

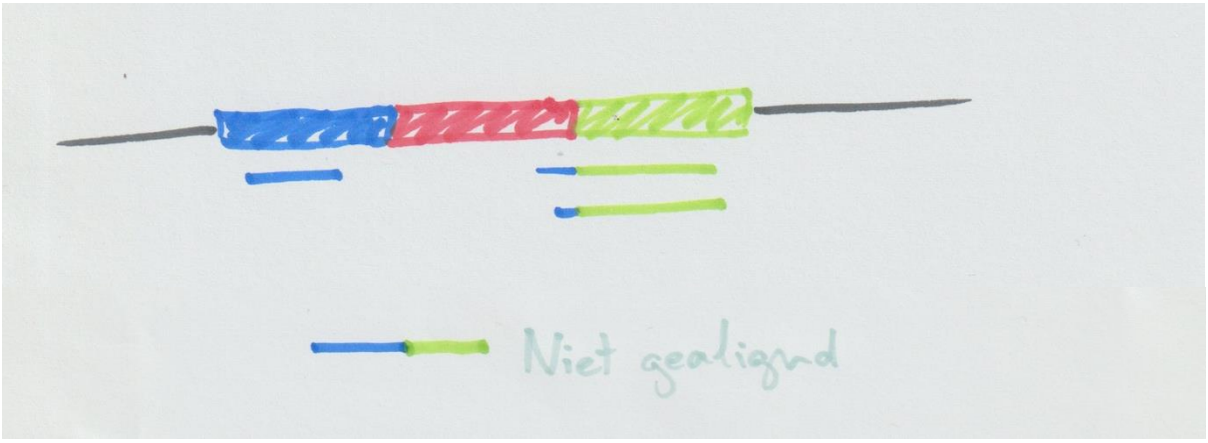
a. 50
b. 1

H9

Opdracht 9.1



Opdracht 9.2



H10

Vraag 10.1

De sequentie van veel kleine RNAs is zo kort dat ze vaak op veel verschillende plekken in het genoom kunnen alignen. Bovendien hoeft de match vaak niet 100% te zijn om toch effect te hebben. Dat laatste vergroot het aantal mogelijke bindingsplaatsen nog meer. Het is niet meteen duidelijk welke van al deze mogelijke bindingsplaatsen relevant zijn.

Vraag 10.2

- Transcriptiefactoren die de expressie van het gen zouden bevorderen, zijn niet of weinig aanwezig in de onderzochte cellen.
- Het chromatine zou ontoegankelijk kunnen zijn op de plaats waar het gen ligt. Dat zou ervoor zorgen dat transcriptiefactoren en RNA-polymerase er niet goed bij kunnen en er dus weinig mRNA van het gen afgeschreven wordt.
- Het gen zou gemethyleerd kunnen zijn. Dat zou ervoor zorgen dat het minder tot expressie komt.

H11

Vraag 11.1

Door gebruik te maken van DNA-markers. De onderzoekster is geïnteresseerd in een specifieke groep organismen (vissen), dus kan een marker gebruiken die alleen op vissen werkt. Door de amplicons met NGS te sequencen, kunnen ook soorten met maar weinig DNA in het sample toch gedetecteerd worden.

Opdracht 11.2

a. Die reads hebben geen match in de database waarmee de data is vergeleken. Het zou kunnen dat de database niet volledig is en de gezochte soorten er niet in staan. Dat kan bijvoorbeeld als de database is opgebouwd met alleen bepaalde organismen. Als de database bijvoorbeeld alleen sequenties van bacteriën en schimmels bevat, zullen de reads van de alg niet gevonden worden. Het is ook mogelijk dat van sommige soorten in het korstmos geen sequenties beschikbaar zijn. Een andere mogelijkheid is dat de betreffende reads te veel foutjes bevatten en daarom niet in de database teruggevonden kunnen worden.
b. Alle reads die met een laag taxonomisch niveau een match hebben (bijv. soort), hebben sowieso ook een match met alle niveaus daarboven (bijv. familie).
c. Er zijn geen reads die met verschillende soorten binnen het *Evernia* genus matchen. Het zou kunnen dat *Evernia prunastri* de enige soort uit het genus *Evernia* is die in de database staat.