

UNIVERZITET U BEOGRADU  
МАТЕМАТИЧКИ ФАКУЛТЕТ



Milica D. Simić

KLASTEROVANJE PODATAKA DOBIJENIH  
TEHNIKAMA PROSTORNE  
TRANSKRIPTOMIKE

master rad

Beograd, 2023.

**Mentor:**

dr Jovana KOVAČEVIĆ, docent  
Univerzitet u Beogradu, Matematički fakultet

**Članovi komisije:**

dr Vladimir KOVAČEVIĆ, naučni saradnik  
Univerzitet u Beogradu, INN Vinča & BGI Research

dr Mladen NIKOLIĆ, vanredni profesor  
Univerzitet u Beogradu, Matematički fakultet

**Datum odrbrane:** \_\_\_\_\_



**Naslov master rada:** Klasterovanje podataka dobijenih tehnikama prostorne transkriptomike

**Rezime:** Tehnike prostorne transkriptomike predstavljaju inovativne laboratorijske procedure pomoću kojih na osnovu uzorka tkiva dobijamo informacije o pojedinačnim ćelijama koje se u datom tkivu nalaze, konkretno o njihovom položaju u prostoru i genima koji su u tim ćelijama ispoljeni. Dobijeni podaci su veoma korisni za dublje razumevanje odnosa između ćelija različitih morfologija i okoline u kojoj se ćelije nalaze. Sa druge strane, u svakom uzorku tkiva eksperimentalno se za svaku ćeliju može utvrditi njen tip. Tip ćelije je bitan jer ukazuje na njenu funkciju u organizmu. Poznavanje tipa ćelije može biti korisno u dijagnostici bolesti, jer neke bolesti utiču samo na određene tipove ćelija u organizmu. Stoga je važno identifikovati koji tip ćelije je zahvaćen kako bi se primenila odgovarajuća terapija i lekovi koji će ciljati baš te ćelije. Cilj rada je da se ispita zavisnost između prostornih koordinata ćelija i gena koji su u tim ćelijama ispoljeni, kao i da se ispita da li se klasterovanjem podataka dobijenih tehnikama prostorne transkriptomike mogu dobiti klasteri koji odgovaraju tipovima ćelija i da se odredi uticaj genskih i koordinatnih komponenti na tip ćelije. Dobijeni rezultati jasno ukazuju na dominantan uticaj genskih komponenti na određivanje tipa ćelija. Međutim, važno je napomenuti da koordinatne komponente, iako ne dominiraju, igraju značajnu ulogu, što ih čini relevantnim faktorima za razmatranje u analizi tipova ćelija.

**Ključne reči:** ekspresija gena, tip ćelije, prostorna transkriptomika, *Stereo-seq* tehnika, *Slide-seq V2* tehnika, klasterovanje, *Leiden* algoritam

# Sadržaj

<b>1</b>	<b>Uvod</b>	<b>1</b>
1.1	Ekspresija gena . . . . .	1
1.2	Prostorna transkriptomika . . . . .	2
<b>2</b>	<b>Tehnike prostorne transkriptomike</b>	<b>6</b>
2.1	<i>Stereo-seq</i> tehnika . . . . .	6
2.2	<i>Slide-seqV2</i> tehnika . . . . .	9
<b>3</b>	<b>Podaci</b>	<b>12</b>
3.1	Podaci dobijeni <i>Stereo-seq</i> tehnikom . . . . .	12
3.2	Podaci dobijeni <i>Slide-seqV2</i> tehnikom . . . . .	18
<b>4</b>	<b>Problemi i metode</b>	<b>23</b>
4.1	Metoda 1 . . . . .	23
4.2	Metoda 2 . . . . .	25
<b>5</b>	<b>Rezultati</b>	<b>29</b>
5.1	Embrion miša . . . . .	29
5.2	Dorzalni srednji mozak miša . . . . .	37
5.3	Medula i korteks ljudskog bubrega i bubreg miša . . . . .	46
5.4	Diskusija . . . . .	53
<b>6</b>	<b>Zaključak</b>	<b>58</b>
	<b>Bibliografija</b>	<b>59</b>

# Glava 1

## Uvod

U ovom poglavlju ćemo detaljnije razmotriti koncept ekspresije gena i prostorne transkriptomike, i predstavićemo glavni cilj istraživanja, kao i rezultate koji su postignuti. Kod koji je kreiran za potrebe ovog istraživanja, zajedno sa svim dobijenim rezultatima, javno je dostupan na Git repozitorijumu [6].

### 1.1 Ekspresija gena

Dezoksiribonukleinska kiselina (DNK) je molekul koji se nalazi u jezgru svake ćelije i prenosi genetske informacije sa jedne generacije na drugu putem reprodukcije. Sastoje od dva komplementarna lanca nukleotida. Nukleotide se razlikuju po azotnoj bazi koja se nalazi u njihovoj strukturi. Postoje četiri azotne baze u DNK nukleotidima i one su adenin (A), timin (T), guanin (G) i citozin (C).

Baze se uparuju prema određenom obrascu, gde adenin formira vezu sa timinom (A-T), a guanin formira vezu sa citozinom (G-C). Ove baze spajaju dva lanca DNK zajedno, a redosled baza na jednom lancu DNK predstavlja genetski kod koji određuje strukturu i funkciju proteina i drugih molekula u organizmu (slika 1.1).



Slika 1.1: Molekul DNK

Proces nastanka proteina (sinteza proteina) se odvija u dve glavne faze: transkripcija i translacija. Ovaj proces uključuje prenos genetske informacije iz DNK u

## *GLAVA 1. UVOD*

---

RNK<sup>1</sup> molekule, koji zatim služe kao šablon za sintezu proteina:

- Transkripcija: Prepisivanje informacija iz molekula DNK u molekul iRNK (informaciona RNK). Odvija se duž molekula DNK, tako što enzim RNK polimeraza čita nukleotide i sintetiše komplementarni lanac iRNK. Kada polimeraza dođe do sekvene za zaustavljanje, proces transkripcije se završava, a molekul iRNK se odvaja od DNK.
- Procesiranje iRNK: Pre nego što napusti jezgro ćelije, iRNK se mora obraditi. U ovoj fazi se dešava dodavanje poli-A repa (sekvenca adenina) na iRNK.
- Translacija: U citoplazmi ćelije, iRNK dolazi do ribozoma, gde se informacije iz iRNK prevode u aminokiselinske sekvene proteina. Molekuli transportne RNK (tRNK) prepoznaju kodone (triplete nukleotida) na iRNK i donose odgovarajuće aminokiseline. Ribozom pomaže u povezivanju aminokiselina, stvarajući polipeptidni lanac koji će se presaviti u funkcionalni protein. Kada se dođe do zaustavnog kodona na iRNK, proces translacije se završava, a novosintetisani protein se oslobađa.

Svaka ćelija istog organizma sadrži isti DNK materijal, a kako su geni delovi molekula DNK, to znači da svaka ćelija sadrži isti skup gena. Ekspresija gena predstavlja kvantifikaciju nivoa transkripta (iRNK) ili proteina koji se proizvode iz tog gena. U nekim ćelijama određeni gen može biti aktivан (njegova ekspresija je različita od nule), dok u drugim ćelijama istog organizma taj gen može biti neaktiviran (njegova ekspresija je nula).

## **1.2 Prostorna transkriptomika**

Kako svaka ćelija istog organizma sadrži isti DNK materijal, odnosno isti skup gena, ćelijska raznolikost i funkcija, koja nastaje tokom embrionalnog razvoja i održava se tokom odraslog života, procenjuje se na nivou proteina, odnosno RNK materijala koji se proizvede iz tih gena.

Tehnika koja omogućava izdvajanje transkriptoma ćelije, odnosno molekula RNK koji su prisutni unutar ćelije u određenom trenutku, naziva se jednoćelijska transkriptomika (eng. *single cell RNA sequencing*). Ova tehnika podrazumeva da se ćelije

---

<sup>1</sup>Za razliku od DNK molekula, RNK je jednolančani molekul, i njegovi nukleotidi umesto timina (T) imaju uracil (U), koji se uparuje sa adeninom (A-U).

## *GLAVA 1. UVOD*

---

iz tkiva sačuvaju neoštećene bez stresa, smrti i/ili agregacije ćelija. Neizvodljivost ovog zahteva za mnoge tipove ćelija onemogućava nam da istražimo te tipove ćelija. Određene ćelije, kao što su na primer imunološke ćelije nisu uvek pričvršćene u tkivima i stoga se relativno lako izoluju iz krvi, limfoidnih organa, perifernih tkiva, itd. Nasuprot tome, mnoge druge vrste ćelija, na primer neuroni u mozgu, manje su podložne ovakvoj analizi jer zahtevaju specijalizovane protokole disocijacije tkiva kako bi se oporavile. Dodatno, prilikom izolacije ćelija iz tkiva gubi se prostorni kontekst ćelija u tom tkivu, koji bi inače obogatio analize identiteta i funkcija ćelija. Zbog ovih nedostataka procesa jednoćelijske transkriptomike postojala je potreba za sprovođenjem transkriptomike ćelija na nivou tkiva. Kako se sada pored transkriptoma ćelije određuje i položaj te ćelije u tkivu, ove tehnike se zovu tehnike prostorne transkriptomike [14].

Tehnike prostorne transkriptomike predstavljaju inovativne labaratorijske procedure pomoću kojih na osnovu uzorka tkiva dobijamo informacije o pojedinačnim ćelijama koje se u datom tkivu nalaze, konkretno o njihovom položaju u prostoru i genima koji su u tim ćelijama ispoljeni.

Prostorne komponente se odnose na raspored i lokaciju ćelija i gena unutar tki-va. Ovde su bitne jer omogućavaju razumevanje kako se ekspresija gena menja u zavisnosti od fizičkog položaja. To je posebno značajno u razvoju, gde određene ćelije igraju ključne uloge u formiranju tkiva i organa. Prostorna komponenta nam omogućava da identifikujemo koje ćelije su prisutne u određenom regionu i kako se njihova ekspresija gena razlikuje. Na primer, otkrivanje koje ćelije su blizu određenih ćelija može razjasniti kako se komunikacija i interakcije između ćelija odvijaju na nivou tkiva.

Genske komponente se odnose na same gene i njihovu ekspresiju. One su bitne jer nam omogućavaju da razumemo koje funkcije date ćelije imaju. Na primer, identifikacija gena koji su posebno aktivni u određenom regionu mozga može otkriti koje biološke procese taj region kontroliše.

Različiti tipovi ćelija u organizmu imaju karakteristične morfološke, strukturalne i funkcionalne osobine. Na primer, neuroni su specijalizovane ćelije nervnog sistema koje se izdvajaju po svojoj izduženoj morfološkoj strukturi, uključujući produžetke poput dendrita i aksona. Ove morfološke osobine omogućavaju neuronima da obavljaju složene funkcije u prenošenju, obradi i interpretaciji informacija u nervnom sistemu. Tako, dendriti služe za prijem signala od drugih neurona ili senzorskih ćelija, dok aksoni prenose te signale do drugih ćelija ili organa. Ovo omogućava

## *GLAVA 1. UVOD*

---

neuronima da igraju ključnu ulogu u prenosu nervnih impulsa, kontroli pokreta, senzacijama, učenju i pamćenju. Osim neurona, postoje i mnoge druge vrste ćelija u organizmu, kao što su epitelne ćelije, koje čine tkiva i organe kao što su koža i creva; mišićne ćelije, koje omogućavaju pokrete tela; imunološke ćelije, koje štite organizam od infekcija, itd. Svaki od ovih tipova ćelija ima svoje karakteristične osobine koje odražavaju njihovu specifičnu funkciju u organizmu.

Određivanje tipova ćelija može se postići eksperimentalno u laboratorijskim uslovima, koristeći tehnike kao što su mikroskopija, analiza genske ekspresije i morfološka ispitivanja. Pored toga, uz pomoć različitih atributa, uključujući genetske, morfološke i prostorne karakteristike, moguće je predvideti tipove ćelija i njihove funkcije. Ćelije koje obavljaju istu funkciju treba da ispolje iste gene, odnosno da imaju slične genske ekspresije. Međutim, moguće je u okviru istog tipa ćelije da se nađu i ćelije koje donekle imaju različite genske ekspresije. Do ovoga može doći usled raznih faktora - spoljašnji uticaji, faza razvoja, itd. Ipak ukoliko ćelije imaju sličnu gensku ekspresiju, velika je verovatnoća da pripadaju istom tipu. Ukoliko imaju različitu, onda nam informacija o genskim ekspresijama nije dovoljna da bi se doneo zaključak da li pripadaju istom tipu ili ne. Neophodno je uzeti u obzir i druge atrubute, pa se tako pored genske komponente, može posmatrati i prostorna komponenta ćelije. Pravilno identifikovanje i klasifikacija različitih tipova ćelija su vrlo bitni jer omogućavaju bolje razumevanje bioloških sistema i procesa, kao i razvoj terapija i lekova usmerenih ka specifičnim tipovima ćelija u svrhu lečenja bolesti.

Cilj ovog rada je da se ispita zavisnost između prostornih koordinata ćelija i gena koji su u tim ćelijama ispoljeni, kao i da se ispita da li se klasterovanjem podataka dobijenih tehnikama prostorne transkriptomike mogu dobiti klasteri koji odgovaraju tipovima ćelija i da se odredi uticaj genskih i koordinatnih komponenti na tip ćelije. Prvi deo je urađen tako što su poređeni grafovi  $G_1$  i  $G_2$  koji su formirani na osnovu koordinata ćelija i genskih ekspresija ćelija. U ovom delu nismo uspeli da dobijemo očekivane rezultate, što nam sugerije da je potrebno primeniti drugačije pristupe prilikom poređenja genskih i prostornih komponenti ćelija. Klasterovanje je uređeno pomoću *Leiden* algoritma nad unijom grafova  $G_1$  i  $G_2$  sa kombinacijom različitih brojeva najbližih suseda za oba grafa, gde maksimalan broj najbližih suseda nije veći od 30. Za svaku kombinaciju je određen *ARI* skor. Iako najveći *ARI* skor nije bio zadovoljavajući, optimalno klasterovanje (ono kod kojeg je *ARI* skor najveći) je iskorišćeno za određivanje optimalanog broj najbližih suseda za  $G_1$  i  $G_2$ , koji je ukazano da je uticaj genske komponenti na tip ćelije dosta veći nego uticaj koordinatne.

## *GLAVA 1. UVOD*

---

Međutim, važno je napomenuti da uticaj koordinatne komponente nije zanemarljiv, što ih čini relevantnim faktorom za razmatranje u analizi tipova ćelija.

# Glava 2

## Tehnike prostorne transkriptomike

Tehnike prostorne transkriptomike koje su korišćene u ovom radu su *Stereo-seq* (eng. *SpaTial Enhanced REsolution Omics - SEQuencing*) i *Slide-seqV2*. U okviru ovih tehnika, ulazni parametar predstavlja uzorak tkiva, dok se na izlazu dobijaju podaci o pojedinačnim celijama koje se nalaze u tom tkivu. Ovi podaci obuhvataju položaj tih celija u tkivu i ekspresiju gena koji su ispoljeni u tim celijama.

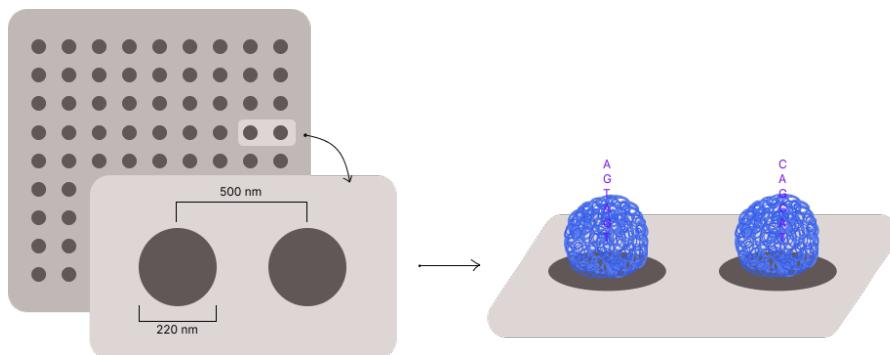
### 2.1 *Stereo-seq* tehnika

*Stereo-seq* tehnika kombinuje uređene nizove DNK nano-kuglica<sup>1</sup> i *in situ* RNK sekvenciranje (sekvenciranje RNK na licu mesta) i može se predstaviti u nekoliko koraka [8, 11]:

1. Priprema čipa uređenih nizova DNK nano-kuglica i formiranje matrice koordinatnih identiteta
  - Koriste se standardni čipovi 13.2 cm x 13.2 cm, sa prostorom za DNK nano-kuglice prečnika od približno 220 nm i rastojanje od centra do centra od 500 nm. Na svako mesto na čipu se postavljaju DNK nano-kuglice koje sadrže nasumične barkodove, tj. sekvence nukleotida (slika 2.1).
  - Izvršava se sekvenciranje kako bi se pročitali nasumični barkodovi prisutni na svakom mestu i onda se na osnovu tih informacija formira matrica koordinatnih identiteta. Svaki element u matrici predstavlja jedinstvenu

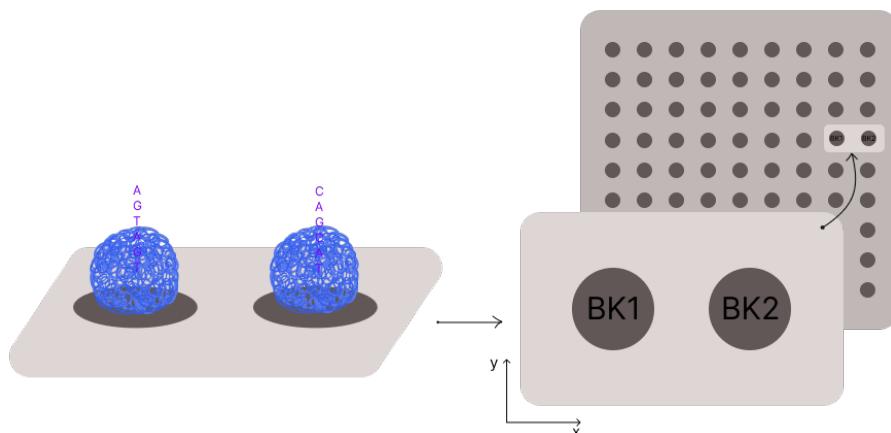
---

<sup>1</sup>DNK nano-kuglice (engl. *DNA nanoballs*) su sferne strukture, izgrađene od kopija jednog molekula DNK, koje su složene u kompaktan i stabilan oblik.



Slika 2.1: Postavljanje DNK nano-kuglica na čip [11]

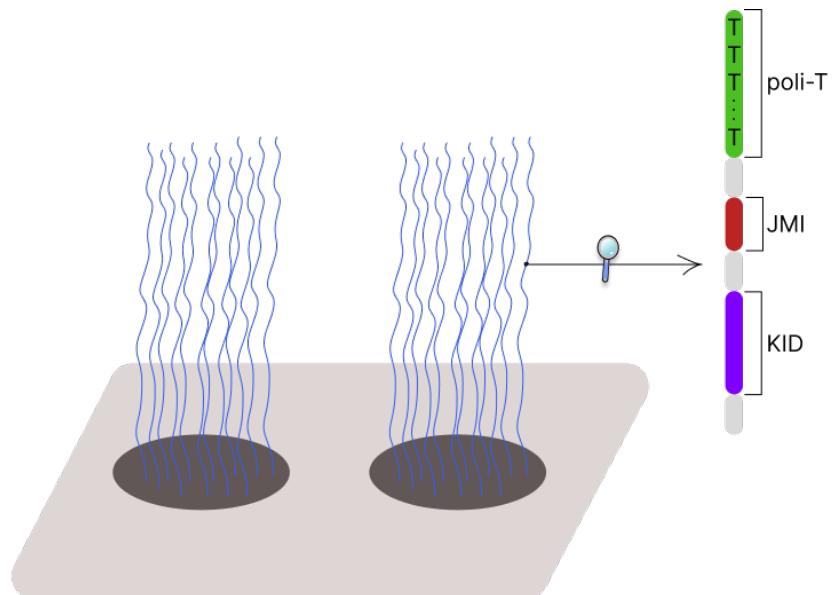
lokaciju na čipu i sadrži odgovarajući barkod, koji ćemo zvati i koordinatni identitet - KID (eng. *Coordinate Identity*) (slika 2.2).



Slika 2.2: Sekvenciranje barkodova koji jedinstveno određuju svako mesto na čipu [11]

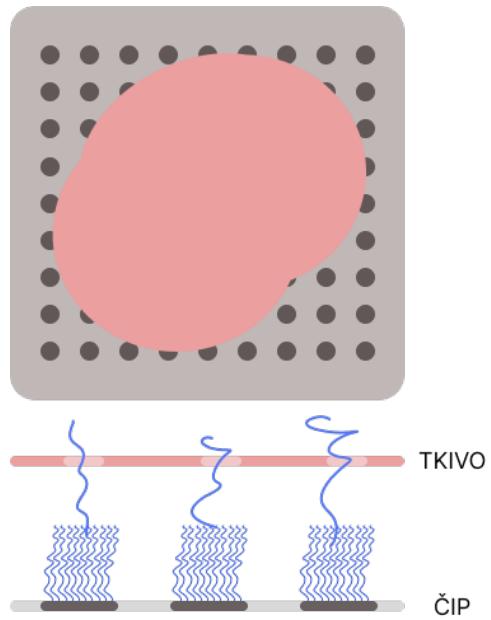
## 2. *In situ* RNK sekvenciranje

- Na svako mesto na čipu za KID (barkod) se vežu:
  - jedinstveni molekularni identifikatori (JMI) koji služe za obeležavanje individualnih molekula, što omogućava istraživačima da tačno kvantifikuju i razlikuju izvorne molekule od onih koji su rezultat amplifikacije i time odrede njegovu ekspresiju i
  - oligonukleotidi sa poli-T sekvencama (sekvenca timina), kako bi se poli-A repovi RNK vezali za njih (slika 2.3)



Slika 2.3: Na čip se dodaju JMI i poli-T sekvenca [11]

- Zamrznuto tkivo se postavlja na čip, fiksira i permeabilizira, omogućavajući difuziju molekula RNK sa poli-A repom iz njega (slika 2.4).



Slika 2.4: Difuzija molekula RNK iz tkiva [11]

- Vezuje se molekul RNK sa poli-A repom za poli-T sekvencu i pomoću reverzne transkripcije na osnovu tog molekula RNK se sintetiše komple-

mentarna DNK - kDNK (slika 2.5). Zatim se vrši amplifikacija, čime se povećava količina materijala potrebnog za sekvenciranje.



Slika 2.5: Sintetisanje kDNK [11]

- Nakon amplifikacije, vrši se sekvenciranje, tj. sakuplja se kDNK (koja sadrži informacije o molekulu RNK), jedinstveni molekularni identifikator (JMI) i koordinatni identitet (KID) koji odgovara položaju na čipu (slika 2.6).



Slika 2.6: Sekvenciranje KID, JMI i kDNK [11]

### 3. Analiza podataka dobijenih *in situ* RNK sekvenciranjem

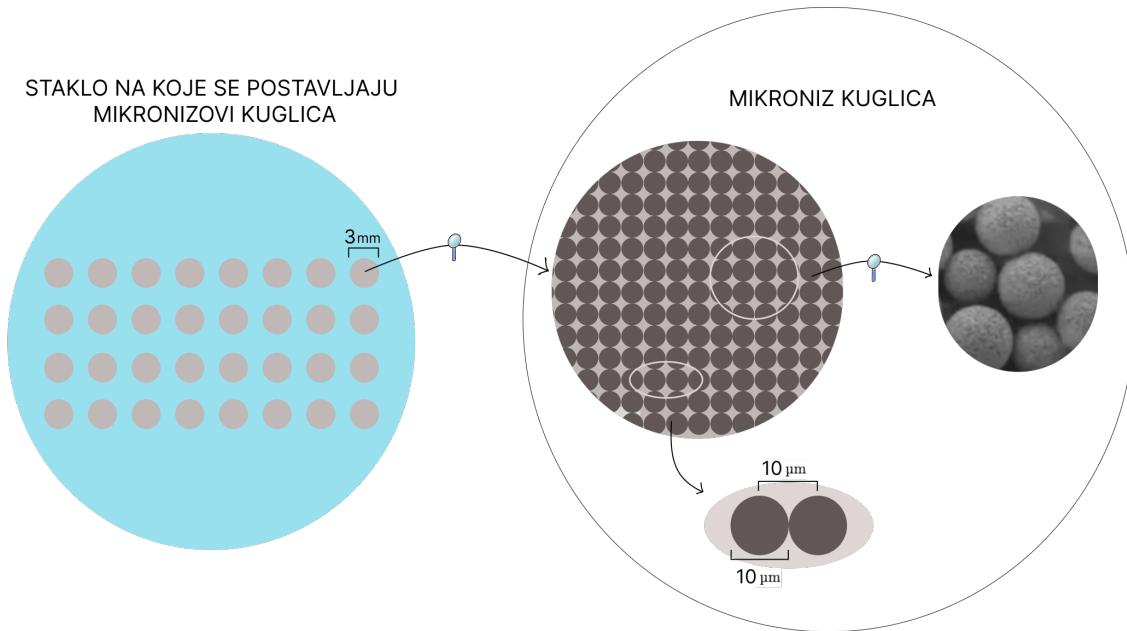
- Na osnovu kDNK molekula i JMI sekvene koja nam pomaže prilikom kvantifikovanja tog kDNK molekula dobijamo nivo ekspresije gena koji odgovara tom kDNK molekulu
- Na osnovu koordinatnog identiteta i matrice koordinatnih identiteta određujemo koordinate u ravni ćelije u kojoj je taj gen ekspresovan

## 2.2 *Slide-seq V2* tehnika

*Slide-seq V2* tehnika radi na sličan način kao i *Stereo-seq* tehnika, osim što kod ove tehnike koristimo mikronizovane kuglice koje nose barkodove (eng. *beads*) umesto čipova DNK nano-kuglica. Mikronizovi su prečnika 3 mm, dok su kuglice prečnika 10 µm postavljene tako da su im rastojanja od centra do centra isto 10 µm (slika 2.7) [8, 10, 9, 12].

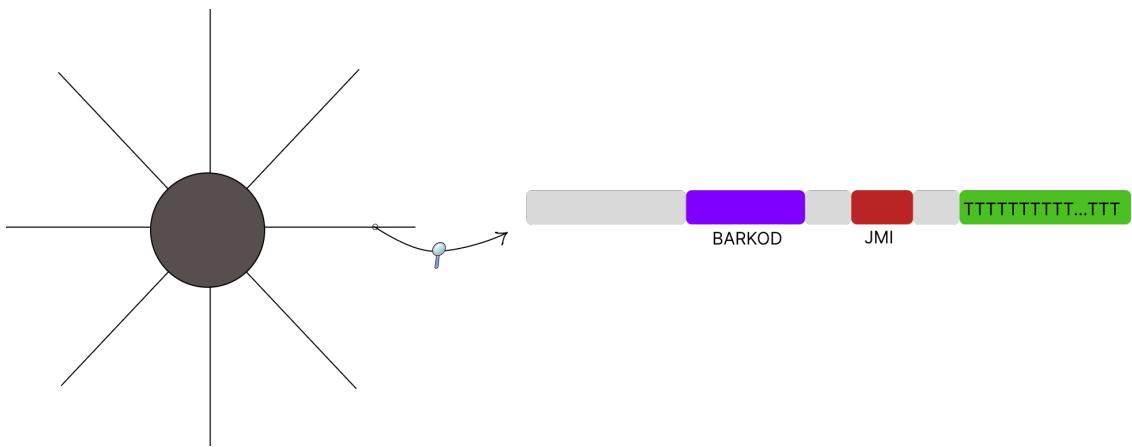
## GLAVA 2. TEHNIKE PROSTORNE TRANSKRIPTOMIKE

---



Slika 2.7: Slide-seq mikroniz na koji su postavljene kuglice koje nose barkodove [8]

Kao što smo kod *Stereo-seq* tehnike korsitili DNK nano-kuglice za barkodove i formiranje matrice koordinatnih identiteta, tako ovde koristimo kuglice. Na njih se kače barkodovi na osnovu kojih formiramo koordinatnu matricu. Svaki element u matrici predstavlja jedinstvenu lokaciju kuglice i sadrži odgovarajući barkod, tj. KID (slika 2.8) [12].



Slika 2.8: Kuglica (eng. *beads*) na koju se kače barkodovi [12]

Ova tehnika je jedna od prvih tehnika prostorne transkriptomike koja je skoro dostigla rezoluciju pojedinačnih ćelija u pristupima koji su zasnovani na barkodo-

## *GLAVA 2. TEHNIKE PROSTORNE TRANSKRIPTOMIKE*

---

vima [8]. Međutim, njena efikasnost nije zadovoljavajuća za korišćenje pojedinačnih celija kao minimalne jedinice zbog same rezolucije koju pružaju date kuglice. Stoga, s obzirom na potrebu za podacima na nivou pojedinačnih celija, ovu tehniku ćemo primenjivati na tkiva koja sadrže celije većih dimenzija [9].

# Glava 3

## Podaci

U ovom istraživanju koristili smo podatke dobijene *Stereo-seq* i *Slide-seq V2* tehnikama. Podaci dobijeni *Stereo-seq* tehnikom obuhvataju podatke za embrion miša i dorzalni srednji mozak miša, dok podaci dobijeni *Slide-seq V2* tehnikom obuhvataju podatake za medulu i korteks zdravog ljudskog bubrega, kao i zdravog bubrega miša.

### 3.1 Podaci dobijeni *Stereo-seq* tehnikom

Podaci koji su dobijeni *Stereo-seq* tehnikom i koji su korišćeni u ovom istraživanju su preuzeti iz baze podataka MOSTA (eng. *Mouse Organogenesis Spatiotemporal Transcriptomic Atlas*) [4]. Ova baza podataka sadrži ukupno 53 sagitalna preseka<sup>1</sup> embriona miša u fazama E9.5, E10.5, E11.5, E12.5, E13.5, E14.5, E15.5 i E16.5<sup>2</sup>. Za faze E9.5 - E15.5, uključeno je četiri do šest preseka iz različitih uzoraka. Što se tiče faze E16.5, analizirano je 18 sagitalnih preseka iz dva biološka uzorka, pri čemu je 13 preseka poteklo iz jednog embriona, omogućavajući pokrivanje svih glavnih tkiva i organa (slika 3.1) [11].

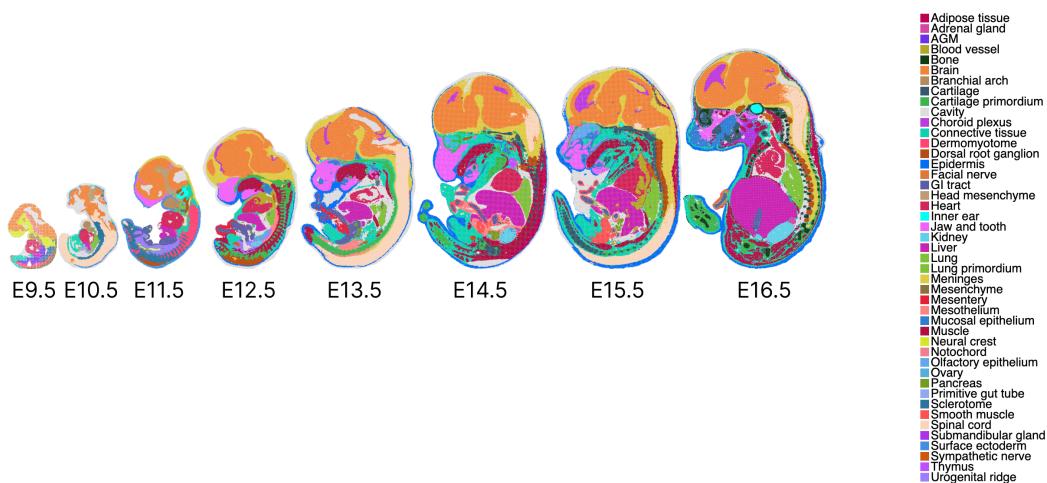
Pored podataka za celokupan embrion miša, u ovoj bazi podataka imamo i izdvojeni skup podataka za dorzalni srednji mozak miša u fazama E12.5, E14.5 i E16.5. Za faze E12.5 i E14.5 imamo po jedan presek, dok za fazu E16.5 imamo tri preseka iz dva uzorka (slika 3.2).

Korišćeni skupovi podataka su u *H5ad* formatu (eng. *Hierarchical Data Format 5*), koji se koristi za skladištenje podataka o sekvenciranju pojedinačnih ćelija

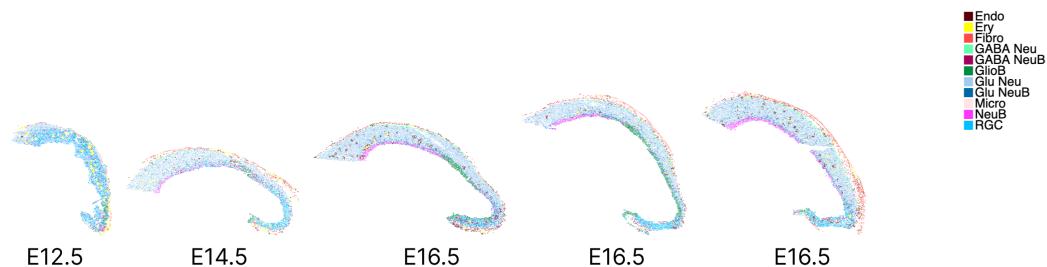
---

<sup>1</sup>Sagitalni presek embriona miša predstavlja presek po sagitalnoj ravni, ravni koja deli embrion na levu i desnu polovicu.

<sup>2</sup>Faze E9.5 - E16.5 se odnose na različite periode embrionalnog razvoja kod miševa. Oznaka „E“ označava embrionalni dan, a brojevi koji slede označavaju broj dana nakon začeća.



Slika 3.1: Tipovi ćelija (zajedno) za embrion miša u različitim fazama [4]



Slika 3.2: Tipovi ćelija (zajedno) za dorzalni srednji mozak miša u različitim fazama [4]

i uključuje matricu ekspresije gena, koordinate ćelija i eksperimentalno utvrđene tipove ćelija, kao i dodatne metapodatke vezane za svaku ćeliju.

## Embrion miša

Za embrion miša smo posmatrali tri preseka, dva u fazi E9.5 iz različitih uzoraka (E1S1 i E2S4) i jedan u fazi E10.5 (E2S1). Radi lakšeg referenciranja, presek embriona miša E1S1 u fazi E9.5 ćemo zvati prvi presek, presek E2S4 u fazi E9.5 drugi presek, a presek E2S1 u fazi E10.5 treći presek embriona miša. Tabelarni prikaz matrice genskih ekspresija za prvi presek embriona miša dat je na slici 3.3. Uočavamo da ovaj skup podataka ima 5913 ćelija i 25568 gena za koje važi da su ispoljeni bar u jednoj ćeliji. Tabelarni prikaz niza koordinata ćelija je prikazan na slici 3.4, dok su eksperimentalno utvrđeni tipovi ćelija prikazani na slikama 3.5 (prikaz datoteke

### GLAVA 3. PODACI

---

Redni broj preseka	Broj ćelija	Broj gena	Broj tipova ćelija
1	5913	25568	12
2	5797	23398	13
3	8494	22385	18

Tabela 3.1: Statistike za embrion miša

sa podacima o tipovima ćelija) i 3.6 (grafički prikaz). Ovde imamo 12 tipova ćelija i redosled tipova ćelija odgovara redosledu indeksa niza koordinata, kao i redosledu redova u matrici genskih ekspresija. Tipovi ćelija za preostala dva preseka prikazani su na slikama 3.7 i 3.8, dok su pojedinačni tipovi ćelija za sva tri preseka prikazani na slici 3.9. Detaljne statistike, odnosno broj ćelija, gena i tipova ćelija, za svaki presek prikazane su u tabeli 3.1.

0	1	2	3	4	5	6	7	8	9	...	25558	25559	25560	25561	25562	25563	25564	25565	25566	25567
0	4.066529	2.620700	1.99778	3.276793	1.99778	2.620700	1.99778	2.6207	2.6207	1.997780	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	2.528584	0.000000	0.00000	3.395828	0.00000	1.912187	0.00000	0.0000	0.0000	1.912187	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	2.759596	0.000000	0.00000	3.244769	0.00000	0.000000	0.00000	0.0000	0.0000	1.780219	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	2.878432	0.000000	0.00000	1.886381	0.00000	0.000000	0.00000	0.0000	0.0000	1.886381	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	2.643301	0.000000	0.00000	2.379051	0.00000	1.450415	0.00000	0.0000	0.0000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5908	2.291063	2.291063	0.00000	3.421593	0.00000	0.000000	0.00000	0.0000	0.0000	1.376702	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5909	1.646878	0.000000	0.00000	1.646878	0.00000	1.646878	0.00000	0.0000	0.0000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5910	1.859816	0.000000	0.00000	1.859816	0.00000	0.000000	0.00000	0.0000	0.0000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5911	1.855589	0.000000	0.00000	0.000000	0.00000	0.000000	0.00000	0.0000	0.0000	1.855589	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5912	0.000000	0.000000	0.00000	0.000000	0.00000	5.050670	0.00000	0.0000	0.0000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5913 rows x 25568 columns

Slika 3.3: Tabelarni prikaz matrice genskih ekspresija za prvi presek embriona miša

	x	y
0	104.0	-147.0
1	105.0	-147.0
2	106.0	-147.0
3	107.0	-147.0
4	108.0	-147.0
...	...	...
5908	110.0	-253.0
5909	111.0	-253.0
5910	112.0	-253.0
5911	113.0	-253.0
5912	114.0	-253.0

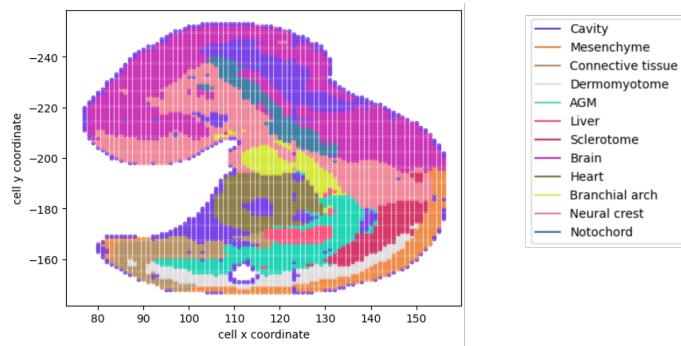
5913 rows x 2 columns

Slika 3.4: Tabelarni prikaz niza koordinata ćelija za prvi presek embriona miša

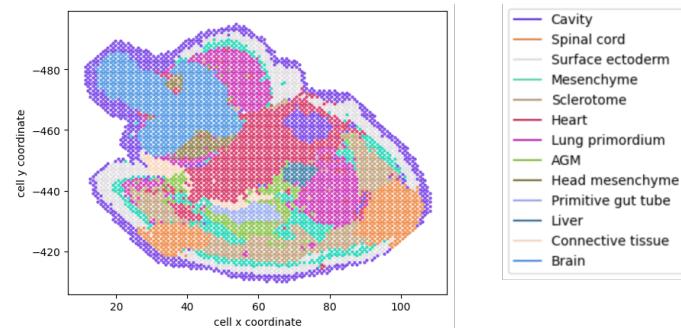
### GLAVA 3. PODACI

```
cell_name
147_104      Cavity
147_105      Cavity
147_106      Cavity
147_107      Cavity
147_108      Mesenchyme
...
253_110      Cavity
253_111      Cavity
253_112      Cavity
253_113      Cavity
253_114      Cavity
Name: annotation, Length: 5913, dtype: category
Categories (12, object): ['AGM', 'Brain', 'Branchial arch', 'Cavity', ..., 'Mesenchyme', 'Neural crest', 'Notochord', 'Sclerotome']
```

Slika 3.5: Tipovi ćelija za prvi presek embriona miša



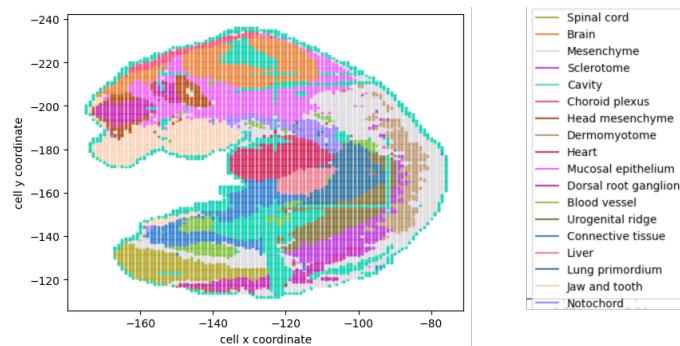
Slika 3.6: Tipovi ćelija (zajedno) za prvi presek embriona miša



Slika 3.7: Tipovi ćelija (zajedno) za drugi presek embriona miša

### Dorzalni srednji mozak miša

Grafički prikaz skupa podataka za dorzalni srednji mozak miša koji je korišćen u ovom radu je dat na slici 3.2. Ovaj skup podataka sadrži matricu ekspresije gena, koordinate ćelija i eksperimentalno utvrđene tipove ćelija za tri faze embrionalnog razvoja miša, po jedan presek iz prve i druge faze i tri preseka iz treće faze. Matrica genskih ekspresija za sve preseke dorzalnog srednjeg mozga miša sadrži 26738 ćelija i



Slika 3.8: Tipovi ćelija (zajedno) za treći presek embriona miša

23955 gena. Sa slike 3.10 gde smo date podatke predstavili tabelarno možemo videti da imamo 24045 gena. Međutim, taj broj je manji jer imamo gene čija je ekspresija nula u svakoj ćeliji.

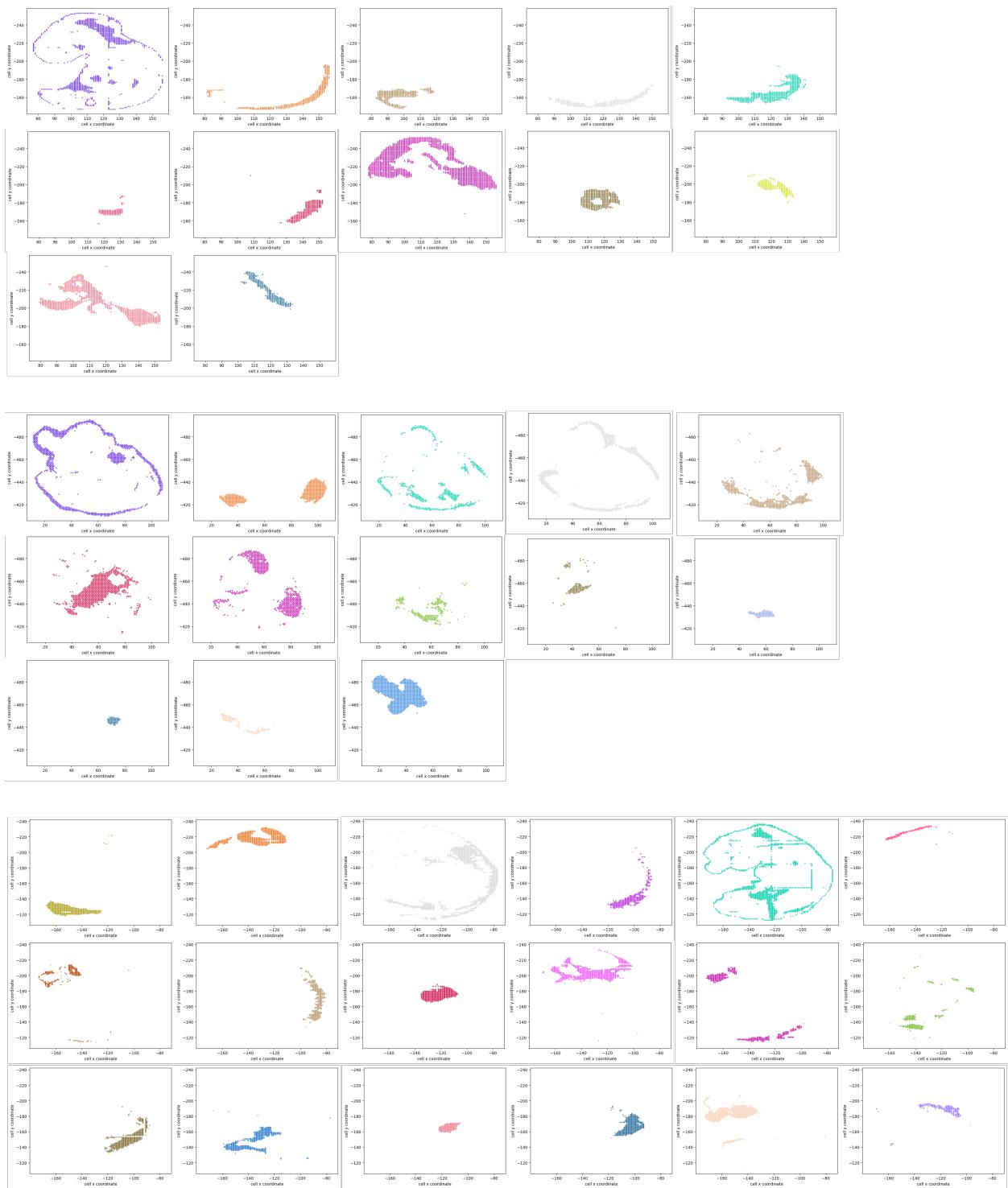
Sa slike 3.2 (grafički prikaz) i 3.11 (priček datoteke sa podacima o tipovima ćelija) uočavamo 11 tipova ćelija za ovaj skup podataka.

Kolona koja nam predstavlja jedinstveni identifikator ćelije nam može pomoći prilikom određivanja koje ćelije pripadaju kom preseku. Ćelije koje imaju oznaku „\_1” na kraju predstavljaju ćelije koje pripadaju prvom preseku, one sa oznakom „\_2” pripadaju drugom preseku, itd. Dodatno, kao što je to bio slučaj kod embriona miša, i ovde redosled eksperimentalno utvrđenih tipova ćelija odgovara redosledu indeksa niza koordinata, kao i redosledu redova u matrici genskih ekspresija. Detaljne statistike za svaki presek su prikazane u tabeli 3.2, dok su na slici 3.12 prikazni pojedinačni tipovi ćelija za svih pet preseka.

Redni broj preseka	Broj ćelija	Broj gena	Broj tipova ćelija
1	3671	17972	11
2	6650	21179	11
3	3648	18442	11
4	5503	20540	11
5	7266	21294	11

Tabela 3.2: Statistike za dorzalni srednji mozak miša

### GLAVA 3. PODACI



Slika 3.9: Tipovi ćelija (pojedinačno) za sva tri preseka embriona miša

### GLAVA 3. PODACI

---

	0	1	2	3	4	5	6	7	8	9	...	24035	24036	24037	24038	24039	24040	24041	24042	24043	24044
0	0.000000	0.0	0.725739	0.725739	1.141836	1.141836	0.725739	0.725739	0.725739	0.725739	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
1	0.000000	0.0	0.000000	1.426898	0.000000	1.748962	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
2	0.000000	0.0	0.000000	2.051748	0.677417	0.000000	0.000000	1.362605	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
3	0.000000	0.0	0.000000	0.934627	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
4	0.000000	0.0	0.000000	2.249852	0.000000	0.992239	0.000000	0.000000	0.992239	1.806943	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
26733	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	1.206665	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
26734	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
26735	0.954521	0.0	0.000000	0.000000	0.000000	0.000000	0.954521	2.195794	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
26736	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	1.006593	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
26737	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.108686

26738 rows x 23955 columns

Slika 3.10: Tabelarni prikaz matrice genskih ekspresija za dorzalni srednji mozak miša

```

CELL.100034_1      RGC
CELL.100035_1      Ery
CELL.100191_1      Ery
CELL.100256_1      Ery
CELL.100264_1      Ery
...
CELL.326357_5      Micro
CELL.326359_5      Fibro
CELL.326384_5      Fibro
CELL.326391_5      Fibro
CELL.326412_5      Fibro
Name: annotation, Length: 26738, dtype: category
Categories (11, object): ['Endo', 'Ery', 'Fibro', 'GABA Neu', ..., 'Glu NeuB', 'Micro', 'NeuB', 'RGC']

```

Slika 3.11: Tipovi ćelija za dorzalni srednji mozak miša

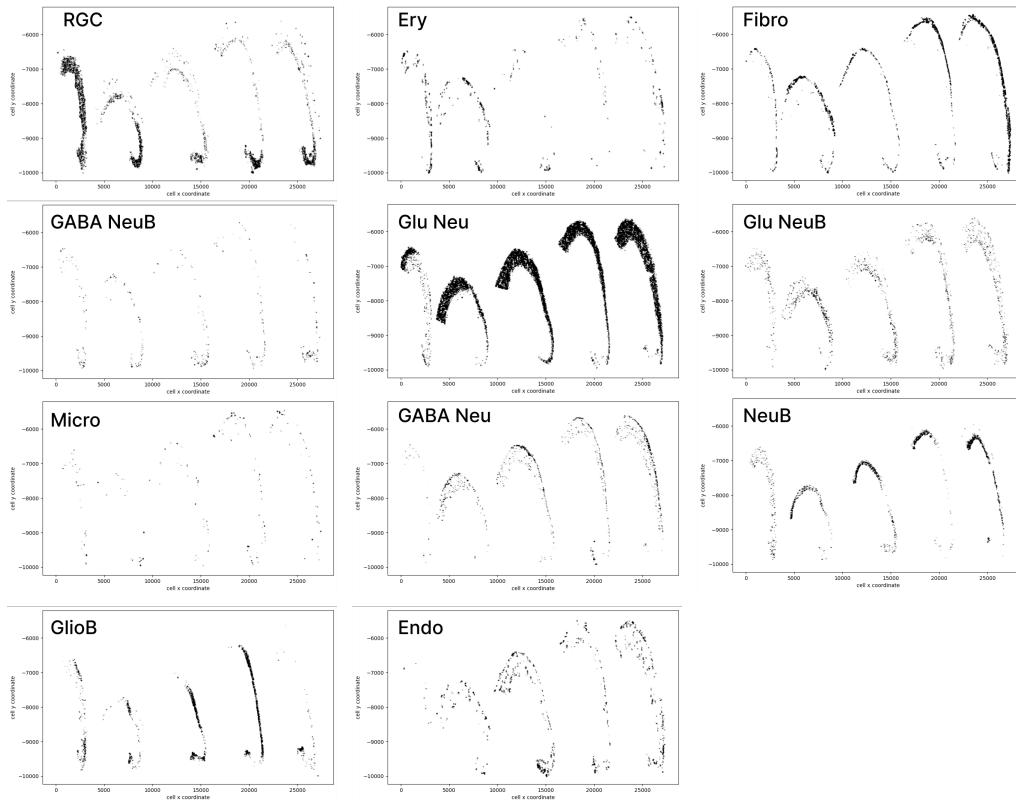
## 3.2 Podaci dobijeni *Slide-seqV2* tehnikom

Skupovi podataka koji su dobijeni *Slide-seqV2* tehnikom i koji su korišćeni u ovom radu su skupovi za medulu i korteks zdravog ljudskog bubrega i zdravog bubrega miša [5]. Ovi podaci su u *H5ad* formatu, kao što je to bio slučaj i za podatke dobijene *Stereo-seq* tehnikom.

Za ovaj skup podataka prvo ćemo analizirati podatke koji sadrže tipove ćelija, budući da je iz njih jasno uočljivo da koristimo jedinstvene barkodove umesto jedinstvenih identifikatora za ćelije (slika 3.13). Kao što je objašnjeno u poglavlju 2.2, ova tehnika ne obezbeđuje da najmanja jedinica bude tačno jedna ćelija. U ovom slučaju će obuhvatati jednu do tri ćelije [9]. S obzirom na to da je većina ćelija u bubregu čoveka i miša veća od 10 µm, ovaj pristup je prikladan u ovom kontekstu.

Datoteke koje sadrže niz koordinata ćelija i matricu genskih ekspresija imaju istu strukturu kao i te iste datoteke sa skupovima podataka koji su dobijeni *Stereo-seq* tehnikom. Takođe, redosled tipova ćelija odgovara redosledu indeksa niza koordinata, kao i redosledu redova u matrici genskih ekspresija. Detaljne statistike za sve

### GLAVA 3. PODACI



Slika 3.12: Tipovi ćelija (pojedinačno) za dorzalni srednji mozak miša

```

AAAAAAAATTAA      endothelial cell
AAAAAAAGCAAAA     endothelial cell
AAAAAAACAAAAAC    endothelial cell
AAAAAAACAAACAC    endothelial cell
AAAAAAATAAATAT    endothelial cell
...
TTTTTTTTGATATT    endothelial cell
TTTTTTTTGTAGTT    parietal epithelial cell
TTTTTTTTGTCCCTT   endothelial cell
TTTTTTTTTAGTTTT   endothelial cell
TTTTTTTTTTTTTTTT  endothelial cell
Name: cell_type, Length: 12531, dtype: category
Categories (17, object): ['native cell', 'endothelial cell', 'mesangial cell', 'podocyte', ...]

```

Slika 3.13: Tipovi ćelija za medulu ljudskog bubrega

skupove podataka su prikazane u tabeli 3.3.

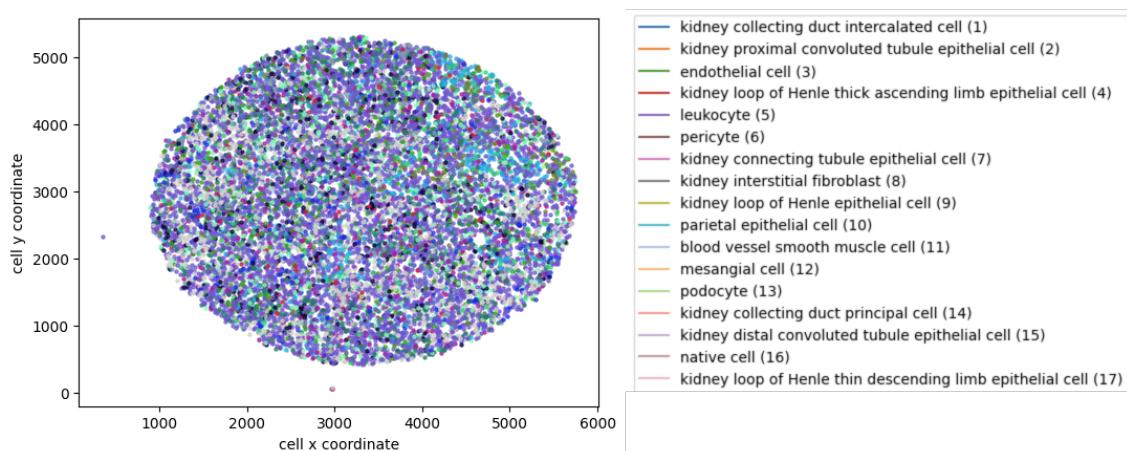
Tipovi ćelija za sve preseke su prikazani na slikama 3.14, 3.15 i 3.17, dok je detaljniji prikaz svakog tipa pojedinačno dat na slikama 3.16 i 3.18.

### GLAVA 3. PODACI

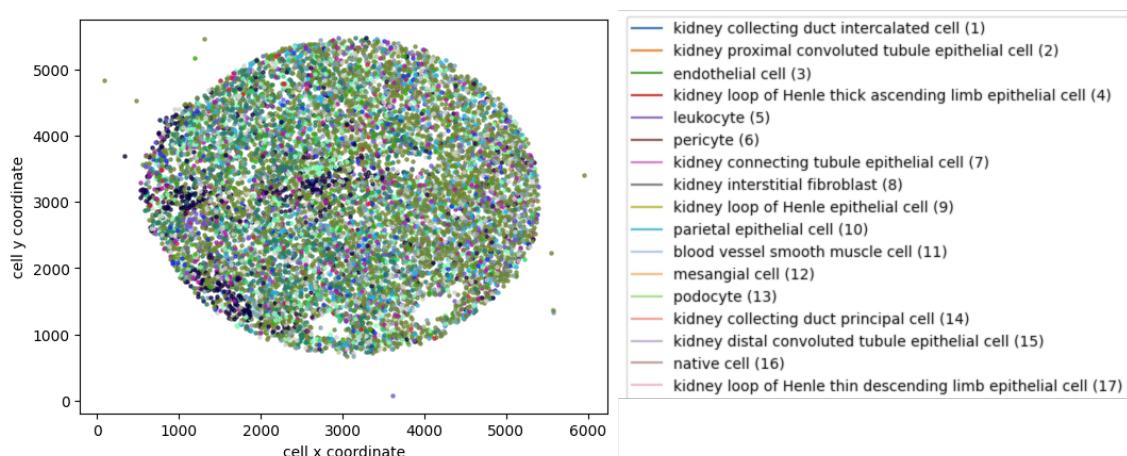
---

Skup podataka	Broj ćelija	Broj gena	Broj tipova ćelija
Medula ljudskog bubrega	12531	16816	17
Korteks ljudskog bubrega	10419	16398	17
Bubreg miša	10888	15718	13

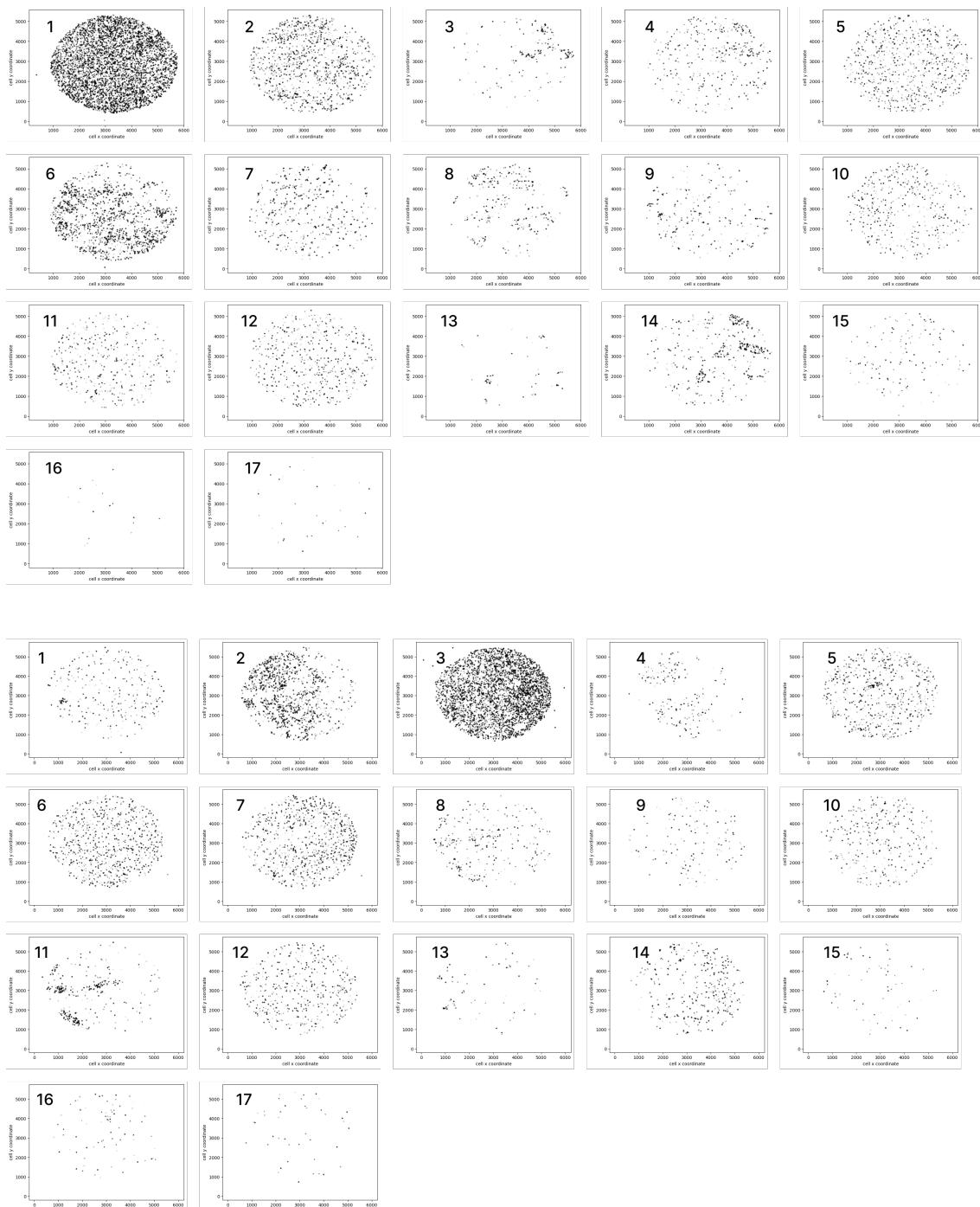
Tabela 3.3: Statistike za medulu i korteks ljudskog bubrega i bubreg miša



Slika 3.14: Tipovi ćelija (zajedno) za medulu ljudskog bubrega

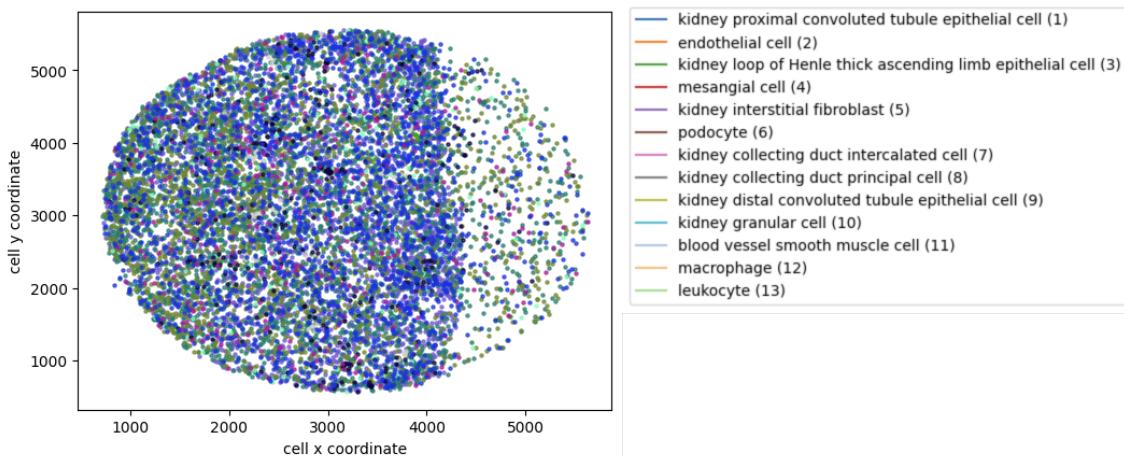


Slika 3.15: Tipovi ćelija (zajedno) za korteks ljudskog bubrega

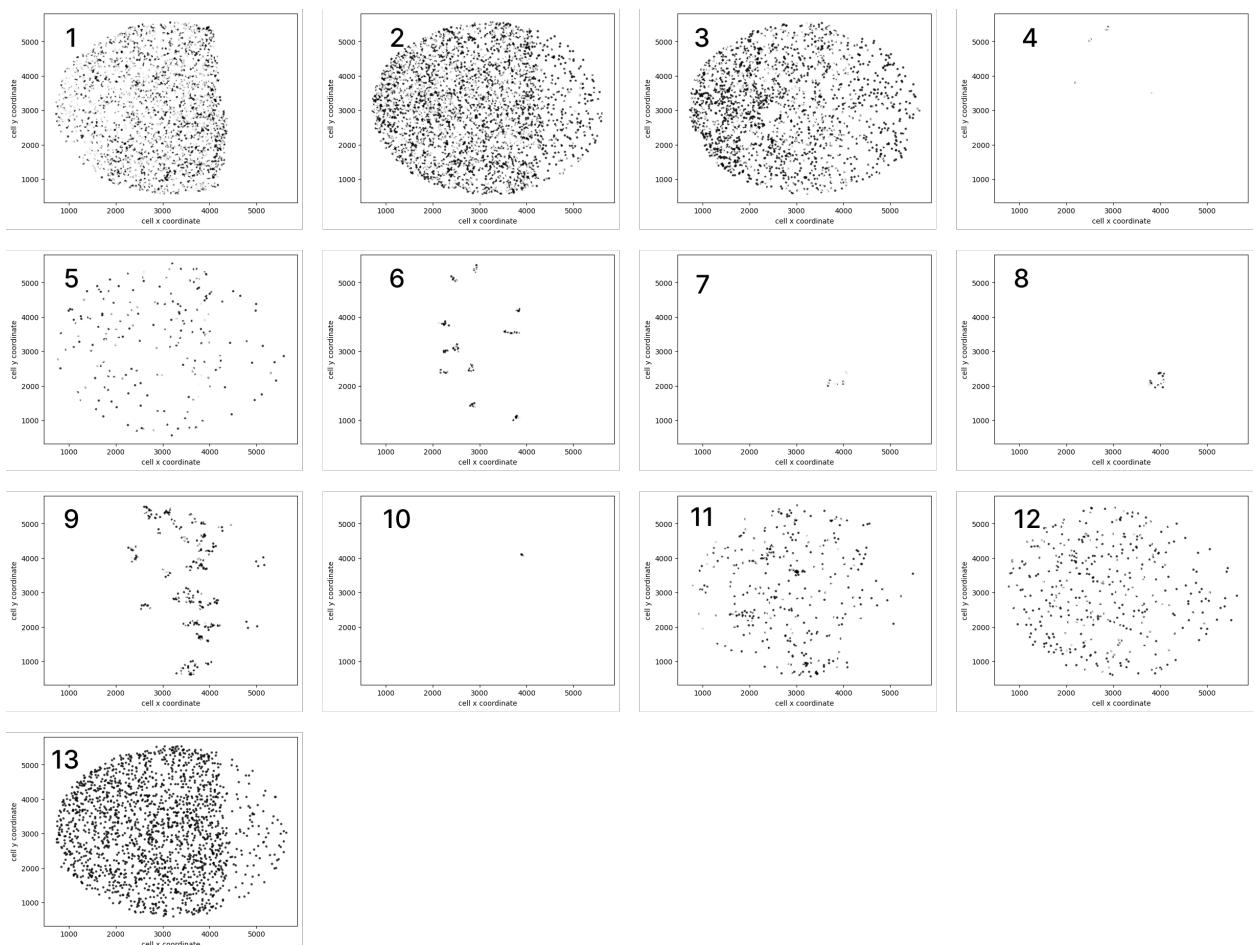


Slika 3.16: Tipovi ćelija (pojedinačno) za medulu i korteks ljudskog bubrega, redom

### GLAVA 3. PODACI



Slika 3.17: Tipovi ćelija (zajedno) za bubreg miša



Slika 3.18: Tipovi ćelija (pojedinačno) za bubreg miša

# Glava 4

## Problemi i metode

Dva glavna problema koja će biti istražena u ovom radu su sledeća:

1. Ispitati zavisnosti između koordinata ćelija i njihovih genskih ekspresija.
  - Metoda: Definisati meru koja će što bolje opisati ovu zavisnost između koordinata ćelija i njihovih genskih ekspresija (u daljem tekstu Metoda 1).
2. Ispitati uticaja prostorne i genske komponente na tip ćelije.
  - Metoda: Primeniti *Leiden* algoritam klasterovanja nad podacima koji su dobijeni tehnikama prostorne transkriptomike (u daljem tekstu Metoda 2).

### 4.1 Metoda 1

Neka su grafovi  $G_1$  i  $G_2$  koordinatni i genski graf, redom, koji se formiraju na osnovu niza koordinata ćelija, odnosno na osnovu matrice genskih ekspresija. Kod oba grafa čvorovi su ćelije i svaka dva čvora su povezana otežanom granom. U koordinatnom grafu  $G_1$  težine grana su euklidska rastojanja između koordinata ćelija, dok su u grenskom grafu  $G_2$  to euklidska rastojanja između vektora redukovanih genskih ekspresija, gde je redukcija genskih ekspresija urađena primenom *PCA* metode (eng. *Principal Component Analysis*).

Neka je  $D^1$  matrica susedstva za graf  $G_1$  gde je:

$$D_{ij}^1 = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

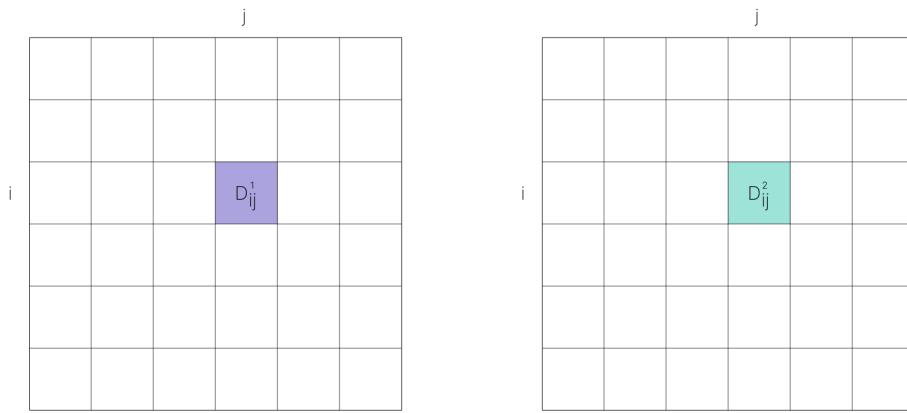
## GLAVA 4. PROBLEMI I METODE

---

rastojanje između koordinata  $(x_i, y_i)$  ćelije  $c_i$  i  $(x_j, y_j)$  ćelije  $c_j$ , a  $D^2$  matrica susedstva za graf  $G_2$  gde je:

$$D_{ij}^2 = \sqrt{\sum_{k=1}^n (pca_{ik} - pca_{jk})^2}$$

rastojanje između vektora redukovanih genskih ekspresija  $\mathbf{pca}_i = (pca_{i1}, \dots, pca_{in})$  ćelije  $c_i$  i  $\mathbf{pca}_j = (pca_{j1}, \dots, pca_{jn})$  ćelije  $c_j$ , gde je  $n$  broj PCA komponenti. Matrice susedstva  $D^1$  i  $D^2$  su prikazane na slici 4.1.



Slika 4.1: Matrice susedstva za grafove  $G_1$  i  $G_2$

Kako bi moglo da se uradi poređenje koordinatnog i genskog grafa, odnosno poređenje rastojanja iz njihovih matrica susedstva  $D^1$  i  $D^2$ , potrebno je uraditi normalizaciju tih vrednosti. Ovde je primenjena min-max normalizacija vrednosti na opseg  $[0, 1]$  tako što je od svih rastojanja iz datih matrica susedstva oduzeta minimalna vrednost, a zatim su takva rastojanja podeljena sa razlikom maksimalne i minimalne vrednosti.

Neka je *CDM* matrica (eng. *Cell Distribution Metric*) razlika redukovanih i normalizovanih matrica susedstva  $D^1$  i  $D^2$ . Njene vrednosti su u opsegu  $[-1, 1]$  i biće blizu nule kada za dve ćelije važi da imaju približno koordinatno i gensko rastojanje, bilo da je ono veliko (ćelije su prostorno udaljene i vektori genskih ekspresija su im različiti) ili malo (ćelije su prostorno bliske i vektori genskih ekspresija su im slični). U tom kontekstu, ova matrica predstavlja poređenje grafova  $G_1$  i  $G_2$ .

Ono što bismo želeli da ispitamo jeste šta nam *CDM* matrica može otkriti o samom odnosu između koordinata ćelija i njihovih genskih ekspresija. Ako bi raspodela vrednosti iz *CDM* matrice bila normalna tako da je njena srednja vrednost

bliska nuli, onda bi to značilo da imamo gomilanje vrednosti oko nule, odnosno da u datom uzorku postoji veliki broj ćelija kod kojih su vrednosti iz matrice  $D^1$  slične odgovarajućim vrednostima iz matrice  $D^2$ . Ovo bi značilo da postoji veliki broj ćelija koje su prostorno bliske i imaju sličnu gensku ekspresiju ili suprotno - prostorno su udaljene i imaju različite genske ekspresije. Ako bi srednja vrednost bila pomerena u levo ili desno, to bi značilo da imamo veliki broj ćelija u datom uzorku koje su prostorno bliske, ali se razlikuju po genskim ekspresijama ili suprotno - prostorno su udaljene, ali su im genske ekspresije slične.

## 4.2 Metoda 2

Rastojanja između čvorova kod grafova  $G_1$  i  $G_2$  su normalizovana pomoću min-max normalizacije, a zatim su oni redukovani tako da svaki čvor zadrži najviše 30 najbližih suseda. Graf  $G$ , koji predstavlja uniju grafova  $G_1$  i  $G_2$ , ima iste čvorove kao i grafovi  $G_1$  i  $G_2$  (ćelije su čvorovi), dok grana između dva čvora  $v_1$  i  $v_2$  postoji ako postoji grana između ta dva čvora u bar jednom od redukovanih grafova  $G_1$  i  $G_2$  i ona je jednaka:

$$k \cdot (t_1 + t_2)$$

gde je  $t_i$  težina grane između  $v_1$  i  $v_2$  u  $G_i$  ako takva grana postoji u tom grafu, inače je  $t_i = 0$  ( $i = 1, 2$ ), dok je  $k = \frac{1}{2}$  ako grana između  $v_1$  i  $v_2$  postoji u oba redukovana grafa  $G_1$  i  $G_2$ , a inače je  $k = 1$ .

Klasterovanje grafa  $G$  će biti urađeno pomoću *Leiden* algoritma. *Leiden* algoritam je algoritam za detekciju klastera unutar grafa, čiji se kvalitet pronađaska gusto povezanih klastera unutar grafa meri pomoću modularnosti. Modularnost u *Leiden* algoritmu se računa na sledeći način:

$$Q = \frac{1}{2m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

gde je  $A_{ij}$  težina grane između čvorova  $i$  i  $j$ ,  $k_i$  i  $k_j$  sume težina grana vezanih za čvorove  $i$  i  $j$ ,  $2m$  suma svih težina grana u grafu,  $c_i$  i  $c_j$  dodeljeni klasteri za čvorove  $i$  i  $j$  i  $\delta(c_i, c_j)$  Kronekerova delta funkcija. Ona može imati vrednosti iz opsega  $[-1, 1]$ , gde su vrednosti manje od  $-0.5$  retke [7, 3]. Visoke pozitivne vrednosti, bliske 1, ukazuju na to da su čvorovi unutar istog klastera gusto povezani, sa manje veza za ostalim klasterima. Suprotno, vrednosti bliske nuli ili manje ukazuju na to da nema jasne podele na klastera, odnosno da podela na klastera nije bolja od nasumične.

## *GLAVA 4. PROBLEMI I METODE*

---

Ovaj algoritam se izvršava u četiri koraka koja se ponavaljaju sve dok je poboljšanje modularnosti moguće [13]:

### 1. Faza brzog lokalnog pomeranja:

- Ulazni parametar za ovu fazu su graf  $G$  i particija  $P$ , koja je na početku takva da je svaki čvor zaseban klaster. Izlazni parametar je unapređena particija  $P$ .
- Algoritam iterira kroz niz svih čvorova  $Q$  u grafu  $G$ .
- Za svaki čvor  $v$  određuje se klaster  $C$  koji će doprineti najvećem povećanju modularnosti particije  $P$ .
- Ako pomeranje čvora  $v$  u klaster  $C$  poboljšava modularnost particije  $P$ , čvor se premešta u taj klaster i na taj način se ažurira particija  $P$ .
- Zatim se određuju susedi čvora  $v$  koji nisu u klasteru  $C$  i koji nisu već u nizu  $Q$  i oni se dodaju u niz  $Q$ .
- Prethodni koraci se ponavljaju sve dok se ne posete svi čvorovi iz niza  $Q$ .

### 2. Faza unapređenja:

- Ulazni parametri za ovu fazu su graf  $G$  i particija  $P$ . Izlazni parametar je unapređena particija  $P_{refined}$ .
- U ovoj fazi algoritam pokušava da dodatno unapredi klastere.
- Na početku, u unapređenu particiju  $P_{refined}$  se stavljuju svi čvorovi grafa  $G$ , koji se ponovo tretiraju kao odvojeni klasteri.
- Prolazi se iterativno kroz klastere  $C$  particije  $P$  i unapređuje se klaster  $C$ :
  - Prolazi se kroz one čvorove  $v$  klastera  $C$  koji su gusto povezani u tom klasteru i za koje važi da nisu već spojeni sa nekim drugim klasterom u ovoj fazi.
  - Zatim se od onih klastera iz unapređene particije  $P_{refined}$  koji su gusto povezani na slučajan način bira klaster  $C'$  i cvor  $v$  se premešta u njega.
  - Prethodni korak povećava modularnost jer se ne razmatraju premeštanja koja bi smanjila modularnost.

## GLAVA 4. PROBLEMI I METODE

---

- Nakon završetka ove faze, klasteri iz particije  $P$  mogu biti podeljeni na vise manjih klastera unutar unapređene particije  $P_{refined}$ .

### 3. Agregacija:

- Ulazni parametri za ovu fazu su graf  $G$  i unapređena particija  $P_{refined}$ . Izlazni parametar je agregirani graf  $G_{agr}$ .
- Nakon faze unapređenja, svaki klaster iz particije  $P_{refined}$  postaje čvor u novom grafu  $G_{agr}$ .
- Grane se formiraju između čvorova (klastera) na osnovu veza između njihovih članova u originalnom grafu  $G$ .
- Rastojanje izmedju dva klastera iz  $G$ , odnosno dva čvora iz  $G_{agr}$ , predstavlja sumu rastojanja između čvorova tih klastera.

### 4. Particija koja je dobijena iz prvog koraka $P$ se ažurira vrednostima za nove čvorove iz agregiranog grafa $G_{agr}$ . Odnosno, ako je skup čvorova iz $G$ sada jedan čvor u $G_{agr}$ , to će biti ažurirano u particiji $P$ .

Tokom klasterovanja probaćemo kombinacije različitog broja najbližih suseda za oba grafa. Broj najbližih suseda za koordinatni graf uzima vrednosti iz skupa  $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20\}$ , dok broj najbližih suseda za genski graf uzima vrednosti iz skupa  $\{5, 10, 15, 20, 25, 30\}$ . Za određivanje optimalanog broja suseda koji će nam pružiti klastera koji najviše odgovaraju eksperimentalno utvrđenim tipovima ćelija koristićemo *ARI* skor (eng. *Adjusted Random Index*). *ARI* skor je statistička metrika koja se koristi za merenje sličnosti između dve particije skupa podataka, obično između stvarne particije (eng. *ground truth*), u ovom slučaju eksperimentalno utvrđenih tipova ćelija i particije koja je dobijena analizom, odnosno klastera koje smo dobili na osnovu *Leiden* algoritma i on se računa na sledeći način:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[ \sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[ \sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \left[ \sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}$$

gde je  $n_{ij}$  broj čvorova koji su istovremeno u klasteru  $i$  u prvoj particiji i u klasteru  $j$  u drugoj particiji,  $a_i$  broj čvorova u klasteru  $i$  u prvoj particiji,  $b_j$  broj čvorova u klasteru  $j$  u drugoj particiji,  $n$  ukupan broj čvorova u grafu,  $\binom{n}{2}$  broj svih mogućih parova čvorova u grafu. On može uzimati vrednosti iz opsega  $[-1, 1]$ , gde su vrednosti manje od  $-0.5$  retke [1]. Visoke pozitivne vrednosti, bliske 1, ukazuju da

## *GLAVA 4. PROBLEMI I METODE*

---

su dve particije (stvarna i dobijena analizom) vrlo slične. Suprotno, vrednosti bliske nuli ili manje ukazuju na to da su dve particije nezavisne, odnosno da između njih nema sličnosti. Optimalan broj suseda za koordinatni i genski graf ukazaće na uticaj koordinata ćelije, odnosno genskih ekspresija ćelije na tip ćelije.

U nastavku ovog rada, ispitaćemo predstavljene metodologije na skupovima podataka za embrion i dorzalni srednji mozak miša, kao i na meduli i korteksu ljudskog bubrega i bubregu miša.

# Glava 5

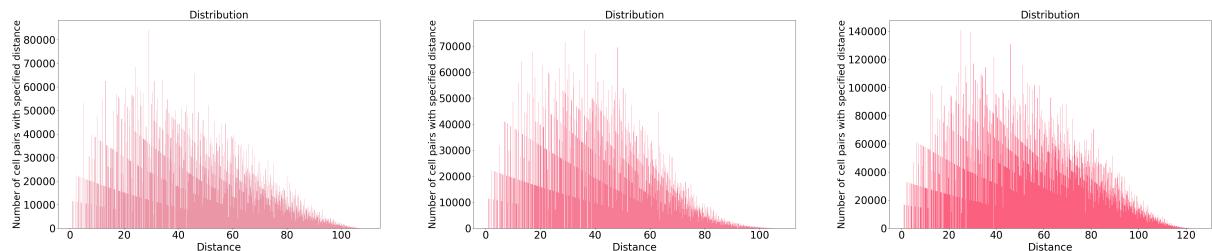
## Rezultati

U ovom poglavlju biće prikazani rezultati koji su dobijeni za embrion i dorzalni srednji mozak miša, kao i za medulu i korteks ljudskog bubrega i bubreg miša.

### 5.1 Embrion miša

#### Problem 1

Raspodele vrednosti iz matrica  $D^1$  za tri preseka embriona miša su prikazane na slici 5.1.



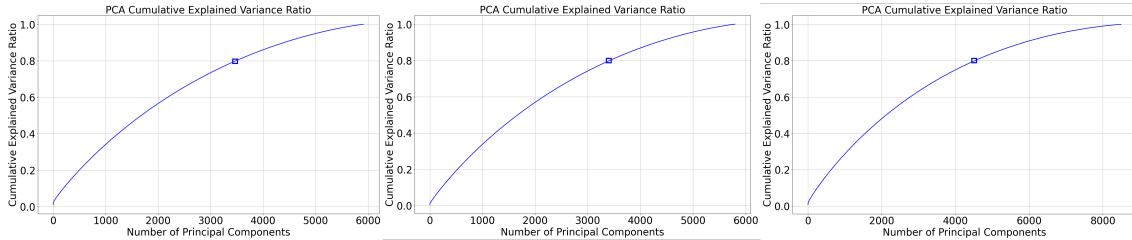
Slika 5.1: Raspodela vrednosti iz matrica susedstva  $D^1$  za tri preseka embriona miša

Pre formiranja matrice susedstva  $D^2$  za ove skupove podataka, iz vektora genskih ekspresija izbačene su vrednosti za gene koji nisu ispoljeni ni u jednoj ćeliji. Zatim je urađena redukcija genskih komponenti pomoću *PCA* metode. Slika 5.2 prikazuje koliko procenta varijanse podataka je pokriveno u odnosu na broj *PCA* komponenti. Kako je potrebno dosta komponenti da se pokrije 80% varijanse podataka (za prvi i drugi presek 3400, dok za treći 4500 *PCA* komponenti), prvo su urađene analize

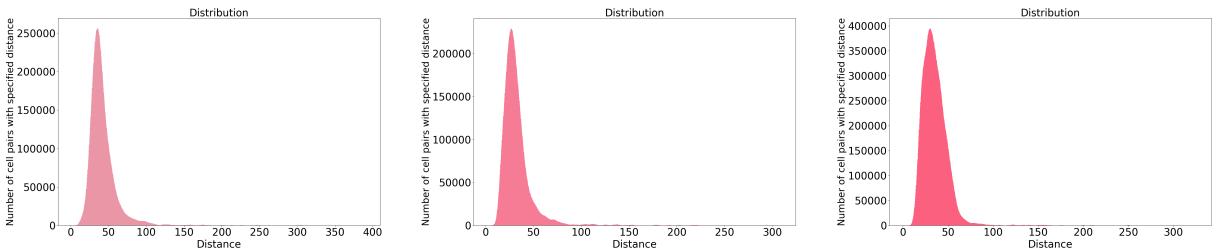
## GLAVA 5. REZULTATI

---

sa manjim brojem *PCA* komponenti. Raspodele vrednosti iz matrica  $D^2$  za embrion miša sa 40 *PCA* komponenti su prikazane na slici 5.3.



Slika 5.2: Procenat varijansi koji je pokriven u odnosu na broj *PCA* komponenti za tri preseka embriona miša



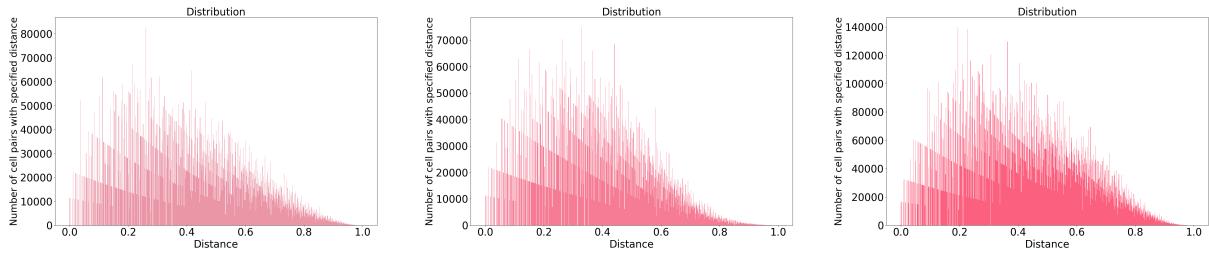
Slika 5.3: Raspodela vrednosti iz matrica susedstva  $D^2$  za tri preseka embriona miša

Kako bi poređenje rastojanja iz matrica susedstva koordinatnog i genskog grafa bilo korektno, potrebno je uraditi njihovu adekvatnu normalizaciju. Prikazane raspodele vrednosti iz matrica susedstva  $D^2$  pokazuju da mali broj parova ćelija ima znatno veća rastojanja vektora genskih ekspresija od ostalih parova (manje od 1%). Ovo povlači raspodele ka desnom kraju, što nas navodi da pre nego što se primeni normalizacija, izbace oni parovi ćelija čija rastojanja nisu u opsegu  $[mean - 3 \cdot std, mean + 3 \cdot std]$ , gde je  $mean$  srednja vrednost, a  $std$  standardna devijacija raspodele vrednosti iz matrice susedstva  $D^2$ . Rastojanja za iste parove ćelija su izbačena i iz matrica  $D^1$ . Nakon ove modifikacije matrica susedstva  $D^1$  i  $D^2$  urađena je min-max normalizacija vrednosti iz ovih matrica. Na slikama 5.4 i 5.5 su prikazane raspodele redukovanih i normalizovanih matrica  $D^1$  i  $D^2$ . Raspodela vrednosti iz redukovane i normalizovane matrice  $D^1$  izgleda slično kao i raspodela vrednosti iz matrice  $D^1$ , dok kod matrice  $D^2$  sada uočavamo ravnomerniju raspodelu vrednosti.

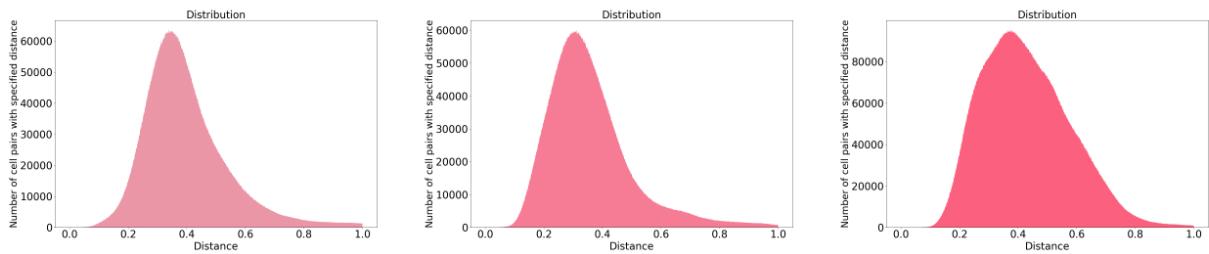
Raspodela vrednosti iz *CDM* matrice, koja predstavlja razliku redukovanih i normalizovanih matrica  $D^1$  i  $D^2$ , predstavljena je na slici 5.6. Prikazane raspodele

## GLAVA 5. REZULTATI

---



Slika 5.4: Raspodela vrednosti iz redukovanih i normalizovanih matrica susedstva  $D^1$  za tri preseka embriona miša



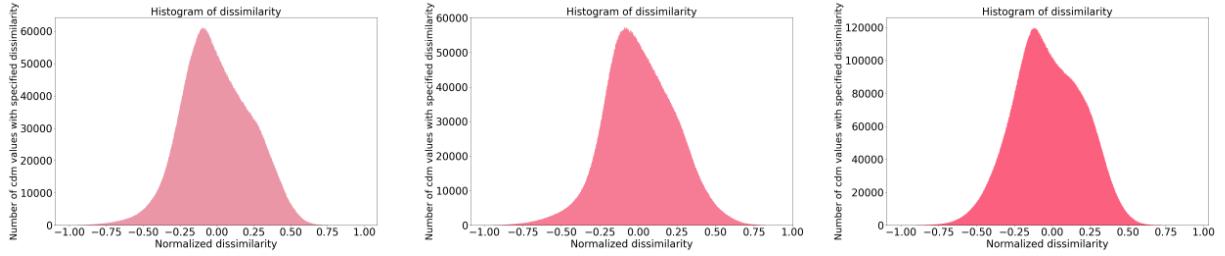
Slika 5.5: Raspodela vrednosti iz redukovanih i normalizovanih matrica susedstva  $D^2$  za tri preseka embriona miša sa 40 PCA komponenti

vrednosti iz *CDM* matrica imaju vrlo slične srednje vrednosti, medijane i moduse za koje važi  $mean > median > mode$  (slika 5.7). Ovo nam ukazuje da su ove raspodele blago desno nagnute, što znači da se rep distribucije proteže prema desno, te da postoje neke više vrednosti koje povlače srednju vrednost u tom smeru. Negativna zakriviljenost za prvi i treći skup podataka ukazuje da su repovi raspodela „lakši”, tj. manje izraženi u odnosu na normalnu raspodelu, dok nam pozitivna zakriviljenost za drugi skup podataka ukazuje da su repovi raspodele „teži”, tj. više izraženi u odnosu na normalnu raspodelu [2]. Dodatno, srednje vrednosti ovih raspodela su bliske nuli što ukazuje da postoji veliki broj ćelija koje su prostorno bliske i imaju sličnu gensku ekspresiju ili suprotno - prostorno su udaljene i imaju različite genske ekspresije.

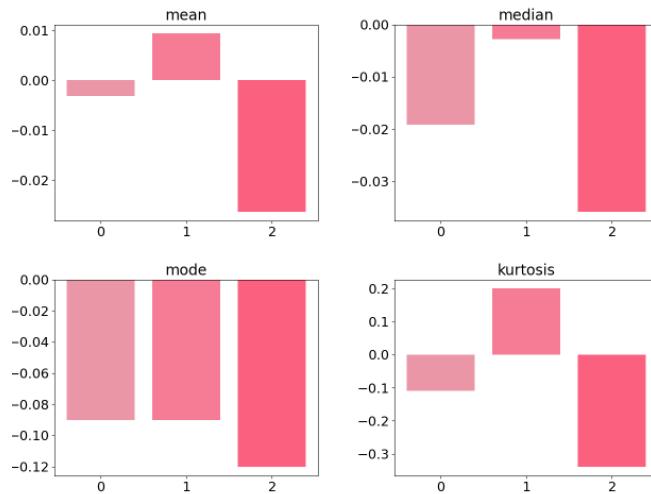
Kako 40 PCA komponenti ne pokriva ni 1% varijanse podataka, prikazaćemo raspodele vrednosti iz matrica  $D^2$  sa brojem PCA komponenti tako da je pokrit veno 80% varijanse podataka (slika 5.8). Kao i za 40 PCA komponenti i ovde su iz matrica susedstva  $D^2$  izbačeni oni parovi ćelija čija rastojanja nisu u opsegu  $[mean - 3 \cdot std, mean + 3 \cdot std]$ . Rastojanja za iste parove ćelija su izbačena i iz matrica  $D^1$ . Zatim je urađena min-max normalizacija vrednosti iz redukovanih matrica susedstva  $D^1$  i  $D^2$ . Na slikama 5.9 i 5.10 su prikazane raspodele redukovanih i

## GLAVA 5. REZULTATI

---



Slika 5.6: Raspodela vrednosti iz  $CDM$  matrice za tri preseka embriona miša sa 40  $PCA$  komponenti



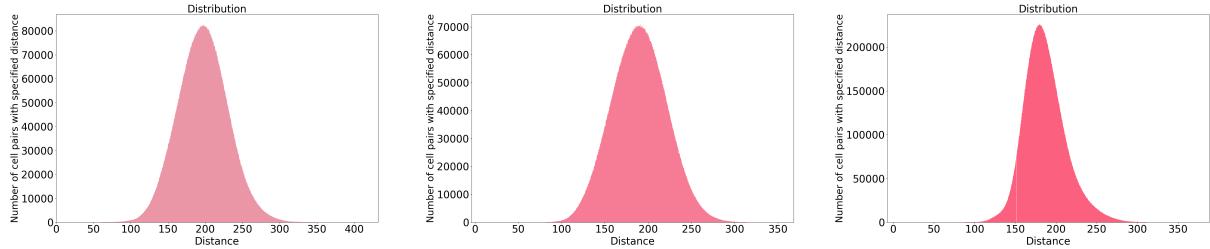
Slika 5.7: Srednja vrednost, medijana, modus i zakrivljenost za tri preseka embriona miša sa 40  $PCA$  komponenti, redom

normalizovanih matrica  $D^1$  i  $D^2$ . Kao što smo imali i kod slučaja sa manjim brojem  $PCA$  komponenti, tako je i ovde raspodela redukovane i normalizovane matrice  $D^1$  slična raspodeli matrice  $D^1$ , dok kod redukovane i normalizovane matrice  $D^2$  sada uočavamo ravnomerniju raspodelu vrednosti.

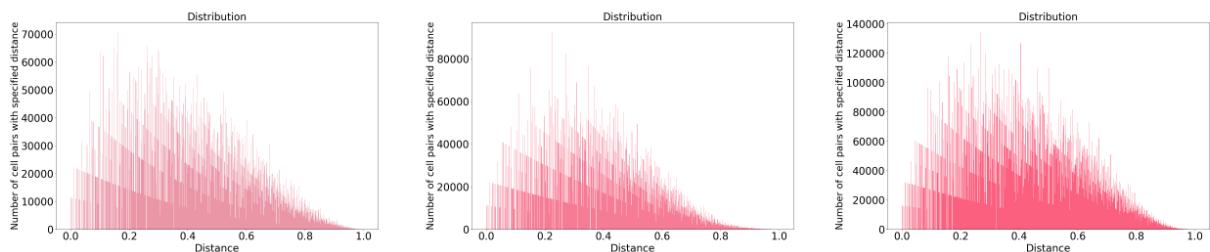
Raspodela vrednosti iz  $CDM$  matrice predstavljena je na slici 5.11. Prikazane raspodele vrednosti iz  $CDM$  matrica imaju vrlo slične srednje vrednosti, medijane i moduse za koje ponovo važi  $mean > median > mode$ , gde su nam sve tri vrednosti bliske (slika 5.12). Ovo nam ukazuje da su ove raspodele skoro simetrične, blago desno nagnute. Negativna zakrivljenost ukazuje da su repovi raspodela „lakši”, tj. manje izraženi u odnosu na normalnu raspodelu. Sada su srednje vrednosti pomerene levo od nule, što ukazuje da ima veliki broj ćelija koje su prostorno bliske, ali se razlikuju po genskim ekspresijama. Ovo je nije u saglasnosti sa rezultatima sa 40

## GLAVA 5. REZULTATI

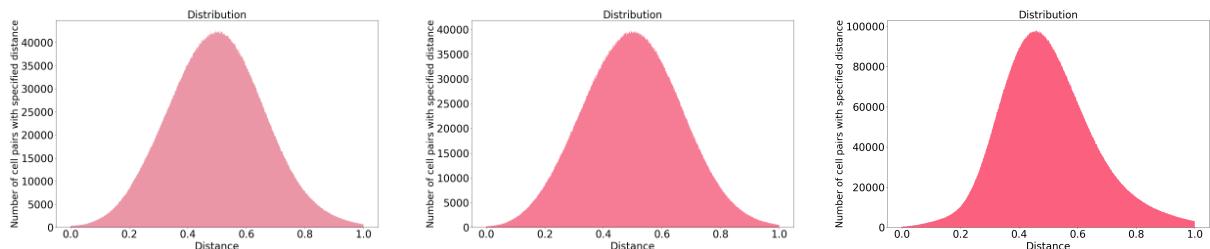
---



Slika 5.8: Raspodela vrednosti iz matrica susedstva  $D^2$  za tri preseka embriona miša sa brojem  $PCA$  komponenti tako da je pokriveno 80% varijanse



Slika 5.9: Raspodela vrednosti iz redukovanih i normalizovanih matrica susedstva  $D^1$  za tri preseka embriona miša sa brojem  $PCA$  komponenti tako da je pokriveno 80% varijanse



Slika 5.10: Raspodela vrednosti iz redukovanih i normalizovanih matrica susedstva  $D^2$  za tri preseka embriona miša sa brojem  $PCA$  komponenti tako da je pokriveno 80% varijanse

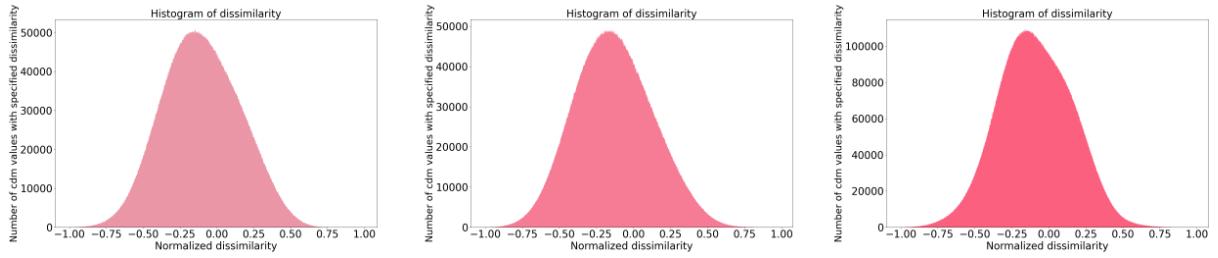
$PCA$  komponenti.

### Problem 2

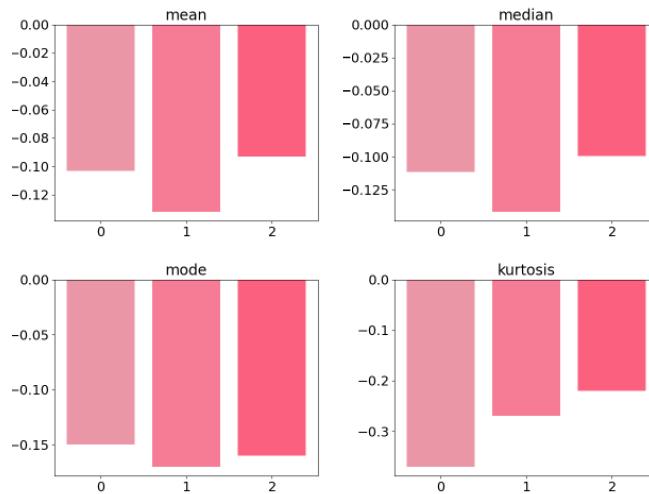
Klasterovanje unije normalizovanih grafova  $G_1$  i  $G_2$  urađeno je pomoću *Leiden* algoritma sa kombinacijom različitih brojeva najbližih suseda za oba grafa sa 40  $PCA$  komponenti. Broj najbližih suseda za koordinatni graf uzima vrednosti iz sku-

## GLAVA 5. REZULTATI

---



Slika 5.11: Raspodela vrednosti iz *CDM* matrice za tri preseka embriona miša sa brojem *PCA* komponenti tako da je pokriveno 80% varijanse



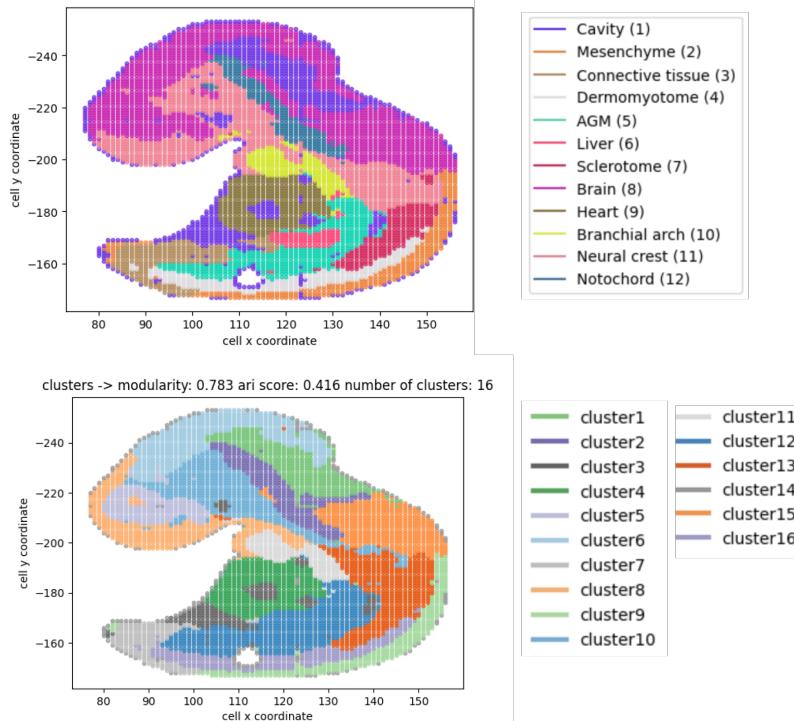
Slika 5.12: Srednja vrednost, medijana, modus i zakriviljenost za tri preseka embriona miša sa brojem *PCA* komponenti tako da je pokriveno 80% varijanse

pa  $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20\}$ , dok broj najbližih suseda za genski graf uzima vrednosti iz skupa  $\{5, 10, 15, 20, 25, 30\}$ . Kako bismo dobili klastere koji u najvećoj meri odgovaraju eksperimentalno utvrđenim tipovima ćelija, kao i optimalan broj najbližih suseda za koordinatni i genski graf, odredićemo *ARI* skor za sve kombinacije.

*ARI* skor za prvi presek embriona miša se kreće u opsegu  $[0.214, 0.416]$ . Klasterovanje sa najmanjim *ARI* skorom od 0.214 ima 20 najbližih suseda za koordinatni graf i 5 najbližih suseda za genski graf. Modularnost za ovo klasterovanje je 0.829, što ukazuje da su su ćelije unutar istih klastera gusto povezane. Dok optimalno klasterovanje za ovaj presek ima *ARI* skor 0.416 i kod njega je broj najbližih suseda za koordinatni graf 8, a za genski 30. Modularnost za ovo klasterovanje ima vrednost 0.783, što opet ukazuje da su ćelije unutar istih klastera gusto povezane. Međutim,

## GLAVA 5. REZULTATI

*ARI* skor nije zadovoljavajući. Dodatno, klasterovanje sa 0 najbližih suseda za koordinatni graf i 30 za genski graf ima *ARI* skor 0.350, koji je približan skoru za optimalno klasterovanje. Ovo ukazuje na to da genska komponenta ima mnogo veći uticaj na tip ćelije, ali i da uticaj koordinatne komponente nije zanemarljiv i da on doprinosi povećanju *ARI* skora.



Slika 5.13: Eksperimentalno utvrđeni tipovi ćelija i tipovi ćelija koji su dobijeni optimalnim klasterovanjem za prvi presek embriona miša

Na slici 5.13 su prikazani eksperimentalno utvrđeni tipovi ćelija i tipovi koje smo dobili klasterovanjem sa najvećim *ARI* skorom. Ako ih uporedimo, videćemo da imamo dosta sličnosti i dobro određenih klastera (vezivno tkivo (3) se u velikoj meri poklapa sa klasterom 11, mezoderm (2) sa klasterom 9, srce (9) sa klasterom 4, itd.), ali i onih koji su rasparčani na manje klastere (mozak (8) koji je rasparčan na klastere 5, 6 i 15), što je u skladu sa dobijenim *ARI* skorom.

Slični rezultati se dobijaju i za preostala dva preseka embriona miša. Detaljne statistike za sve preseke su prikazane u tabeli 5.1. Uočavamo da je kod svih preseka uticaj genske komponente na tip ćelije veći nego uticaj koordinatne, ali i da uticaj koordinatne komponente nije zanemarljiv i da doprinosi poboljšanju klasterovanja, odnosno preklapanja klastera sa eksperimentalno utvrđenim tipovima ćelija. *ARI*

## GLAVA 5. REZULTATI

---

Redni broj preseka	Broj suseda za $G_1$	Broj suseda za $G_2$	$ARI$ skor	Modularnost	Broj klastera	Broj tipova celija
1	8	30	0.416	0.783	16	12
	0	30	0.350	0.829	19	12
	20	5	0.214	0.854	17	12
2	3	15	0.573	0.745	13	13
	0	15	0.507	0.770	14	13
	15	5	0.272	0.776	16	13
3	8	30	0.583	0.757	15	18
	0	30	0.489	0.810	18	18
	15	5	0.269	0.817	15	18

Tabela 5.1: Statistike za optimalno klasterovanje, klasterovanje na osnovu samo genskog grafa i klasterovanje sa najmanjim  $ARI$  skorom za sve preseke embriona miša, redom

Broj PCA komponenti	Minimalni $ARI$ skor	Maksimalni $ARI$ skor
40	0.214	0.416
3400	0.026	0.282

Tabela 5.2: Statistike za 40 i 3400 PCA komponenti za prvi presek embriona miša

skorovi su nešto veći za drugi i treći presek, odnosno kod njih postoji više preklapanja klastera sa eksperimentalno utvrđenim tipovima celija.

Povećanje broja PCA komponenti doprinosi još lošijem  $ARI$  skoru, tako da klasterovanja sa većim broj PCA komponenti ovde neće biti prikazana (tabela 5.2).

## Zaključak

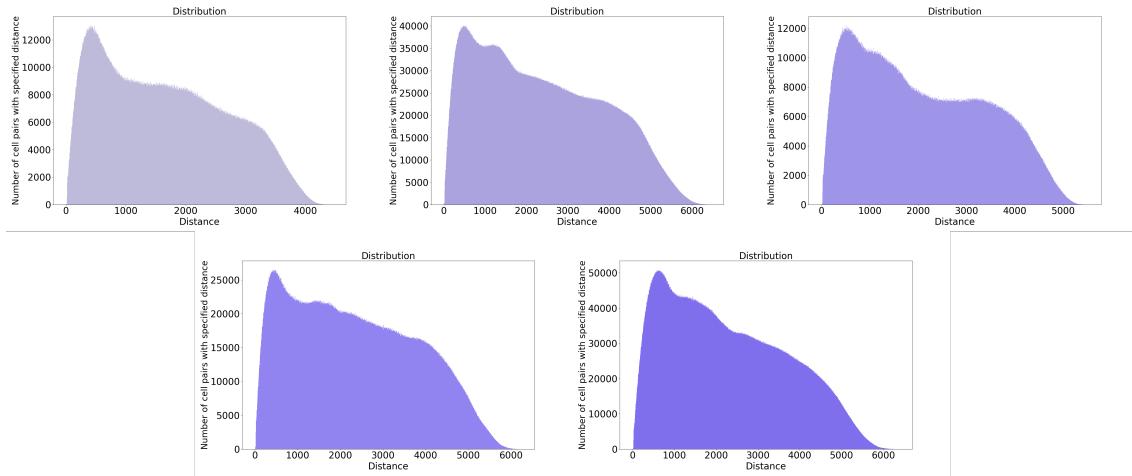
Iz priloženih rezultata uočavamo:

- da se na osnovu srednje vrednosti  $CDM$  matrice sa 40 PCA komponenti i sa brojem PCA komponenti tako da je pokriveno 80% varijanse podataka dobijaju rezultati koji nisu u saglasnosti, i
- da klasteri, dobijeni *Leiden* algoritmom klasterovanja nad unijom koordinatnog i genskog grafa, nemaju zadovoljavajuće poklapanje sa eksperimentalno utvrđenim tipovima celija

## 5.2 Dorzalni srednji mozak miša

### Problem 1

Raspodele vrednosti iz matrica  $D^1$  za pet preseka dorzalnog srednjeg mozga miša su prikazane na slici 5.14.



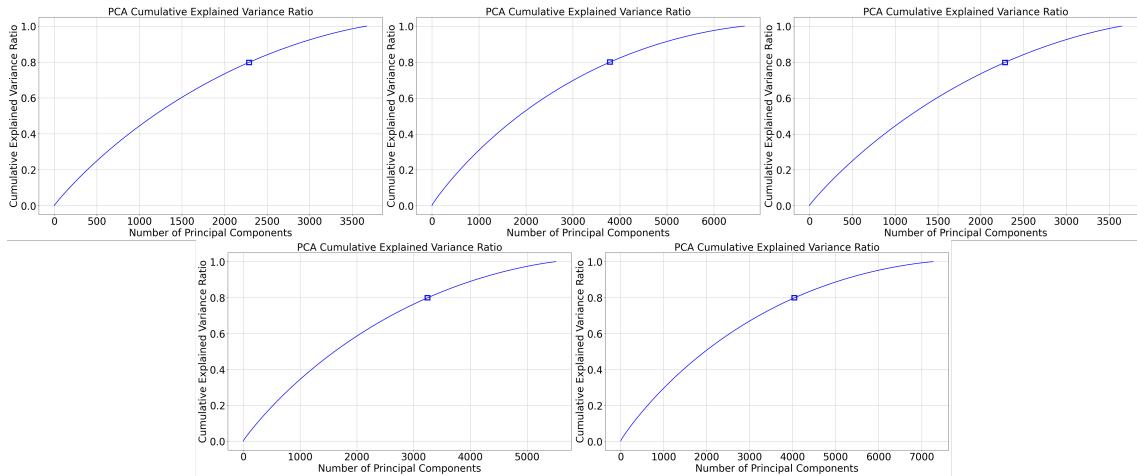
Slika 5.14: Raspodela vrednosti iz matrica susedstva  $D^1$  za 5 preseka dorzalnog srednjeg mozga miša

Pre formiranja matrice susedstva  $D^2$  za ove skupove podataka, iz vektora genskih ekspresija izbačene su vrednosti za gene koji nisu ispoljeni ni u jednoj ćeliji. Zatim je urađena redukcija genskih komponenti pomoću *PCA* metode, kao što smo to uradili i za embrion miša (slika 5.15). Kako je potrebno dosta komponenti da se pokrije 80% varijanse podataka (za prvi i treći presek 2300, drugi 3800, četvrti 3200, i za peti 4000 *PCA* komponenti), prvo su urađene analize sa manjim brojem *PCA* komponenti. Raspodele vrednosti iz matrica  $D^2$  za embrion miša sa 40 *PCA* komponenti su prikazane na slici 5.16.

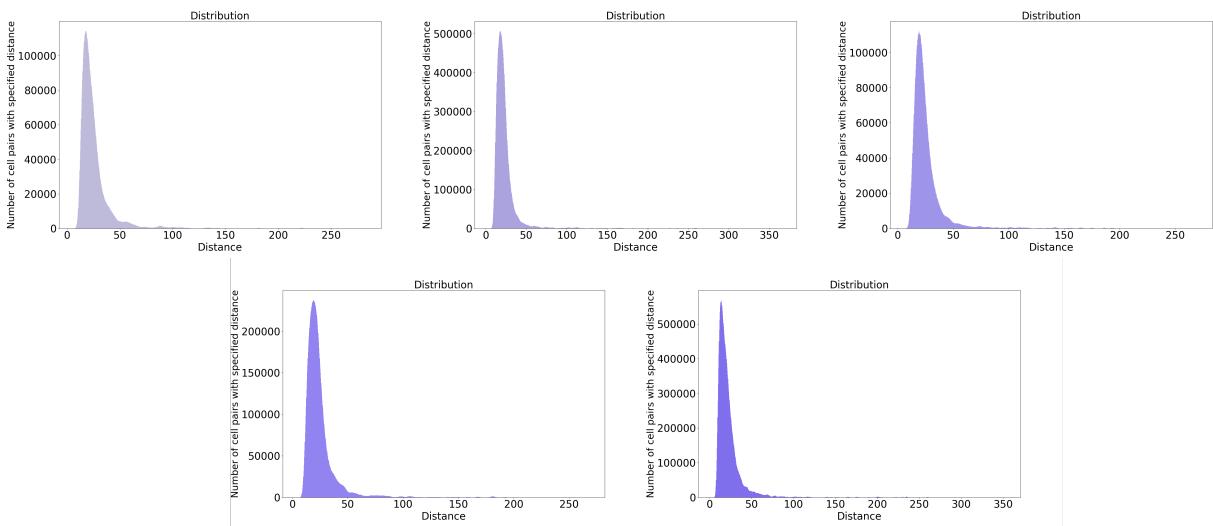
Kako bi poređenje rastojanja iz matrica susedstva koordinatnog i genskog grafa bilo korektno, potrebno je uraditi njihovu adekvatnu normalizaciju. Prikazane raspodele vrednosti iz matrica susedstva  $D^2$  pokazuju da mali broj parova ćelija ima znatno veća rastojanja vektora genskih ekspresija od ostalih parova, kao i kod embriona miša. Ovo ponovo povlači raspodelu ka desnom kraju, i zato se i ovde radi izbacivanje onih parova ćelija čija rastojanja nisu u opsegu  $[mean - 3 \cdot std, mean + 3 \cdot std]$ , gde je  $mean$  srednja vrednost, a  $std$  standardna devijacija raspodele vrednosti iz matrice susedstva  $D^2$ . Rastojanja za iste parove ćelija su izbačena i iz matrica  $D^1$ .

## GLAVA 5. REZULTATI

---



Slika 5.15: Procenat varijansi koji je pokriven u odnosu na broj *PCA* komponenti za 5 preseka razvoja dorzalnog srednjeg mozga miša



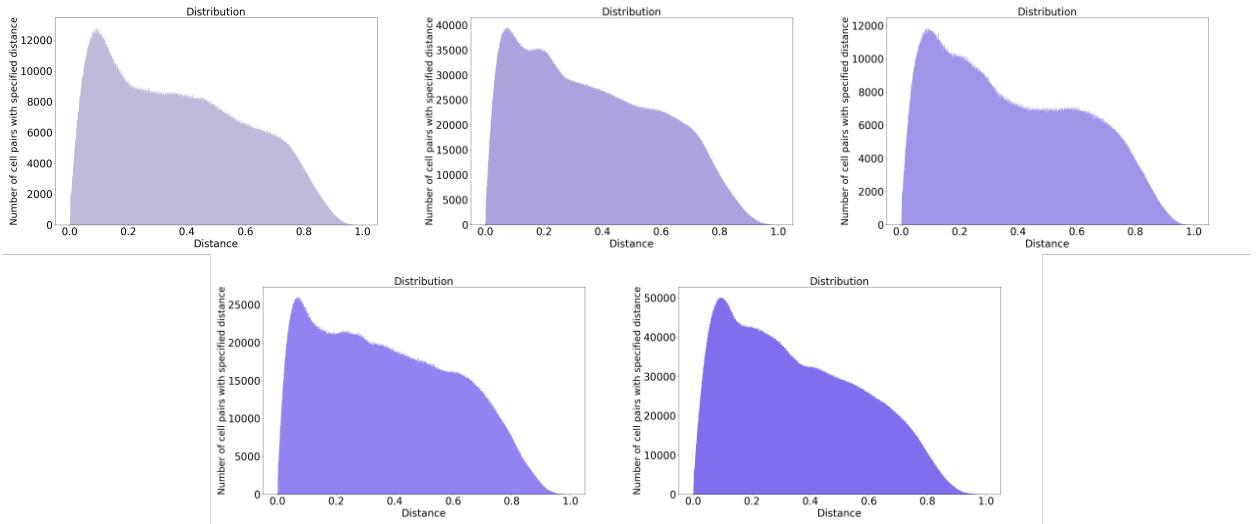
Slika 5.16: Raspodela vrednosti iz matrica susedstva  $D^2$  za 5 preseka dorzalnog srednjeg mozga miša sa 40 *PCA* komponenti

Kao i kod embriona miša, nakon redukcije matrica susedstva urađena je min-max normalizacija vrednosti iz tih matrica. Na slikama 5.17 i 5.18 su prikazane raspodele redukovanih i normalizovanih matrica  $D^1$  i  $D^2$ . Raspodela vrednosti iz redukovane i normalizovane matrice  $D^1$  izgleda slično kao i raspodela vrednosti iz matrice  $D^1$ , dok kod redukovane i normalizovane matrice  $D^2$  sada uočavamo ravnomerniju raspodelu vrednosti.

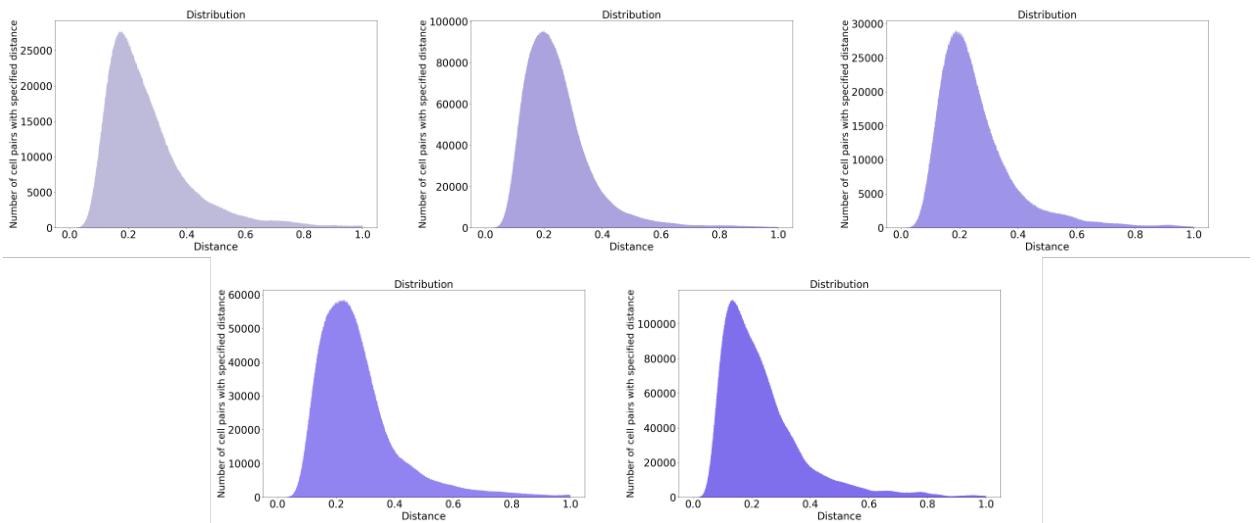
Raspodela vrednosti iz *CDM* matrice predstavljena je na slici 5.19. Sa datim

## GLAVA 5. REZULTATI

---



Slika 5.17: Raspodela vrednosti iz redukovanih i normalizovanih matrica susedstva  $D^1$  za 5 preseka dorzalnog srednjeg mozga miša sa 40 PCA komponenti



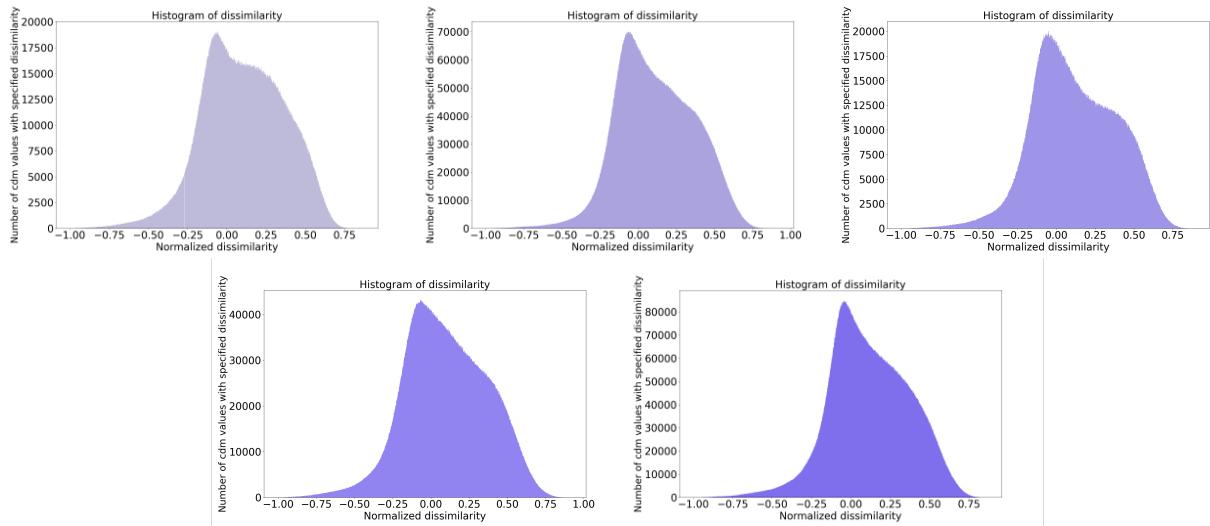
Slika 5.18: Raspodela vrednosti iz redukovanih i normalizovanih matrica susedstva  $D^2$  za 5 preseka dorzalnog srednjeg mozga miša sa 40 PCA komponenti

slika uočavamo da su prva i treća raspodela bimodalne, odnosno da imaju dva jasno izražena maksimuma. Prikazane raspodele vrednosti iz *CDM* matrica imaju vrlo slične srednje vrednosti, medijane i moduse za koje važi  $mean > median > mode$  (slika 5.20). Ovo nam ukazuje da su ove raspodele desno nagnute, što znači da se rep distribucije proteže prema desno, te da postoje neke više vrednosti koje povlače srednju vrednost u tom smeru. Pozitivna zakriviljenost kod poslednjeg preseka

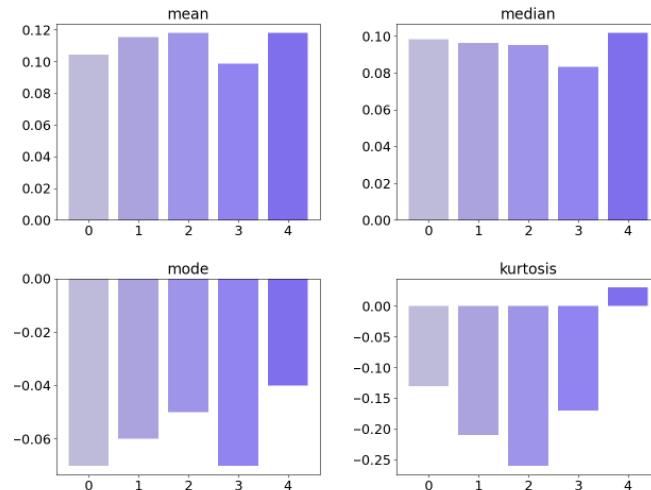
## GLAVA 5. REZULTATI

---

ukazuje da su repovi raspodela „teži”, tj. malo više izraženi u odnosu na normalnu raspodelu. Dok, negativna zakrivljenost kod svih ostalih preseka ukazuje da su repovi raspodela „lakši”, tj. manje izraženi u odnosu na normalnu raspodelu. Dodatno, srednje vrednosti datih raspodela su pomerene desno od nule što ukazuje da postoji veliki broj ćelija koje su prostorno udaljene, ali su im genske ekspresije slične.



Slika 5.19: Raspodela vrednosti iz *CDM* matrice za 5 preseka dorzalnog srednjeg mozga sa 40 *PCA* komponenti

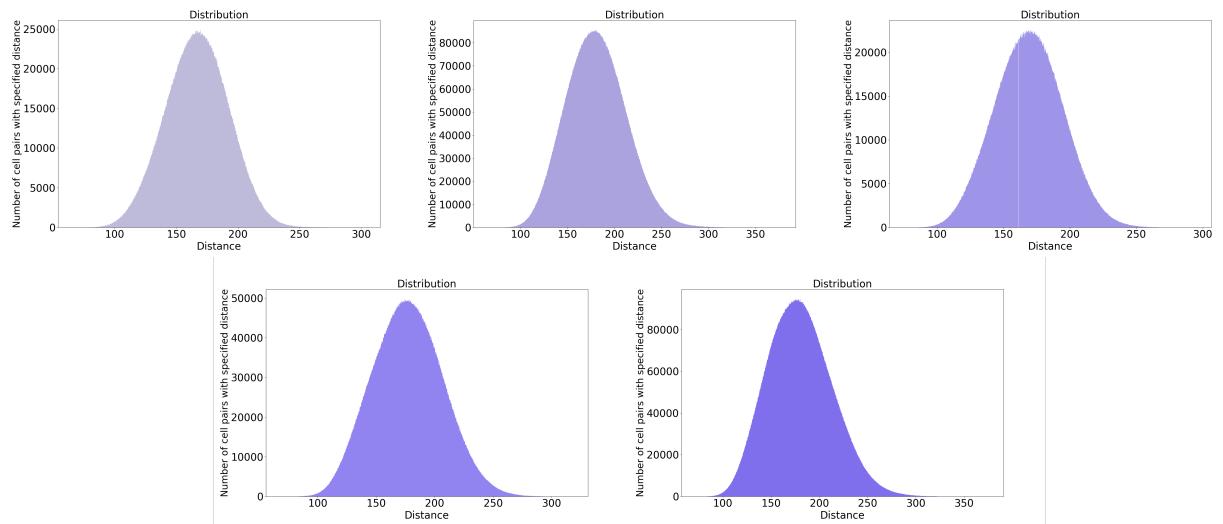


Slika 5.20: Srednja vrednost, medijana, modus i zakrivljenost za 5 preseka dorzalnog srednjeg mozga sa 40 *PCA* komponenti

## GLAVA 5. REZULTATI

---

Kako 40 *PCA* komponenti ne pokriva ni 1% varijanse podataka, prikazaćemo raspodele vrednosti iz matrica  $D^2$  za dorzalni srednji mozak miša sa brojem *PCA* komponenti tako da je pokriveno 80% varijanse podataka, kao što smo to uradili i kod embriona miša (slika 5.21). Kao i za 40 *PCA* komponenti i ovde su iz matrica susedstva  $D^2$  izbačeni parovi ćelija čija rastojanja nisu u opsegu  $[mean - 3 \cdot std, mean + 3 \cdot std]$ . Rastojanja za iste parove ćelija su izbačena i iz matrica  $D^1$ . Zatim je urađena min-max normalizacija vrednosti iz redukovanih matrica susedstva  $D^1$  i  $D^2$ . Na slikama 5.22 i 5.23 su prikazane raspodele redukovanih i normalizovanih matrica  $D^1$  i  $D^2$ . Kao što smo imali i kod slučaja sa manjim brojem *PCA* komponenti, tako je i ovde raspodela redukovane i normalizovane matrice  $D^1$  slična raspodeli matrice  $D^1$ , dok kod redukovane i normalizovane matrice  $D^2$  sada uočavamo ravnomerniju raspodelu vrednosti.

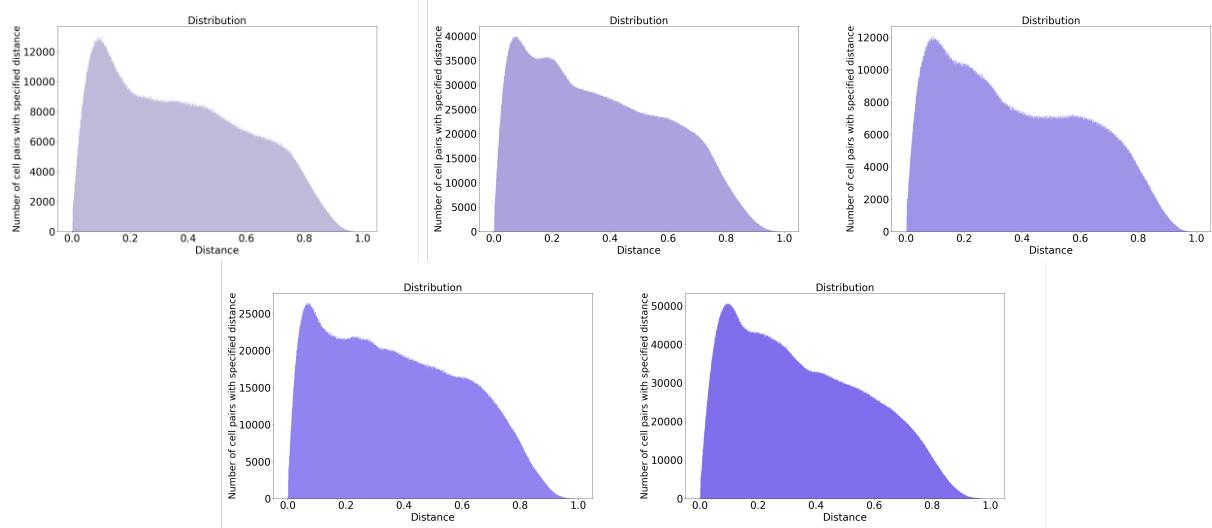


Slika 5.21: Raspodela vrednosti iz matrica susedstva  $D^2$  za 5 preseka dorzalnog srednjeg mozga miša sa brojem *PCA* komponenti tako da je pokriveno 80% varijanse

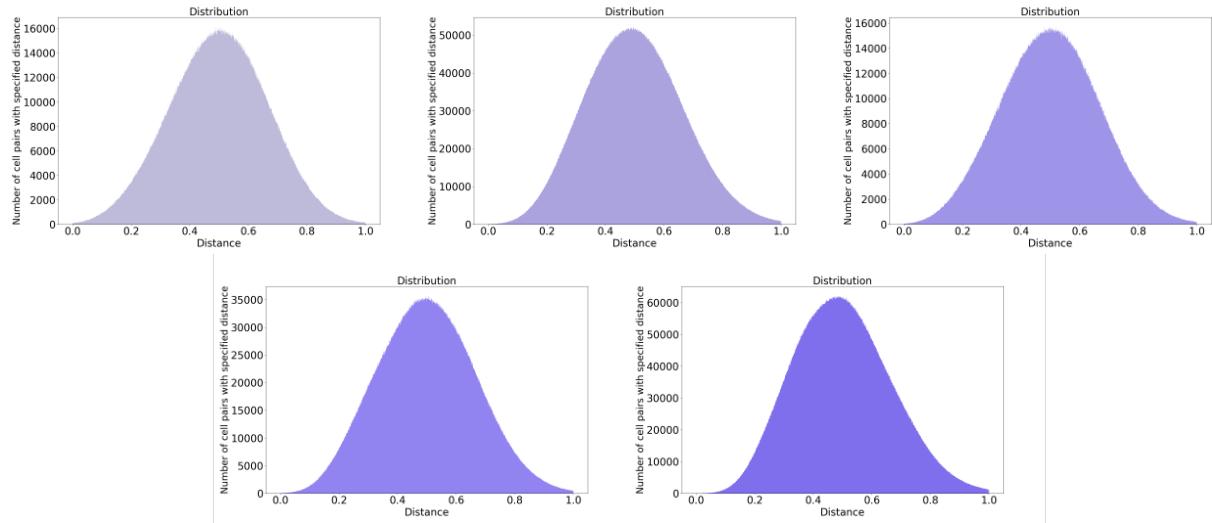
Raspodela vrednosti iz *CDM* matrice predstavljena je na slici 5.24. Uočavamo da su i ovde vrlo slične srednje vrednosti, medijane i modusi za koje ponovo važi  $mean > median > mode$  (slika 5.25), gde su srednje vrednosti i medijane bliske, a zakriviljenost negativna. Stoga, i ovde su raspodele desno nagnute, a njihovi repovi „lakši”, tj. manje izraženi u odnosu na normalnu raspodelu, kao što je bilo i kod raspodela sa 40 *PCA* komponenti. Kod prvog preseka sve tri vrednosti su vrlo bliske, što ukazuje da je ova raspodela skoro simetrična, blago desno nagnuta. Ovde su srednje vrednosti pomerene levo od nule, što ukazuje da postoji veliki broj ćelija koje

## GLAVA 5. REZULTATI

---



Slika 5.22: Raspodela vrednosti iz redukovanih i normalizovanih matrica susedstva  $D^1$  za 5 preseka dorzalnog srednjeg mozga sa brojem  $PCA$  komponenti tako da je pokriveno 80% varijanse

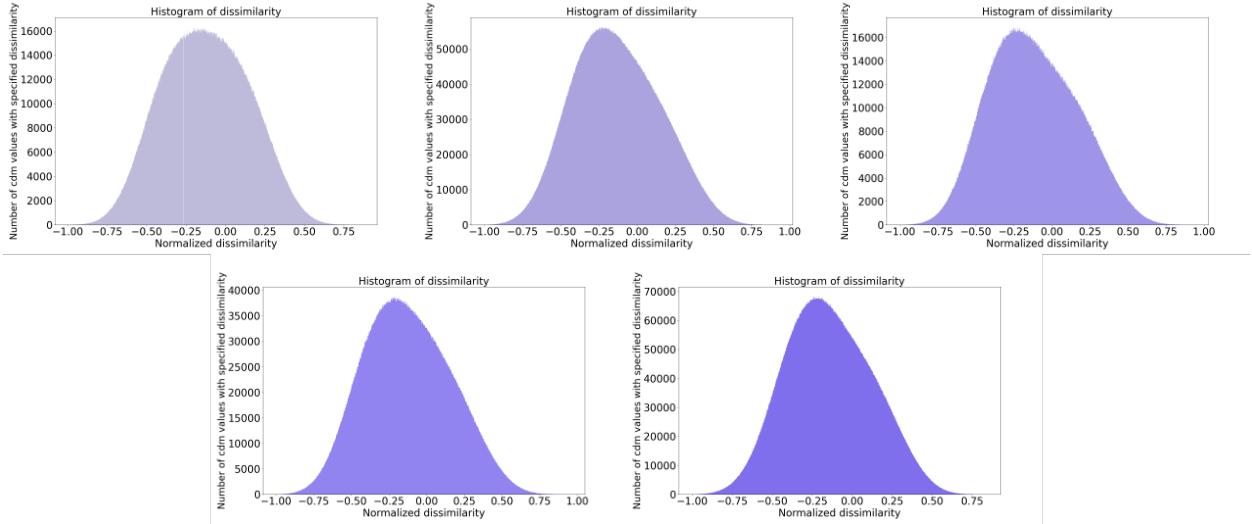


Slika 5.23: Raspodela vrednosti iz redukovanih i normalizovanih matrica susedstva  $D^2$  za 5 preseka dorzalnog srednjeg mozga sa brojem  $PCA$  komponenti tako da je pokriveno 80% varijanse

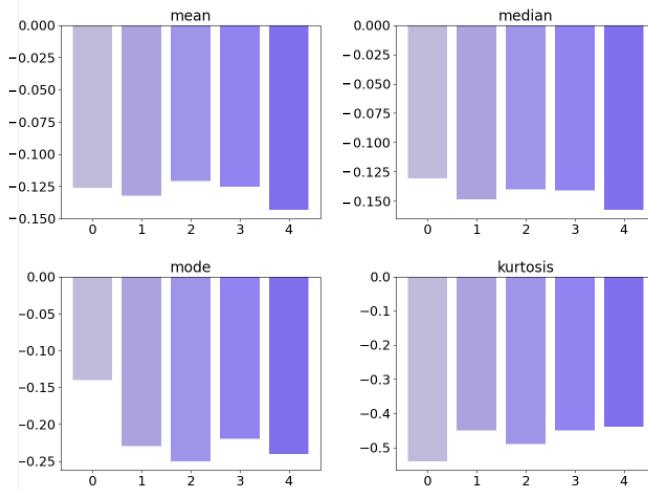
su prostorno bliske, ali se razlikuju po genskim ekspresijama. Ovo nije u saglasnosti sa rezultatima sa 40  $PCA$  komponenti.

## GLAVA 5. REZULTATI

---



Slika 5.24: Raspodela vrednosti iz *CDM* matrice za 5 preseka dorzalnog srednjeg mozga miša sa brojem *PCA* komponenti tako da je pokriveno 80% varijanse



Slika 5.25: Srednja vrednost, medijana, modus i zakriviljenost za 5 preseka dorzalnog srednjeg mozga miša sa brojem *PCA* komponenti tako da je pokriveno 80% varijanse

### Problem 2

Klasterovanje unije normalizovanih grafova  $G_1$  i  $G_2$  urađeno je pomoću *Leiden* algoritma sa kombinacijom različitih brojeva najbližih suseda za oba grafa sa 40 *PCA* komponenti. Broj najbližih suseda za koordinatni graf uzima vrednosti iz skupa  $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20\}$ , dok broj najbližih suseda za genski graf uzima vrednosti iz skupa  $\{5, 10, 15, 20, 25, 30\}$ . Kako bismo dobili klastere koji u najvećoj

## *GLAVA 5. REZULTATI*

---

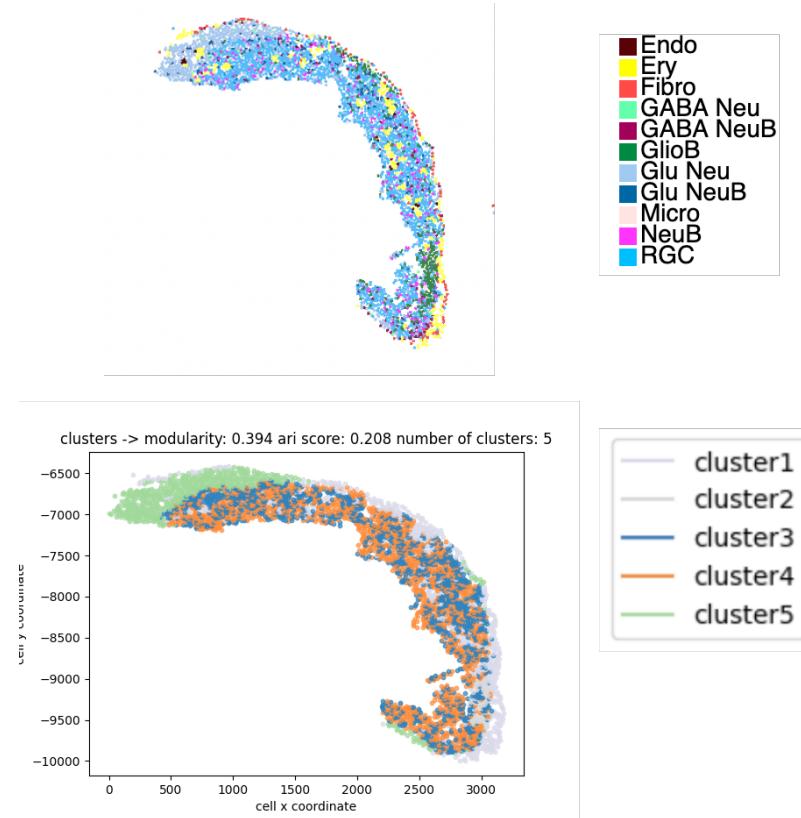
meri odgovaraju eksperimentalno utvrđenim tipovima ćelija, kao i optimalan broj najbližih suseda za koordinatni i genski graf, i ovde ćemo odrediti *ARI* skor za sve kombinacije.

*ARI* skor za prvi presek dorzalnog srednjeg mozga miša se kreće u opsegu [0.058, 0.208]. Klasterovanje sa najmanjim *ARI* skorom od 0.058 ima 9 najbližih suseda za koordinatni graf i 5 najbližih suseda za genski graf. Modularnost za ovo klasterovanje je 0.621, što ukazuje da su su ćelije unutar istih klastera nisu dovoljno gusto povezane. Dok optimalno klasterovanje za ovaj presek ima *ARI* skor 0.208 i kod njega je broj najbližih suseda za koordinatni graf 5, a za genski 25. Modularnost za ovo klasterovanje ima vrednost 0.394, što opet ukazuje da ćelije unutar istih klastera nisu gusto povezane. Dakle, osim što *ARI* skor nije zadovoljavajući, ni modularnost nije zadovoljavajuća. Dodatno, klasterovanje sa 0 najbližih suseda za koordinatni graf i 25 za genski graf ima *ARI* skor 0.172, koji je približan skoru za optimalno klasterovanje. Ovo ukazuje na to da genska komponenta ima mnogo veći uticaj na tip ćelije, ali i da uticaj koordinatne komponente nije zanemarljiv i da on doprinosi povećanju *ARI* skora, kao što smo imali i kod embriona miša. Kod ovog preseka postoje klasterovanja koja imaju vrednosti *ARI* skora koje su veće od 0.2, što je vrlo blizu optimalnog *ARI* skora. To su klasterovanja (4, 15), (7, 20), (8, 30), gde nam je prvi element iz para broj najbližih suseda za koordinatni graf, a drugi broj najbližih suseda za genski graf. Uočavamo da je kod svih klasterovanja uticaj genske komponente veći, ali i da je uticaj koordinatne komponente značajan, kao što smo imali i kod optimalnog klasterovanja.

Na slici 5.26 su prikazani eksperimentalno utvrđeni tipovi ćelija i tipovi koji su dobijeni klasterovanjem sa najvećim *ARI* skorom. Ako ih uporedimo, videćemo da postoje sličnosti (GluNeu ima poklapanja sa klastrom 5, kao i RGC i klastrom 4), ali i dosta klastera koji se ne poklapaju (postoji 11 tipova ćelija, a 5 klastera, što znači da se više tipova ćelija nalazi u istom klastru), što je u skladu sa dobijenim *ARI* skorom.

Slični rezultati se dobijaju i za preostala četiri preseka dorzalnog srednjeg mozga miša. Detaljne statistike za sve preseke su prikazane u tabeli 5.3. Uočavamo da je kod svih preseka uticaj genske komponente na tip ćelije veći nego uticaj koordinatne, ali i da uticaj koordinatne komponente nije zanemarljiv i da doprinosi poboljšanju *ARI* skora. Kod prvog preseka, ovo poboljšanje je znatno manje nego kod ostalih preseka. Modularnosti za sva klasterovanja, osim za klasterovanja na osnovu samo genskog grafa, nisu zadovoljavajuće. *ARI* skorovi su nešto veći za preseke tri i četiri, odnosno

## GLAVA 5. REZULTATI



Slika 5.26: Eksperimentalno utvrđeni tipovi ćelija i klasteri koji su dobijeni optimalnim klasterovanjem za prvi presek dorzalnog srednjeg mozga miša sa 40 *PCA* komponenti

za njih postoji više preklapanja klastera sa eksperimentalno utvrđenim tipovima ćelija, ali nedovoljno.

Povećanje broja *PCA* komponenti doprinosi još lošijem *ARI* skoru kao i kod embriona miša. Stoga, klasterovanja sa većim broj *PCA* komponenti i ovde neće biti prikazana (tabela 5.4).

### Zaključak

Iz priloženih rezultata uočavamo:

- da se na osnovu srednje vrednosti *CDM* matrice sa 40 *PCA* komponenti i sa brojem *PCA* komponenti tako da je pokriveno 80% varijanse podataka dobijaju rezultati koji nisu u saglasnosti, i

## GLAVA 5. REZULTATI

---

Redni broj preseka	Broj suseda za $G_1$	Broj suseda za $G_2$	$ARI$ skor	Modularnost	Broj klastera	Broj tipova ćelija
1	5	25	0.208	0.394	5	11
	0	25	0.170	0.430	5	11
	9	5	0.058	0.621	20	11
2	5	10	0.396	0.560	6	11
	0	10	0.202	0.644	7	11
	20	5	0.027	0.746	16	11
3	8	30	0.495	0.490	3	11
	0	30	0.307	0.543	4	11
	20	5	0.067	0.716	16	11
4	10	25	0.479	0.518	5	11
	0	25	0.276	0.596	7	11
	20	5	0.040	0.735	20	11
5	9	20	0.367	0.519	6	11
	0	20	0.191	0.630	9	11
	20	5	0.053	0.743	18	11

Tabela 5.3: Statistike za optimalno klasterovanje, klasterovanje na osnovu samo genskog grafa i klasterovanje sa najmanjim  $ARI$  skorom za sve preseke dorzalnog srednjeg mozga miša, redom

Broj PCA komponenti	Minimalni $ARI$ skor	Maksimalni $ARI$ skor
40	0.058	0.208
2300	0	0.04

Tabela 5.4: Statistike za 40 i 2300 PCA komponenti za prvi presek dorzalnog srednjeg mozga miša

- da klasteri nemaju zadovoljavajuće poklapanje sa eksperimentalno utvrđenim tipovima ćelija, čak imaju i lošije rezultate od embriona miša

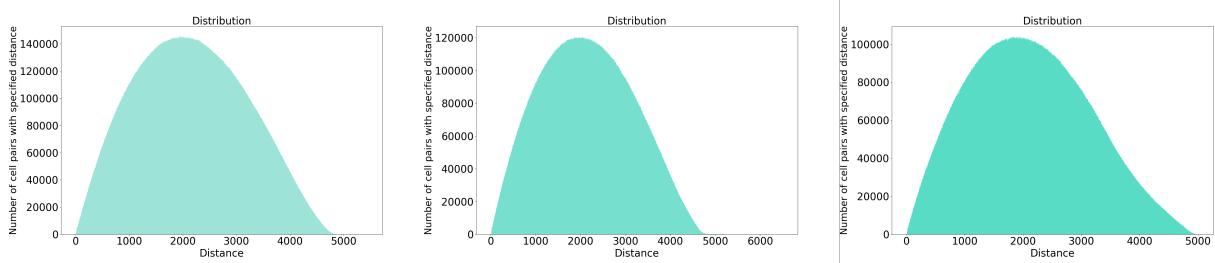
### 5.3 Medula i korteks ljudskog bubrega i bubreg miša

#### Problem 1

Raspodele vrednosti iz matrica  $D^1$  za ove skupove podataka su prikazane na slici 5.27.

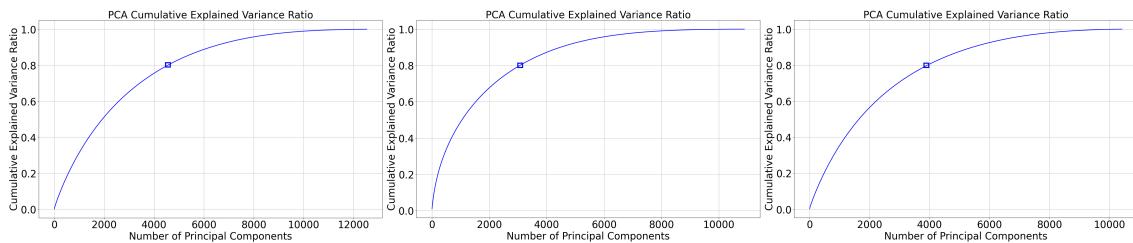
## GLAVA 5. REZULTATI

---

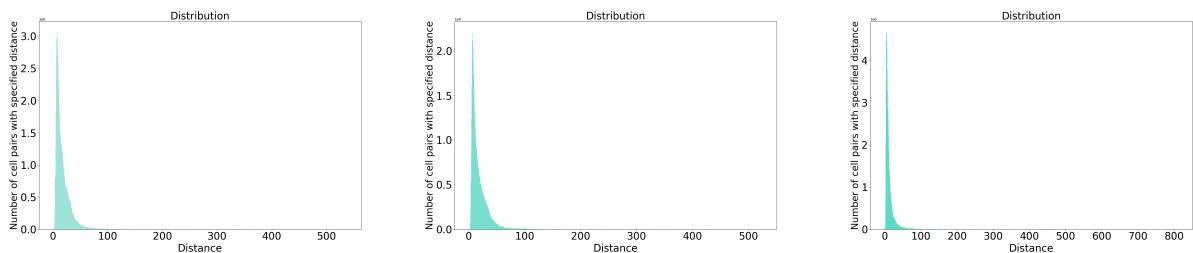


Slika 5.27: Raspodela vrednosti iz matrica susedstva  $D^1$  za medulu i korteks ljudskog bubrega i bubreg miša, redom

Pre formiranja matrice susedstva  $D^2$  za ove skupove podataka, iz vektora genskih ekspresija izbačene su vrednosti za gene koji nisu ispoljeni ni u jednoj ćeliji. Zatim je urađena redukcija genskih komponenti pomoću *PCA* metode, kao što smo to uradili i za embrion i dorzalni srednji mozak miša (slika 5.28). Kako je i ovde je potreban veliki broj *PCA* komponenti da bismo pokrili 80% varijanse podataka (4500 za prvi, 3000 za drugi i 4000 za treći skup podataka), prvo su urađene analize sa manjim brojem *PCA* komponenti. Raspodele vrednosti iz matrica  $D^2$  za ove skupove podataka sa 40 *PCA* komponenti su prikazane na slici 5.29.



Slika 5.28: Procenat varijansi koji je pokriven u odnosu na broj *PCA* komponenti za medulu i korteks ljudskog bubrega i bubreg miša, redom

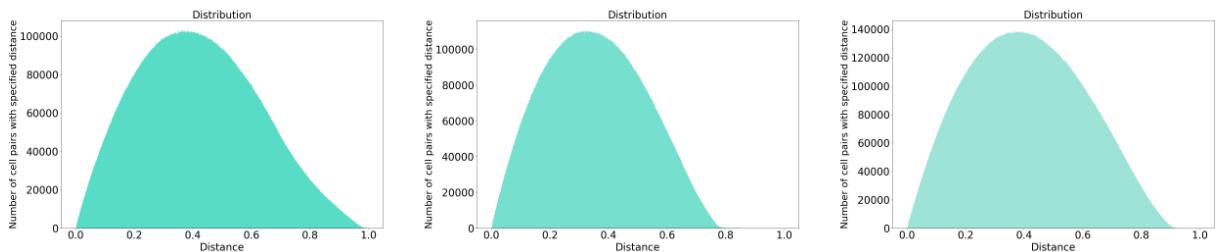


Slika 5.29: Raspodela vrednosti iz matrica susedstva  $D^2$  za medulu i korteks ljudskog bubrega i bubreg miša sa 40 *PCA* komponenti, redom

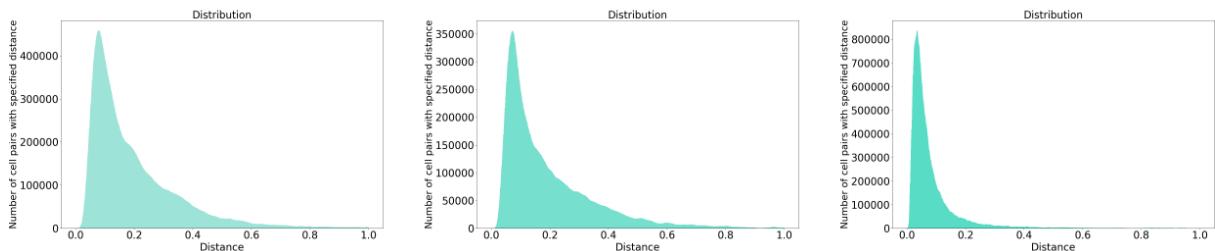
## GLAVA 5. REZULTATI

---

Kod prikazanih raspodela vrednosti iz matrica susedstva  $D^2$  možemo uočiti da postoje parovi ćelija koji imaju znatno veća rastojanja vektora genskih ekspresija od ostalih parova, kao što je bio slučaj kod embriona i dorzalnog srednjeg mozga miša. Analogno, i ovde to dovodi do toga da su raspodele pomerene ka desno, što nas navodi da pre nego što se primeni normalizacija, izbacuje oni parovi ćelija čija rastojanja nisu u opsegu  $[mean - 3 \cdot std, mean + 3 \cdot std]$ , gde je  $mean$  srednja vrednost, a  $std$  standardna devijacija raspodele vrednosti iz matrice susedstva  $D^2$ . Rastojanja za iste parove ćelija su izbačena i iz matrica  $D^1$ , kao i u prethodnim analizama. Zatim je urađena min-max normalizacija vrednosti iz redukovanih matrica susedstva  $D^1$  i  $D^2$ . Raspodele redukovanih i normalizovanih matrica  $D^1$  i  $D^2$  su prikazane na slikama 5.32 i 5.31. Kao i kod embriona i dorzalnog srednjeg mozga miša i ovde raspodela vrednosti iz redukovane i normalizovane matrice  $D^1$  izgleda slično kao i raspodela vrednosti iz matrice  $D^1$ , dok kod matrice  $D^2$  sada imamo ravnomerniju raspodelu vrednosti.



Slika 5.30: Raspodela vrednosti iz redukovanih i normalizovanih modifikovanih matrica susedstva  $D^1$  za medulu i korteks ljudskog bubrega i bubreg miša sa 40 PCA komponenti, redom

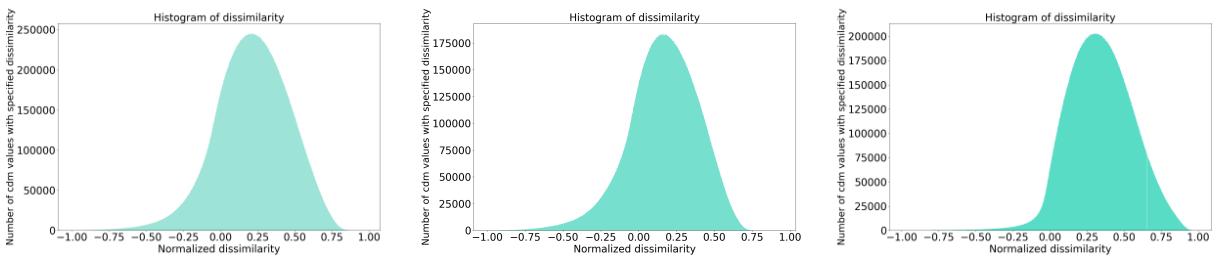


Slika 5.31: Raspodela vrednosti iz redukovanih i normalizovanih modifikovanih matrica susedstva  $D^2$  za medulu i korteks ljudskog bubrega i bubreg miša sa 40 PCA komponenti, redom

## GLAVA 5. REZULTATI

---

Raspodela vrednosti iz  $CDM$  matrice predstavljena je na slici 5.32. Prikazane raspodele vrednosti iz  $CDM$  matrica imaju vrlo slične srednje vrednosti, medijane i moduse. Kod medule ljudskog bubrega važi  $mean < median < mode$ , gde su sve tri vrednosti vrlo bliske što ukazuje da je ova raspodela skoro simetrična, blago levo nagnuta. Kod korteksa ljudskog bubrega važi  $mean < mode < median$ , gde su opet sve tri vrednosti vrlo bliske što ukazuje da je ova raspodela skoro simetrična, blago levo nagnuta. Dok kod bubrega miša važi  $mean > median > mode$ , gde su sve tri vrednosti opet vrlo bliske, što nam ukazuje da je ova raspodela skoro simetrična, blago desno nagnuta. Pozitivna zakrivljenost ukazuje da su repovi raspodele „teži”, tj. više izraženi u odnosu na normalnu raspodelu (slika 5.33). Dodatno, srednje vrednosti datih raspodela su pomerene desno od nule, što ukazuje da ima veliki broj ćelija koje su prostorno udaljene, ali su im genske ekspresije slične.



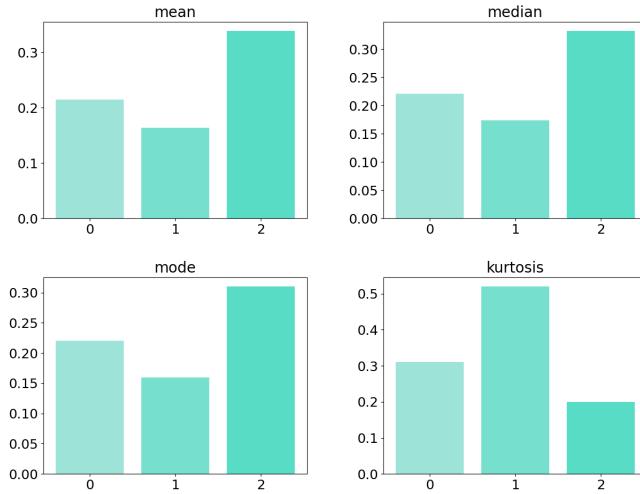
Slika 5.32: Raspodela vrednosti iz  $CDM$  matrice za medulu i korteks ljudskog bubrega i bubreg miša sa 40  $PCA$  komponenti, redom

Kako 40  $PCA$  komponenti ne pokriva ni 1% varijanse podataka, prikazaćemo raspodele vrednosti iz matrica  $D^2$  sa brojem  $PCA$  komponenti tako da je pokriveno 80% varijanse podataka, kao što smo to uradili i kod embriona i dorzlanog srednjeg mozga miša (slika 5.34).

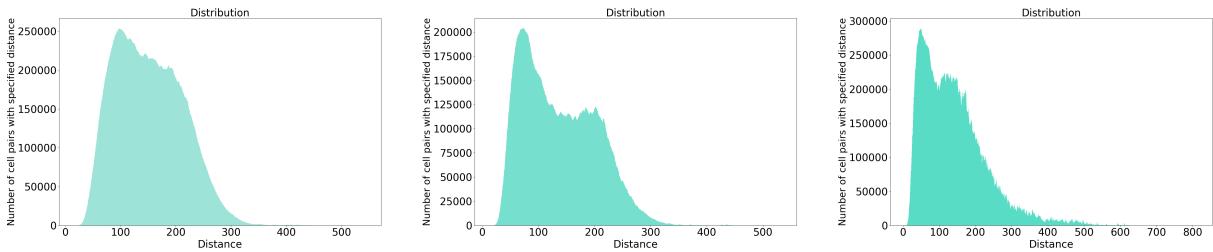
Kao i za 40  $PCA$  komponenti i ovde su iz matrica susedstva  $D^2$  izbačeni oni parovi ćelija čija rastojanja nisu u opsegu  $[mean - 3 \cdot std, mean + 3 \cdot std]$ . Rastojanja za iste parove ćelija su izbačena i iz matrica  $D^1$ . Zatim je urađena min-max normalizacija vrednosti iz redukovanih matrica susedstva  $D^1$  i  $D^2$ . Na slikama 5.35 i 5.36 su prikazane raspodele redukovanih i normalizovanih matrica  $D^1$  i  $D^2$ . Kao što smo imali i kod slučaja sa manjim brojem  $PCA$  komponenti, tako je i ovde raspodela redukovane i normalizovane matrice  $D^1$  slična raspodeli matrice  $D^1$ , dok kod redukovane i normalizovane matrice  $D^2$  sada imamo ravnomerniju raspodelu vrednosti.

## GLAVA 5. REZULTATI

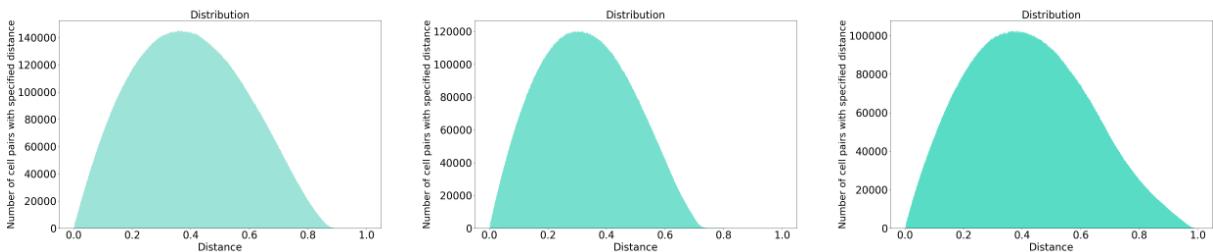
---



Slika 5.33: Srednja vrednost, medijana, modus, i zakriviljenost za medulu i korteks ljudskog bubrega i bubreg miša sa 40 *PCA* komponenti, redom



Slika 5.34: Raspodela vrednosti iz matrica susedstva  $D^2$  za medulu i korteks ljudskog bubrega i bubreg miša sa brojem *PCA* komponenti tako da je pokriveno 80% varijanse podataka, redom

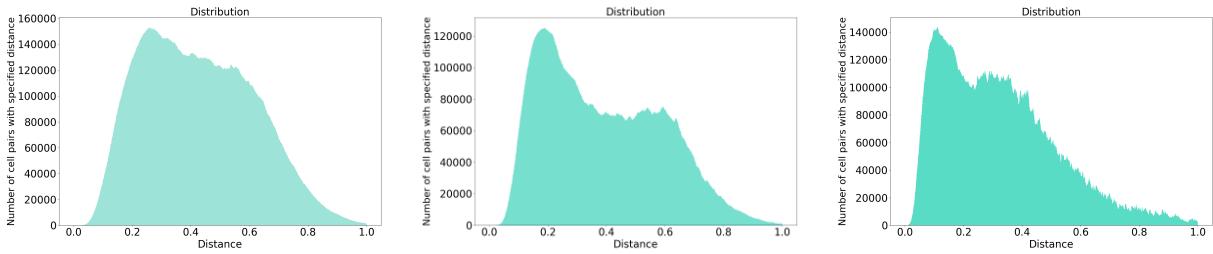


Slika 5.35: Raspodela vrednosti redukovanih i normalizovanih matrica susedstva  $D^1$  za medulu i korteks ljudskog bubrega i bubreg miša sa brojem *PCA* komponenti tako da je pokriveno 80% varijanse podataka, redom

Raspodela vrednosti iz *CDM* matrice predstavljena je na slici 5.37. Prikazane raspodele vrednosti iz *CDM* matrica imaju opet vrlo slične srednje vrednosti, me-

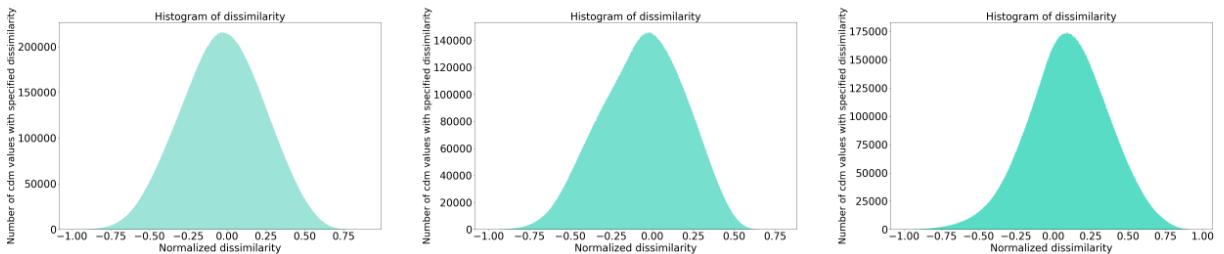
## GLAVA 5. REZULTATI

---



Slika 5.36: Raspodela vrednosti iz redukovanih i normalizovanih matrica susedstva  $D^2$  za medulu i korteks ljudskog bubrega i bubreg miša sa brojem  $PCA$  komponenti tako da je pokriveno 80% varijanse podataka, redom

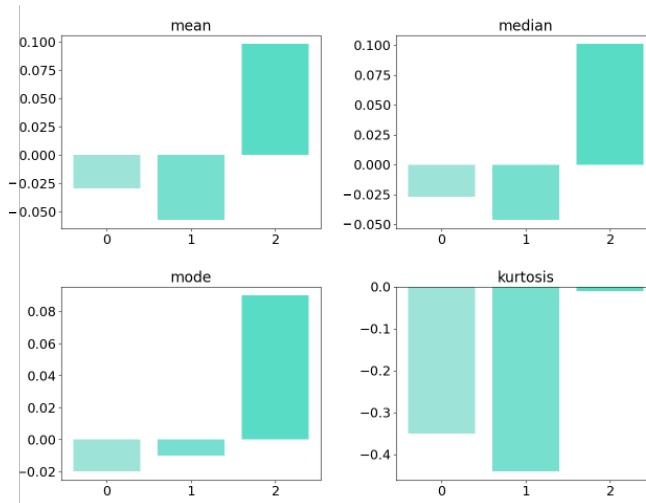
dijane i moduse. Imamo da je kod medule i korteksa ljudskog bubrega  $mode > median > mean$ , gde su nam srednja vrednost i medijana vrlo bliske. Ovo ukazuje da su ove raspodele blago nagnute u levo, odnosno da postoje neke više vrednosti koje povlače srednju vrednost u tom smeru. Dok kod bubrega miša važi da je  $median > mean > mode$ , gde su sve tri vrednosti vrlo bliske, što ukazuje da je ova raspodela skoro simetrična, blago desno nagnuta. Negativna zakrivljenost ukazuje da su repovi raspodela „lakši”, tj. manje izraženi u odnosu na normalnu raspodelu (slika 5.38). Sada su srednje vrednosti za medulu i korteks ljudskog bubrega bliske nuli, što ukazuje da postoji veliki broj ćelija koje su prostorno bliske i imaju sličnu gensku ekspresiju ili suprotno - prostorno su udaljene i imaju različite genske ekspresije. Dok je srednja vrednost kod bubrega miša pomeren desno do nule, što ukazuje da ima veliki broj ćelija koje su prostorno udaljene, ali su im genske ekspresije slične. Za medulu i korteks ljudskog bubrega ovo nije u saglasnosti, dok za bubreg miša ovo jeste u saglasnosti sa razultatima sa 40  $PCA$  komponenti.



Slika 5.37: Raspodela vrednosti iz  $CDM$  matrice za medulu i korteks ljudskog bubrega i bubreg miša sa brojem  $PCA$  komponenti tako da je pokriveno 80% varijanse podataka, redom

## GLAVA 5. REZULTATI

---



Slika 5.38: Srednja vrednost, medijana, modus, i zakriviljenost za medulu i korteks ljudskog bubrega i bubreg miša sa brojem  $PCA$  komponenti tako da je pokriveno 80% varijanse podataka, redom

### Problem 2

Klasterovanje unije normalizovanih grafova  $G_1$  i  $G_2$  urađeno je pomoću *Leiden* algoritma sa kombinacijom različitih brojeva najbližih suseda za oba grafa sa 40  $PCA$  komponenti. Broj najbližih suseda za koordinatni graf uzima vrednosti iz skupa  $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20\}$ , dok broj najbližih suseda za genski graf uzima vrednosti iz skupa  $\{5, 10, 15, 20, 25, 30\}$ . Kako bismo dobili klastera koji u najvećoj meri odgovaraju realnim tipovima ćelija, kao i optimalan broj najbližih suseda za koordinatni i genski graf, i ovde smo odredili  $ARI$  skor za sve kombinacije. Međutim, ovde su vrednosti  $ARI$  skora vrlo male (manje od 0.06), pa samim tim nećemo moći da upotrebimo dati skor za upoređivanje uticaja genskih i kordinantnih komponenti na tip ćelije, kao što smo to uradili kod embriona i dorzalnog srednjeg mozga miša. Slični rezultati se dobijaju i za preostale skupove podataka. Povećanje broja  $PCA$  komponenti ne doprinosi poboljšanju  $ARI$  skora, tako da klasterovanja sa većim broj  $PCA$  komponenti i ovde neće biti prikazana.

### Zaključak

Iz priloženih rezultata uočavamo:

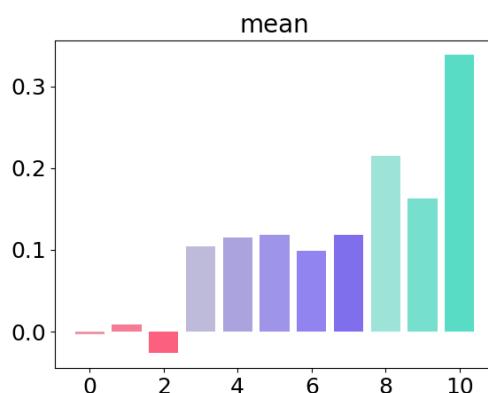
- da se na osnovu srednje vrednosti  $CDM$  matrice sa 40  $PCA$  komponenti i sa brojem  $PCA$  komponenti tako da je pokriveno 80% varijanse podataka

dobijaju rezultati koji nisu u saglasnosti za medulu i korteks ljudskog bubrega, a koji su u saglasnosti za bubreg miša, i

- da dobijeni klasteri nisu u skladu sa eksperimentalno utvrđenim skupovima podataka, čak se zbog izuzetno male vrednosti *ARI* skora ne može odrediti uticaj koordinata i genskih ekspresija ćelije na tip ćelije, kao što je to urađeno kod embriona i dorzalnog srednjeg mozga miša

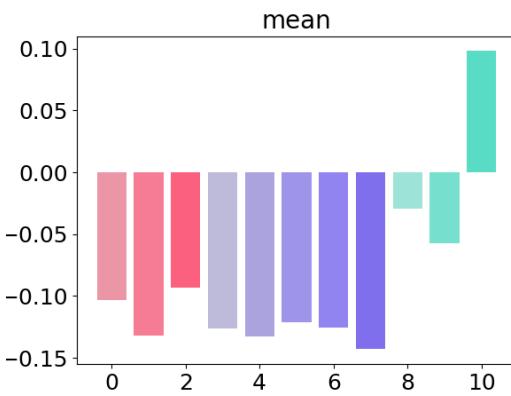
## 5.4 Diskusija

Srednje vrednosti *CDM* matrice za 40 PCA komponenti su prikazane na slici 5.39, dok su iste vrednosti za broj PCA komponenti tako da je pokriveno 80% varijanse podataka prikazane na slici 5.40.



Slika 5.39: Srednje vrednosti *CDM* matrice sa 40 PCA komponenti za embrion miša, dorzalni srednji mozak miša, medulu i korteks ljudskog bubrega i bubreg miša, redom

Sa slika 3.9, 3.12, 3.16 i 3.18 gde su nam prikazani tipovi ćelija pojedinačno uočavamo da kod embriona miša, pre svega kod prvog i trećeg preseka, postoji više ćelija nego kod preostalih skupova za koje važi da su prostorno bliske i imaju sličnu gensku ekspresiju ili suprotno - prostorno su udaljene i imaju različite genske ekspresije. Konkretno, kod prvog preseka embriona miša to su ćelije tipa 3, 5, 6, 7, 9, 10 i 12, dok kod trećeg preseka embriona miša to su ćelije tipa 1, 2, 4, 6, 8, 9 i 13 – 18 (slika 3.9). Na osnovu ovog zapažanja, može se očekivati da su srednje vrednosti *CDM* matrice za ove preseke embriona miša znatno bliže nuli u poređenju sa srednjim vrednostima *CDM* matrice za ostale skupove.



Slika 5.40: Srednje vrednosti  $CDM$  matrice brojem PCA komponenti tako da je pokriveno 80% varijanse podataka za embrion miša, dorzalni srednji mozak miša, medulu i korteks ljudskog bubrega i bubreg miša, redom

Kada analiziramo podatke sa 40 PCA komponenti (slika 5.39), primećujemo da su srednje vrednosti  $CDM$  matrice za sve preseke embriona miša dosta manje nego za ostale skupove. Ovo jeste u skladu sa očekivanim stanjem. Međutim, srednja vrednost  $CDM$  matrice za drugi presek embriona miša je bliža nego srednja vrednost  $CDM$  matrice za treći presek embriona miša. Ovo nije u skladu sa očekivanim stanjem, jer očekivano bi bilo da su srednje vrednosti  $CDM$  matrice za prvi i treći presek embriona miša bliže nuli u poređenju sa srednjim vrednostima  $CDM$  matrice za preostale preseke, a samim tim i za drugi presek embriona miša. Ovo ukazuje na to da se na osnovu srednje vrednosti  $CDM$  matrice sa 40 PCA komponenti ne može meriti korelacija između koordinata i genskih ekspresija ćelija.

Kada koristimo dovoljan broj PCA komponenti kako bismo pokrili 80% varijanse podataka (slika 5.40), primećujemo da su srednje vrednosti  $CDM$  matrice za embrion miša i dorzalni srednji mozak miša bliske i pomerene levo od nule. Ovo nije u skladu sa očekivanim stanjem, jer očekivano bi bilo da su srednje vrednosti  $CDM$  matrice za embrion miša bliže nuli u poređenju sa srednjim vrednostima  $CDM$  matrice za dorzalni srednji mozak miša. Takođe, srednje vrednosti  $CDM$  matrice za medulu i korteks ljudskog bubrega su bliže nuli nego srednje vrednosti  $CDM$  matrice za embrion miša. Ovo isto nije u skladu sa očekivanim stanjem. Mala razlika između srednjih vrednosti  $CDM$  matrice za embrion miša i dorzalni srednji mozak miša, kao i srednje vrednosti  $CDM$  matrice za medulu i korteks ljudskog bubrega koje su najbliže nuli ukazuju na to da se na osnovu srednje vrednosti  $CDM$  matrice sa brojem PCA komponenti tako da je pokriveno 80% varijanse podataka ne može

## *GLAVA 5. REZULTATI*

---

meriti korelaciju između koordinata i genskih ekspresija ćelija.

Ako uporedimo dobijene rezultate sa manjim i većim brojem *PCA* komponenti, uočavamo da su rezultati sa 40 *PCA* komponenti približniji očekivanom stanju, ali se ipak ne poklapaju u potpunosti.

*CDM* matrice za sve skupove podataka sa manjim i većim brojem *PCA* komponenti za koje smo prikazali rezultate su dobijene kao razlika vrednosti iz redukovanih i normalizovanih matrica susedstva koordinatnog i genskog grafa. Redukcija matrica susedstva je rađena tako što su iz njih izbačena rastojanja za one parove ćelija za koje su rastojanja u matrici susedstva za genski graf van opsega  $[mean - 3 \cdot std, mean + 3 \cdot std]$ , gde je *mean* srednja vrednost, a *std* standardna devijacija raspodele vrednosti iz matrice susedstva genskog grafa. Pored ovakvog formiranja *CDM* matrice probano je još par pristupa, gde je *CDM* matrica formirana na isti način, samo uz različitu redukciju matrica susedstva genskog i koordinatnog grafa. Prvo je probano bez bilo kakve redukcije ovih matrica susedstva, a zatim i sa redukcijom tako da su iz matrica susedstva za genski i koordinatni graf izbačena rastojanja za one parove ćelija za koje su rastojanja u matrici susedstva za genski graf:

- van opsega  $[mean - 2 \cdot std, mean + 2 \cdot std]$
- van opsega  $[mean - std, mean + std]$
- među 10% najvećih rastojanja
- među 20% najvećih rastojanja
- među 40% najvećih rastojanja
- među 60% najvećih rastojanja
- među 80% najvećih rastojanja
- među 90% najvećih rastojanja

Cilj ovih redukcija je zadržavanje parova ćelija koje dele slične genske ekspresije kako bi se istražila moguća korelacija između genske sličnosti i prostornog položaja ovih ćelija. Međutim, ni pomoću ovih pristupa nije postignuta očekivana razlika u srednjim vrednostima *CDM* matrica za analizirane skupove podataka. Zbog toga, ovi rezultati nisu prikazani u ovom radu, ali su dostupni u datoteci nazvanoj „Rezultati/Zadatak1” na Git repozitorijumu [6].

## *GLAVA 5. REZULTATI*

---

Dodatno, rezultati dobijeni za srednju vrednost  $CDM$  matrice pokazuju zavisnost od izbora metrike, u ovom slučaju, euklidske udaljenosti. Ovo postavlja pitanje mogućeg unapređenja rezultata putem promene metrike. Posebno, razmatranje rangiranja vrednosti matrica susedstva umesto samih vrednosti može pružiti nezavisnost od izbora metrike, čime se otvara perspektiva za potencijalno poboljšanje analize. Takođe, važno je napomenuti da u visokodimenzionalnim prostorima euklidska rastojanja između tačaka pokazuju relativno malu varijaciju. Ovo dodatno naglašava značaj problema izbora metrike, posebno u okruženju sa većim brojem PCA komponenti, i pruža dodatni kontekst za razmatranje alternativnih pristupa analizi genskih grafova.

U tabeli 5.5 su predstavljene statistike klasterovanja korišćenjem *Leiden* algoritma sa najvišim  $ARI$  skorom za embrion i dorzalni srednji mozak miša, primenjujući 40 PCA komponenti. Za medulu i korteks ljudskog bubrega, kao i za bubreg miša, ovi podaci nisu prikazani jer je maksimalna vrednost  $ARI$  skora manja od 0.1.

Primećujemo da kod analiziranih skupova podataka genska komponenta ima veći uticaj na tip ćelije u poređenju sa koordinatnom komponentom. Međutim, koordinatne komponente doprinose poboljšanju  $ARI$  skora i samim tim i njihov uticaj je značajan. Važno je istaći da  $ARI$  skorovi, koji mere koliko se klasteri podudaraju sa eksperimentalno utvrđenim tipovima ćelija, nisu dostigli zadovoljavajuće vrednosti, kako u slučaju embriona miša, tako i u slučaju dorzalnog srednjeg mozga miša.

Lošiji rezultati, naročito u kontekstu povećanja broja  $PCA$  komponenti, i ovde bi mogli biti posledica odabira metrike za udaljenost, s obzirom na činjenicu da u visokodimenzionim prostorima euklidska rastojanja između tačaka malo variraju. Ovo opet sugerše potrebu za daljim istraživanjem i eksperimentisanjem sa različitim metrikama udaljenosti kako bi se poboljšala preciznost određivanja rastojanja u genskom grafu.

## *GLAVA 5. REZULTATI*

---

Presek	Broj suseda za $G_1$	Broj suseda za $G_2$	<i>ARI</i> skor	Modularnost	Broj klastera	Broj tipova celija
Embrion miša P1	8 0	30 30	0.416 0.350	0.783 0.829	16 19	12 12
Embrion miša P2	3 0	15 15	0.573 0.507	0.745 0.770	13 14	13 13
Embrion miša P3	8 0	30 30	0.583 0.489	0.757 0.810	15 18	18 18
Dorzalni srednji mozak miša P1	5 0	25 25	0.208 0.172	0.394 0.430	5 5	11 11
Dorzalni srednji mozak miša P2	5 0	10 10	0.396 0.202	0.560 0.644	6 7	11 11
Dorzalni srednji mozak miša P3	8 0	30 30	0.495 0.307	0.490 0.543	3 4	11 11
Dorzalni srednji mozak miša P4	10 0	25 25	0.479 0.276	0.518 0.596	5 7	11 11
Dorzalni srednji mozak miša P5	9 0	20 20	0.367 0.191	0.519 0.630	6 9	11 11

Tabela 5.5: Statistike za optimalna klasterovanja i klasterovanje na osnovu samo genskog grafa za embrion i dorzalni srednji mozak miša

# Glava 6

## Zaključak

Ovo istraživanje donosi više važnih saznanja. Prvenstveno, tehnike prostorne transkriptomike pružaju moćan alat za istraživanje pojedinačnih ćelija u tkivu, omogućavajući nam da razumemo njihove genske ekspresije i prostornu organizaciju. Međutim, ovaj rad ukazuje na izazove u analizi ovih podataka i interpretaciji rezultata.

Na osnovu analize srednjih vrednosti *CDM* matrice, nastojali smo da istražimo vezu između koordinata i genskih ekspresija u ćelijama. Očekivali smo da će preseci kod kojih ima više ćelija koje su prostorno bliske i imaju sličnu gensku ekspresiju ili suprotno - prostorno su udaljene i imaju različite genske ekspresije imati srednje vrednosti *CDM* matrice koje su dosta bliže nuli. Rezultati nisu u potpunosti potvrdili ta očekivanja, što nam ukazuje da je neophodno razmotriti druge mere i pristupe kako bi se bolje razumele prostorne i genske karakteristike pojedinačnih ćelija.

Tokom klasterovanja primenom *Leiden* algoritma, *ARI* skor je korišćen kako bi se dobili klasteri koji bi što bolje odgovarali eksperimentalno utvrđenim tipovima ćelija. Iako dobijene vrednosti nisu bile zadovoljavajuće, ovaj skor je poslužio kako bismo analizirali uticaj koordinatnih i genskih komponenata na formiranje klastera kod embriona i dorzalnog srednjeg mozga miša. Zaključak koji se može izvući iz ovih rezultata jeste da genske komponente imaju dominantan uticaj na strukturu klastera, iako je evidentan i značajan uticaj prostornih koordinata. Takođe, rezultati *ARI* skora koji nisu ispunili očekivanja sugerisu potrebu za primenom drugih metrika udaljenosti i algoritama klasterovanja.

# Bibliografija

- [1] Adjusted Random Index. on-line at: [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted\\_rand\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html).
- [2] Kurtosis and Skewness. on-line at: <https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistics/basic-statistics/supporting-topics/data-concepts/how-skewness-and-kurtosis-affect-your-distribution/>.
- [3] Modularity. on-line at: <https://sparkling-graph.readthedocs.io/en/latest/modularity.html>.
- [4] MOSTA Database. on-line at: <https://db.cngb.org/stomics/mosta/>.
- [5] SlideSeqV2 data. on-line at: <https://cellxgene.cziscience.com/collections/8e880741-bf9a-4c8e-9227-934204631d2a>.
- [6] Git repozitorijum, 2023. on-line at: <https://github.com/milicas19/SpatialTranscriptomicsDataClustering>.
- [7] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188, 2008.
- [8] Mengnan Cheng, Yujia Jiang, Jiangshan Xu, Alexios-Fotios A. Mentis, Shuai Wang, Huiwen Zheng, Sunil Kumar Sahu, Longqi Liu, and Xun Xu. Spatially resolved transcriptomics: A comprehensive review of their technological advances, applications, and challenges. *Journal of Genetics and Genomics*, 2023.
- [9] J. L. Marshall, T. Noel, Q. S. Wang, H. Chen, E. Murray, A. Subramanian, K. A. Vernon, S. Bazua-Valenti, K. Liguori, K. Keller, R. R. Stickels, B. McBean, R. M. Heneghan, A. Weins, E. Z. Macosko, F. Chen, and A. Greaka. High-resolution Slide-seqV2 spatial transcriptomics enables discovery of disease-specific cell neighborhoods and pathways. *IScience*, 2022.

## BIBLIOGRAFIJA

---

- [10] Samuel G. Rodriques, Robert R. Stickels, Aleksandrina Goeva, Carly A. Martin, Evan Murray, Charles R. Vanderburg, Joshua Welch, Linlin M. Chen, Fei Chen, and Evan Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.
- [11] Robert R. Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L. Marshall, Daniela Di Bella, Paola Arlotta, Evan Z. Macosko, and Fei Chen. Sensitive spatial genome wide expression profiling at cellular resolution. *BioRxiv*, 2020.
- [12] Robert R. Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L. Marshall, Daniela Di Bella, Paola Arlotta, Evan Z. Macosko, and Fei Chen. Sensitive spatial genome wide expression profiling at cellular resolution. *bioRxiv*, 2020.
- [13] V. A.A Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, mar 2019.
- [14] Cameron G. Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 2022.

# Biografija autora

Milica Simić, rođena 05. novembra 1997. godine u Valjevu, odrasla je u Osečini. Već u ranim školskim danima pokazala je talenat za matematiku. U sedmom razredu, odlučila je da započne svoje putovanje u svet znanja upisavši Osnovnu školu pri Matematičkoj gimnaziji u Beogradu, a potom i Matematičku gimnaziju, koju je završila kao nosilac Vukove diplome.

Njen akademski put vodio je kroz izazovno okruženje Matematičkog fakulteta Univerziteta u Beogradu, gde je stekla temeljno obrazovanje sa prosekom ocena od 9.28. Decembra 2021. godine, započela je svoju profesionalnu karijeru u kompaniji Endava, gde je prvo provela tri meseca obavljanjem stručne prakse, da bi zatim postala stalni član tima.

Ono što posebno izdvaja Milicu je njena strast prema matematici, odbojci i umetnosti. Osim što je talentovana matematičarka, Milica je i strastvena odbojkašica koja je uspešno kombinovala sportske i akademske izazove tokom svog puta. Takođe, umetnost je deo njenog života, pružajući joj kreativni izlaz i inspiraciju.

S obzirom na svoju želju da se bavi oblastima koje imaju pozitivan uticaj na ljude, Milica je odabrala da istražuje područje bioinformatike kao temu svog master rada.