Prediktivno modeliranje retroviralnih integracija virusa HIV-1 u aktivirane CD4+ T stanice

Martinović, Moreno

Moreno Martinović

Prediktivno modeliranje retroviralnih integracija virusa HIV-1 u aktivirane $CD4^+$ T stanice

DIPLOMSKI RAD

Predan Sveučilištu u Zagrebu Farmaceutsko-biokemijskom fakultetu

Zagreb, 2020.

Ovaj diplomski rad je prijavljen na kolegiju Molekularna biologija s genetičkim inženjerstvom Sveučilišta u Zagrebu Farmaceutsko-biokemijskog fakulteta i izrađen u Grupi za bioinformatiku na Zavodu za molekularnu biologiju Sveučilišta u Zagrebu Prirodoslovno-matematičkog fakulteta pod stručnim vodstvom prof. dr. sc. Gordana Lauca i suvoditeljstvom prof. dr. sc. Kristiana Vlahovičeka.

Zahvaljujem mentoru prof. dr. sc. Gordanu Laucu i komentoru prof. dr. sc. Kristianu Vlahovičeku na pruženoj prilici za izradu diplomskog rada u Grupi za bioinformatiku Prirodoslovno-matematičkog fakulteta.
Hvala cijeloj Grupi za bioinformatiku na ugodnoj radnoj (i neradnoj) atmosferi (prije no što je Zagreb postao niskobudžetni film katastrofe).

Posebno hvala (sada već dr. sc.) Maji koja me naučila svemu potrebnom za izradu diplomskog rada i više, bila tu za svako moje pitanje i omogućila mi prilike izvan okvira ovog rada.

Hvala mojoj obitelji i prijateljima na pruženoj potpori, kako kroz studiranje, tako i kroz cijeli život.

Najveće hvala Rachel bez čijih poticaja i vjere u mene se ne bih usudio ići putem kojim jesam.

1. Uvod

Retrovirusi su velika i raznolika porodica jednolančanih RNA virusa s ovojnicom, zajedničkih odrednica kao što su struktura, sastav i replikacijska svojstva (Vogt, 1997). Na stanicu domaćina se vežu putem površinskih glikoproteina na specifične membranske receptore, što dovodi do fuzije virusa i stanične membrane i posljedičnog ulaska virusa u stanicu (Vogt, 1997). Ono što retroviruse određuje i razlikuje od drugih virusa su procesi obrnute transkripcije i integracije koji se odvijaju nakon ulaska virusa u stanicu (Craigie i Bushman, 2012). U replikacijskom ciklusu retrovirusa obrnuta transkripcija prethodi integraciji i uključuje sintezu dvolančane DNA molekule po predložku jednolančanog RNA genoma virusa aktivnošću enzima reverzne transkriptaze. Procesom integracije se novonastala DNA kopija genoma kovalentno veže i umeće unutar DNA stanice organizma domaćina (Craigie i Bushman, 2012). Broj mogućih mjesta integracije retrovirusa u genom je velik i široko raspodijeljen. Nakon integracije se provirus eksprimira djelovanjem stanične RNA-polimeraze II i replicira istim staničnim mehanizmima kao i geni domaćina (Vogt, 1997).

Virus humane imunodeficijencije (*Human immunodeficiency virus* - HIV) je retrovirus roda *Lentivirus*, uzročnik Stečenog sindroma imunodeficijencije (*Acquired immunodeficiency syndrome* - AIDS) (Gallo i sur., 1984). Infekcije lentivirusima imaju određena zajednička svojstva, kao što su dugi i varijabilni inkubacijski period, perzistentna viralna replikacija, neurološke manifestacije i smanjenje broja određenih hematoloških i imunoloških stanica (Desrosiers i Letvin, 1987). Svi imaju sličnu morfologiju, pokazuju tropizam prema makrofagima i veliku genetsku i antigenu varijabilnost. Također posjeduju dodatne regulatorne gene koji nisu nađeni kod ostalih retrovirusa (Fauci i Desrosiers, 1997). Karakterizirana su dva tipa, HIV-1 i HIV-2 kao jedini poznati humani lentivirusi. HIV-1 je virulentniji i infektivniji od HIV-2 i uključuje većinu slučajeva infekcija HIV-om u svijetu (Gilbert i sur., 2003). Uspjeh antiretroviralne terapije (ARV) doveo je do drastičnog smanjenja broja smrtonosnih slučajeva AIDS-a, no perzistencija virusa HIV-1 i dalje predstavlja problem (Martin i Siliciano, 2016).

Nakon integracije, viralni genom HIV-a može biti eksprimiran ili ući u dormantnu fazu i uspostaviti spremnik latentno inficiranih stanica. Stanice s latentnom infekcijom se ne mogu razlikovati od neinficiranih, stoga ih imunološki sustav ne može eliminirati i na njih ne djeluje ARV (Sengupta i Siliciano, 2018; Churchill i sur., 2016). Mirujući memorijski $CD4^+$ T limfociti su glavni latentni spremnik virusa HIV-1 (Chomont i sur., 2009), no još nije razjašnjeno kako dolazi do uspostave takvih spremnika, s obzirom da HIV-1 neefikasno inficira mirujuće T stanice zbog različitih prepreka na integracijskoj i pred-integracijskoj razini (Sengupta i Siliciano, 2018; Zack i sur., 2013; Pace i sur., 2012; Dai i sur., 2009; Agosto i sur., 2009). Jedan od potencijalnih mehanizama je povratak aktiviranih $CD4^+$ T limfocita u mirujuće stanje, čime nastaju spremnici transkripcijski utišanih, ali vijabilnih viralnih genoma (Sengupta i Siliciano, 2018). Genomi HIV-1 mogu desetljećima ostati skriveni tijekom ARV-a prije ponovne pojave simptoma viralnog opterećenja i relapsa AIDS-a. Posljedično tome, unatoč preko 30 godina napretka u kontroli viralne replikacije i prevenciji AIDS-a, transkripcijska latencija predstavlja barijeru prema izlječenju (Margolis i sur., 2016).

HIV-1 u aktivne $CD4^+$ T limfocite ulazi preko kompleksa nuklearne pore (*Nuclear pore complex* - NPC) (Di Nunzio i sur., 2013; Koh i sur., 2013; Ocwieja i sur., 2011). Proteini NPC-a su važni čimbenici samog ulaska virusa u jezgru stanice, ali i posljedične lokalizacije i integracije u genom (Lusic i Siliciano, 2017; Marini i sur., 2015; Lelek i sur., 2015; Di Nunzio i sur., 2013; Koh i sur., 2013; Ocwieja i sur., 2011; Suzuki i Craigie, 2007). Integracija nije nasumičan proces. Pokazano je da protein viralna integraza kroz interakciju s LEDGF/p75 (Singh i sur., 2015; Ciuffi i sur., 2005; Cherepanov i sur., 2003) usmjerava integraciju virusa u tijela aktivnih gena, najčešće u genski-bogatim regijama (Schröder i sur., 2002). Kad je LEDGF/p75 smanjen, obrazac integracije se pomiče prema 5' kraju gena odnosno prema genski-siromašnim regijama. Protein kapside također pridonosi lokaliziranju viralnog genoma kroz interakciju s poliadenilacijskim faktorom (engl. *cleavage and polyadenylation specificity factor 6* - CPSF6) (Sowd i sur., 2016; Vranckx i sur., 2016; Singh i sur., 2015). Nedostatak CPSF6 zaustavlja nadolazeće viralne čestice na razini NPC ili preusmjeravaju viralnu DNA na heterokromatinske domene asocirane s laminom (Bejarano i sur., 2019; Achuthan i sur,. 2018).

Poznato je da se HIV-1 usmjerava u područja otvorenog kromatina s aktivnom transkripcijom i područja koja imaju pojačivačke epigenomske oznake (Chen i sur., 2017; Wang i sur., 2007, Schröder i sur., 2002). U ljudskom genomu postoje takozvani super pojačivači (engl. *super enhancers* - SE) koje odlikuje visoka razina acetiliranog lizina 27 histona 3 (H3K27ac) i vezanje transkripcijskih koaktivatora, kao što su protein koji sadrži bromodomenu 4 (*bromodomain-containing protein 4* - BRD4) i medijatorski kompleks (Whyte i sur., 2013; Hnisz i sur., 2013). U odnosu na tipične pojačivače, SE su veći, imaju veću gustoću transkripcijskih faktora i često su asocirani s genima ključnima za identitet stanice koji kontroliraju stanični stadij i diferencijaciju somatskih stanica (Parker i sur., 2013; Hnisz, Day i Young, 2013). SE kontroliraju citokine, njihove receptore i transkripcijske faktore čime reguliraju za T stanice specifične transkripcijske profile (Witte, O'Shea i Vahedi, 2015). Sukladno tome, zanimljivo je da je gen BACH2 jedan od najčešće HIV-om integriranih gena i ujedno je pod utjecajem jednim od najjačih imuno-aktivirajućih SE (Roychoudhuri i sur,. 2013; Tsukumo i sur., 2013). Pokazano je da su SE elementi gena koji određuju identitet stanice povezani s NPC proteinima koji reguliraju njihovu ekspresiju i lokaliziraju ih u periferiju jezgre (Toda i sur., 2017; Ibarra i sur., 2016). SE također imaju ulogu u organizaciji kromatina općenito kroz kromatinske strukture višeg reda i kromatinske petlje (Olley i sur,. 2018; Rao i sur., 2017; Beagrie i sur., 2017).

Razvoj Hi-C metode (Lieberman-Aiden i sur., 2009) je kroz posljednje desetljeće doveo do saznanja da kromosomsko smatanje i stvaranje petlji organizira genom u prostorno razdvojene odjeljke (Bonev i Cavalli, 2016). Visoko transkribirani geni međusobno preferabilno ostvaruju prostorne kontakte u odnosu na druge gene, čineći tako tzv. odjeljak A (Rao i sur., 2014; Lieberman-Aiden i sur., 2009). Na isti način utišani geni i intergenske regije čine odjeljak B. Lokusi odjeljka B se obično nalaze u periferiji jezgre u dodiru s nuklearnom laminom i odlikuju se niskom razinom genske ekspresije i represivnim heterokromatinskim oznakama (Noordermeer i sur., 2011). HIV-1 se gotovo uopće ne integrira u regije B odjeljka (Vranckx i sur., 2016; Marini i sur., 2015), već većinom regije otvorenog kromatina, koje se prema nekim radovima mapiraju u blizini NPC (Marini i sur., 2015; Lelek i sur., 2015).

Najnoviji radovi pokazali su da je za mjesta HIV-1 integracije ključna lokalizacija SE u odjeljcima genoma gdje je integracija olakšana, odnosno da se žarišta HIV-1 integracija grupiraju u trodimenzionalnom nuklearnom prostoru i u kontaktu su sa SE (Lucic, Chen, Kuzman i sur., 2019). Također, aktivnost SE indirektno utječe na integraciju reorganizacijom genoma aktiviranih T stanica (Lucic, Chen, Kuzman i sur., 2019).

Iz navedenog se može zaključiti da je integracija virusa HIV-1 u genom složen proces na koji direktno i indirektno utječe mnogo čimbenika, odnosno da se radi o složenoj međuigri virusa, kromatina stanice domaćina i dinamičke nuklearne organizacije. Iz tog razloga, unatoč brojnim istraživanjima, cjelokupna slika tog procesa još je uvijek nejasna. S obzirom na broj potencijalnih čimbenika, u ovom radu se pokušala kvantificirati njihova relativna značajnost pomoću računalnih metoda.

2. Obrazloženje

HIV-1 je retrovirus za koji je poznato da kao mete integracije ponajprije cilja transkripcijski aktivne gene i ugrađuje se u blizini odjeljka nuklearne pore. Najnovija istraživanja su pokazala da se općenito ugrađuje u susjedstvu super-pojačivačkih genomskih elemenata, i to u specifičnim prostornim nuklearnim odjeljcima aktiviranih $CD4^+$ T stanica te da grupe gena unutar istih odjeljaka stječu svoju lokalizaciju tijekom procesa aktivacije T stanica. Dakle, grupiranje gena u odjeljke u blizini super-pojačivača i transkripcijska aktivnost koja je određena i epigenetskim markerima su glavne odrednice integracije virusa HIV-1 u $CD4^+$ T stanice. No, relativna značajnost svake od tih odrednica još nije razjašnjena. U ovom radu je razvijen prediktivni model mjesta integracija HIV-1 u ljudskom genomu pomoću metoda strojnog učenja kako bi se kvantificiralo koliko prostorni podaci doprinose mogućnosti predviđanja integracija HIV-a u odnosu na epigenetske modifikacije i stanje kromatina, koristeći Hi-C podatke kao noviju metodu koja je još uvijek slabo istražena u kontekstu infekcije virusom HIV-1. Ovdje razvijene i opisane računalne metode bi se mogle primijeniti na različitim tipovima stanica za daljnje istraživanje mnogih aspekata retroviralne DNA integracije kako bi se potencijalno razjasnile okolnosti pri nastanku latentnog spremnika virusa, što bi imalo implikacije u razvoju terapije za AIDS.

3. Metode

Definicija rekurentno integriranih gena (RIG)

Koordinate verzije hg38 ljudskog genoma su preuzete pomoću R paketa biomaRt (Durinck i sur., 2009). Za svaki gen je određeno u koliko od 8 eksperimenata uzetih u analizu (Lucic, Chen, Kuzman i sur., 2019; Kok i sur., 2016; Cohn i sur., 2015; Wagner i sur., 2014; Maldarelli i sur., 2014; Brady i sur., 2009; Ikeda i sur., 2007; Han i sur., 2004) ima barem jednu HIV-1 integraciju, upotrebom data.table paketa (Dowle i Srinivasan, 2019). Zatim je modeliran očekivani broj eksperimenata u ovisnosti o veličini gena koristeći Poissonovu regresiju implementiranu u R funkciji glm (R Core Team, 2020).

Kao rekurentno integrirani geni (RIG) su definirani samo oni koji imaju barem jednu HIV-1 integraciju u 2 ili više eksperimenta uzeta u analizu i broj stvarnih eksperimenata veći od srednje vrijednosti očekivane Poissonove distribucije za svaki eksperiment.

U daljnju analizu su uzeti samo geni koji kodiraju za proteine.

Kromatinska imunoprecipitacija sa sekvenciranjem (ChIP-seq)

ChIP-seq je metoda kojom se mogu odrediti regije genoma obogaćene vezama s proteinom od interesa (Johnson i sur., 2007). ChIP-seq podaci primarnih $CD4^+$ T stanica su dostupni na Gene Expression Omnibus (Clough i Barrett, 2016) pod pristupnim brojem GSE122826 (Lucic, Chen, Kuzman i sur., 2019). Sljedovi Illumina PCR adaptera izrezani su iz ChIP-seq sljedova pomoću alata BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/) s parametrom ref=adapters. Za poravnanje sljedova upotrijebljen je alat minimap2 (Li, 2018) s parametrom -ax sr za kratke sljedove i referentnim ljudskim genomom verzije hg38. Duplikati poravnatih sljedova su uklonjeni pomoću alata sambamba (Tarasov i sur., 2015) i zatim su istim alatom sljedovi sortirani i indeksirani te replikati spojeni u jedinstvenu BAM (Li i sur., 2009) datoteku.

Stvarni ChIP-seq signal, značajno jači od pozadinskog šuma, je agregiran i pronađeni su maksimumi signala koristeći macs2 (Zhang i sur., 2008) s parametrima --broad --broad-cutoff 0.1 -p 1e-9 -g hs -B.

Broj ChIP-seq vrhova za svaku od histonskih modifikacija je agregiran po regijama gena koristeći R (R core team, 2020) odnosno data.table paket (Dowle i Srinivasan, 2019).

Test za kromatin dostupan transpozazi sa sekvenciranjem (ATAC-seq)

ATAC-seq je metoda kojom se određuje dostupnost kromatina transkripcijskom aparatu (Buenrostro, Chang i Greenleaf, 2015). ATAC-seq sljedovi su tretirani na isti način kao i ChIP-seq sljedovi. Korišteni su podaci dostupni na GEO (GSM3557584).

RNA sekvenciranje (RNA-seq)

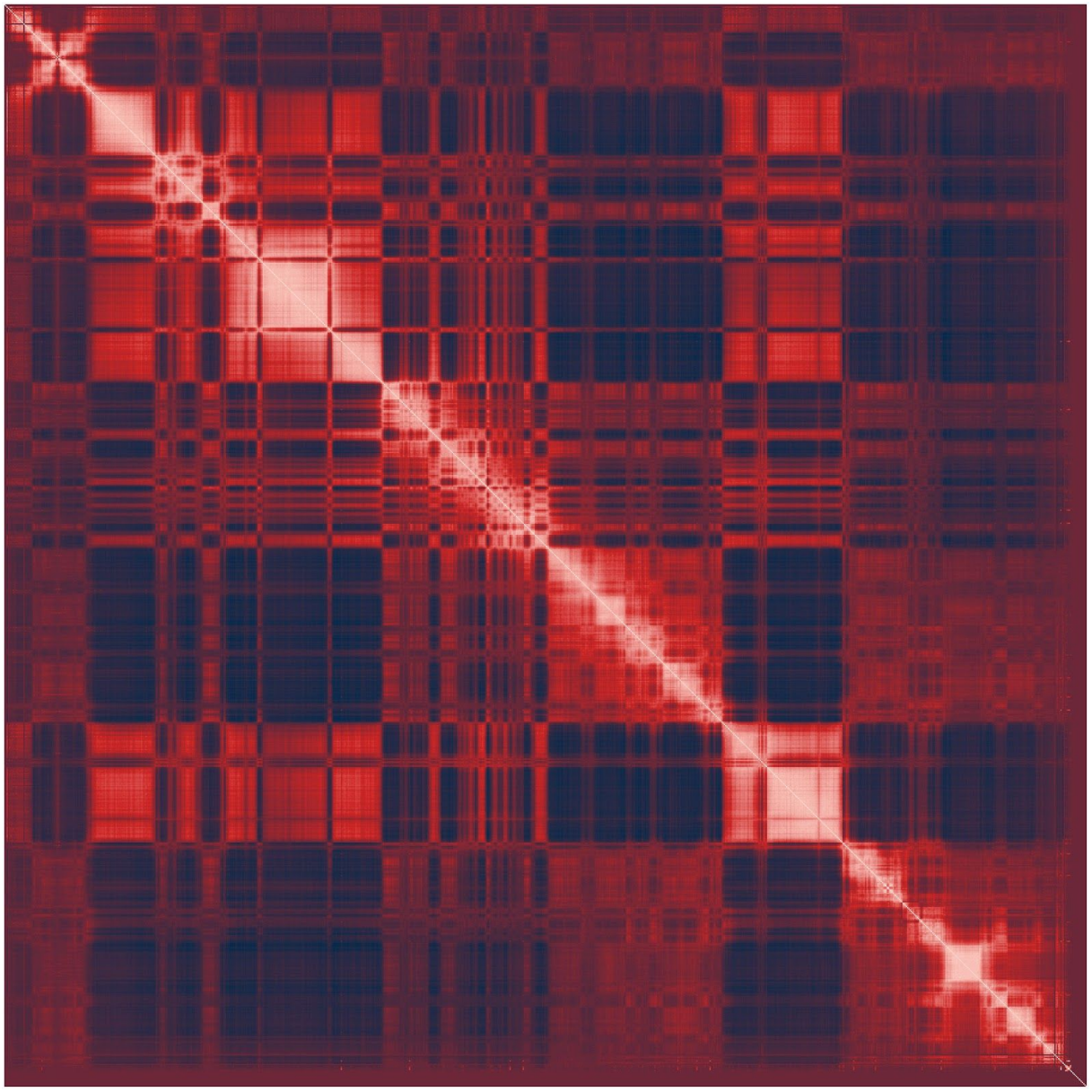
RNA-seq je metoda koja kvantificira relativnu ekspresiju gena sekvenciranjem DNA komplementarne genomskim transkriptima (Lister i sur., 2008). Korišteni su podaci RNA sekvenciranja na aktiviranim $CD4^+$ T limfocitima iz Lucic, Chen, Kuzman i sur., 2019 dostupni na GEO (GSE122735). Sljedovi su poravnati i broj sljedova po svakoj genskoj regiji hg38 genoma je agregiran upotrebom STAR alata (Dobin i sur., 2013) s parametrima --outSAMtype BAM SortedByCoordinate --quantMode GeneCounts. Broj sljedova koji se mapira na pojedini gen je normaliziran po ukupnoj veličini egzona unutar gena.

Poravnanje Hi-C podataka

Hi-C je metoda koja daje informacije o trodimenzionalnoj genomskoj arhitekturi, temeljeno na ligaciji i masivno-paralelnom sekvenciranju prostorno bliskih genomskih fragmenata (Lieberman-Aiden i sur., 2009). U radu su korišteni Hi-C podaci iz studije Lucic, Chen, Kuzman i sur., 2019 na jurkat stanicama, dostupni na GEO (GSE122958). Prvi korak u analizi Hi-C podataka je poravnanje parova sljedova na referentni ljudski genom. Za poravnanje korišten je alat minimap2 (Li, 2018) s opcijom -ax sr za kratke sljedove i --no-pairing kako bi se uspješno mogli mapirati udaljeni upareni sljedovi, s verzijom hg38 ljudskog genoma kao referentnim genomom. Dobiveni poravnati sljedovi su spojeni i filtrirani su oni s kvalitetom poravnanja manjom od 10 pomoću alata sambamba (Tarasov i sur., 2015).

Pronalaženje Hi-C kontakata

Restrikcijski fragmenti su grupirani u regije duljine 500 kilobaza i zatim su parovi Hi-C veza između takvih regija ekstrahirani pomoću vlastitog alata napisanog u programskom jeziku nim (https://nim-lang.org/). Za daljnju analizu matrice korišten je programski jezik R (R Core Team, 2020) kako bi se uklonile inherentne tehničke i biološke sklonosti Hi-C interakcijske mape. Postoje dva pristupa korekciji matrice - eksplicitni i implicitni. Eksplicitni modeli uzimaju u obzir poznate čimbenike kao što su GC sadržaj, duljina restrikcijskih fragmenata i mogućnost mapiranja genomske regije (Lajoie, Dekker i Kaplan, 2015). S obzirom da nije moguće znati sve čimbenike koji utječu na interakcije, primijenjena je implicitna metoda iterativne korekcije (engl. *iterative correction* - IC) (Imakaev i sur., 2012). Procedura se zasniva na Sinkhorn-Knopp algoritmu za balansiranje matrica (Sinkhorn i Knopp, 1967). IC algoritam je također implementiran u programskom jeziku R. Sastoji se od dva iterativna koraka - prvo se svaki red matrice podijeli sa srednjom vrijednosti reda pa zatim svaki stupac sa srednjom vrijednosti stupca. Takav proces se ponavlja do konvergencije (Lajoie, Dekker i Kaplan, 2015). Prije IC metode, iz matrice su izbačeni svi neinformativni redovi i stupci koji se nalaze u 2% onih s najmanjom varijancom. Također je 1% najvećih vrijednosti matrice svedeno na razinu 99. kvantila i izračunata je Pearsonova korelacija matrice.



Slika 1. Detalj korelacijske matrice kromosoma 14. Jasno su vidljivi aktivni i neaktivni A i B odjeljak i topološki asocirajuće domene (Dixon i sur., 2012).

Grupiranje Hi-C veza

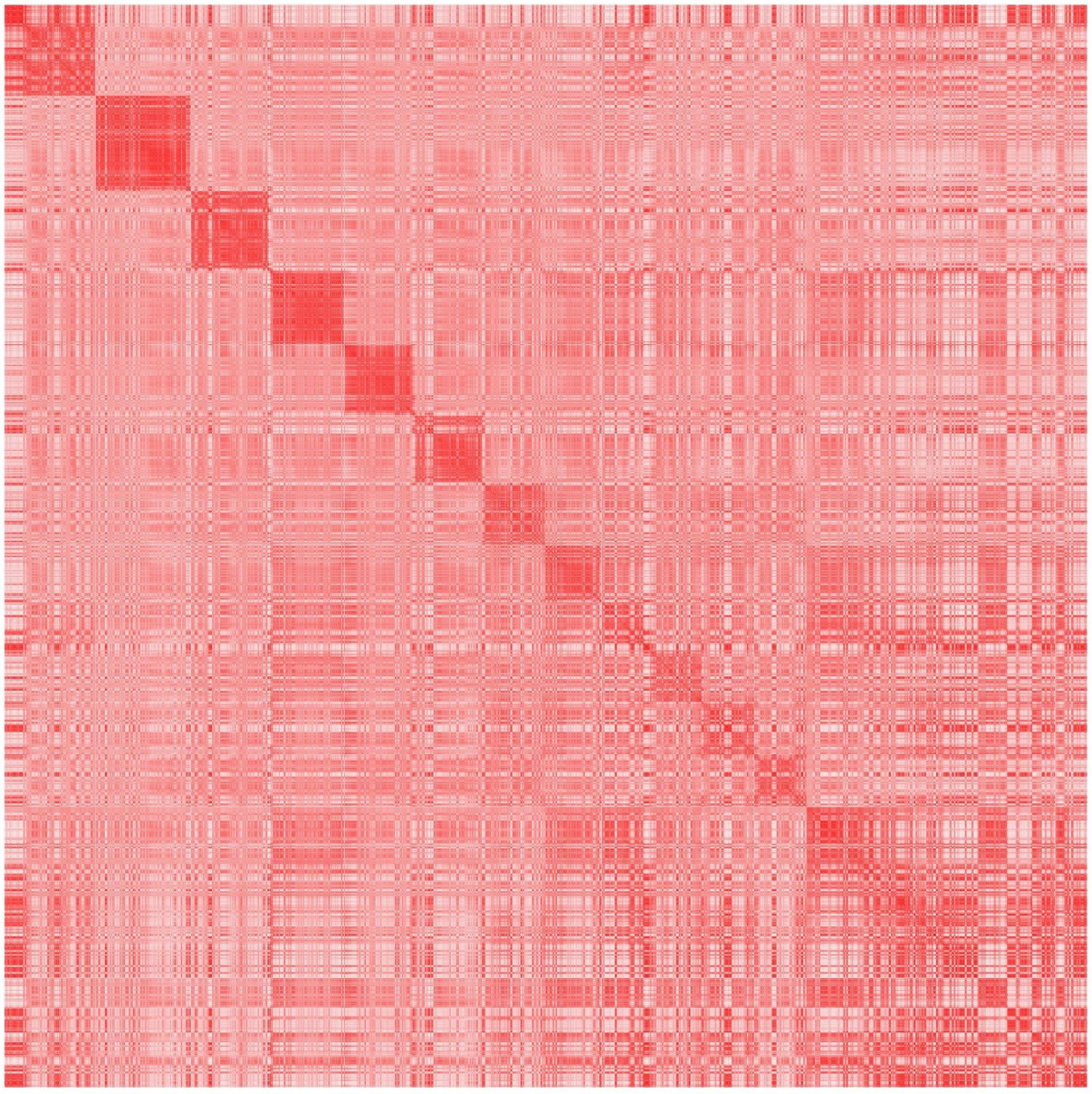
Prvi pokušaji grupiranja Hi-C interakcijske matrice u klastere su bili pomoću analize principalnih komponenti (Principal component analysis - PCA) (Lieberman-Aiden i sur., 2009). U tom pristupu predznak prve principalne komponente matrice Hi-C veza dijeli kontakte u dvije skupine od kojih jedna skupina odgovara genima s više obilježja otvorenog kromatina, dok druga skupina sadrži više zatvorenog kromatina (Rao i sur., 2014).

Kasniji pristupi (Imakaev i sur., 2012; Yaffe i Tanay, 2011) su istražili modele temeljene na više odjeljaka i takvi modeli su bolje odgovarali uzorcima u interkromosomalanim Hi-C matricama (Rao i sur., 2014). Grupiranje je izvršeno samo na interkromosomskim vezama kako bi se uklonio utjecaj kromosomskih teritorija, tj. činjenice da su kromosomi fizički odvojeni i iz tog razloga čine najjače uzorke u matrici (Lajoie, Dekker i Kaplan, 2015).

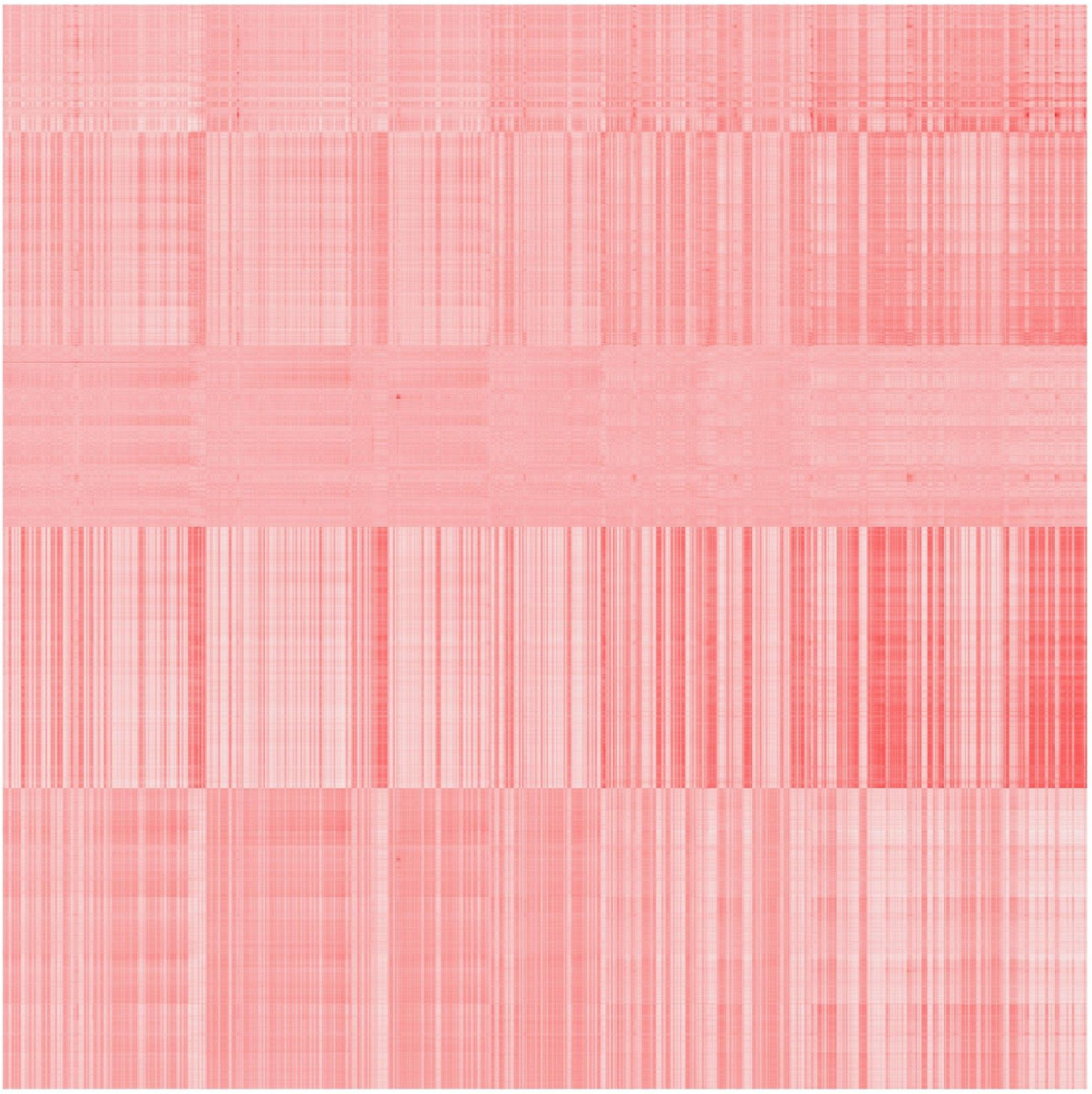
Za grupiranje Hi-C veza u odjeljke primijenjen je pristup sličan onome iz studije Rao i sur., 2014. Prvo je konstruirana normalizirana korelacijska matrica interkromosomskih veza (M) rezolucije od 500 kilobaza, takva da se lokusi neparnih kromosoma duljine 500 kilobaza nalaze u redovima, a parnih u stupcima. Broj redova i stupaca je u tom slučaju približno isti. $M_{i,j}$ je broj normaliziranih kontakata između i-tog lokusa od 500 kilobaza na neparnom kromosomu i j-tog na parnom kromosomu. Redovi i stupci s više od 30% vrijednosti 0 su izbačeni iz matrice i daljnje analize.

Rao i sur. su zatim grupirali podatke koristeći z-score transformiranu matricu kao ulazne podatke za Gauss skriveni Markovljev model za grupiranje bez nadzora (Pedregosa i sur., 2011). Umjesto toga korišten je k-means algoritam (Lloyd, 1982) implementiran u R stats paketu (R Core Team, 2020) sa zadanim postavkama. K-means algoritmom su grupirani redovi matrice M odnosno neparni kromosomi. Parametar *k* kmeans funkcije određuje broj grupa u koji će se podijeliti redovi. Nakon vizualne inspekcije raspona parametra k algoritma k-means od k = 2 do k = 7 i temeljeno na Rao i sur., 2014 i Lucic, Chen, Kuzman i sur., 2019, k = 5 je određeno kao optimalna vrijednost.

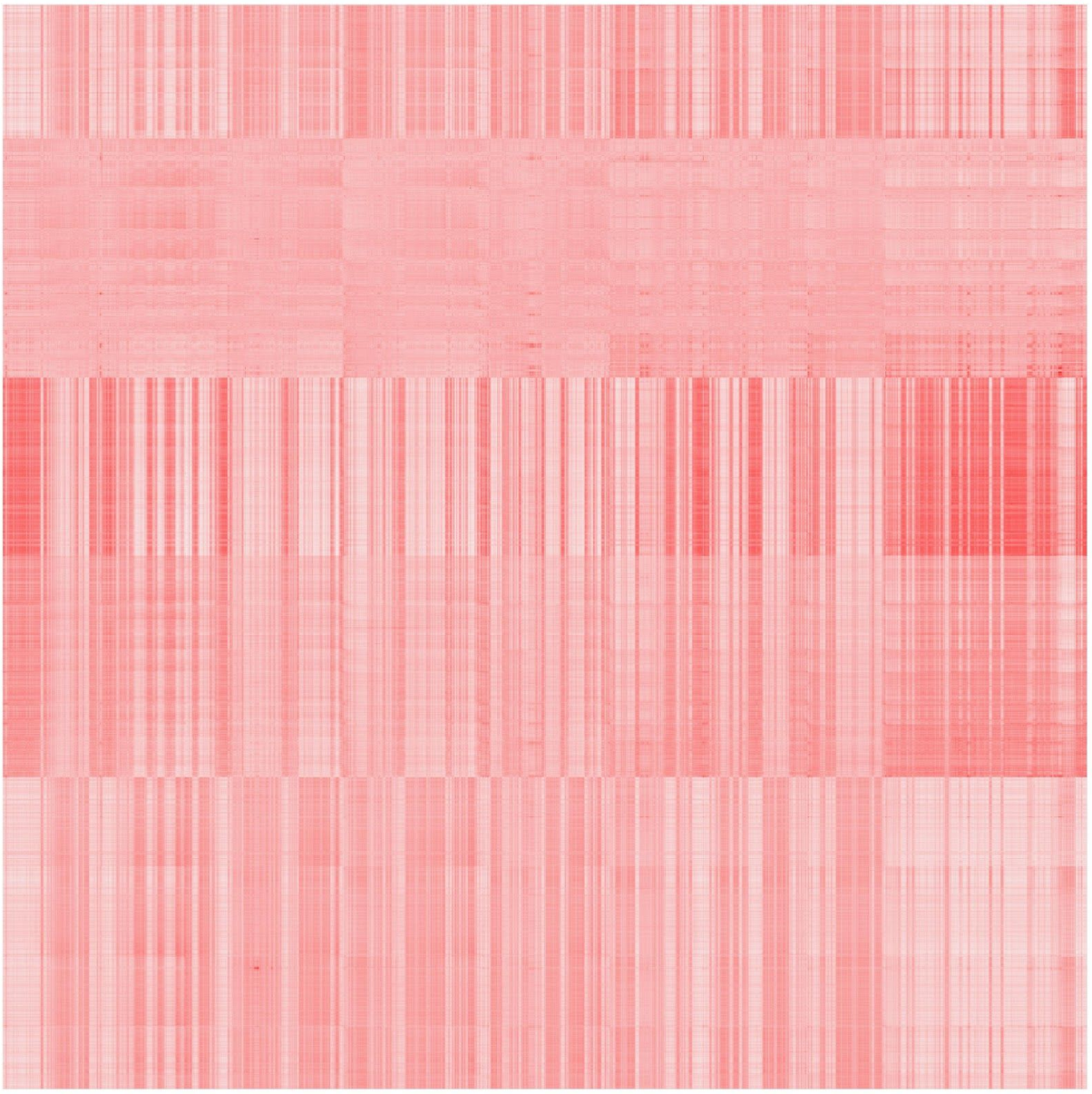
Kako bi se grupirali i parni kromosomi, isti koraci su primijenjeni nakon transpozicije matrice M. Kao i u Rao i sur., 2014, pronađeno je da svaka od grupa neparnih kromosoma preferabilno ostvaruje interakcije s po jednom grupom parnih kromosoma. Takvi parovi su spojeni i predstavljaju pododjeljke genoma.



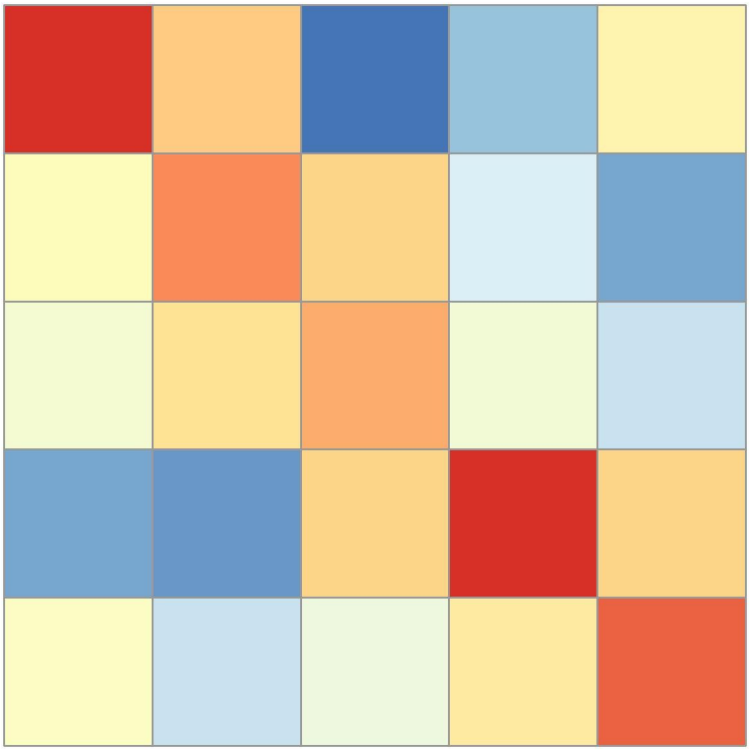
Slika 2. Korelacijska Hi-C matrica cijelog genoma podijeljenog u regije od 500 kilobaza. Vide se granice kromosoma, ali i interkromosomski uzorci.



Slika 3. Grupirani redovi interkromosomske matrice koji predstavljaju odjeljke na neparnim kromosomima. Vizualno se očituje homogenost uzorka unutar pojedinih odjeljaka.



Slika 4. Transpozicija matrice grupiranih stupaca interkromosomske matrice koji predstavljaju odjeljke na parnim kromosomima. Vide se slični uzorci kao i na neparnim kromosomima.



Slika 5. Matrica srednje vrijednosti broja veza između odjeljaka na neparnim kromosomima (redovi) i odjeljaka na parnim kromosomima (stupci). Stupci i redovi su normalizirani dijeljenjem svakog reda sa zbrojem tog reda i svakog stupca sa zbrojem tog stupca. Matrica je zatim standardizirana na način da se svake vrijednosti oduzme srednja vrijednost matrice te zatim podijeli standardnom devijacijom. Svaki odjeljak na neparnim kromosomima preferabilno ostvaruje veze s jednim odjeljkom na parnima. Takvi odjeljci su spojeni. Grafički prikaz matrice je generiran pomoću R paketa corrplot (Wei i Simko, 2017).

Super-pojačivači

Koordinate SE uzorka $CD4^+$ T stanica iz krvi su preuzete s *The comprehensive human Super-Enhancer database* (SEdb) podatkovne baze (Yong i sur., 2019). Za svaki gen je nađen najbliži SE pomoću GenomicRanges paketa (Lawrence i sur., 2013) u programskom jeziku R (R core team, 2020). Geni unutar kojih se nalazi SE i oni od kojih je najbliži SE udaljen 10 ili manje kilobaza su određeni kao geni u blizini SE.

Korelacija i prediktivna snaga varijabli

Prediktivna snaga varijabli (https://github.com/8080labs/ppscore) je implementirana u programskom jeziku R (R Core Team, 2020). Za modeliranje korištena je metoda slučajnih šuma (Breiman, 2001; Ho, 1995). Kod za implementaciju je dostupan na https://github.com/mormart/hiv_integrations. Prikazi korelacija i prediktivna snaga varijabli su generirani pomoću R paketa corrplot (Wei i Simko, 2017).

Konstrukcija modela

Kako bi se izbjegli lažno pozitivni rezultati zbog eksperimentalnog šuma odnosno očekivane češće integracije velikih gena, prije modeliranja su uzorkovani geni iz istih kategorija veličine. Nasumično uzorkovanje je ponovljeno 10 puta i model nanovo treniran i kros-validiran kako bi se izbjegli slučajni učinci uslijed uzorkovanja. Ovakvo uzorkovanje je omogućilo i upotrebu ROC krivulja za usporedbu modela, koje u slučaju neuravnoteženih klasa nisu mjerodavne (Saito i Rehmsmeier, 2015).

Za podjelu podataka na trening set i set test set te za treniranje i kros-validaciju korišten je R paket caret (Kuhn, 2008). Za treniranje je uzeto 75% izvornih podataka, a za testiranje 25%, na način da omjer klasa RIG-ova i gena bez integracija ostane jednak u oba seta.

Za modeliranje je odabrana metoda slučajnih šuma (Breiman, 2001; Ho, 1995) koja se temelji na principu treniranja mnoštva stabala odlučivanja na nasumičnom dijelu podataka kako bi se smanjila varijanca i dekorelirala stabla (Hastie, Tibshirani i Friedman, 2008). Nakon treniranja stabala na različitim uzorcima podataka, za predviđanja na test setu se uzima prosjek predviđanja svih individualnih stabala u slučaju regresije:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

B - broj stabala,

f_b - stablo trenirano na dijelu podataka uzorkovanom B puta,

x' - nova opservacija za koju se predviđa vrijednost,

ili mod predviđanja odnosno najčešće predviđena klasa u slučaju klasifikacije.
Značajnost varijabli unutar modela je određena permutacijskim testom na način da se izračuna razlika u točnosti modela prije i nakon nasumičnog miješanja varijable čija značajnost se pokušava odrediti. Permutacijski test je računalno zahtjevan, ali su rezultati pouzdani i u slučaju koreliranih varijabli značajnost se dijeli između njih (https://explained.ai/rf-importance/#6.2).

Model slučajnih šuma je implementiran pomoću R paketa ranger (Wright i Ziegler, 2017) s parametrima num.trees = 1000, importance = "permutation", sample.fraction = 0.63, metric = "ROC". Deseterostrukom kros-validacijom su određeni najbolji hiperparametri mtry = 6, min.node.size = 1, splitrule = "extratrees".
Naučenim modelom je predviđena binarna varijabla koja označava je li gen RIG ili nije RIG u test setu. Kao mjera točnosti predviđanja konstruirana je *receiver operating characteristic* (ROC) krivulja (Hajian-Tilaki, 2013) i izračunata površina ispod krivulje (engl. *area under the curve* - AUC) koristeći R paket pROC (Robin i sur., 2011).

Da bi se odredile razlike u modelima i procijenila značajnost određenih varijabli, osim modela s uključenim svim prediktorima, izrađeni su i modeli bez prostornih i SE podataka (bez Hi-C/SE), bez histonskih oznaka i kromatinskog stanja (bez ChIP-seq), bez podataka o ekspresiji gena (bez RNA-seq) i samo s histonskim oznakama i kromatinskim stanjem (samo ChIP-seq). Za usporedbu modela korišten je DeLong test (DeLong i sur., 1988) kako bi se odredile značajne razlike između AUC-a ROC krivulja. Sve ROC su prikazane na istoj slici. Konstruirane su i *precision-recall* (PR) krivulje pomoću R paketa PRROC (Keilwagen, Grosse i Grau, 2015) za iste modele, izračunate AUC i krivulje također prikazane na jednoj slici.

Čitava konstrukcija modela je ponovljena 10 puta za svako od uzorkovanja po veličini gena i izračunate su prosječne AUC vrijednosti za ROC i PR krivulje.

Prediktorska varijabla	**Tip prediktora**	**Opis varijable**
H3K27ac	Numerička	Broj vrhova ChIP-seq signala za

		acetilaciju lizina 27 histona 3
H3K36me3	Numerička	Broj vrhova ChIP-seq signala za trimetilaciju lizina 36 histona 3
H3K4me3.TSS	Numerička	Pokrivenost ChIP-seq signala za trimetilaciju lizina 4 histona 3 na 200 parova baza oko mjesta početka transkripcije
H3K9me2	Numerička	Broj vrhova ChIP-seq signala za dimetilaciju lizina 9 histona 3
H4K20me1	Numerička	Broj vrhova ChIP-seq signala za metilaciju lizina 20 histona 4
ATACseq	Numerička	Broj vrhova ChIP-seq signala za kromatin dostupan transpozazi
readCountsLog	Numerička	Logaritam srednje vrijednosti broja RNA-seq sljedova koji se mapiraju u gen uvećane za 1
distToSE	Numerička	Udaljenost gena do najbližeg SE u parovima baza
SE	Kategorička	Binarna varijabla, je li najbliži SE udaljen manje ili više od 10 kilobaza
compartment	Kategorička	Genomski odjeljak
cmptsize	Numerička	Veličina genomskog odjeljka u parovima baza
gene.size	Numerička	Veličina gena u parovima baza
numberOfExons	Numerička	Broj egzona po genu

Tablica 1. Prikaz svih prediktora u modelu, tip svake varijable i opis varijabli.

4. Rezultati

Integracijska mjesta u genima

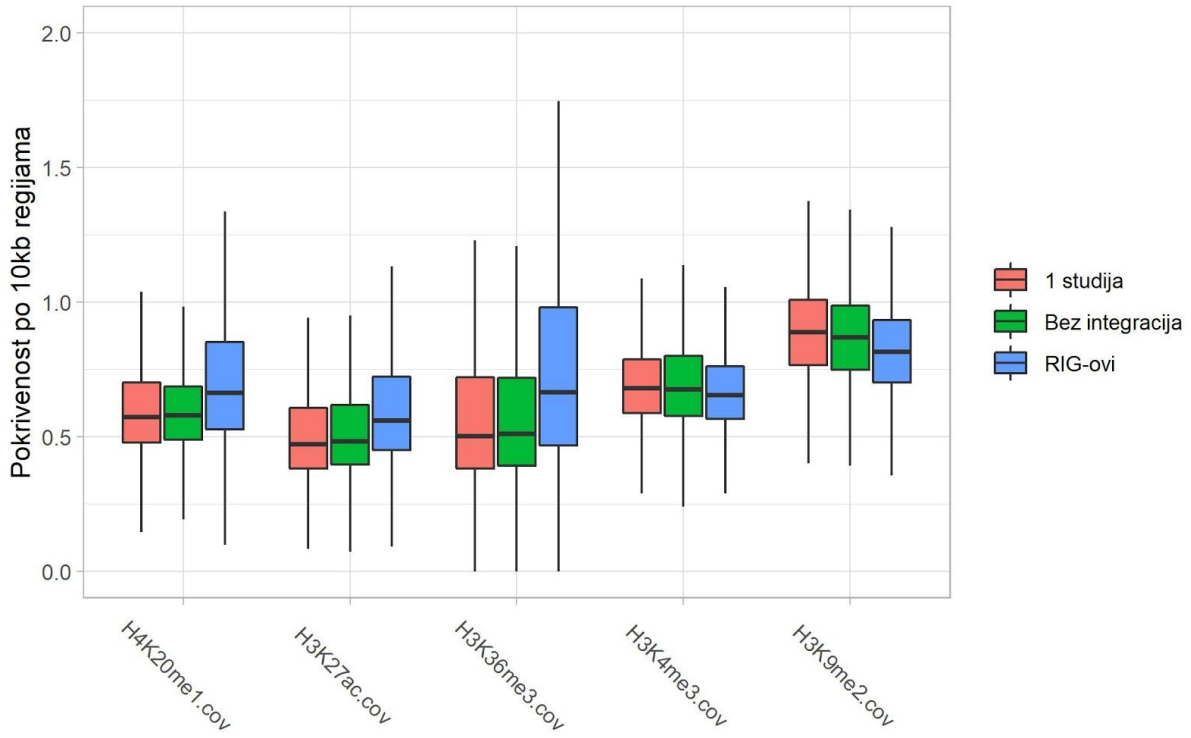
Za analizu integracija HIV-1 u gene uzeti su dostupni podaci iz 8 različitih studija (Lucic, Chen, Kuzman i sur., 2019; Kok i sur., 2016; Cohn i sur., 2015; Wagner i sur., 2014; Maldarelli i sur., 2014; Brady i sur., 2009; Ikeda i sur., 2007; Han i sur., 2004) na primarnim $CD4^+$ T limfocitima. Od 13544 ukupnih integracijskih mjesta, 9688 se nalazi u 14249 gena koji kodiraju proteine.

	Ukupno IM	**Broj IM u genima**	**% IM u genima**	**Ukupno gena**	**Broj gena s IM**	**% gena s IM**
Brady	862	692	80.28	14249	600	4.21
Coh	6414	4026	62.77	14249	2437	17.10
Han	74	59	79.73	14249	60	0.42
Ikeda	366	298	81.42	14249	264	1.85
Kok	497	421	84.71	14249	402	2.82
Lucic	3167	2431	76.76	14249	1864	13.08
Maldarelli	1723	1410	81.83	14249	1044	7.33
Wagner	441	351	79.59	14249	293	2.06

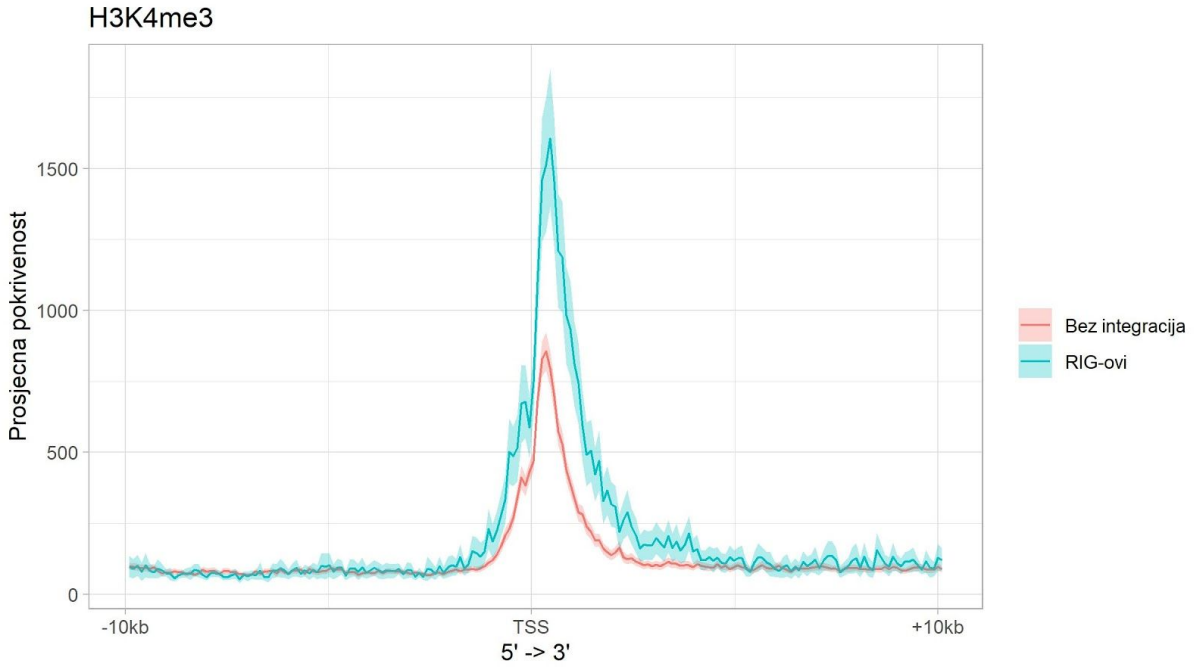
Tablica 2. Sažetak broja integracija po studijama: Ukupan broj jedinstvenih integracijskih mjesta po studiji (Ukupno IM), broj i postotak integracijskih mjesta iz svake studije koji se nalazi unutar gena (Broj i % IM u genima), ukupan broj gena koji kodiraju za proteine

(Ukupno gena), broj i postotak gena koji imaju barem jedno integracijsko mjesto za svaku studiju (Broj i % gena s IM).

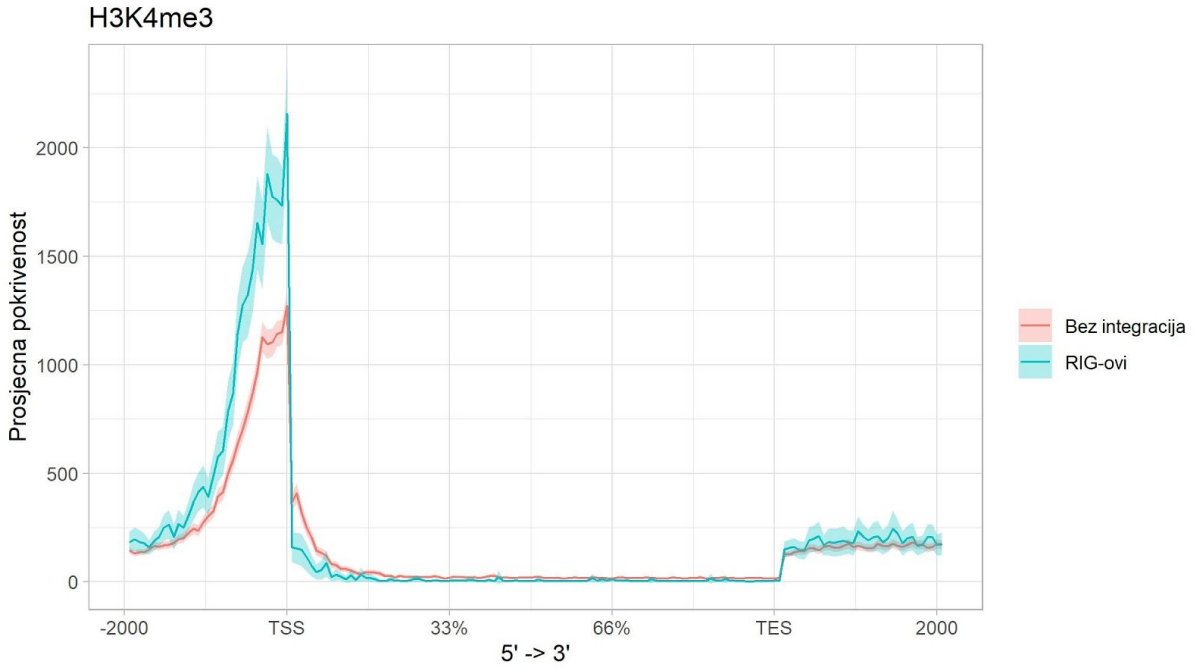
RIG-ovi su obogaćeni određenim histonskim modifikacijama



Slika 6. Normalizirana pokrivenost regija veličine 10 kilobaza različitim histonskim modifikacijama, podijeljenih prema tome preklapaju li se s RIG-ovima, genima koji samo u jednoj od 8 studija imaju integracije ili s genima bez integracija.

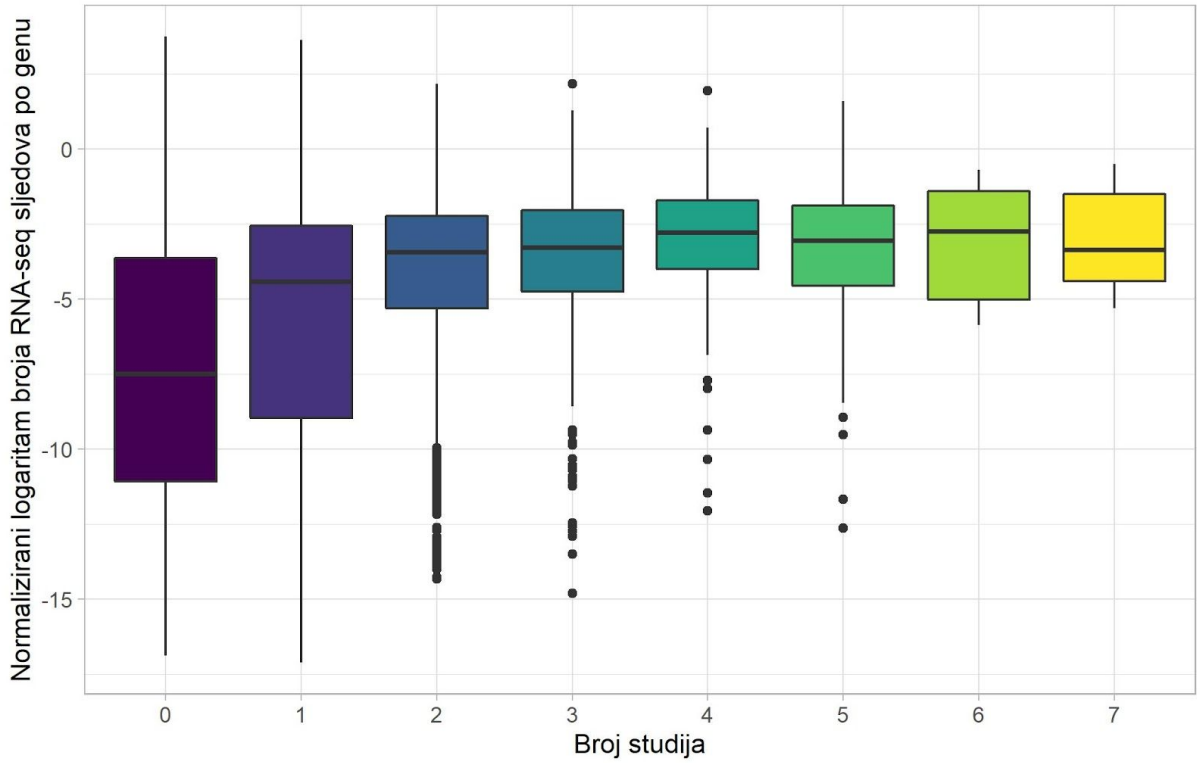


Slika 7. Profil pokrivenosti gena histonskom oznakom H3K4me3 u području 10 kilobaza prije i poslije mjesta početka transkripcije. Obojeno područje označava 95% interval pouzdanosti.



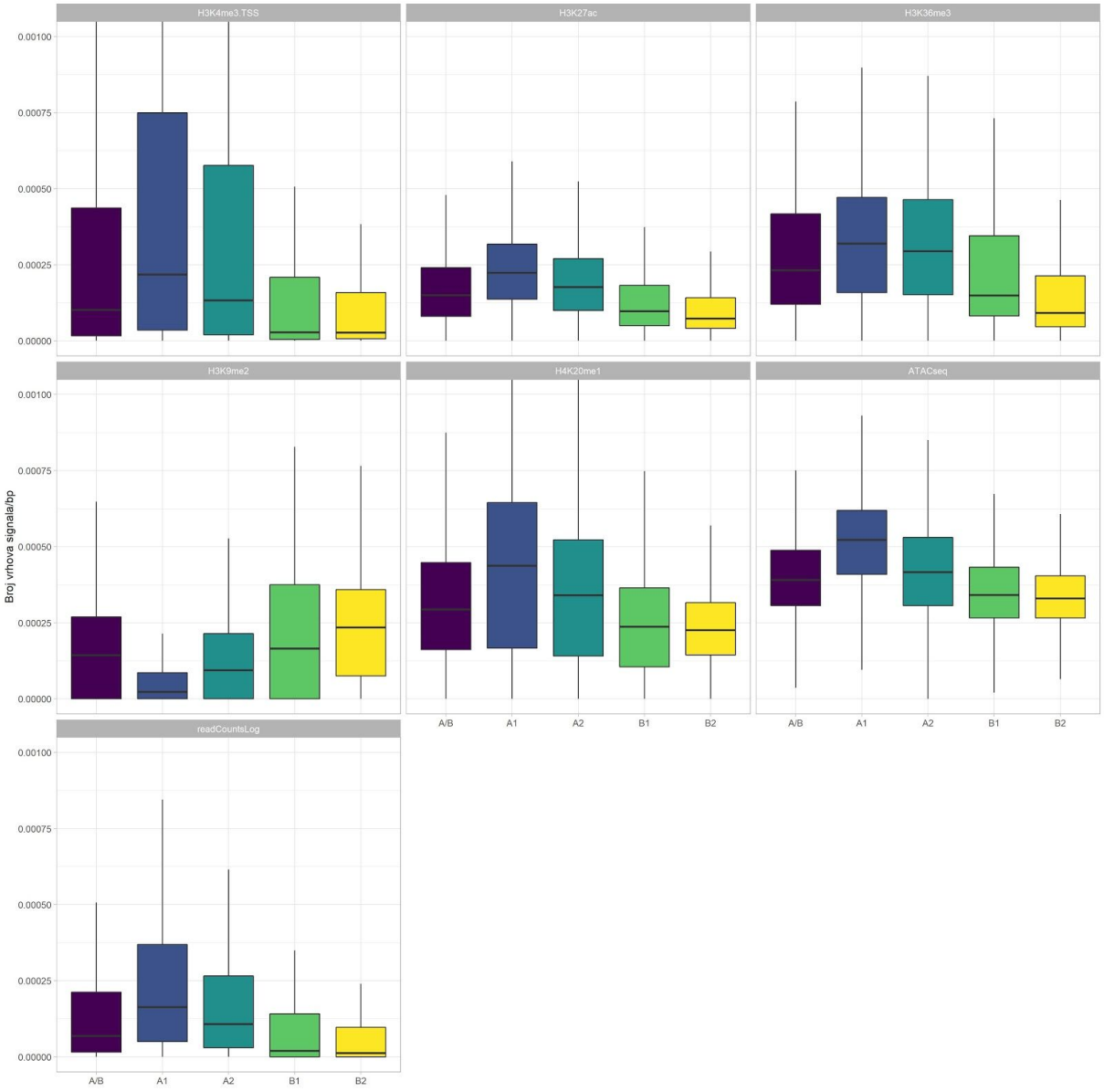
Slika 8. Profil pokrivenosti gena histonskom oznakom H3K4me3 u područjima 2 kilobaze prije mjesta početka transkripcije, tijelu gena podijeljenom na 100 dijelova i 2 kilobaze poslije mjesta završetka transkripcije. Obojeno područje označava 95% interval pouzdanosti.

Geni u više studija imaju višu razinu genske ekspresije



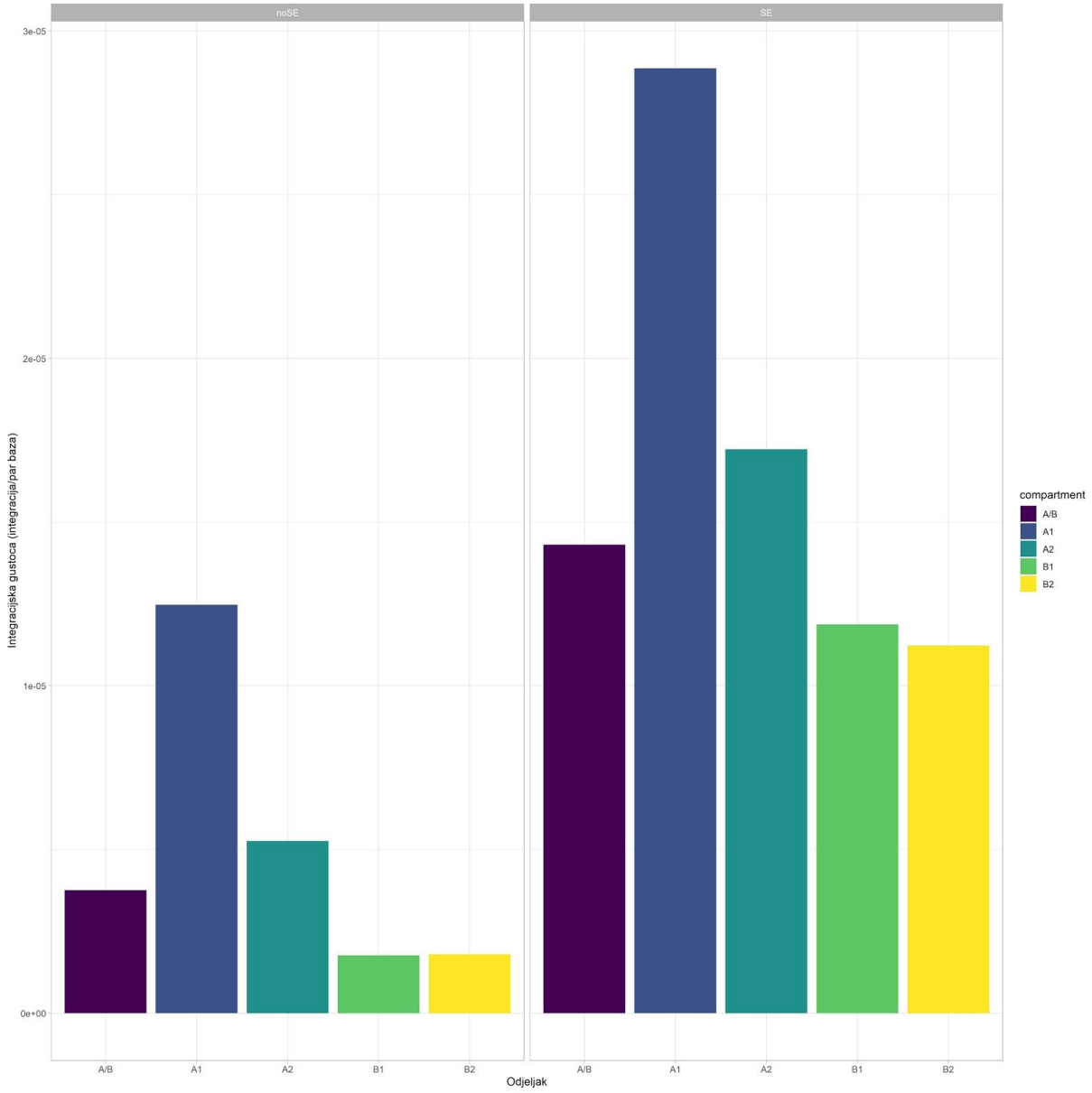
Slika 9. Prikaz razine genske ekspresije RNA-seq podataka po broju studija u kojima se nalaze geni.

Odjeljci genoma korelirani su s histonskim modifikacijama i genskom ekspresijom

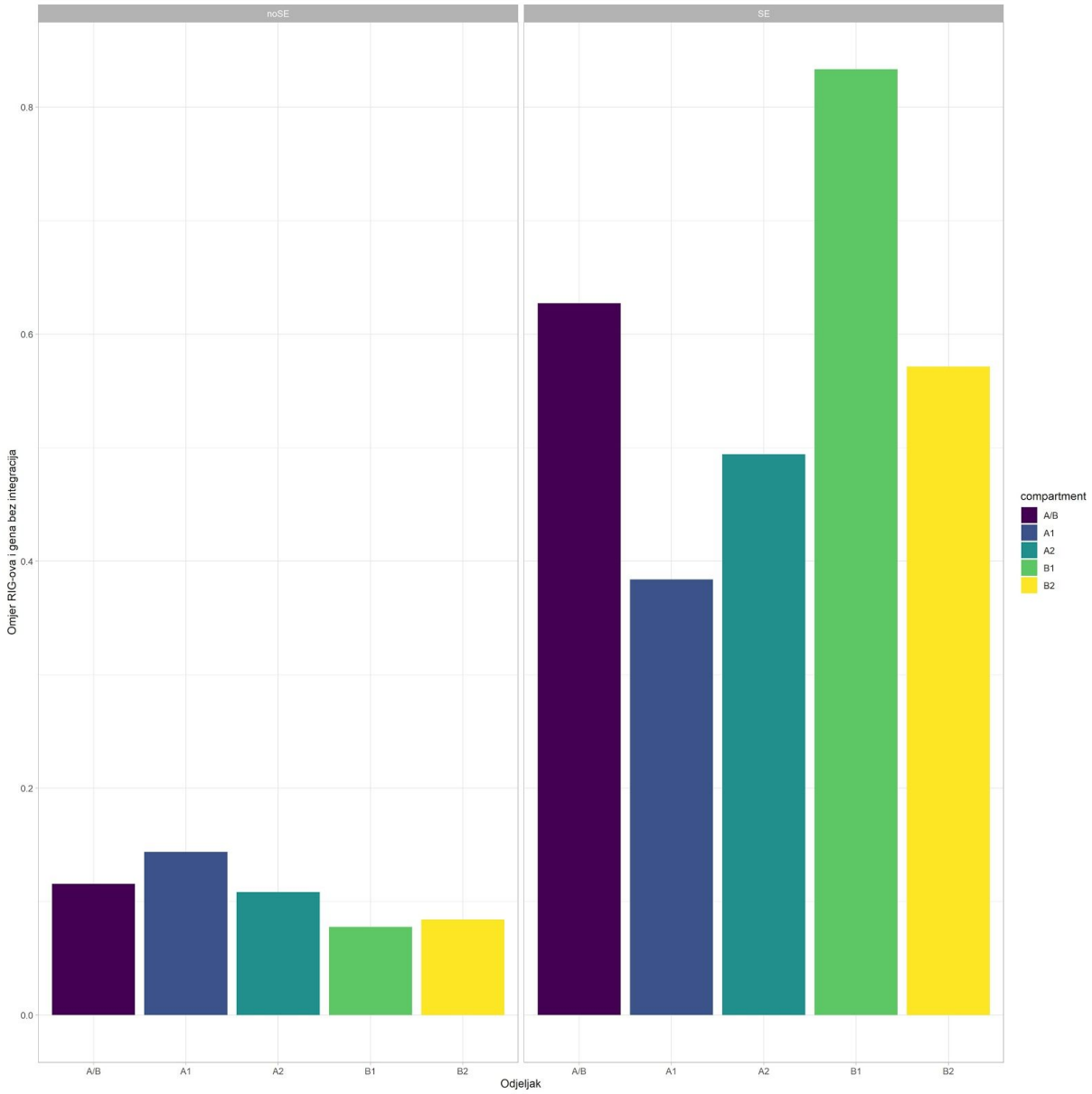


Slika 10. Profil pokrivenosti gena u različitim genomskim odjeljcima histonskim modifikacijama, otvorenim kromatinom i genskom ekspresijom.

Blizina SE ima veći utjecaj na RIG-ove nego na integracijsku gustoću

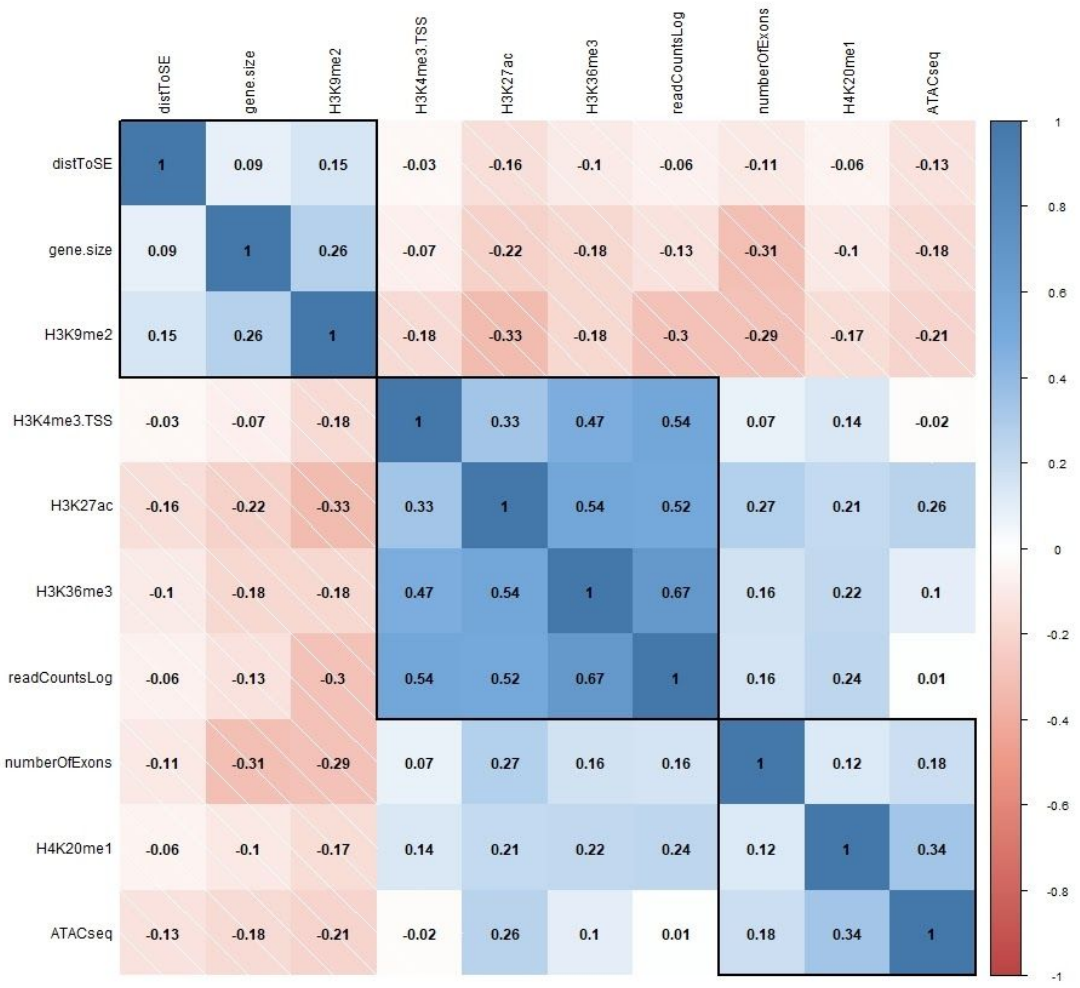


Slika 11. Srednja vrijednost integracijske gustoće (broj integracija po paru baza) unutar gena po genomskim odjeljcima, odvojeno za gene udaljene od SE više od 10 kilobaza (lijevo) i manje od 10 kilobaza (desno).

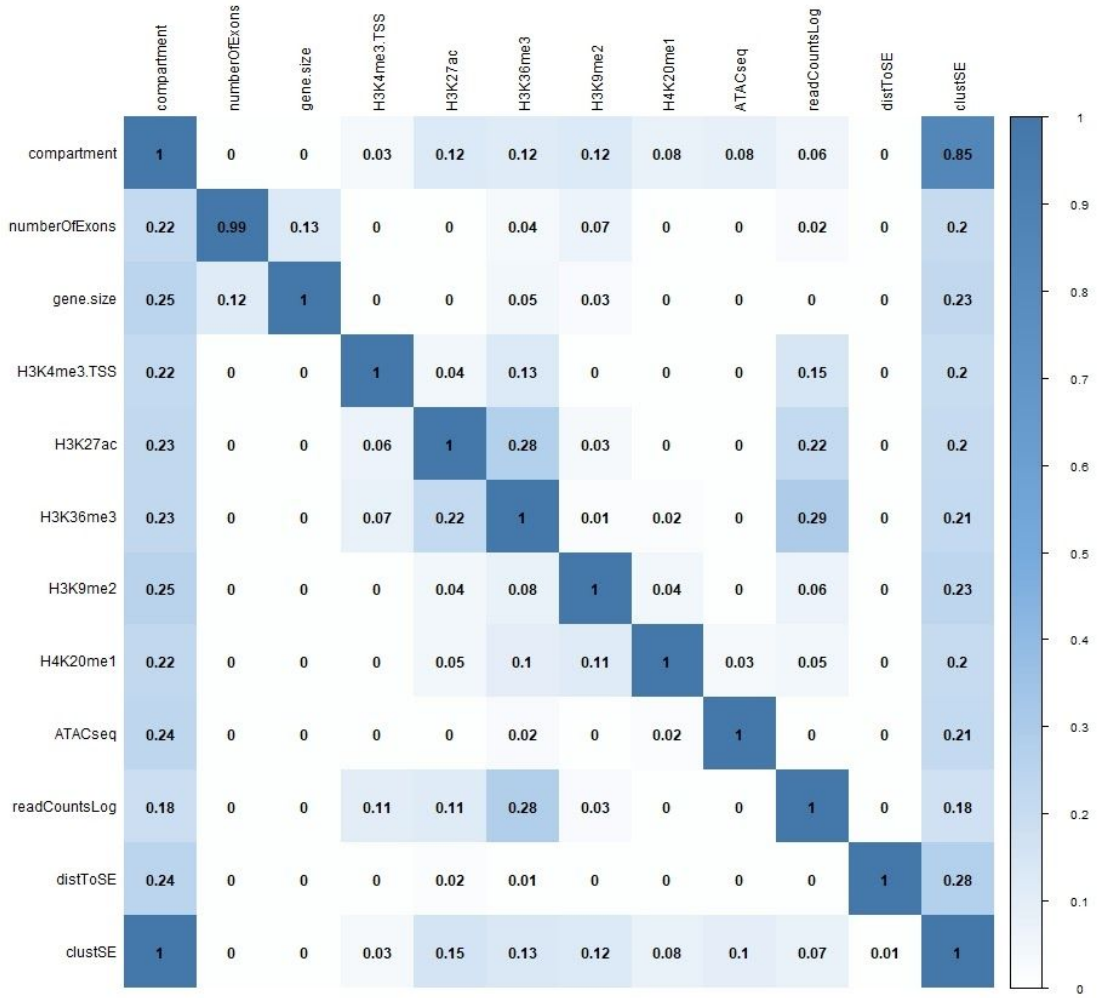


Slika 12. Srednja vrijednost omjera RIG-ova i gena bez integracija po genomskim odjeljcima, odvojeno za gene udaljene od SE više od 10 kilobaza (lijevo) i manje od 10 kilobaza (desno).

Korelacija i prediktivna snaga varijabli

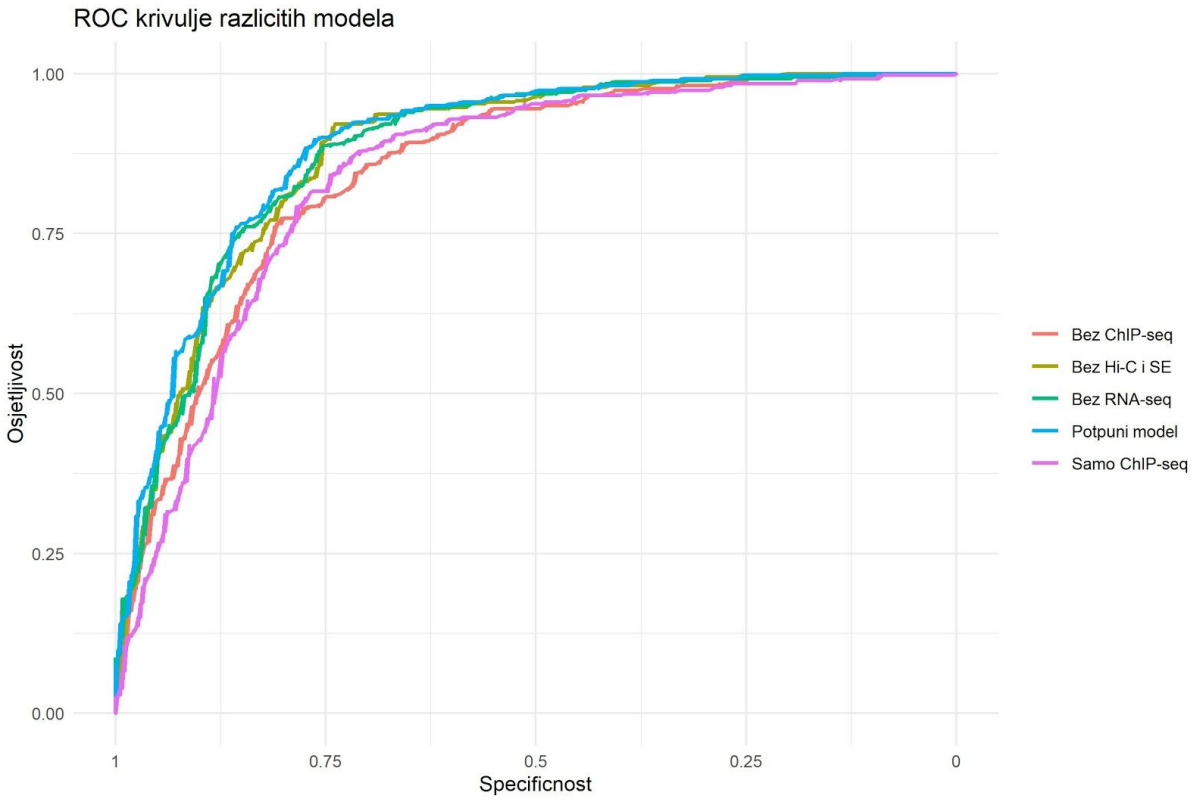


Slika 13. Matrica korelacija između numeričkih varijabli. Unutar kvadrata su grupirane varijable na temelju hijerarhijskog grupiranja (Johnson, 1967).



Slika 14. Matrica prediktivne snage varijabli. Vrijednost (i, j) matrice označava prediktivnu sposobnost i-te varijable s j-tom varijablom kao odgovorom.

ROC i PR krivulje



Slika 15. ROC krivulje za modele s različitim prediktorima. Ovdje je prikazan samo jedan reprezentativni set krivulja od 10 ponavljanja.

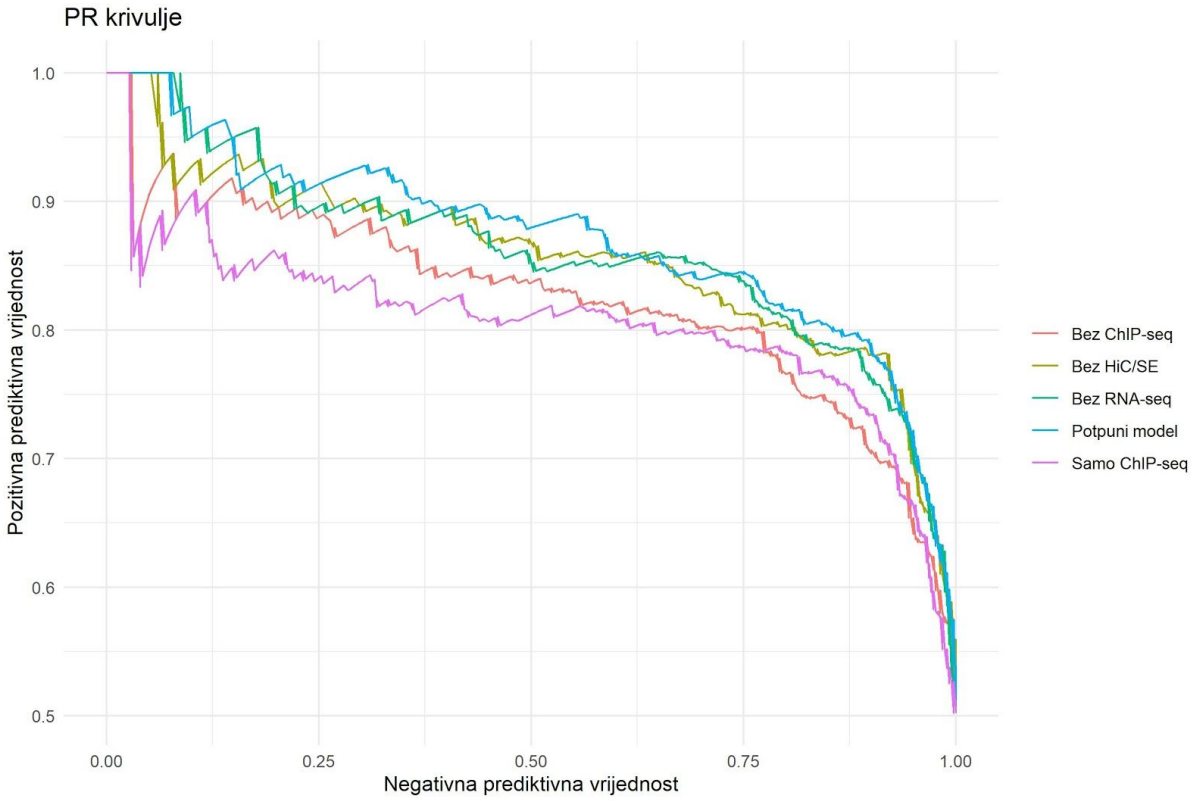
	Potpuni model	**Samo ChIP-seq**	**Bez Hi-C/SE**	**Bez ChIP-seq**	**Bez RNA-seq**
AUC	0.899	0.844	0.885	0.852	0.895

Tablica 3. Srednja vrijednost površine ispod krivulja za različite modele nakon 10 ponavljanja s različitim uzorcima gena bez integracija.

	Potpuni model	**Samo ChIP-seq**	**Bez Hi-C/SE**	**Bez ChIP-seq**	**Bez RNA-seq**
Potpuni model	**1**	**0**	**0.021**	**0**	**0.098**

Samo ChIP-seq	**0**	**1**	**0**	**0.6**	**0**
Bez Hi-C/SE	**0.021**	**0**	**1**	**0.005**	**0.113**
Bez ChIP-seq	**0**	**0.6**	**0.005**	**1**	**0**
Bez RNA-seq	**0.098**	**0**	**0**	**0**	**1**

Tablica 4. Matrica p-vrijednosti DeLong testa korelacija parova ROC krivulja različitih modela (DeLong i sur., 1988) nakon 10 ponavljanja s različitim uzorcima gena bez integracija. AUC potpunog modela se značajno razlikuje od svih drugih modela osim modela bez RNA-seq podataka na razini značajnosti $p < 0.05$.



Slika 16. PR krivulje za modele s različitim prediktorima nakon 10 ponavljanja s različitim uzorcima gena bez integracija.

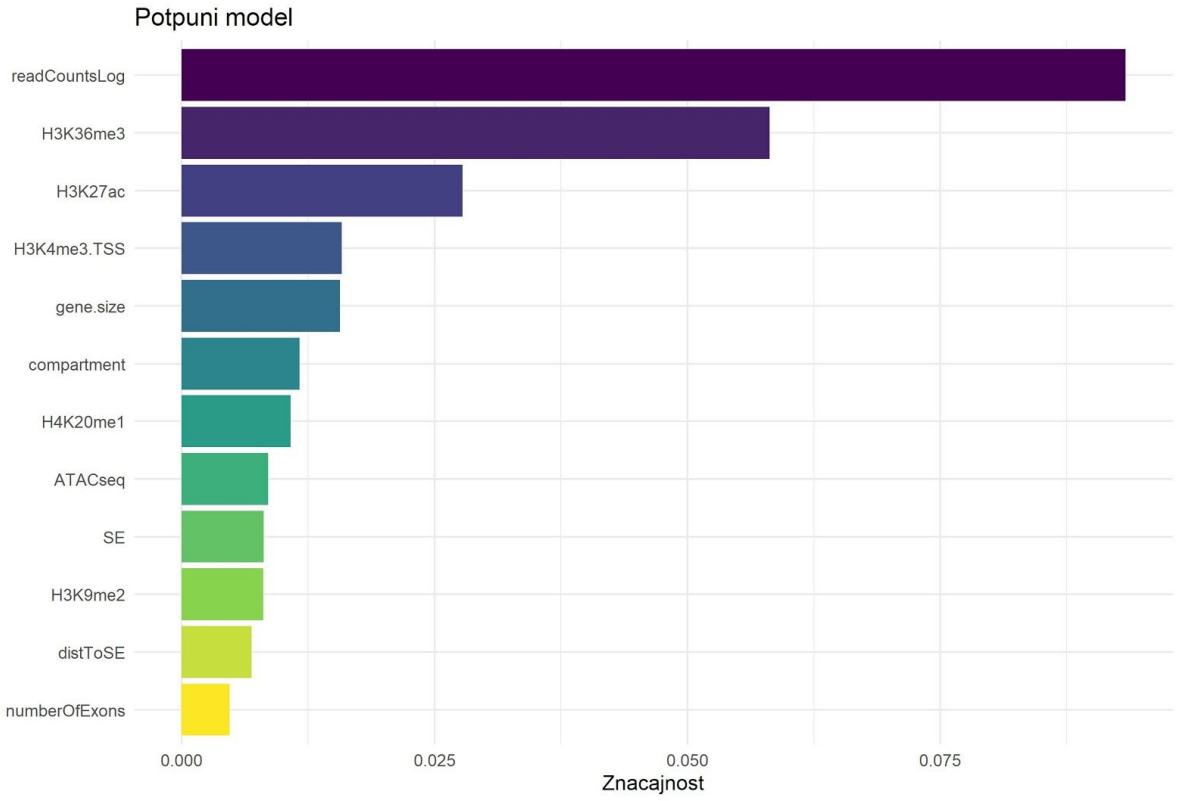
Matrica konfuzije

```
Confusion Matrix and Statistics

                Reference
Prediction       integrated not_integrated
  integrated            309              86
  not_integrated         71             290

               Accuracy : 0.7923
                 95% CI : (0.7616, 0.8207)
    No Information Rate : 0.5026
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5846

 Mcnemar's Test P-Value : 0.2639

              Precision : 0.7823
                 Recall : 0.8132
                     F1 : 0.7974
             Prevalence : 0.5026
         Detection Rate : 0.4087
   Detection Prevalence : 0.5225
      Balanced Accuracy : 0.7922

       'Positive' Class : integrated
```

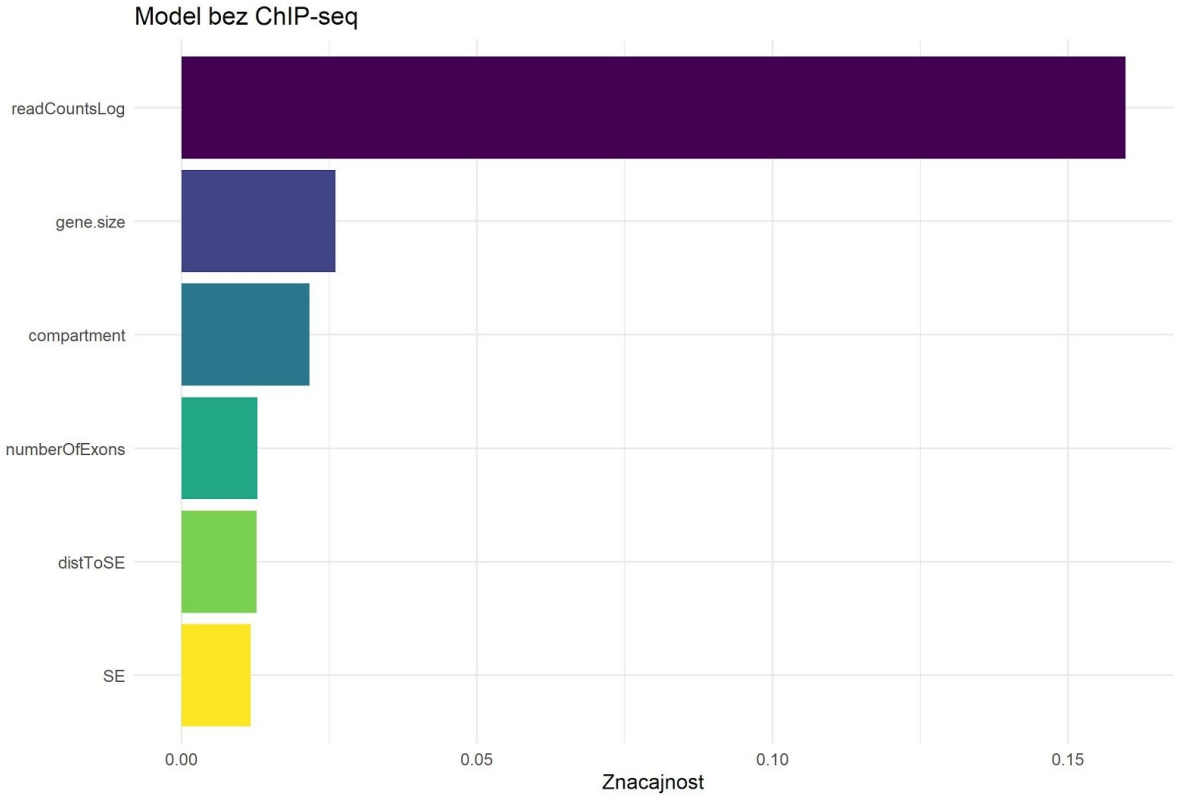
Slika 17. Matrica konfuzije (Stehman, 1997) za potpuni model, konstruirana pomoću R paketa caret (Kuhn, 2008).

Hijerarhija značajnosti pojedinih varijabli

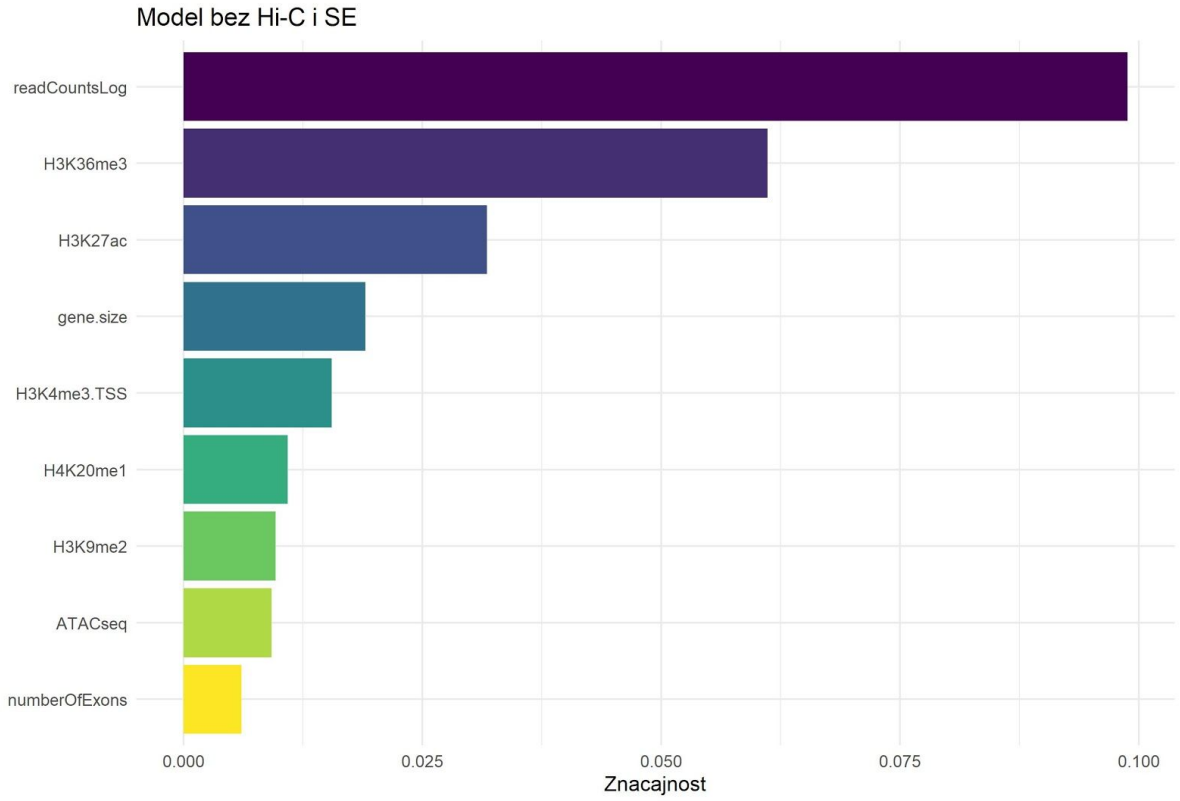
Značaj pojedinih varijabli u modelu je izračunat permutacijskim testom implementiranim u R paketu ranger (Wright i Ziegler, 2017).



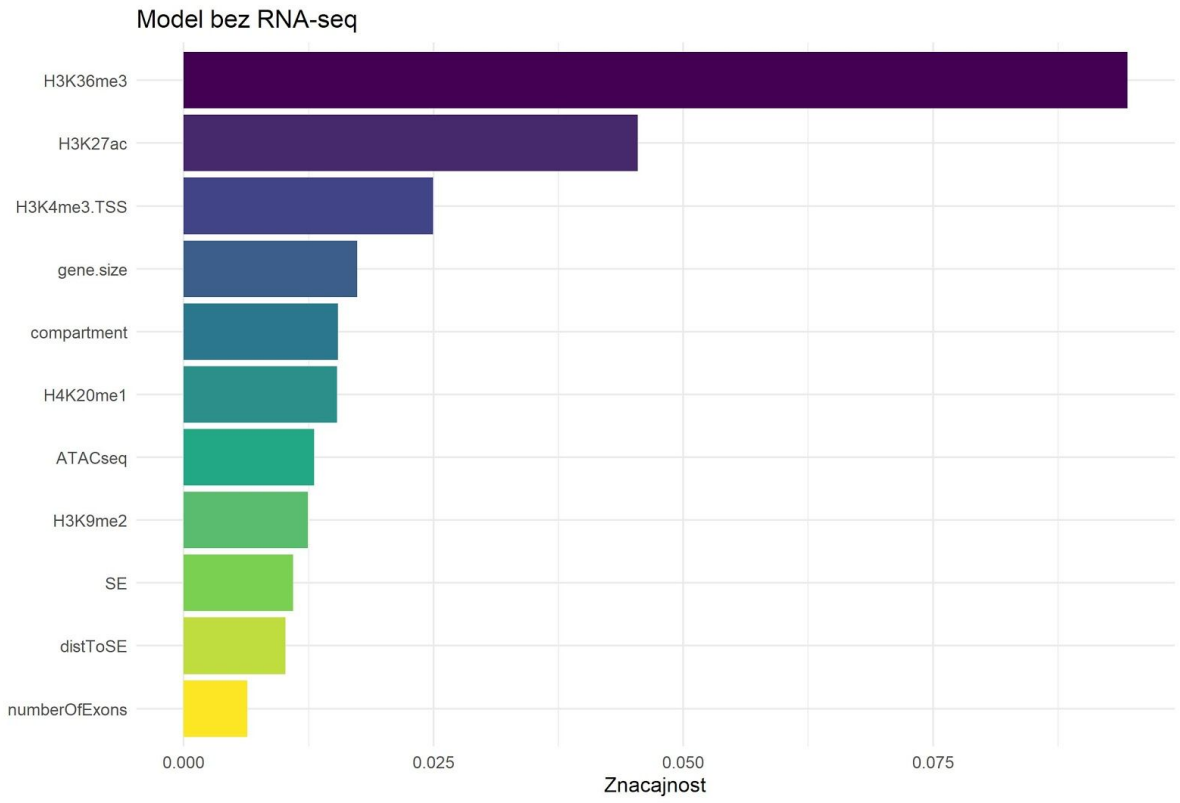
Slika 18. Hijerarhija značajnosti pojedine varijable za model sa svim prediktorima.



Slika 19. Hijerarhija značajnosti pojedine varijable za model bez ChIP-seq podataka (H3K27ac, H3K36me3, H3K9me2, H4K20me1, H3K4me3, ATAC-seq).



Slika 20. Hijerarhija značajnosti pojedine varijable za model bez Hi-C i SE podataka.



Slika 21. Hijerarhija značajnosti pojedine varijable za model bez RNA-seq podataka.

5. Rasprava

Poznato je da HIV-1 integrira aktivno prepisujuće gene bogate aktivnim histonskim oznakama (Turner i Margolis, 2017; Wang i sur.. 2007) i prostorno povezane odjeljke genoma s više SE (Lucic, Chen, Kuzman i sur., 2019), no nije kvantificirano koliko trodimenzionalna organizacija prostora u aktiviranim $CD4^+$ T limfocitima bolje objašnjava mjesta integracije u odnosu na jednodimenzionalna genomska svojstva, s obzirom da se i sama trodimenzionalna organizacija genoma stanice može predvidjeti pomoću histonskih modifikacija (Fortin i Hansen, 2015).

Također do sada nije bilo pokušaja implicitnog istraživanja, prediktivnim modeliranjem, svih veza između mjesta integracije virusa HIV-1 i velikog broja genomskih čimbenika koji mogu utjecati na integraciju. Ovdje se pokušala kvantificirati njihova relativna značajnost upotrebom računalnih metoda strojnog učenja.

S obzirom da su integracije HIV-1 rijedak proces, tj. događaju se u malom postotku stanica, skupljeni su podaci mjesta integracija iz 8 studija (Lucic, Chen, Kuzman i sur., 2019; Kok i sur., 2016; Cohn i sur., 2015; Wagner i sur., 2014; Maldarelli i sur., 2014; Brady i sur., 2009; Ikeda i sur., 2007; Han i sur., 2004). RIG su definirani kao geni koji imaju barem jednu integraciju u 2 ili više od tih 8 studija.

U radu Lucic, Chen, Kuzman i sur., 2019 pokazano je da je različit proces posrijedi pri predviđanju gena s najvećom gustoćom integracija (broj integracija po paru baza), koji su gotovo u potpunosti određeni transkripcijskom aktivnošću, i žarišta integracija, koja imaju multifaktorijalne značajne odrednice, uključujući i genomske odjeljke. Stoga je za modeliranje izabran binarni odgovor je li gen RIG ili nije i u model su uz razinu genske ekspresije, histonske modifikacije, otvorenost kromatina i blizinu SE, uključeni i Hi-C prostorni podaci.

Po epigenetskom profilu i veličini odjeljaka vidi se da odjeljci u ovom radu odgovaraju odjeljcima u Lucic, Chen, Kuzman i sur., 2019, stoga su referirani A-B oznakama kao u navedenom radu, prema profilu na slici 10. A1 je najbogatiji aktivnim histonskim oznakama i visoko eksprimiranim genima i ima najizraženiji interakcijski obrazac. A2 je sličan A1, ali ima manju gustoću gena unutar odjeljka i slabiji interakcijski obrazac. A/B odjeljak pokazuje srednje vrijednosti histonskih oznaka. B1 ima interakcijski obrazac komplementaran A1. B2 je sličan B1, ali je najveći odjeljak, ima jako malu gensku gustoću i najbogatiji je represivnom oznakom H3K9me2. A1 odjeljak je također obogaćen genima koji u blizini imaju SE elemente, što je u skladu s pronalascima iz Lucic, Chen, Kuzman i sur., 2019.

Na različitu važnost genomskih odjeljaka u predviđanju gena s najvećom gustoćom integracija i RIG-ova upućuju i rezultati ovog rada. A1 odjeljak je najbogatiji otvorenim kromatinom i aktivnim histonskim oznakama te ima najveću integracijsku gustoću i omjer RIG-ova u odnosu na gene bez integracija. Jako je zanimljivo opažanje da gustoća integracija po odjeljcima ostaje očuvana neovisno o blizini SE genima, ali omjer RIG-ova unutar odjeljaka se mijenja u blizini SE. Vidljivo je da A1 odjeljak u blizini SE ima najmanji omjer RIG-ova i gena bez integracija, a B1 i B2 najveći, obrnuto od gena bez SE u blizini. Dakle, SE su značajni prediktori RIG-ova u odjeljcima genoma siromašnim integracijama s manje

SE ukupno. To upućuje na međuigru prostornog rasporeda genoma i lokalizacije SE unutar odjeljaka.

Pokrivenost aktivnim histonskim oznakama (H3K27ac, H3K36me3, H4K20me1) koje koreliraju s otvorenim kromatinom dostupnim transpozazi (ATAC-seq) je očekivano veća za RIG-ove nego za gene bez integracija i obrnuto za represivnu oznaku H3K9me2 (Turner i Margolis, 2017). H3K4me3 je oznaka aktivne transkripcije koja nije obogaćena unutar RIG-ova, ali s obzirom da se većinom nalazi u području aktivnih pojačivača (Spicuglia i Vanhille, 2012) očekivano je da će biti obogaćena za RIG-ove u susjedstvu mjesta početka transkripcije, što se pokazalo točnim. Stoga su za modeliranje korišteni brojevi vrhova signala kromatinskih oznaka po paru baza gena, osim za H3K4me3 za koji je izračunata pokrivenost unutar 2 kilobaze oko mjesta početka transkripcije gena.

Za modeliranje je odabrana metoda slučajnih šuma jer korigira sklonost stabala odlučivanja tzv. *overfittingu*, odnosno hvatanju slučajnih uzoraka u trening setu koji ne postoje u test setu podataka (Hastie, Tibshirani i Friedman, 2008). Metoda daje modele točnosti slične kao i kompliciraniji modeli, robusna je prema iznimkama, eksperimentalnom šumu i koreliranim varijablama, brza je i daje dobre procjene pogreške, značajnosti i korelacije varijabli, prihvaća numeričke i kategoričke varijable kao prediktore (Breiman, 2001).

Za modeliranje su uzeti samo geni koji kodiraju proteine i njihova integracijska mjesta zbog brže i preciznije analize. Manje podataka ubrzava treniranje modela, što je poželjno s obzirom na broj ponavljanja uzorkovanja i zahtjevnost permutacijskog testa. Ovime nije izgubljeno puno informacija s obzirom da je većina integracija u protein kodirajućim genima. Stoga uključivanje ostalih tipova gena ne bi značajno doprinjelo zaključcima a dovelo bi do teže interpretabilnog modela.

Učinak svih modela na test setu je dobar i postiže se prilično točna diskriminacija RIG-ova i gena bez integracija. Može se primijetiti da se relativna značajnost varijabli potpunog modela ne poklapa s razlikom ROC krivulja modela, najviše očitovano kod ekspresije gena. Ekspresija gena je prema permutacijskom testu najbolji prediktor RIG-ova u potpunom modelu, no AUC modela bez ekspresije gena se od potpunog modela razlikuje manje nego

model bez histonskih oznaka. Takvi rezultati su najvjerojatnije posljedica koreliranosti varijabli odnosno mogućnosti ekspresije gena da samostalno predvidi histonske modifikacije H3K36me3 i H3K27ac i obrnuto, kao i same koreliranosti aktivnih histonskih modifikacija (što se vidi iz matrice prediktivne snage varijabli), a poznato je da se uslijed koreliranosti varijabli značajnost kod permutacijskog testa dijeli između tih varijabli (https://explained.ai/rf-importance/#6.2). Iz istih razloga relativne značajnosti SE i prostornih odjeljaka pojedinačno nisu visoke, no skupa su informativne što se i potvrdilo ROC analizom. Još jedno neočekivano opažanje je da je veličina gena relativno značajan prediktor unatoč uzorkovanju, što može biti povezano s alternativnim prekrajanjem u pozadini HIV-1 integracije (Sertznig i sur., 2018) i potrebno je dalje istražiti.

Ovakav način prediktivnog modeliranja potvrđuje gensku ekspresiju i epigenetske histonske oznake kao glavnu odrednicu RIG-ova, ali je metoda ujedno i dovoljno osjetljiva da pronađe utjecaj ostalih varijabli i pruži uvid u njihovu važnost. Bitan pronalazak je utjecaj prostornih odjeljaka i SE neovisno o epigenetskim oznakama i potvrda da sama ekspresija gena ne objašnjava u potpunosti diskriminaciju RIG-ova i gena bez integracija. Ipak, i dalje je teško donijeti biološke zaključke o kauzalnosti pri kompartmentalizaciji integracija s obzirom da su sami odjeljci korelirani s određenim profilom histonskih oznaka. Da bi se to razjasnilo, potrebna su dodatna istraživanja u kontekstu rearanžmana nuklearnog prostora pri aktivaciji T limfocita. Ostali zaključci su u skladu s dosadašnjim pronalascima što potvrđuje ovakve računalne metode analize kao valjan pristup koji dovodi do konzistentnih i reproducibilnih zaključaka.

Nastavno na ovaj rad slične metode bi se mogle primijeniti za analizu T limfocita u različitim fazama infekcije i u različitim uvjetima kako bi se potencijalno razjasnile genomske odrednice nastanka latentnog spremnika HIV-a (Chen i sur., 2017). Takva saznanja bi se mogla primijeniti u razvoju terapije za AIDS.

Također, ovdje razvijene i opisane računalne analitičke metode mogu biti korisne za daljnje istraživanje mnogih aspekata retroviralne DNA integracije.

6. Zaključci

1. Nuklearni prostor je podijeljen na barem pet odjeljaka unutar kojih se očituje homogenost preferabilnih interkromosomalnih interakcijskih uzoraka.

2. Interakcije i lokalizacija super-pojačivača unutar nuklearnih odjeljaka ljudskog genoma koreliraju s HIV-1 integracijskim profilom.

3. Razina genske ekspresije i epigenetski profil histonskih modifikacija su najizraženije odrednice gena u koje se HIV-1 rekurentno integrira, a očituje se povećanom pokrivenošću acetilacije lizina 27 histona 3 i trimetilacije lizina 36 histona 3 unutar cijelih gena i pokrivenošću trimetilacije lizina 4 histona 3 u području oko mjesta početka transkripcije.

4. Prediktivno modeliranje otkriva da geni u koje se virus HIV-1 rekurentno integrira imaju multifaktorijalne značajne odrednice, prvenstvno aktivne histonske modifikacije i povišenu razinu genske ekspresije, ali uključuju i genomske odjeljke i prisutnost super-pojačivača.

5. SE su bolji prediktori RIG-ova nego gena s najvećom gustoćom integracija i bitnija su odrednica RIG-ova unutar odjeljaka siromašnih aktivnim histonskim modifikacijama i siromašnih genima od onih unutar transkripcijski aktivnih odjeljaka.

6. Metodama strojnog učenja je moguće postići dobru diskriminaciju između gena u koje se HIV-1 rekurentno integrira i gena bez integracija, prilikom čega kao prediktore koristimo razinu genske ekspresije, histonske modifikacije, prostorne Hi-C podatke i udaljenost gena do super-pojačivača.

7. Relativna značajnost prediktora integracija, određena permutacijskim testom, nije nužno reproducibilna, no usporedbom različitih modela i opreznim odabirom možemo odrediti reproducibilno značajne prediktore.

8. Računalne metode i prediktivno modeliranje dovode do konzistentnih zaključaka i koristan su alat za proučavanje retroviralne integracije u genom.

7. Literatura

Achuthan V, Perreira JM, Sowd GA, i sur. Capsid-CPSF6 Interaction Licenses Nuclear HIV-1 Trafficking to Sites of Viral DNA Integration. *Cell Host Microbe*, 2018, 24(3), 392-404.

Agosto LM, Yu JJ, Liszewski MK, i sur. The CXCR4-tropic human immunodeficiency virus envelope promotes more-efficient gene delivery to resting CD4+ T cells than the vesicular stomatitis virus glycoprotein G envelope. *J Virol*, 2009, 83(16), 8153-8162.

Beagrie RA, Scialdone A, Schueler M, i sur. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 2017, 543(7646), 519-524.

Bejarano DA, Peng K, Laketa V, i sur. HIV-1 nuclear import in macrophages is regulated by CPSF6-capsid interactions at the nuclear pore complex. *Elife*, 2019.

Bonev B, Cavalli G. Organization and function of the 3D genome. *Nat Rev Genet*, 2016, 17(11), 661-678.

Brady T, Agosto LM, Malani N, Berry CC, O'Doherty U, Bushman F. HIV integration site distributions in resting and activated CD4+ T cells infected in culture. *AIDS*, 2009, 23(12), 1461-1471.

Breiman L. Random Forests. *Machine Learning*, 2001, 45, 5–32.

Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol*, 2015, 109, 21.29.1-21.29.9.

Chen HC, Martinez JP, Zorita E, Meyerhans A, Filion GJ. Position effects influence HIV latency reversal. *Nat Struct Mol Biol*, 2017, 24(1), 47-54.

Cherepanov P, Maertens G, Proost P, i sur. HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J Biol Chem*, 2003, 278(1), 372-381.

Chomont N, El-Far M, Ancuta P, i sur. HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat Med*, 2009, 15(8), 893-900.

Churchill MJ, Deeks SG, Margolis DM, Siliciano RF, Swanstrom R. HIV reservoirs: what, where and how to target them. *Nat Rev Microbiol*, 2016, 14(1), 55-60.

Ciuffi A, Llano M, Poeschla E, i sur. A role for LEDGF/p75 in targeting HIV DNA integration. *Nat Med*., 2005, 11(12), 1287-1289.

Clough E, Barrett T. The Gene Expression Omnibus Database. *Methods Mol Biol*, 2016, 1418, 93-110.

Cohn LB, Silva IT, Oliveira TY, i sur. HIV-1 integration landscape during latent and active infection. *Cell*, 2015, 160(3), 420-432.

Craigie R, Bushman FD. HIV DNA integration. *Cold Spring Harb Perspect Med*, 2012, 2(7).

Dai J, Agosto LM, Baytop C, i sur. Human immunodeficiency virus integrates directly into naive resting CD4+ T cells but enters naive cells less efficiently than memory cells. *J Virol*, 2009, 83(9), 4528-4537.

DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 1988, 44(3), 837-845.

Desrosiers RC, Letvin NL. Animal models for acquired immunodeficiency syndrome. *Rev Infect Dis*, 1987, 9(3), 438-446.

Di Nunzio F, Fricke T, Miccio A, i sur. Nup153 and Nup98 bind the HIV-1 core and contribute to the early steps of HIV-1 replication. *Virology*, 2013, 440(1), 8-18.

Dobin A, Davis CA, Schlesinger F, i sur. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 2013, 29(1), 15-21.

Dowle M, Srinivasan A. data.table: Extension of `data.frame`, R package version 1.12.8, https://CRAN.R-project.org/package=data.table, pristupljeno 9.5.2020.

Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*, 2009, 4(8), 1184-1191.

Fauci AS, Desrosiers RC. Pathogenesis of HIV and SIV. U: Retroviruses. Coffin JM, Hughes SH, Varmus HE, urednici, New York, Cold Spring Harbor Laboratory Press, 1997).

Fortin JP, Hansen KD. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol*, 2015, 16(1), 180.

Gallo RC, Salahuddin SZ, Popovic M, i sur. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*, 1984, 224(4648), 500-503.

Gilbert PB, McKeague IW, Eisen G, Mullins C, Guéye‐NDiaye A, Mboup S, Kanki PJ. Comparison of HIV‐1 and HIV‐2 infectivity from a prospective cohort study in Senegal. *Statist. Med*, 2003, 22, 573-593.

Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med*, 2013, 4(2), 627-635.

Han Y, Lassen K, Monie D, i sur. Resting CD4+ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. *J Virol*, 2004, 78(12), 6122-6133.

Hastie T, Tibshirani R, Friedman J. Random Forests. U: The Elements of Statistical Learning. New York, Springer New York Inc., 2001, str. 587-603.

Hnisz D, Abraham BJ, Lee TI, i sur. Super-enhancers in the control of cell identity and disease. *Cell*, 2013, 155(4), 934-947.

Hnisz D, Day DS, Young RA. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell*, 2016, 167(5), 1188-1200.

Ho TK. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, 1, 278-282.

Ibarra A, Benner C, Tyagi S, Cool J, Hetzer MW. Nucleoporin-mediated regulation of cell identity genes. *Genes Dev*, 2016, 30(20), 2253-2258.

Ikeda T, Shibata J, Yoshimura K, Koito A, Matsushita S. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. *J Infect Dis*, 2007, 195(5), 716-725.

Imakaev M, Fudenberg G, McCord RP, i sur. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, 2012, 9(10), 999-1003.
Johnson SC. Hierarchical clustering schemes. *Psychometrika*, 1967, 32(3), 241-254.

Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 2007, 316(5830), 1497-1502.

Grau J, Grosse I, Keilwagen J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, 2015, 31(15), 2595-2597.

Koh Y, Wu X, Ferris AL, i sur. Differential effects of human immunodeficiency virus type 1 capsid and cellular factors nucleoporin 153 and LEDGF/p75 on the efficiency and specificity of viral DNA integration. *J Virol*, 2013, 87(1), 648-658.

Kok YL, Vongrad V, Shilaih M, i sur. Monocyte-derived macrophages exhibit distinct and more restricted HIV-1 integration site repertoire than CD4(+) T cells. *Sci Rep*, 2016, 6, 24157.

Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 2008, 28(5), 1-26.

Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods*, 2015, 72, 65-75.

Lawrence M, Huber W, Pagès H, i sur. Software for computing and annotating genomic ranges. *PLoS Comput Biol*, 2013, 9(8).

Lelek M, Casartelli N, Pellin D, i sur. Chromatin organization at the nuclear pore favours HIV replication. *Nat Commun*, 2015, 6, 6483.

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 2018, 34(18), 3094-3100.

Li H, Handsaker B, Wysoker A, i sur. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009, 25(16), 2078-2079.

Lieberman-Aiden E, van Berkum NL, Williams L, i sur. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 2009, 326(5950), 289-293.

Lister R, O'Malley RC, Tonti-Filippini J, i sur. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 2008, 133(3), 523-536.

Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 1982, 28(2), 129-137.

Lucic B, Chen HC, Kuzman M, i sur. Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration. *Nat Commun*, 2019, 10(1), 4059.

Lusic M, Siliciano RF. Nuclear landscape of HIV-1 infection and integration. *Nat Rev Microbiol*, 2017, 15(2), 69-82.

Maldarelli F, Wu X, Su L, i sur. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science*, 2014, 345(6193), 179-183.

Margolis DM, Garcia JV, Hazuda DJ, Haynes BF. Latency reversal and viral clearance to cure HIV-1. *Science*, 2016, 353(6297).

Marini B, Kertesz-Farkas A, Ali H, i sur. Nuclear architecture dictates HIV-1 integration site selection. *Nature*, 2015, 521(7551), 227-231.

Martin AR, Siliciano RF. Progress Toward HIV Eradication: Case Reports, Current Efforts, and the Challenges Associated with Cure. *Annu Rev Med*, 2016, 67, 215-228.

Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D. The dynamic architecture of Hox gene clusters. *Science*, 2011, 334(6053), 222-225.

Ocwieja KE, Brady TL, Ronen K, i sur. HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog*, 2011, 7(3).

Olley G, Ansari M, Bengani H, i sur. BRD4 interacts with NIPBL and BRD4 is mutated in a Cornelia de Lange-like syndrome. *Nat Genet*., 2018, 50(3), 329-332.

Pace MJ, Graf EH, Agosto LM, i sur. Directly infected resting CD4+T cells can produce HIV Gag without spreading infection in a model of HIV latency. *PLoS Pathog*, 2012, 8(7).

Parker SC, Stitzel ML, Taylor DL, i sur. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A*, 2013, 110(44), 17921-17926.

Parr T, Turgutlu K, Csiszar C, Howard J. Beware Default Random Forest Importances, 2018, https://explained.ai/rf-importance/#6.2, pristupljeno 1.7.2020.

Pedregosa F, Varoquaux G, Gramfort A, i sur. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011, 12, 2825-2830.

Picheta D. Nim Programming Language, verzija 1.2.0, 2019, https://nim-lang.org/, pristupljeno 15.4.2020.

R Core Team. R: A Language and Environment for Statistical Computing, 2020, preuzeto s https://www.R-project.org/, pristupljeno 9.5.2020.

Rao SS, Huntley MH, Durand NC, i sur. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 2014, 159(7), 1665-1680.
Rao SSP, Huang SC, Glenn St Hilaire B, i sur. Cohesin Loss Eliminates All Loop Domains. *Cell*, 2017, 171(2), 305-320.

Robin X, Turck N, Hainard A, i sur. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 2011, 12, 77.

Roychoudhuri R, Hirahara K, Mousavi K, i sur. BACH2 represses effector programs to stabilize T(reg)-mediated immune homeostasis. *Nature*, 2013, 498(7455), 506-510.

Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, 2015, 10(3).

Schröder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, 2002, 110(4), 521-529.

Sengupta S, Siliciano RF. Targeting the Latent Reservoir for HIV-1. *Immunity*, 2018, 48(5), 872-895.

Sertznig H, Hillebrand F, Erkelenz S, Schaal H, Widera M. Behind the scenes of HIV-1 replication: Alternative splicing as the dependency factor on the quiet. *Virology*, 2018, 516, 176-188.

Singh, PK, i sur. LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes Dev.*, 2015, 29, 2287–2297.

Sinkhorn R, Knopp P. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math*., 1967, 21(2), 343-348.

Sowd GA, Serrao E, Wang H, i sur. A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc Natl Acad Sci U S A*., 2016, 113(8).

Spicuglia S, Vanhille L. Chromatin signatures of active enhancers. *Nucleus*, 2012, 3(2), 126-131.

Stehman SV. Selecting and interpreting measures of thematic classification accuracy. *Remote Sens. Environ.*, 1997, 62, 77-89.

Suzuki Y, Craigie R. The road to chromatin - nuclear entry of retroviruses. *Nat Rev Microbiol*, 2007, 5(3), 187-196.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 2015, 31(12), 2032-2034.

Toda T, Hsu JY, Linker SB, i sur. Nup153 Interacts with Sox2 to Enable Bimodal Gene Regulation and Maintenance of Neural Progenitor Cells. *Cell Stem Cell*, 2017, 21(5), 618-634.

Tsukumo S, Unno M, Muto A, i sur. Bach2 maintains T cells in a naive state by suppressing effector memory-related genes. *Proc Natl Acad Sci U S A*, 2013, 110(26), 10735-10740.

Turner AW, Margolis DM. Chromatin Regulation and the Histone Code in HIV Latency. *Yale J Biol Med*, 2017, 90(2), 229-243.

Vogt, PK. The Place of Retroviruses in Biology. U: Retroviruses. Coffin, JM, Hughes, SH, Varmus, HE, urednici, New York, Cold Spring Harbor Laboratory Press, 1997.

Vranckx LS, Demeulemeester J, Saleh S, i sur. LEDGIN-mediated Inhibition of Integrase-LEDGF/p75 Interaction Reduces Reactivation of Residual Latent HIV. *EBioMedicine*, 2016, 8, 248-264.

Wagner TA, McLaughlin S, Garg K, i sur. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science*, 2014, 345(6196), 570-573.

Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res*, 2007, 17(8), 1186-1194.

Wei T, Simko V. R package "corrplot": Visualization of a Correlation Matrix (Version 0.84), 2017, preuzeto s https://github.com/taiyun/corrplot, pristupljeno 9.5.2020.

Wetschoreck F, Krabel T, Krishnamurthy S. ppscore - a Python implementation of the Predictive Power Score (PPS), 2020, https://github.com/8080labs/ppscore, pristupljeno 23.4.2020.

Whyte WA, Orlando DA, Hnisz D, i sur. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 2013, 153(2), 307-319.

Witte S, O'Shea JJ, Vahedi G. Super-enhancers: asset management in immune cell genomes. *Trends Immunol*, 2015, 36, 519–526.

Wright M, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 2017, 77(1), 1-17.

Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, 2011, 43(11), 1059-1065.

Jiang Y, Qian F, Bai X, i sur. SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res*, 2019, 47(D1), D235-D243.

Zack JA, Kim SG, Vatakis DN. HIV restriction in quiescent $CD4^+$ T cells. *Retrovirology*, 2013, 10, 37.

Zhang Y, Liu T, Meyer CA, i sur. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 2008, 9(9), R137.

8. Sažetak/Summary

Virus humane imunodeficijencije 1 (HIV-1) je uzročnik stečenog sindroma imunodeficijencije kod ljudi. Unatoč uspjesima antiretroviralne terapije, perzistencija virusa uslijed uspostave transkripcijske latencije i dalje predstavlja barijeru prema izlječenju. Integracija HIV-1 u ljudski genom je složena međuigra virusa, kromatina stanice domaćina i nuklearne organizacije na koju utječe mnogo čimbenika. HIV-1 integrira aktivno prepisujuće gene bogate aktivnim histonskim oznakama i prostorno povezane odjeljke genoma s više super-pojačivača. Metode strojnog učenja su dobar izbor za analizu takvih problema jer omogućuju istraživanje velikog broja čimbenika, mogu pronaći složene veze i uzorke u velikim biološkim skupovima podataka i pružiti uvid u relativnu značajnost čimbenika. Model razvijen u ovom radu postiže dobar učinak diskriminacije gena koje HIV-1 rekurentno

integrira, potvrđuje dosadašnje spoznaje i kvantificira koliko prostorni podaci doprinose mogućnosti predviđanja integracija u odnosu na gensku ekspresiju, epigenetske modifikacije i stanje kromatina. Analizom značajnosti varijabli permutacijskim testom, aktivne histonske oznake (H3K27ac i H3K36me3 unutar gena i H3K4me3 oko mjesta početka transkripcije) i razina genske ekspresije su određene kao glavni faktori diskriminacije takvih gena u odnosu na gene bez integracija. Analizom površina ispod ROC krivulja, ustanovljeno je da prostorna podjela genoma u odjeljke i lokalizacija super-pojačivača unutar istih, neovisno o histonskim oznakama, značajno doprinose diskriminaciji ($p < 0.05$). Također je primijećeno da isključenje podataka o genskoj ekspresiji iz modela ne uzrokuje gubitak informacija kakav bi se očekivao na temelju relativne značajnosti određene permutacijskim testom. Računalne metode i prediktivno modeliranje dovode do konzistentnih zaključaka i koristan su alat za proučavanje retroviralne integracije. Slične metode bi se mogle primijeniti u različitim fazama infekcije kako bi se potencijalno otkrile razlike u mehanizmu odabira integracijskih mjesta i objasnila perzistencija HIV-1 i nastanak latentnog spremnika, što bi moglo biti primjenjivo u razvoju antiretroviralne terapije. Također, ovdje razvijene i opisane računalne analitičke metode mogu biti korisne za daljnje istraživanje mnogih aspekata retroviralne DNA integracije i drugih bioloških problema.

Human immunodeficiency virus (HIV-1) is responsible for acquired immune deficiency syndrome in humans. Despite the advances of antiretroviral therapy, persistence of the virus due to establishment of transcriptional latency still remains a barrier for the cure. HIV-1 integration presents as a complex interplay between the virus, cell chromatin and nuclear organisation and is influenced by a myriad of factors. HIV-1 preferentially integrates actively transcribed genes rich in active histone marks and spatially connected genomic compartments with more super-enhancers. Machine learning methods are a good choice for analyses of such problems as they allow for an exploration of a multitude of factors and they can discern complex patterns in big biological datasets, as well as provide insight into relative importance of factors. Model developed in this paper achieves good discrimination of recurrently integrated genes, confirms previous findings and quantifies the importance of spatial data in predicting HIV-1 integrations compared to gene expression, histone modifications and chromatin state. Variable importance analysis, using a permutation test, determines higher gene expression levels and active histone marks (H3K27ac and H3K36me3 in gene bodies,

and H3K4me3 in transcription start site neighbourhood) as main factors in discrimination of recurrently integrated genes from genes without integrations. ROC curve analysis reveals that the spatial division of genome into compartments and localisation of super-enhancers inside them significantly contribute to discrimination ($p < 0.05$), independently of histone marks and gene expression. It is also found that removal of gene expression data from the model does not lead to information loss that would be expected from relative importances determined by the permutation test. Computational methods and predictive modeling lead to consistent conclusions and provide useful tools for retroviral integration study. Similar methods could be used in order to infer differences in mechanisms of integration site selection in different infection phases and clarify the formation of a latent HIV-1 reservoir, which could subsequently be used to advance antiretroviral therapy development. Also, herein developed and described computational analysis methods can be useful for further research of many aspects of retroviral DNA integrations, as well as other biological problems.

Temeljna dokumentacijska kartica

Sveučilište u Zagrebu
Farmaceutsko-biokemijski fakultet
Studij: Farmacija
Zavod za biokemiju i molekularnu biologiju
A. Kovačića 1, 10000 Zagreb, Hrvatska

Diplomski rad

Prediktivno modeliranje retroviralnih integracija virusa HIV-1 u aktivirane CD4+ T stanice

Moreno Martinović

SAŽETAK

Virus humane imunodeficijencije 1 (HIV-1) je uzročnik stečenog sindroma imunodeficijencije kod ljudi. Unatoč uspjesima antiretroviralne terapije, perzistencija virusa uslijed uspostave transkripcijske latencije i dalje predstavlja barijeru prema izlječenju. Integracija HIV-1 u ljudski genom je složena međuigra virusa, kromatina stanice domaćina i nuklearne organizacije na koju utječe mnogo čimbenika. HIV-1 integrira aktivno prepisujuće gene bogate aktivnim histonskim oznakama i prostorno povezane odjeljke genoma s više super-pojačivača. Metode strojnog učenja su dobar izbor za analizu takvih problema jer omogućuju istraživanje velikog broja čimbenika, mogu pronaći složene veze i uzorke u velikim biološkim skupovima podataka i pružiti uvid u relativnu značajnost čimbenika. Model razvijen u ovom radu postiže dobar učinak diskriminacije gena koje HIV-1 rekurentno integrira, potvrđuje dosadašnje spoznaje i kvantificira koliko prostorni podaci doprinose mogućnosti predviđanja integracija u odnosu na gensku ekspresiju, epigenetske modifikacije i stanje kromatina. Analizom značajnosti varijabli permutacijskim testom, aktivne histonske oznake (H3K27ac i H3K36me3 unutar gena i H3K4me3 oko mjesta početka transkripcije) i razina genske ekspresije su određene kao glavni faktori diskriminacije takvih gena u odnosu na gene bez integracija. Analizom površina ispod ROC krivulja, ustanovljeno je da prostorna podjela genoma u odjeljke i lokalizacija super-pojačivača unutar istih, neovisno o histonskim oznakama, značajno doprinose diskriminaciji ($p < 0.05$). Također je primijećeno da isključenje podataka o genskoj ekspresiji iz modela ne uzrokuje gubitak informacija kakav bi se očekivao na temelju relativne značajnosti određene permutacijskim testom. Računalne metode i prediktivno modeliranje dovode do konzistentnih zaključaka i koristan su alat za proučavanje retroviralne integracije. Slične metode bi se mogle primijeniti u različitim fazama infekcije kako bi se potencijalno otkrile razlike u mehanizmu odabira integracijskih mjesta i objasnila perzistencija HIV-1 i nastanak latentnog spremnika, što bi moglo biti primjenjivo u razvoju antiretroviralne terapije. Također, ovdje razvijene i opisane računalne analitičke metode mogu biti korisne za daljnje istraživanje mnogih aspekata retroviralne DNA integracije i drugih bioloških problema.

Rad je pohranjen u Središnjoj knjižnici Sveučilišta u Zagrebu Farmaceutsko-biokemijskog fakulteta.

Rad sadrži: 49 stranica, 21 grafičkih prikaza, 4 tablice i 96 literaturnih navoda. Izvornik je na hrvatskom jeziku.

Ključne riječi: retroviralna integracija, hiv-1 spremnik, latentna infekcija, genomski odjeljci, super-pojačivači, histonske modifikacije, slučajne šume

Mentor: **Dr. sc. Gordan Lauc**, *redoviti profesor Sveučilišta u Zagrebu Farmaceutsko-biokemijskog fakulteta.*

Ocjenjivači: **Dr. sc. Gordan Lauc**, *redoviti profesor Sveučilišta u Zagrebu Farmaceutsko-biokemijskog fakulteta.*
Dr. sc. Kristian Vlahoviček, *redoviti profesor Sveučilišta u Zagrebu Prirodoslovno-matematičkog fakulteta.*
Dr. sc. Gordana Maravić Vlahoviček, *izvanredni profesor Sveučilišta u Zagrebu Farmaceutsko-biokemijskog fakulteta.*

Rad prihvaćen: rujan 2020.

Basic documentation card

University of Zagreb
Faculty of Pharmacy and Biochemistry
Study: Pharmacy
Department of Biochemistry and Molecular Biology
A. Kovačića 1, 10000 Zagreb, Croatia

Diploma thesis

Predictive modelling of retroviral HIV-1 integration in activated CD4+ T cells

Moreno Martinović

SUMMARY

Human immunodeficiency virus (HIV-1) is responsible for acquired immune deficiency syndrome in humans. Despite the advances of antiretroviral therapy, persistence of the virus due to establishment of transcriptional latency still remains a barrier for the cure. HIV-1 integration presents as a complex interplay between the virus, cell chromatin and nuclear organisation and is influenced by a myriad of factors. HIV-1 preferentially integrates actively transcribed genes rich in active histone marks and spatially connected genomic compartments with more super-enhancers. Machine learning methods are a good choice for analyses of such problems as they allow for an exploration of a multitude of factors and they can discern complex patterns in big biological datasets, as well as provide insight into relative importance of factors. Model developed in this paper achieves good discrimination of recurrently integrated genes, confirms previous findings and quantifies the importance of spatial data in predicting HIV-1 integrations compared to gene expression, histone modifications and chromatin state. Variable importance analysis, using a permutation test, determines higher gene expression levels and active histone marks (H3K27ac and H3K36me3 in gene bodies, and H3K4me3 in transcription start site neighbourhood) as main factors in discrimination of recurrently integrated genes from genes without integrations. ROC curve analysis reveals that the spatial division of genome into compartments and localisation of super-enhancers inside them significantly contribute to discrimination ($p < 0.05$), independently of histone marks and gene expression. It is also found that removal of gene expression data from the model does not lead to information loss that would be expected from relative importances determined by the permutation test. Computational methods and predictive modeling lead to consistent conclusions and provide useful tools for retroviral integration study. Similar methods could be used in order to infer differences in mechanisms of integration site selection in different infection phases and clarify the formation of a latent HIV-1 reservoir, which could subsequently be used to advance antiretroviral therapy development. Also, herein developed and described computational analysis methods can be useful for further research of many aspects of retroviral DNA integrations, as well as other biological problems.

The thesis is deposited in the Central Library of the University of Zagreb Faculty of Pharmacy and Biochemistry.

Thesis includes: 49 pages, 21 figures, 4 tables and 96 references. Original is in Croatian language.

Keywords: retroviral integration, HIV-1 reservoir, latent infection, genomic compartments, super-enhancers, histone modifications, random forests

Mentor: **Gordan Lauc**, **Ph.D.** *Full Professor,* University of Zagreb Faculty of Pharmacy and Biochemistry

Reviewers: **Gordan Lauc, Ph.D.** Full Professor**,** University of Zagreb Faculty of Pharmacy and Biochemistry
Kristian Vlahoviček, **Ph.D.** *Full Professor,* University of Zagreb Faculty of Science
Gordana Maravić Vlahoviček, **Ph.D.** *Associate Professor,* University of Zagreb Faculty of Pharmacy and Biochemistry

The thesis was accepted: September 2020