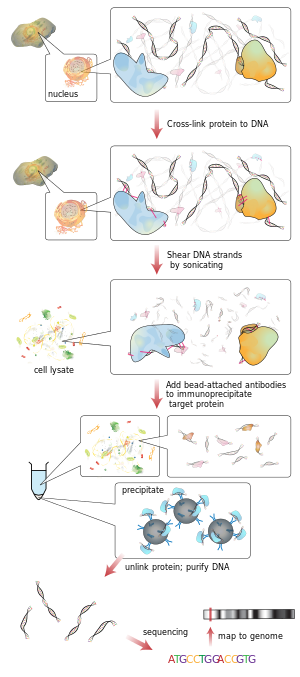
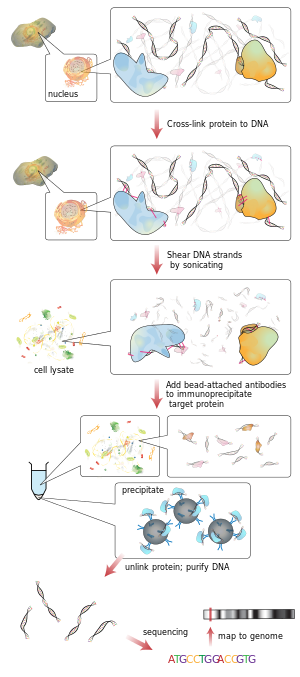
ChIP секвениране

Кирил Кирилов

ChIP-секвенирането, известна още като ChIP-seq, е метод, използван за анализ на протеиновите взаимодействия с ДНК. ChIP-seq комбинира имунопреципитация на хроматин (ChIP) с масивно успоредно ДНК секвениране, за да идентифицира свързващите места на ДНК-асоциираните протеини. Може да се използва за картографиране на глобални сайтове за свързване точно за всеки белтък, който представлява интерес. Преди това ChIP-on-chip беше най-разпространената техника, използвана за изследване на протеин-ДНК взаимоотношения

ChIP-seq се използва главно за определяне как транскрипционните фактори и други свързани с хроматина протеини влияят върху биологичните механизми. Както и определянето на взаимодействието на протеините с ДНК е от съществено значение за пълно разбиране на генната експресия , биологични процеси и болестни състояния.



Изчислителен анализ

ChIP-seq генерира изключително големи набори от данни, за които са необходими подходящи методи за изчислителен анализ. За да се предскаже ДНК-свързващи места от ChIP-seq данни, са разработени методи за анализ на пикове. Най-популярният метод е MACS, който емпирично моделира размерите на изместване на ChIP-Seq таговете и го използва за подобряване на пространствената разделителна способност на прогнозираните сайтове за свързване.

Друг подходящ изчислителен проблем е диференциалното нализ на пикове, което идентифицира значителни разлики в два ChIP-последователни сигнала от различни биологични условия. Диференциалните пикови повикващи сегментират два ChIP-seq сигнала и идентифицират диференциални пикове с помощта на скрити Марков модели.

Анализ на данни от последователността на ChIP

1. предварително четене на секвенции чете
2. Картиране на риидовете
3. данни след обработка
4. оценка на качеството и силата на ChIP-сигнал
5. показват диаграми за покритие в геном браузери
6. търсене на ChIP пикове с MACS2
7. инспектира получените резултати
8. потърсете мотиви за последователност в рамките на наречените върхове
9. анализ на разпределението на обогатените региони в гените.

[**https://galaxyproject.org/tutorials/chip/**](https://galaxyproject.org/tutorials/chip/)

[**https://www.sciencedirect.com/science/article/pii/S0092867411013511**](https://www.sciencedirect.com/science/article/pii/S0092867411013511)

Задача 1

Влезте във вашия профил на Галакси, създайте си нова история.

Reb1 ChIP-exo

За този анализ ще използваме набори от данни ChIP-exo. Получени чрез експеримент на имунопреципитация с антитела срещу Reb1. Reb1 разпознава специфична последователност (TTACCCG) и участва в много аспекти на регулацията на транскрипцията от трите дрождени РНК полимерази и спомага образуването на свободни от нуклеозоми региони (NFRs) (Hartley & Madhani: 2009; Raisner: 2005).

Описание на данните

Ще използваме следните 4 набори с данни

| **Dataset** | **Description** |
| --- | --- |
| Reb1\_R1 | ChIP experiment, Replicate 1 |
| Input\_R1 | Input DNA, Replicate 1 |
| Reb1\_R2 | ChIP experiment, Replicate 2 |
| Input\_R2 | Input DNA, Replicate 2 |

Тези набори с данни са налични в [Galaxy library](https://usegalaxy.org/library/list" \l "folders/F050cbba300e2dbed) <https://vimeo.com/212753639>

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/lib.png |
| (R1 и R2). Това са данни от един край. Качете набори от данни в нова история, като изберете всички набори от данни и щракнете върху бутона История. Дайте име на новата история и изберете Импортиране  <https://vimeo.com/212753639> |

Creating a dataset collection | Single end data

След като качим набори от данни в историята на Galaxy, ние ще комбинираме всички набори от данни в един набор от данни. Това ще опрости обработката на данните надолу по веригата. Процесът за създаване на колекция за този урок е показан тук.

<https://vimeo.com/212757252>

Картиране и пост-обработка (Mapping and Post-processing)

Mapping (картиране)

В този конкретен случай данните са с много високо качество и не е необходимо да се подрязват или да се обработват по никакъв начин преди картографирането. Ще продължим, като картографираме всички данни спрямо последната версия на дрождевия геном sacCer3:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/mapping.png |

Картиране на всички данни наведнъж. Обърнете внимание, че Избор на тип въвеждане е зададен на Single fastq и чрез избиране на бутона папка () можете да изберете като цяла колекция от набори от бързи данни. Важно: тук също задаваме автоматично групи за четене чрез превключване на падащото меню за информация за четене на групи за задаване на групи за четене (спецификация SAM / BAM) и настройване на всички бутони за автоматично присвояване на Да.

Пускането на BWA върху колекция ще генерира друга колекция от BAM файлове. Дайте име на данните от тази колекция

https://vimeo.com/212758694

**Последваща обработка**

За последваща обработка ще премахнем всички нееднозначно картографирани показания. Това може да стане чрез просто филтриране на всички четения с качество на картографиране по-малко от 20

**NGS: SAMtools → Filter SAM or BAM**:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/bam_filter.png |

Филтриране на многопосочни четения чрез ограничаване на данните до четения с качество на картографиране над 20. Имайте предвид, че като изберете бутона папка (), можете да изберете като цяла колекция от набори от данни за BAM, които да филтрирате наведнъж.

Изпълнението на филтър SAM или BAM в колекция ще генерира друга колекция от BAM файлове. Наименувайте филтрираните данни от тази колекция <https://vimeo.com/212758694>

(Оценка на качеството на ChIP) Assessment of ChIP quality

След като картографирахме и филтрирахме показанията, време е да направим някои изводи за това колко добри са основните данни.

Корелация между пробите

В експеримента има две реплики, всяка от които съдържа обработка и входни (контролни) набори от данни. Първото нещо, което можем да проверим, е дали пробите са свързани (с други думи, ако пробите за обработка и контрол в двете реплики съдържат същия вид сигнал). За целта първо генерираме матрица за броене на четене с помощта на NGS:

**NGS: DeepTools → multiBamSummary**.

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/multibamsummary.png |

Изпълнение на **multiBAMsummary** за колекция от BAM набори от данни (както преди можете да изберете колекция чрез натискане на( папка ) бутон).

Този инструмент разбива генома на участъци с фиксиран размер (10 000 bp в нашия пример) и изчислява броя на показанията, попадащи във всеки кош. Ето фрагмент от неговата продукция:

#'chr' 'start' 'end' 'Reb1\_R1' 'Input\_R1' 'Input\_R1' 'Reb1\_R2'

chrVI 0 1000 19.0 41.0 3.0 6.0

chrVI 1000 2000 29.0 30.0 13.0 5.0

chrVI 2000 3000 0.0 0.0 0.0 0.0

chrVI 3000 4000 0.0 2.0 0.0 0.0

chrVI 4000 5000 7447.0 139.0 7.0 2645.0

След това генерираме HeatMap s **NGS: DeepTools → plotCorrelation**

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/plotcorr.png |
| **A.** Стартираме plotCorrelation от изхода на multiBamSummary. |
| https://galaxyproject.org/tutorials/chip/corr.png |
| **B.** Heatmap от четири проби: обработки (Rab1) и контроли (вход) са добре свързани помежду си. |

Тук можем да видим, че има добри корелации между реплики (между Reb1\_R1 и Reb1\_R2, и между input\_R1 и input\_R2), докато корелациите между обработките (Reb1) и контролите (вход) са слаби. Това е добър знак, подсказващ, че има някакъв сигнал към нашите данни.

Оценка на силата на сигнала

Как да кажем дали имаме сигнал, идващ от обогатяването на ChIP? Един от начините за това е скалирането на извличане на сигнала (SES), предложено от Diaz: 2012. SES работи по следния начин. Да предположим, че имаме два набора от данни: ChIP и Input DNA. Разделяме генома на N не-припокриващи се прозорци (N = 10 в примера по-долу) и за всеки прозорец изчисляваме броя на показанията. По този начин завършваме с два списъка: единият списък се чете за ChIP (ChIP списък), а другият за Input (Input list):

Window ChIP-count Input-count

-------------------------------

1 3 3

2 4 3

3 2 1

4 1 3

5 3 3

6 27 2

7 18 3

8 2 2

9 45 3

10 8 3

След това сортираме списъка ChIP във възходящ ред и преместваме елементи от входния списък, за да съответстваме на този ред:

Window ChIP-count Input-count

-------------------------------

4 1 3

3 2 1

8 2 2

1 3 3

5 3 3

2 4 3

10 8 3

7 18 3

6 27 2

9 45 3

Сега нека добавим още две колони към този набор от данни. Тези колони ще показват процент на четенията, обобщаващи до всеки ред за ChIP и входни данни. Например 0,044 на ред 3 е (1 + 2 + 2) / 113 = 0,044.

1 2 3 4 5

-------------------------------

4 1 3 0.008 0.115

3 2 1 0.026 0.153

8 2 2 0.044 0.230

1 3 3 0.070 0.346

5 3 3 0.097 0.230

2 4 3 0.132 0.576

10 8 3 0.203 0.692

7 18 3 0.362 0.807

6 27 2 0.601 0.884

9 45 3 1.000 1.000

------------------------

113 26

Where:

1 = Window, 2 = read count in ChIP, 3 = read count in Input

4 = % or read to this point in ChIP 5 = % of read to this point

В матрицата над голяма част от отчитанията на ChIP (колона 4) е концентрирана в няколко контейнера близо до дъното. Това не е така за входните показания (колона 5). Ако начертаем две последни колони на тази матрица, ще получим такаwa крива:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/ses.png |

SES сюжет за нашия пример. Повечето "четения" в експеримента ChIP са концентрирани в последните три проби.

DeepTools предоставят инструмент как може да се прецени успехът на ChIP експеримента въз основа на сюжетите на SES (also called *fingerprint*) plots:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/fingerprint_dt.png |

Обяснение на DeepTools за сюжетите на SES.

Така че нека да приложим DeepTools към собствените ни данни с помощта на **NGS: DeepTools → plotFingerprint**:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/plotfingerprint.png |
| **A.** Стартирайте plotFingerprint върху филтрирани данни (15.). |
| https://galaxyproject.org/tutorials/chip/plotfingerprint_out.png |
| **B.** SES пръстов отпечатък от четири проби: Третиранията (Rab1) показват характерна форма, показваща ChIP-сигнал. Приблизително 30% от показанията се съдържат в няколко% от генома. |

Генериране на набори от bigWig за показване

В този раздел ще конвертираме BAM файлове, генерирани с bwa, във формат bigWig, който ще ни позволи да гледаме разпределението на покритието на четене в генома..

За генериране на набори от bigWigще използваме **NGS: DeepTools → bamCoverage**:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/bam_cov_1.png |
| Изпълнение на bamCoverage върху колекция от филтрирани набори от данни за BAM (както преди можете да изберете колекция чрез натискане на бутона папка ()). Тук задаваме размера на **Bin size** 25. След това задаваме **Effective genome size** според зададения от потребителя и въвеждаме 12000000 (приблизителен размер на генома Saccharomyces cerevisiae). |
| https://galaxyproject.org/tutorials/chip/bam_cov_3.png |
| И накрая, ние зададохме Extend четене до дадения среден размер на фрагмента до 150. Това е така, защото в този конкретен експеримент ДНК беше избран размер, който да бъде между 120 и 170 bp за подготовка на библиотеката. |

Изпълняването на bamCoverage върху колекция от BAM набори от данни ще генерира колекция от набори от bigWig. Наименувайте на тази колекция coverage.

Визуализиране на multiple datasets in IGV

Сега можем да покажем набори от данни bigWig, генерирани в предишния раздел в браузър за геноми. Има различни налични браузъри. В този урок ще използваме IGV браузър

<https://vimeo.com/123414437>

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/cvrg_collection.png |

Колекция от bigWigs произведени от bamCoverage по-горе. Обърнете внимание, че в един разширен набор от данни (Reb\_R2) има дисплей на IGV` връзка.

Кликването на тази връзка във всичките четири набора от данни (ще трябва да разширите всеки набор от данни, като щракнете върху нея. Това ще разкрие IGV връзка) и фокусирането на браузъра върху MPH1 (YIR002C) гена ще генерира следното изображение:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/igv1.png |
| **Coverage** разпространение в рамките на IGV. Тук репликите на ChIP са оцветени в оранжево, а контролите - в синьо. И четирите песни бяха зададени на максимална стойност 70. |

Calling peaks

Въпреки че върховете, показани на екрана на браузъра по-горе, са доста ясни и последователни в двете реплики, разглеждането на целия геном в браузъра едва ли е устойчив начин за идентифициране на всички пикове. Има няколко начина за идентифициране на свързващи събития в геном. Те са обобщени на фигурата по-долу:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/t15_peak_calling.jpg |

Очертаване на три метода за откриване на събитие за свързване на ChIP. Методите за намиране на пик обикновено или изместват местата на таг за ChIP-seq в 3 'посока наполовина от очакваната дължина на фрагмента, или разширяват дължината на тага в 3' посока, за да бъде равна на очакваната дължина на фрагмента. Маркерите от противоположни направления са обединени, за да създадат необвързани пейзажи на плътност на таговете, а местата за свързване на събитията се прогнозират от местоположенията с максимално покритие на тагове във всеки регион, които съдържат значително обогатяване на ChIP-seq тагове (т.е. върховата среща на върха). Методи за върхово сдвояване [например GeneTrack изгражда подобни пейзажи за плътност на маркери, но запазва информация за нанизаност и обикновено не измества или разширява местата на маркерите. Локалите на върховете се определят на всеки кичур поотделно, а близките върхове в правилната верижна ориентация в рамките на дадено разстояние са сдвоени заедно. Местата на свързване на събитието се прогнозират от местата на средната точка на върховата двойка. Вероятни методи за откриване на свързване имат за цел да оценят местоположенията на свързващи събития, които биха могли да доведат до наблюдаваните локации на ChIP-seq тагове. Тези методи започват обучение с първоначални предположения за местата на свързване на събитията и модел за това как се очаква да се разпределят етикетите около реални ChIP-seq свързващи събития. По време на всяка стъпка на обучение всеки ChIP-seq маркер е вероятно свързан с близки свързващи събития, в зависимост от разстоянието между маркера и мястото на събитието. Като се имат предвид тези вероятностни назначения на маркери, местата за свързване на събитията се актуализират, за да се постигне по-добро съответствие с техните асоциирани маркери, а моделът за това как се разпределят етикетите около свързващите събития се актуализира, за да отразява натрупването на тагове около всички текущи събития на свързване. По време на процеса на обучение, свързващи събития с малко присвоени маркери са изсечени от модела и процесът в крайна сметка се сближава до набор от крайни места за свързване. (Фигура Mahony and Pugh: 2015).

Ние ще използваме MACS2 peak caller

Как работи MACS?

MACS (или текущата му версия MACS2) изпълнява няколко стъпки за peak caller от сдвоени набори данни за лечение / контрол:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/t15_macs_workflow.jpg |
| **Стъпки от работния процес на** (From [Feng:2012](http://www.nature.com/nprot/journal/v7/n9/full/nprot.2012.101.html)). |

Ето кратко описание на тези стъпки:

**Removing redundancy**- MACS запазва уникално картографирани четения и премахва четения, които многократно са картографирани на едно и също място. Това намалява ефектите от пристрастия на PCR усилването по време на подготовката на библиотеката.

**Build model and estimate fragment size** - един от входовете на MACS е размерът на фрагмента или широчината на честотната лента, което е приблизителният размер на ДНК фрагменти, генерирани по време на етапа на фрагментиране на библиотечната подготовка. MACS първо плъзга прозорец с размер, двойно по-широк от честотната лента в генома, и намира случаи, при които броя на четените се обогатява с 10 до 30 пъти спрямо фона на генома. След това на случаен принцип изважда 1000 от такива региони и изгражда модела. За да изгради модела, той отделя четене на карти на всяка от направленията и изгражда две разпределения (два режима). Междинната точка между двата режима е средата на размера на свързване, а разстоянието между режимите е размер на фрагмента d (вижте фигурата по-долу).

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/t15_macs_model.png |
| Пиковете, картографирани на две нишки, се обработват отделно, за да се създадат два профила на плътност на покритието - два два режима. Разстоянието между режимите е размерът на фрагмента d. Този профил е изграден от 1000 произволно подбрани обогатени региони (От Zhang: 2008). |
|  |

**Generate peaks**- сега, когато d е дефинирано, MACS плъзга прозорец с размер 2d през генома, за да идентифицира региони, значително обогатени в ChIP пробата. MACS приема, че фонът на четене се подчинява на разпределението на Poisson. По този начин, като се има предвид броят на четенията в даден интервал в рамките на контролната проба, можем да изчислим вероятността да сме наблюдавали брой четения в пробата ChIP (например, вижте примера наводнение тук). Тази процедура се провежда за няколко интервала около изследваното място (2d, 1kb, 5kb, 10kb и целия геном) и се избира максималната стойност. Един проблем при този подход е, че той работи само ако и двете проби (ChIP и контрол) са секвенирани до дълбочината, което обикновено не се случва на практика. За да коригирате с това MACS мащабира по-голямата проба.

Compute False Discovery Rate (FDR) – (Feng: 2012) обяснява изчисляването на FDR в MACS, както следва: „Когато е налична контролна извадка (и наистина трябва винаги да я използвате - AN), MACS може също да оцени емпиричен FDR за всеки пик, като размени ChIP-seq и контролни проби и идентифициране на пикове в контролната проба, като се използва един и същ набор от параметри, използвани за ChIP-seq пробата. Тъй като контролната проба не трябва да показва обогатяване на четене, всички такива пикове, открити от MACS, могат да се разглеждат като невярно положителни . За определен праг на P стойност, емпиричният FDR след това се изчислява като броя на контролните пикове, преминаващи прага, разделен на броя ChIP-последователни пикове, преминаващи същия праг. "

Намиране на върхове

В нашия случай имаме две реплики, всяка от които съдържа ChIP и входни ДНК проби. Първо ще стартираме MACS2 върху обединени данни (комбинирайки съответно две ChIP проби и два входа). След това ще стартираме MACS2 на всяка реплика поотделно. Накрая ще изберем стабилен набор от пикове, присъстващи и в трите телефонни апарата.

Разделяне на данни в отделни проби

Едно усложнение с начина, по който обработваме всички данни е, че сме комбинирали всичко в една колекция от база данни. MACS обаче ще трябва да разделим ChIP проби и контроли. За щастие за нас сме създали групи за четене, когато сме картографирали показанията към генома на дрождите. Това ще ни е удобно в момента, защото ще:

обединете цялата колекция от картографирани и филтрирани BAM в единичен набор от данни за BAM

разделете този набор данни на четири отделни BAM файла, като използвате групи за четене

стартирайте MACS върху получените файлове.

Първо, за да обединим колекция от картографирани, филтрирани BAM файлове в един набор от данни, който ще използваме **NGS: Picard → MergeSamFiles**:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/merge_bam.png |

Обединяване на колекция с **MergeSamFiles**. Тук използваме параметри по подразбиране.

След това ще използваме **NGS: SAMtools → Split**, за да разделим обединения файл в отделни BAM файлове. Всеки получен BAM файл ще съдържа подравнени показания, съответстващи на първоначалните четири набора от данни:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/bam_split.png |
| **Splitting BAM dataset** на групи за четене. Това ще доведе до четири BAM набора от данни. |
| https://galaxyproject.org/tutorials/chip/split_data.png |

Резултатни набори от данни. Всяка съдържа подравнени четения от четирите оригинални условия.

Стартиране MACS2

Стартиране MACS2. Първо ще използваме NGS: Peak call → MACS2 предсказуем инструмент от пакета MACS2. Този инструмент ще ни помогне да намерим оптимални параметри за изпълнение на пикова функция на MACS2:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/predictd1.png |
| Прогноза за оценка на параметъра d. Тук задаваме Ефективния размер на генома на стойност, специфична за дрождите, задаваме ширината на лентата на 150 (дължината на фрагмента след избора на размер) и увеличаваме Задаване на горния mfold свързан на 100 (това е ChIP-exo експеримент, където очакваме да има остър, значително обогатяване на региони. Оставянето на този параметър по подразбиране 50 може да не успее да намери пикове, тъй като тези конкретни набори от данни бяха генерирани от ChIP-exo протокол). |

Направете това и на другата реплика!

Тази процедура ще ни помогне да оценим d параметъра чрез извършване на [кръстосана корелация] анализ между четене картиране на + и - направления. Нека да разгледаме тези резултати:

| **Replicate 1** | **Replicate 2** |
| --- | --- |
| https://galaxyproject.org/tutorials/chip/d_r1.png | https://galaxyproject.org/tutorials/chip/d_r2.png |
| Peak model и изоставане между нишките. |  |

В случая на тези данни пиковете са много остри и имат тясна междина между тях: 27 и 33 bp за реплики 1 и 2, съответно. Ще използваме средно тези стойности, 30, като --extsize параметър за извикване на пикове с помощта **NGS: Peak calling → MACS2 callpeak**:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/macs1.png |
| Извикване на пикове с MACS2 върху обединени данни. Тук избираме множество входове чрез натискане на бутон и избор на двата набора от данни ChIP в обработващия файл на ChIP-Seq и двата набора от данни за ДНК в контролния файл на ChIP-Seq. След това избираме генома Saccharomyces cerevisiae като Ефективния размер на генома. Интерфейсът на MACS2s е дълъг и на тази фигура го разделихме на няколко части. Вижте и долния раздел - важно е! |
| https://galaxyproject.org/tutorials/chip/macs2.png |

В тази долна част на MACS2 интерфейс задайте Build model но не изграждайте модела на изместване (вече направихме това с preductd в предишната стъпка) и \*Set extension size to 30 (числото, което сме оценили в предишната стъпка). И накрая, ще поискаме само MACS2 да произведе два изхода: Пикови срещи и този, който е произведен по подразбиране, който съдържа пикови координати.

Ако зададете параметри, както е показано по-горе, MACS2 ще произведе два изхода (ако той произведе повече просто намерете тези, наречени тесни пикове и срещи). Нека кликнете върху иконата на молив (), съседна на summits and narrow peak на набори от данни и преименувайте след това, както е показано по-долу:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/macs_out_renamed.png |
| **MACS2 output** with summits and narrow peak datasets renamed. | . |

След това ще стартираме MACS2 на BAM набори от данни само за реплика 1:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/macs3.png |
| Calling peaks with MACS2 on R1 набори от данни, всички останали параметри трябва да бъдат зададени както на предишната фигура |

Сега направете това сами:

преименувайте получените набори от данни като R1 срещи и R1 върхове

стартирайте MACS2 стартирайте на Replicate 2

преименувайте получените върхове и тесни набори от пикови връзки като R2 срещи и R2 върхове.

В крайна сметка трябва да имате нещо подобно:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/all_macs_datasets.png |
| Всички набори от данни MACS2. След като стартирахме MACS2 три пъти, трябваше да имаме анкети, R1 и R2. |

Проверка на върхове

Разглеждайки данните на MACS2, ние сме получили следните числа на върховете:

| **Pooled** | **Replicate 1** | **Replicate 2** |
| --- | --- | --- |
| 974 | 955 | 784 |

Peaks data is generated in the following format:

1 2 3 4 5 6 7 8 9 10

--------------------------------------------------------------------

chrI 35 491 MACS2\_peak\_1 176 . 10.35332 21.51081 17.68957 101

chrI 87135 87212 MACS2\_peak\_2 127 . 7.71763 15.89278 12.78060 26

chrI 92612 92793 MACS2\_peak\_3 153 . 9.22373 18.72748 15.31966 49

chrI 119739 119782 MACS2\_peak\_4 78 . 6.08885 10.52482 7.82302 25

Където колоните са:

1. Chromosome
2. Start
3. End
4. Iterative id given by MACS2
5. Integer score for display
6. Strand (irrelevant in this case)
7. Fold-change (fold enrichment for this peak summit against random Poisson distribution with local [lambda](https://en.wikipedia.org/wiki/Poisson_distribution))
8. -log10*P*-value (e.g., 17.68 is 1 x 10-17)
9. -log10*Q*-value from [Benjamini–Hochberg–Yekutieli procedure](https://en.wikipedia.org/wiki/False_discovery_rate" \l "Benjamini.E2.80.93Hochberg.E2.80.93Yekutieli_procedure.)
10. Relative summit position to peak start

Колко върхове са общи между репликите?

За да видите колко върхове са общи между обединените набори от данни и двете реплики, ще използваме инструмента

**Operate on Genomic Intervals → Join** tool twice.

Първо ще се присъединим към Peaks, обединен с Peaks R1:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/join1.png |

Присъединяване към групирани и R1 резултати с инструмент за присъединяване. Имайте предвид, че тъй като преименувахме наборите от данни, те вече лесно се избират.

След това ще се присъединим към резултата от предишната операция с Peaks R2:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/join2.png |
| **Joining Pooled/R1 with R2** с инструмент за присъединяване. Имайте предвид, че тъй като преименувахме наборите от данни, те вече лесно се избират |

Резултатите от това в 723 региона са споделени между пиковете, изследвани R1 и R2. Нека наречем този набор **High confidence set**. Преди да можем да го използваме обаче, нека изрежем само съответните колони. Тъй като създадохме този набор от данни, като се присъединихме към три други набора от данни, той е три пъти по-широк (30 колони). За да изрежем първите три колони, можем да използваме инструмент **Text Manipulation → Cut columns** tool:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/cut.png |

Изрязване на колони от изход за присъединяване.

Преименувайте последния набор от данни като **High confidence set**. Това ще улесни намирането, докато продължаваме.

Използването на инструмента за изрязване на колони създава набор от данни от табличен тип. Въпреки това, чрез изрязване на първите десет колони създадохме набор от данни във формат BED. Ето защо трябва да уведомим Galaxy за това, като нулираме метаданните, както е показано по-долу.

След това трябва да се уверим, че изходът на инструмента за изрязване на колони има типа BED. За целта ще редактираме метаданните му, както е показано по-долу

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/bed_type.png |
| Задаване на метаданни на тип BED. Щракнете върху иконата на ( молив) в съседство с набора от данни и изберете раздела Datatype. Там ще можете да го зададете на BED. |

Нека разгледаме всичко в браузъра

Нека визуализираме обединените върхове, както и тесни пикове и срещи на върха, произведени от MACS2 в IGV, като щракнете върху дисплей с IGV локални връзки в съседство с обединените пикове и набори от данни с високо доверие (вече трябва да имате отворен браузър):

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/igv2.png |
| Преглед в IGV. Тук можете да видите оригинални набори от данни на bigWig заедно с прогнозираните пикове. |

Какви последователности се намират в пикове

В този експеримент антитела срещу Reb1 протеин са използвани за имунопреципитация. Мястото за разпознаване на Reb1 е TTACCCG (Badis: 2008 и Harbison: 2004). За да разберем кои мотиви на последователността се намират в нашите върхове, първо трябва да преобразуваме координати в основни последователности. Това става с помощта **Fetch Alignments/Sequences → Extract Genomic DNA** tool:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/extract_dna.png |

Екстрахиране на геномна ДНК, съответстваща на ChIP-seq пикове. Тук използваме набор от данни за обединени пикове, генерирани няколко стъпки по-рано.

На следващо място, трябва да се уверим, че всички последователности са достатъчно дълги за намиране на модели. MEME, инструментите, които ще използваме за намиране на мотиви, изискват последователности да са дълги поне 8 нуклеотида. Така че ще премахнем кратки последователности, използвайки FASTA manipulation → Filter sequences by length tool:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/fa_length.png |

Филтриране на FASTA по дължина. Тук премахваме всички секвенции, къси от 8 нуклеотида.

Сега можем да използваме Motif Tools → MEME:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/meme1.png |

Изпълнение на MEME върху филтрирани по дължина FASTA последователности от предишния етап. Обърнете внимание, че конфигурацията на Опции е зададена на Разширено, а Проверка на обратното допълване е на Да.

MEME генерира редица изходи. Най-интересният е HTML Report. Показва, че 620 региона съдържат TTACCCG мотив:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/meme2.png |

MEME мотив, открит в 620 последователности, съответстващи на общи пикови участъци.

Обобщаване обогатяването на ChIP сигнала във всички гени

Колко гена съдържат предходни региони, обогатени в ChIP маркери. Това често се представя като heatmap:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/plotHeatmap_example.png |
| **Heatmap example** from [DeepTools documentation](https://deeptools.readthedocs.io/en/latest/). |

За да генерираме **Heatmap**, първо трябва да нормализирме набори от данни за двете репликирани, които имаме. Това става с помощта **NGS: DeepTools → bamCompare** tool:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/bamCompare1.png |

Стартираме bamCompare на реплика 1. Тук зададохме Метод, който да използваме за мащабиране на най-голямата проба до най-малката до SES

Правим един и същ анализ на Replicate 2 набора от данни и преименувайте двата получени елемента като R1 normalized and R2 normalized.

Тъй като искаме да начертаем обогатяване около гени, трябва да изтеглим анотацията на гените. Ще използваме Get Data → UCSC Main за това:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/ucsc1.png |
| Получаване на данни от UCSC. Тук не забравяйте да изберете монтажа, наречен sacCer3 и избирате SGD гени **Get output** ще покаже следващия екран, показан по-долу |
| https://galaxyproject.org/tutorials/chip/ucsc2.png |
| Izbirame **Send query to Galaxy**. |

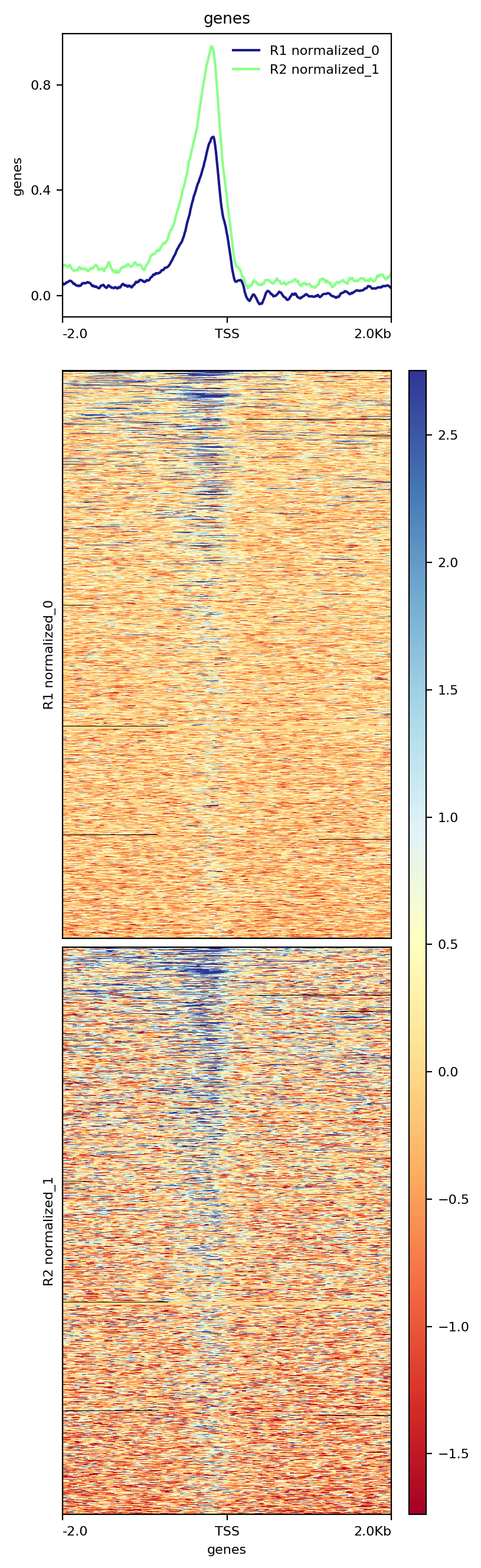
На следващо място, за да подготвим данни, необходими за изготвяне на heatmap, ще използваме heatmap **NGS: DeepTools → computeMatrix** utility:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/computeMatrix1.png |
| Изчислителна матрица - данните, от които ще бъде изградена Heatmap . Тук и двата нормализирани набора от данни са избрани в полето Score файл, гените за дрожди, които току-що изтеглихме от UCSC, се избират като Regions to plot . 'референтната точка е зададена като **computeMatrix main option** и накрая разстоянията нагоре и надолу по веригата са зададени на 2000 bp. Очевидно сте добре дошли да играете с тези параметри. |

И накрая, можем да визуализираме топлинната карта, като използваме **NGS: DeepTools → plotHeatmap** tool:

|  |
| --- |
| https://galaxyproject.org/tutorials/chip/plotHeatmap.png |
| **Drawing heatmap** with plotHeatmap tool. |

Полученото изображение показва, че значителна част от 6 692 гена, присъстващи в данните за поясненията, които използваме, съдържат Reb1 свързващи сайтове в техните предходни региони:



Целият анализ е на разположение като Galaxy history [here](https://usegalaxy.org/u/aun1/h/reb1-yeast-tutorial). Импортирайте го.

Линк: <https://usegalaxy.org/u/stanislav99/h/imported-reb1-yeast-tutorial>

