SARS-COV-2 raðgreining

Elías Bjartur Einarsson - ebe19@hi.is
Þórhallur Auður Helgason - thh114@hi.is

Verkefnið fólst í að mynda samsemdarstreng úr gríðarstóru safni raðgreindra búta af erfðaefni SARS-COV-2. Viðmiðunarstrengur var gefinn en handfylli stökkbreytinga skildu strengina tvo að. Bútarnir voru af flestir stærðargráðunni 150 niturbasar þó einhverjir voru nær því að vera 40-50 nitubasar. Markmið vinnunnar var að skila nákvæmri staðsetningu þessara stökkbreytinga og hvernig niturbasar breyttust milli samsemdarstrengsins og viðmiðunarstrengsins.

Til að byrja með var viðmiðunarerfðamengið forunnið í orðasafn (e. dictionary) 21-mera á þann hátt að lyklarnir í safninu voru allir þeir basabútar af lengd 21 sem finnast í viðmiðinu og gildin voru listar staðsetninga þeirra í viðmiðinu.

Bútastærðin 21 var valin strax í upphafi. Það er u.þ.b. 15% af dæmigerðri stærð bútanna í safninu, sem skilur eftir 85% bútsins til að ganga úr skugga um hvort mátið gangi á hverjum stað. 21 basaröð er ekki það stutt að hún passi víða en þó ekki svo löng að hún geti innihaldið mörg frávik milli strengjanna og þannig passað hvergi nákvæmlega í heild sinni. Einnig hafði Páll sagt að 21 væri góð tala.

Því næst tókum við hvern lestursbút (kallaður 'lest' í kóðanum), skoðuðum alla 21-mera hans og athuguðum hvar í viðmiðinu þeir fundust og hvaða svæði það benti til að þessi lestur ætti að vera á. Bútarnir gátu verið úr erfðaefni vírusins (sem er einstrent RNA) eða strengir sem mátuðust við það. Því voru báðar myndir hvers 21-mers skoðaðar - fyrst sá sem lesinn var og því næst mótstrengurinn í öfuga stefnu. Fyrsti 21-mer bútsins var þannig lesinn og allar staðsetningar þar sem hann passaði merktar í safn. Þá var næsti lesinn en við merkingu á staðsetningum hans var tekið tillit til að hann ætti að koma á eftir þeim fyrsta. Því var staðsetningin bakfærð um einn. Fyrir þann þriðja voru allar staðsetningar bakfærðar um tvo og þannig fyrir alla 21-mera bútsins. Ef strengurinn passaði fullkomlega á einhvern stað, höfðu því allir bútarnir kosið um sömu staðsetningu og sá fyrsti. Því var í lok lesturs allra 21-mera framkvæmd meirihlutakosning á staðsetningu bútsins í heild.

Þetta er þó ekki skotheld nálgun, því þarna er miðað við að leggja strenginn frá fyrsta basa til þess seinasta við þessa staðsetningu. Í tilfelli þess að öfugur strengur hafi fundið fleiri mát var strengurinn lagður í öfuga stefnu og endaði þá á staðnum sem flest atkvæði fékk. Þessi óháða mátun á hverjum 21-mer gerir það að verkum að jafnvel þótt villa leynist snemma í röð bútsins, mun besta staðsetning bútsins finnast þegar seinni 21-merar fá að kjósa. Ólíklegt er að tvær villur leynist í röð niturbasa sem er ekki lengri er 150bp að lengd en í einhverjum tilvikum gætu stystu strengirnir innihaldið villu í miðju strengsins sem gerði það að verkum að enginn 21-mer mátaðist við réttan stað. Þar sem bútarnir eru gífurlega margir og meðaldýpt raðgreiningar var yfir 4000, var talið ásættanlegt að þeir röðuðust vitlaust.

Þegar besti upphafsstaður mátunar hafði verið fundinn fyrir safn 100 búta, kom í ljós að flestir þeirra pössuðu hárrétt eða innihéldu aðeins eina stökkbreytingu. Í

búti af lengd L er fjöldi k-mera, $L_k = L - k + 1$. Ef gert er ráð fyrir því að 1 villa sé í bútnum og hún sé stökkbreyting sem valdi ekki hliðrun á bútnum með tilliti til viðmiðunarstrengs, er í minnsta lagi 1 k-mer sem ekki passar. Það er í því tilviki að villan sé á öðrum enda bútsins. Hæsti fjöldi k-mera sem ekki passar er þá k/L_k.

Í meirihlutakosningunni fæst því ágætisdómur á það hvort nauðsyn sé á staðbundinni sérröðun bútsins (*local alignment*) með tilliti til eyðingar eða innsetningar niturbasa. Látum fjölda atkvæða um besta stað vera m og því fjölda þeirra k-mera sem ekki passa vera $M = 1 - (m/k)$. Þá er ljóst að ef $M > k/L_k$, þá eru villurnar á svæðinu fleiri en 1 eða mátunin hefur hliðrast á einhverjum tímapunkti vegna innsetningar eða eyðingar. Mögulegt er að hliðrunin sé innan við 21 sæti frá upphafi og falli þá ekki inn í þetta skilyrði en ákveðið var að kanna slík tilvik ekki sérstaklega, þar sem þau myndu hafa lítil áhrif í svona stóru safni búta.

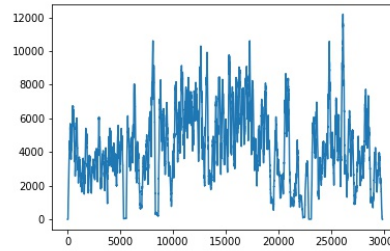

FIGURE 1. *Dýpt raðgreiningar við mátun 1.8 milljóna búta*

Þeir bútar sem stóðust ekki þennan þröskuld voru mátaðir sérstaklega með staðbundinni skoðun. Þar var gert ráð fyrir því að lesinn bútur gæti verið styttri en viðkomandi svæði á viðmiðunarstreng en ekki að hann gæti verið lengri. Fyrirmæli um verkefnið tóku sérstaklega fram að innsetning og eyðing kæmu ekki fyrir í raunverulegum samsemdarstreng bútanna. Þar sem mátun lengri búts krefðist þess að lengja allan viðmiðunarstrenginn og uppfæra viðmið um staðsetningar í samræmi við það, var ákveðið að eyða frekar þeim basa í lesnum búti sem ekki var umfram í þeim streng.

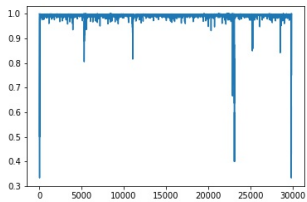

FIGURE 2. *Hlutfall vinningshafa í kosningum*

Hver bútur var þannig lesinn inn og færður í viðeigandi stað fylkis með dálkafjölda lengdar viðmiðunarstrengsins og línur fyrir möguleikana `A`,`C`,`G`,`T` og `-`, þar sem síðasta táknið stóð fyrir úrfellingu í bútnum. Búturinn var ekki notaður framar og minni losað fyrir næsta bút. Þannig gekk keyrslan frekar hratt og óx línulega með stærð safns búta.

Mest voru 900.000 bútar keyrðir úr tveimur söfnum búta og því alls 1.800.000 bútar mátaðir við strenginn. Meðaldýpt raðgreiningar, þ.e. meðalfjöldi búta sem kusu um hvern stað, var þá 4231 og hafði hver einasti staður á strengnum fengið kosningu. Erfitt reyndist að fylla fremsta hluta strengsins og aftasta og því þurfti að fjölga

keyrslum upp í þennan mikla fjölda. Einnig var bútur á svæðinu 22890 - 23130 í viðmiðunarstrengnum sem erfitt reyndist að fylla af einhverjum ástæðum en á endanum tókst að fá einhverja þekju á hann.

Þó upphaf og endir strengins hafi fengið einhverja þekju var hún yst ekki einu sinni 100 raðgreiningar á dýpt og því ekki ýkja áreiðanleg. Á því svæði var áreiðanleiki meirihlutakosningar áberandi minni en annars staðar. Ekki er óviðbúið að fá gisna kafla í endum strengsins við meðhöndlun raðgreiningargagna og því var ákveðið að miða samanburð við það að endaþekjan hafði náð raðgreiningardýpt að lágmarki 100 eða frá og með 39 niturbasa að framan til og með 53 niturbasa að aftan.

Ómeðhöndlaðir voru strengirnir með Hamming-fjarlægð upp á 49 en eftir snyrtingu á endum féll sú fjarlægð niður í 9. Í öllum tilvikum höfðu bútarnir kosið í yfir 95% tilvika um sama niturbasann, sem var ólíkur viðmiðunarstrengnum, og meirihluti þeirra yfir 99%. Eftirfarandi stökkbreytingar fundust við gerð samsemdarstrengsins út 1.800.000 bútum.

Staðsetning	Viðmiðunarstr.	Samsemdarstr.	Kosning
884	C	T	99.6%
1397	G	A	99.4%
8653	G	T	99.0%
11083	G	T	98.3%
12357	C	A	99.6%
13506	C	T	98.8%
26447	C	T	99.7%
28688	T	C	95.2%
29742	G	T	99.7%

TABLE 1. *Staðsetningar stökkbreytinga, miðað við að talning hefjist á ‚1‘*

Áhugavert er að sjá að í öllum tilfellum nema einu var stökkbreytingin úr C eða G niturbasa í viðmiðunarstreng yfir í A eða T í samsemdarstrengnum. Eitthvað hefur verið rætt um óstöðugleika á milli C og G í afritunarferli baktería en hér virðast niturbasarnir báðir gjarnir á að stökkbreytast yfir í T eða A. Það er áhugavert en við treystum okkur ekki til að draga frekar ályktanir um það.

Elías Bjartur Einarsson
Þórhallur Auður Helgason