転写サイクル　初級者向けインフォマティクス講習会

「ウェブツールで始められる RNAseq、ChIPseq、発現アレイデータの解析・比較」

日時：２０１５年１１月１４日（土）９：３０~１７：００（一日目）

および１５日（日）９：００~１３：００（二日目）

場所：筑波大学　医学地区（http://www.md.tsukuba.ac.jp/top/access/）

講習会、解析内容についての質問先：村谷匡史（筑波大学、muratani@md.tsukuba.ac.jp）

日程：

１１月１４日（土）

８：３０　　会場を開放

（到着した参加者から）インターネット接続の確認、UCSC ブラウザ、Galaxy のアカウント作成

テキスト、練習用データの配布（データを持っていない参加者用）

９：３０~１１：００　　RNA 発現解析、ChIPseq 解析結果を用いたパスウェイ解析

パスウェイ解析（DAVID、REACTOME、PANTHER、GREAT）

Galaxy ツール、遺伝子名と GeneID の変換、ベン図の作成

１１：００~１２：００　　UCSC ブラウザの基本操作と様々なデータセット

UCSC ブラウザへのデータのアップロード

気になる遺伝子周辺の ChIPseq ピークの可視化、フィギュアパネルの作り方

１２：００~１３：００　　昼休み

（午後からの参加者のための UCSC ブラウザと Galaxy の説明リピート）

１３：００~１５：００　　プロモーターや ChIPseq ピークのモチーフ解析

ChIPseq ピーク領域のモチーフ解析、転写因子結合サイト解析（DREME、TOMTOM）

アレイ解析結果などの遺伝子リストからプロモーター領域の塩基配列を取得

RNAseq データを用いた場合（アイソフォームや lincRNA を考慮したプロモーター解析）

１５：３０~１６：３０　　３末端 UTR 配列の miRNA ターゲット解析

アレイ解析結果などの遺伝子リストから３末端 UTR の塩基配列を取得

RNAseq データを用いた場合（アイソフォームを考慮した３末端 UTR 解析）

TargetScan データベースとの比較による miRNA ターゲット解析

１６：３０~１７：００　　ENCODE、IHEC 等の公開データセットの利用とその他のツール紹介

１７：００~　　自由解析、Q&A、解散

１８：００　　会場施錠

１１月１５日（日）

９：００~１３：００　　一日目午前の内容のリピート、および自由解析と Q&A

大まかな解析の流れについて

RNAseq や ChIPseq では、定量解析やピークコールなどの結果が半自動化された「解析パイプライン」から提供される。「初級編」では、その後の解析について解説する。実験条件に特異的なイベントを探るには、コントロールや比較条件データとの「足し算・引き算」を行い、その後、得られた遺伝子リストやゲノム領域についてパスウェイ解析やモチーフ解析を行うが、それぞれのステップにウェブツールがあるので、組み合わせて「解析フロー」を作る。

	足し算、引き算	解析
遺伝子リスト（テキストファイル）	VENNY ベン図	DAVID GREAT パスウェイ解析
ゲノム領域リスト（BEDファイル）	Galaxy Interval tool	MEME Suite モチーフ解析

目次

1. ウェブツールアカウントの作成 4

1.1 UCSC ブラウザアカウントの作成

1.2 Galaxy アカウントの作成

2．パスウェイ解析 5

2.1a 遺伝子リストを作成する（Prdm14 論文から）

2.1b 遺伝子リストを作成する（アジレントマイクロアレイの例）

2.1c 遺伝子リストを作成する（RNAseq の場合の対応）

2.2 DAVID でパスウェイ解析を行う（https://david.ncifcrf.gov/）

Official Gene Symbol 以外の遺伝子リストの変換

Gene ID Conversion Tool による遺伝子名の変換

2.3 REACTOME を使う（http://www.reactome.org/）

2.4 PANTHER を使う（http://pantherdb.org/）

2.5 GREAT ツールを使った ChIPseq ピーク領域のパスウェイ解析

3．UCSC ブラウザと Galaxy を用いたデータの操作 15

3.1 Mock ChIPseq ピークの処理

3.2 ChIPseq ピーク領域を-/+50kb 前後に延長し、BED フォーマットに加工する

3.3 RefSeq トランスクリプトデータベースを Galaxy 上で加工する

3.4 ChIPseq ピークの BED ファイルを Galaxy にアップロードする

3.5 Galaxy の Interval ツールを使って 2 つの BED ファイルの重複領域を見つける

3.6 Venny でベン図を作成して、Prdm14 の Direct target の候補遺伝子を探す

4．UCSC ブラウザと Galaxy を用いたデータの可視化 20

4.1 可視化に使うデータの種類

4.2 bedGraph ファイルから必要な部分を切り取って小さなファイルを作成し、ヘッダーを付ける

4.3 UCSC ブラウザに可視化したいファイルをアップロードする

4.4 UCSC ブラウザの画像をダウンロードする

５．ChIPseq ピークのモチーフ解析 26

5.1　ChIPseq ピーク領域の塩基配列の長さの揃った BED ファイルを作成する

5.2　ChIPseq ピーク領域中心 300 ベースの FASTA ファイルを作成する

5.3　DREME でモチーフ解析、TOMTOM でモチーフ検索を行う

６．プロモーター領域のモチーフ解析 29

6.1　プロモーター領域の塩基配列の長さの揃った BED ファイルを作成する（Gata2 遺伝子付近で練習）

6.2　アレイ解析で変化していた遺伝子のプロモーター領域の取得

７．マイクロ RNA ターゲットのダウンロードとアレイデータとの比較 38

８．ノンコーディング RNA 遺伝子のプロモーター解析 42

９．遺伝子周辺のエンハンサー等の制御領域の解析 44

１０．その他のデータベース、解析ツール 45

10.1　ENCODE (https://genome.ucsc.edu/ENCODE/)、IHEC (http://ihec-epigenomes.net/)

10.2　Galaxy の FTP サイト

10.3　Lift-over tool（Galaxy）

10.4　EBI データベース（https://www.ebi.ac.uk/）

10.5　Gene-set enrichment analysis (GSEA)

10.6　Gene Regulatory Network tools

終わりに 47

便利な基本操作など 48

A1.　FASTA ファイルと BED ファイル形式

A2.　エクセルのキーボードショートカットと便利な機能

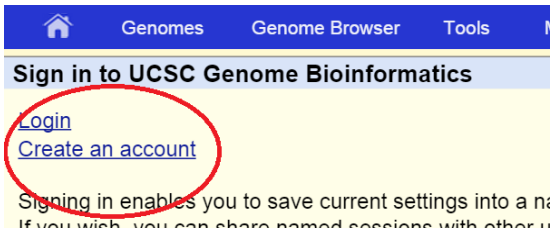
A3.　講習で使用する Prdm14 論文の Figure 2a のようなグラフをエクセルで作る

1. ウェブツールアカウントの作成

1.1 UCSC ブラウザアカウントの作成

https://genome.ucsc.edu/index.html

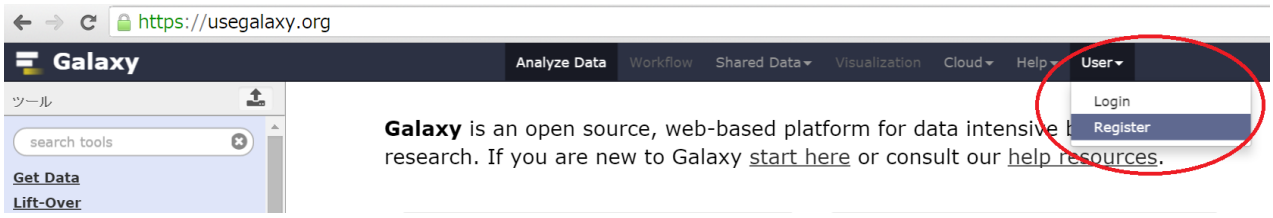
My Data のプルダウンメニューから、



Create an account を選択、メールアドレスなどを入力すると、リンクを含む確認メールが送られてきます。次回からは Login をクリックしてスタートすると前回のセッションが継続できます。

1.2 Galaxy アカウントの作成

https://usegalaxy.org/



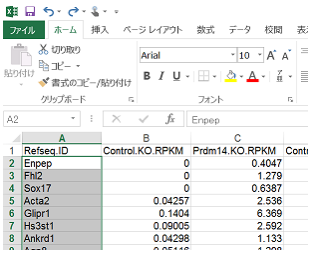
User から Register を選択、メールアドレスとパスワードを入力する。

２．パスウェイ解析

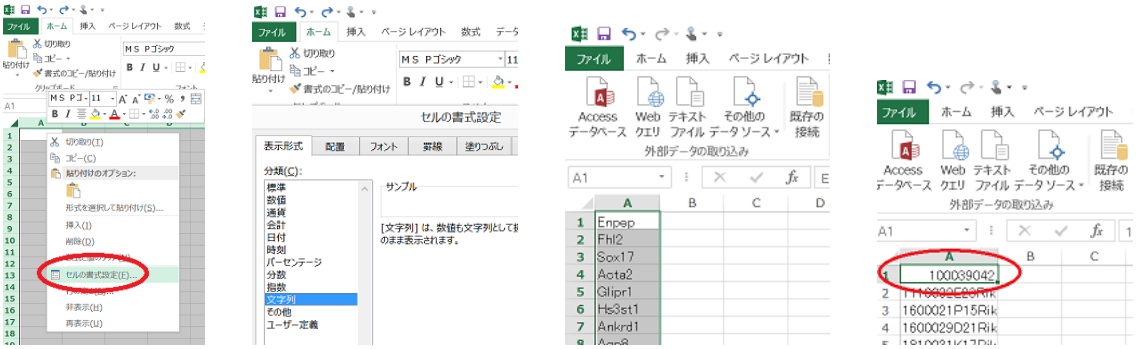
目標：アレイの委託解析などを用いた場合、「2 倍以上変化したプローブ」などのリストが納品される。このような多数の遺伝子名を含むリストから、ゲノム全体に比べて出現頻度の高い遺伝子の機能カテゴリを見つけ出す。また、機能的に重要そうな遺伝子やヒト疾患に関連した遺伝子を探す。

2.1a 遺伝子リストを作成する（Prdm14 論文から）

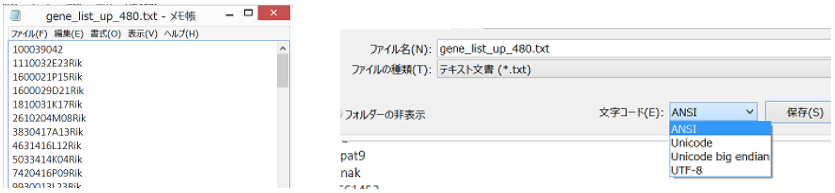
Table S2 の「Upregulated」のタブを開いて、「Refseq.ID」のカラムを選択、「Control+shift+↓」で 2 行目以降のすべての行のカラム A を選択、「Control+C」でコピーする。



新しいエクセル文書を開き、すべてのカラムについて「書式設定＞文字列」に変更した上で、「Control+V」でペーストする。「データ」メニューから「並べ変え＞A→Z」として一部の遺伝子名が日付などになっていないか再確認。ここで行数を数えておく。



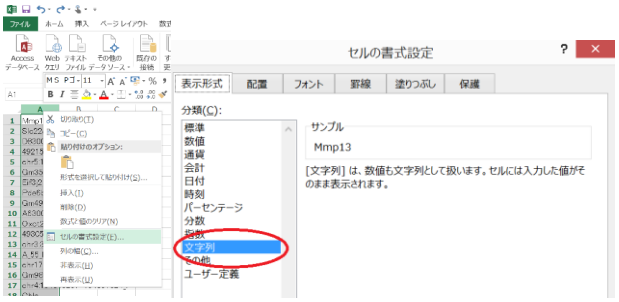
先頭に「100039042」という遺伝子名があるが、Table S2 もこうなっているので、これで良いと思われるが、エクセルによる自動変換に注意する。論文発表の時にも注意。このエクセルシートをタブ区切りのテキストファイルで保存するか、テキストファイルに、「Control+shift+↓」、「Control+C」、「Control+V」でコピー＆ペーストしてからテキストファイルとして保存する（gene_list_up.txt など）。同様に発現が下がる遺伝子のリストも作成する（gene_list_down.txt）



Windows での文字コードは「ANSI」に設定する。

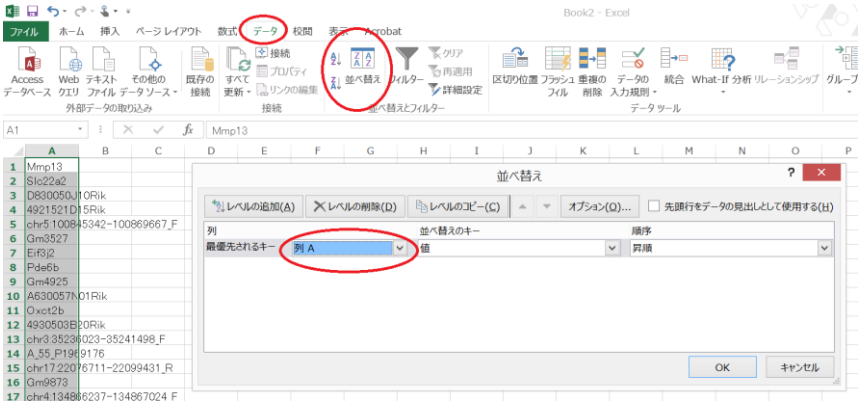
2.1b 遺伝子リストを作成する（アジレントマイクロアレイの例）

アレイ解析などでは、一つの遺伝子、またはトランスクリプトに対して複数のプローブがあるので、遺伝子リストのテキストファイルをエクセルに読み込んでフォーマットを整える。エクセルの標準設定では、数字のみのカラムは日付などに自動変換されてしまうので、列を選択して、（右クリックで）書式設定を「文字列」に変更する。



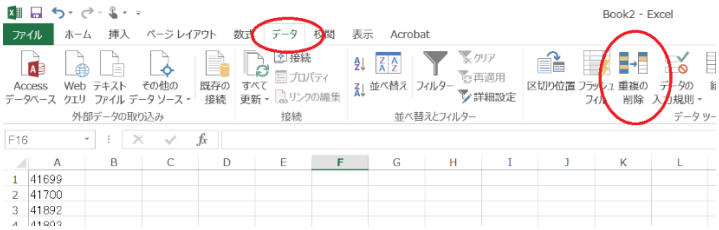
「文字列」に設定。

「データ」メニューから「並べ替え」を選ぶ。列 A を選択する。



ここで行の数を確認しておく（「Control+↓」で一番下まで行って行数を見る）。

「データ」から「重複の削除」で重複する遺伝子名を削除する。（多くのツールでは自動で重複を削除してくれるので、この操作は不要だが、行数を把握するためにやっておくとよい。）



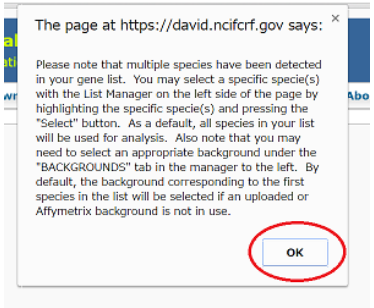
ここで遺伝子が少なすぎると、エンリッチメント解析ができないので、100 個程度はあるとよい。このシートをテキストファイルとして保存する（2x_up.txt など）。

2.1c 遺伝子リストを作成する（RNAseq の場合の対応）

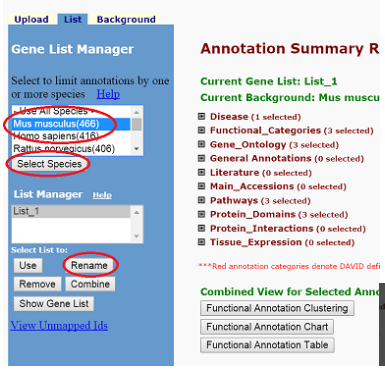
RNAseq では、トランスクリプトバリアントごとの定量結果がでるため、RefSeq ID（NM_XXXXX の形式）でも、「Gata2」などの標準遺伝子名（Standard gene name）でも解析できるが、ここでは標準遺伝子名を使って練習する。該当するカラムが無い場合は、RefSeq ID を使う。RefSeq ID と標準遺伝子名の変換方法については、DAVID を用いたパスウェイ解析の項で説明する。

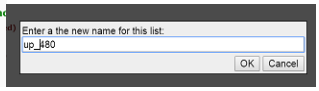
2.4 DAVID でパスウェイ解析を行う（https://david.ncifcrf.gov/）

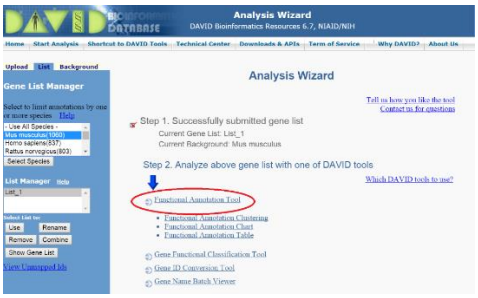
「Shortcut to DAVID Tools」から、「Functional Annotation」をクリック。



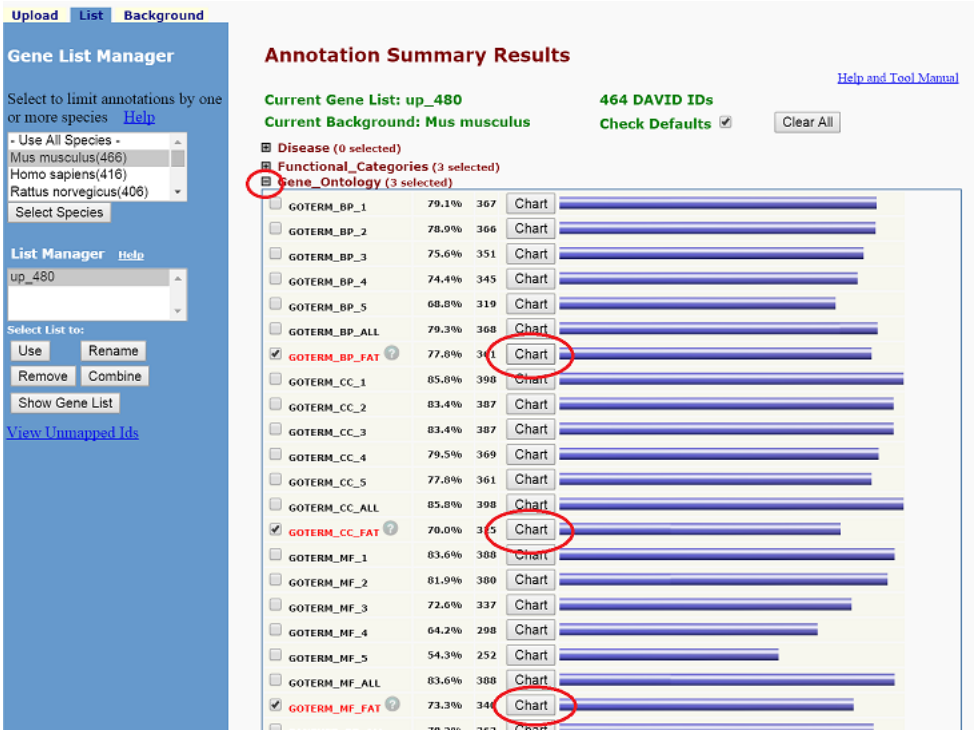
遺伝子名のリストはテキストからコピーする。（エクセルのセルではエラーになる）「Official Gene Symbol」、「Gene ID s」を選択して「Submit List」。上右図の 2 つの例のようにエラーが出る時もあるが（マップされない ID があるため）、「Continue… 」をクリックして進む。この場合、マッピングが成功した遺伝子のみで解析する。Refseq ID (NM_XXXX)を入力するときは、「Refseq_mRNA」などを選ぶ。



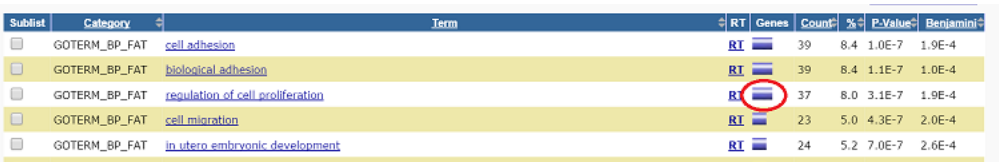




Prdm14 論文はマウスデータなので、「Mus musculus」を選んで「Select species」をクリック。「Rename」をクリックして、先ほどアップロードした遺伝子リストを「up_480」とリスト名を付ける。遺伝子リストだけをアップロードした場合は、右上のように次の解析を選択できるので、「Functional annotation tool」を選択する。

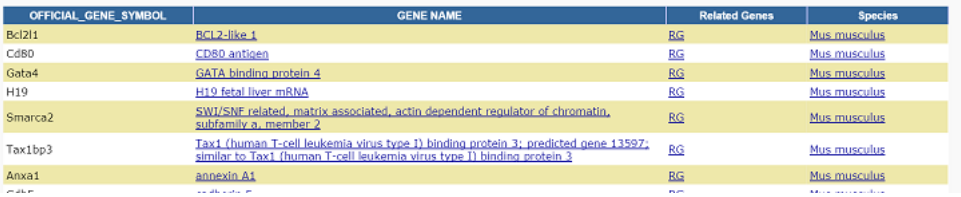


チェックボックスをクリックすると、選択肢が新たなウィンドウに表示される。Gene Ontology では、BP (Biological Process)、CC (Cellular Component)、MF (Molecular Function)の GO term のエンリッチメントが出てくる。FAT は DAVID で使われている少し使いやすくなった GO カテゴリーリストの意味。「Chart」をクリックすると機能カテゴリが表示される。「Biological Process」では、細胞接着や細胞増殖などの GO term がトップヒットになっている。



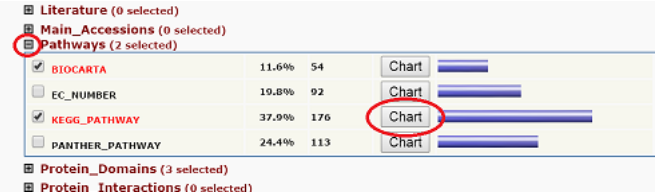
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_BP_FAT	cell adhesion	RT		39	8.4	1.0E-7	1.9E-4
	GOTERM_BP_FAT	biological adhesion	RT		39	8.4	1.1E-7	1.0E-4
	GOTERM_BP_FAT	regulation of cell proliferation	RT		37	8.0	3.1E-7	1.9E-4
	GOTERM_BP_FAT	cell migration	RT		23	5.0	4.3E-7	2.0E-4
	GOTERM_BP_FAT	in utero embryonic development	RT		24	5.2	7.0E-7	2.6E-4

さらに「Genes」をクリックすると、GO term に含まれる遺伝子が表示される。例えば細胞増殖制御...



OFFICIAL_GENE_SYMBOL	GENE NAME	Related Genes	Species
Bcl2l1	BCL2-like 1	RG	Mus musculus
Cd80	CD80 antigen	RG	Mus musculus
Gata4	GATA binding protein 4	RG	Mus musculus
H19	H19 fetal liver mRNA	RG	Mus musculus
Smarca2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2	RG	Mus musculus
Tax1bp3	Tax1 (human T-cell leukemia virus type 1) binding protein 3; predicted gene 13597; similar to Tax1 (human T-cell leukemia virus type 1) binding protein 3	RG	Mus musculus
Anxa1	annexin A1	RG	Mus musculus

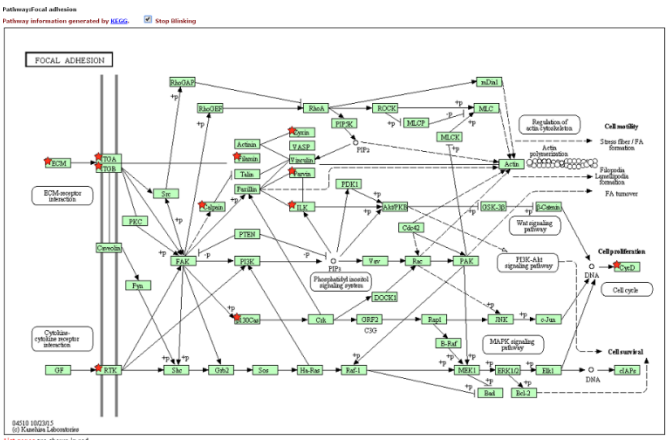
パスウェイ解析でも同様な傾向が出ているか確認する。

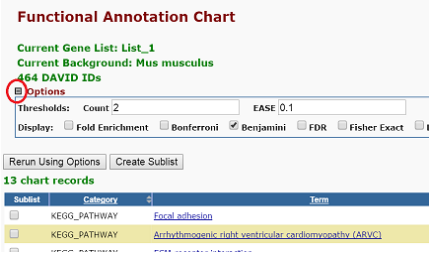


KEGG パスウェイの例。細胞接着、細胞外マトリックスなどのパスウェイが出ているので、GO 解析の結果と似ている。

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Focal adhesion	RT		22	4.7	4.6E-7	5.5E-5
	KEGG_PATHWAY	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	RT		12	2.6	1.4E-5	8.2E-4
	KEGG_PATHWAY	ECM-receptor interaction	RT		12	2.6	3.7E-5	1.5E-3
	KEGG_PATHWAY	Hypertrophic cardiomyopathy (HCM)	RT		10	2.2	9.6E-4	2.8E-2
	KEGG_PATHWAY	Small cell lung cancer	RT		10	2.2	1.0E-3	2.4E-2
	KEGG_PATHWAY	Dilated cardiomyopathy	RT		10	2.2	1.8E-3	3.5E-2

機能カテゴリをクリックするとマップが表示され、リストにあった遺伝子がハイライトされる。

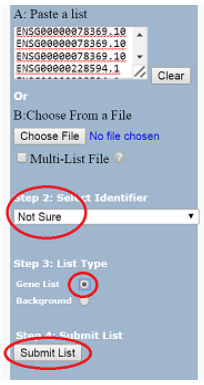


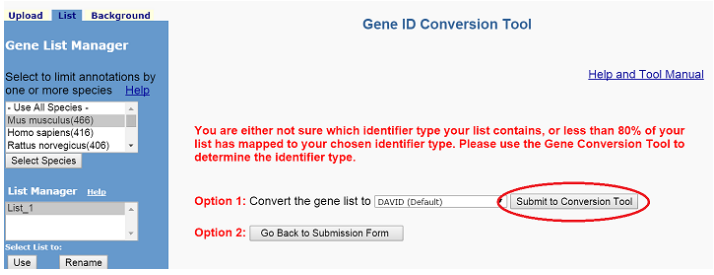


マウスの細胞を用いた実験データを解析する場合、マウスの遺伝子にマッピングするのが正しいが、ヒトのデータベースのほうが疾患関連遺伝子などを検索するのに便利である。探索的な作業では、ヒトの遺伝子にマッピングを行って、どんな機能の遺伝子があるのかを見ても良い。また、上右図のように、表示するヒットの閾値を変えられるので、遺伝子リストから機能分類をとりあえず見てみたい、というような場合にもこのツールが使える。

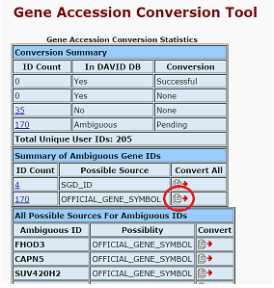
Official Gene Symbol 以外の遺伝子リストの変換

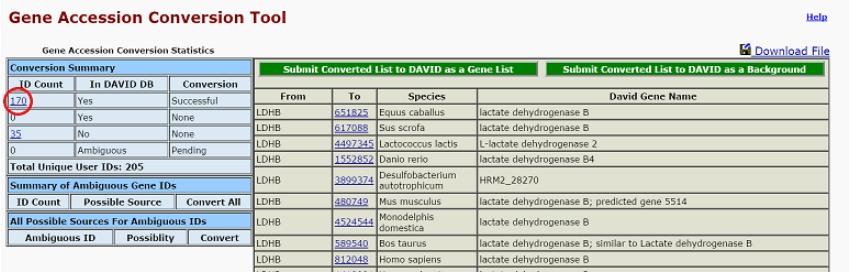
遺伝子名のリストの形式がよくわからない場合や、複数の形式の寄せ集めの場合は、「Not sure」を選択する。次のように Gene ID Conversion の画面が出るので、





Gene Conversion Tool をクリックすると、





変換の候補がグループになって表示される。赤矢印のアイコンをクリックすると、変換されたリストに加えられる。

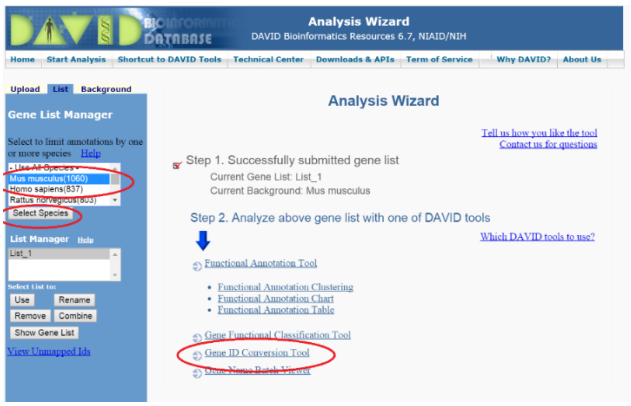
Gene ID Conversion Tool による遺伝子名の変換

アジレントアレイ解析の納品ファイルの Probe ID および Gene ID、プローブ情報の例。こららのカラムをそのまま使うこともできる。

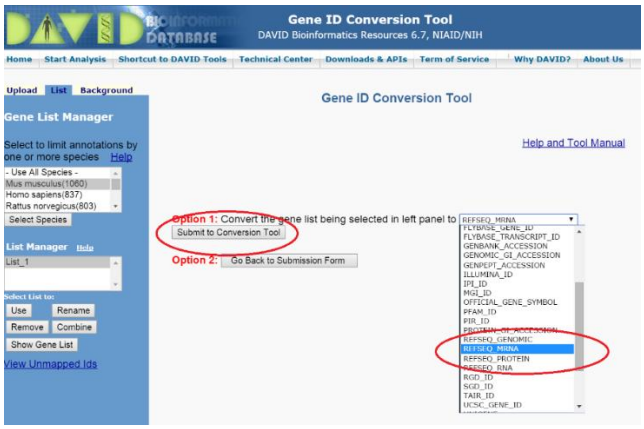
	A	B	C	数式バー	E
1	#フィーチャ(スポット)番号	#Agilent社が定義したプローブ名	#プローブがハイブリするよう設計されたターゲット遺伝子	#プローブがハイブリするよう設計されたターゲット配列のID	#公的データベースのaccession番号
2					
3					
4	FeatureNu	ProbeName	GeneName	SystematicNa	accessions
8	4	A_55_P2051983	Spink7	NM_001001803	ref|NM_001001803|ens|ENSMUST00000076194|ref|XM_006526082|gb|BC148575
9	5	A_52_P169082	Dhil5	NM_021294	ref|NM_021294|ens|ENSMUST00000040806|gb|AK006528|gb|S83465

	F	G	H	I	J	K	L	M	N	O	P
1	#プローブの染色体位置	#プローブのスタート位置	#プローブの詳細情報	#フィーチャのコントロールタイプ 0 = コントロール以外のフィーチャ 1,-1 = コントロールフィーチャ	#ノーマライズ後のシグナル値	#実測のシグナル値	#ノーマライズ後のシグナル値	#実測のシグナル値	#Normalized値のテスト/コントロールの変動倍率	#シグナルがサチュレーションしているか判定 0 = Not Saturated 1 = Saturated	#フィーチャの不均一性の判定 0 = 規定閾値内 1 = Non-Uniform OL
2					(グループ比較の場合、しプリケートの平均値となります)					(グループ比較の場合、しプリケート分足し算した数に	
3					Mus-Control		Mus-Suspension			FEFlags; Mus-Control	
4	chr_coord	Start	Description	ControlType	Normal	Raw	Normal	Raw	FoldChange	gIsSaturated	gIsFeatNonUni
8	mm|chr18:62753954-62753895	151	ref|Mus musculus	0	0.017498	4.74736	0.021635	4.663463	1.236414004	0	0
9	mm|chr11:76031955-76032014	362	ref|Mus musculus	0	0.032924	8.932634	0.080646	17.3838	2.449473528	0	0

パスウェイ解析の手順で遺伝子名のリストをアップロードし、マウス「Mus musculus」を選択「Select Species」をクリック。次に「Gene ID Conversion Tool」をクリック。



変換する先の ID の種類を指定する。例えば「RefSeq mRNA」など。



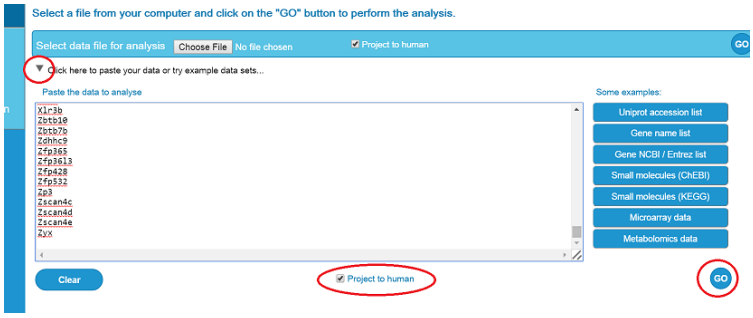
変換された遺伝子名が表で出力されるので、ダウンロードする。

2.5 REACTOME を使う（http://www.reactome.org/）

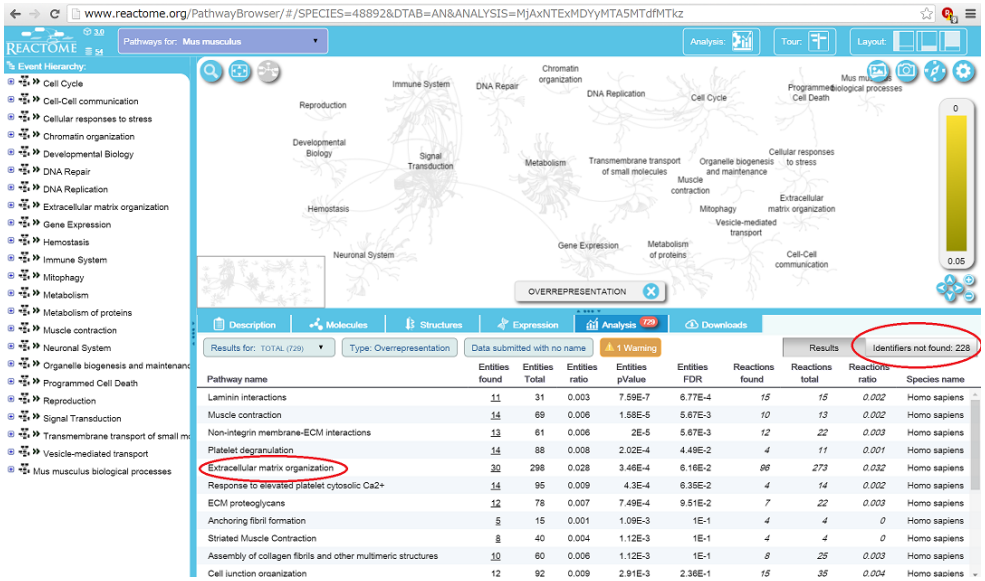
パスウェイ解析ツールの多くは GO、KEGG、BIOCARTA などの共通のデータベースを使っているため、解析結果に大きな違いはないが、使い勝手やアウトプットのビジュアルのデザインが異なるので、Google などで検索していろいろなツールを試すとよい。REACTOME は、パスウェイの階層構造が見やすい他、アレイデータなどの数値が入ったものも解析できる。一方で、ヒト以外の生物種ごとのマッピングに不便。

「Analyze Data」をクリック。

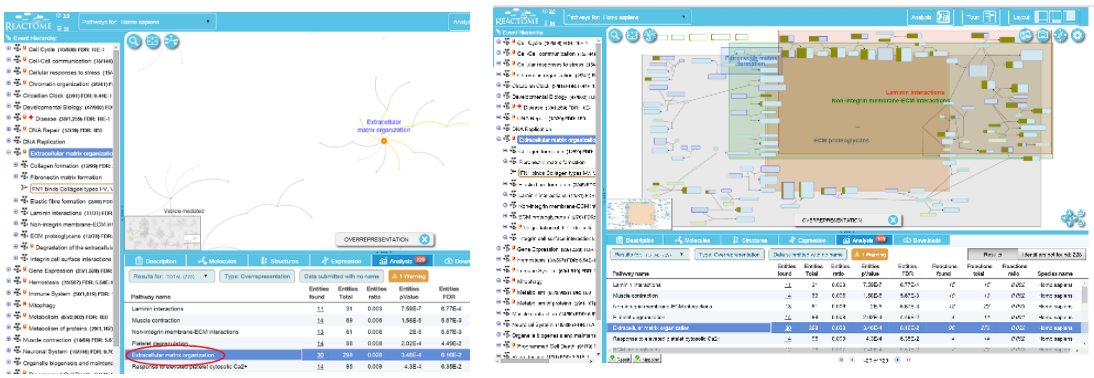
下図の三角をクリックすると入力ウィンドウが現れるので、遺伝子リスト（標準名など）をペースト。エクセルから直接コピー&ペーストしても良いが、余計な文字コードなどが入ることもあるので、ソート、重複の削除を行ってから、一度テキストファイルにコピーすると安全で、同時に行数を確認することもできる。



ひとまず「Project to human」を選択して、「Go」をクリック。

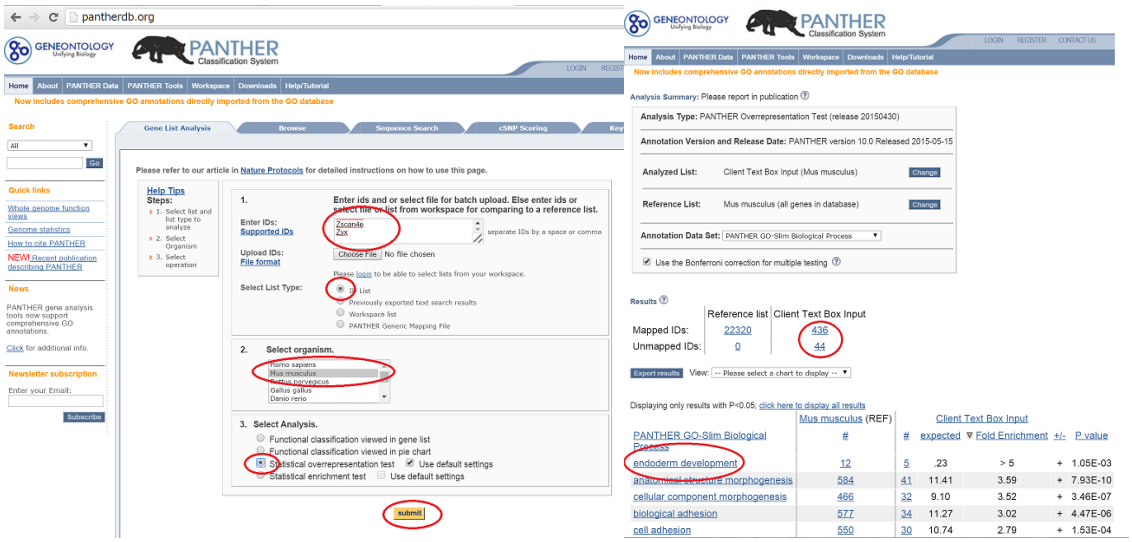


マッピングされなかった遺伝子が多く出るので、注意しておく。



結果の例。「Pathway Name」のところに出現頻度の高いカテゴリから表示される。下段のパスウェイの行をクリックすると、上図のように選択された機能カテゴリの階層（左ウィンドウ)、パスウェイのマップ上の位置（右上段）が表示される。マップ上の「ノード」(円の中心）をクリックすると、右上図のような詳しいマップが表示される。

2.4 PANTHER を使う（http://pantherdb.org/）

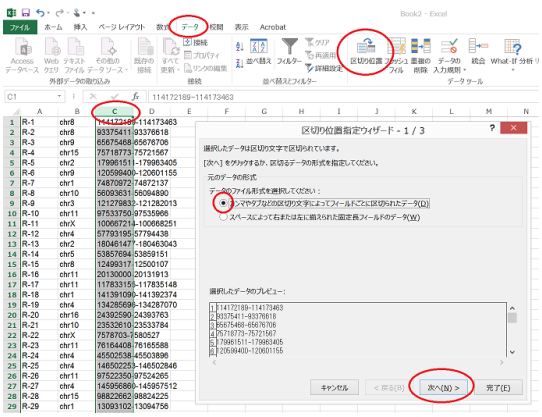


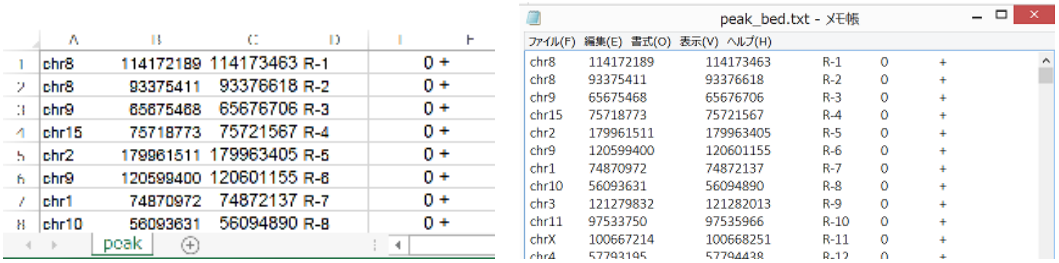
遺伝子リストを入力して、設定を選んで「Submit」をクリック。結果が表で表される。グラフで表す機能なども結果画面から選択できる。Prdm14 ノックダウンで発現が上昇している遺伝子を入力すると、そのまま「endoderm development」のパスウェイが出てくる。　このように、リストを目で見ただけではつかめない傾向を出してくれるのがパスウェイツールの便利な所。

2.5 GREAT ツールを使った ChIPseq ピーク領域のパスウェイ解析

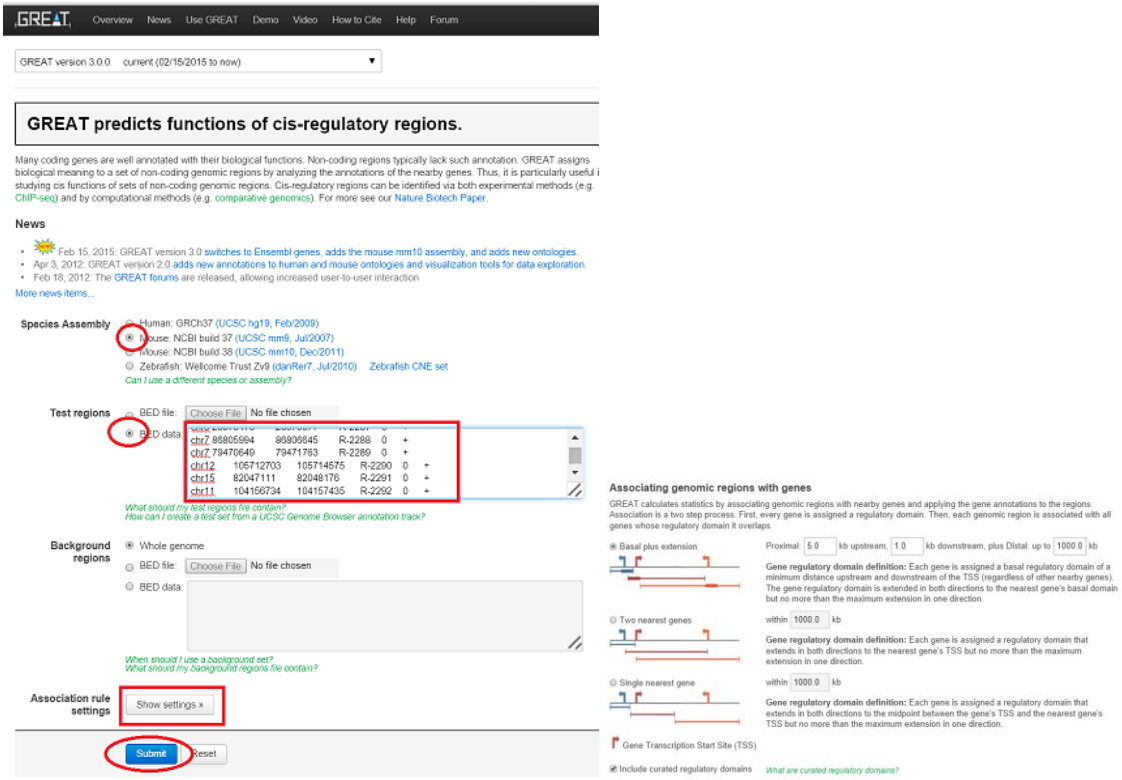
ChIPseq ピーク領域などは、機能分類ができないため、付近の遺伝子を見つけてから、その遺伝子の機能分類を解析することになる。GREAT ツールはそのような付近の遺伝子へのマッピングを行ってくれる。この場合、遺伝子リストではなく、ゲノム領域のリストがインプットとなるが、ゲノム領域は「BED フォーマット」で表される。まずは Prdm14 の ChIP ピーク領域を BED フォーマットに変換する。

エクセルで Table S3 を開き、新しいファイルにピーク領域をコピーする。

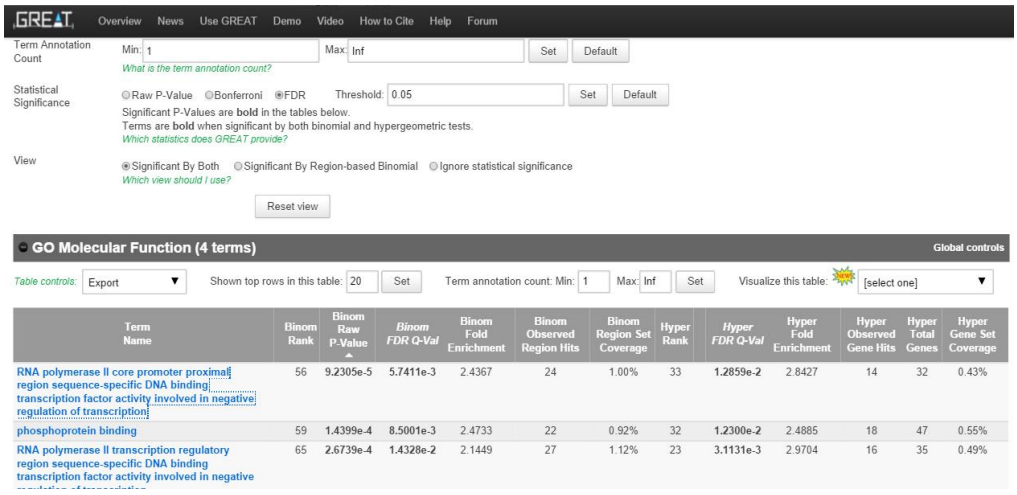




ピーク領域の Start と End がまとまっているので、カラム C を選択してから、「データ」＞「区切り位置」＞「カンマや…」と進み、「-」でカラムに分ける。ここで、ピーク領域を BED フォーマット（カラム 1 から 6 に、「chr」「start」「end」「name」「score」「strand」の順に並べる）に手作業で Cut & paste を繰り返しながら順序を整える。Score は「0」、Strand はすべて「＋」とする。これを「Peak」タブと名前を付ける。このタブをテキストファイルで保存する（peak_bed.txt）。



Google で「GREAT」＋「Genome」＋「Stanford」と検索し、GREAT ツールウェブサイトを開く。（http://bejerano.stanford.edu/great/public/html/index.php）mm9 を選択、BED データをペーストして、Submit する。Settings を変えても良い。



Prdm14 ピーク領域のアウトプット。

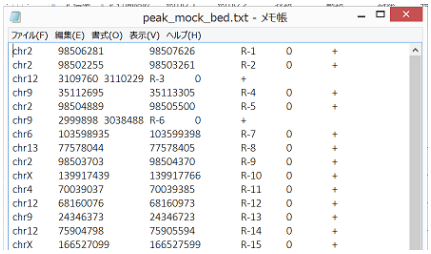
General control をクリックすると、結果表示の閾値を調節できる。TSS からの距離のヒストグラムの下のほうの表の「Term Name」をクリックして開くと、BED の領域のリストと遺伝子のリストが表示される。GREAT 解析ツールは様々なゲノム領域セットをインプットとして用いることができるが、オーバーラップしている領域を含むリストなどを用いるとオーバーカウントされるので、注意して使う。講習の後半で使う Galazy の Interval ツールを使って、重複している領域をまとめたりすることができる。

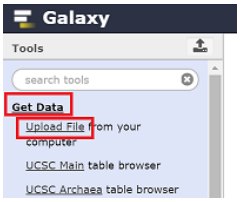
３．UCSC ブラウザと Galaxy を用いたデータの操作

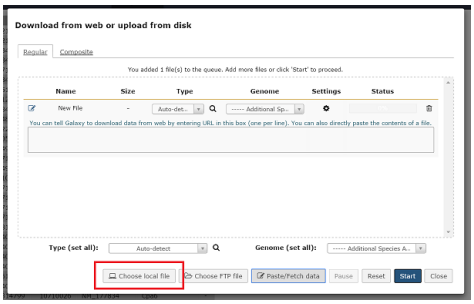
目標：RefSeq 遺伝子・トランスクリプトデータベースをダウンロードし、BED 形式で表されたゲノム上の位置情報を扱う。ChIPseq ピーク付近にある遺伝子を見つけ、RNAseq 解析の遺伝子セットと比較する。

3.1　Mock ChIPseq ピークの処理

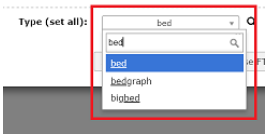
Prdm14 論文には、Flag-HA-Prdm14 を発現していない細胞で行った ChIPseq ピークが提供されていた。Prdm14 の ChIPseq ピークから、これらの領域を除いたものが特異的なピーク領域といえる。GREAT 解析のために Prdm14 ChIPseq ピークの BED ファイルを作成したのと同じ方法で、「mock_peak_bed.txt」ファイルを作成する。

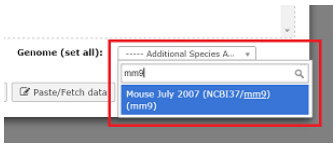


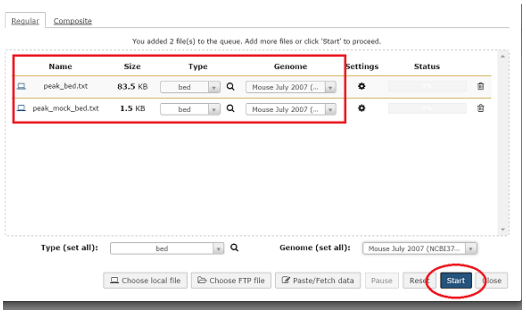




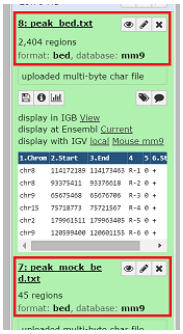
次に Galaxy ウェブサイトを開いてログインして、「Get data」＞「Upload files」を選択。「choose local file」で peak_bed.txt と mock_peak_bed.txt を選択する。

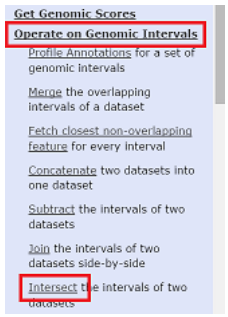


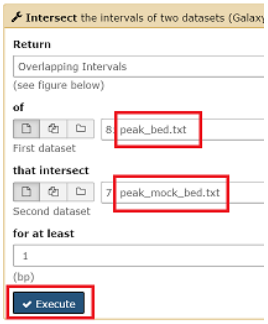


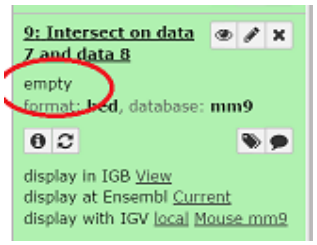


この時、Type を「bed」、Genome を「mm9」に設定し、「start」をクリック。





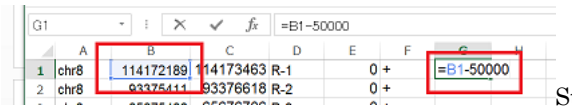




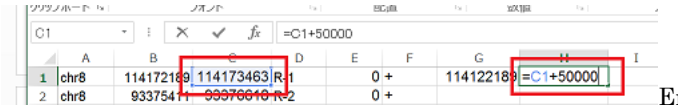
2 つの BED ファイルがアップロードされたら、行数を確認する。「Operate on Genomic Intervals」から「Intersect」を選択する。peak_bed.txt と mock_peak_bed.txt を選んで「Excute」をクリック。結果は「empty」なので、バックグラウンドのピークは無さそう。

3.2　ChIPseq ピーク領域を-/+50kb 前後に延長し、BED フォーマットに加工する

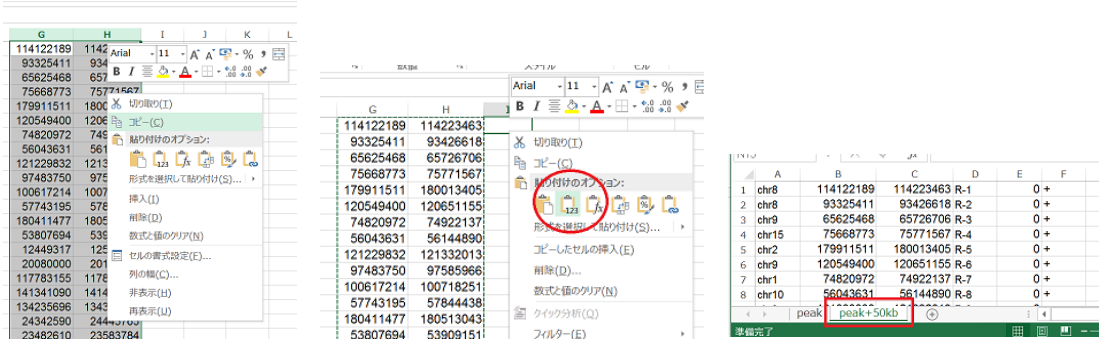
Prdm14 の BED ファイルを作成したエクセルファイルで、新しいタブ（peak+50kb）にコピーして、ピーク領域を前後に 50kb ずつ延長した BED データを作る。



Start 位置を-50,000 ベース移動。

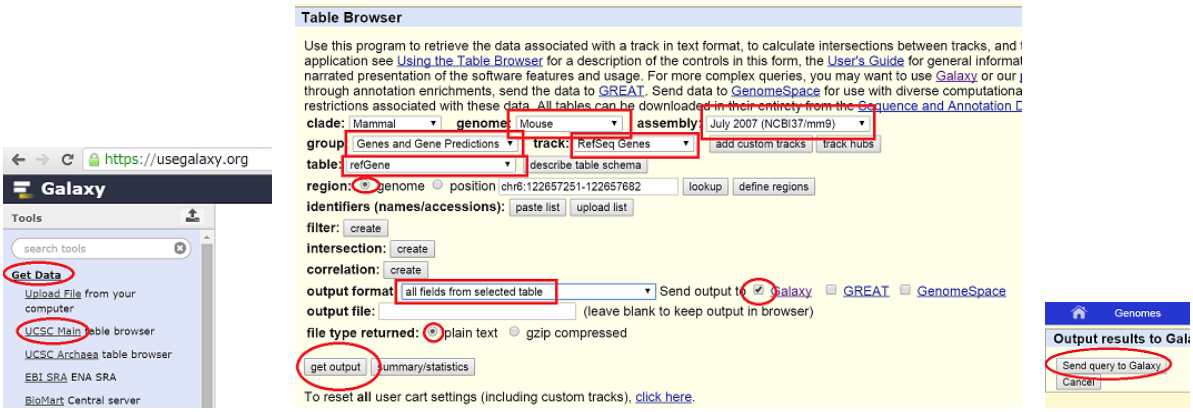


End 位置を+50,000 ベース移動。

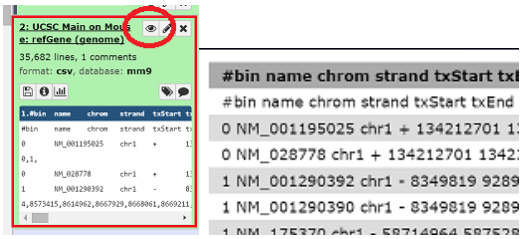


Cut & Paste を繰り返しながら、-/＋50kb 延長した BED データが出来た。このタブをテキストファイルで保存する（peak_50kb_bed.txt）。

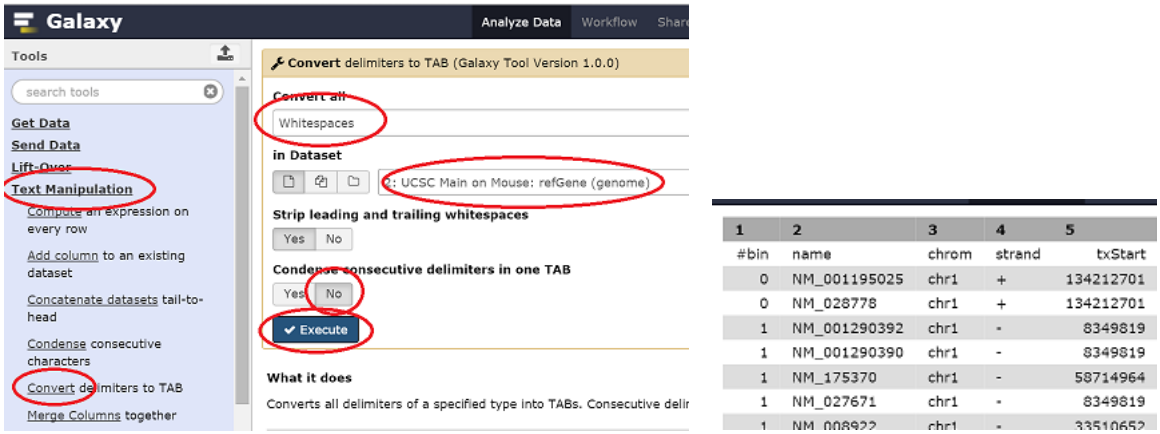
3.3　RefSeq トランスクリプトデータベースを Galaxy 上で加工する



左上のメニューから、「Get data」＞「UCSC main」を選択する。上右図のように RefSeq の全遺伝子について、データ全体を Galaxy にダウンロードする。

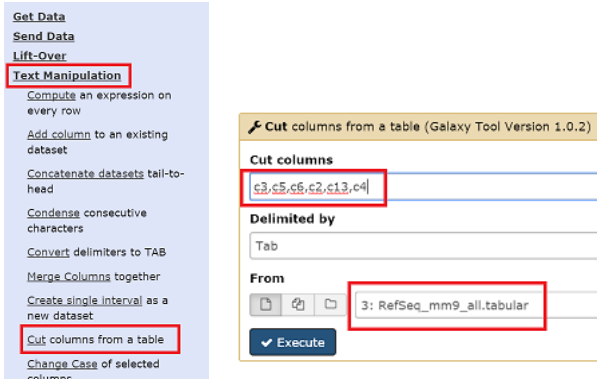


エントリーが緑色になったら、ダウンロード完了。データ名をクリックすると、先頭行のプレビューが見える。35,682 行のそれぞれがトランスクリプトのアノテーションを表す。さらに目のマークをクリックすると、中央のパネルにデータが表示される。

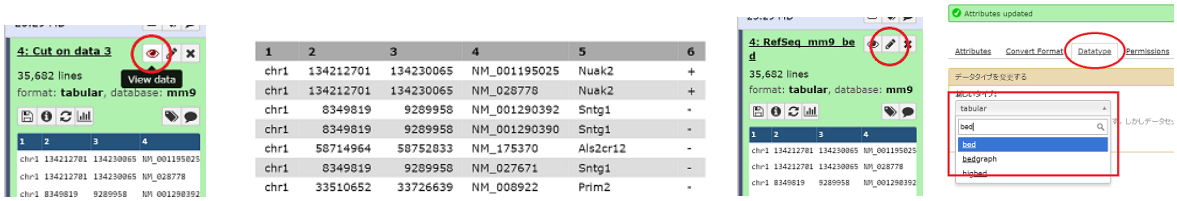


「Text manipulation」＞「Convert」を選び、RefSeq トランスクリプトデータのすべての Whitespace を Tab に変換する。再びアイマークをクリックして、変換が成功したことを確認する。

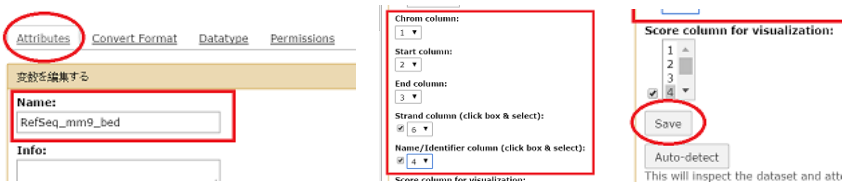
鉛筆マークをクリックして、名前を「RefSeq_mm9_all.tabular」などに変更する。ここでデータをスクロールしてみると、カラム２、３、４、５、６、１３が「RefSeq ID」、「chr」、「strand」「txStart」、「txEnd」、「Name2（標準名）」であることがわかる。こちらのカラムを「BED フォーマット」の形式に加工する。



「Text manipulation」＞「Cut」で、「RefSeq_mm9_all.tabular」を「c3,c5,c6,c2,c13,c4」で並べ替える。

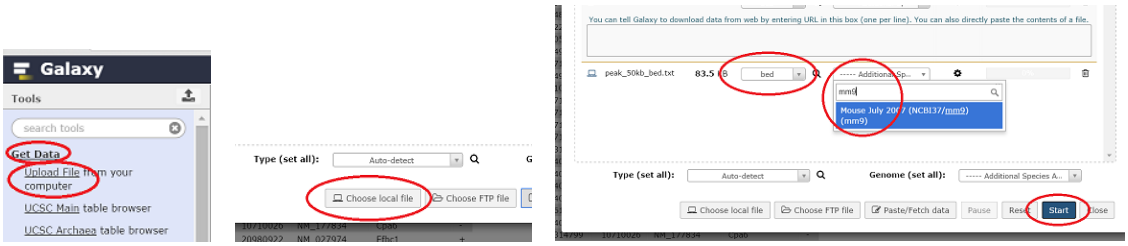


Cut が終わったら、アイマークをクリックして内容を確認する。このデータは「Tab」で区切られた表データ「Tabular」と認識されているので、「BED ファイル」であることを認識させる。「鉛筆マーク」をクリックし、「Datatype」を「bed」に変更する。

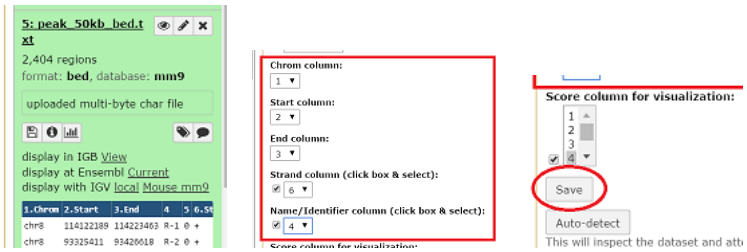


次に「Attributes」のタブを選択して、「Name」を新しくして、BED フォーマットの各カラムが正しく認識されるように意味を割り振っていく。最後に「Save」する。

3.4 ChIPseq ピークの BED ファイルを Galaxy にアップロードする



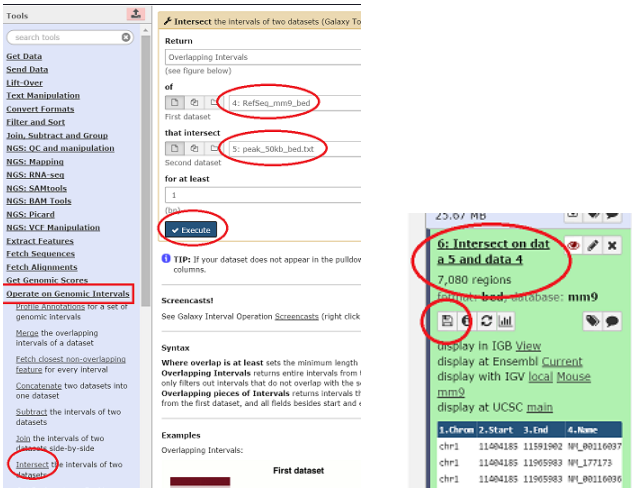
Galaxy メニューの「Get Data」＞「Upload file」から、peak_50kb_bed.txt を選び、bed ファイルであること、mm9 のゲノム位置であることを指定してアップロードする。



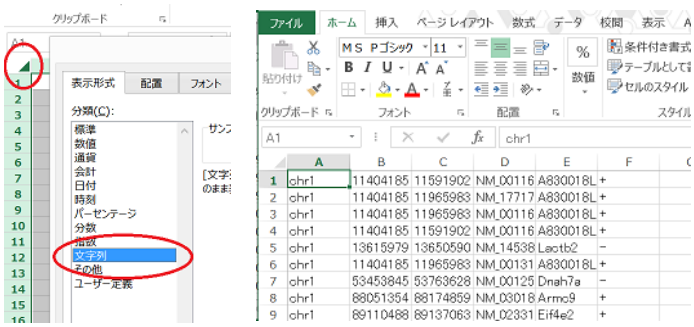
データがアップロードされたら、行数（2404）、フォーマット（bed）、アセンブリ（mm9）を再び確認する。

3.5 Galaxy の Interval ツールを使って 2 つの BED ファイルの重複領域を見つける

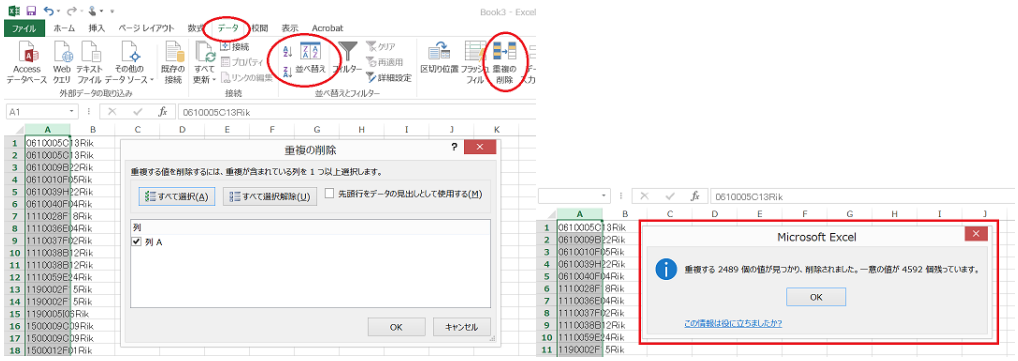
RefSeq トランスクリプトと 50kb ずつ延長した ChIP ピーク領域を比較して、重複するものがあれば、「ピークの 50kb 以内にある遺伝子」のリストが得られる。



結果が出たら、行数を確認する。7080 個のトランスクリプトがピーク付近にある。このファイルをディスクマークをクリックしてダウンロードする。ダウンロードしたファイルは、拡張子を「.txt」に変更して、テキストファイルとして開く。

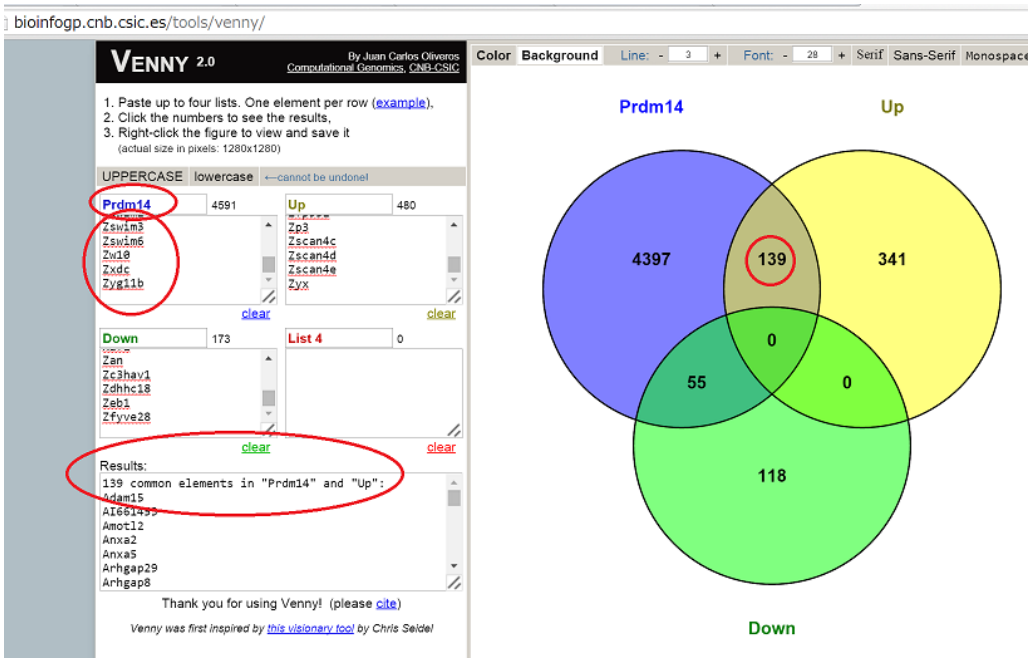


新しいエクセルドキュメントを開いて、<u>最初にすべてのセルの書式を「文字列」に変更する。</u>その後、7080 行のテキストファイルの内容をペーストする。



遺伝子名のカラムを残して、他の部分を消去する。「データ」から「並び替え」、「重複の削除」を行い、行数を確認する。ここで、ChIPseq ピークとオーバーラップする遺伝子をテキストファイルに保存しても良い。

3.6 Venny でベン図を作成して、Prdm14 の Direct target の候補遺伝子を探す



Venny (http://bioinfogp.cnb.csic.es/tools/venny/)を開き、Prdm14 の ChIPseq ピークの-/+50kb 以内にある遺伝子、Prdm14 のノックダウンで発現が上下に変化する遺伝子のリストをエクセルやテキストファイルからペーストする。ベン図の重なりの部分をクリックすると、「Results」の部分にその部分に該当する遺伝子リストが表示される。Venny では、重複するエントリーは自動的に除かれるが、エクセルなどで「データ」＞「重複の削除」を行っておくと行数も確認できて安心。今回は 50kb 以内で重複を見たが、条件を変化させて試してみる。

４．UCSC ブラウザと Galaxy を用いたデータの可視化

目標：パスウェイ解析やベン図で見つかった遺伝子について、ChIPseq ピークの形や RNAseq タグの様子を目視で確認する。また、UCSC ブラウザにピーク位置などをアップロードし、様々な情報を表示したイメージを作成する。

4.1　可視化に使うデータの種類

bedGraph ファイル：BED 形式のカラム１、２、３と４番目のカラムに「スコア」が付く。

「chr1　200　212　82」というタブ区切りの一行で、UCSC ブラウザ上では 1 番染色体の 200 から 212 まで、82 の高さにグラフが描かれる。下に示した例のように、先頭に「header」を付けないと UCSC ブラウザに正しく認識されないので注意する。解像度が維持されるので、RNAseq の Exson-Intron 境界などを見たい時には Wig ファイルではなく bedGraph ファイルが使われる。Galaxy の「Operate on Genomic Intervals」の「Intersect」などのツールを使って、見たい範囲だけの bedGraph ファイルを作るなど、加工に便利。

1	2	3	4
track type=bedGraph name=K27ac color=100,0,0			
chr17	3002096	3002357	1
chr17	3004153	3004177	1
chr17	3004177	3004414	2
chr17	3004414	3004438	1
chr17	3004463	3004602	1
chr17	3004602	3004724	2
chr17	3004724	3004828	1
chr17	3004828	3004863	2

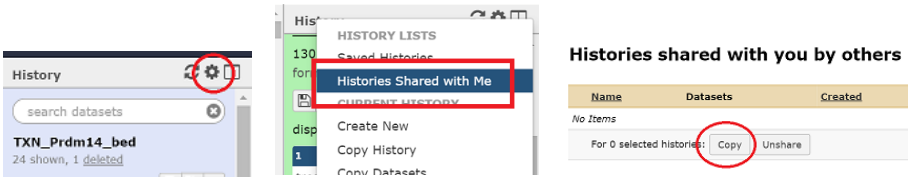
Wig ファイル：染色体ごとにヘッダーが付き、開始位置とスコアのみが各行に示される。等幅のステップを指定したものは、等間隔のスコアだけが並ぶので、bedGraph 形式よりもファイルサイズが小さい。テキストからバイナリ（ヒトには読めない形式）にしたものは BigWig ファイル形式と呼ばれ、さらにサイズが小さくできる。

1	2
track type=wiggle_0 name="Prdm14_chr17" description="counts for every 50 bp"	
variableStep chrom=chr17 span=50	
3002251	1
3002301	1
3002351	1
3004301	1.98425
3004351	1.98425
3004401	1.98425

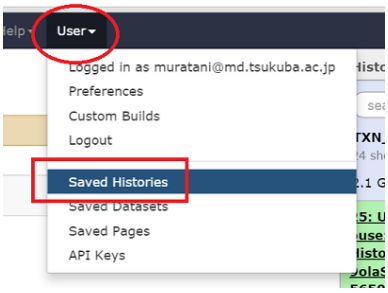
4.2　bedGraph ファイルから必要な部分を切り取って小さなファイルを作成し、ヘッダーを付ける

Prdm14 ChIPseq の全ゲノム分の bedGraph ファイルは 500Mb 程度もある。この場合、WWW サーバーからデータを読み込んだりする必要があり、やや複雑な操作が必要になる。200Mb 程度までのファイルであれば、「4.3」の項目で説明する方法で直接アップロードできる。

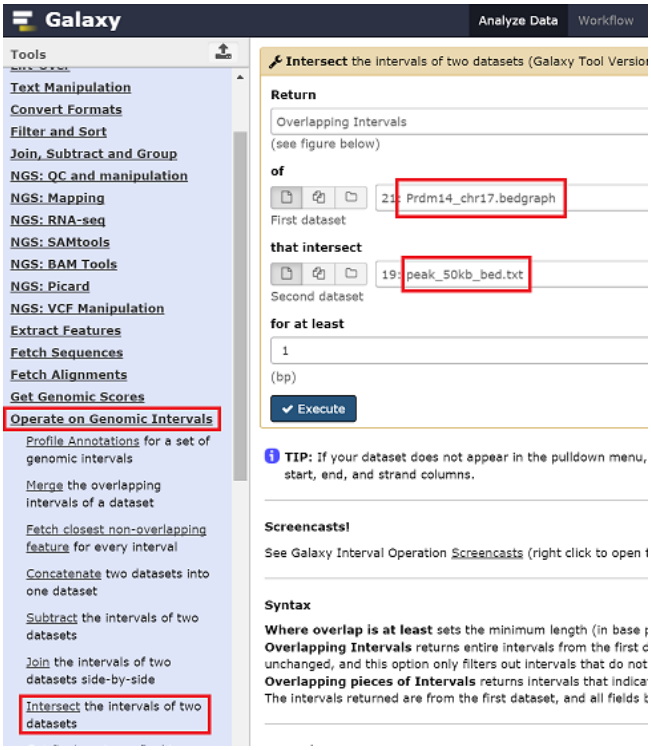
共同研究者などに、「ヘッダー無しの全ゲノム分の bedGraph ファイルを Galaxy にアップロードして、ヒストリーを共有してください」と伝えれば、その後は以下の作業ができる。この講習会では、速く処理できるように chr17 の部分だけを使うが、実際はファイルサイズを見ながら対応する。シークエンシングタグから bedGraph や wig ファイルを作るツールも Galaxy ツールに含まれている（Bedtools の GenomeCoverage、bedGraphToBigWig など）。



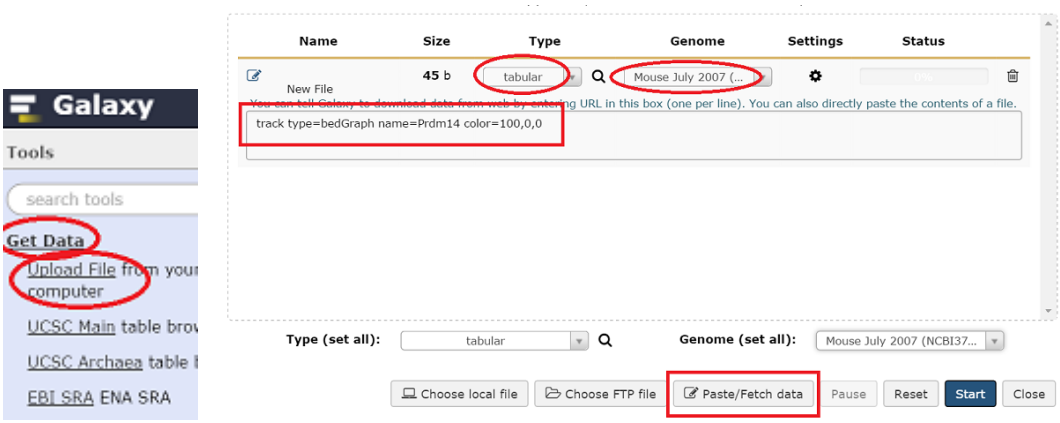
Galaxy の右パネルの上にある Setting アイコンをクリック、「Histories shared with me」をクリック。「Prdm14_bedgraph」があるはずなので、コピーする。



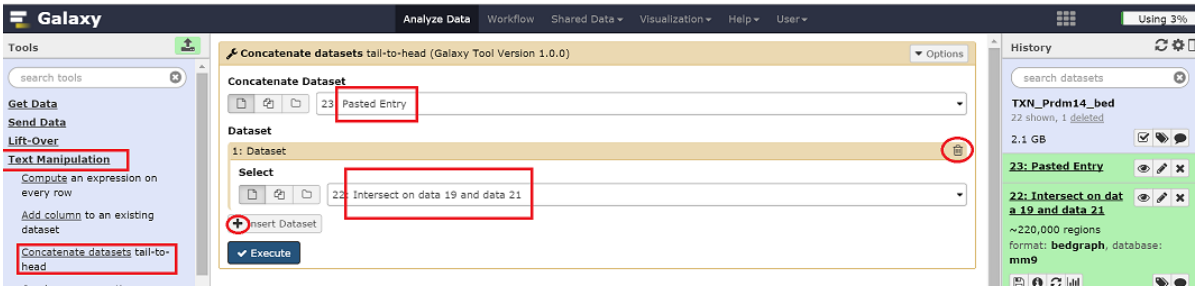
「User」＞「Saved Histories」と進むと、「Prdm14_bedgraph」がある。プルダウンメニューで「Switch」を選ぶと Prdm14_chr17.bedgraph のあるヒストリーに移動する。ここでもう一度「peak_50kb_bed.txt」（ベン図の時に作ったもの）をアップロードする。（他のヒストリーからコピーしても良い）



「Prdm14_chr17.bedgraph」から「peak_50kb_bed.txt」と重なる部分を選ぶ。「Operate on Genomic Intervals」＞「Intersect」ツールを使う。



「Get Data」＞「Upload File」を選び、「Paste/Fetch Data」をクリック。「track type=bedGraph name=Prdm14 color=100,0,0」と入力し、「Start」。bedGraph の「G」は大文字、余計なスペースは入れない。

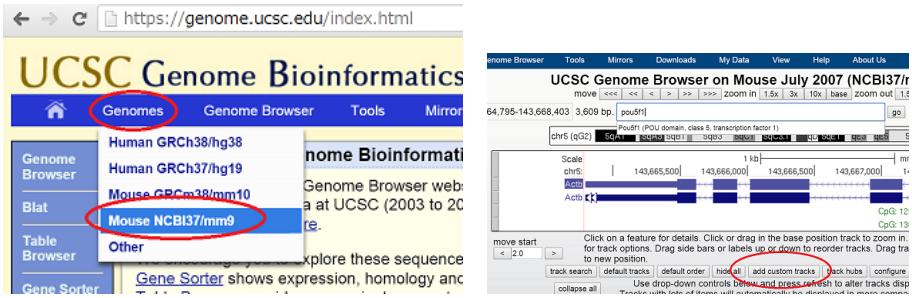


「Text manipulation」＞「Concatenate...」を選び、ヘッダーのライン「Pasted entry」の次に「Intersect」で作成した bedGraph ファイルを指定する。

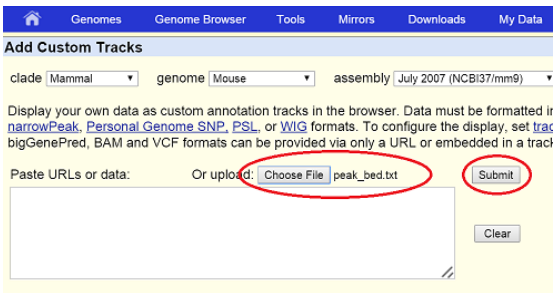
1	2	3	4
track type=bedGraph name=Prdm14 color=100,0,0			
chr17	5043286	5043439	1
chr17	5043460	5043617	1
chr17	5043617	5043721	2
chr17	5043721	5043878	1
chr17	5044111	5044196	1
chr17	5044196	5044258	3
chr17	5044258	5044372	4
chr17	5044372	5044457	3

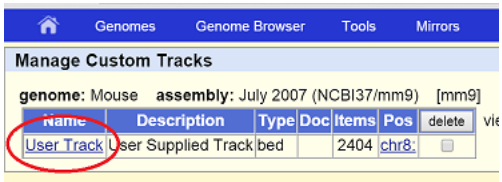
アイマークをクリックして内容を確認する。ヘッダーが一行、その下に 4 カラムの bedgraph データ。ディスクマークをクリック、ファイルをダウンロードして「Prdm14_chr17_peak.bedgraph」と名前を変更する。

4.3 UCSC ブラウザに可視化したいファイルをアップロードする

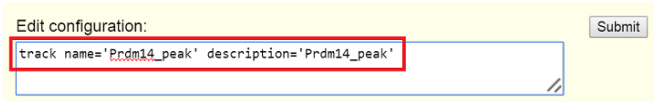


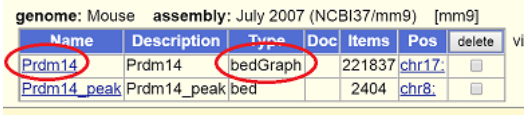
UCSC ブラウザを開き、「Genomes」から「mm9」アセンブリを選ぶ。「Add custom tracks」をクリク。



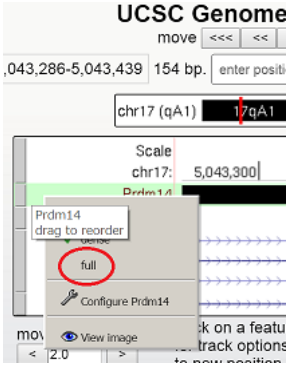


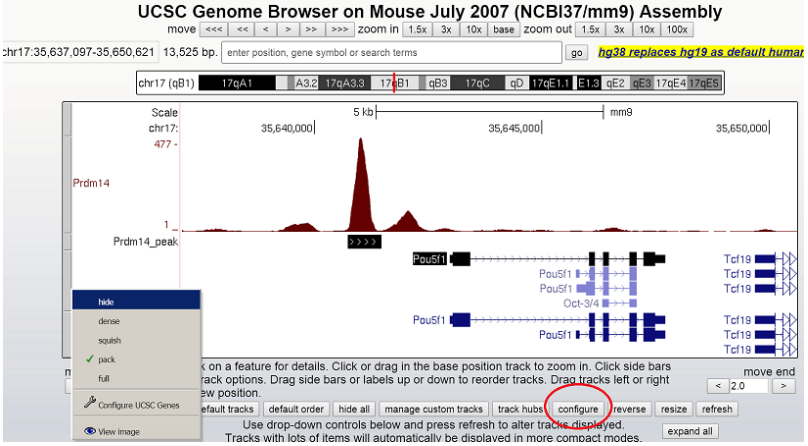
「Choose file」で「peak_bed.txt」を選び、「Submit」。ファイルが合ううぷロードされたら、「User Track」をクリックして、名前（track name='User Track'）を「track name='Prdm14_peak'」に変更する。



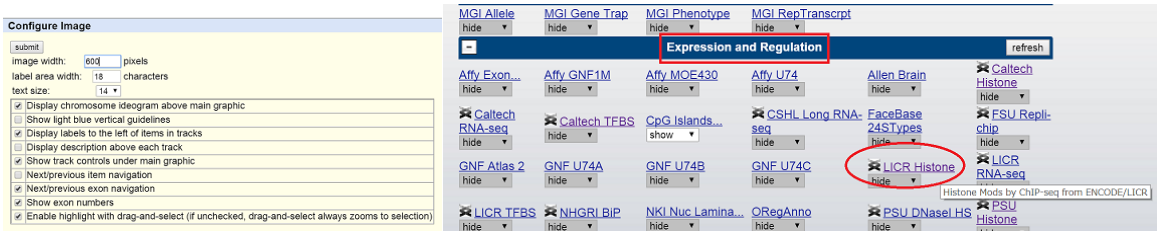


同様に「peak_50kb_chr17.bedgraph」もアップロードする。今度は「Type」が「bedGraph」になっていることを確認し、名前を「Prdm14」に変更。「Genome browser > Go」をクリックしてブラウザ画面に戻る。

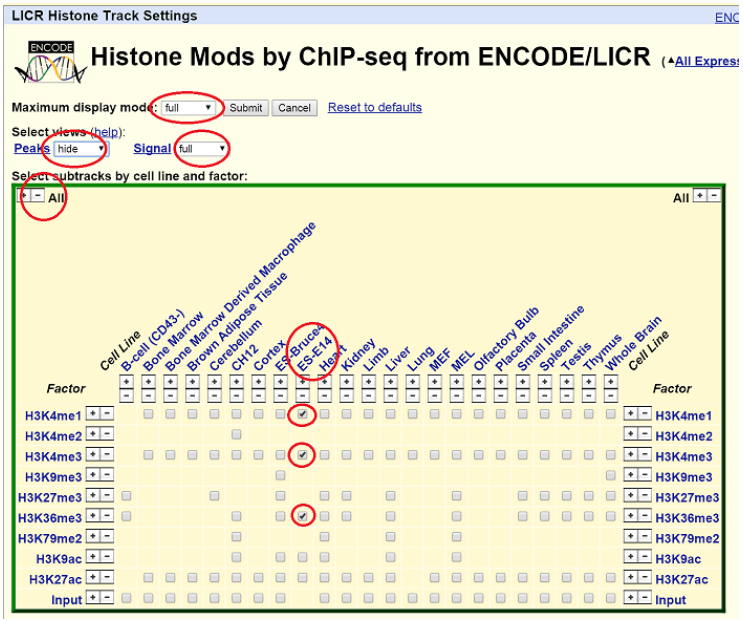


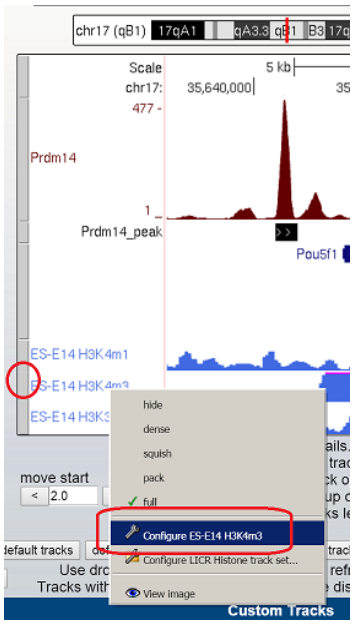


左端のグレーの枠を右クリックすると、トラックの見え方を変えられる。「Prdm14」のトラックを「Full」に。「Pou5f1」（Oct4）遺伝子を検索して、ズームアウトしていく。さらに各トラックの左端を右クリックで必要のないものを「hide」にしていくと、上の図のようになる。「Configure」をクリックすると、ブラウザのウィンドウサイズや青の縦線、フォントサイズなどを変更できる。

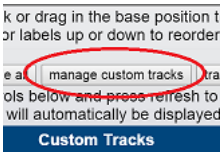


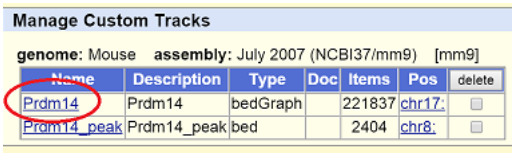
トラックコントロールを下のほうへ見ていくと、「Expression and Regulation」のところに「LICR Histone」がある。文字の部分をクリックすると、ENCODE の ChIPseq データセットがある。

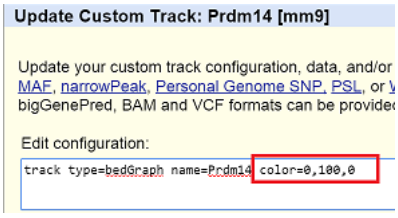




Prdm14 データはマウス ES 細胞なので、ヒストン修飾と比較してみる。ブラウザ画面の左端を右クリックで「Confiure」からスケールなどを調整できる。

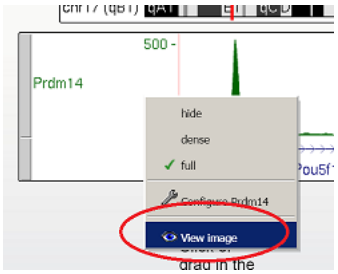


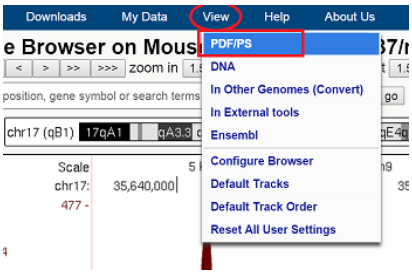


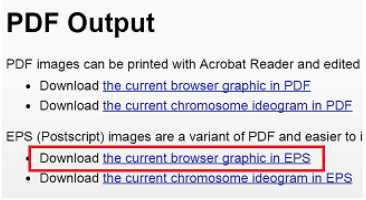


「Manage custom track」から Prdm14 トラックをクリック、「color=100,0,0」を「color=0,100,0」に変更すると、トラックが緑になる。「color=R,G,B」（数値は 0-255）の意味。

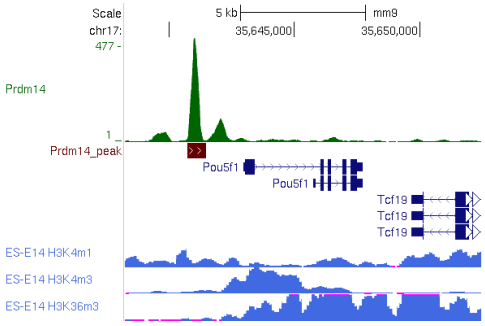
4.4 UCSC ブラウザの画像をダウンロードする

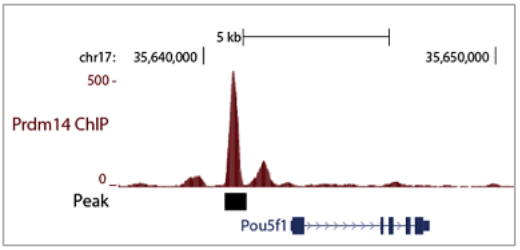






ブラウザ画面を右クリックすると、アイマークが出るので、そこからスクリーンショットが得られる。イラストレーターで編集できる画像を得るには、「View」＞「PDF/PS」、EPS ファイルを選択。





イラストレーターで「Ungroup」するとすべての要素が分解できるので、デザインを整えてフィギュアパネルに使う。

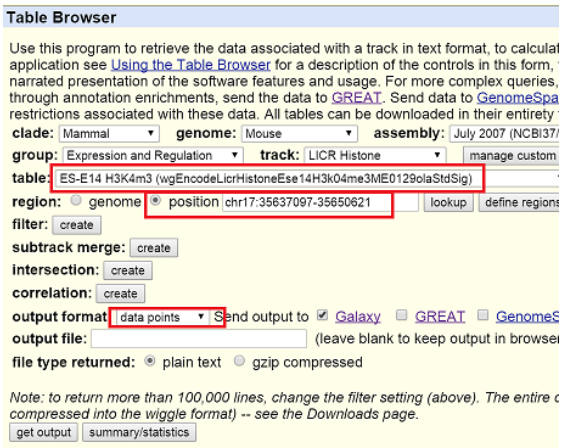
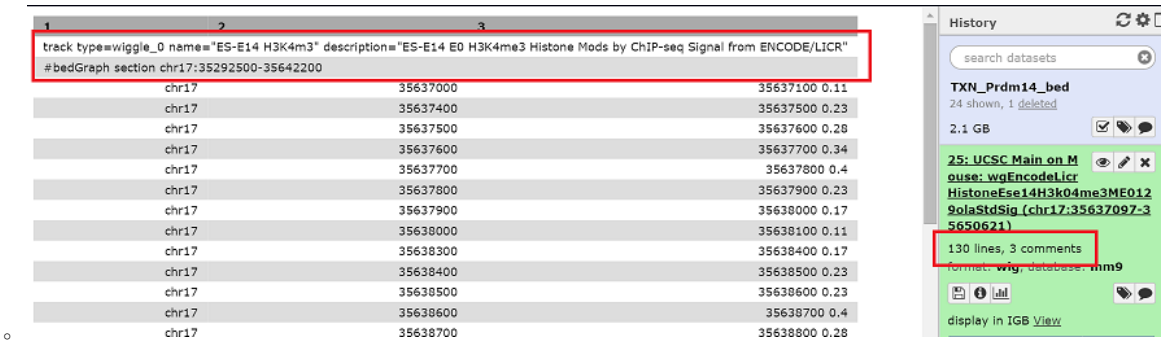


Table browser から ENCODE データをダウンロード



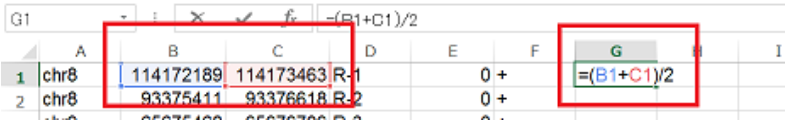
この部分の ENCODE ヒストン ChIPseq データは 130 行の bedGraph データなので、そのままテキストエディタでヘッダーを編集して、再度アップロードすれば、色の指定などが自在にコントロールできる。

５．ChIPseq ピークのモチーフ解析

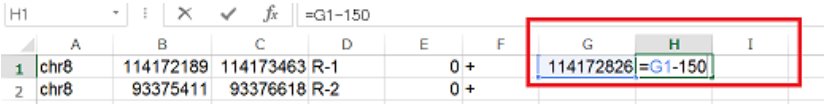
目標：ChIPseq ピーク領域に高頻度で見られる配列モチーフを検索する。モチーフ検索ツールへのインプットとして、長さの揃った塩基配列（最大 500 ベース程度）の FASTA ファイルを用いる。

5.1　ChIPseq ピーク領域の塩基配列の長さの揃った BED ファイルを作成する

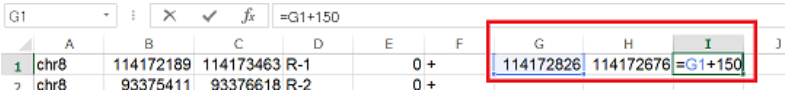
peak_bed.txt ファイルは、それぞれの領域の長さが異なるのでモチーフ解析には使えない。そこで、ピークの中心（summit）から両方向に 150 ベースずつ延長した、300 ベースの BED ファイルを作成する。



空きカラムに Start と End の平均を計算する

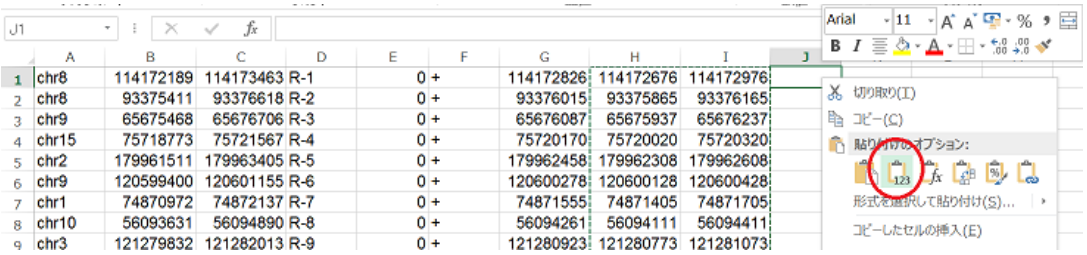


その隣に-150 ベースの Start 位置

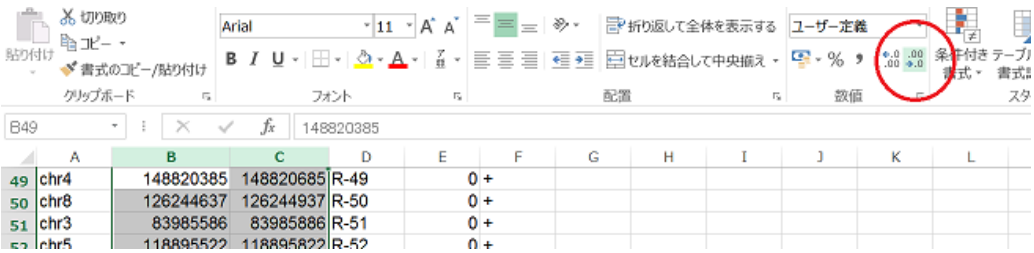


+150 ベースの End 位置を計算

この 3 つのカラムを「Control+C」でコピーしてから、BED データの最終列に「Control+↓」で 2404 行目まで移動して、カラム G、H、I 全体に数式をペーストする。



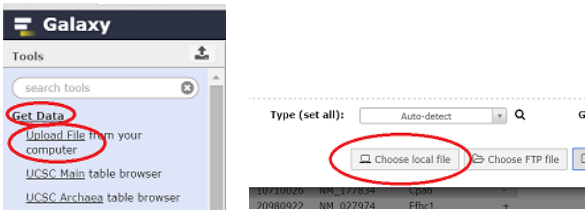
数値だけをコピー&ペーストして、

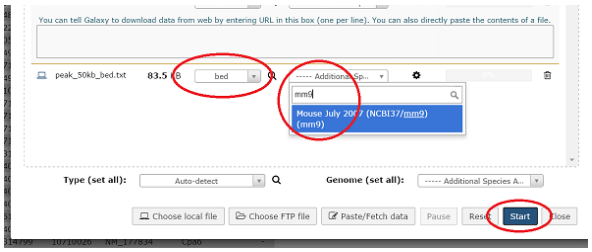


最終的に各ピークの中心 300 ベースを指定する BED ファイルを作成する。Start と End のカラムが整数のみになるように、小数点以下を切り捨ててから、「peak_summit_300_bed.txt」で保存する。

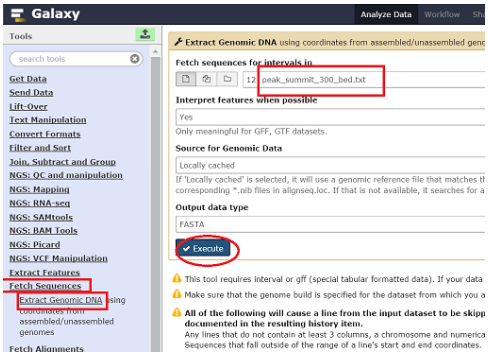
5.2 ChIPseq ピーク領域中心 300 ベースの FASTA ファイルを作成する

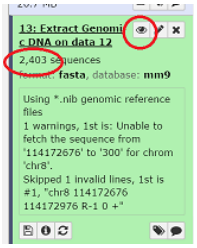
Galaxy に「peak_summit_300_bed.txt」をアップロードする。

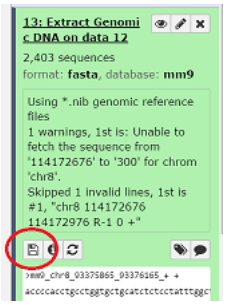




Galaxy メニューの「Get Data」＞「Upload file」から、peak_summit_300_bed.txt を選び、bed ファイルであること、mm9 のゲノム位置であることを指定してアップロードする。データがアップロードされたら、行数（2404）、フォーマット（bed）、アセンブリ（mm9）を再び確認する。



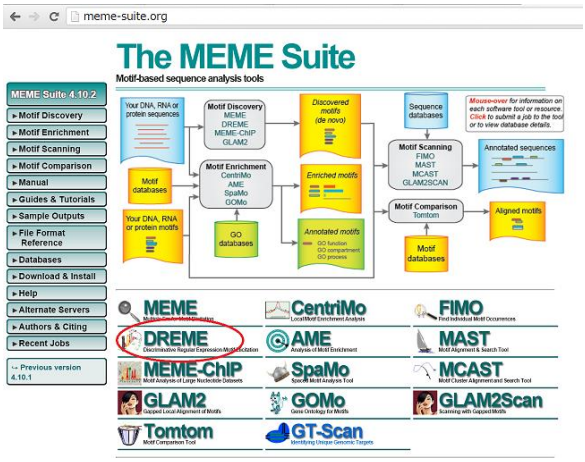


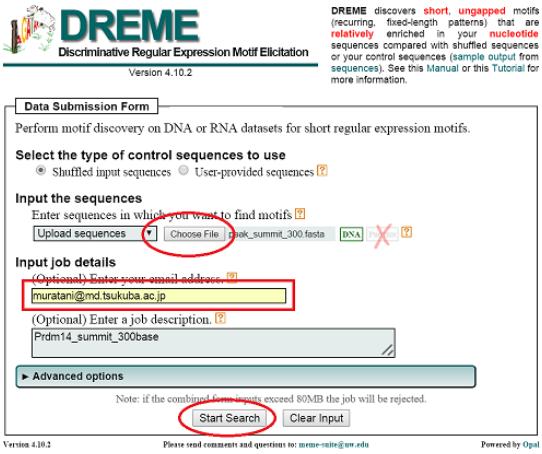


「Fetch Sequences」の「Extract Genomic DNA」を選択。アップロードした「peak_summit_300_bed.txt」を選択して「Execute」。リファレンスゲノム配列から塩基配列を抽出して、FASTA ファイルが出来上がる。行数が 1 行少ないが、この原因は不明。問題は 1 行目なので、エクセルのコードが残っていた可能性がある。このような場合は、テキストファイルを作り直す、などで対処してみる。ディスクマークをクリックして、このファイルをダウンロードする。「peak_summit_300.fasta」と名前を変更する。

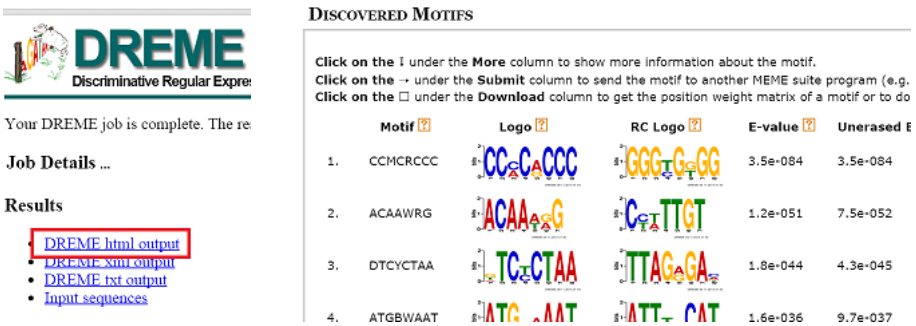
5.3 DREME でモチーフ解析、TOMTOM でモチーフ検索を行う

MEME Suite（http://meme-suite.org/ 又は、http://alternate.meme-suite.org/）を開く。

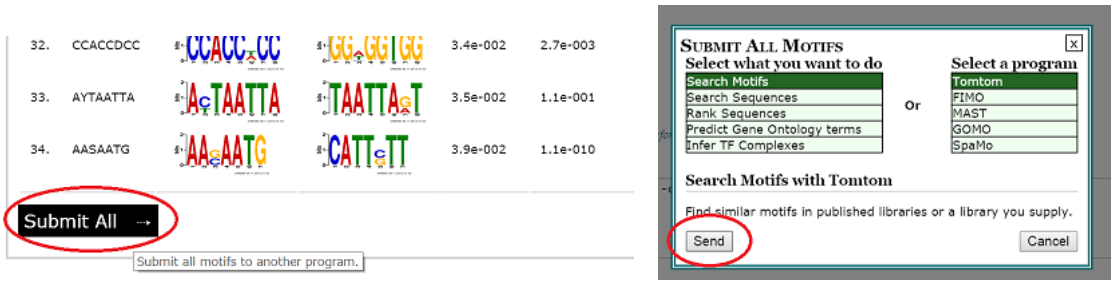




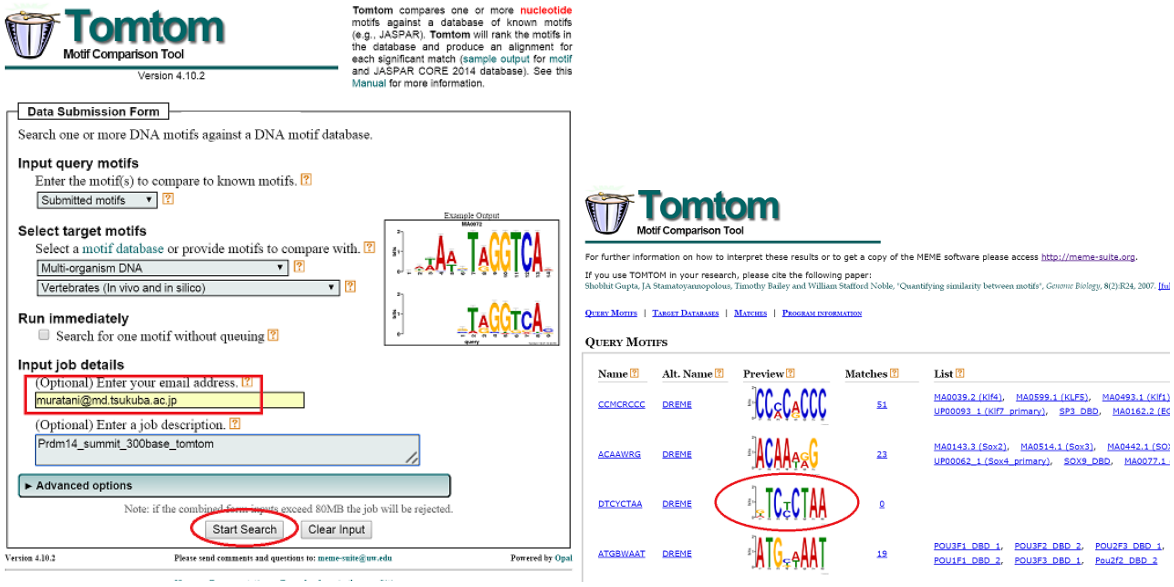
DREME をクリックして、FASTA ファイルをアップロード、メールアドレスを入力して、「Start Search」。



しばらく待つと結果のページが現れる。「html output」をクリックすると、モチーフロゴが見える。



これだけではどんな転写因子なのかわからないので、このモチーフにマッチする転写因子を検索する。結果出力ページの下のほうに、「Submit All」があるのでクリックする。TOMTOM をハイライトして「Send」をクリック。



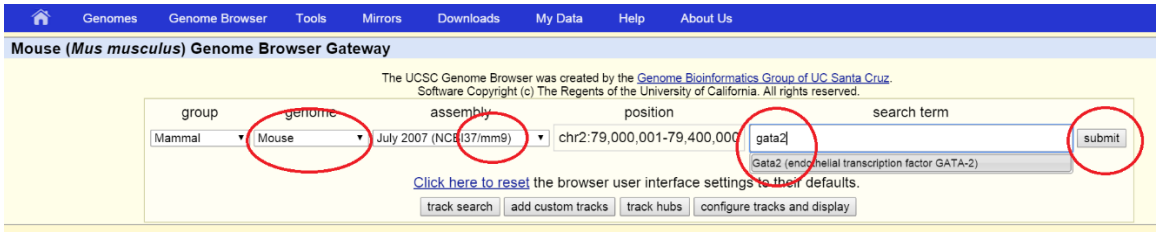
DREME のアウトプットが既に入っているので、メールアドレスを入力して「Start Search」をクリック。結果が返ってくる。Oct、Sox、Klf モチーフと未知のモチーフがトップヒットとして出てくる。このデータベースに無いモチーフ（TCTCTAA）を Prdm14 論文は Figure 6 で解析している。モチーフロゴはイメージなので、そのままパワーポイントスライド等にペーストできる。

６．プロモーター領域のモチーフ解析

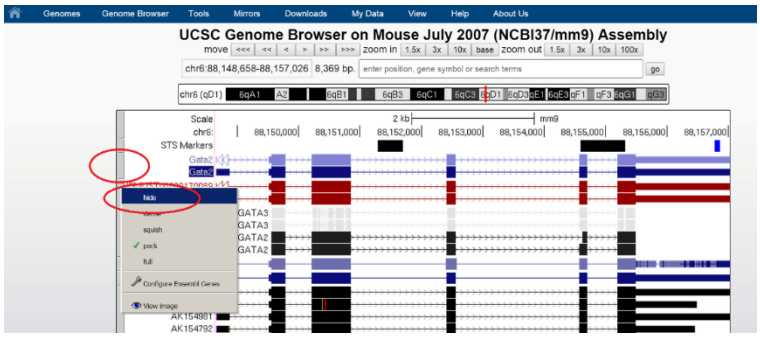
目標：RNA 発現解析で変化した遺伝子群のプロモーター領域に高頻度で見られる配列モチーフを検索する。トランスクリプトデータベースをもとに、転写開始点付近の BED ファイルを作成し、さらに FASTA ファイルを作成する。

6.1　プロモーター領域の塩基配列の長さの揃った BED ファイルを作成する（Gata2 遺伝子付近を使った練習）

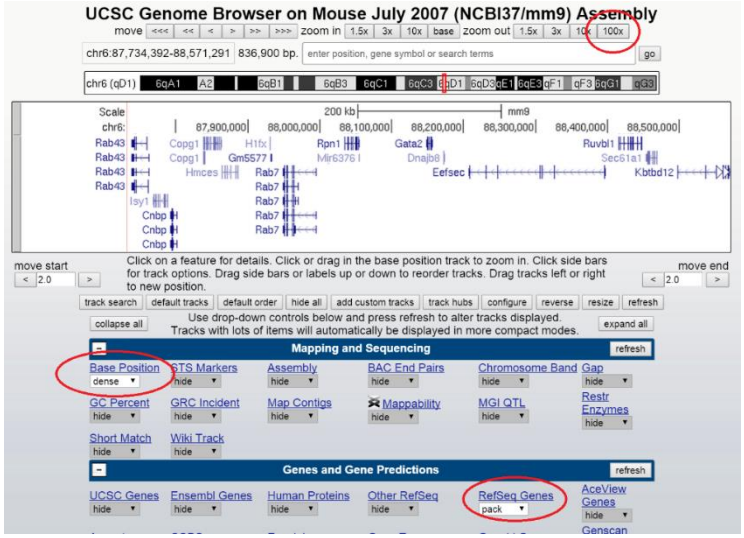
UCSC ブラウザを開き、「Genomes」をクリックすると、下図のプルダウンメニューが現れるので、「Mouse」「mm9」を選択、「Search term」のところに遺伝子名を入れると候補が表示されるので選択する。ここでは「gata2」と入力。Gata2 遺伝子の名前が出るので、マウス操作で選択して、「Submit」をクリック。



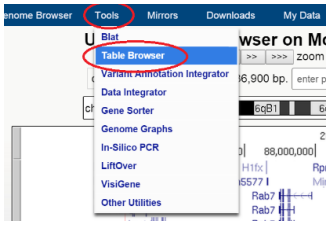
Gata2 遺伝子の付近が表示される。



いろいろなトラックが表示されているので、左端のグレーのバーを右クリック、「Hide」を選択して消していく。



トラックの表示は下のほうのプルダウンメニューでも操作できる。「Base Position」を「Dense」、「RefSeq Genes」を「Pack」にして、2 つのトラックだけが表示されたところで、「Zoom out」の「100x」をクリックして Gata2 周辺領域も見てみる。いくつかのトランスクリプトが見える。

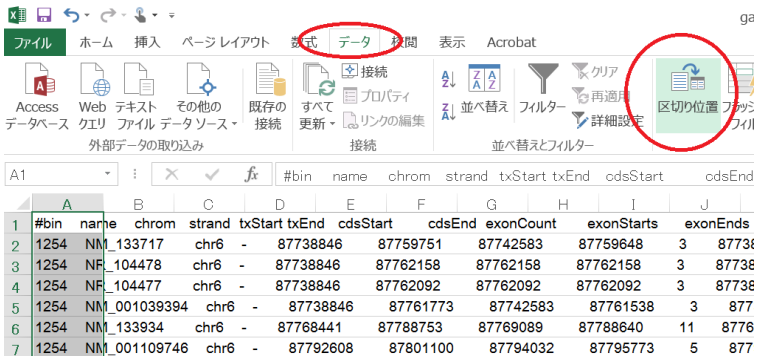


「Tools」から「Table Browser」を選ぶと、下図のプルダウンメニューが現れる。

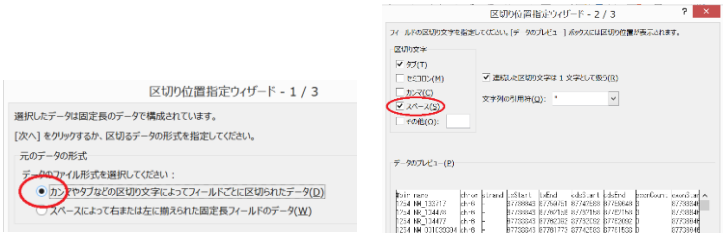
「RefSeq Genes」の「refGene」テーブルを、今表示されている範囲で取得するため、「Position」をチェック。「Galaxy」のチェックは、ここでは外しておく。ブラウザ上で出力するため、ファイル名は空欄に。「Plain text」で出力。

```
#bin    name    chrom   strand  txStart txEnd   cdsStart        cdsEnd  exonCount
1254    NM_133717       chr6    -       87738846        87759751        87742583
cmpl    1,0,0,
1254    NR_104478       chr6    -       87738846        87762158        87762158
unk     -1,-1,-1,
1254    NR_104477       chr6    -       87738846        87762092        87762092
unk     -1,-1,-1,
1254    NM_001039394    chr6    -       87738846        87761773        87742583
cmpl    1,0,0,
1254    NM_133934       chr6    -       87768441        87788753        87769089
87768441,87769550,87771501,87772333,87775364,87777577,87783635,87784421,87784936,8
87769197,87769637,87771623,87772456,87775482,87777690,87783678,87784487,87784988,8
1254    NM_001109746    chr6    -       87792608        87801100        87794032
87794150,87795314,87795551,87795787,87801100,   0       Cnbp    cmpl    cmpl    2,
```

画面いっぱいにテキストが出てくるので、「Control+A」、「Control+C」で全文選択、コピー、新しいエクセルファイルに「Control+V」でペースト。テキストラインとして認識されているので、「データ」、「区切り位置」メニューを選んでカラムに変換する。



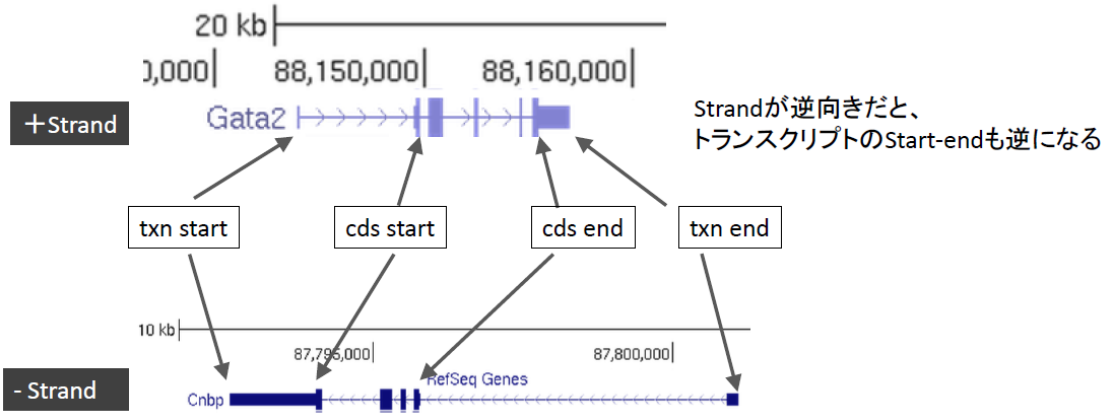
スペースで区切る。



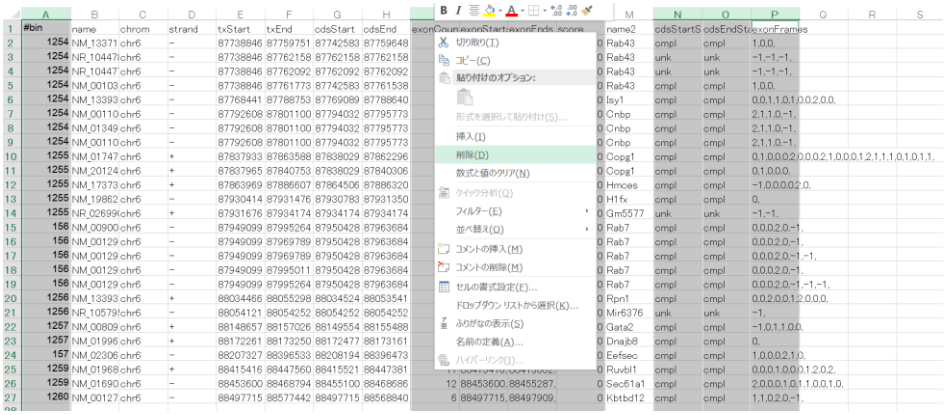
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	#bin	name	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	exonCoun	exonStart	exonEnds	score	name2	cdsStartS	cdsEndSta	exonFrames	
2	1254	NM_13371	chr6	-	87738846	87759751	87742583	87759648	3	87738846,	87742834,	0	Rab43	cmpl	cmpl	1,0,0,	
3	1254	NR_10447	chr6	-	87738846	87762158	87762158	87762158	3	87738846,	87742834,	0	Rab43	unk	unk	-1,-1,-1,	
4	1254	NR_10447	chr6	-	87738846	87762092	87762092	87762092	3	87738846,	87742834,	0	Rab43	unk	unk	-1,-1,-1,	
5	1254	NM_00103	chr6	-	87738846	87761773	87742583	87761538	3	87738846,	87742834,	0	Rab43	cmpl	cmpl	1,0,0,	
6	1254	NM_13393	chr6	-	87768441	87788753	87769089	87788640	11	87768441,	87769197,	0	Isy1	cmpl	cmpl	0,0,1,1,0,1,0,0,2,0,0,	
7	1254	NM_00110	chr6	-	87792608	87801100	87794032	87795773	5	87792608,	87794150,	0	Cnbp	cmpl	cmpl	2,1,1,0,-1,	
8	1254	NM_01349	chr6	-	87792608	87801100	87794032	87795773	5	87792608,	87794150,	0	Cnbp	cmpl	cmpl	2,1,1,0,-1,	
9	1254	NM_00110	chr6	-	87792608	87801100	87794032	87795773	5	87792608,	87794150,	0	Cnbp	cmpl	cmpl	2,1,1,0,-1,	

今回使うのは赤線で囲ったカラム。ここから先はしばらくエクセルの操作になる。

最終的に、プロモーター（転写開始点から上流 200 ベース、下流 100 ベース）、または、３末端 UTR（cDNA の End からトランスクリプトの End まで）の場所を BED ファイル形式にまとめる。



１．B-H、M 以外のカラムを削除する。



２．シートを複製して、「Promoter」と名前を付ける

３．「Strand」のカラムを選択して、ソートする。「 - 」が上半分、「＋」が下半分、など。

４．ワークシートの右側の空いている部分に BED フォーマットのカラムを作っていく。

５．最初に「chr」のカラムをコピー&ペースト。

６．次に「Start」のカラムを作る。「＋」の側では、「Start」の位置が転写開始点なので、新しいカラムに「＝D2-200」とすると、スタートの位置が転写開始点の 200 ベース上流、となる。

J2 =D2-200

	A	B	C	D	E	F	G	H	I	J	K
1	name	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	name2	chrom		
2	NM_01747	chr6	+	87837933	87863588	87838029	87862296	Copg1	chr6	=D2-200	
3	NM_20124	chr6	+	87837965	87840753	87838029	87840306	Copg1	chr6		
4	NM_17373	chr6	+	87863969	87886607	87864506	87886320	Hmces	chr6		
5	NR_02699(	chr6	+	87931676	87934174	87934174	87934174	Gm5577	chr6		
6	NM_13393	chr6	+	88034466	88055298	88034524	88053541	Rpn1	chr6		

7．「End」のカラムを作り、「＝D2+100」とする。

8．残りの「＋」Strand の部分に同じ式をコピーする。

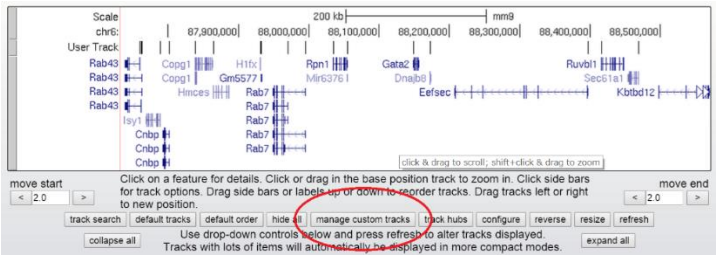
9．「 - 」のストランドも同様に新しいカラムを作成し、残りの行にコピーする。

１０．BED フォーマットの残りの４つのカラムをコピペで作成する。

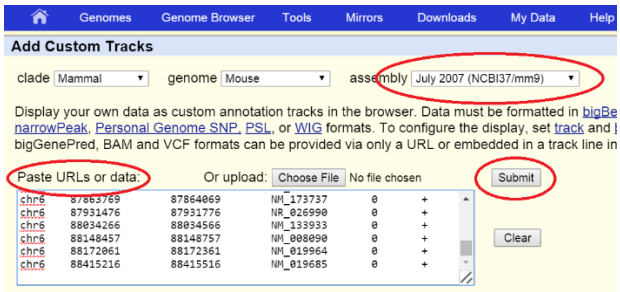
出来上がった BED フォーマット。（テキストファイルに保存すれば BED ファイル）

	A	B	C	D	E	F
1	chrom	start	end	name		strand
2	chr6	87759651	87759951	NM_13371	0	-
3	chr6	87762058	87762358	NR_10447	0	-
4	chr6	87761992	87762292	NR_10447	0	-
5	chr6	87761673	87761973	NM_00103	0	-
6	chr6	87788653	87788953	NM_13393	0	-
7	chr6	87801000	87801300	NM_00110	0	-
8	chr6	87801000	87801300	NM_01349	0	-
9	chr6	87801000	87801300	NM_00110	0	-
10	chr6	87931376	87931676	NM_19862	0	-
11	chr6	87995164	87995464	NM_00900	0	-
12	chr6	87969689	87969989	NM_00129	0	-
13	chr6	87969689	87969989	NM_00129	0	-
14	chr6	87994911	87995211	NM_00129	0	-
15	chr6	87995164	87995464	NM_00129	0	-
16	chr6	88054152	88054452	NR_10579	0	-
17	chr6	88396433	88396733	NM_02306	0	-
18	chr6	88468694	88468994	NM_01690	0	-
19	chr6	88577342	88577642	NM_00127	0	-
20	chr6	87837733	87838033	NM_01747	0	+
21	chr6	87837765	87838065	NM_20124	0	+
22	chr6	87863769	87864069	NM_17373	0	+
23	chr6	87931476	87931776	NR_02699	0	+
24	chr6	88034266	88034566	NM_13393	0	+
25	chr6	88148457	88148757	NM_00809	0	+
26	chr6	88172061	88172361	NM_01996	0	+
27	chr6	88415216	88415516	NM_01968	0	+
28						

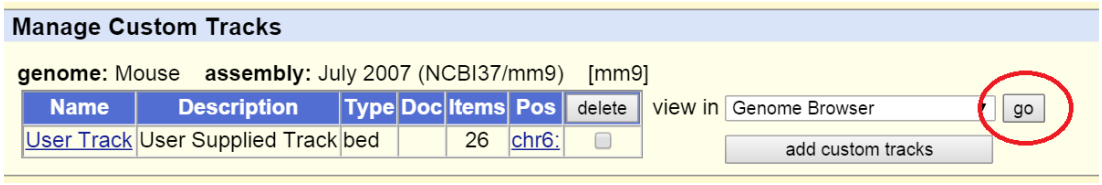
UCSC ブラウザの「Manage custom tracks」をクリック。



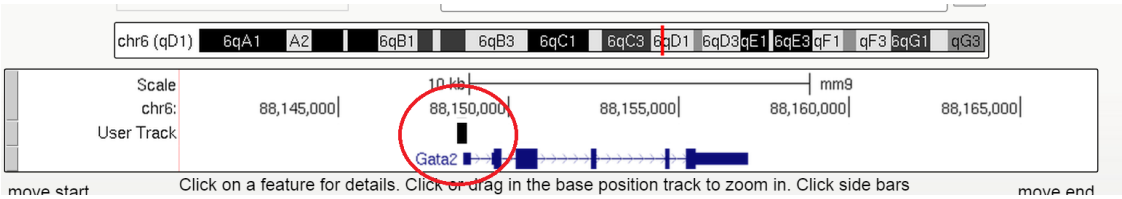
最初の行を除いた部分をエクセルからコピー＆ペースト。mm9 のアセンブリを確認して、「Submit」。



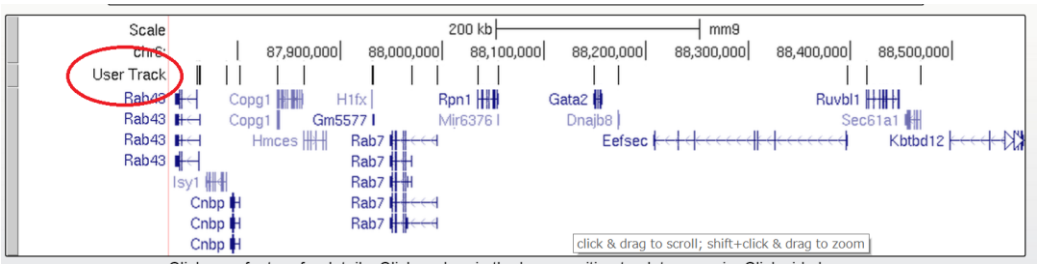
データがアップロードされる。



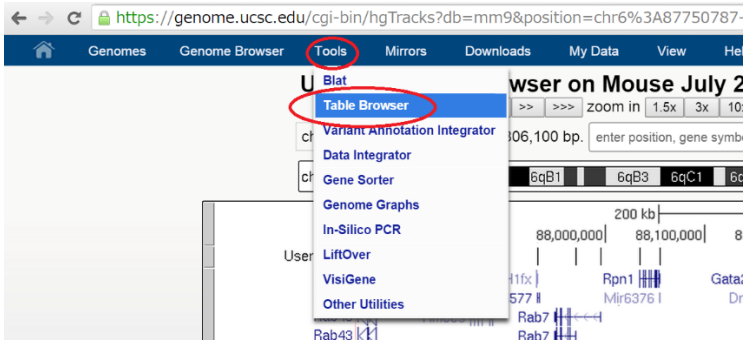
「Genome browser」を指定して「Go」をクリックするとランダムな場所がブラウザに現れるので、Gata2 遺伝子周辺を見てみる。BED 形式で指定された場所に黒いバーが見える。



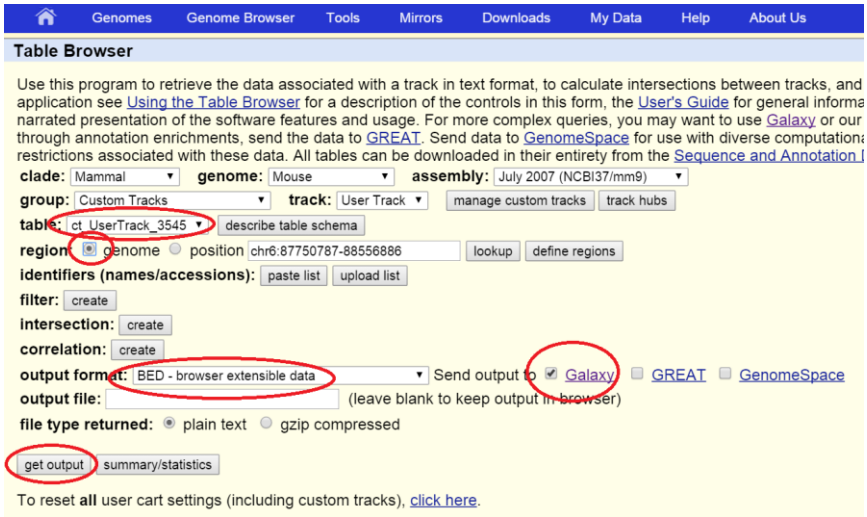
さらにズームアウトして、Strand が逆の遺伝子も正しくプロモーター領域が指定されているかを確認する。



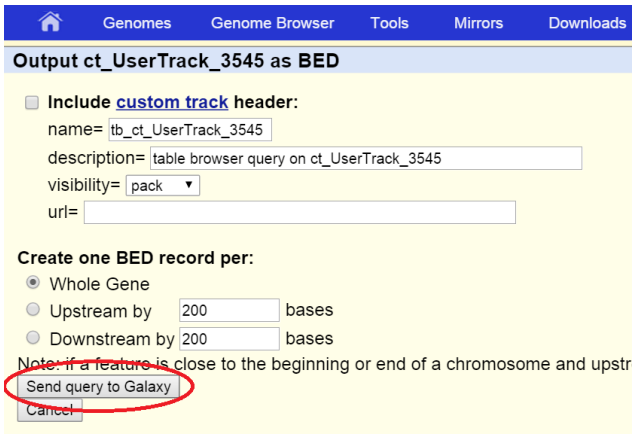
「User Track」に入っている自分のデータを「Table Browser」からダウンロードしてみる。



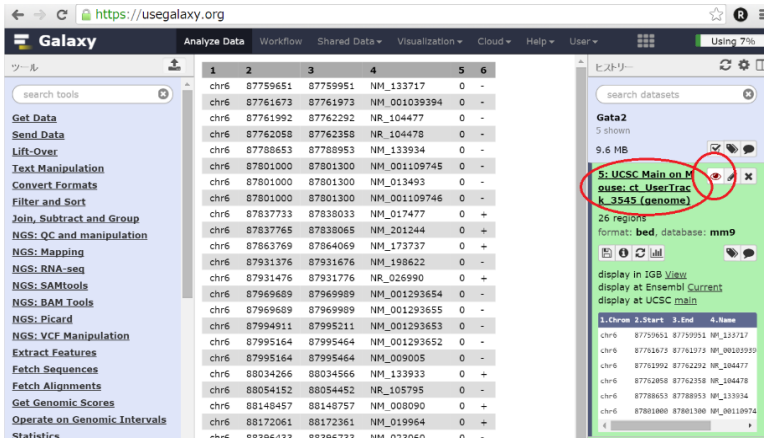
プルダウンメニューで下図のように指定。



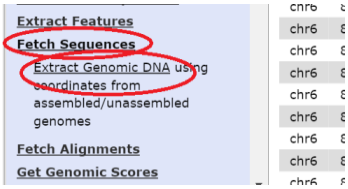
クエリを Galaxy に送る。



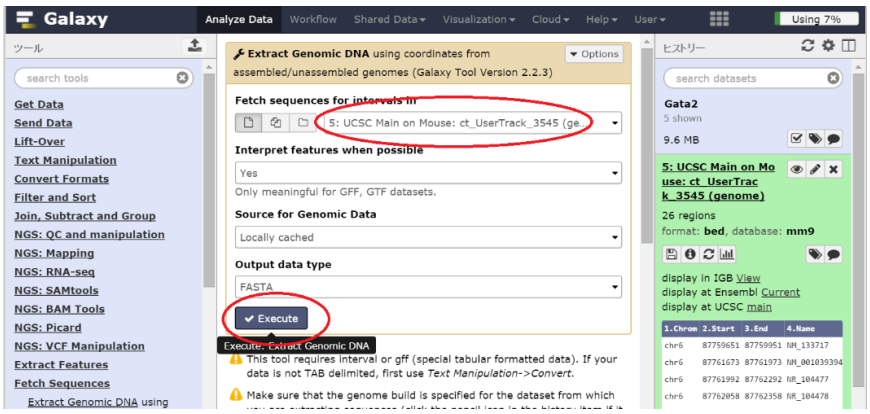
Galaxy の右パネルにデータが現れる。



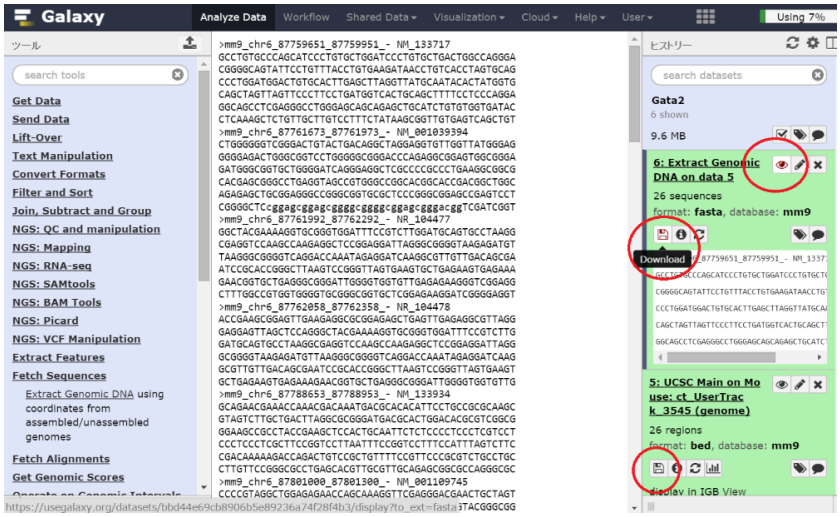
データ名をクリックして開き、「アイマーク」をクリックするとメインパネルにデータが表示される。このデータを利用して、ゲノム塩基配列を「Fetch Sequences」の「Extract Genomic DNA」で取得してみる。



下図のように設定する。



FASTA ファイルデータが得られる。ディスクマークをクリックするとデータがダウンロードできる。

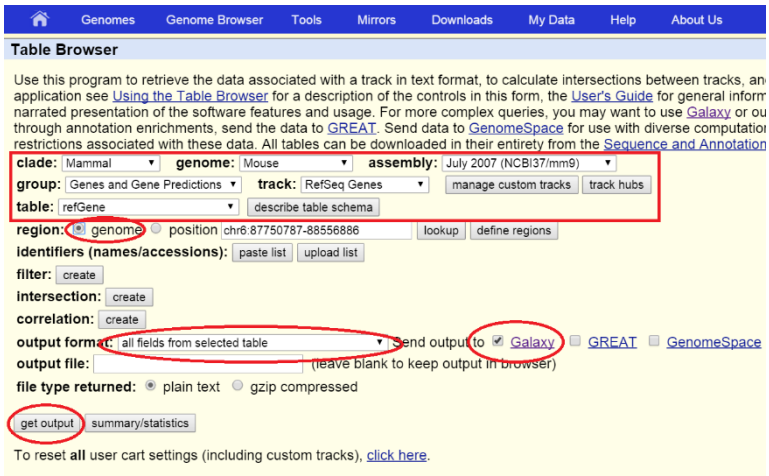


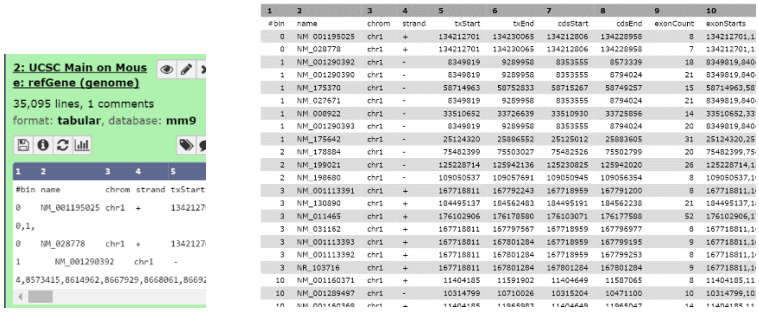
BED ファイルと FASTA ファイルがダウンロードされる。ここまでが練習。6.2 では行数が増すので、注意して作業する。

6.2　アレイ解析で変化していた遺伝子のプロモーター領域の取得

- RefSeq Genes レコードのダウンロード

  UCSC ブラウザの「Tools」＞「Table Browser」から mm9 の RefSeq Genes の全ゲノム分のレコードを Galaxy にダウンロードする。

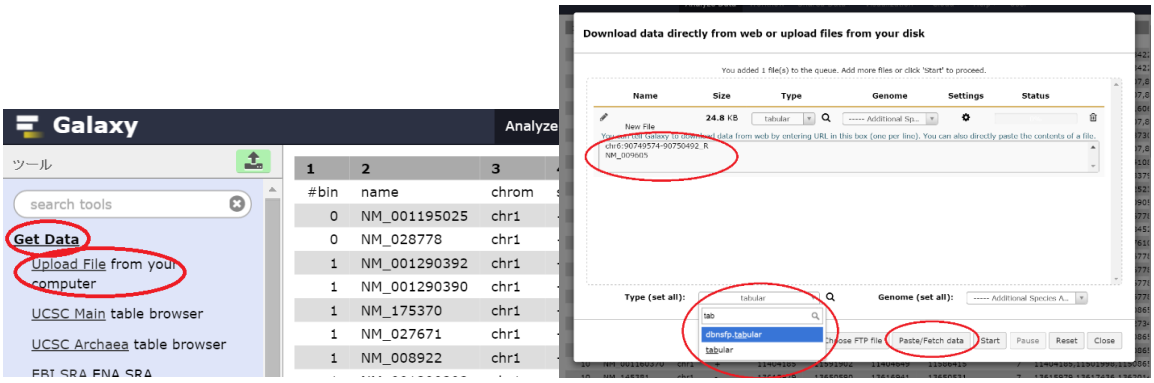




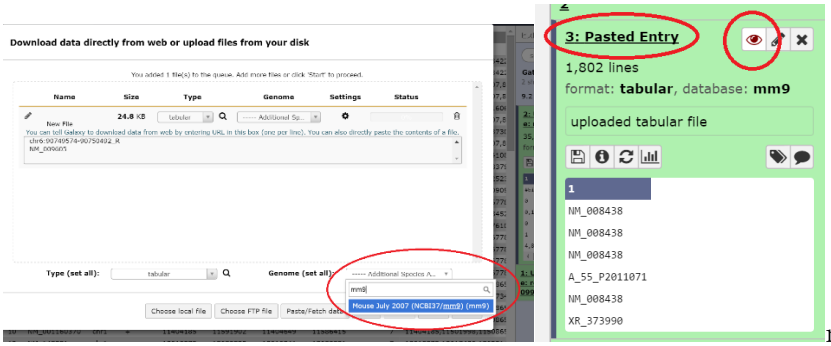
35682 行のファイルが Galaxy に入ったことを確認する。（RefSeq データベースは頻繁にアップデートされるので、行数を確認して保存しておき、同一のプロジェクトでは同じバージョンのデータベースを使う。キャプチャーは少し古い。）

- アレイデータをもとに RefSeq Genes の行を選ぶ

  遺伝子リストを Tabular file としてアップロードする。

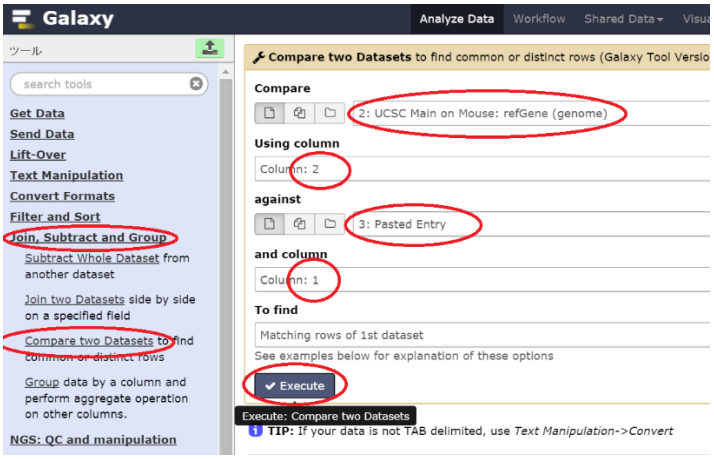


「Upload files」から、「Paste/Fetch data」でリストをコピー＆ペースト。「Tabular」を選択。

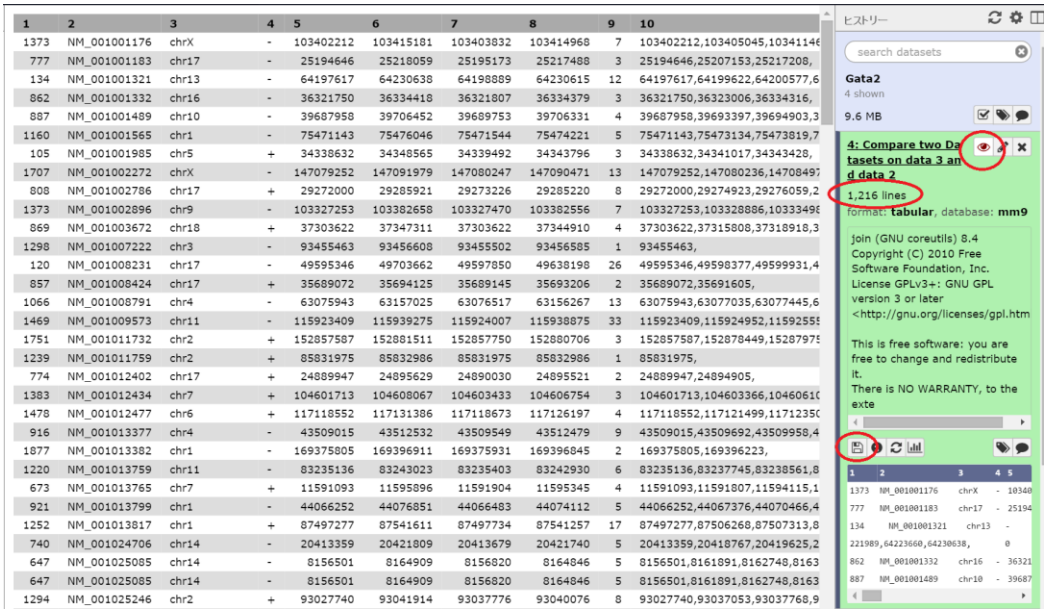


mm 9 を選択する。

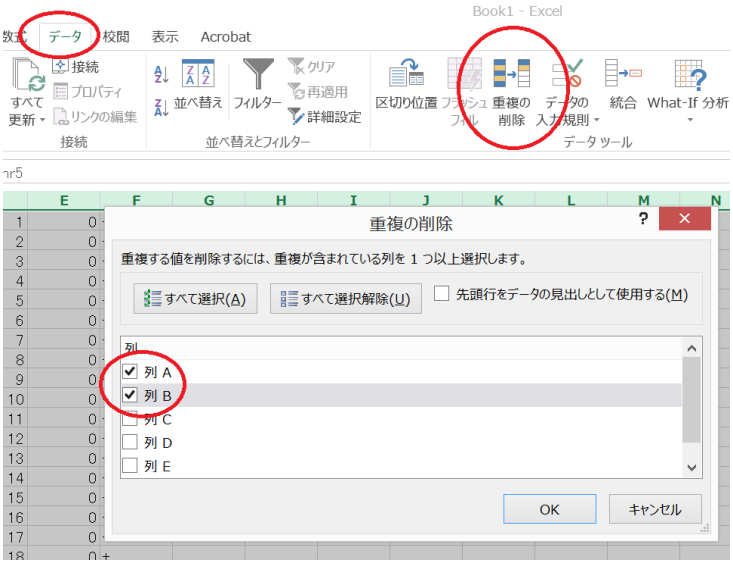
「Join, Subtract and Group」から、「Compare two datasets」、RefSeq Genes の Column 13（RefSeq ID の場合は Column 2）と遺伝子リスト（Column 1）を比較して、マッチする RefSeq Genes データの行を取得する。



結果が新しいデータとして現れる。行数を確認し、データをダウンロードする。このファイルはマイクロ RNA 標的の解析でも使うので、エクセルファイルとして保存する。



Gata2 遺伝子付近のデータで練習した方法で、プロモーター領域の BED ファイルを作成する。プロモータ領域は、転写開始点から - 200 ベース、＋100 ベースとしているが、いろいろな範囲を試しても良い。次に行うモチーフ解析では、重複する塩基配列を削除する必要がある。



「データ」＞「重複の削除」で同じスタート位置の行を削除する。この BED ファイルをテキスト形式で保存し、「Get Data」＞「Upload files」で Galaxy にアップロードする。

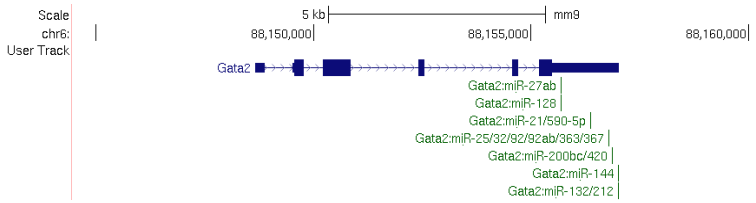
- プロモーター領域の FASTA ファイルを取得する

  Fetch Sequences で BED ファイルに指定されたプロモーター領域の塩基配列を FASTA ファイルで取得して、ChIPseq ピークのモチーフ解析と同様に DREME にアップロードする。

７．マイクロ RNA ターゲットのダウンロードとアレイデータとの比較

目標：RNAseq で変化していた遺伝子の 3 末端 UTR 領域にあるマイクロ RNA の標的配列を検索し、転写後調節に関わりそうなマイクロ RNA を見つける。ゲノムワイドと比較しての標的配列の出現頻度の解析や、同一のマイクロ RNA の制御を受けている遺伝子群のパスウェイ解析などへと展開する。

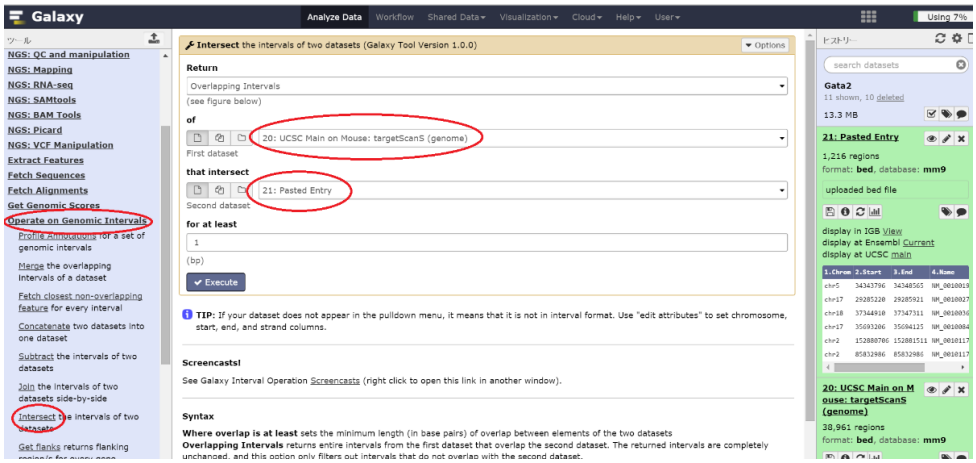
- miRNA target scan のデータベース
  mm9 のゲノムブラウザの「Expression and Regulation」の項目から「TS miRNA sites」を「Pack」にして miRNA ターゲットを表示してみる。下図は Gata2 遺伝子の例。



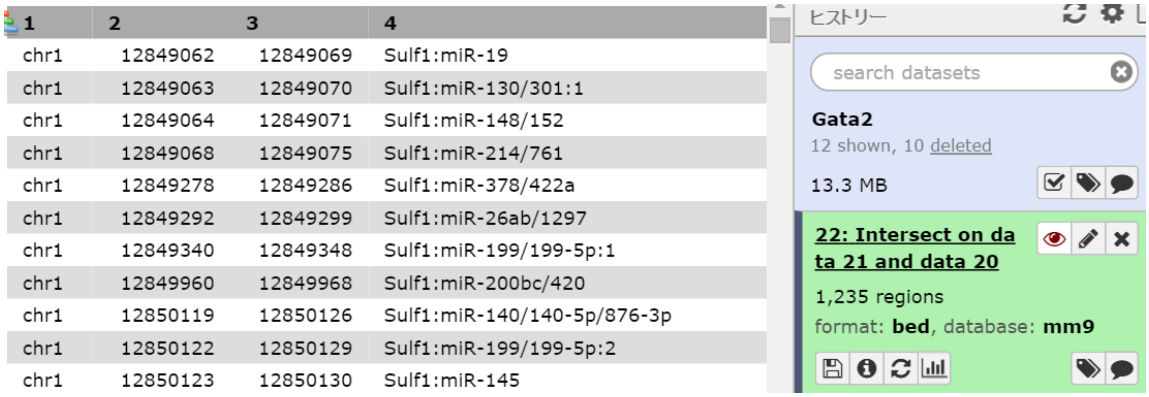
- テーブルブラウザで TS miRNA sites のデータを全ゲノム分、BED ファイルとして Galaxy にダウンロードする。
  Gata2 遺伝子とオーバーラップするものは、BED ファイルではこのように見える。

```
chr6    88155678        88155686        Gata2:miR-27ab  95      +
chr6    88155678        88155685        Gata2:miR-128   53      +
chr6    88156367        88156374        Gata2:miR-21/590-5p     35      +
chr6    88156775        88156783        Gata2:miR-25/32/92/92ab/363/367 89      +
chr6    88156861        88156868        Gata2:miR-200bc/420     46      +
chr6    88156992        88156999        Gata2:miR-144   84      +
chr6    88156994        88157001        Gata2:miR-132/212       82      +
```

  全ゲノムでは 38,961 行。
- アレイデータの RefSeq 遺伝子セットから、3 末端 UTR の BED ファイルを作り、Galaxy にアップロードする。
  「Strand」で場合分けして「cds End」と「txn End」の間を指定した BED ファイルをエクセルで作成する。重複の削除はこの場合は必要ない。
- アレイリストから選んだ 3 末端 UTR にオーバーラップする miRNA ターゲットを探す。
  Galaxy の「Operate on Genomic Intervals」から、「Intersect」を選び、マイクロ RNA ターゲットを最初のデータセット、3 末端 UTR の BED ファイルを 2 番目のデータセットに指定して、「Execute」をクリック。

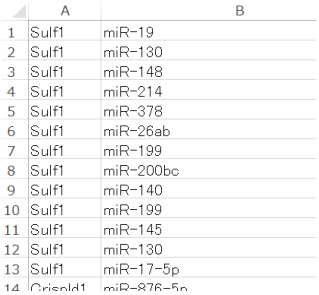


アウトプットは下図のようになる。



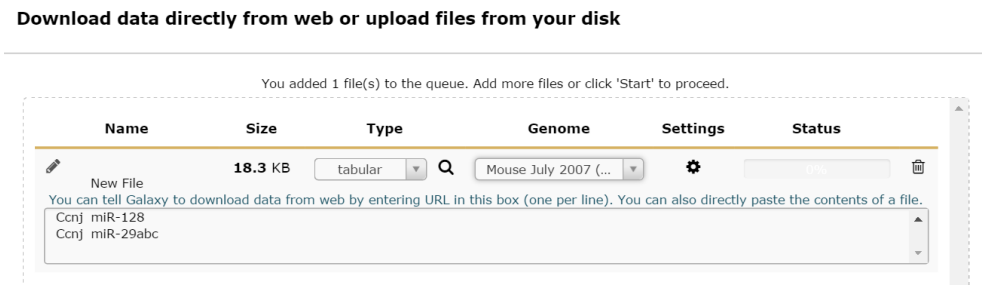
1	2	3	4
chr1	12849062	12849069	Sulf1:miR-19
chr1	12849063	12849070	Sulf1:miR-130/301:1
chr1	12849064	12849071	Sulf1:miR-148/152
chr1	12849068	12849075	Sulf1:miR-214/761
chr1	12849278	12849286	Sulf1:miR-378/422a
chr1	12849292	12849299	Sulf1:miR-26ab/1297
chr1	12849340	12849348	Sulf1:miR-199/199-5p:1
chr1	12849960	12849968	Sulf1:miR-200bc/420
chr1	12850119	12850126	Sulf1:miR-140/140-5p/876-3p
chr1	12850122	12850129	Sulf1:miR-199/199-5p:2
chr1	12850123	12850130	Sulf1:miR-145

テーブルをダウンロードし、エクセルでフォーマットを整える。「:」で区切って、次に「/」で区切り、ターゲット遺伝子名と miRNA の名前を別のカラムに分ける。

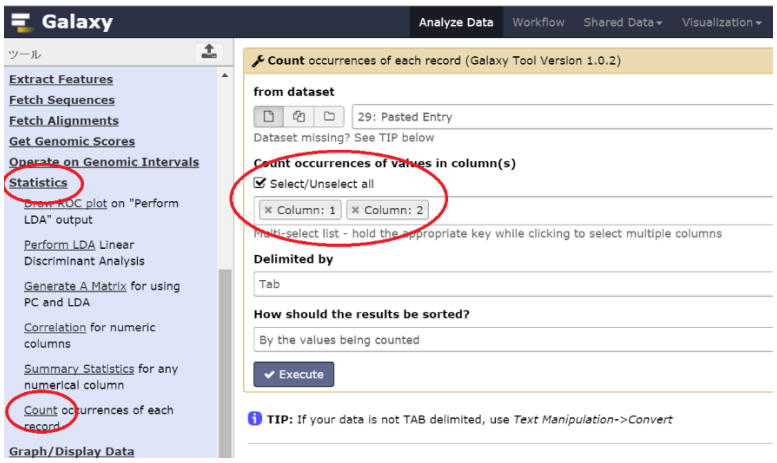


	A	B
1	Sulf1	miR-19
2	Sulf1	miR-130
3	Sulf1	miR-148
4	Sulf1	miR-214
5	Sulf1	miR-378
6	Sulf1	miR-26ab
7	Sulf1	miR-199
8	Sulf1	miR-200bc
9	Sulf1	miR-140
10	Sulf1	miR-199
11	Sulf1	miR-145
12	Sulf1	miR-130
13	Sulf1	miR-17-5p
14	Crispld1	miR-876-5p

2 つのカラムを Galaxy にアップロードする。



Galaxy の左パネルのメニューから、「Statistics」の「Count」を選択、「Select」のチェックボックスをクリックして 2 つのカラムが認識されていることを確認する。Column 1、2、両方の 3 通りでカウントしてみる。



下図は両方でカウントした結果。同じ遺伝子に複数の miRNA 標的があることは多くなさそう。

1	2	3
1	1300007F04Rik	miR-299
1	2610027C15Rik	miR-29abc
1	2700060E02Rik	miR-18ab
1	3110003A22Rik	miR-106
1	3110003A22Rik	miR-130

下図は遺伝子名でカウントした結果。ターゲットの多い遺伝子と少ない遺伝子がある。3 末端 UTR の長さで補正しても良いかもしれない。

1	2
1	1300007F04Rik
1	2610027C15Rik
1	2700060E02Rik
5	3110003A22Rik
1	4930417M19Rik
1	8430419L09Rik
1	9030425E11Rik
1	Abca2
12	Adam23

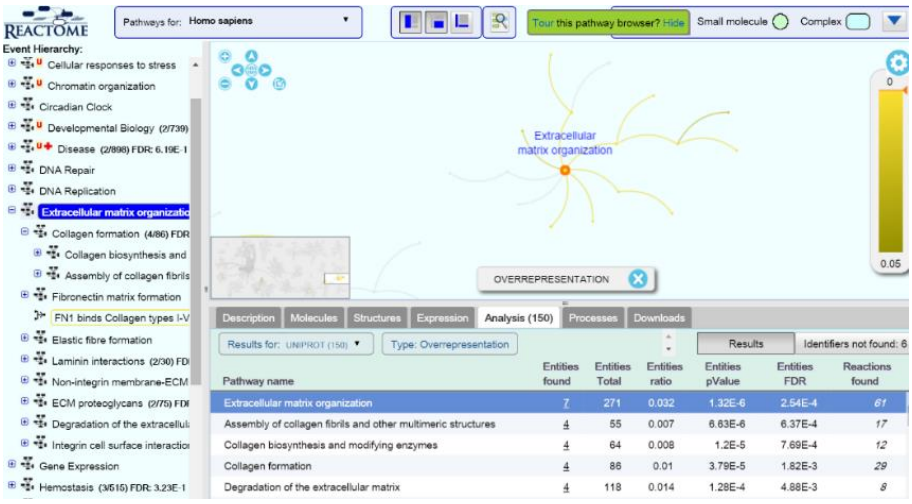
下図は miRNA でカウントした結果。アレイで変化していた複数の遺伝子の制御に関わる miRNA もある。

1	2
23	let-7
8	miR-1
3	miR-10
13	miR-101
9	miR-103
13	miR-106
1	miR-122
27	miR-124

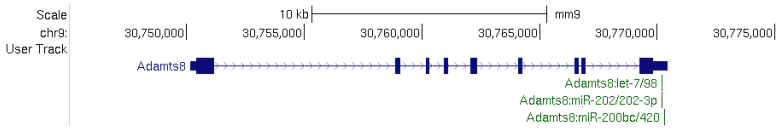
Galaxy メニューの「Filter and Sort」＞「Sort」（又はエクセル）を使ってカラム 1 でソートすれば、ターゲットの多い順に並べ替えられる。ゲノムワイドに比べて特に出現頻度が高い miRNA を探すには、Normalize するためのデータが必要になるので、全ゲノムやランダムに選んだ遺伝子セットを用いる方法がある。

- miRNA の標的遺伝子のパスウェイ解析

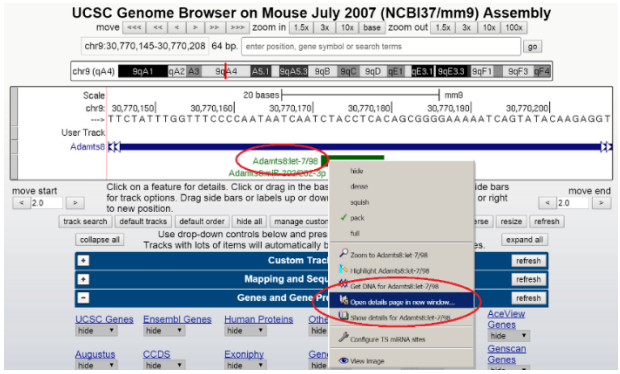
  遺伝子の数が少なく、うまくいかないツールもあるが、REACTOME などは使える。例えば、let-7 の標的遺伝子で、アレイデータで変化していた 19 個の遺伝子には（23 か所のターゲットは 19 個の遺伝子上にあった）、細胞外マトリックスに関連する遺伝子が 7 個あった。



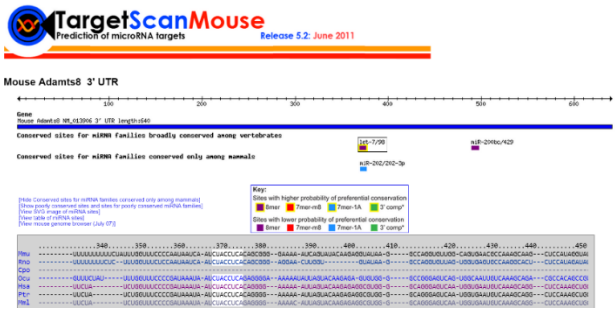
ヒットした遺伝子のいくつかについては、実際に let7 のターゲットがあるか、ブラウザ上で確認する。



さらにクローズアップして塩基配列レベルで確認、「Detail」を新しいタブで開く、



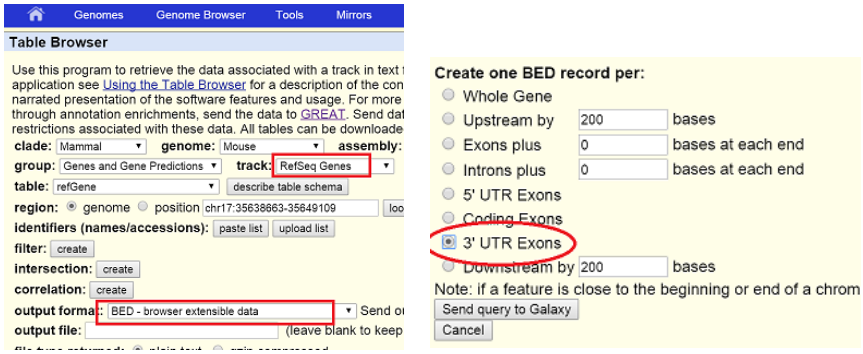
Target scan のページへのリンクを開く。



また、miRNA に関する文献検索でも発生分化や環境応答に関連した miRNA の文献が見つかるので、いろいろ生物学的な意味を考えてみる。

- もっと「簡単な」やり方について

Table browser にはフィルタリングや範囲の指定方法がいろいろある。



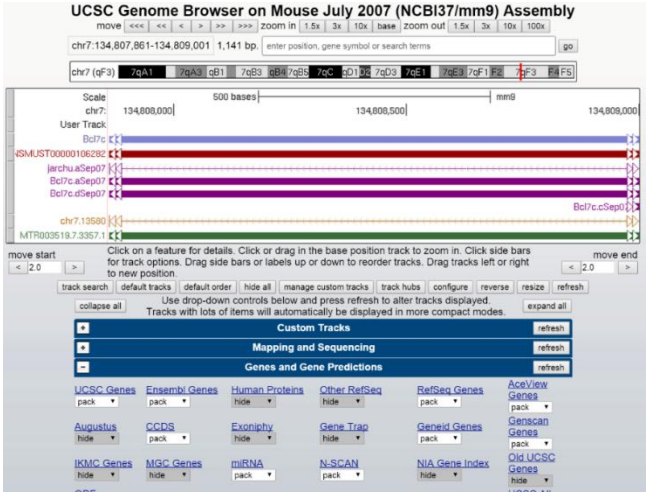
1	2	3	4	5	6
chr1	134228958	134230065	NM_001195025_utr3_7_0_chr1_134228959_f	0	+
chr1	134228958	134230065	NM_028778_utr3_6_0_chr1_134228959_f	0	+
chr1	8349819	8353555	NM_001290392_utr3_0_0_chr1_8349820_r	0	-
chr1	8349819	8353555	NM_001290390_utr3_0_0_chr1_8349820_r	0	-
chr1	58714964	58715267	NM_175370_utr3_0_0_chr1_58714965_r	0	-
chr1	8349819	8353555	NM_027671_utr3_0_0_chr1_8349820_r	0	-
chr1	33510653	33510830	NM_008922_utr3_0_0_chr1_33510653_r	0	-

BED ファイルをリクエストすると、3 末端 UTR だけを取ってくることもできる。が、名前のカラムで余計なものを付けてくれるので、遺伝子リストとのマッチングに結局またひと手間かかる。

８．ノンコーディング RNA 遺伝子のプロモーター解析

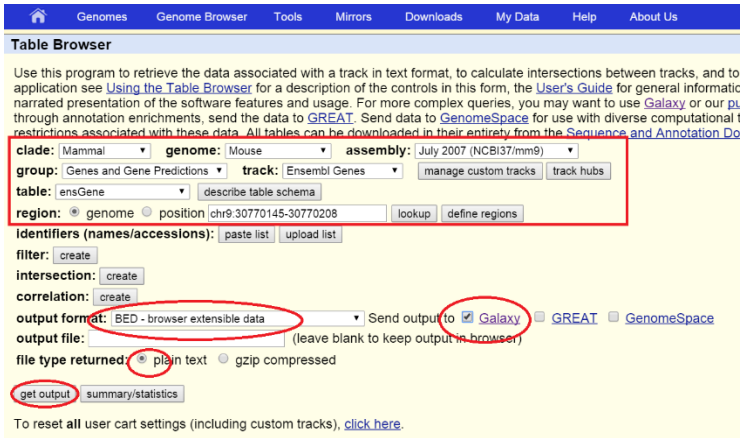
目標：RefSeq や UCSC Gene のデータベースに収録されていない遺伝子をターゲットとしたプローブについてもトランスクリプトの位置からプロモーターの位置を割り出す。

- ノンコーディング RNA 遺伝子の情報から、データベース上のトランスクリプトを見つける。
  例えば「lincRNA:chr7:134807861-134809001 reverse strand」の位置を mm9 で見ると、



  RefSeq 遺伝子セットには含まれていないが、UCSC Gene や UNSEMBLE データベースにはトランスクリプトがのっていることがわかる。

- Ensemble データベースの遺伝子リストを取得する
  UCSC ブラウザのテーブルブラウザから下図のように遺伝子リストを Galaxy に BED ファイルでダウンロードする。95833 行のデータが Galaxy にアップロードされる。



- アレイデータのノンコーディング RNA 遺伝子の BED ファイルの作成

エクセルを使って RefSeqID の付いていない遺伝子の位置を BED 形式に変換する。「Accessions」が空欄のものを、ソートして Spreadsheet の上位に持ってくると変換しやすい。

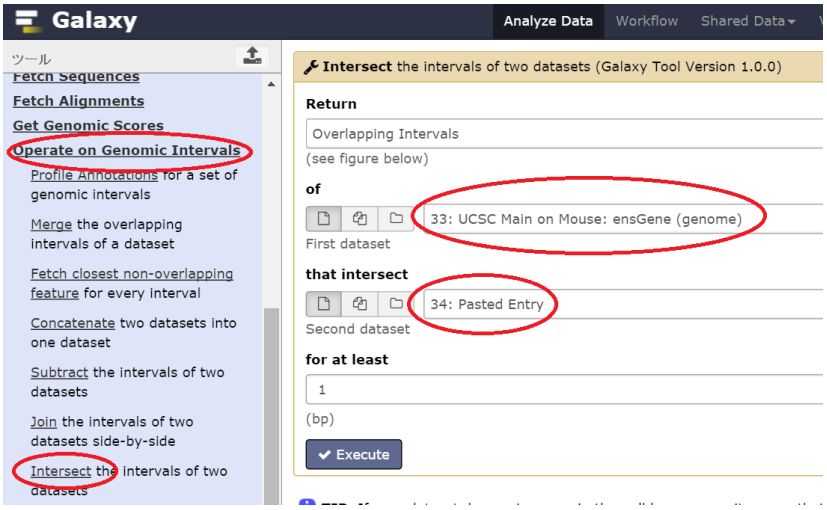
「mm|chr13:52071833-52071892」をエクセルメニューの「データ」＞「区切り位置」を使って mm chr13 52071833 52071892 の 4 つのカラムに変換してから、BED 形式に整える。空欄もあるので、カラム A でソートしておく。「Name」のカラムには、適当な通し番号を付ける。

	A	B	C	D	E	F
1	chr1	122980930	122980989	linc1	0	+
2	chr1	92152485	92152426	linc2	0	+
3	chr1	138528983	138528924	linc3	0	+
4	chr1	160888317	160888258	linc4	0	+
5	chr1	122660083	122660142	linc5	0	+
6	chr1	195332744	195332685	linc6	0	+
7	chr1	63426115	63426056	linc7	0	+
8	chr1	40113731	40113672	linc8	0	+
9	chr1	182183923	182183864	linc9	0	+
10	chr1	93815364	93815423	linc10	0	+
11	chr1	135956678	135956619	linc11	0	+

Galaxy にアップロードする。

- ノンコーディング RNA 遺伝子にマッチするレコードを Ensemble データベースから探す

Galaxy には Ensemble 遺伝子リスト（95833 行のデータ）とアレイ解析からのノンコーディング RNA 遺伝子の両方が BED 形式でアップロードされているので、Galaxy の「Operate on Genomic Intervals」の「Intersect」でマッチするものを探す。



3 分の一程度にマッチする Ensemble 遺伝子があるはず。より登録遺伝子が多い EST データベースなどを用いるとより多くのノンコーディング RNA 遺伝子がマップされるが、転写開始点などの情報が不明なものも多いので、Ensemble データベースにのっていないものは解析から除外する。

1	2	3	4	5	6	7	8	9
chr1	84989264	85242972	ENSMUST00000159114	0	-	85242972	85242972	0
chr1	85024528	85038533	ENSMUST00000160831	0	-	85038533	85038533	0
chr2	48805029	49115659	ENSMUST00000112745	0	+	49101613	49113870	0
chr2	48805029	49101257	ENSMUST00000127085	0	+	49101257	49101257	0
chr2	48805037	48974532	ENSMUST00000154653	0	+	48974532	48974532	0
chr2	48805662	49101077	ENSMUST00000147055	0	+	49101077	49101077	0

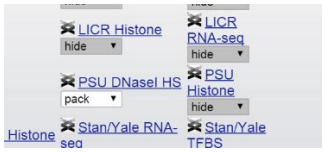
アウトプット。転写開始点がわかるので、プロモーター解析などに RefSeq 遺伝子セットと合わせて用いるなどの用途がある。

９．遺伝子周辺のエンハンサー等の制御領域の解析

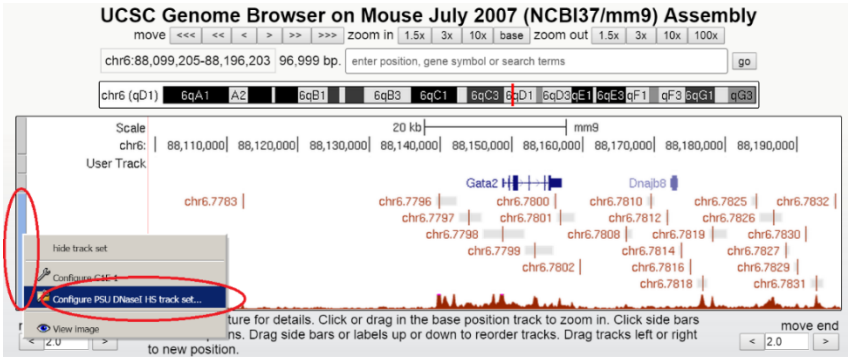
目標：DNase-hypersensitive site など、エピゲノムデータベースとの比較による解析により、アレイで見つかった遺伝子群の付近にあるエンハンサー等の遺伝子制御領域について塩基配列を得て、モチーフ検索に使用する。これにより、プロモーター以外の制御領域の転写因子モチーフ解析を行う。

- DNAHS データ

  DNase-HS は転写因子の結合サイトや遺伝子プロモーターの可能性が高い。USCS ブラウザのプルダウンメニュー「Regulation and Expression」の「PSU DNaseI HS」を「Pack」に設定、「Refresh」をクリック。



  ブラウザに DNaseI HS トラックが現れる。左端のバーを右クリックすると、「Configure…」のメニューがあり、細胞株やデータの種類を変更できる。



  全てのカラムの情報。ピーク位置はブラウザウィンドウでは縦線で表されており、表では Start からの塩基数で表される。「Table Browser」からデータを BED 形式でそのままダウンロードすると、Start-end の位置になるので、ピーク周辺を使いたい場合は自分で加工して BED 形式を作る必要がある。

#bin	chrom	chromStart	chromEnd	name	score	strand	signalValue	pValue	qValue	peak
1257	chr6	88112087	88112384	chr6.7783		100	.	0.0427	-1	-1	154
1257	chr6	88139360	88142385	chr6.7796		100	.	0.2819	-1	-1	377
1257	chr6	88142534	88145788	chr6.7797		100	.	0.0621	-1	-1	1398
1257	chr6	88145860	88151755	chr6.7798		100	.	0.3576	-1	-1	2769
1257	chr6	88151863	88155859	chr6.7799		100	.	0.1232	-1	-1	1356
1257	chr6	88155899	88156247	chr6.7800		100	.	0.0374	-1	-1	170

- アレイデータからの遺伝子リストの加工

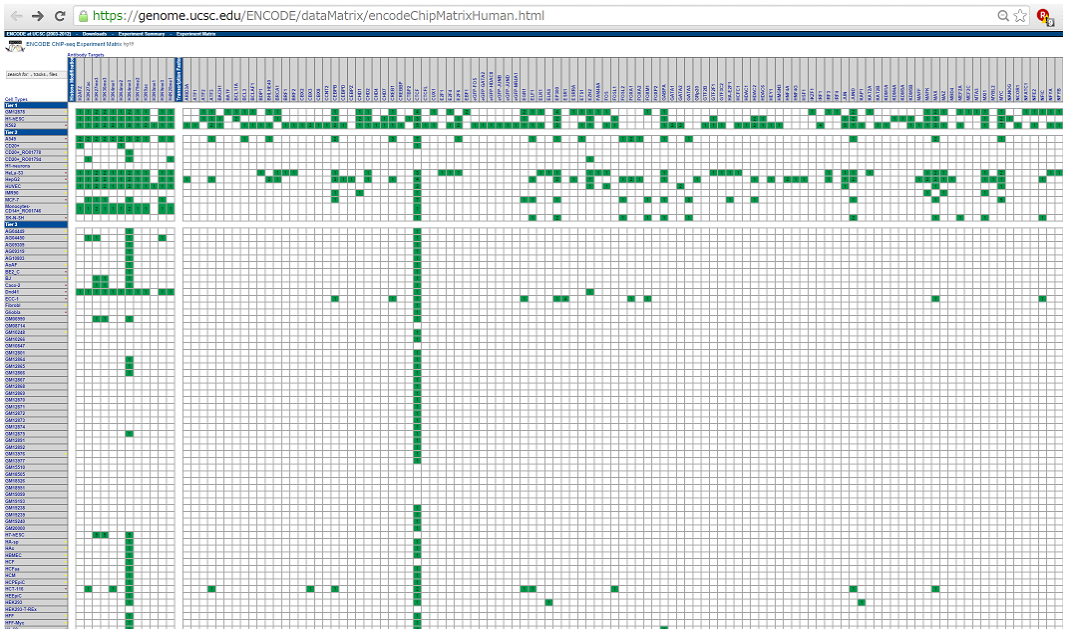
  「2x_up.bed」をエクセルで開き、「2x_up_20kb.bed」など別名で保存する。遺伝子の「Start」と「End」の位置をそれぞれ上流と下流に 20kb ずつ伸ばす。このファイルを Galaxy にアップロードし、「Operate on Genomic Intervals」の「Intersect」ツールを用いて、アレイ遺伝子の付近にある DNaseHS のみを選び出す。

- アレイで変化していた遺伝子付近の DNaseHS を 400 ベースのサイズに統一してから「Fetch Sequences」で塩基配列の FASTA ファイルを取得して、「DREME」でモチーフ解析を行う。

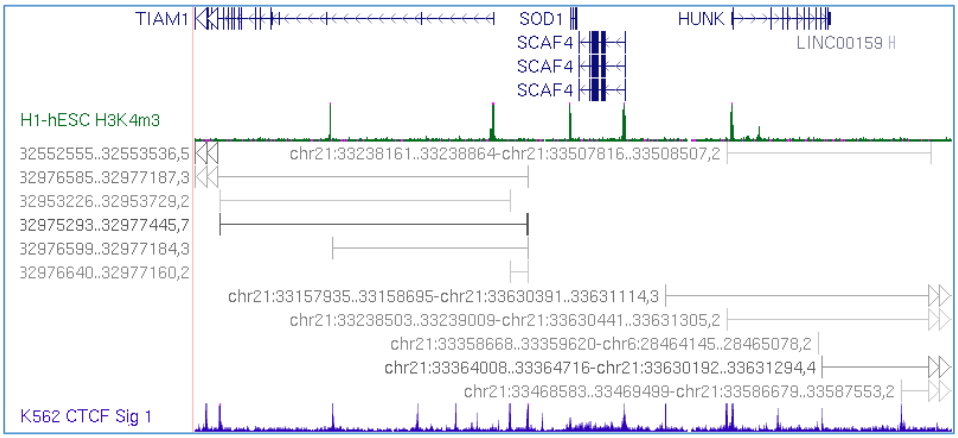
１０．その他のデータベース、解析ツール

10.1　ENCODE (https://genome.ucsc.edu/ENCODE/)、IHEC (http://ihec-epigenomes.net/)

既にヒストン修飾の ChIPseq データなどを用いたが、特にヒトの組織、細胞のエピゲノムデータが今後増えると期待される。RNAseq、DNaseHS、ChIA-PET、5C などのデータもある。マトリックスの各カラムをクリックすると、UCSC ブラウザへリンクがあるので、Table browser からダウンロードする。



例えば、CTCF の ChIA-PET のデータは…



2 つのピークを結ぶスパンとして可視化されている。このデータは、

1	2	3	4	5
#bin	chrom	chromStart	chromEnd	name
0	chr7	64299266	153110240	chr7:64299266..64299795-chr7:153109659..153110240,2
22	chr7	115778002	116571133	chr7:115778002..115778876-chr7:116570600..116571133,2
183	chr7	115812164	116229742	chr7:115812164..115814136-chr7:116228770..116229742,9
22	chr7	115812797	116443182	chr7:115812797..115813710-chr7:116442637..116443182,2

スパン全体の BED データ、Name のところにピークが 2 つテキスト行として入っている。

H3K4me3 ピークとオーバーラップしているものと、していないものを分けて、モチーフ解析を行う、付近の RNAseq の Prdm14KO に対する応答はどうか？また、CTCF ループの中にある、または外にある RefSeq 遺伝子群の発現量を自分の RNAseq データから比較して Boxplot を描くなど、シンプルなツールを駆使して様々な解析ができそう。

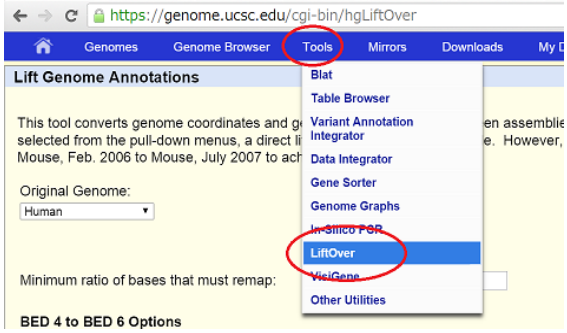
10.2　Galaxy の FTP サイト

FTP サイト usegalaxy.org にユーザー名とパスワードでログインしてファイルをアップロードすると、多数のファイル、大きめのファイルの転送に便利。詳しくは Galaxy の「Get Data」＞「Upload File」を参照。

10.3　Lift-over tool（Galaxy）

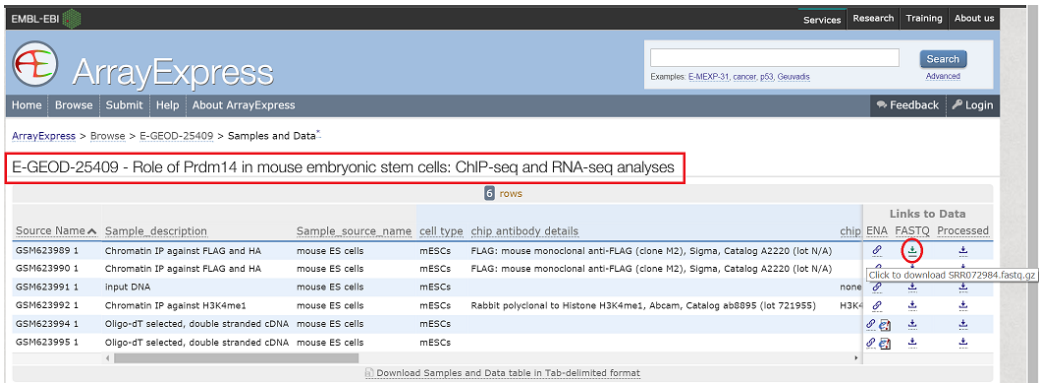
https://genome.ucsc.edu/cgi-bin/hgLiftOver

mm9、mm10、hg19 などは、ゲノムのアセンブリのバージョンを表している。マッピングのアセンブリが異なると、ゲノム上の位置も異なる。自分が持っているデータが mm10 で、mm9 のデータと比較したい時、mm10 から mm9 の位置に「Lift-over」する。

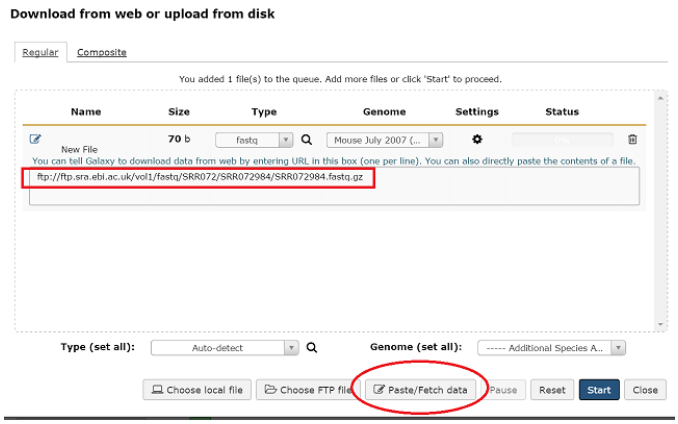


10.4　EBI データベース（https://www.ebi.ac.uk/）

Google で論文に書いてあるデータの「Accession number」を検索すると、GEO（Gene Expression Omniubs）と EBI のデータベースのどちらからでも同じデータがダウンロードできるリンクが見つかる。GEO サイトは「SRA ツール」でデータを解凍するステップがあり、簡単だが初級者にはその一歩が大きな障壁に。EBI サイトは素直に Zip 圧縮データをダウンロードできる。



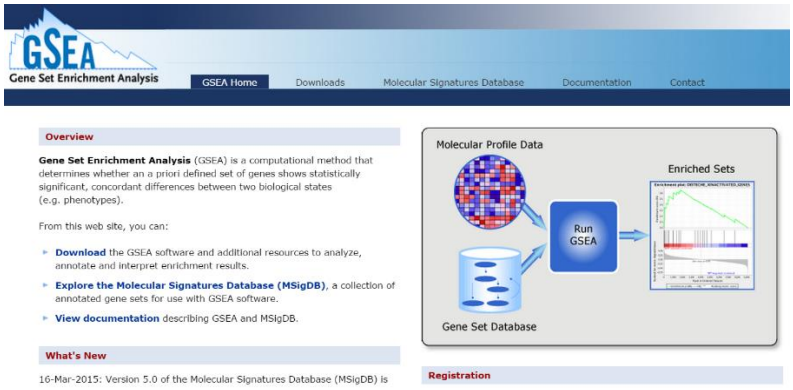
右クリックで URL をコピーしたら、Galaxy の「Upload files」で「Fetch Data」、



そのままアドレスを貼り付ける。

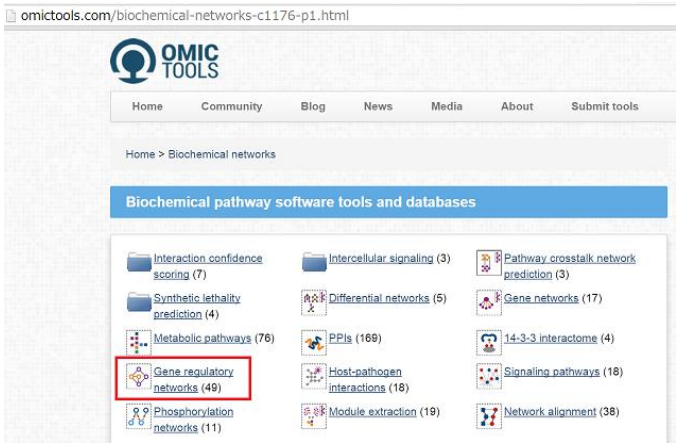
10.5　Gene-set enrichment analysis (GSEA)

パスウェイ解析の結果を、閾値を切らずにより視覚的に表現できる。ソフトウェアのダウンロードが必要。http://www.broadinstitute.org/gsea/index.jsp



10.6　Gene Regulatory Network tools

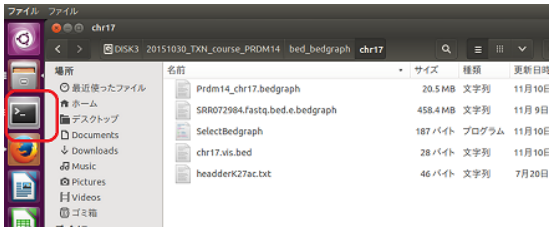
論文でよく見る毛玉のようなものから、簡単な骨組みのようなものまで、遺伝子制御ネットワークを試したい時はこの辺から探す。Linux 上で動くものが多いが、ウェブツール、Windows、Mac にインストールできるものもある。論文発表されているものは公開されている。

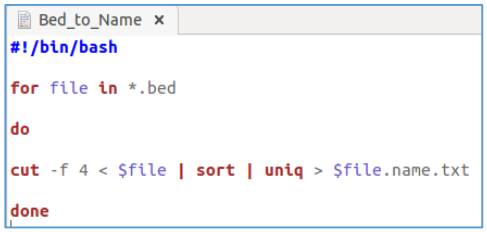


終わりに

- もっとたくさんのファイルをバッチで解析したい
- Galaxy ではファイルのアップロードや解析に時間がかかりすぎる
- ウェブインターフェースのない新しいツールを使ってみたい

などと思った人は、ぜひ Mac や Linux のコマンドにトライしてみてください！





Linux (Ubuntu)の画面。普段のファイル操作などはマウスで。必要な時だけターミナルを開けばよい。右の例では、このスクリプトを流すフォルダの中のすべての「.bed」ファイルについて、カラム 4 を選んで、ソート、重複を除いて、「.name.txt」をファイル名に付け足した新しいファイルを作ってくれる（テキストの 2.1 の作業）。手作業によるミスも防ぐことができる。

便利な基本操作など

A1.　FASTA ファイルと BED ファイル形式

- FASTA ファイル（塩基配列とその名前などを表す）

  塩基配列のファイル。最初の行が「＞」で始まるテキスト情報、2 行目から次の「＞」までが塩基配列。

```
>mm9_chr6_87759651_87759951_- NM_133717
GCCTGTGCCCAGCATCCCTGTGCTGGATCCCTGTGCTGACTGGCCAGGGA
CGGGGCAGTATTCCTGTTTACCTGTGAAGATAACCTGTCACCTAGTGCAG
CCCTGGATGGACTGTGCACTTGAGCTTAGGTTATGCAATACACTATGGTG
CAGCTAGTTAGTTCCCTTCCTGATGGTCACTGCAGCTTTTCCTCCCAGGA
GGCAGCCTCGAGGGCCTGGGAGCAGCAGAGCTGCATCTGTGTGGTGATAC
CTCAAAGCTCTGTTGCTTGTCCTTTCTATAAGCGGTTGTGAGTCAGCTGT
>mm9_chr6_87761673_87761973_- NM_001039394
CTGGGGGGTCGGGACTGTACTGACAGGCTAGGAGGTGTTGGTTATGGGAG
GGGGAGACTGGGCGGTCCTGGGGGCGGGACCCAGAGGCGGAGTGGCGGGA
GATGGGCGGTGCTGGGGATCAGGGAGGCTCGCCCCGCCCTGAAGGCGGCG
CACGAGCGGGCCTGAGGTAGCCGTGGGCCGGCACGGCACCGACGGCTGGC
AGAGAGCTGCGGAGGGCCGGGCGGTGCGCTCCCGGGCGGAGCCGAGTCCT
CGGGGCTCcggagcggagcggggcggggcggagcgggacggTCGATCGGT
>mm9_chr6_87761992_87762292_- NR_104477
GGCTACGAAAAGGTGCGGGTGGATTTCCGTCTTGGATGCAGTGCCTAAGG
CGAGGTCCAAGCCAAGAGGCTCCGGAGGATTAGGGCGGGGTAAGAGATGT
TAAGGGCGGGGTCAGGACCAAATAGAGGATCAAGGCGTTGTTGACAGCGA
```

- BED ファイル（ゲノム上の位置情報を表す）

  ゲノム上の位置を、「chr」、「start」、「end」で一か所につき一行で表す。「Name」「Score」「Strand」を加えた 6 カラムが標準。

1	2	3	4	5	6
chr6	87759651	87759951	NM_133717	0	-
chr6	87761673	87761973	NM_001039394	0	-
chr6	87761992	87762292	NR_104477	0	-
chr6	87762058	87762358	NR_104478	0	-
chr6	87788653	87788953	NM_133934	0	-

1 行につき 1 カ所のゲノム領域を表すが、「#」で始まる行は次の改行までコメント行として扱われる。

A2.　エクセルのキーボードショートカットと便利な機能

- 連続している最後の行、カラムに移動する

  Control+矢印キー

- カラムをまとめて選択する

  Control+Shift+矢印キー

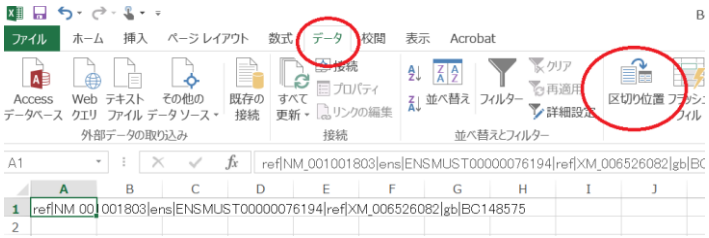
- テキストからカラムへの変換

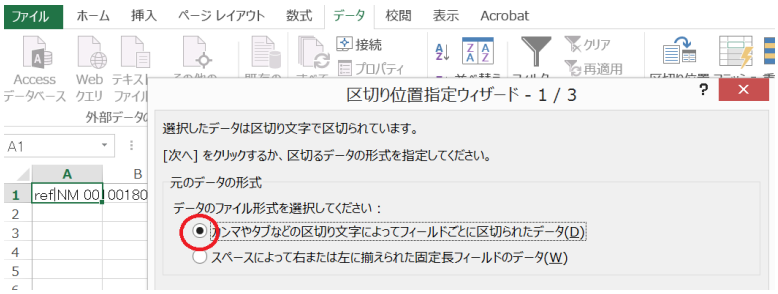
  例えばアレイの納品データの GeneID カラムや、

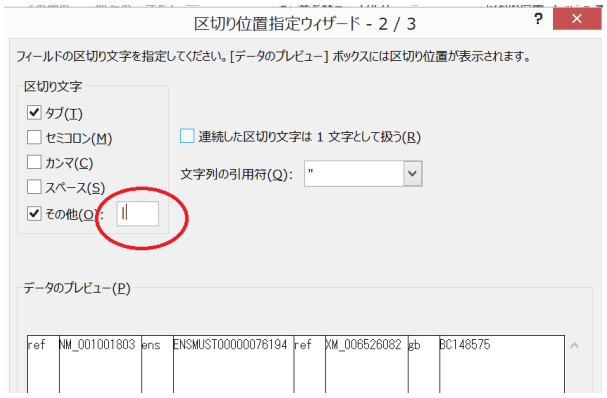
  ref|NM_001001803|ens|ENSMUST00000076194|ref|XM_006526082|gb|BC148575

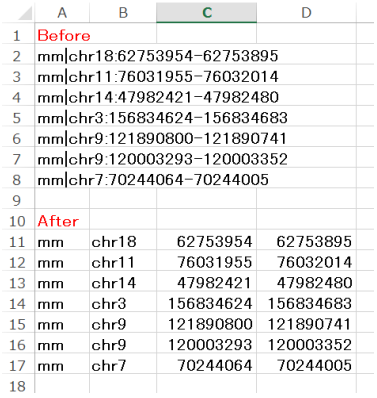
  プローブ位置のテキストをカラムに分割したい時

  mm|chr7:56862455-56862514







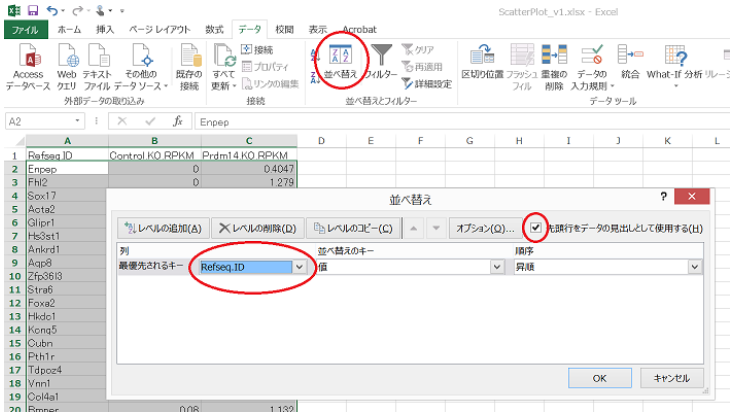


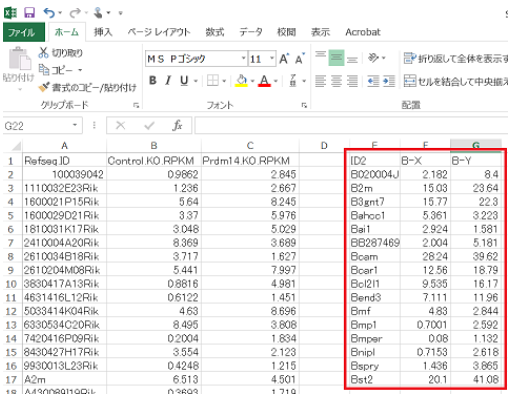
	A	B	C	D
1	Before			
2	mm|chr18:62753954-62753895			
3	mm|chr11:76031955-76032014			
4	mm|chr14:47982421-47982480			
5	mm|chr3:156834624-156834683			
6	mm|chr9:121890800-121890741			
7	mm|chr9:120003293-120003352			
8	mm|chr7:70244064-70244005			
9				
10	After			
11	mm	chr18	62753954	62753895
12	mm	chr11	76031955	76032014
13	mm	chr14	47982421	47982480
14	mm	chr3	156834624	156834683
15	mm	chr9	121890800	121890741
16	mm	chr9	120003293	120003352
17	mm	chr7	70244064	70244005
18				

この場合はタテ線（パイプ）で区切りたいので、「その他」をチェックして空欄に「｜」を入れる。プレビューで予想通りの区切りになっていることを確認。プローブ位置のカラムでは、「：」や「-」で一歩ずつ区切っていけば、プローブ位置を Chr、Start、End に分割してカラムを作ることができる。プローブリストから BED ファイルを作りたい時などに便利。

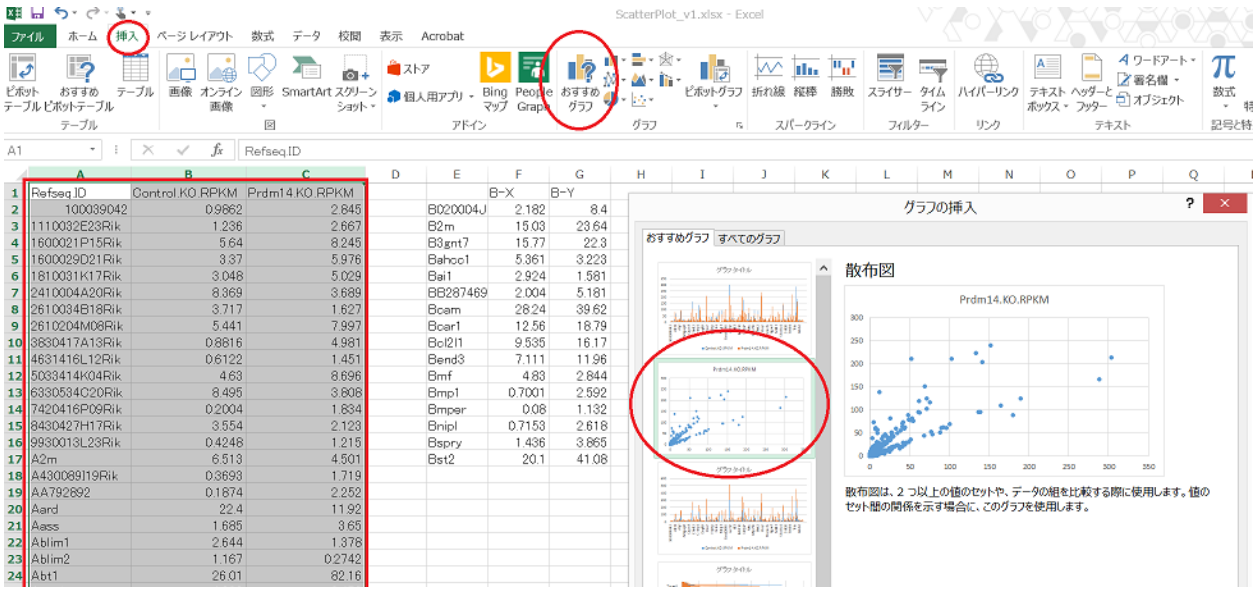
A3. 講習で使用する Prdm14 論文の Figure 2a のようなグラフをエクセルで作る

Excel 2013 で新しいファイルを作成し、Table S2 から遺伝子名、Control、Prdm14KO の RNAseq 定量結果をコピーする。Up が 480 遺伝子、Down が 173 遺伝子、ラベル行 1 、の合計 653 行、 3 カラムの表ができる。（Table S2 のファイルは Excel の形式が古いので、互換モードで作業するとラベルを付けるところがうまくいかない。）

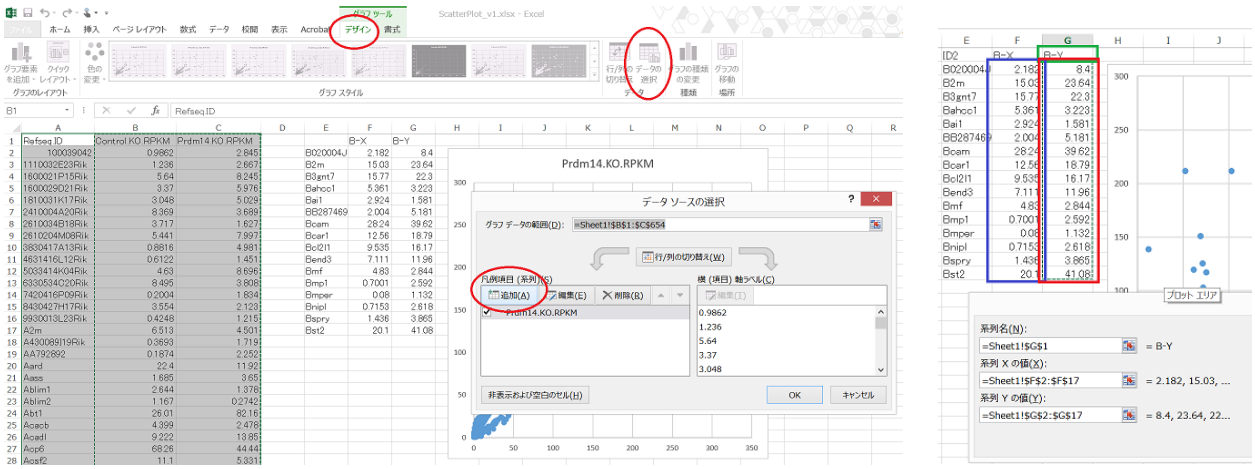




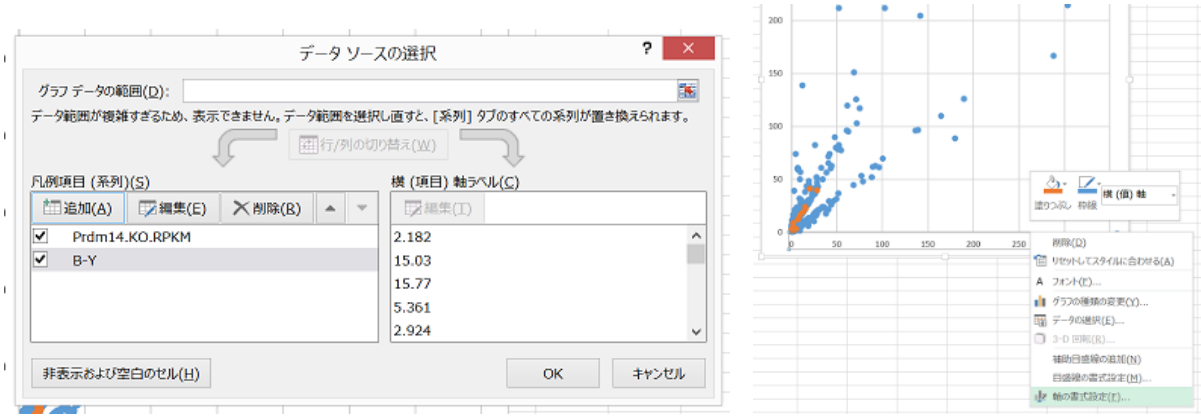
例えば、B で始まる遺伝子「B 遺伝子群」が目立つように、色を変えてラベルを付けたい場合、カラム A でソートした後、右上図の赤線部分ように 16 個の B 遺伝子群の行を別の表に分ける。一行目は ID2、B-X、B-Y と名前を付ける。実際には、パスウェイ解析で出てきた遺伝子などを選べばよい。



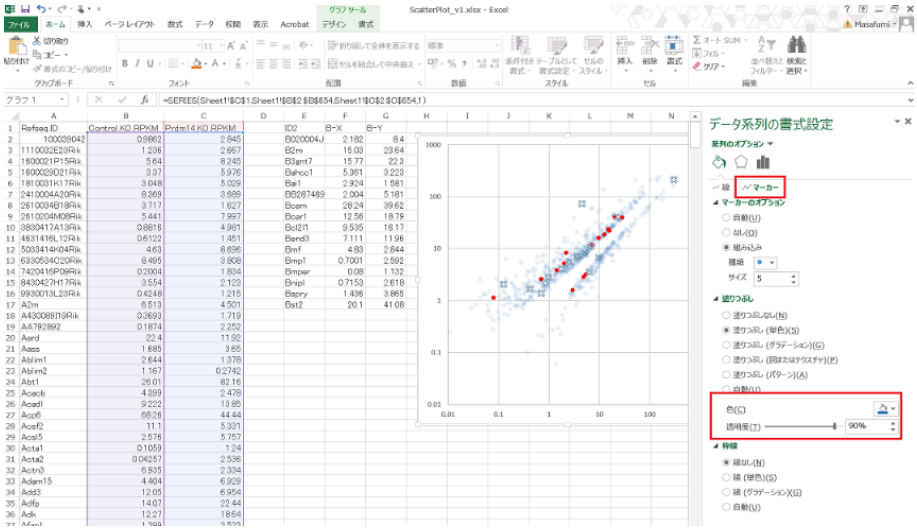
まずは大きいリストのほうで Scatter plot を作成する。



1 つ目のプロットを選択した状態で、「グラフツール」＞「デザイン」＞「データの選択」＞左側パネルの「追加」と進み、ラベルを B-Y のセル、X と Y の値をそれぞれ選択する。

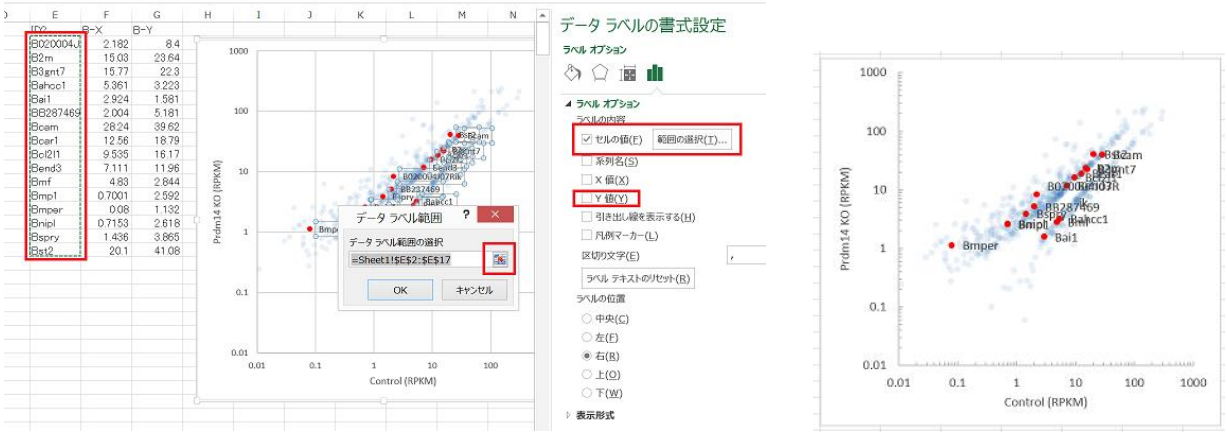


ここまでで、2 つ目のデータ系列（B 遺伝子群）が入り、プロットに新しい点が加わる。軸を対数にする。



プロットの点のフォーマットを変える。濃い目の塗りつぶし、枠線無し、透明度を上げると点の濃淡が分かりやすい。

B 遺伝子群の点を右クリックしてラベルを入れる。このままでは右上図のように数字が表示されるので、「グラフツール」＞「書式」から、左上のプルダウンメニュー、「系列“B-Y”データラベル」を選んで「選択対象の書式設定」をクリックする。



「セルの値」にチェックを入れて B 遺伝子群の名前を選択する。Y 値のチェックを外して、プロットのフォーマットをさらに整えれば出来上がり（右上図）。名前を一つずつダブルクリックすると、位置を移動できるので、あとは手作業で微調整すると重なりが無くなり、見やすくなる。線を引きたい場合、それに乗るような透明なプロットを重ねて近似線だけを入れる。