はじめに

- ChIP-seq解析関連のworkflow、解説はこちら
  - https://github.com/suimye/NGS_handson2015
- Black list等の置き場
  - http://tinyurl.com/gu4ov48

エピゲノム解析・中級編

Epigenomic analysis for advanced researchers

p227-p259

を同業者として眺めてみる

DRY本 + αの情報

Masaki Suimye Morioka

@suimye

Contents

今日は的をしぼって

- ChIP-seqのQCとデータクレンジング
- データの統合
  - エピゲノムデータをマイクロアレイデータのように

はじめに

基本的に私の解析手法の紹介は、
後出しじゃんけんなので、間違って私の方法が良いと思っても著者をdisるのはやめるように

どんな場合でもパイオニアというのは尊敬に値するものです。

QCする時の選択

一般的なQC

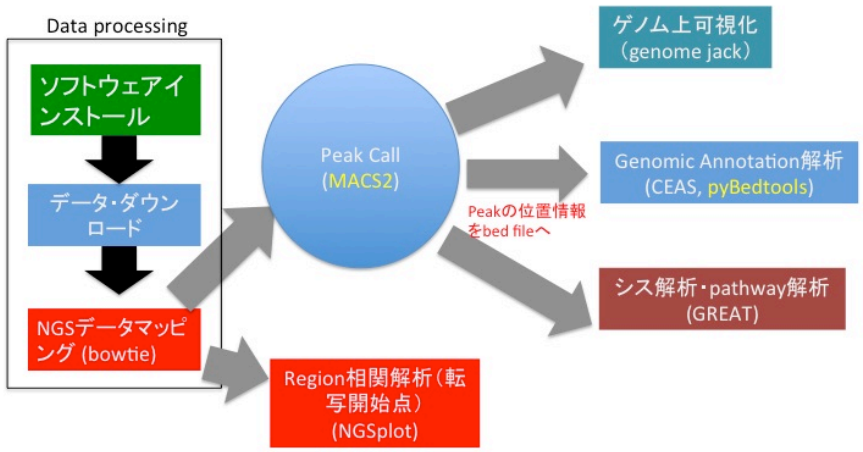
- Bioanalyzer: 各ステップでのライブラリ質の調査
  - 免疫沈降など濃縮後のDNA library
  - RNA 抽出後のクオリティ check
  - Sample prep後のクオリティ
- Qbit: Sample prep以前のDNAライブラリの定量
- KAPA Q Kit: シーケンシングのstarting material参考濃度
- Phred Score QC and Mapping QC

以上のQCでクリアできる項目

1. 特定サイズのDNAサンプルがあるかないか
2. Adopter 配列のあるDNA
3. DNA以外のcontaminationの評価
4. 適度なloading DNA concentration
5. 目的生物種のDNA contentの評価

DRY本のエピゲノム解析workflow

をまとめて、眺めてみよう

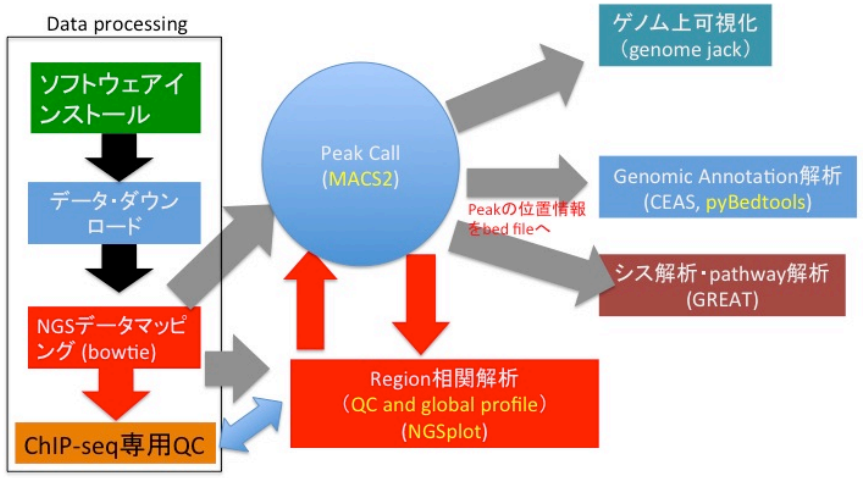


不足しているQC

- 目的標本以外のcontamination評価
  - Mappingされないreadってなに？ (実験環境でのコンタミや、サンプル取り間違え)
- ChIP特有のartifact
- DNA sharingの不均一性
  - 不安定な実験手技
- GC含有量が生物種、regionによって違う
- PCR biasがある
- 失敗か成功かの判断が難しい

多くの場合、ポジコン、(よい)ネガコンが無い。
微妙なデータが出た場合に解釈が困っている人が多い

今回提案するworkflow



私のpipeline

Suimye's super strict

その1

```
#!/bin/sh
source ~/.bashrc

FILENAME=$1
HG19="/mnt/data/bio/genome/Homo_sapiens/NCBI/build37.2/Sequence/BWAIndex/genome.fa"
#time bwa index -a bwtsw ref/human/hg19p9_masked.fa
fastq_quality_filter -q 25 -p 90 -i $FILEANME -o $FILENAME.fastq

bwa aln -t 24 $HG19 $FILENAME.qc.fastq > $FILENAME.sai

bwa samse /mnt/data/bio/genome/Homo_sapiens/NCBI/build37.2/Sequence/BWAIndex/genome.fa $FILENAME.sai $FILENAME.qc.fastq > $FILENAME.sam

cat $FILENAME.sam |perl -e 'while(<>){print $_ if(/\@SQ|\@PG/);}' >$FILENAME.header.txt

cat $FILENAME.sam |grep "X0:i:1" >$FILENAME.uq.sam.old
cat $FILENAME.header.txt $FILENAME.uq.sam.old >$FILENAME.uq.sam

samtools view $FILENAME.uq.sam -Shb >$FILENAME.uq.bam
samtools sort $FILENAME.uq.bam $FILENAME.uq.sort

java -jar ~/tools/picard-tools-1.119/MarkDuplicates.jar INPUT=$FILENAME.uq.sort.bam OUTPUT=$FILENAME.drm.bam METRICS_FILE=$FILENAME.out.metrics AS=true REMOVE_DUPLICATES=true VALIDATION_STRINGENCY=SILENT

intersectBed -a $FILENAME.drm.bam -b /mnt/data/database/wgEncodeDacMapabilityConsensusExcludable.bed -v >$FILENAME.drm.blst.bam
intersectBed -a $FILENAME.drm.blst.bam -b /mnt/data/database/hg19.rmsk.2.bed -v > $FILENAME.drm.blst.rmsk.bam
samtools index $FILENAME.drm.rmsk.blst.bam
bamToBed -i $FILENAME.drm.blst.rmsk.bam > $FILENAME.drm.blst.rmsk.bed
#bedToBam -i $FILENAME.drm.rmsk.blst.bed -g /home/admin/src/bedtools2/genomes >$FILENAME.drm.rmsk.blst.bam
```

Suimye's pipline

Suimye's super strict

filtering categories

- Fastq
  - Phred Score
- Mapping
  - MAQ
  - Unique mapped read
- PCR duplicates
  - MarkDuplicates
- サンプル間、リード分布QC
  - Ngsplotのdistribution
- 経験的filtering
  - BLACK List
  - Repeat masked region by repeatMasker

Mapping後のデータのクレンジング

Suimye's super strict

注意点

- MAQ値 or SAM情報
  - Mappingが正常にされているreadのみを用いる（主にmultihit）
  - SAMの情報はalignerによって違うので注意する

Multi-mapped readを除く

Bowtie: XSタグを使う

```
cat sample.sam |perl -e 'while(<>){print $_ if(/\@SQ|\@PG/);}' >sample.header.txt
grep -v "XS" sample.sam.old > sample.uq.sam.old
cat sample.header.txt sample.uq.sam.old >sample.uq.sam
```

BWA: X0タグを使う

```
cat sample.sam |perl -e 'while(<>){print $_ if(/\@SQ|\@PG/);}' >sample.header.txt
grep "X0:i:1" sample.sam.old >sample.uq.sam.old
cat sample.header.txt sample.uq.sam.old >sample.uq.sam
```

https://github.com/suimye/NGS_handson2015/wiki/NGS_beginner

Mapping の時の注意

Suimye's super strict

注意点

- masked genomeを使わない
  - 本来maskedされている部分に張り付くはずだったreadが、他の領域に張り付くことになる
- Multihit readは必ず除く
  - bowtieのオプションやmapping後のMAQ値を利用しても除ける。
  - ただし、SAM/BAMファイルの情報はmapperによって変わるのでマニュアルをしっかり見る。

Artifact cleaning

Suimye's super strict

その3

BedtoolsのintersectBedを利用したArtifactの除去 (オプションが –abamの場合と –aのままで良いバージョンあり)

- **intersectBed -abam $FILENAME.drm.bam -b /mnt/data/database/wgEncodeDacMapabilityConsensusExcludable.bed -v > $FILENAME.drm.blst.bam**
- **intersectBed -abam $FILENAME.drm.blst.bam -b /mnt/data/database/hg19.rmsk.2.bed -v > $FILENAME.drm.blst.rmsk.bam**
- **samtools index $FILENAME.drm.rmsk.blst.bam**

ChIP-seqをはじめとするDNA-seqでは、repeat配列や、構造上、性質上残りやすいDNA断片が存在する。

- Blacklist (ENCODEプロジェクトで利用しているArtifactの領域情報)
- Masked genomic sequences by RepeatMasker

repeatMaskerによるrepeat領域等の除去は、時に転写因子のpeakを壊してしまうことがあります。Repeat領域（またはその周囲）に結合する転写因子もあるので、最後のintersectBed処理はあり/なし両方のバージョンを作成した方がよいです。

Repeat情報の作り方: https://github.com/suimye/NGS_handson2015/wiki/repeat-region-from-UCSC_table_browser

https://github.com/suimye/NGS_handson2015/wiki/NGS_beginner

私のpipeline

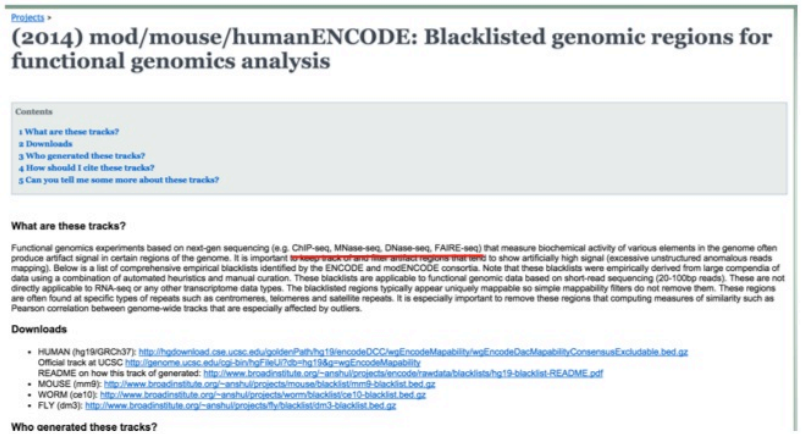
Suimye's super strict

その2

- #!/bin/sh
- source ~/.bashrc
- FILENAME=$1
- HG19="/mnt/data/bio/genome/Homo_sapiens/NCBI/build37.2/Sequence/BWAIndex/genome.fa"
- #time bwa index -a bwtsw ref/human/hg19/t9_masked.fa
- fastq_quality_filter -q 25 -p 90 -i $FILEANME -o $FILENAME.fastq
- bwa aln -t 24 $HG19 $FILENAME.qc.fastq > $FILENAME.sai
- bwa samse /mnt/data/bio/genome/Homo_sapiens/NCBI/build37.2/Sequence/BWAIndex/genome.fa $FILENAME.sai $FILENAME.qc.fastq > $FILENAME.sam
- **cat $FILENAME.sam |perl -e 'while(<>){print $_ if(/\@SQ|\@PG/);}' >$FILENAME.header.txt**
- **cat $FILENAME.sam |grep "grep "X0:i:1"" >$FILENAME.uq.sam.old**
- **cat $FILENAME.header.txt $FILENAME.uq.sam.old >$FILENAME.uq.sam**
- **samtools view $FILENAME.uq.sam -Shb >$FILENAME.uq.bam**
- **samtools sort $FILENAME.uq.bam $FILENAME.uq.sort**
- java -jar ~/tools/picard-tools-1.119/MarkDuplicates.jar INPUT=$FILENAME.uq.sort.bam OUTPUT=$FILENAME.drm.bam METRICS_FILE=$FILENAME.out.metrics AS=true REMOVE_DUPLICATES=true VALIDATION_STRINGENCY=SILENT
- intersectBed -a $FILENAME.drm.bam -b /mnt/data/database/wgEncodeDacMapabilityConsensusExcludable.bed -v >$FILENAME.drm.blst.bam
- intersectBed -a $FILENAME.drm.blst.bam -b /mnt/data/database/hg19.rmsk.2.bed -v > $FILENAME.drm.blst.rmsk.bam
- samtools index $FILENAME.drm.rmsk.blst.bam
- bamToBed -i $FILENAME.drm.blst.rmsk.bam > $FILENAME.drm.blst.rmsk.bed
- #bedToBam -i $FILENAME.drm.rmsk.blst.bed -g /home/admin/src/bedtools2/genomes >$FILENAME.drm.rmsk.blst.bam

THE BLACK LIST

ENCODE projectのblack list
ChIP-seq Mnase-seq, DNase-seq FAIRE-seqなどに利用。



Projects »

(2014) mod/mouse/humanENCODE: Blacklisted genomic regions for functional genomics analysis

Contents

1 What are these tracks?
2 Downloads
3 Who generated these tracks?
4 How should I cite these tracks?
5 Can you tell me some more about these tracks?

What are these tracks?

Functional genomics experiments based on next-gen sequencing (e.g. ChIP-seq, MNase-seq, DNase-seq, FAIRE-seq) that measure biochemical activity of various elements in the genome often produce artifact signal in certain regions of the genome. It is important to keep track of such [illegible] to show artifically high signal (excessive unstructured anomalous reads mapping). Below is a list of comprehensive empirical blacklists identified by the ENCODE and modENCODE consortia. Note that these blacklists were empirically derived from large compendia of data using a combination of automated heuristics and manual curation. These blacklists are applicable to functional genomic data based on short-read sequencing (20-100bp reads). These are not directly applicable to RNA-seq or any other transcriptome data types. The blacklisted regions typically appear uniquely mappable so simple mappability filters do not remove them. These regions are often found at specific types of repeats such as centromeres, telomeres and satellite repeats. It is especially important to remove these regions that computing measures of similarity such as Pearson correlation between genome-wide tracks that are especially affected by outliers.

Downloads

- HUMAN (hg19/GRCh37): http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz
  Official track at UCSC http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability
  README on how this track of generated: http://www.broadinstitute.org/~anshul/projects/encode/rawdata/blacklists/hg19-blacklist-README.pdf
- MOUSE (mm9): http://www.broadinstitute.org/~anshul/projects/mouse/blacklist/mm9-blacklist.bed.gz
- WORM (ce10): http://www.broadinstitute.org/~anshul/projects/worm/blacklist/ce10-blacklist.bed.gz
- FLY (dm3): http://www.broadinstitute.org/~anshul/projects/fly/blacklist/dm3-blacklist.bed.gz

Who generated these tracks?

https://sites.google.com/site/anshulkundaje/projects/blacklists

PCR duplicates

Suimye's super strict

その3

- java -jar ~/tools/picard-tools-1.119/MarkDuplicates.jar INPUT=$FILENAME.uq.sort.bam OUTPUT=$FILENAME.drm.bam METRICS_FILE=$FILENAME.out.metrics AS=true REMOVE_DUPLICATES=true VALIDATION_STRINGENCY=SILENT
- intersectBed -a $FILENAME.drm.bam -b /mnt/data/database/wgEncodeDacMapabilityConsensusExcludable.bed -v >$FILENAME.drm.blst.bam
- intersectBed -a $FILENAME.drm.blst.bam -b /mnt/data/database/hg19.rmsk.2.bed -v > $FILENAME.drm.blst.rmsk.bam
- samtools index $FILENAME.drm.rmsk.blst.bam
- bamToBed -i $FILENAME.drm.blst.rmsk.bam > $FILENAME.drm.blst.rmsk.bed
- #bedToBam -i $FILENAME.drm.rmsk.blst.bed -g /home/admin/src/bedtools2/genomes >$FILENAME.drm.rmsk.blst.bam

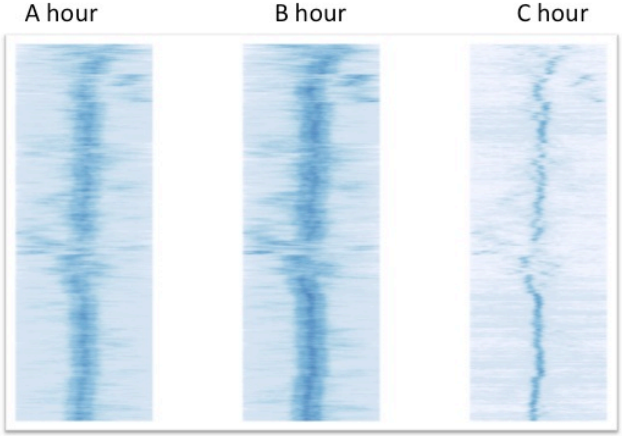
Peak Callの時のMACSでは、標準でduplicateをカウントしないが、その後の可視化や、tagの再カウントなど様々な状況で要求されるので、必ず実施する。

MACS option: **--keep-dup**
It controls the MACS behavior towards duplicate tags at the exact same location -- the same coordination and the same strand. The default 'auto' option makes MACS calculate the maximum tags at the exact same location based on binomal distribution using 1e-5 as pvalue cutoff; and the 'all' option keeps every tags. If an integer is given, at most this number of tags will be kept at the same location. **Default: 1.**

https://github.com/suimye/NGS_handson2015/wiki/NGS_beginner

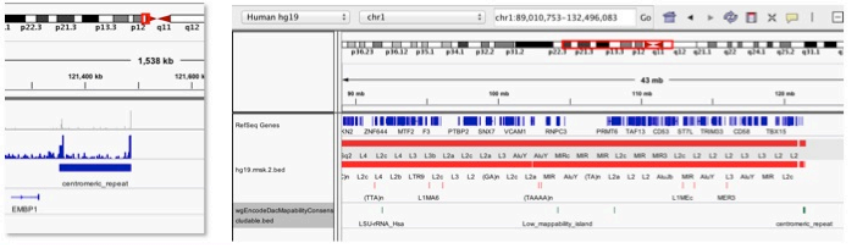
TSSを中心としたリードの分布

某NGS dataのtime series



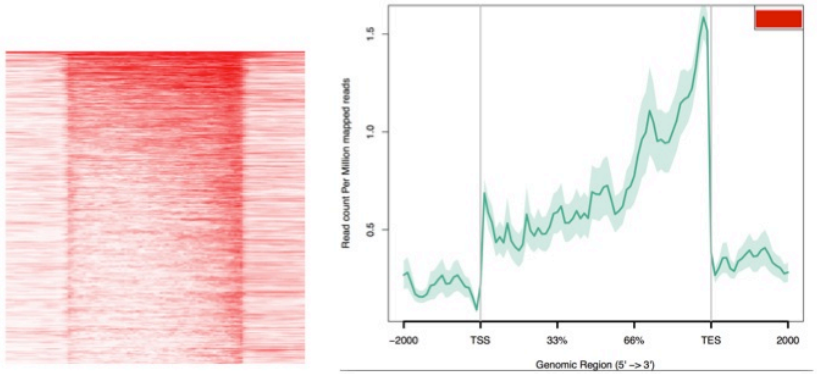
TSSを中心として、+- 1000のtag densityをheatmapでみている

BLACK LISTとrepeatMaskerの領域情報



例: chr1 セントロメア付近のBLACK LIST

polyA RNA-seqのNgsplot



polyA-seqの場合、3' biasがかならず観測されるはず。

Ngsplotを用いたQC

Dry本: p251
https://github.com/suimye/NGS_handson2015/wiki/NGSplotsOnBiolinux8

IGVの使い方

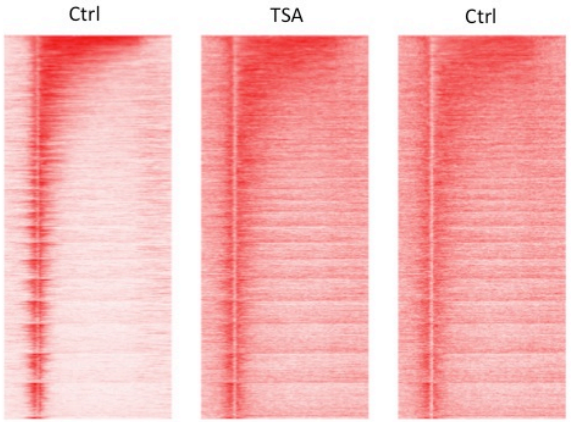
Integrative Genomics Viewer

1. 綺麗
2. local
3. DATA serverへのHTTP経由でのアクセスが可能
   - (https://www.broadinstitute.org/igv/DataServer)
4. タグextensionが容易
5. 軽快 (特にtdf)
   - read/millionでの規格化でデータを表示してくれる
   - 複数のデータを一気に可視化できる
6. Bam,bed,BW,VCF,GFF,GTFなどあらゆるファイルを可視化可能

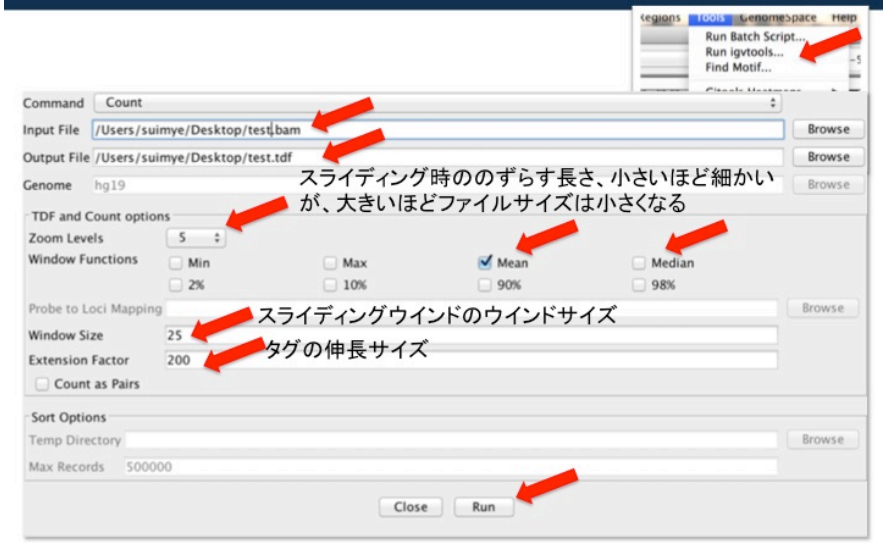
Total read数での規格化方法

P265-7のデータでNgsplotしてみました

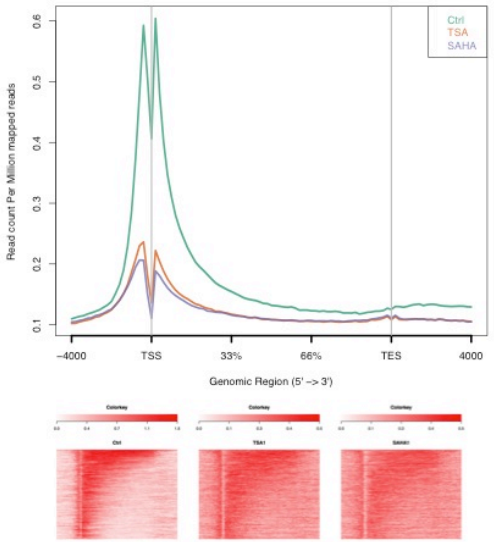
Gene bodyのtag density



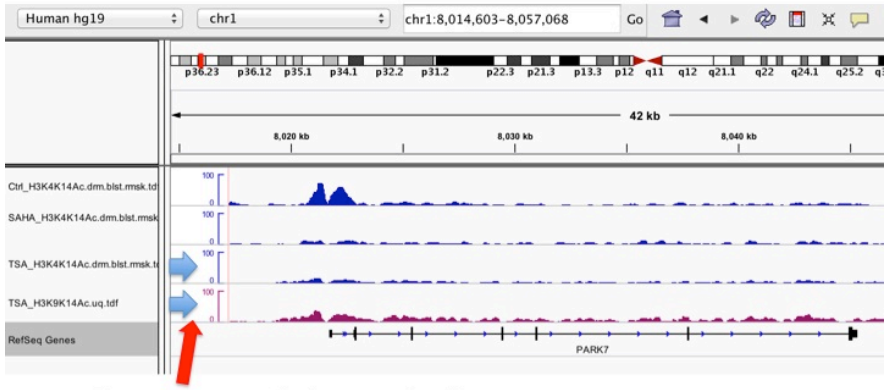
IGVの使い方(tdfファイルの作り方(GUI))



Heatmapに騙されないように



QC前後で比較



Y軸はtotal readによる規格化していない描画

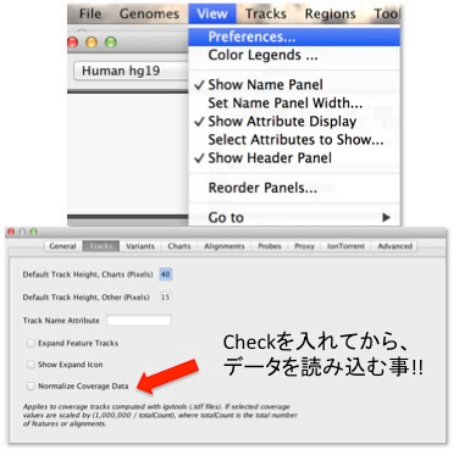
H3K4me3のデータをIGVで可視化したもの (Y軸は5bp毎の平均tag数を上限100として表示)（規格化済み、データ間の比較ができるようになった）。

IGVの使い方

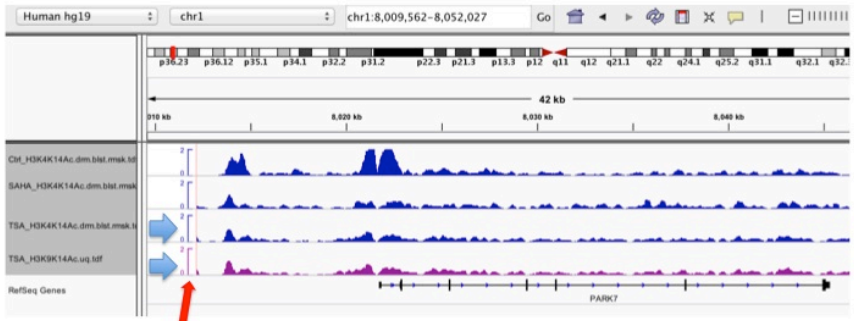
Integrative Genomics Viewer

1. 綺麗
2. DATA serverへのHTTP経由でのアクセスが可能
   - (https://www.broadinstitute.org/igv/DataServer)
3. タグextensionが容易
4. 軽快 (特にtdf)
   - read/millionでの正規化でデータを表示してくれる
5. Bam,bed,BW,VCF,GFF,GTFなどあらゆるファイルを可視化可能

tdfファイルに変換する事で、より現実的なChIP-seq Peakをみることができる。



QC前後で比較



Y軸はtotal mapped readによる規格化を実施

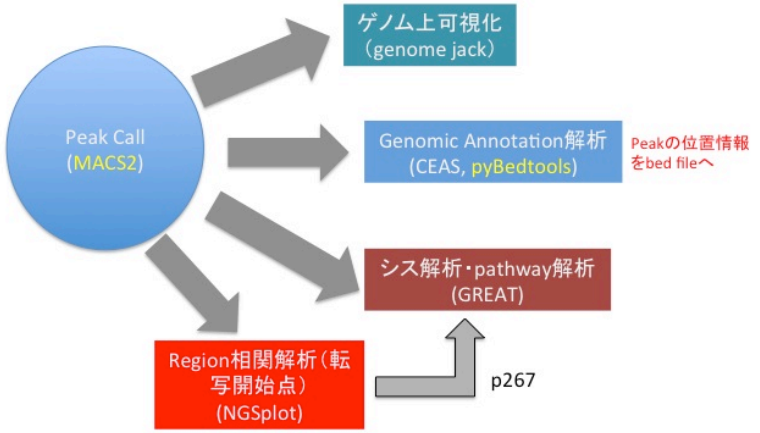
H3K4me3のデータをIGVで可視化したもの (Y軸を上限2として表示)（規格化済み、データ間の比較ができるようになった）。

以上をふまえてデータを比較してみる

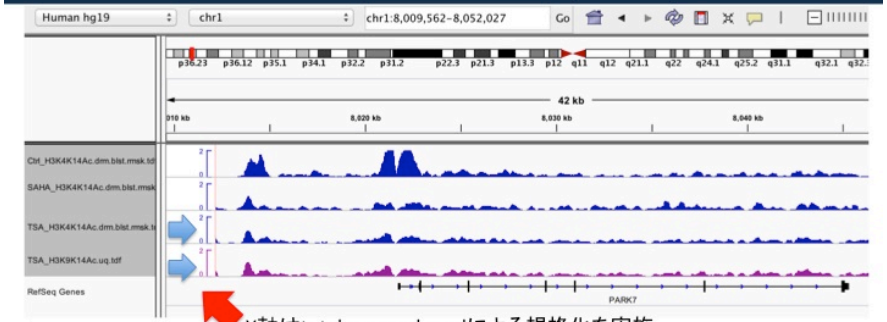
QC and クレンジング v.s. DRY本 pipeline

DRY本のエピゲノム解析workflow

をまとめて、眺めてみよう



QCとデータクレンジング後の結果



Y軸はtotal mapped readによる規格化を実施

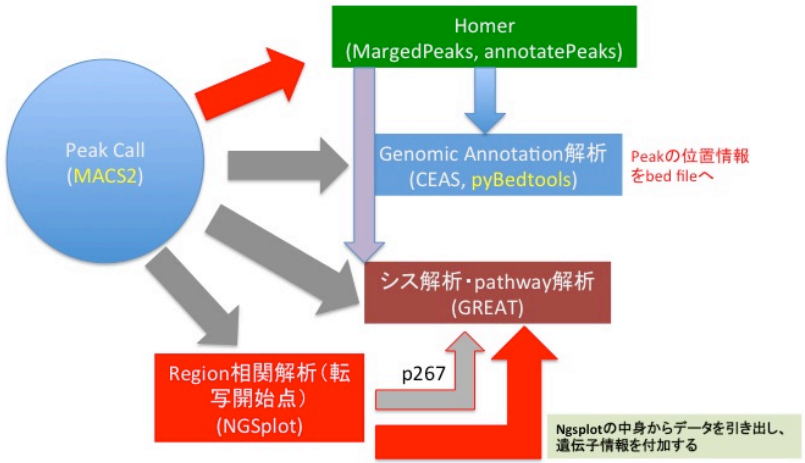
H3K4me3のデータをIGVで可視化したもの (Y軸を上限2として表示)（規格化済み、データ間の比較ができるようになった）。

The number of identified H3K9K14Ac Peaks (same MACS parameters)

	Ctrl	TSA	SAHA
p263	38,211	41,367	40,483
Suimye's pipeline	34,686	16,638	12,777

Suimye's解析workflow

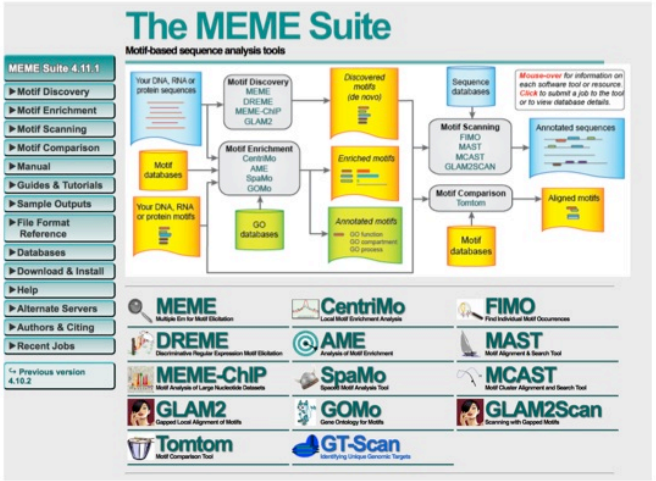
データの統合と解析



ChIP-seqデータの統合解析について

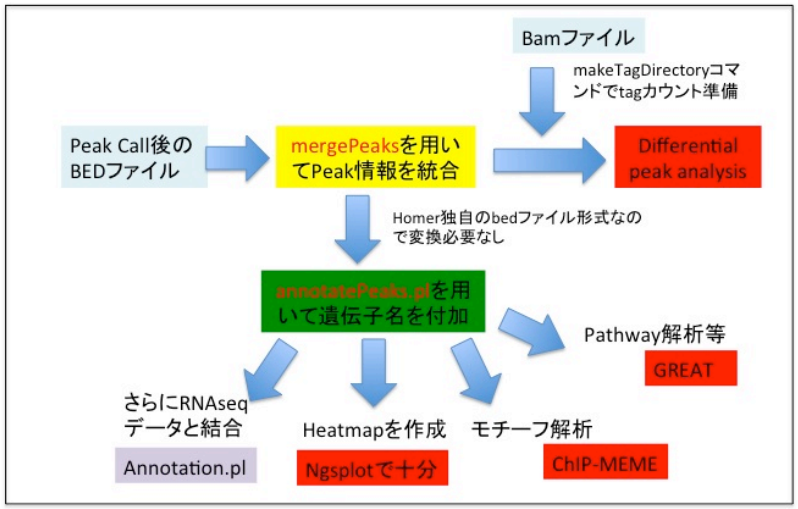
MEME Suite

モチーフ発見関連全てを網羅する



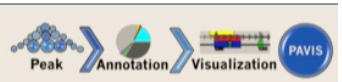
https://github.com/suimye/NGS_handson2015/wiki/PeakCallAndMDA

Homerを用いた統合

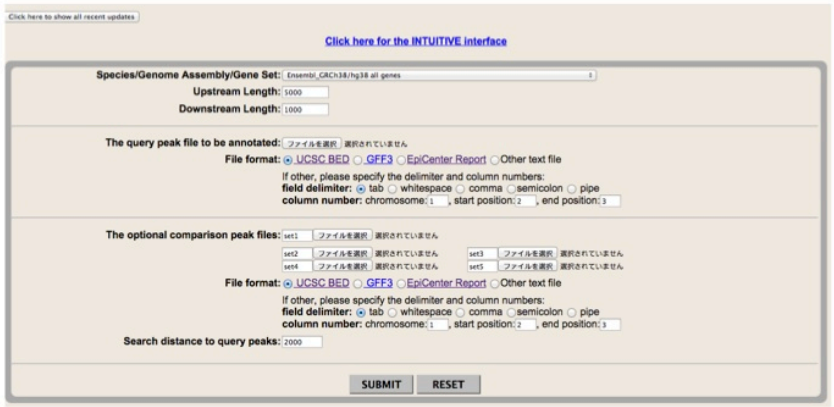


https://github.com/suimye/NGS_handson2015/wiki/Homer_Data_integration

PAVIS2



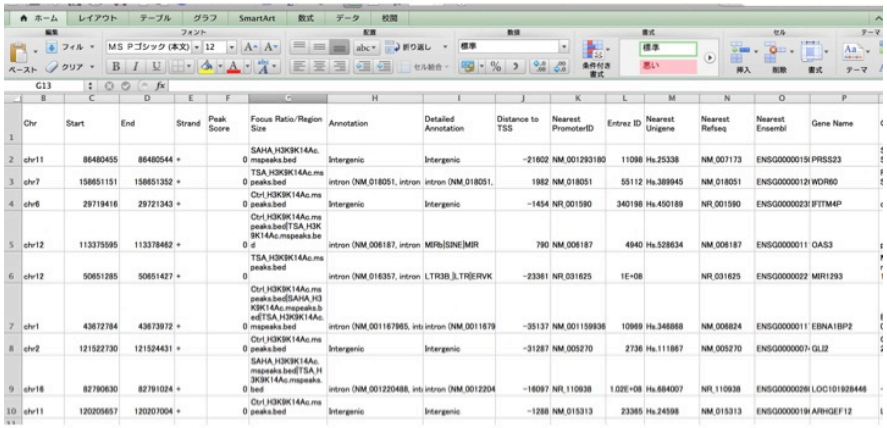
BEDファイルからベン図



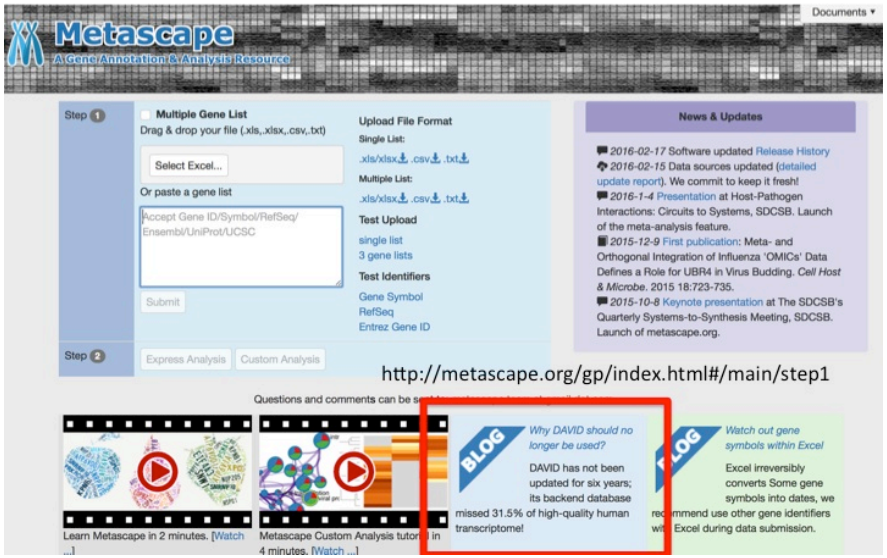
https://github.com/suimye/NGS_handson2015/wiki/PeakCallAndMDA

統合したデータの中身

Githubのページで説明



Metascape



nemui