VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

BIOINFORMATIKOS BAKALAURO STUDIJŲ PROGRAMA

Tbx5 transkripcijos faktoriaus tyrimas *Mus musculus* širdies ląstelėse

Research of Tbx5 transcription factor in *Mus musculus* heart cells

Kursinis darbas

Autorius: Danielė Stasiūnaitė
VU el. p.: (daniele.stasiunaite@mif.stud.vu.lt)

Darbo vadovas: J. m. d. Kotryna Kvederavičiūtė

Vilnius
2022

Turinys

Santrauka

Regeneracijos procesų tyrinėjimas yra svarbi sritis, galinti prisidėti prie įvairių ligų bei traumų gydymo, todėl yra svarbu išsiaiškinti šį procesą valdančius mechanizmus bei juose dalyvaujančius įvairių genų produktus - baltymus.

Šiame darbe naudojantis R programavimo kalbos bibliotekomis bei bioinformatiniais komandinės eilutės bei internetiniai duomenų apdorojimo bei analizės įrankiais išanalizuota, kokiuose ChIP sekoskaitos metodu gautų naminės pelės (lot. *Mus musculus*) ląstelių regionų mėginiuose Tbx5 transkripcijos faktoriaus motyvų yra daugiausiai bei kokie veiksniai gali įtakoti skirtingą faktoriaus motyvų skaičių naminių pelių genomo sekose.

Atliktos analizės metu mėginiuose identifikuota daug skirtingų motyvų. Daugiausiai tyrinėjamo Tbx5 transkripcijos faktoriaus motyvų identifikuota naminių pelių embrionų fibroblastų ląstelių, veiktų serino/treonino kinaze 1 (Akt1) bei kardiogeniniais transkripcijos faktoriais GATA4, HAND2 ir MEF2C antroje chromosomoje. Mažiausias Tbx5 transkripcijos faktoriaus motyvų skaičius būdingas fibroblastams, kurie buvo veikti ne pilnu kardiogeninių faktorių rinkiniu.

Raktiniai žodžiai: ChIP sekoskaita, transkripcijos faktorius, Tbx5, regionas, motyvas, R.

Summary

The investigation of regeneration processes is an important field of research that plays a significant role in the treatment of various diseases and injuries. Therefore, it is mandatory to determine and apprehend the mechanisms and gene products that regulate regeneration processes.

In this work, the analysis of the ChIP sequencing peaks samples that were retrieved from different house mouse (lat. *Mus musculus*) cells was conducted using functions from R programming language libraries and performing analysis steps using bioinformatics command-line tools and online services in order to determine what type of house mouse cells has the greatest number of Tbx5 transcription factor motifs and what factors might have an impact on total Tbx5 motif hit count differences among house mouse genome sequences.

The undertaken analysis identified an ample number of motifs. An abundance of key interest Tbx5 transcription factor was identified in the second chromosome of house mouse embryonic fibroblasts that were treated with serine/threonine kinase 1 (Akt1) and GATA4, HAND2, MEF2C cardiogenic transcription factors. The least count of Tbx5 transcription factor motifs is common for fibroblasts that were not treated with complete cardiogenic transcription factor set.

Keywords: ChIP-seq, transcription factor, Tbx5, peak, motif, R.

1 Įvadas

Darbo temos aktualumas

Spartėjanti mokslo raida, įvairūs atradimai bei išradimai stipriai paspartino ir pagerino įvairių ligų diagnostikos bei prevencijos tyrimus, praplėtė žinias ląstelės biologijos, genetikos, fiziologijos ir kitose srityse. Viena iš medicinos sričių, kuri pradėta tyrinėti dar XVIII a., kai buvo nustatyta, kad kai kurie organizmai geba atsiauginti prarastas arba sužeistas galūnes bei kitus kūno audinius[1], yra organizmų audinių bei organų regeneracija.

Šiais laikais regeneracijos tyrimai atliekami su hidromis, planarijomis, tritonais bei zebražuvėmis[2], siekiant išsiaiškinti šių organizmų audinių regeneracijos mechanizmus bei pritaikyti žmonėms, patyrusiems traumas ar turintiems specifinių kūno audinių pažeidimų. Regeneracijos procese pagrindinę funkciją atlieka kamieninės ląstelės bei įvairūs transkripcijos reguliavimo faktoriai, gebantys prisijungti prie DNR chromatino ir skatinti arba slopinti specifinių genų transkripciją.

Sėkmingam regeneracijos procesų mechanizmų supratimui būtina išsiaiškinti, kokie transkripcijos faktoriai dalyvauja šiame procese bei kokias funkcijas jie atlieka.

Darbo tikslas

Pagrindinis šio darbo tikslas yra palyginti Tbx5 transkripcijos faktoriaus regionų skirtumus skirtinguose *Mus musculus* ląstelių mėginiuose, gautuose panaudojus ChIP-seq metodą.

Uždaviniai

- Nustačius Tbx5 transkripcijos faktoriaus regionų skaičių mėginiuose įvertinti, kuriuose mėginiuose praturtintų genominių regionų buvo didžiausias ir mažiausias.
- Išsiaiškinti, kaip skiriasi regionų skaičius skirtingose genominėse pozicijose (chromosomose).
- Apskaičiavus tarp mėginių persidengiančių regionų procentinę dalį nustatyti, kurie mėginiai yra panašiausi.
- Identifikavus Tbx5 transkripcijos faktoriaus sekos motyvą nustatyti, kurių mėginių ląstelės turi didžiausią transkripcijos faktoriaus motyvo kiekį.
- Atlikus *de novo* motyvų paiešką nustatyti, kokios biologinės funkcijos yra būdingos identifikuotiems motyvams.

2 Duomenų bazės ir duomenys

2.1 GTRD duomenų bazė

Tyrimui naudoti duomenys atsisiųsti iš GTRD (Gene Transcription Regulation Database)[3] duomenų bazės, saugančios informaciją apie transkripcijos sekų ir atviro chromatino regionus. Taip pat duomenų bazėje saugomi nekartografuojamų regionų duomenys bei potencialūs žmonių bei naminių pelių regionai, prie kurių gali jungtis transkripcijos faktoriai.

Ši duomenų bazė pasirinkta dėl sistemiškai surinktų ChIP-seq eksperimentų, kurių metu gauti rezultatai yra unifikuotai apdoroti ir paruošti tyrėjų meta-analizėms.

GTRD duomenų bazėje duomenys saugomi binariniu anotacijų formatu *bigBed*, leidžiančiu atvaizduoti pasirinktą chromosomos regioną interaktyviose genominės informacijos vizualizavimo naršyklėse (pavyzdžiui, UCSC Genome Browser[4]) efektyviau nei tekstinis BED formatas.

2.2 Pasirinktų mėginių charakteristika

Analizė atlikta, naudojantis 4 nepriklausomais eksperimentais, kuriuos iš viso sudarė 7 biologinės replikos. Pirmoje lentelėje pateikta informacija apie tyrimui atlikti naudotus duomenis, surinktus iš naminės pelės (lot. *Mus musculus*) ląstelių.

1 lentelė. Mėginių charakteristikos

GTRD ID	Ląstelių tipas	Kamienas	Poveikis	Antikūnai	PubMed ID
EXP030898	HL - 1 (širdies raumens)	C57BL/6J	TRE promotorius (2 d.)	-	21415370[5]
EXP058852	Širdies prieširdžių	C57BL/6	-	Tbx5 (sc-17866)	31080136[6]
EXP062056	Pelių naujagimių širdies fibroblastų, ekspresuojančių didelį kiekį T antigeno, linija	CD1	sb431542, xav939	anti-TBX5 (sc-17866x)	31271750[7]
EXP058843	MEF (embrionų fibroblastai)	C57BL/6	AGHMT (2 d.)	anti-Tbx5 (sc-17866)	31080136[6]
EXP058847	MEF (embrionų fibroblastai)	C57BL/6	GHMT (2 d.)	Tbx5 (sc-17866)	31080136[6]
EXP058850	MEF (embrionų fibroblastai)	C57BL/6	GMT (2 d.)	Tbx5 (sc-17866)	31080136[6]
EXP058856	MEF (embrionų fibroblastai)	C57BL/6	vienas faktorius (2 d.)	Tbx5 (sc-17866)	31080136[6]

2.3 Santrumpų bei pavadinimų paaiškinimai

- **HL - 1**: pelių širdies raumens ląstelės, išgautos iš navikinių prieširdžių kardiomiocitų linijos. Šios ląstelės gali betarpiškai dalintis ir spontaniškai keisti savo formą, vykstant širdies raumens susitraukimo/ atsipalaidavimo procesams.

- **MEF**: pelių embrionų fibroblastai (angl. *Mouse Embryonic Fibroblast*). Šiai ląstelių linijai būdingas ląstelių gyvybingumo apribojimas, reiškiantis, jog šios ląstelės greitai pasensta ir miršta.

- **C57BL/6**: inbrydingo (angl. *inbreeding*) būdu išvestų naminių pelių veislė. Šios veislės pelėms būdingas itin tamsus kailis, padidėjęs jautrumas garsams, kvapams, skausmui ir žemai temperatūrai. Ši veislė dažnai naudojama nutukimą ir imuninę sistemą tiriančiuose tyrimuose.

- **C57BL/6J**: prie naminių pelių veislės pavadinimo pridėtos raidės patikslina, kurioje laboratorijoje veislės išvestos. 'J' raidė nurodo, kad pelių veislė išvesta Meino valstijoje (JAV) įsikūrusioje Džeksono laboratorijoje[8].

- **CD1**: autbrydingo (angl. *outbreeding*) būdu išvestų naminių pelių veislė. Šios veislės pelėms būdingas baltas kailis. Taip pat CD1 pelės dažnai naudojamos genetiniuose, toksikologiniuose, farmakologiniuose ir senėjimo tyrimuose.

- **TRE**: tetraciklino atsako elementas (angl. *Tetracycline Response Element*). Tai yra 7 DNR sekos fragmentai, sudaryti iš 19 nukleotidų ir atskirti trumpesniais sekų fragmentais.

- **sb431542**: stipriai veikianti, selektyvi cheminė medžiaga; transformuojančio augimo faktoriaus β (TGF-β) inhibitorius.

- **xav939**: stipriai veikianti cheminė medžiaga; tankirazės inhibitorius. Tankirazė slopina TERF1 baltymo, stabdančio telomerazės veiklą, jungimąsi prie telomerinių DNR sekų.

- **AGHMT**: AKT1 - serino/treonino kinazė 1; GATA4, HAND2, MEF2C, Tbx5 - kardiogeniniai transkripcijos faktoriai.

- **GHMT**: GATA4, HAND2, MEF2C, TBX5 transkripcijos faktorių komplektas.

- **GMT**: GATA4, MEF2C, Tbx5 transkripcijos faktorių komplektas.

- **sc-17866x/ sc-17866**: iš ožkų išskirti antikūnai, atpažįstantys žmonių, pelių ir žiurkių Tbx5 antigeną.

2.4 Pasirinktų eksperimentų apžvalga

- **EXP030898**: vienas mėginys iš septyniolikos eksperimento metu tirtų mėginių. Eksperimente buvo siekiama patvirtinti arba atmesti hipotezę apie širdies stipriklių (angl. *enhancer*) identifikiavimą prie chromatino jungiantis keliems transkripcijos faktoriams.

  HL - 1 širdies raumens ląstelės buvo infekuotos su adenovirusu, ekspresuojančiu troponiną T, kuris skatina *rtTA* ir *BirA* genų ekspresiją, bei TRE promotoriumi, skatinančiu Tbx5 transkripcijos faktoriaus geno raišką. Sąlygos taikytos 48 valandas.

- **EXP058852**: prieširdžių ląstelės buvo du kartus po 24 valandas laikytos mišinyje su retrovirusais. Papildomi poveikiai nebuvo taikyti.

- **EXP062056**: eksperimente ląstelės buvo infekuotos su GATA4, Mef2c, ir Tbx5 transkripcijos faktorius sintetinančiais retrovirusais. Ląstelės augintos terpėje, kurioje buvo Tgfβ inhibitoriaus sb431542, skatinančio kardiomiocitų diferenciaciją iš pliuripotentinių kamieninių ląstelių, ir Wnt inhibitoriaus xav939, stabdančio nediferencijuotų ląstelių sintezę ir skatinančio progenitorinių ląstelių kardiomiogenezę.

Pasirinktų duomenų rinkinyje naudoti vieno eksperimento, kuriame buvo tirti širdies ląstelių atsinaujinimo ir diferenciacijos mechanizmai, keturiais skirtingais poveikiais tirti mėginiai:

- **EXP058843**: embrionų fibroblastai veikti AGHMT. Esant kardiogeniniams transkripcijos faktoriams, AKT1 skatina fibroblastų diferenciaciją į širdies ląsteles - kardiomiocitus.

- **EXP058847**: neįtraukta AKT1 serino/treonino kinazė 1.

- **EXP058850**: į kardiogeninių transkripcijos faktorių mišinį neįtrauktas HAND2 transkripcijos faktorius.

- **EXP058856**: ląstelės veiktos tik vienu faktoriumi, kuris straipsnyje nebuvo specifikuotas.

3 Tyrimo metodai

Tbx5 transkripcijos faktoriaus regionų tyrimo analizė atlikta su R programavimo kalba[9] (4.2.0 versija).

Tarpiniams analizės rezultatams pateikti naudotas komandinės eilutės įrankis Scikick[10] (0.2.0 versija), leidžiantis generuoti R Markdown (Rmd) ataskaitas *html* formatu bei kurti struktūrizuotus puslapius, apjungiant iš daugelio Rmd failų gautus HTML ataskaitų failus.

3.1 Regionų skaičiaus nustatymas mėginiuose

Tbx5 regionų skaičius skirtinguose mėginiuose apskaičiuotas su standartine R ilgio funkcija *length()*, kuri pritaikyta *GRanges* objektui, aprašančiam genomines pozicijas bei su jomis susijusias anotacijas. Objektas sukurtas su *rtracklayer*[11] bibliotekos funkcija *import()*.

Regionų skaičių mėginiuose atvaizduojanti stulpelinė diagrama sukurta su *ggplot2*[12] bibliotekos *geom_ bar()* funkcija.

3.2 Regionų skaičiaus nustatymas chromosomose

Transkripcijos faktoriaus regionų skaičius skirtingose chromosomose kiekvienam mėginiui apskaičiuotas, naudojantis standartine R funkcija *length()*, pritaikyta atskiroms chromosomoms, kurių pozicijos aprašytos *GRanges* objekte. Kiekvieno mėginio Tbx5 transkripcijos faktoriaus pasiskirstymas chromosomose atvaizduotas su *ggplot()* ir papildoma funkcija *facet_ wrap()*, sukuriančia atskirus grafikus pagal pasirinktą elementą - chromosomas.

3.3 Persidengiančių regionų procentinė dalis

Persidengiančių regionų tarp mėginių procentinė dalis nustatyta su modifikuota *Jaccard()* funkcija, apskaičiuojančia, kiek yra sutampančių regionų tarp dviejų mėginių poros. Naudojantis nemodifikuota funkcija, Jaccard koeficientas apskaičiuojamas pagal išraišką:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Jaccard koeficientas gaunamas iš rinkiniams A ir B bendrų duomenų ilgio padalinus dviejų duomenų rinkinių bendrą duomenų ilgį.

Modifikavus *Jaccard* koeficiento gavimo funkciją, koeficientas apskaičiuojamas pagal išraišką, kur sutampančių A ir B rinkinių duomenų ilgis padalinamas iš A rinkinio ilgio:

$$J(A,B) = \frac{|A \cap B|}{|A|}$$

Jaccard koeficiento skaičiavimo funkcija modifikuota, nes skaičiuojant koeficientą su standartine *Jaccard* funkcija, gaunamas itin didelis regionų sąjungos skaičius, o persidengiančių regionų skaičius gaunamas mažas, todėl persidengiančių regionų skaičių padalinus iš regionų sąjungos gaunamas itin mažas koeficientas, kurio apskaičiuota procentinė dalis neretai neviršijo 1%.

Tam, jog būtų galima patikimiau įvertinti, kokia pirmojo mėginio procentinė regionų dalis persidengia su antruoju mėginiu, persidengiančių regionų skaičius padalintas iš pirmojo mėginio regionų skaičiaus.

Gauti rezultatai atvaizduoti spalvų intensyvumo grafike (angl. *heatmap*), sukurtame su *ggplot()* ir papildoma funkcija *geom_tile()*.

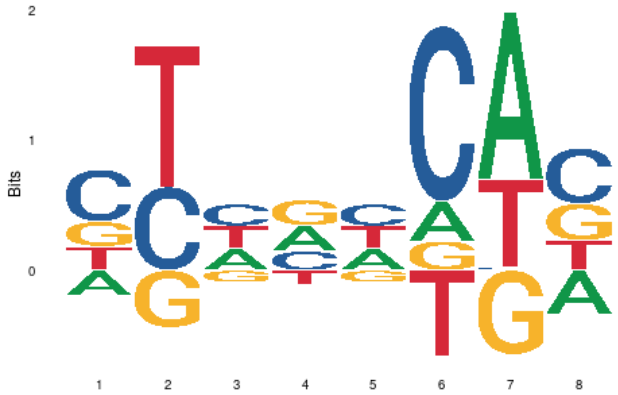
3.4 Tbx5 motyvo nustatymas

Šiame etape genomines pozicijas aprašantys *bigBed* formato failai konvertuoti į BED formato failus, pasinaudojus UCSC komandinės eilutės programa *bigBedToBed*[13].

Sugeneruoti BED formato failai panaudoti pikus atitinkančių sekų iš naminės pelės genomo gavimui FASTA formatu. Sekos iš genomo išgautos, pasinaudojus komandinės eilutės įrankio BEDTools[14] (2.30.0 versija) programa *getfasta*[15].

Kiekviename mėginyje esančio Tbx5 transkripcijos faktoriaus motyvo procentinė dalis apskaičiuota susumavus *Biostrings*[16] bibliotekos funkcijos *countPWM()* rezultatus. Funkcijai *countPWM()* kaip argumentas pateikta Tbx5 transkripcijos faktoriaus pozicinė svorių matrica bei mėginio nukleotidų sekų rinkinys FASTA formatu. Gauta funkcijos reikšmė padalinta iš bendro regionų skaičiaus.

Pozicinė svorių matrica atsisiųsta iš HOCOMOCO[17] (11.0 versija) (angl. *HOmo sapiens COmprehensive MOdel COllection*) duomenų bazės *Homo sapiens* ir *Mus musculus* organizmų transkripcijos faktorių kolekcijos. Pozicinę svorių matricą atitinkantis sekos logotipas vaizduojamas pirmame paveiksle (1 pav.).



1 pav. Tbx5 transkripcijos faktoriaus sekos logotipas

Identifikuotų Tbx5 transkripcijos faktoriaus motyvų skaičius vizualizuotas su pagrindinėmis *ggplot()* ir *geom_ bar()* funkcijomis.

3.5 Motyvų paieška *de novo*

Praturtintų sekų radimui panaudota komandinės eilutės įrankio HOMER[18] (v4.11 versija) programa *findMotifsGenome.pl*, analizuojanti BED formato failus (faile specifikuotas pozicijas), ir ieškanti praturtintų sekų atitikimo anotuotame naminės pelės *mm10* referentiniame genome. Tarp mėginių persidengiantys motyvai nustatyti, naudojantis R biblioteka *UpSetR*[19].

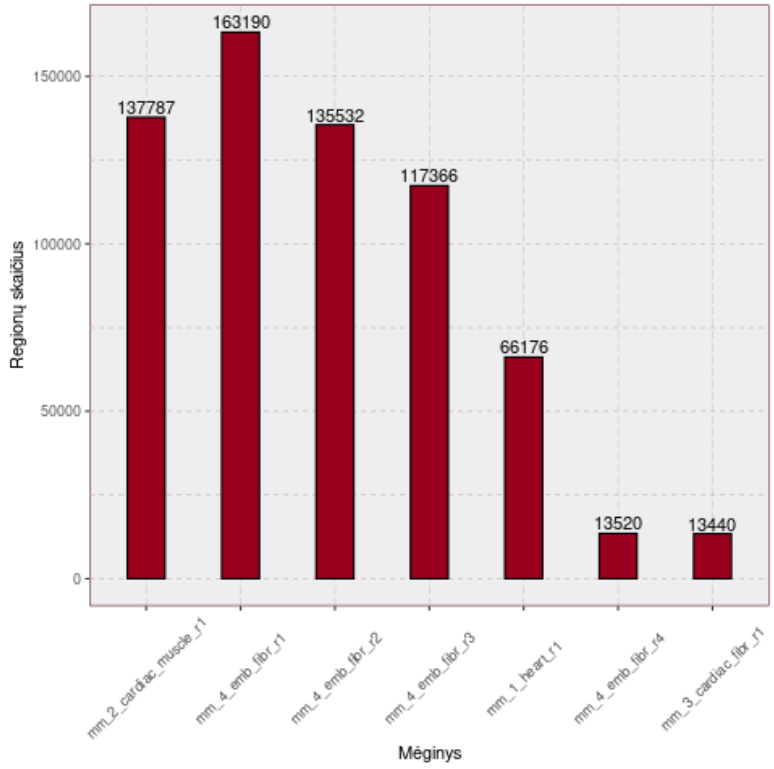
3.6 Praturtintų sekų biologinių funkcijų nustatymas

Identifikuotų motyvų biologinės funkcijos nustatytos, pasinaudojus UniProt[20] duomenų bazės genų ontologijos (angl. *Gene Ontology (GO)*) biologinių procesų, ląstelinių komponentų ir molekulinių funkcijų klasifikacija.

4 Rezultatai ir jų aptarimas

4.1 Regionų skaičiaus skirtumai tarp mėginių

Pirmajame analizės etape kiekviename mėginyje nustatytas bendras regionų skaičius pavaizduotas pirmoje stulpelinėje diagramoje (2 pav.).



2 pav. Regionų skaičių mėginiuose vaizduojanti stulpelinė diagrama

Remiantis diagrama didžiausias Tbx5 transkripcijos faktoriaus regionų skaičius nustatytas eksperimento *mm_4_emb_fibr_r1* techninėje replikoje, kurioje pelių embrionų fibroblastų ląstelės dvi dienas veiktos AGHMT faktoriais. Šį rezultatą palyginus su kitomis biologinėmis replikomis, kuriose tirtas tas pats pelių embrionų fibroblastų ląstelių kamienas, tačiau ląstelės veiktos tik kai kuriais faktoriais, pastebimas gradualus Tbx5 transkripcijos faktoriaus regionų skaičiaus mažėjimas diagramoje *mm_4_emb_fibr_r2*, *mm_4_emb_fibr_r3* ir *mm_4_emb_fibr_r4* pavaizduotuose stulpeliuose.

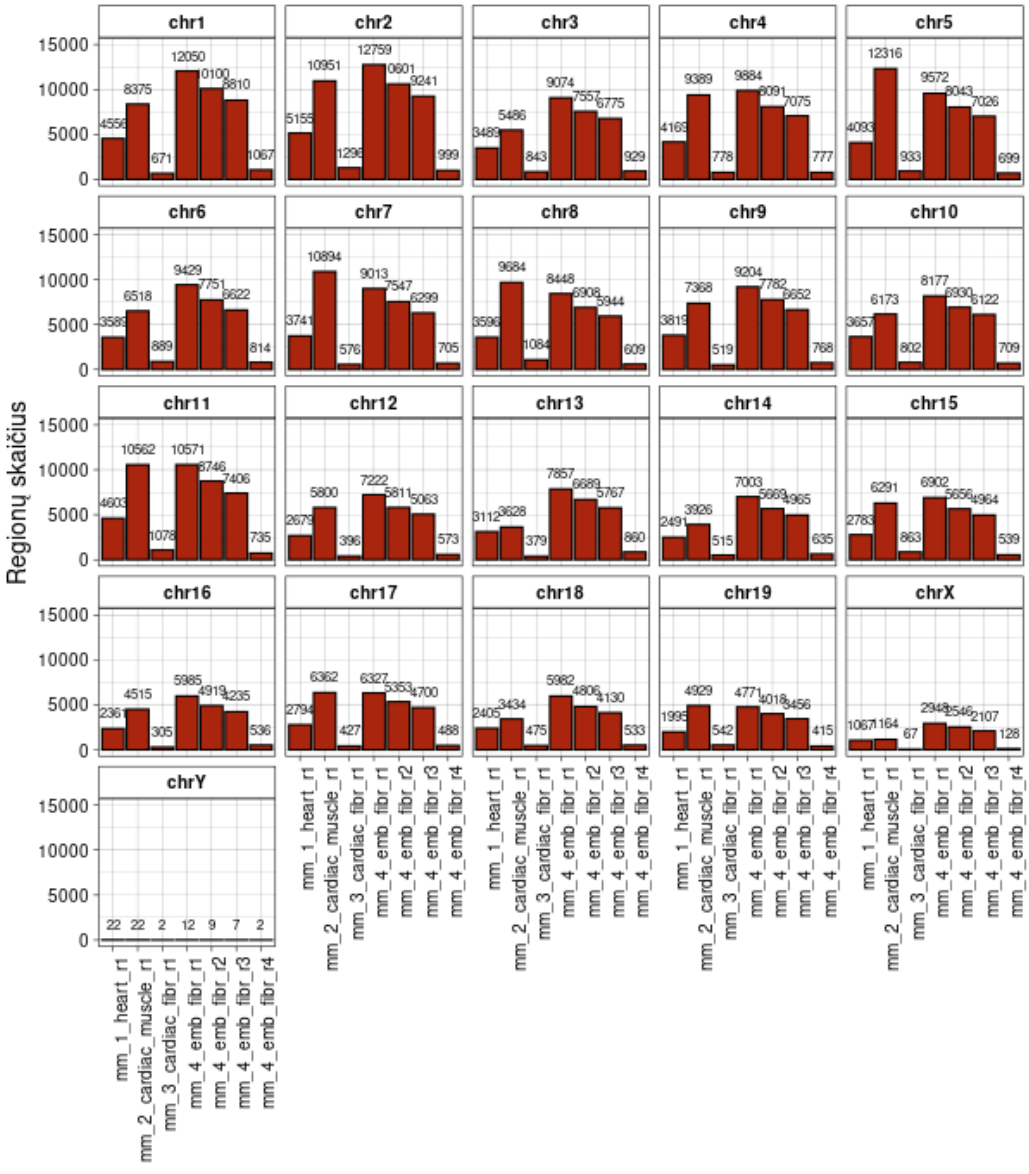
Mėginyje, kuriame embrionų fibroblastai veikti tik vienu faktoriumi, Tbx5 transkripcijos faktoriaus regionų nustatyta nedaug - 13520.

Mažiausiai regionų nustatyta mėginyje, kuriame tirta pelių naujagimių širdies fibroblastų, ekspresuojančių T antigeną ir paveiktų inhibitoriais: sb431542 ir xav939. Nepaisant to, kad abu inhibitoriai skatina širdies ląstelių diferenciaciją[22], itin mažas transkripcijos faktoriaus regionų skaičius rodo, kad papildomas veikimas inhibitoriais daro mažą įtaką transkripcijos faktoriaus jungimuisi prie DNR sekų.

4.2 Regionų pasiskirstymas chromosomose

Nustačius Tbx5 transkripcijos faktoriaus regionų pasiskirstymą eksperimentų mėginiuose, kitame analizės etape patikrinta, kaip faktoriaus regionai pasiskirstę atskirose chromosomose.

Vaizduojamuose grafikuose (3 pav.) didžiausias regionų skaičius nustatomas pirmoje, antroje ir penktoje chromosomose. Naminės pelės pirmoji chromosoma yra pati didžiausia, turinti 195 milijonų bazių porų, antroji chromosoma sudaryta iš 182 megabazių, penktoji chromosoma - 152 milijonų bazių porų, todėl didesnis regionų skaičius šiose chromosomose nėra neįprastas reiškinys. Kitose chromosomose regionų skaičius yra mažesnis. Ypač mažas regionų skaičius nustatytas devynioliktoje (61 Mbp), X (169 Mbp) ir Y (91 Mbp) chromosomose.



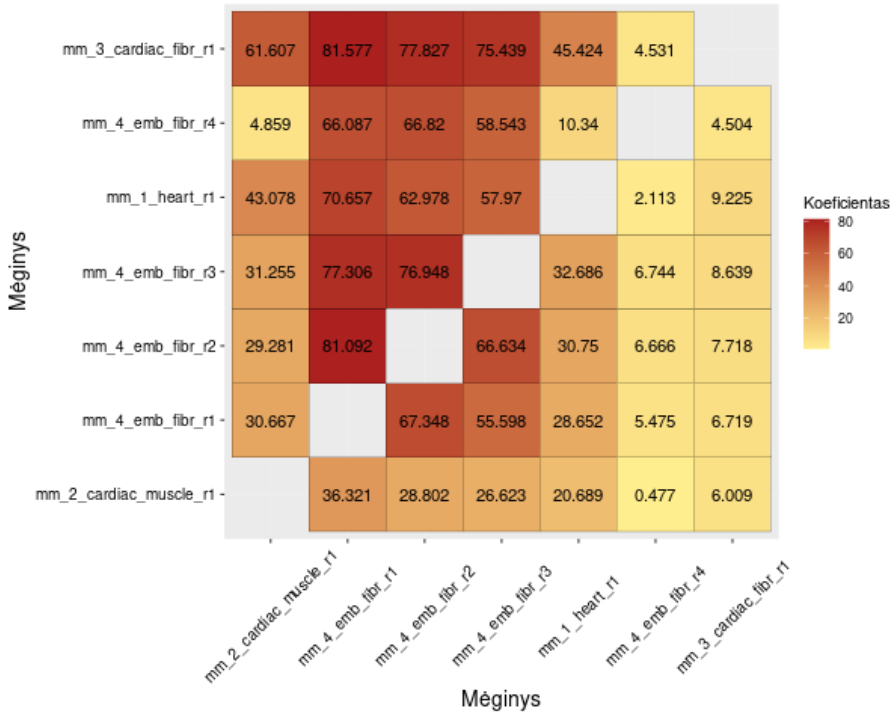
3 pav. Regionų pasiskirstymas chromosomose

Biologinių replikų mėginiuose didžiausias regionų skaičius nustatytas antroje chromosomoje. Taip pat grafikuose išsiskiria kontrolinis HL - 1 širdies ląstelių mėginys, kuriame didžiausias regionų skaičius nustatytas penktojoje chromosomoje.

Remiantis pavaizduotomis regionų skaičiaus pasiskirstymo chromosomose stulpelinėmis diagramomis, itin išsiskiriantis atrankumas chromosomų atžvilgiu nenustatytas, todėl galima teigti, jog šiame analizės etape duomenų problematiškumas nepastebimas arba jo nėra.

4.3 Tarp mėginių persidengiantys regionai

Dažnai siekiant nustatyti mėginių panašumą, yra tiriama, kokia mėginių duomenų dalis persidengia.



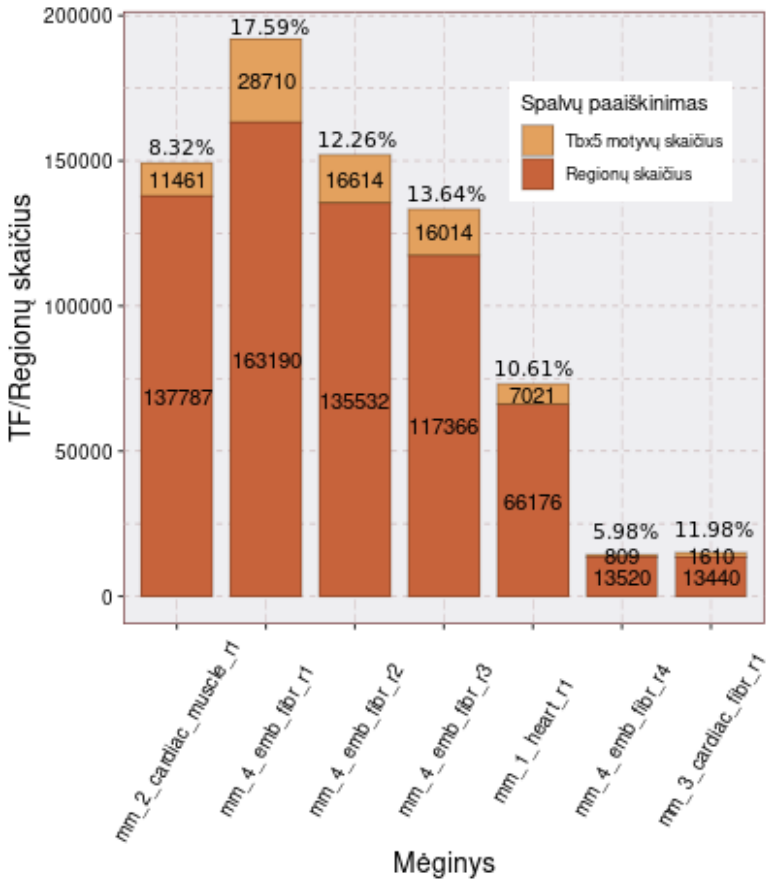
4 pav. Persidengiančių regionų procentinės dalies spalvų intensyvumo grafikas

Remiantis pavaizduoto spalvų intensyvumo grafiko (4 pav.) duomenimis, didžiausi persidengiančių regionų procentai nustatyti tarp šių mėginių:

- **81.577 %** - tarp mėginio, kuriame buvo tiriamos širdies fibroblastų ląstelės, ekspresuojančios T antigeną, ir mėginio, kuriame tirti embrionų fibroblastai, veikiant AGHMT.
- **81.092 %** - tarp mėginio, kuriame tirti embrionų fibroblastai ir mėginio, kuriame nebuvo AKT1.
- **77.827 %** - tarp mėginio su T antigeną ekspresuojančiomis širdies fibroblastų ląstelėmis ir mėginio, kuriame nebuvo AKT1.
- **76.948 %** - tarp mėginio, kuriame nebuvo HAND2 faktoriaus, ir mėginio, kuriame nebuvo AKT1.
- **75.439 %** - tarp mėginio su T antigeną ekspresuojančiomis širdies fibroblastų ląstelėmis ir tarp mėginio, kuriame nebuvo HAND2 faktoriaus.

4.4 Tbx5 motyvo pasiskirstymas mėginiuose

Penktajame grafike (5 pav.) pavaizduota, kiek Tbx5 motyvo atitikimų nustatyta skirtinguose mėginiuose.



5 pav. Tbx5 motyvų atitikimų skaičiaus palyginimo sudėtinė diagrama

Daugiausiai Tbx5 motyvo sekos (AGGTGTCA) atitikimų (39820) nustatyta mėginyje, kuriame embrionų fibroblastai veikti serino/treonino kinaze 1 (Akt1) bei keliais transkripcijos faktoriais (GATA4, HAND2, MEF2C, Tbx5).

Mažiausias Tbx5 motyvo sekų skaičius nustatytas mėginiuose, kuriuose embrionų fibroblastai veikti tik vienu transkripcijos faktoriumi (*mm_4_emb_fibr_r4*), ir mėginyje, kuriame pelių naujagimių širdies fibroblastai veikti sb431542 ir xav939 inhibitoriais (*mm_3_cardiac_fibr_r1*). Nepaisant to, kad šiuose mėginiuose motyvų skaičius mažiausias, remiantis procentine motyvų mėginiuose dalimi, mėginys (*mm_3_cardiac_fibr_r1*) Tbx5 motyvo sekos fragmentų turi daug (1610), atsižvelgus į bendrą šio mėginio sekų kiekį.

Mėginio, kuriame tirtos širdies raumens ląstelės (*mm_2_cardiac_muscle_r1*), Tbx5 motyvų sekų nustatyta mažai, palyginus identifikuoto transkripcijos faktoriaus motyvo sekų skaičių su bendru šio mėginio sekų rinkinio dydžiu.

4.5 *De novo* identifikuoti motyvai

De novo motyvų paieškos programos įvykdymas buvo ilgiausiai trukęs analizės etapas, lyginat su kitais tyrimo žingsniais. Šio etapo metu buvo sugeneruoti HTML formato failai, kuriuose buvo pateiktas identifikuotų motyvų sąrašas, išrikiuotas pagal p-vertę didėjančia tvarka, motyvų sekų logotipai, nuorodos į puslapius su pozicinėmis motyvų svorių matricomis bei identifikuotų motyvų procentinę dalį visame mėginio sekų rinkinyje, pateiktame FASTA formatu.

Trečioje lentelėje (3 lentelė) kiekvienam mėginiui pavaizduoti trys motyvai, turintys mažiausią p-vertę bei apimantys didžiausią mėginių pilno sekų rinkinio dalį (procentiškai). Penktame trečios lentelės stulpelyje nurodyta, kokią procentinę dalį mėginyje sudaro identifikuotas Tbx5 motyvas.

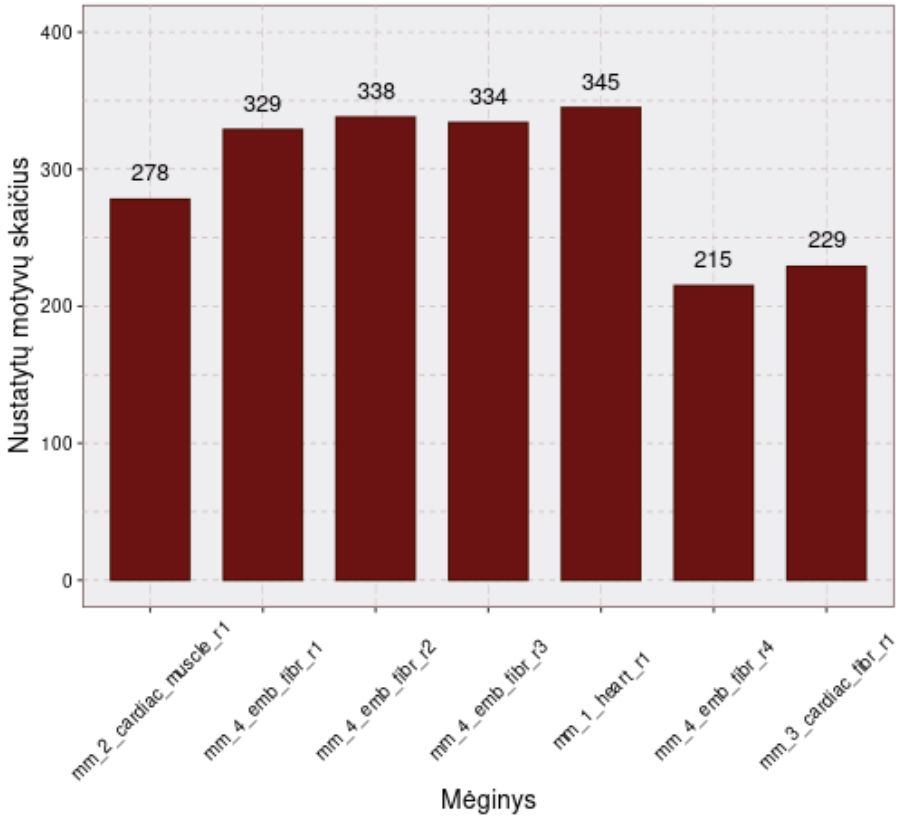
3 lentelė. Identifikuotų motyvų pavyzdžiai

Mėginys	**Pavadinimas**	**p vertė**	**Procentinė dalis**	***Tbx5* motyvas**
mm_2_cardiac_muscle_r1	Tbx6(T-box)	1e-1881	20.28%	1e-3266; 54.87%
	Tbet(T-box)	1e-1472	16.48%	
	Eomes(T-box)	1e-1332	25.49%	
mm_4_emb_fibr_r1	Mef2b(MADS)	1e-3037	15.60%	1e-2359; 44.22%
	TRPS1(Zf)	1e-2983	31.99%	
	GATA3(Zf)	1e-2936	25.49%	
mm_4_emb_fibr_r2	Fos(bZIP)	1e-2912	13.22%	1e-1873; 42.97%
	Fra1(bZIP)	1e-2889	12.66%	
	Fra2(bZIP)	1e-2855	11.36%	
mm_4_emb_fibr_r3	GATA3(Zf)	1e-1895	25.06%	1e-1391; 41.04%
	TRPS1(Zf)	1e-1872	31.55%	
	Fos(bZIP)	1e-1857	11.66%	
mm_1_heart_r1	Mef2c(MADS)	1e-1226	8.45%	1e-438; 39.08%
	Mef2b(MADS)	1e-1174	12.63%	
	Mef2d(MADS)	1e-1174	5.42%	
mm_4_emb_fibr_r4	TRPS1(Zf)	1e-1404	63.53%	1e0; 26.57%
	GATA3(Zf)	1e-1271	52.43%	
	GATA4(Zf)	1e-1024	38.94%	
mm_3_cardiac_fibr_r1	Tbx6(T-box)	1e-1431	38.28%	1e-1474; 69.71%
	Tbet(T-box)	1e-1077	33.27%	
	Tbx21(T-box)	1e-1034	30.60%	

Remiantis trečios lentelės duomenimis bei naudojantis statistinių hipotezių testavimu itin mažos p vertės rodo, jog tikimybė, kad identifikuoti motyvai atsitiktiniai, yra labai maža, todėl gauti duomenys yra statistiškai reikšmingi - juos galima toliau analizuoti ir daryti įvairias išvadas.

Pirmajame ir paskutiniame mėginiuose, kurie lentelėje pažymėti *mm_2_cardiac_muscle_r1* ir *mm_3_cardiac_fibr_r1*, Tbx5 motyvo p vertės buvo mažiausios, o procentinė dalis - didžiausia. Palyginus šių mėginių rezultatus su 4.4 dalyje gautais tų pačių mėginių rezultatais pastebimas didelis procentinės dalies neatitikimas, tačiau šie procentai apskaičiuoti, naudojantis skirtingais metodais. Taip pat 4.4 dalyje naudota pozicinė svorių matrica neatitinka šioje dalyje identifikuoto Tbx5 motyvo pozicinės svorių matricos.

Pateiktoje stulpelinėje diagramoje (6 pav.) vaizduojamas bendras identifikuotų motyvų skaičius kiekvienam mėginiui.

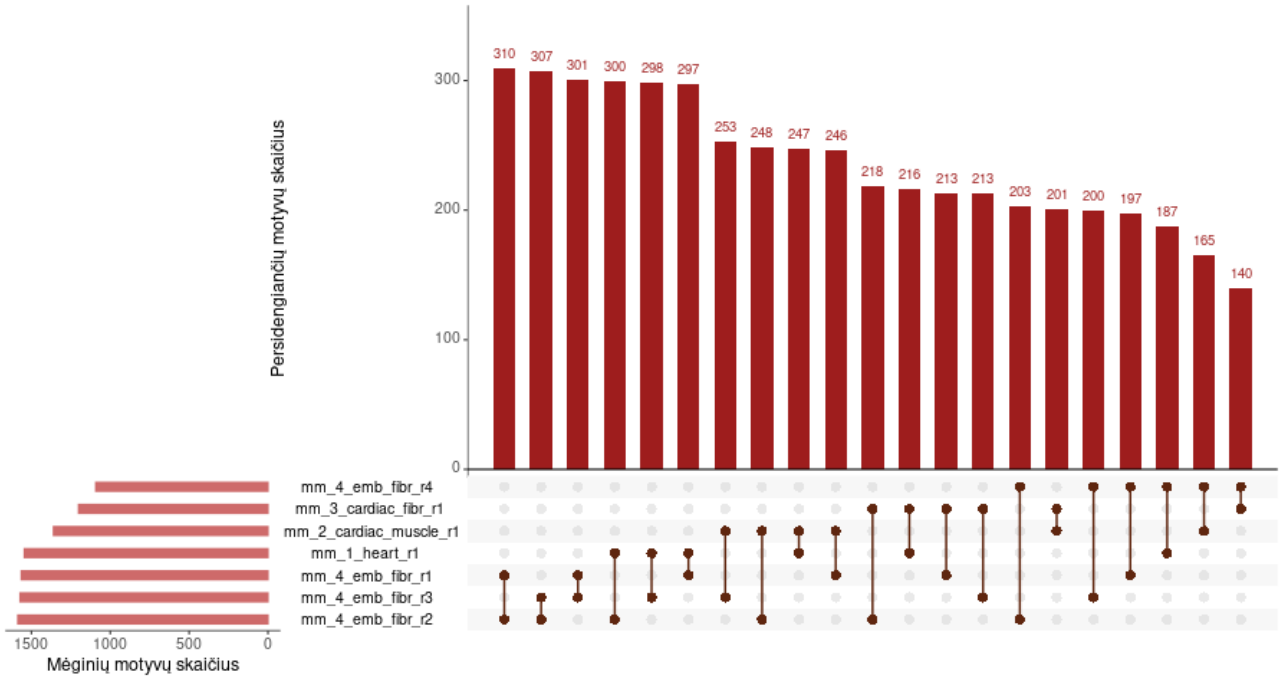


6 pav. Identifikuotų motyvų skaičiaus mėginiuose stulpelinė diagrama

Nėra neįprasta, kad paskutinių mėginių (*mm_4_emb_fibr_r4* ir *mm_3_cardiac_fibr_r1*) identifikuotų motyvų skaičius yra pats mažiausias - šie mėginiai turi mažiausią regionų skaičių bei mažiausią juos atitinkančių sekų rinkinį.

Nepaisant šio atitikimo, mėginys *mm_1_heart_r1*, turintis mažiausią regionų skaičių po *mm_4_emb_fibr_r* ir *mm_3_cardiac_fibr_r1* mėginių, turi didžiausią identifikuotų motyvų skaičių (345 motyvai). Nustatyti motyvai yra ∼12 nukleotidų ilgio.

Toliau pateikiamame grafike (7 pav.) vaizduojama, kiek vienodų motyvų buvo nustatyta skirtinguose mėginiuose.



7 pav. Tarp mėginių persidengiančių motyvų kompleksinė diagrama

Remiantis gautu grafiku, galima pastebėti, kad tarp mėginių porų buvo nustatytas didelis visiems mėginiams būdingų motyvų skaičius. Patikrinus, kiek motyvų aptinkami visuose mėginiuose, buvo nustatyta, kad 131 motyvas yra būdingas visiems analizuojamiems mėginiams.

4.6 *De novo* nustatytų motyvų biologinės funkcijos

Išsiaiškinus, kiek motyvų pateiktų mėginių regionų failuose buvo identifikuota, buvo nustatytos su šiais motyvais susijusių genų funkcijos.

Pateiktame funkcijų sąraše akronimas **MF** nurodo molekulinę funkciją, **BP** - biologinį procesą, **CC** - ląstelės komponentą. Naudojantis UniProtKB[21] duomenų bazės rezultatais, visi trečioje lentelėje (3 lentelė) pateikti motyvai:

- **MF:** atlieka jungimosi prie DNR funkciją bei dalyvauja transkripcijos procese.
- **BP:** dalyvauja ląstelių diferenciacijos procesuose.
- **CC:** yra aptinkami branduolyje.

Unikalios motyvų funkcijos aprašytos sąraše:

1. **Tbx6(T-box)[23] - T-box transkripcijos faktorius 6**
    - **BP:** dalyvauja ląstelių proliferacijos ir organizacijos procesuose, signalinių kelių valdyme, kardioblastų diferenciacijoje.
    - **CC:** nėra prisijungęs prie membranų, aptinkamas branduolyje.
2. **Tbet(T-box)[24] - T-box transkripcijos faktorius 21**
    - **BP:** dalyvauja imuninės sistemos, baltymų metabolinių kelių procesuose. Taip pat pasireiškia organizmui reaguojant į dirgiklius.
    - **CC:** aptinkamas branduolyje, neuroninių ląstelių kūne.
3. **Eomes(T-box)[25] - Eomesoderminas**
    - **BP:** dalyvauja įvairių ląstelių (pavyzdžiui, kardiomiocitų) diferenciacijoje, neurogenezėje, kamieninių ląstelių populiacijos palaikyme
    - **CC:** aptinkamas branduolyje, chromatine.
4. **Mef2b(MADS)[26] - Miocitams specifiškas transkripcijos aktyvatorius 2B**
    - **BP:** dalyvauja įvairių ląstelių diferenciacijoje, transkripcijos proceso aktyvavime.
    - **CC:** aptinkamas citozolyje, branduolyje, nukleoplazmoje, ląstelių jungtyse.
5. **Mef2c(MADS)[27] - Miocitams specifiškas transkripcijos aktyvatorius 2C**
    - **BP:** dalyvauja ląstelių apoptozėje, kraujagyslių formavimęsi, pradinės embrionų širdies vystymęsi.
    - **CC:** aptinkamas citozolyje, branduolyje, nukleoplazmoje, sarkomerose, sarkoplazmoje.
6. **Mef2d(MADS)[28] - Miocitams specifiškas transkripcijos aktyvatorius 2D**
    - **BP:** dalyvauja suaugusių organizmų širdies vystymęsi, kremzlinių bei kaulinių ląstelių diferenciacijoje.
    - **CC:** aptinkamas citoplazmoje, branduolyje, nukleoplazmoje.
7. **TRPS1(Zf)[29] - Cinko „pirštelio" transkripcijos faktorius**
    - **BP:** dalyvauja kremzlinių ląstelių diferenciacijoje, skeleto vystymęsi, būdingas neigiamas transkripcijos reguliavimas.
    - **CC:** aptinkamas branduolyje, nukleoplazmoje, chromatine, baltymų kompleksuose.
8. **GATA3(Zf)[30] - T ląstelėms specifiškas transkripcijos faktorius GATA-3**

- **BP:** dalyvauja aortos vožtuvų formavimęsi, širdies prieširdžių morfogenezėje, embrionų organų vystymęsi, eritrocitų diferenciacijoje.
- **CC:** aptinkamas branduolyje, nukleoplazmoje, chromatine.

9. **GATA4(Zf)[31] - Transkripcijos faktorius GATA-4**
    - **BP:** dalyvauja širdies ląstelių diferenciacijoje, širdies raumens regeneracijoje, embrionų širdies formavimęsi.
    - **CC:** aptinkamas branduolyje, nukleoplazmoje, chromatine.
10. **Fos(bZIP)[32] - AP-1 transkripcijos faktoriaus subvienetas**
    - **BP:** dalyvauja atsako į jonus (kadmio, kalcio), citokinus bei progesteroną procesuose. Taip pat aktyvus nervų sistemos vystymosi metu.
    - **CC:** aptinkamas citozolyje, branduolyje, nukleoplazmoje, endoplazminiame tinkle, sinaptosomose.
11. **Fra1(bZIP)[33] - Onkogenas, AP-1 transkripcijos faktoriaus subvienetas**
    - **BP:** dalyvauja ląstelės ciklo valdyme, apoptoziniuose procesuose bei embrionų vystymosi gimdoje procesuose.
    - **CC:** aptinkamas citozolyje, branduolyje, nukleoplazmoje, presinapsinėje membranoje.
12. **Fra2(bZIP)[34] - AP-1 transkripcijos faktoriaus subvienetas**
    - **BP:** dalyvauja teigiamoje fibroblastų proliferacijoje, atsako į estradiolio - moteriško lytinio hormono - procesuose. Taip pat būdingas teigiamas transkripcijos reguliavimas.
    - **CC:** aptinkamas branduolyje ir nukleoplazmoje.

5 Išvados

Atlikus Tbx5 transkripcijos faktoriaus analizę su skirtingomis naminės pelės ląstelėmis, kurioms buvo taikyti skirtingi poveikiai, gauti rezultatai buvo apibendrinti:

- Didžiausias praturtintų genominių regionų skaičius nustatytas naminės pelės embrionų fibroblastų ląstelėse (163190 regionų), kurios veiktos AGHMT. Mažiausias praturtintų genominių regionų skaičius nustatytas pelių naujagimių širdies fibroblastų, veiktų sb431542 ir xav939 inhibitoriais, linijoje (13440 regionai).

- Skirtingose genominėse pozicijose (chromosomose) Tbx5 faktoriaus regionų pasiskirstymas yra susijęs su naminių pelių chromosomų dydžiais. Didžiausias Tbx5 regionų skaičius nustatytas pirmoje ir antroje chromosomose, kurios yra sudarytos iš daugiausiai nukleotidų. Mažiausias regionų skaičius nustatytas lytinėse chromosomose - X ir Y, kur Y chromosomoje Tbx5 regionų skaičius neviršijo 22 regionų skaičiaus.

- Didžiausias mėginių persidengimo procentas nustytatas tarp mėginio, kuriame tirti širdies fibroblastai veikti inhibitoriais ir AGHMT veiktų fibroblastų mėginio (81.577%). Taip pat didelis panašumas (81.092%) nustatytas tarp mėginių, kuriuose naminių pelių embrionai veikti pilnu serino/treonino kinazės 1 (Akt1) ir transkripcijos faktorių GATA4, HAND2, MEF2C, TBX5 rinkiniu bei rinkiniu, kuriame nebuvo Akt1, tačiau buvo transkripcijos faktorių.

Literatūra

[1] Kate MacCord, Jane Maienschein (2019) Philosophy of Biology: Understanding regeneration at different scales eLife 8:e46569 https://doi.org/10.7554/eLife.46569

[2] Mehta AS, Singh A. Insights into regeneration tool box: An animal model approach. Dev Biol. 2019 Sep 15;453(2):111-129. doi: 10.1016/j.ydbio.2019.04.006. Epub 2019 Apr 13. PMID: 30986388; PMCID: PMC6684456.

[3] GTRD: an integrated view of transcription regulation. Kolmykov S, Yevshin I, Kulyashov M, Sharipov R, Kondrakhin Y, Makeev VJ, Kulakovskiy IV, Kel A, Kolpakov F Nucleic Acids Res. 2021 Jan 8;49(D1):D104-D111.

[4] UCSC Genome Browser: Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006.

[5] He A, Kong SW, Ma Q, Pu WT. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc Natl Acad Sci U S A. 2011 Apr 5;108(14):5632-7. doi: 10.1073/pnas.1016959108. Epub 2011 Mar 17. PMID: 21415370; PMCID: PMC3078411.

[6] Hashimoto H, Wang Z, Garry GA, Malladi VS, Botten GA, Ye W, Zhou H, Osterwalder M, Dickel DE, Visel A, Liu N, Bassel-Duby R, Olson EN. Cardiac Reprogramming Factors Synergistically Activate Genome-wide Cardiogenic Stage-Specific Enhancers. Cell Stem Cell. 2019 Jul 3;25(1):69-86.e5. doi: 10.1016/j.stem.2019.03.022. Epub 2019 May 9. PMID: 31080136; PMCID: PMC6754266.

[7] Stone NR, Gifford CA, Thomas R, Pratt KJB, Samse-Knapp K, Mohamed TMA, Radzinsky EM, Schricker A, Ye L, Yu P, van Bemmel JG, Ivey KN, Pollard KS, Srivastava D. Context-Specific Transcription Factor Functions Regulate Epigenomic and Transcriptional Dynamics during Cardiac Reprogramming. Cell Stem Cell. 2019 Jul 3;25(1):87-102.e9. doi: 10.1016/j.stem.2019.06.012. PMID: 31271750; PMCID: PMC6632093.

[8] Jackson Laboratory (RRID:SCR_004633).

[9] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[10] Scikick. Utility for executing collections of computational notebooks.
URL https://petronislab.camh.ca/pub/scikick/stable/docs/report/out_html/introduction.html

[11] M. Lawrence, R. Gentleman, V. Carey: "rtracklayer: an R package for interfacing with genome browsers". Bioinformatics 25:1841-1842.

[12] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[13] BigWig and BigBed tools: Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed data sets. Bioinformatics. 2010 Sep 1;26(17):2204-7.

[14] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841-2. doi: 10.1093/bioinformatics/btq033. Epub 2010 Jan 28. PMID: 20110278; PMCID: PMC2832824.

[15] BEDTools komandinės eilutės įrankis. Programų rinkinio programa *getfasta*.
Prieiga per https://bedtools.readthedocs.io/en/latest/content/tools/getfasta.html [žiūrėta 2022-06-03].

[16] Pagès H, Aboyoun P, Gentleman R, DebRoy S (2022). _Biostrings: Efficient manipulation of biological strings_. R package version 2.64.0, <https://bioconductor.org/packages/Biostrings>.

[17] HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis Ivan V. Kulakovskiy; Ilya E. Vorontsov; Ivan S. Yevshin; Ruslan N. Sharipov; Alla D. Fedorova; Eugene I. Rumynskiy; Yulia A. Medvedeva; Arturo Magana-Mora; Vladimir B. Bajic; Dmitry A. Papatsenko; Fedor A. Kolpakov; Vsevolod J. Makeev Nucl. Acids Res., Database issue, gkx1106 (11 November 2017) doi: 10.1093/nar/gkx1106

[18] Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589. PMID: 20513432.

[19] Jake R. Conway, Alexander Lex, Nils Gehlenborg. UpSetR: An R Package For The Visualization Of Intersecting Sets And Their Properties Bioinformatics, 33(18): 2938-2940, doi:10.1093/bioinformatics/btx364, 2017.

[20] The UniProt Consortium UniProt: the universal protein knowledgebase in 2021 Nucleic Acids Res. 49:D1 (2021).

[21] Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot Methods Mol. Biol. 406:89-112 (2007)

[22] Drowley L, Koonce C, Peel S, et al. Human Induced Pluripotent Stem Cell-Derived Cardiac Progenitor Cells in Phenotypic Screening: A Transforming Growth Factor-β Type 1 Receptor Kinase Inhibitor Induces Efficient Cardiac Differentiation. Stem Cells Transl Med. 2016;5(2):164-174. doi:10.5966/sctm.2015-0114

[23] UniProtKB duomenų bazė. *Tbx6 aprašymas* (2022).
Prieiga per https://www.uniprot.org/uniprot/P70327 [žiūrėta 2022-06-12].

[24] UniProtKB duomenų bazė. *Tbet aprašymas* (2022).
Prieiga per https://www.uniprot.org/uniprot/Q9JKD8 [žiūrėta 2022-06-12].

[25] UniProtKB duomenų bazė. *Eomes aprašymas* (2022).
Prieiga per https://www.uniprot.org/uniprot/O54839 [žiūrėta 2022-06-12].

[26] UniProtKB duomenų bazė. *Mef2b aprašymas* (2022).
Prieiga per https://www.uniprot.org/uniprot/O55087 [žiūrėta 2022-06-12].

[27] UniProtKB duomenų bazė. *Mef2c aprašymas* (2022).
Prieiga per https://www.uniprot.org/uniprot/Q8CFN5 [žiūrėta 2022-06-12].

[28] UniProtKB duomenų bazė. *Mef2d aprašymas* (2022).
Prieiga per https://www.uniprot.org/uniprot/Q63943 [žiūrėta 2022-06-12].

[29] UniProtKB duomenų bazė. *TRPS1 aprašymas* (2022).
Prieiga per https://www.uniprot.org/uniprot/Q925H1 [žiūrėta 2022-06-12].

[30] UniProtKB duomenų bazė. *GATA3 aprašymas* (2022).
Prieiga per https://www.uniprot.org/uniprot/P23772 [žiūrėta 2022-06-12].

[31] UniProtKB duomenų bazė. *GATA4 aprašymas* (2022).
Prieiga per https://www.uniprot.org/uniprot/Q08369 [žiūrėta 2022-06-12].

[32] UniProtKB duomenų bazė. *Fos aprašymas* (2022).
Prieiga per https://www.uniprot.org/uniprot/P01101 [žiūrėta 2022-06-12].

[33] UniProtKB duomenų bazė. *Fra1 aprašymas* (2022).
Prieiga per https://www.uniprot.org/uniprot/P48755 [žiūrėta 2022-06-12].

[34] UniProtKB duomenų bazė. *Fra2 aprašymas* (2022).
Prieiga per https://www.uniprot.org/uniprot/P47930 [žiūrėta 2022-06-12].

6 Priedas

Priedų sąraše pateikiamos tarpinių rezultatų puslapio, sugeneruoto su Scikick, bei Git repozitorijos, kurioje saugomi analizei naudoti duomenų failai, parašyti skriptai bei pagrindinė R programa, nuorodos.

- **Tarpinių rezultatų Scikick puslapis:**
  https://karklas.mif.vu.lt/~dast6577/KursinisDarbas/v1.1/peaks_MM.html
- **Analizės Git repozitorija:**
  https://github.com/dansta0804/TF_analysis.git