**რეფერატი**

ახალი ზელანდიის მღვიმეებსა და მდინარეების ტყით დაფარულ ხეობებში შებინდებისას ადვილად შესამჩნევია ვარსკვლავის მსგავსად მანათობელი არსებები. ადგილობრივებში ციცინათელას სახელით ცნობილი მწერის სრული მეცნიერული დასახელებაა Arachnocampa Luminosa. მისი მანათობელი თვისება გამოწვეულია კუდის ბოლოს მოთავსებული პატარა ორგანოს მიერ. იგი პატარა მწერებით იკვებება, რომლებსაც სინათლე იზიდავთ, მასთან მიახლოვების შემდეგ კი arachnocampa-ს მიერ დაკიდებულ, წებოვან აბრეშუმის ძაფებში ეხვევიან.

საინტერესო ისაა, რომ მათი სახელის მიუხედავად, ციცინათელები მწერების ერთ-ერთ ტიპს წარმოადგენენ - (რიგი Diptera, ოჯახი Keroplatidae). თუმცა კარგად შესწავლილი მანათობელი ხოჭოები სხვა რიგის წარმომადგენლები არიან - Coeloptera (Elateroidea-ს სუპეროჯახი). დიპტერასა და კოლეოპტერას გზები ევოლუციის პროცესში 330 მილიონი წლის წინ გაიყო და მას შემდეგ მათ შორის არანაირი თანაკვეთა არაა ცნობილი, რაც მიუთითებს რომ ამ სახეობებში ეს თვისება დამოუკიდებლად განვითარდა.

საზოგადოდ, ბიოლუმინესცენციის განვითარება რთული და კომპლექსური პროცესი იყო. ერთ-ერთი ახალი კვლევის თანახმად, ამ თვისებამ ჩვენამდე მოღწეულ ორგანიზმებში მინიმუმ 40-ჯერ განიცადა ცვლილება, შესაძლოა 50-ზე მეტჯერაც. მიუხედავად განსხვავებებისა, ყველა ბიოლუმინესცენტური სისტემა სინათლეს გამოყოფს სინათლის გამომყოფელი სუბსტრატის დაჟანგვით. სუბსტრატს უწოდებენ ლუციფერინს, ამ რეაქციის კატალიზატორი კი არის ენზიმი, ლუციფერაზა.

კონკრეტულად არაქნოკამპას შემთხვევაში ძალიან ბევრი დეტალი არის ბუნდოვანი, მათ შორის გადაუჭრელი რჩება კონკრეტულად ამ ლუციფერაზის და ლუციფერინის იდენტიფიკაციის პრობლემა. მიუხედავად იმისა რომ სხვა სახეობების მსგავსად, ციცინათელებშიც ატფ გამოიყენება ბიოლუმინესცენციისთვის, ქიმიური მექანიზმები მათში განსხვავებულია. მაგალითად, როცა firefly სისტემის ლუციფერინი და glowworm სისტემის ლუციფერაზა ერთმანეთს შეურიეს, სინათლე არ წარმოიქმნა. რაც იმაზე მეტყველებს, რომ ისინი განსხვავებულ რეაგენტებსა და კატალიზატორებს იყენებენ. ასევე უნიკალურია ზემოაღნიშნული ორგანო ფიზიოლოგიურად. იგი შედგება ოთხი მალპიგის მილაკის წვერისგან. მალპიგის ტუბულები მწერების გამომყოფი სისტემის ნაწილს წარმოადგენენ, და რაც ყველაზე საინტერესოა, არ მონაწილეობენ არც ერთი სხვა მწერის ბიოლუმინესცენციაში.

**ექსპერიმენტი**

სექვენირებისთვის გამოყენებული იყო მოწყობილობა HiSeq-2000. შედეგად ჯამში 6 cDNA ბიბლიოთეკა მიიღეს, რომლიდანაც 3 იყო მანათობელი ორგანოდან, 3 კი - არამანათობელი.

**პრეპროცესინგი**

პირველ რიგში, სანამ უშუალოდ მონაცემების ესემბლინგზე გადავიდოდით, ზემოხსენებულ “ბიბლიოთეკებს”, რომლებიც რეალურად ძალიან დიდი რაოდენობის ნუკლეოტიდების მიმდევრობებს, read-ებს წარმოადგენენენ, პრეპროცესინგი ჩაუტარდა.

პირველ რიგში, მონაცემები გავასუფთავეთ ე.წ. ადაპტერ მიმდევრობებისგან. ეს არის პატარა დნმ მიმდევრობები, რომლებიც სექვენირების პროცესში მის დაწყებამდე ხელოვნურად არიან ჩართული და რამდენიმე მიზანს ემსახურებიან, როგორებიცაა ბიბლიოთეკის მომზადება (ისინი ემაგრებიან დნმ ფრაგმენტების ბოლოებს და შეიცავენ ისეთ მიმდევრობებს, რომლებიც საჭიროა სექვენირების პლატფორმასთან მისამაგრებლად და ამ პროცესის დასაჩქარებლად), ასევე ფრაგმენტების იდენტიფიკაცია (თითოეული read-ის ბოლოებზე ადაპტერ მიმდევრობები გვაწვდიან უნიკალურ მაიდენტიფიცირებლებს, რაც საბოლოოდ პლატფორმას აძლევს საშუალებას ისინი გამოარჩიოს და სხვებისგან განასხვავოს. ასევე ისინი უერთდებიან პრაიმერებს, რაც სექვენირების ინიციალიზაციისთვის ძალიან მნიშვნელოვანია და ა.შ.

თუმცა ადაპტერ მიმდევრობები ზემოხსენებული მიზეზის გამო არ არიან საწყისი, ორიგინალი გენომის ნაწილი, ამიტომ მონაცემებში ისინი უბრალოდ ამოიშალა. ადაპტერ მიმდევრობებად გამოვიყენეთ ორი მიმდევრობა, რომელსაც ილუმინა ზოგადად იყენებს მუშაობის პროცესში (PreProcessing/adapters.fa).

გარდა ამისა, “მოიჭრა” დაბალი ხარისხის ფუძეები, ანუ რომელთა Phred სქორი იყო 20-ზე ნაკლები. ეს მნიშვნელოვანი ნაბიჯია, რადგან შეფასების ეს სისტემა გამოიყენება ფუძეების “სანდოობის” შესაფასებლად. დაბალი “სანდოობის” ფუძეებისგან გათავისუფლება მონაცემებს უფრო მეტ სიზუსტეს სძენს.

ფორმულა, რომლითაც Phred Score გამოითვლება, შემდეგნაირად გამოიყურება:

Phred Score (Q) = -10 \* log10(P) , სადაც Q არის Phred Score, ხოლო P იმის ალბათობა, რომ ფუძე შეიძლება არასწორად იყოს “გამოცნობილი”. თავად ამ ალბათობის გამოთვლა სექუენირების პროცესში ხდება და სპეციფიკურია მოწყობილობისა და ალგორითმის მიხედვით, ძირითადად სიგნალების ინტენსიურობას ეფუძნება.

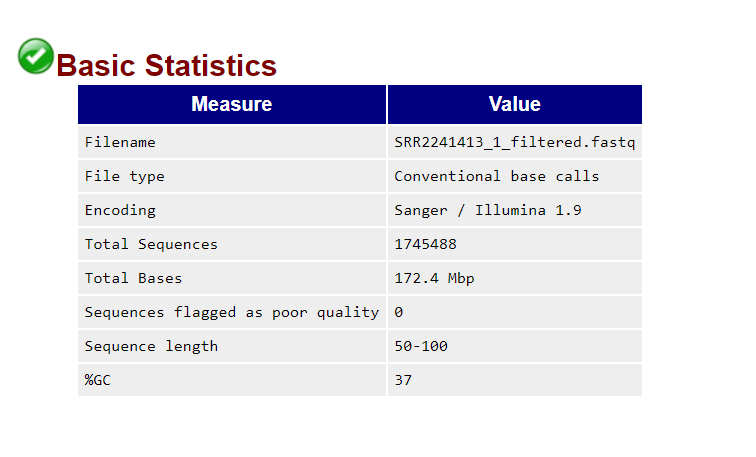
Phred Score = 20 ნიშნავს, რომ არასწორი ფუძის ალბათობა არის 1%. რაც უფრო იზრდება Phred Score, იზრდება სანდოობაც. მაგალითად, 30-სთვის შეცდომის ალბათობა 0.1%-ა და ა.შ. შესაბამისად, ფაქტობრივად, ამოვარდა ისეთი ფუძეები, რომლებისთვისაც შეცდომის ალბათობა იყო 1%-ზე მეტი.

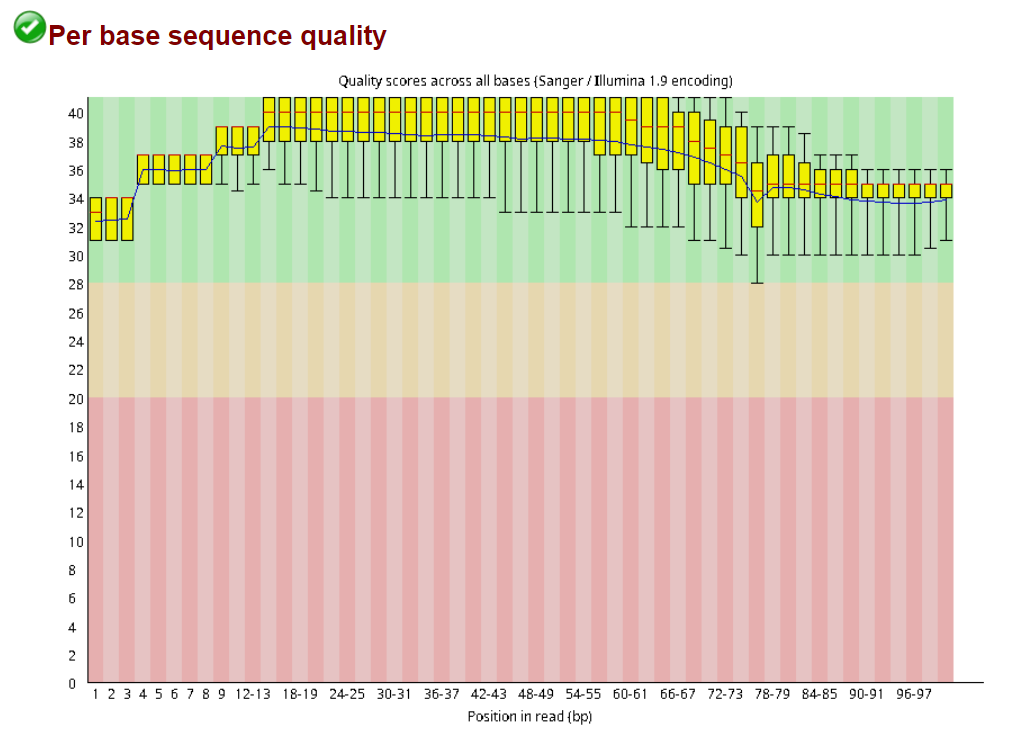
ამის გარდა, ამოვარდა ისეთი read-ებიც, რომელთა სიგრძე რომელიმე სთრენდზე მაინც 50 ნუკლეოტიდზე ნაკლები იყო.

აღსანიშნავია, რომ ექსპერიმენტის მონაცემების ლეიაუთი იყო PAIRED, ანუ რიდები გაყოფილი იყო ორ ფაილში დნმ ჯაჭვების მიხედვით. შესაბამისად, პრეპროცესინგის ნაბიჯები ორივე ფაილისთვის ერთდროულად უნდა მომხდარიყო, წინააღმდეგ შემთხვევაში მივიღებდით შედეგებს, რომ ერთი და იგივე ექსპერიმენტის საბოლოო შედეგი ორი ჯაჭვისთვის განსხვავებული იქნებოდა.

ადაპტერების ამოსაგდებად და ფრედ სქორების მიხედვით გასაფილტრად გამოვიყენეთ fastq-mcf პაკეტი, ხოლო საბოლოოდ სიგრძის მიხედვით გასაფილტრად - fastp (PreProcessing/trim\_adapters.sh, PreProcessing/discard\_shorts.sh).

ასევე, ამ დამუშავებულ მონაცემებზე გავუშვით ე.წ. Quality Assessment, რომლისთვისაც fastqc გამოვიყენეთ. (PreProcessing/quality\_assessment.sh).

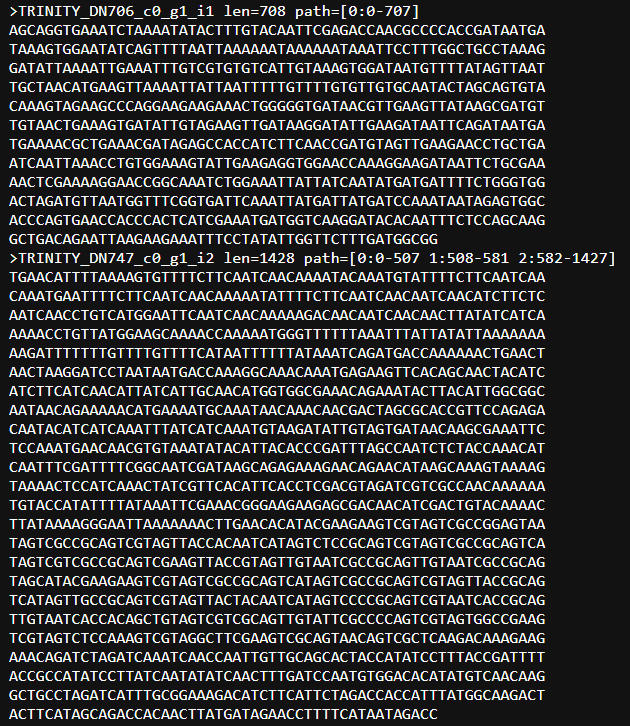




**Assembling**

არაქნოკამპა ლუმინოსასთვის არ არსებობს ე.წ. რეფერენს გენომი, შესაბამისად ესემბლინგისთვის დე ნოვო მიდგომა უნდა გამოგვეყენებინა. გაფილტრულ და დამუშავებულ მონაცემებზე გავუშვით Trinity სოფთვეარი, რომელმაც საბოლოო შედეგები დაგვიბრუნა .fasta ფორმატში.

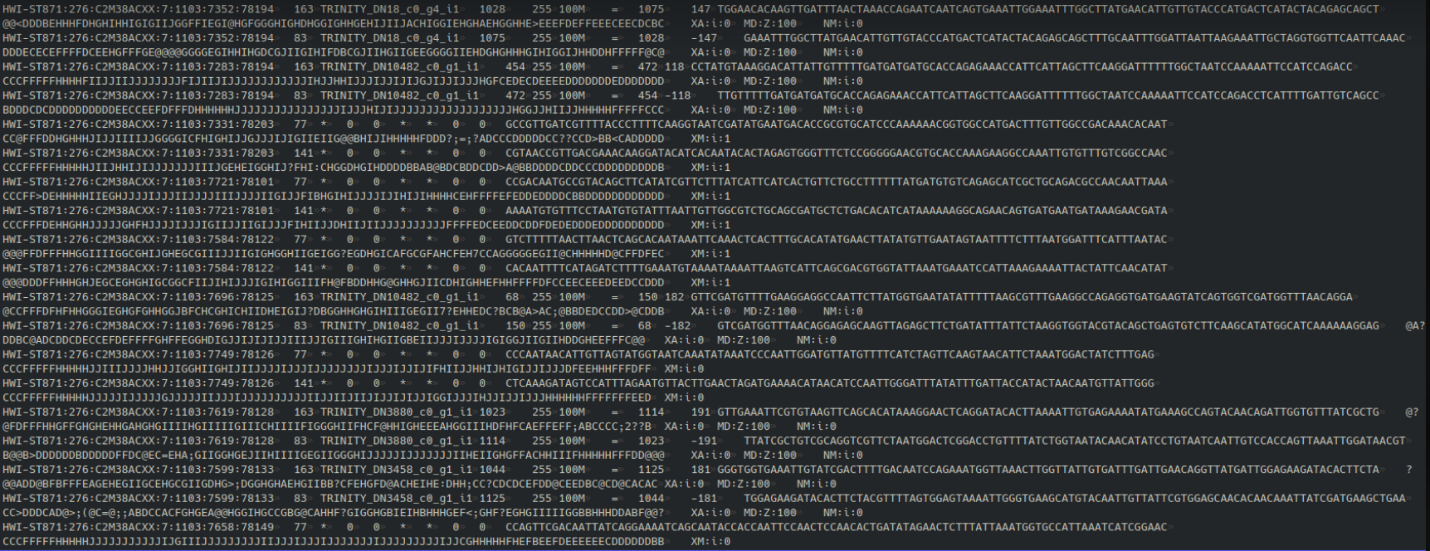
(“assemble and map”/assemble\_trinity.sh).



**Mapping**

ამის შემდეგ, მოხდა თითოეული რიდის მეპინგი Bowtie -ის გამოყენებით. აქ უნდა აღინიშნოს, რომ თავად კვლევაში Bowtie 2 იყო მოხსენიებული. საწყის ეტაპზე ჩვენც Bowtie 2 გამოვიყენეთ, თუმცა შედეგად გეფებიანი ელაინმენთები მივიღეთ, რამაც ტრანსკრიპტების სიხშირეების და სტატისტიკების გამოთვლის შემდეგ ეტაპზე პრობლემები გამოიწვია (RSEM პაკეტს არ აქვს გეფებიანი ელაინმენთების მხარდაჭერა). შესაბამისად, გადავედით Bowtie–ზე, რამაც დამეპილი მიმდევრობების წილი შეამცირა, მაგრამ სამაგიეროდ ზემოხსენებული პრობლემა გააქრო. Bowtie 2 -ით საშუალოდ 95% იმეპებოდა წარმატებით, Bowtie-ით კი 65-70%. (მათ შორის მთავარი განსხვავებაც ისაა, რომ Bowtie 2 გეფებსაც ითვალისწინებს).

ამ ნაბიჯის პასუხი დაგვიბრუნდა sam ფორმატში. შემდეგი ნაბიჯისთვის გვჭირდებოდა მისი binary ვერსია, ამიტომ გადავაკონვერტირეთ .bam ტიპის ფაილებში.



(“assemble and map”/map\_using\_bowtie.sh).

**შედეგების სტატისტიკური ანალიზი**

**Transcript Abundancy**

პირველ რიგში დავთვალეთ ტრანსკრიპტების სიხშირეები RSEM პაკეტის გამოყენებით. შედეგები არის ასახული “\*\_rsem\_output\_genes.results” ტიპის ფაილებში ცხრილის სახით. გენისა და ტრანსკრიპტის უნიკალურ აიდისთან ერთად, სვეტებად გვაქვს სიგრძეები, სიხშირეები და ასევე TPM და FPKM.

TPM (Transcripts Per Million) არის ნორმალიზაციის მეთოდი, რომელიც მხედველობაში იღებს ტრანსკრიპტის სიგრძესაც. იგი ისეთი რიდების რაოდენობას, რომელიც ამ კონკრეტულ ტრანსკრიპტზე დაიმეპა, ყოფს ტრანსკრიპტის სიგრძეზე და მილიონზე ამრავლებს. ოღონდ, აქ საუბარია ე.წ. effective სიგრზეზე. ეს არის სიგრძის მიახლოება, რომელიც ითვალისწინებს სექვენირების პროცესში ბაიასს და მის კომპენსირებას ცდილობს. TPM გენების ექსპრესიის დონეზე სამსჯელოდ უკეთესი საშუალებაა, ვიდრე უბრალოდ სიხშირეების დათვლა, რადგან იგი ბიბლიოთეკისა და ტრანსკრიპტის ზომებში არსებულ განსხვავებასაც ითვალისწინებს.

რაც შეეხება FPKM (Fragments Per Kilobase Milion), ისიც ნორმალიზაციის მეთოდია, ოღონდ შედარებით უფრო ძველი. მათ შორის განსხვავება არის დასკალირებაში. FPKM ასკალირებს read-ების სიხშირეებს ტრანსკრიპტის სიგრძეზე დაყრდნობით, და შემდეგ ახდენს მათ ნორმალიზებას. ამ შემთხვევაში ფორმულა შედარებით უფრო ჩახლართულია: წინა ფორმულისგან განსხვავებით, აქ ტრანსკრიპტზე დამეპილი რიდების რაოდენობა ჯერ იყოფა ტრანსკრიპტის სიგრძის მეათასედზე, შემდეგ შედეგი მრავლდება საერთო ჯამში დამეპილი რიდების რაოდენობაზე და იყოფა მილიონზე.

(transcript\_abundancy/rsem.sh, transcript\_abundancy/\*\_rsem\_output.genes.results)

გარდა RSEM-ისა, ასევე გამოვიყენეთ სხვა ფექიჯიც ანალიზისთვის - salmon, მათ შორის განსხვავება ისაა, რომ salmon-ს მეპინგსაც თავად აკეთებს და არ სჭირდება Bowtie -ს აუთფუთი. იგი პირდაპირ Trinity-ს შედეგზეც შეგვიძლია გავუშვათ.

(transcript\_abundancy/salmon.sh, transcript\_abundancy/salmon\_analysis\_output\_\*)

ასევე გამოვიყენეთ ნორმალიზების TMM (trimmed mean of M values) მეთოდი, რომელმაც საშუალება მოგვცა, რომ დაგვეთვალა მონაცემების საშუალო ისე, რომ ექსტრემალურ მნიშვნელობებს ნაკლები გავლენა მოეხდინა შედეგზე.

ამის დასათვლელად გამოვიყენეთ edgeR package, რომელიც იმპლემენტირებულია R-ში.

ამისთვის მივყევით შემდეგ ნაბიჯებს:

1. გამოვითვალეთ თითოეული სემფლის ბიბლიოთეკის ზომა.

2. გამოვითვალეთ საშუალო გეომეტრიული ყველა სემფლისთვის.

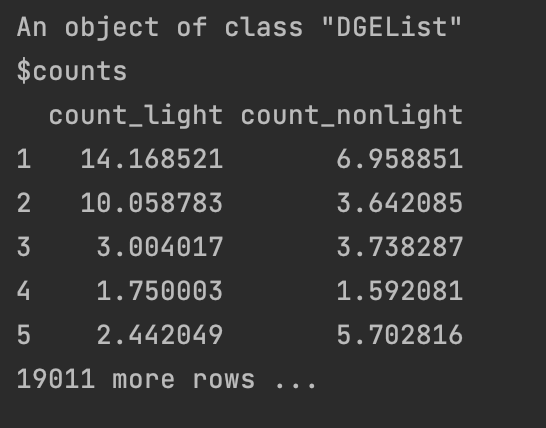
3. (log-ratio)ლოგარითმი(expression value/ საშუალო გეომეტრიული). 4.გამოვთვალეთ M value-ები შემდეგნაირად i-ური გენისთვის j sample ში:

M = |log-ratio(i, j) - median(log-ratio i across all j samples)|.

5. გამოვაკელით ექსტრემალური მნიშვნელობები

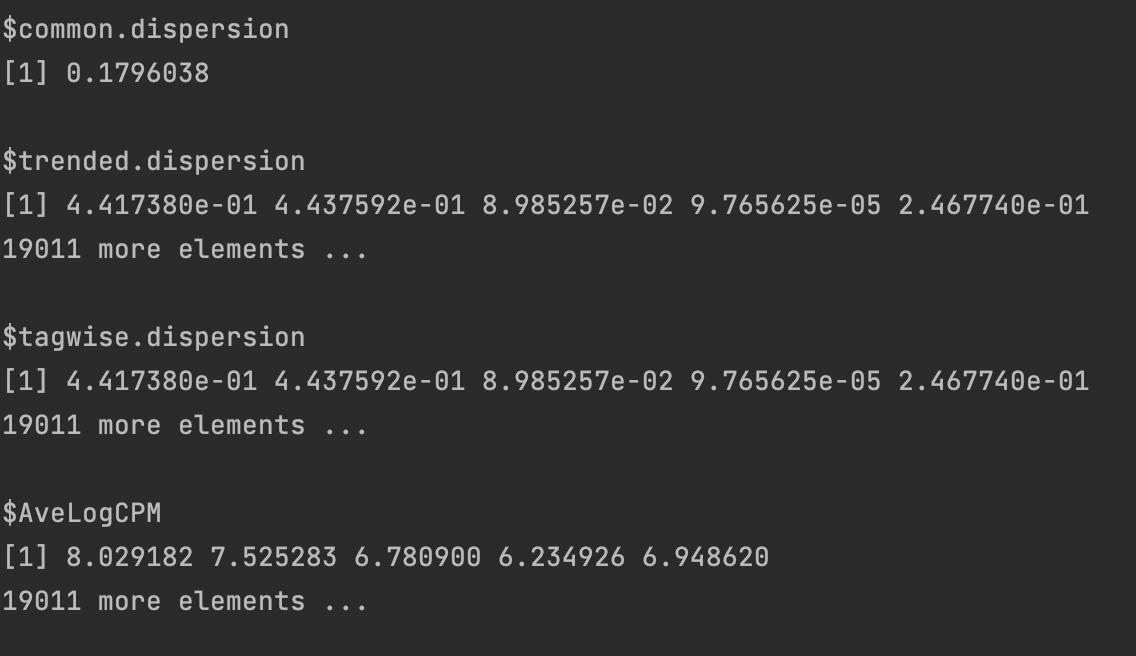
6.სკალირების ფაქტორის გამოთვლა და დასკალირება.

ეს მეთოდი მნიშნელოვანია განსაკუთრებით რნმ ტრანსკრიპტების ანალიზისთვის რომლებიც სხვადასხვა სემფლებიდან არის მიღებული და მათგან სექუენირებულ ბიბლიოთეკებს აქვთ სხვადასხვა ზომა. (მოკლედ რომ ვთქვათ, bias-ს ამცირებს). კოდი მოცემულია TMM.R ფაილში.



ანალიზისთვის ასევე გამოვიყენეთ edgeR package-ის მეთოდები, რათა შეგვეფასებინა დისპერსია მოცემული მონაცემებისთვის:

ამ შემთხვევაში დისპერსია თითოეული გენის შემთხვევაში წარმაოდგენს თუ რამდენადაა გადახრილი მოცემული გენის ექსპერსია მისი მოსალოდნელი ექსპრესიის დონისგან.

როგორც სურათიდან ჩანს Common dispersion ჩვენი მონაცემებისთვის არც ისე მაღალია რაც მიუთითებს, რომ ვარიაცია მონაცემების დაბალია, ეს კი იმის მანიშნებელია, რომ მონაცემები არ არის  შემთხვევითად დაგენერირებული და გენების ექსპრესიებს შორის განსხვავებას აქვს ბიოლოგიური შინაარსი.

Trended dispersion და tagwise dispersion არის common dispersion-ის გავრცობა და გვაძლევს ინფორმაციას თითოეული გენის ჭრილში.

**Differential Expression Analysis**

ბოლო საფეხურზე ჩავატარეთ დიფერენციალური ექსპრესიის ანალიზი. გვქონდა RSEM-ის შედეგები ოთხ სხვადასხვა ექსპერიმენტზე, რომელთაგან ორი არის მანათობელი ორგანოდან მიღებული, ორი - არამანათობელი. ეს მონაცემები გავაერთიანეთ ერთ ცხრილში, პითონში, panda-ს DataFrame-ში. საბოლოო ცხრილი იყო შემდეგი სახის:  
ტრანსკრიპტი —-- სიხშირე მანათობელ ორგანოში —----- სიხშირე არამანათობელ ორგანოში

ამის შემდეგ, ჩავატარეთ ორი განსხვავებული ანალიზი.

1. საწყის ჯერზე საკმაოდ მარტივი და საბაზისო ნაბიჯები გადავდგით. ზემოაღნიშნული ცხრილიდან დავტოვეთ მხოლოდ ის რიდები, რომლებიც მანათობელ ორგანოში უფრო მეტჯერ გამოვლინდნენ, ვიდრე არამანათობელში და შემდეგ დავსორტეთ მათ შორის სხვაობის მიხედვით, კლებადობით. ეს ნაბიჯები არის “differential expression analysis”/simple\_frequency\_analysis.py-ში, ხოლო მისი შედეგები ცხრილის სახით შენახულია იმავე დირექტორიის “sorted.csv” ფაილში.
2. იგივე ტიპის დატაფრეიმზე ჩავატარეთ სტატისტიკური ანალიზი, რომელიც ბენჯამინი-ჰოჩბერგის ანალიიზის სახელწოდებითაა ცნობილი.

ეს პროცედურა გამოიყენება ჰიპოთეზების ტესტირებაში, როცა false discovery rate-ის კონტროლი სჭირდებათ რამდენიმე ჰიპოთეზის გატესტვისას. უფრო კონკრეტულად, ჰიპოთეზების ტესტირების დროს მკვლევარები ხშირად ერთსა და იმავე ტესტს ერთდროულად ატარებენ, რასაც მივყავართ ფოლს პოზიტივის შედეგების წილის ზრდამდე (Type 1 errors). ბენჯამინი-ჰოჩბერგის პროცედურა ამ პრობლემას უმკლავდება False Discovery Rate-ის კონტროლით, რაც მოსალოდნელია რომ false rejection-ების წილის ტოლი იქნება საერთო ჯამში დარეჯექტებულ ჰიპოთეზებს შორის.

ამ პროცედურისთვის პარამეტრად FDR = 0.1 ავიღეთ, როგორც ეს თავად ორიგინალი კვლევაში იყო მოხსენიებული.

პროცედურა შემდეგნაირად მიმდინარეობს:

პირველ რიგში ხდება p-value მნიშვნელობების მიღება და სორტირება ზრდადობით. ეს მნიშვნელობა წარმოადგენს ალბათობას იმისა, რომ ნალ ჰიპოთეზის ჭეშმარიტების შემთხვევაში მსგავსი შედეგი მიგვეღო. ამის შემდეგ ხდება rejection cutoff მნიშვნელობის გამოთვლა. k = (i/m) \* alpha, სადაც i არის p-value-ს რანკი (ინდექსი უკვე დასორტილ მონაცემებში), m ტესტების საერთო რაოდენობა, ხოლო alpha კი FDR-ს შეესაბამება. საბოლოო ეტაპზე კი ხდება იმ ჰიპოთეზების უარყოფა, რომელთათვისაც p-value-ები არის <= k.

კოდი არის “differential expression analysis”/Benjamini-Hochberg\_analysis.py ფაილში, ხოლო საბოლოო შედეგი იმავე დირექტორიის Benjamini-Hochberg\_sorted.csv-ში.

ორივე შემთხვევაში, ცხადია, მიზანი იყო ისეთი ტრანსკრიპტების გამოყოფა, რომელთაც გაცილებით უფრო ხშირად ვხვდებით მანათობელი ორგანოდან მიღებულ მიმდევრობებში, ვიდრე არამანათობელში.

**Blast**

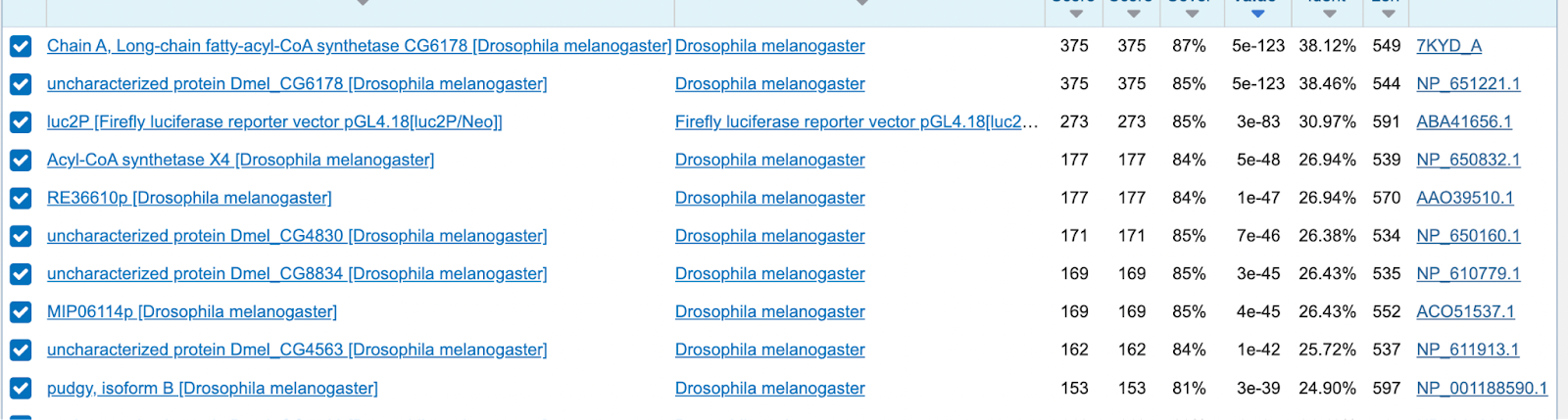
აღნიშნული ანალიზის შედეგად გამოყოფილ ტრანსკრიპტებზე და drosophila melanogaster, firefly luciferase(Firefly luciferase is the light-emitting enzyme responsible for the bioluminescence of fireflies and click beetles) რეფერენს გენომზე გავუშვით ბლასტის ალგორითმი. დროზოფილას რეფერენს გენომის არჩევის მიზანი იყო ის, რომ ამ ორგანიზმსაც გააჩნია სინათლის ციკლი და ლუმინესცენციური თვისებები, თუმცა არაქნოკამპასგან განსხვავებით უფრო მეტადაა გამოკვლეული. ბლასტის ალგორითმით კვლევა გავაგრძელეთ მხოლოდ იმ ტრანსკრიპტებზე რომელიც 1კბ-ზე მეტი იყო სიგრძეში, რათა უშუალოდ არაქნოკამპას სპეციფიურ გენებზე მიგვეღო დამატებითი ინფორმაცია.

ორიგინალი კვლევის შედეგების მიხედვით ტრანსკრიპტებს ყველას გააჩნდათ motifs ANL სუპეროჯახიდან, რომლის 3 მთავარი ტიპია:

* AcylCoA synthetases,
* the NRPS adenylation domains,
* the beetle (firefly) Luciferase enzymes.

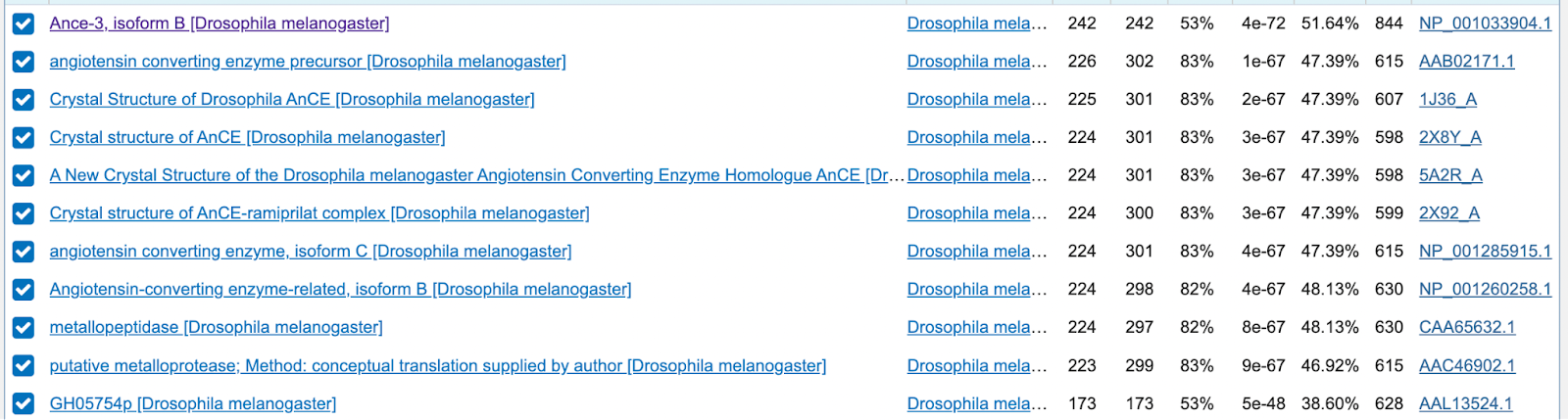
ჩვენი ანალიზის შედეგად ყველაზე მეტად გამოხატული ტრანსკრიპტის ბლასტის შედეგშიც გვხვდება Acyl-CoA, ისევე, როგორც ორიგინალ კვლევაში. მოცემული ენზიმი ბლასტმა შეუსაბამა როგორც დროზოფილას ასევე ლუციფერაზას ორგანიზმში ნაპოვნ Acyl-CoA-ს.

1- TRINITY\_DN10482\_c0\_g1\_i1



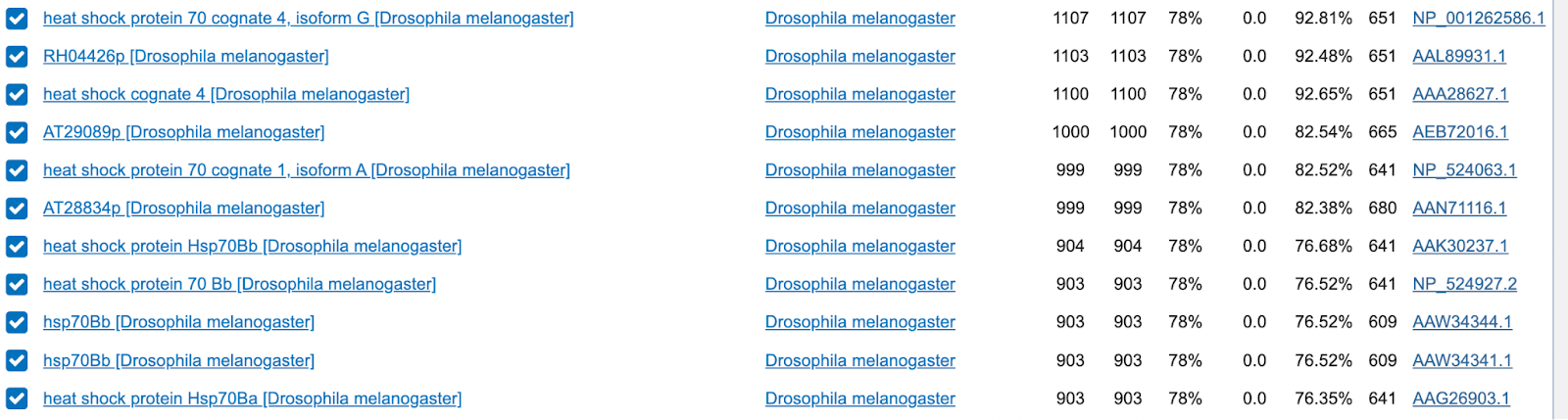
2 - TRINITY\_DN16723\_c0\_g1\_i1

მეორე ყველაზე ხშირი ტრანსკრიპტის ბლასტის შედეგები:



[angiotensin converting enzyme precursor [Drosophila melanogaster]](https://blast.ncbi.nlm.nih.gov/Blast.cgi#alnHdr_AAB02171) - არის ენზიმი, რომლიც სისხლის წნევის და კარდიოვასკულარული ჰომეოსტაზის რეგულაციაში იღებს მონაწილეობას. Ance ენზიმები არ არის შესწავლილი ლუმინისცენციური კუთხით.

3 - TRINITY\_DN362\_c0\_g1\_i1



heat shock protein 70 cognate 4, isoform G [Drosophila melanogaster] - კონსერვირებულია ბევრ ორგანიზმში და პასუხისმგებელია სითბურ სტრესზე რეაქციაზე. ორიგინალ კვლევაშიც აღნიშნული იყო რამდენიმე ენზიმი რომელიც ტემპერატურის ცვლილებაზე რეაგირებდა, როგორც შესაძლო მონაწილე არაქნოკამპას ლუმინისცენციისა.

საერთო ჯამში შევამოწმეთ 10 ყველაზე მეტად გამოხატული ტრანსკრიპტი, რომლებიც მიღებული იყო დიფერენციალური ანალიზით, ბლასტის ალგორითმზე. მიუხედავად იმისა, რომ სხვა თულები გამოვიყენეთ ანალიზისთვის და პრეპროცესინგის ზოგიერთი ეტაპისთვის მაინც შევამჩნიეთ ჩვენს შედეგებში რამდენიმე ძირეული დამთხვევა. მაგალითად: heat shock protein და Acyl-CoA-სთან თანხვედრები. ასევე ზოგადად 1 კბ-ზე მეტი სიგრძის ტრანსკრიპტებიდან დიდი ნაწილი ბლასტმა შეუსაბამა drosophila melanogaster-ის რეფერენს გენომს, რაც ასევე აღნიშნულია ორიგინალი კვლევის შედეგებშიც.