Plan upravljanja istraživačkim podacima-EpiTranSat

Meštrović, Nevenka

Data management plan / Plan upravljanja istraživačkim podacima

Plan upravljanja istraivačkim podacima

Opć e informacije		
	Ime i prezime predlagatelja	dr. sc. Nevenka Meštrović Radan upišite ime i prezime glavnog istraživača
	Matič na organizacija	Institut Ruđer Bošković
	Naziv projekta	(Epi)genomika i transkriptomika eukromatinskih satelitnih DNA u embriogenezi i razvoju
	Upravitelj podacima	dr. sc. Nevenka Meštrović Radan, nevenka@irb.hr upišite ime prezime te e-adresu osobe koja je odgovorna za upravljanje podacima i *Planom upravljanja istraživačkih podataka*
I	Prikupljanje podataka i dokumentacija	
	Koje ć ete podatke prikupljati, obrađ ivati, stvarati ili se ponovno njima koristiti? (navedite format, vrstu i opseg podataka)	U tijeku projekta najvećim dijelom će se prikupljati rezultati sekvenciranja (do sada pretežito vezano uz aktivnost A1.6, A.1.7, A3.1, A4.1, A5.3) koji obuhvaćaju Nanopore sekvenciranje dugih odsječaka DNA i Illumina RNA sekvenciranje različitih tkiva i spolova kukca *Tribolium castaneum*. Navedeni sirovi podaci će se daljnje obrađivati te će se analize biti bazirane na obradi posloženog genoma (eng. *genome assembly*). Sirovi Nanopore podaci su u .fast5 formatu koji predstavljaju HDF5 (eng. hierarchical data format 5) tip datoteke koji je sa specifičnim shemom definiran od strane Oxford Nanopore Technologies (ONT) za pohranjivanje sirovih strujnih signala generiranih od strane ONT uređaja (u našem slučaju MinION prijenosni sekvencer). Zatim će se iz .fast5 datoteka pozivati baze uz specifično razvijene algoritme (npr. Guppy) te dobivati .fastq podaci. Oni će zatim biti dalje spojeni u jedne velike datoteke koje odgovaraju pojedinom uzorku. Na posloženom genomu će se tražiti specifične sekvence čija će lokacija biti zapisana u .gff formatu. S obzirom kako .fast5 format zadržava sve podatke o Nanopore sekvenciranju on predstavlja najveće grupe podataka čina se veličina kreće 5-10 TB. Ostale .fasta datoteke proizašle iz sirovih podataka su do 1 TB, a posloženi genom i liste sekvenci/gena/svojstava su u 100 Mb rasponu. Rezultati RNA sekvenciranja su grupe od kratkih sekvenci zapisane u obliku .fastq datoteka koji u sebi zadržavaju kvalitetu očitanja i opsega su 5-10 Gb po uzorku. U međukoracima različitih analiza se pretežno koriste tablice zapisane u jednom od prikladnih formata za korišteni algoritam (vrlo često .csv format), dok su grafički prikazi zapisani u vektorskom obliku (.svg format) te finalno pohranjeni u .png ili .tiff formatu te su najčešće u rasponu od nekoliko Mb po tablici/slici.
	Kako ć e se podaci prikupljati, obrađ ivati ili stvarati? (ukratko navedite metodologiju i procese osiguranja kvalitete te nač ine organiziranja podataka)	Nanopore sekvenciranje se odvija uz pomoć MinION uređaja koji nam omogućava dobivanje dugih očitanja uz provedbu svih koraka unutar našeg laboratorija. Program koji prati ONT uređaje su zove MinKnow i on omogućava validaciju ćelija na kojima se provodi sekvenciranje, osigurava kvalitetu sekvenci, zapisuje dobivene signale u .fast5 formatu te omogućava pozivanje baza i bazičnu razinu analize podataka. U ovom koraku se imenuju i organiziraju istraživački podaci unutar strukturiranih mapa koje odgovaraju različitim životnim stadijima/tkivima/spolovima uzoraka. Također MinKnow omogućava stvaranje report datoteke koja u sebi sadrži pregled svih informacija o svakom sekvenciranju, parametre, kumulativni zbroj sekvenci, njihove duljine i potencijalne poruke stvorene od strane programa za vrijeme rada tako osiguravajući pregled kvalitete podataka. S obzirom kako postoji nekolicina razvijenih alata za pozivanje baza od velikog je značaja pohraniti sirove podatke u ovom trenutku na nekoliko lokacija. Za naše potrebe, korišten je Guppy alat za

		pozivanje baza te osnovne komande za filtriranje/povezivanje datoteka s kojima smo dalje ulazili u proces poslagivanja genoma. Za ovaj korak korištena je Canu aplikacija koja je komputacijski vrlo zahtjevna i iz tog razloga je pokretana u skopu računalnog klastera Isabella. Za tu namjenu je otvoren profil istraživačkog projekta pod šifrom HRZZ-IP-2019-04-4910 uz 1 TB skladišnog prostora. To nam omogućava da sortiramo i validiramo izlazne datoteke prije nego ih trajno pohranim na drugim tvrdim diskovima.Ostale analize uključuju analize podataka uz novo stvorene skripte u R programskom sučelju, analize sekvenci u Geneious programu te uređivanje slika u Adobe Photoshopu, Inkscape-u i Gimpu. Navedeni proizvedeni rezultati lokalno na računalima svih suradnika na projektu se redovito pohranjuju na unutarnjim diskovima računala uz dodatno pohranjivanje na jednoj od Cloud lokacija i zajedničkim vanjskim diskovima te na NAS sustavu pohrane podataka.
	Koju ć ete dokumentaciju i metapodatke ustupiti osim podataka? (navedite koje su informacije potrebne korisnicima kako bi mogli č itati i interpretirati podatke u buduć nosti te koji ć e se standardi koristiti pri tumač enju podataka)	Pri izvođenju istraživačkog projekta i ostvarivanju ciljeva planiranim u random plan zapisuju se protokoli razvijeni za eksperimentalno izvođenje te bioinformatičke analize u obliku skripti. Dalje se navedeno priprema za prezentaciju u sklopu objave radova (kao što je primjer publikacija 1) gdje su detaljno opisani svi koraci napravljeni u eksperimentima ili su priložene točne komande s kojima su pokretani algoritmi. Prilikom objave radova za koje su bile potrebne kompleksnije razvijene kodne knjige iste će biti dostupne u skopu Github online alata koji omogućava i verzioniranje. Sirove podatke koje budemo pohranjivali su uz priložene protokole/kodne knjige dovolji za interpretaciju istih. S obzirom da se radi o malom broju uzoraka a velikom broju sekvenci za svaki, pravilno imenovanje i pohrana u jedan od javno dostupnih baza podataka (NCBI, ENA…) koji zahtijevaju detaljan opis svakog uzorka/stadija/tkiva/tipa sekvenciranja osiguravaju pravilno raspoznavanje i tumačenje podataka. Nadalje, skupovi podataka su u repozitorijima opisani navedenim metapodacima i trajnim identifikatorima te na taj način omogućavaju pretraživanje u databazama. U slučaju novo posloženog genoma ili opisanih sekvenci iste će biti pripremljene kao dodatak radu u časopis na način na budu pohranjeni u repozitoriju. Prilikom pripreme ovih podataka potrebno je navesti sve potrebne karakteristike i algoritme s verzijama koji su korišteni u analizama što će biti vezano uz same podatke i dobiti trajni identifikacijski broj.
II	Pravna i sigurnosna pitanja	
	Jeste li ogranič eni sporazumom o povjerljivosti? Imate li potrebna	S obzirom da se radi na o podacima koji su prikupljeni na životinjskoj vrsti (beskralježnjaci, kukci) koja je nabavljena i uzgaja se već dugi niz godina uz svu potrebnu dokumentaciju, nikakav sporazum o povjerljivosti nije primjenjiv u ovom slučaju.

	dopuštenja za prikupljanje, obradu, č uvanje i dijeljenje podataka? Jesu li osobe č iji se podaci pohranjuju informirani o tome i jesu li dali privolu? Kojim ć ete se metodama koristiti u svrhu zaštite osjetljivih podataka (GDPR - posebne kategorije osobnih podataka)?	
	Kako ć e se regulirati pristup podacima i njihova sigurnost? Koji su potencijalni rizici koje treba uzeti u obzir? Kako ć ete osigurati sigurnost pohrane osjetljivih podataka?	Pristup podacima uključenim u izračun i dobivanje rezultata istraživanja objavljenih putem radova biti će javno dostupan s obzirom na standarde u polju istraživanja putem nekih od javnih baza podataka za istraživanje (i.e. FigShare). Pristup podacima koji su se koristili kako bi se dobili rezultati biti će omogućeni po objavi rada u jednom u repozitoriju ili na zahtjev istraživačkoj grupi s obzirom na veliku količinu podataka koja onemogućuje jednostavno dijeljenje preko mrežnih infrastruktura. Trajna pohrana velikih setova biti će izvršena na lokalnoj pohrani u najmanje dvije kopije kako bi se zaštitio integritet podataka u slučaju nesreće. Drugih potencijalnih rizika vezanih uz podatke poput klasificiranih informacija ili GDPR zaštićenih informacija nema tako da osiguravanje visoke razine sigurnosti u ovom slučaju nije primjenjivo.
	Kako ć ete upravljati zaštitom autorskih prava i intelektualnog vlasništva? Tko ć e biti vlasnik podataka? Koje ć e se licence primjenjivati na podatke? Koja ć e se ogranič enja primjenjivati na ponovnu uporabu osobnih podataka?	Sve podatke dobivene u sklopu istraživanja biti će licencirane CC BY 4.0 licencom po preporuci Europske komisije kako bi omogućili ponovno i ispravno korištenje podataka od strane drugih znanstvenika. Sam vlasnik intelektualnog vlasništva je laboratorij u kojem je projekt i rezultati producirani: Laboratorij za nekodirajuće DNA, Institut Ruđer Bošković. S obzirom da nema osobnih i povjerljivih podataka osim intelektualnih dostignuća nikakva ograničenja neće biti primjenjiva na podatke.
III	Pohrana i č uvanje podataka	
	Kako ć e podaci biti pohranjeni i kako ć e biti napravljena sigurnosna kopija podataka (*backup*) tijekom istraž ivanja? Koji su kapaciteti č uvanja podataka kojim raspolaž ete? Kojim se procedurama koristite za sigurnosnu kopiju (*backup*)?	Obzirom na veliku količinu konačnih prikupljenih podataka (>10-12TB) efektivno skladištenje na platformi MojOblak ili PUH nije efektivno moguće. Zbog toga je u sklopu projekta nabavljeno nekoliko (3) vanjskih tvrdih diskove veličine of 4TB, nekoliko unutarnjih tvrdih diskova (4) veličine 2- 4TB te Synology NAS sustav za pohranu podataka sa mogućnošću nadogradnje do 48TB memorije (trenutno 8TB), što nam trenutno omogućuje do preko 40TB prostora za pohranu a maksimalno 80TB. Ta količina prostora omogućuje nam da pohranimo sve podatke u najmanje dvije kopije, a oni važniji ili manji u veličini u najmanje tri kopije kako bi se mogli vjerno replicirati rezultati istraživanja, poglavito .fast5 datoteke za ONT očitanja te .fastq datoteke iz RNA-seq eksperimenata koji služe kao početne točke proračunima. Ti podaci skupa sa pohranjenim kodnim knjigama će omogućiti replikaciju eksperimenata i podataka unatoč gubitku nekih od međupodataka. Uz to u vrijeme trajanja istraživanja jedna kopija svih podataka će biti pohranjena na lokalnim računalima kako bi se omogućio brz pristup i manipulacija sa podacima.

	Koji je vaš plan č uvanja podataka? U kojim ć e se formatima č uvati?	Većina podataka biti će čuvana u .fast5 obliku bez potrebe za kompresijom s obzirom da taj oblik već koristi složene knjižnice kako bi optimizirao kompresiju te se standardnim metodama veličina neznatno smanji. Za pohranu .fastq datoteka koristiti će se standardni adaptivno Lempel-Ziv kodiranje putem *gunzip* programa kako bi se komprimirali podaci u .fastq.gz format te tako drastično smanjila njihova veličina (standardni omjer za.fastq u fastq.gz konverziju je 6:1). Kodne knjige biti će verzionirane i pohranjene skupa sa podacima. Podaci će se čuvati na neodređeni vremenski period skupa sa metapodacima i kodnim knjigama kao i verzijama korištenih programa za obradu kako bi omogućili ponovno korištenje podataka u bilo kojoj točki u vremenu.
IV	Dijeljenje i ponovna uporaba podataka	
	Kako i gdje ć e se podaci dijeliti? Na kojem repozitoriju planirate dijeliti podatke? Kako ć e potencijalni korisnici doznati za podatke?	Temeljni i sirovi podaci neće biti dijeljeni na javnim repozitorijima zbog njihove veličine. Temeljni i sirovi podaci biti će ustupljeni svim znanstvenicima koji traže dopuštenje korištenje, osobno ili na zahtjev preko za to osiguranih protokola za dijeljenje podataka (direktni FTP/SFTP transfer, ili putem programa za prijenos poput WinSCP ili FileZilla). Podaci dobiveni izračunima (tablice, genomska sravnjenja) će biti podijeljeni na javnim repozitorijima primjenjenima za tu svrhu u polju znanosti (National Center for Biotechnology Information, FigShare, European Nucleotide Archive) dok će korisnici o postojanju tih podataka biti informirani u sklopu objavljenih publikacija.
	Ako postoje podaci koji se ne smiju dijeliti (prijavitelji vezani zakonskim, etič kim, autorskim pravila, povjerljivošć u i sl.), pojasnite razloge ogranič enja.	Svi podaci uključeni u istraživanje se smiju dijeliti uz pripisivanje zasluga u sklopu CC BY 4.0 licence.
	Potvrdite da ć ete se koristiti digitalnim repozitorijem koji je u skladu s nač elima *FAIR data*.	Digitalni repozitorij koji će biti korišten prati sva 4 FAIR načela, Findable - dostupan na internetu poveznicama sa objavljenih publikacija, Accessible - moguće je pristupiti uz stabilnu internetsku vezu, Interoperable - podaci se preuzimaju u standardnim formatima za pohranu podataka poput .zip, Reusable - svi metapodaci biti će objavljeni skupa sa glavnim dijelom podataka.
	Potvrdite da ć ete se koristiti digitalnim repozitorijem koji održ ava neprofitna organizacija (ako ne, objasnite zašto ne mož ete dijeliti podatke na digitalnom repozitoriju koji nije komercijalan).	Jedna kopija digitalnih podataka će biti pohranjena na javno dostupnim nekomercijalnim baza Hrvatske znanstvene infrastrukture (Dabar, PUH) uz Figshare platformu.

Ref:

[1] Celjak, D., Dorotić Malič, I., Matijević, M., Poljak, Lj., Posavec K. i Turk, I.: „Istraživački podaci–što s njima?“ Istraživački podaci - što s njima? : priručnik o upravljanju istraživačkim podacima | Digitalni repozitorij Srca (unizg.hr)