Primena veštačke inteligencije u analizi genetičkih podataka dobijenih sekvenciranjem nove generacije

Simpozijum Veštačka inteligencija i medicina

Dr Biljana Stanković, viši naučni saradnik

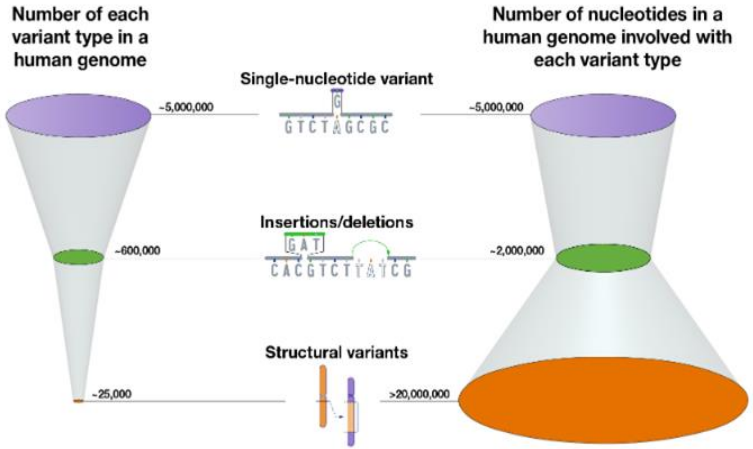
Dr Sonja Pavlović, naučni savetnik

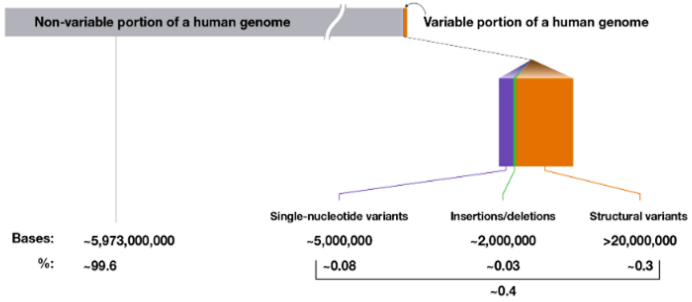
Institut za molekularnu genetiku i genetičko inženjerstvo, Univerzitet u Beogradu

Humana molekularna genetika i genomika, Grupa za molekularnu biomedicinu

Genomika - izučava ceo genetički materijal jednog organizma

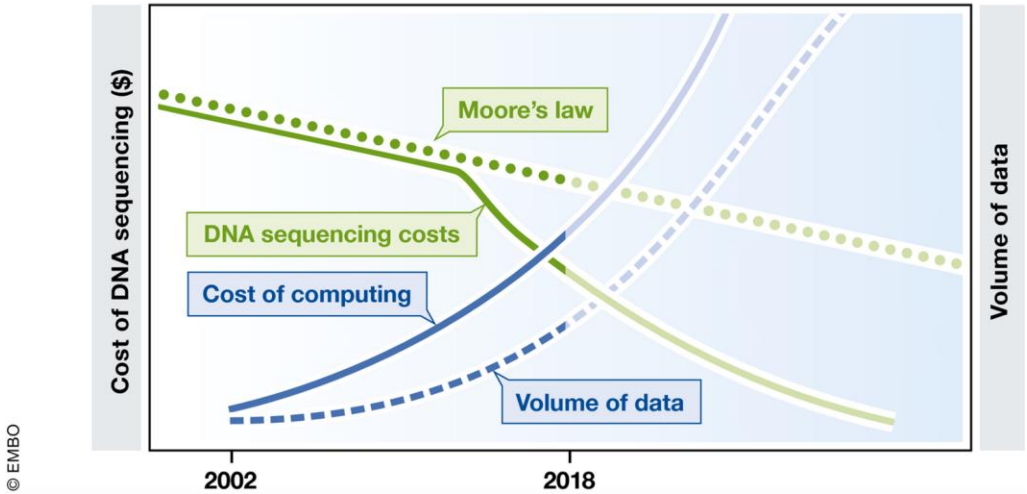
- Razumevanje funkcije genoma i na koji način je genetička varijabilnost povezana sa fenotipom
- Genetička varijabilnost predstavlja razlike u DNK sekvenci između individua/populacija jedne vrste
- U proseku, genom dve osobe je **~99.6% identičan** a **~0.4% čine razlike** (~27 miliona varijabilnih nukleotida!)
- Genetičke varijante: SNV (varijante na nivou pojedinačnih nukleotida), indel (umetnute ili nedostajuće nukleotidne sekvence dužine do 50bp), CNV (promena broja kopija ponavljanja)





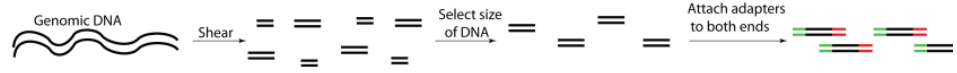
Sekvenciranje nove generacije (NGS)

- Detekcija genetičke varijabilnosti omogućena je sekvenciranjem genoma
- Razvoj modernih visokopropusnih tehnologija (NGS) - masivno paralelno sekvenciranje kako bi se odredila sekvenca nukleotida u DNK ili RNK
- Brzo i jeftino sekvenciranje humanih genoma (sati/dani, cena < 1,000$) u poređenju sa starim metodama (Projekat Humani Genom trajao 13 godina i koštao 3,000,000,000$)

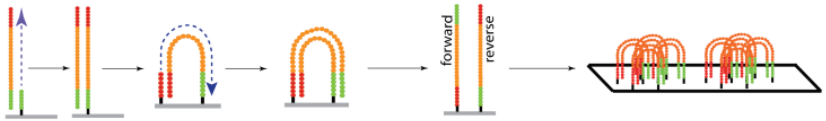


Todd J Treangen and Mihai Pop. EMBO Rep (2018)19: e47036 https://doi.org/10.15252/embr.201847036

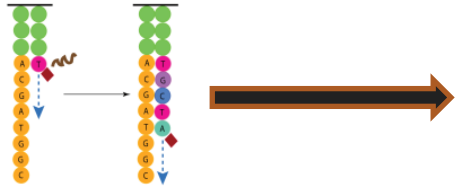
SEKVENCIRANJE NOVE GENERACIJE KRATKIH OČITAVANJA



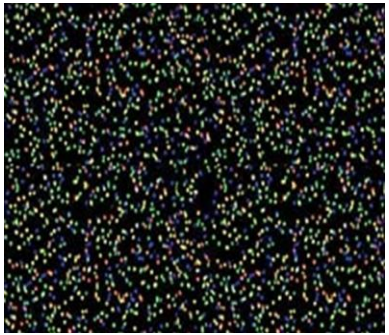
(a) Library preparation



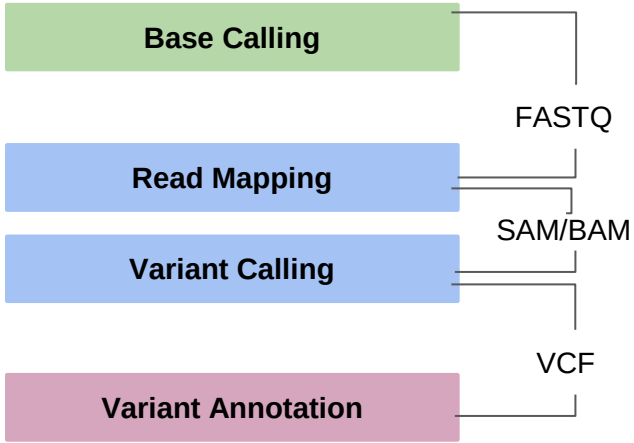
(b) Cluster generation

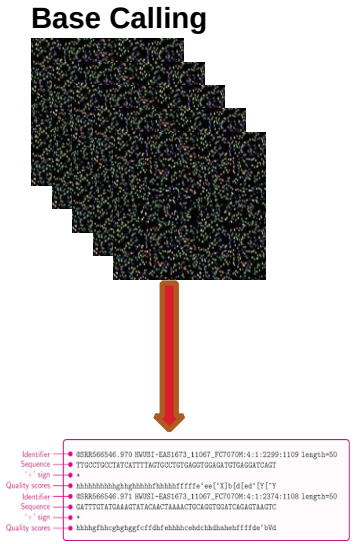


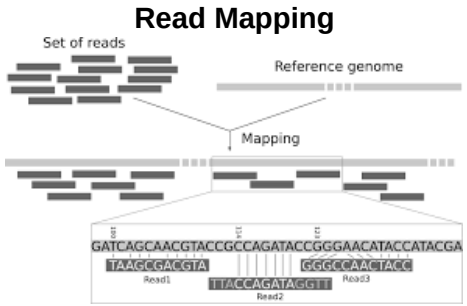
(c) Sequencing by synthesis

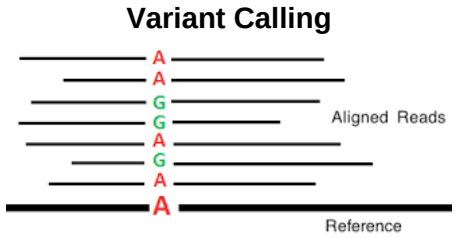


BIOINFORMATIČKA OBRADA PODATAKA









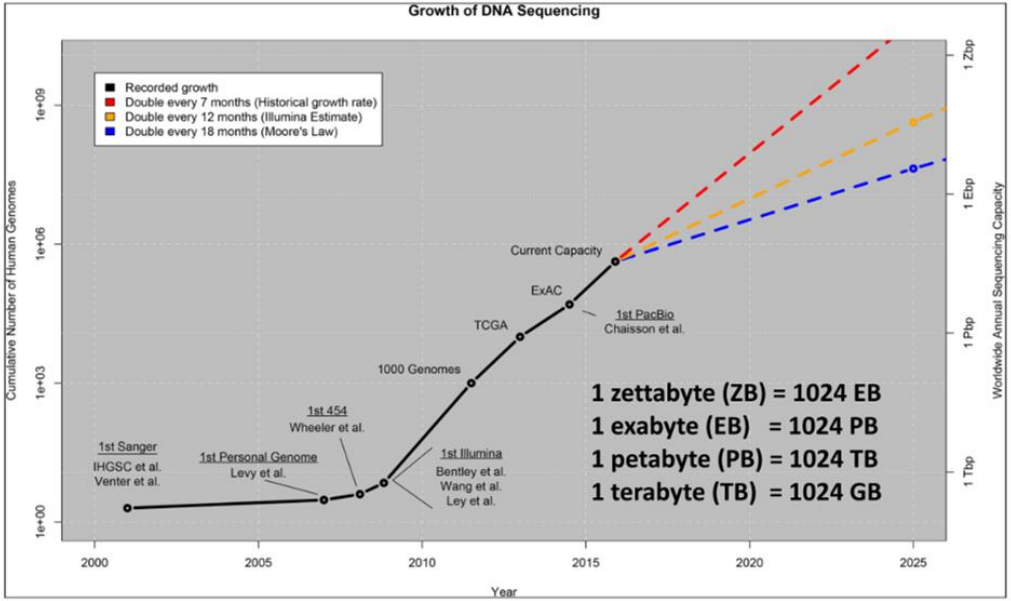
ANALIZA I INTERPRETACIJA PODATAKA

Klinička interpretacija

Naučni doprinos: genomska baza podataka

- Znanje o genomskoj varijabilnosti se sve više koristi u medicini u cilju poboljšanja dijagnostike, prognoze, individualizacije terapije (personalizovana/genomska medicina)
- Nacionalne i internacionalne inicijative koje imaju za cilj da proizvedu i omoguće siguran pristup genomskim podacima kako bi unapredili personalizovanu medicinu (National Institutes of Health (NIH) ‘All of Us’, European 1+ Million Genomes Initiative, Genome of Europe, Emirati Genome Programme…)

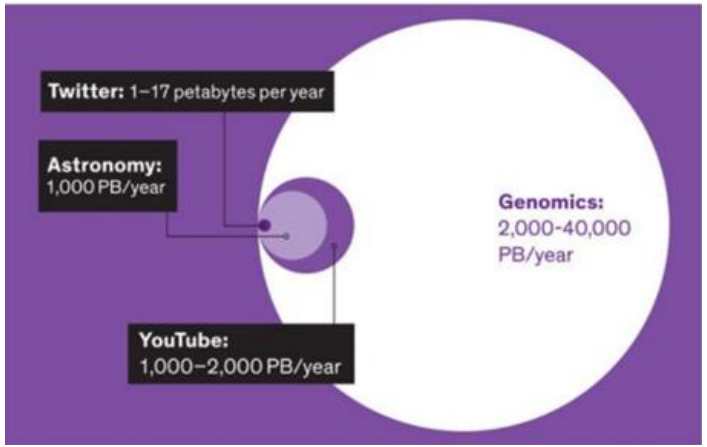
Genomics Data is Big Data

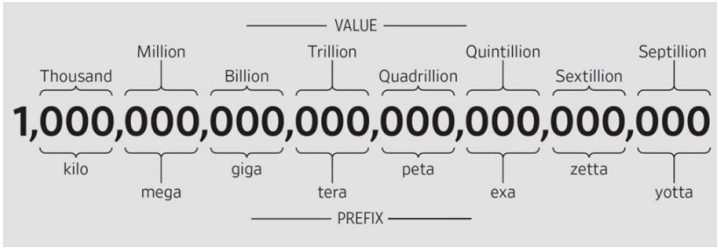


Stephens, et al. *PLOS Biology*, 2015.

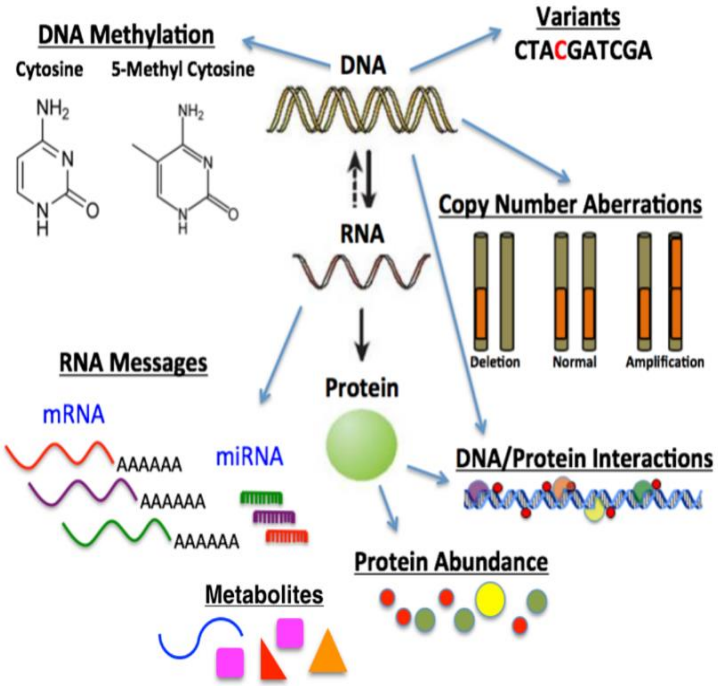
10.1371/journal.pbio.1002195

~~Astronomical~~ 'Genomical' Data:





BIOLOŠKA VARIJABILNOST

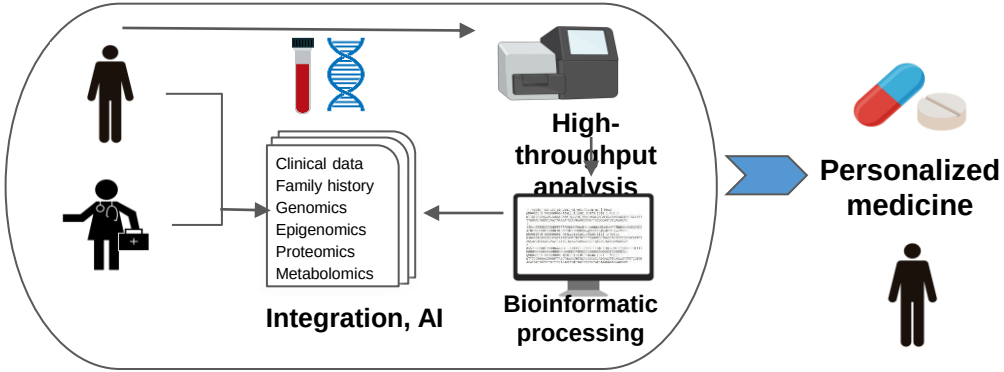


KOMPLEKSNA PITANJA (NASTANAK BOLESTI, BIOMARKERI, PROGNOZA BOLESTI?)

+

MULTI-DIMENZIONALNI HETEROGENI VELIKI PODACI

= MAŠINSKO UČENJE !

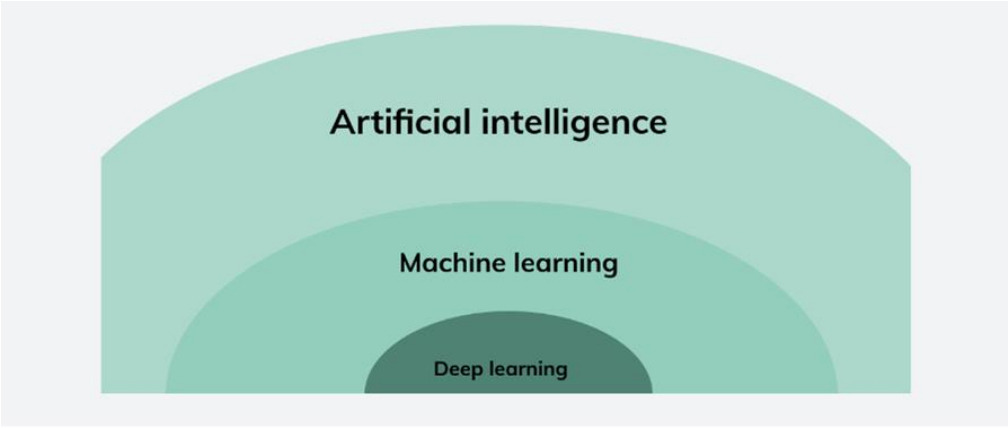


VI/MU/DU

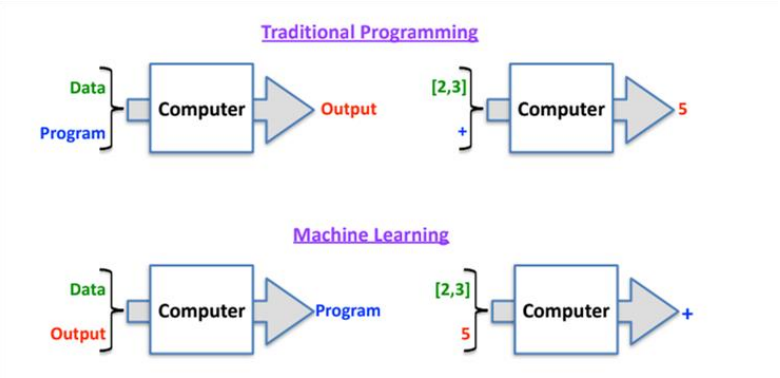
Veštačka inteligencija (VI) - sposobnost kompjuterskog sistema da oponaša humane kognitivne funkcije kao sto su učenje i rešavanje problema

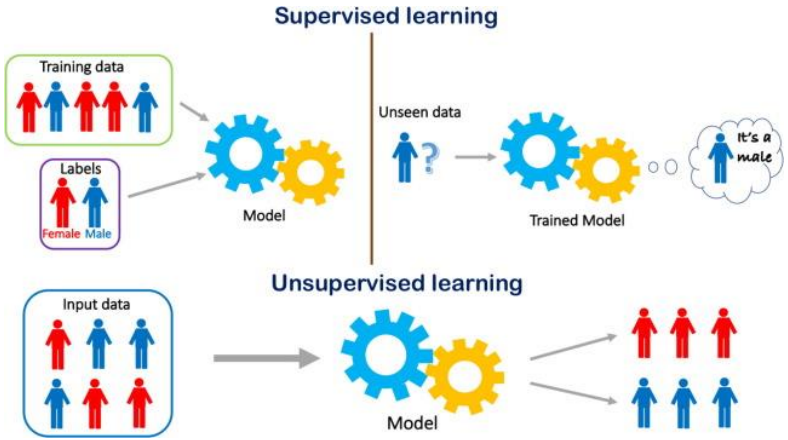
Mašinsko učenje (MU) - podvrsta VI koja koristi matematičke modele kako bi iz podataka “učio” bez direktnih instrukcija kako da uči već algoritam sam otkriva svoja pravila učenja koja može da poboljša kroz iskustvo. MU analizira obrasce u podacima i odnose između njih. MU ne zahteva prethodno znanje, može da uči kroz nove podatke (fleksibilnost)

Duboko učenje (DU) - podvrsta MU koja se bazira na korišćenju dubokih neuralnih mreža - pristup koji se sastoji iz manjih kompjutacionih jedinica/čvorova koje su među sobom povezane na način koji podseća na veze između neurona u mozgu



Proces MU





Nadgledano učenje	**Nenadgledano učenje**
Ulazni podaci su obeleženi	Ulazni podaci su neobeleženi
Postoji faza treninga	Ne postoji faza treninga
Podaci se modeluju na osnovu trening seta	Koristi osobine datih podataka za klasifikaciju
Najčešće podeljeno u 2 tipa: klasifikacija, regresija	Najpopularnije: Klasterovanje i redukcija dimenzionalnosti
Poznat broj klasa	Nepoznat broj klasa

Primena MU u genomici

1. **Detekcija varijanti**

    - Pozivanje baze/nukleotida nakon NGS - predikcija pozivanja ispravne nukleotidne sekvence iz sirovih podataka
    - Pozivanje varijanti - bioinformatička analiza tokom koje se utvrđuje koje pozicije u genomu se razlikuju u poređenju sa referentnom sekvencom (Google DeepVariant, Octopus); detekcija somatskih varijanti, detekcija CNV varijanti

1. **Interpretacija varijanti**

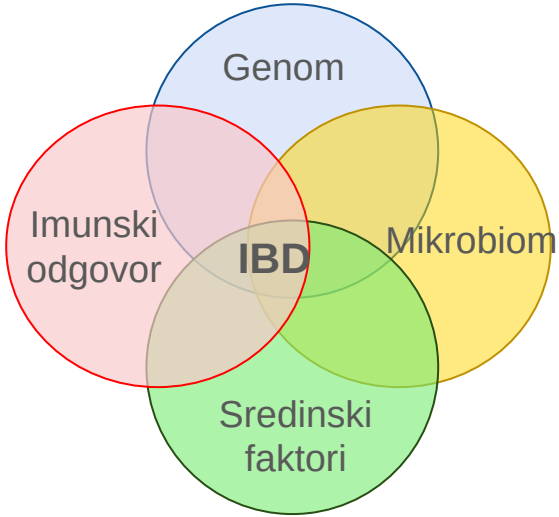
    - Predikcija efekta genetičke varijante na strukturu i funkciju proteina (Polyphen, Mutation Taster, CADD, AlphaMissense)
    - Razumevanje genetičke varijacije u nastanku bolesti (identifikacija genotip-fenotip korelacija, identifikacija biomarkera, polygenic risk scores)

1. **Anotacija genomskih i proteinskih sekvenci**

    - Klasifikacija genomskih sekvenci (promotori, mesta splajovanja, enhenseri, mesta vezivanja transkripcioniih faktora, itd) - važno za razumevanje funkcionalnih, strukturalnih i regulatornih mehanizama u genomu
    - Predikcija savijanja proteina bazirana na nukleotidnim sekvencama (AlphaFold)

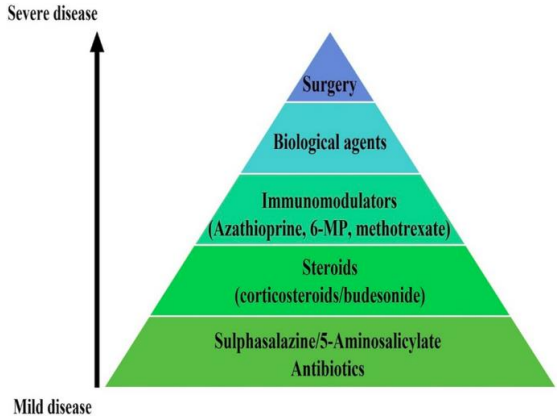
Inflamatorne bolesti creva

- Inflamatorne bolesti creva (*Inflammatory bowel disease,* IBD) - nastaju kao neadekvatan imunološki odgovor na prisustvo crevnih bakterija kod genetički podložnih osoba
- **6,8 miliona ljudi širom sveta (1,3 miliona u Evropi) boluje od IBD**, ¼ se dijagnostikuje kod dece
- Neizlečiva, tretman različitim lekovima kako bi se simptomi držali pod kontrolom
- Visoka heritablinost
- Bolest je heterogena, može imati različite kliničke manifestacije koje se najčešće odnose na lokaciju i stepen zahvaćenosti crevnog epitela (Kronova bolest, ulcerozni kolitis).

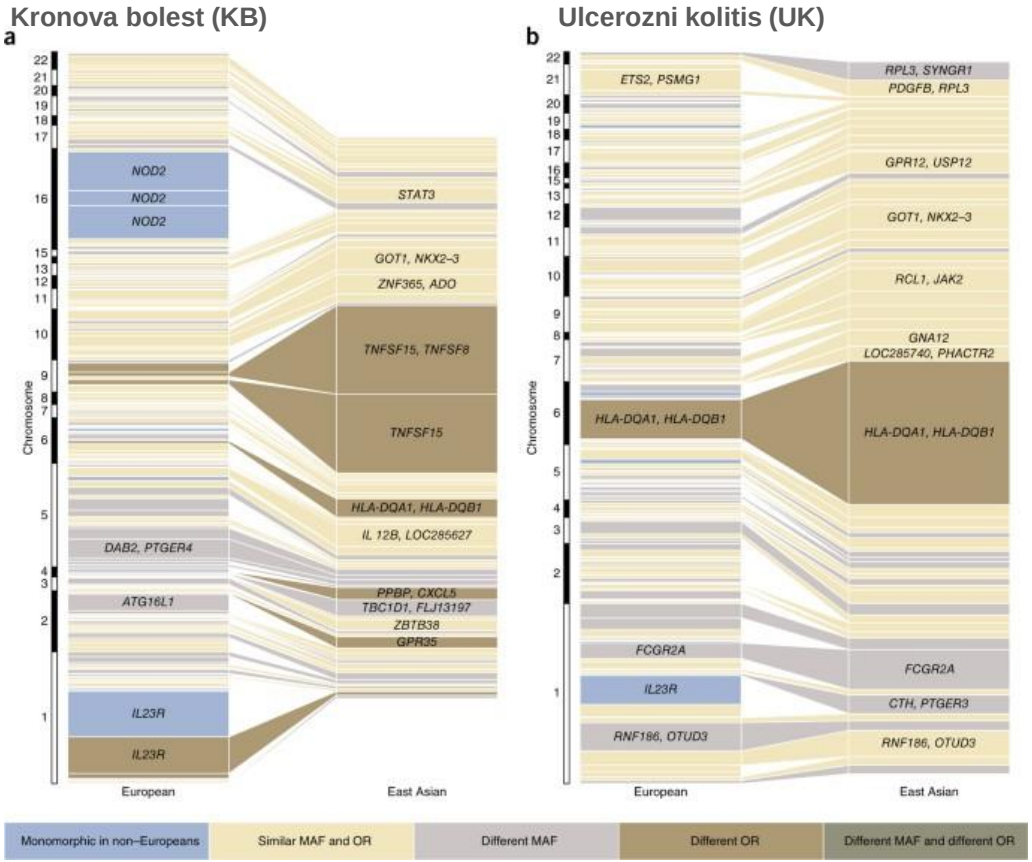


IBD spektar

Gastroenterology 2022; 10.1053/j.gastro.2021.12.246



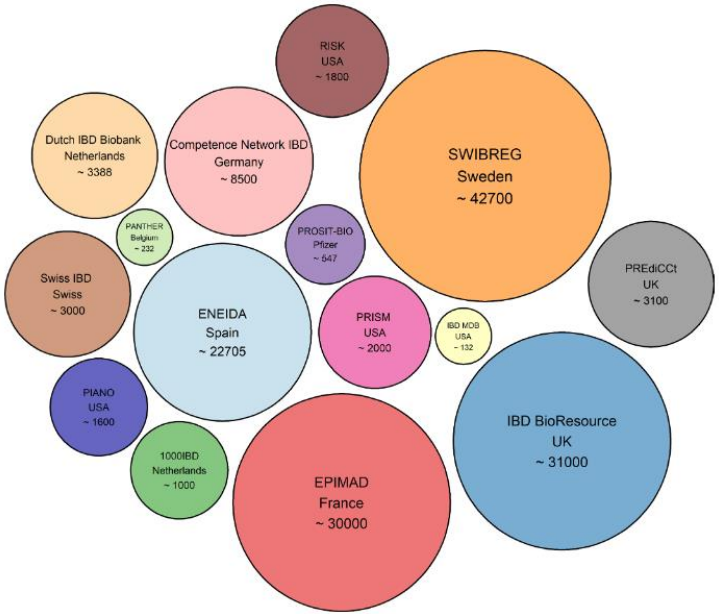
- Do sada > 200 gena asocirano sa IBD (uglavnom povezani sa inflamacijom, odgovorom imunskog sistema, autofagijom, funkcijom intestinalne epitelijalne barijere)



Nature Genetics 2015; 10.1038/ng.3359

- *linkage-based studies*
- *candidate gene association studies*
- *high coverage technologies [DNA arrays and next-generation sequencing (NGS)]*

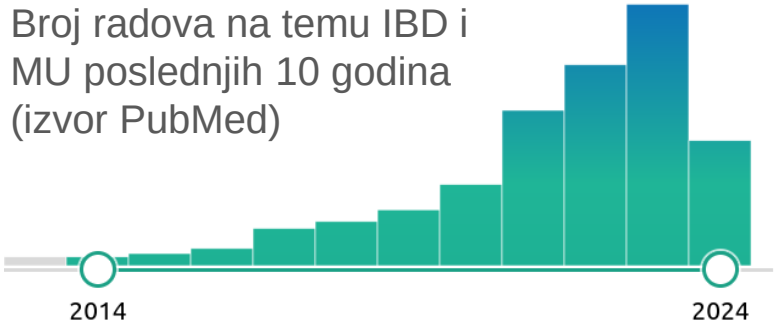
- Koji pacijenti su u većem riziku da razviju bolest?
- Koji su rani biomarkeri bolesti?
- Koji pacijenti su u većem riziku da imaju komplikovanije forme bolesti?
- Od koje terapije će pacijent imati najviše koristi/minimum neželjenih efekata?



Gut 2020. 10.1136/gutjnl-2019-320065

VELIKI OBIM IBD PODATAKA
+
MU
= ODGOVORI ?

Broj radova na temu IBD i MU poslednjih 10 godina (izvor PubMed)



First Author and Year [ref]	Machine Learning Algorithm	Predictors/Prediction	Performance	Tested on Independent Cohort	Subjects
Chen 2017 [65]	Bayesian mixture approach	GWAS or Immunochip SNPs data/ IBD risk score	CD AUC: 0.75, UC AUC: 0.70	yes	The IIBDGC) cohort—over 68,000 IBD patients and 29,000 healthy controls (4:5 ratio for training and testing, respectively)
Wei 2013 [66]	L1 penalized logistic regression, SVM, gradient boosted trees	Immunochip SNPs data/CD and UC distinction from healthy controls	CD AUC 0.86, UC AUC 0.83	yes	The IIBDGC cohort—~17,000 CD, ~13,000 UC, and ~22,000 controls (randomly divided into 3 folds of equal size for preselection, training and testing, respectively)
Romagnoni 2019 [37]	Logistic regression, gradient boosted trees, neural network and ensemble method	Immunochip SNPs data/probability of CD	AUC 0.8	yes	The IIBDGC cohort—train dataset (34,634 samples), test dataset (17,317 samples)
Pal 2017 [51]	Naïve Bayes	Exome data/CD status	AUC 0.81	yes	Training set: 64 CD and 47 controls (CAGI4); Testing set: 51 CD and 15 controls (CAGI3)
Raimondi 2020 [63]	Neural network	Whole exomes/to distinguish between CD and healthy controls	AUC 0.74–0.83 AUPRC 0.81–0.93	yes	CAGI2, CAGI3, CAGI4 datasets (training and testing)
Wang 2019 [64]	SVM	Whole exomes/to distinguish between CD and healthy controls	AUC 0.7–0.75 AUPRC 0.73–0.80	yes	CAGI4 (training set), CAGI3 (testing set)

Stankovic et al. Genes 2021. Machine Learning Modeling from Omics Data as Prospective Tool for Improvement of Inflammatory Bowel Disease Diagnosis and Clinical Classifications. Doi: 10.3390/genes12091438

- Najčešće korišćene metode u IBD studijama: penalized regression models, random forest, support vector machines, Bayesian approach and neural networks
- AUC u opsegu 0.7 to 0.95

- Klasifikacija IBD i zdrave kontrole iz Srbije bazirana na upotrebi MU koje koristi genetičke podatke kao ulazna obeležja
- 167 IBD pacijenata, 101 zdravih ispitanika
- 10 genetičkih varijanti u genima *NOD2*, *TLR4*, *TNF-α*, *IL-6*, *IL-1β* i *IL-1RN*
- Najbolje rezultate su dali modeli koji su uključivali sve analizirane genetičke varijante
- Limitacija: Relativno mali broj ispitanika
- Nepostojanje nezavisne kohorte za testiranje modela, korišćena 10-struka kros-validacija

Model name	CD dataset				UC dataset			
	Sensitivity	Specificity	AUC	AUC CI	Sensitivity	Specificity	AUC	AUC CI
Elastic net	0.236	0.950	0.746	(0.674, 0.819)	0.453	0.733	0.640	(0.566, 0.715)
Neural net	0.500	0.950	0.832	(0.773, 0.891)	0.579	0.762	0.727	(0.660, 0.794)
Random forest	0.583	0.901	0.870	(0.818, 0.923)	0.789	0.619	0.770	(0.704, 0.835)
Linear SVM	0.417	0.881	0.733	(0.655, 0.810)	0.705	0.515	0.641	(0.564, 0.718)
Rbf SVM	0.542	0.911	0.864	(0.810, 0.917)	0.695	0.653	0.742	(0.673, 0.811)

Stankovic et al. Journal of Digestive Diseases 2015. Variations in inflammatory genes as molecular markers for prediction of inflammatory bowel disease occurrence

Prospective, Observational Study of Pharmacogenomics Markers of bio-naïve patients treated with anti-TNFs or vedolizumab with Crohn Disease: comprehensive genomics and machine learning approach in Serbia (PI: dr Sonja Pavlovic)

- Saradnja sa farmaceutskom kompanijom Takeda
- Sekvenciranje celokupnog genoma (WGS) pacijenata sa Kronovom bolesšću
- Ciljevi studije: **identifikacija farmakogenomskih markera** odgovora na biološku terapiju, **predikcija lošeg odgovora** na terapiju korišćenjem MU na WGS podacima
- Dugoročni ciljevi: individualizacija biološke terapije i stratifikacija pacijenata prilikom uvođenja terapija
- Strategija za selekciju ulaznih varijabli: 1. Preselekcija genetičkih varijanti koje su prepoznate u literaturi; 2. GWAS analiza koja će indentifikovati značajne genetičke signale; 3. Penalizirajući model koji će sam birati informativna obeležja

Izazovi i ograničenja

- **Kvalitet anotacija** podataka za treniranje. Korišćenje podataka koji nisu dobro opisani za trening modela može dovesti do niskih performansi i male sposobnosti modela za generalizaciju
- **Genomski podaci su visoko-dimenzionalni** i kao takvi predstavljaju izazov za algoritme MU. Odabir obeležja i redukcija dimenzionalnosti su tehnike koje se često koriste da bi se ovaj problem rešio; odabir informativnih obeležja iz velikih genomskih podataka nije jednostavan i često uključuje više strategija.
- **Interpretacija i transparentnost modela MU.** Naročito kod DU, teško je razumeti mehanizme koji su u osnovi njihovih predikcija. U genomici je interpretacija važna za identifikaciju biomarkera i razumevanje bioloških procesa
- **Genomski podaci često nemaju balansirane klase**, broj instanci u različitim klasama (npr u grupama bolesnih, zdravih) se značajno razlikuje. Nebalansirani podaci mogu voditi ka pristrasnim modelima koji favorizuju većinsku klasu, rezultujući u slabim performansama kada su u pitanju manjinske klase.
- **Genomski podaci su osetljivi na serijski efekat (batch effect), heterogenost uzoraka, konfundirajuče varijable**. Prisustvo oviih “ometajućih“ faktora u podacima za treniranje MU može voditi ka pristrasnim modelima i pogrešnim predikcijama
- **Pristup HPC serverima i ekspertiza** na polju biologije, bioinformatike i MU

Hvala na pažnji!

Grupa za molekularnu biomedicinu

Branka Zukić
Sonja Pavlović
Nataša Tosić
Teodora Karan-Đurašević
Biljana Stanković
Nikola Kotur
Vladimir Gašić
Sanja Srzentić Dražilov
Irena Marjanovic
Ivana Grubiša
Bojan Ristivojević
Đorđe Pavlović
Marina Jelovac
Isidora Ćurić
Katarina Krstajić

INSTITUT ZA MOLEKULARNU GENETIKU
I GENETIČKO INŽENJERSTVO
Univerzitet u Beogradu

INSTITUTE FOR MOLECULAR GENETICS
AND GENETIC ENGINEERING
University of Belgrade

imgge.bg.ac.rs

@IMGGE

IMGGE

imggi

Institut za molekularnu genetiku i genetičko inženjerstvo

Institut za molekularnu genetiku i genetičko inženjerstvo