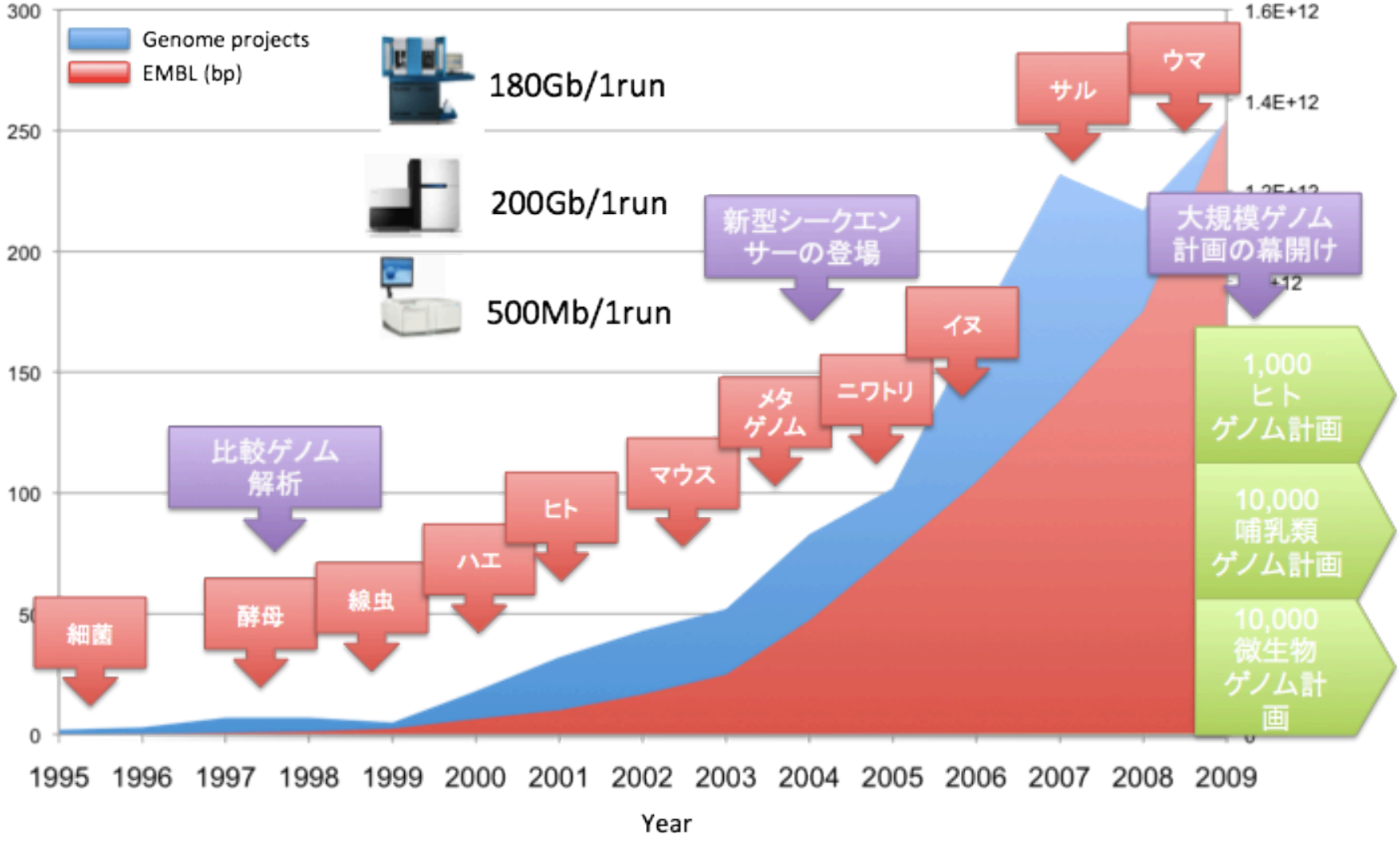
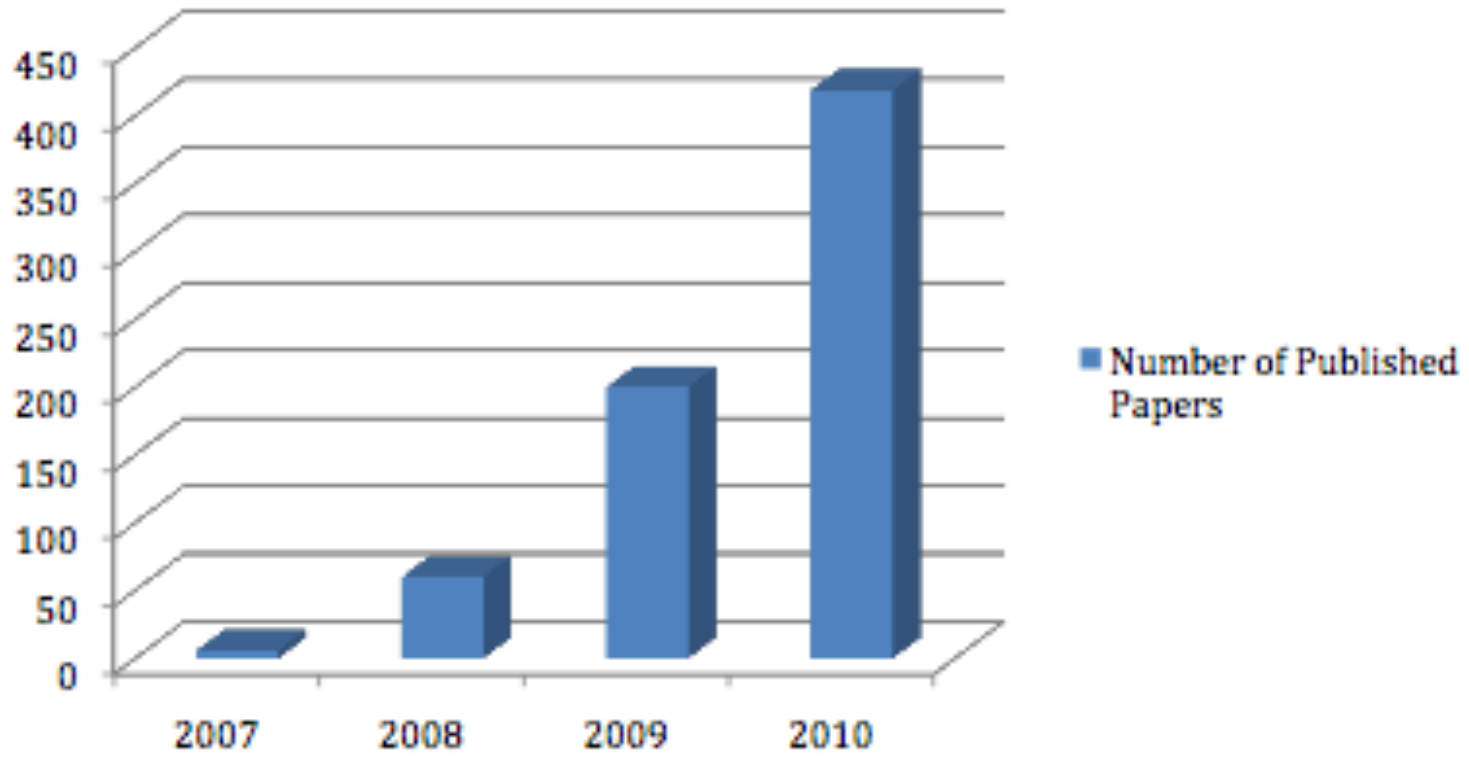
ChIP-seq:実験のデザインと落とし穴

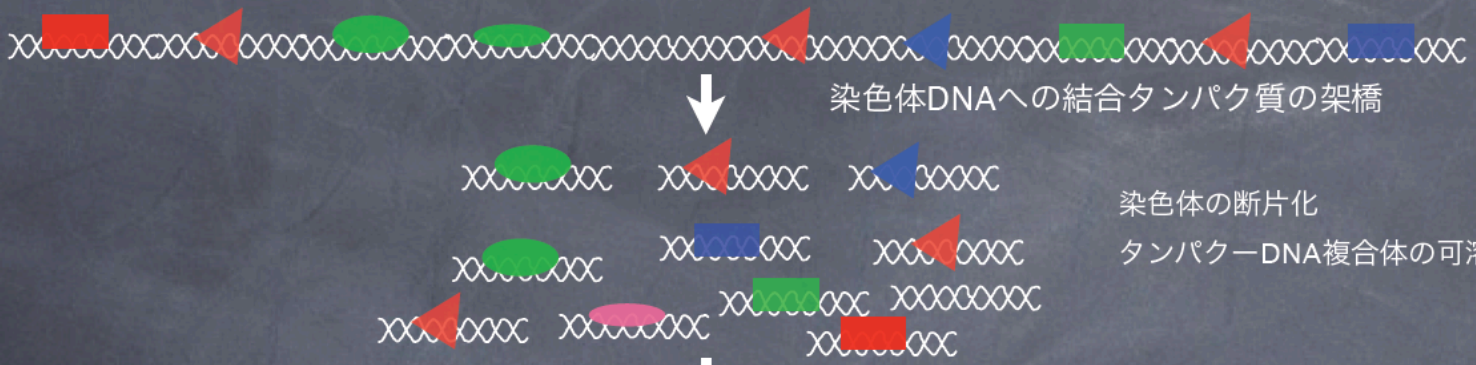
はじめに

ゲノム科学の爆発的発展



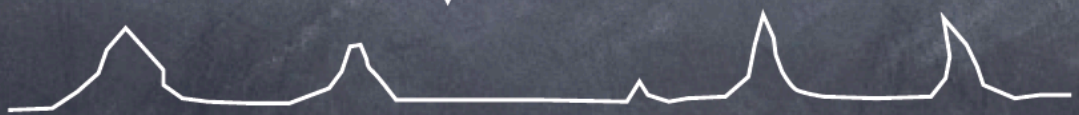


次世代シークエンサー関連の論文



染色体免疫沈降(ChIP)

DNA配列の決定(Sequence)

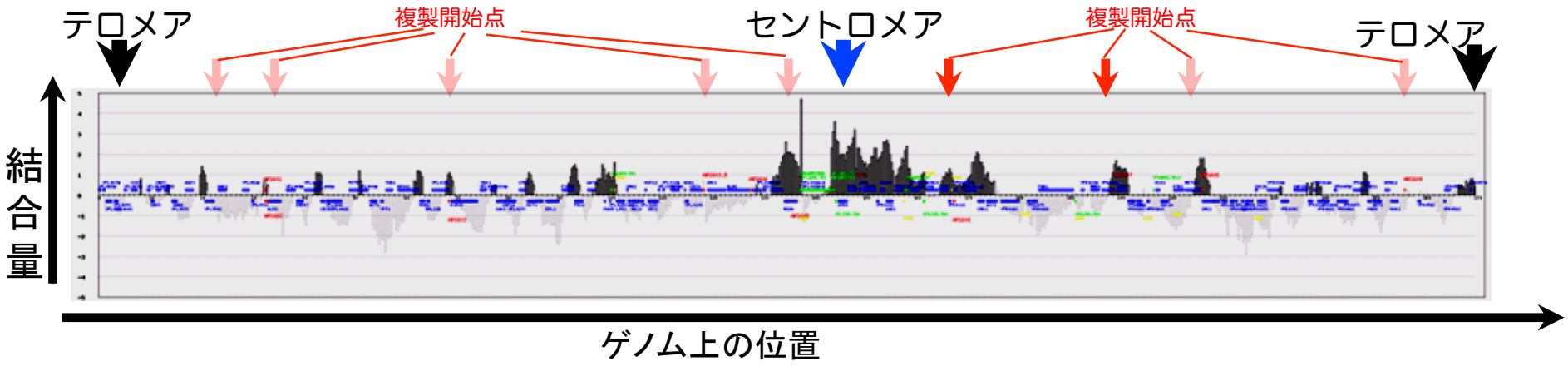


蛋白結合プロファイル

ゲノムへの
マッピング

全ゲノムレベルでタンパク結合位置解明のためのChIP-seq法
タンパクがいつどこでどのように働いているかを知る

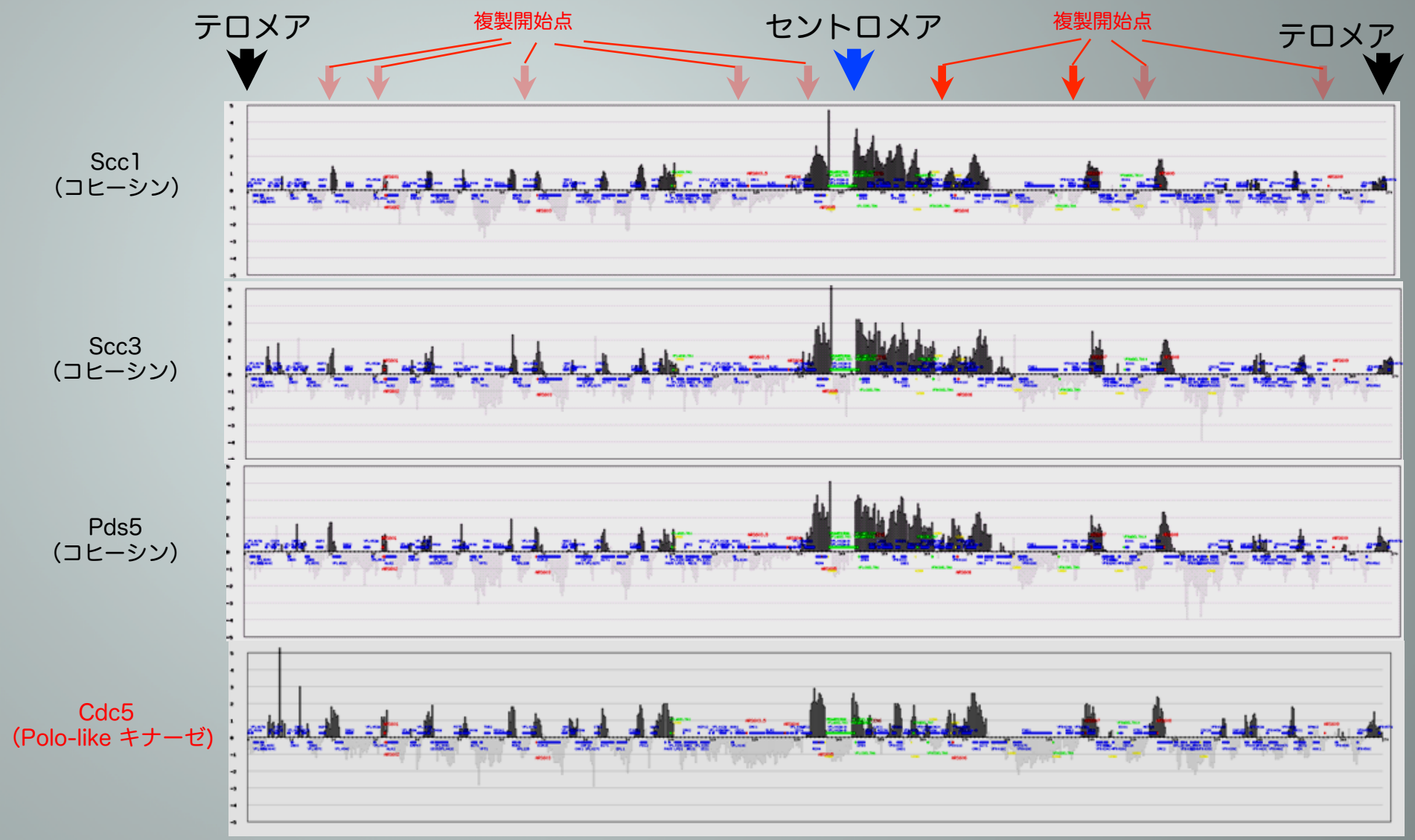
ChIP-seq法によるタンパク質の動態解析
ーゲノムのどの場所にいつ結合するのかを明らかにするー



- タンパクの位置情報は様々な機能情報をもたらす
  - 既知の機能配列との相関。既知の機能タンパクとの相関
    - 例えば二つのタンパク結合プロファイルが一致する
      - 二つは同じ機能を持つ複合体に帰属する可能性、制御関係にある可能性を示唆
  - 健常と疾患由来細胞におけるタンパクの振る舞いの違い
    ✳分子病態の解明

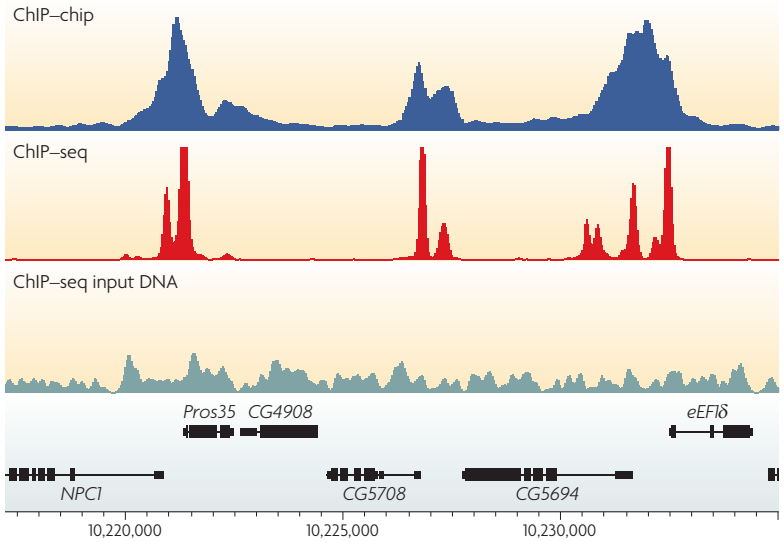
ChIP-seq法による制御系とタンパク複合体の予測例

コヒーシン複合体と Polo-like キナーゼ



なぜChIP-seqをわざわざ選択するのか？

- 網羅性
  - GeneChipの限界
    - ▶25bp probe (Repetitiveは×) 約半分のみをCover
  - Human
    - ▶25bp でuniq 79.6%(Rep. 63.4%)
    - ▶35bp でuniq 85.7%(Rep. 75.1%)
    - ▶50bp でuniq 91.1%(Rep. 85.3%)
  - シークエンス情報があれば良い

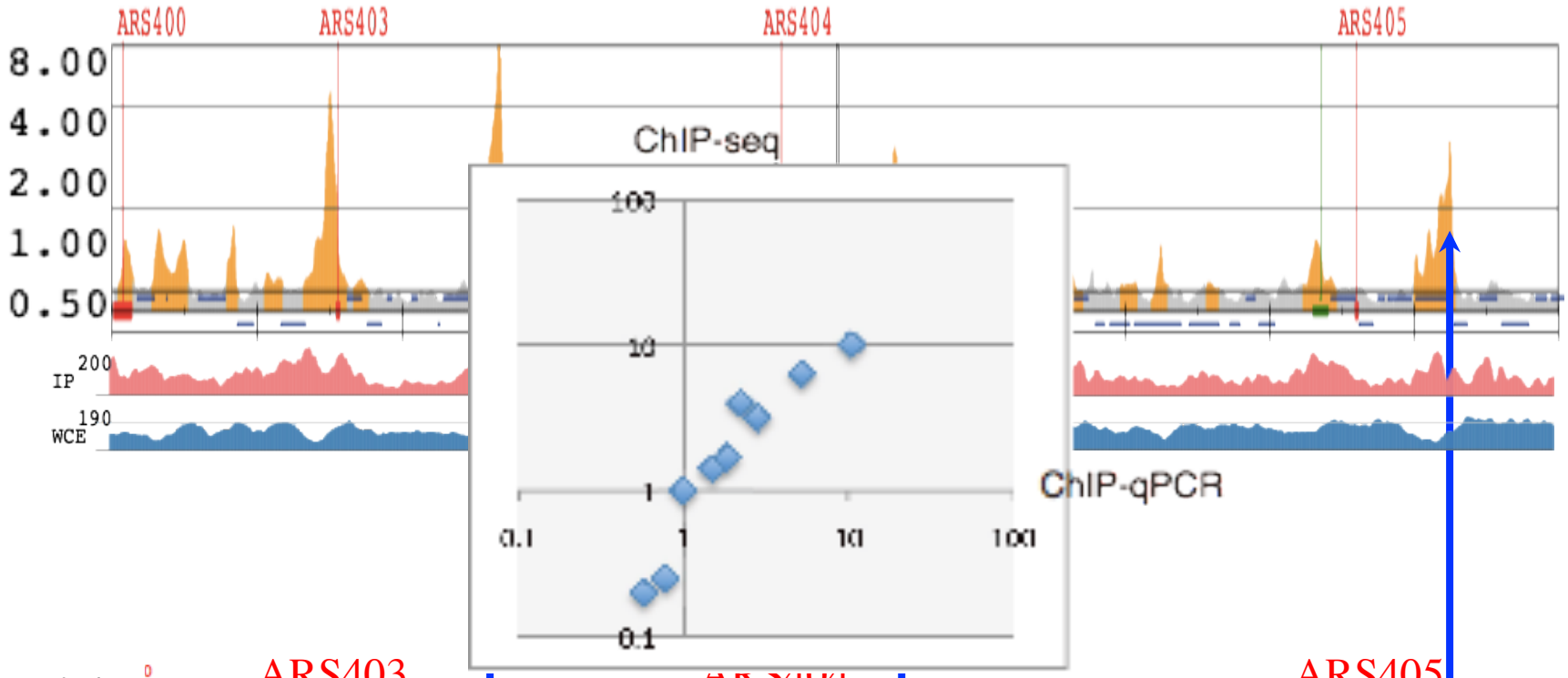


- 解析精度
  - 塩基配列を決定すること、リードを数えることはハイブリダイゼーションによる測定よりも定量的。
  - 解像度も高い

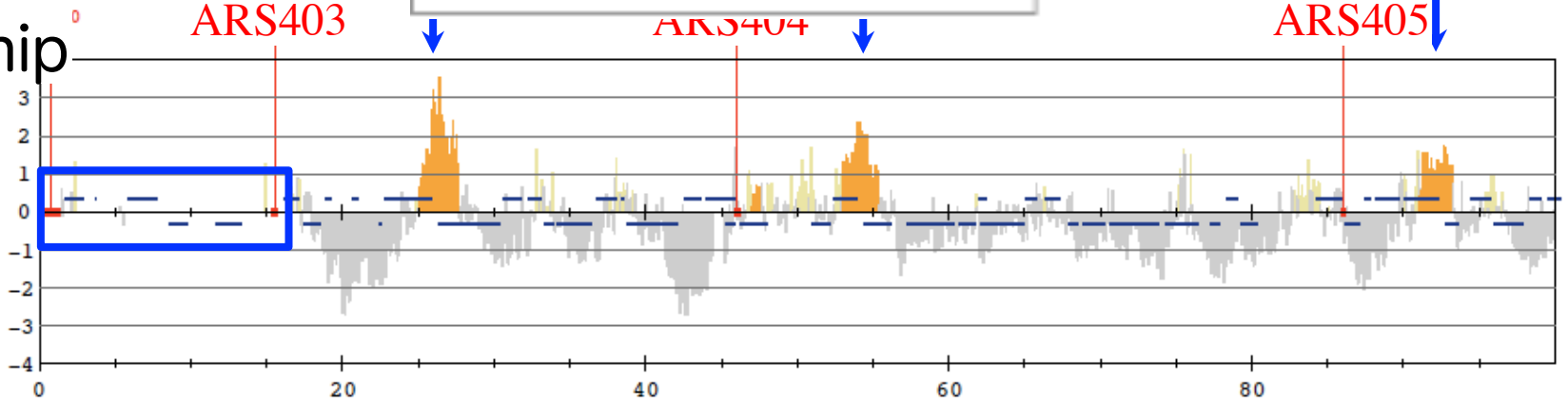
- 解析コスト
  - ChIP-chip 140万円／サンプル　ChIP-seq？？

ChIP-chip vs ChIP-seq (酵母の場合)

ChIP-seq



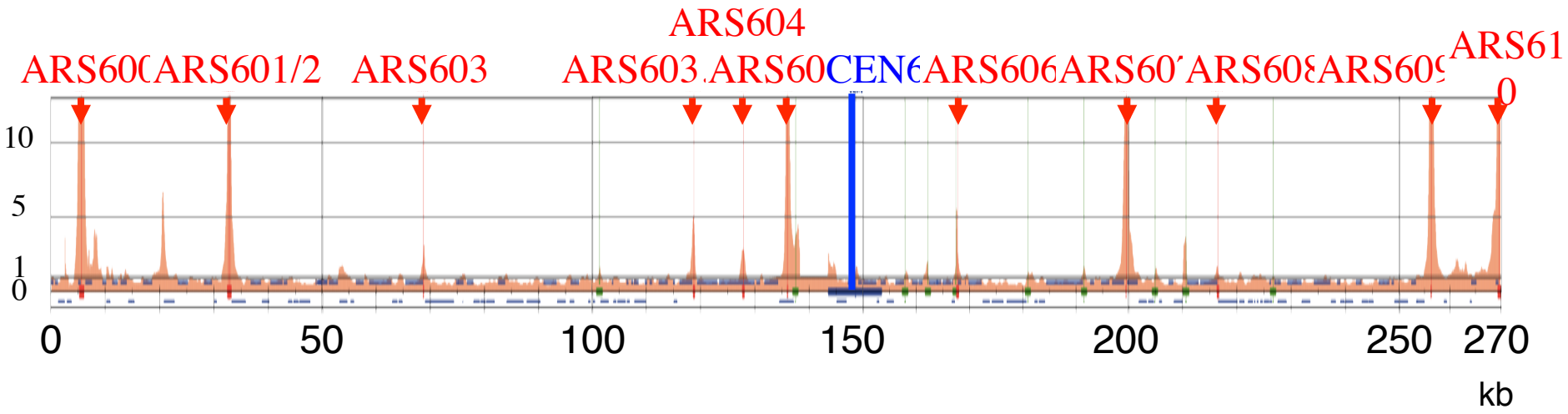
ChIP-chip

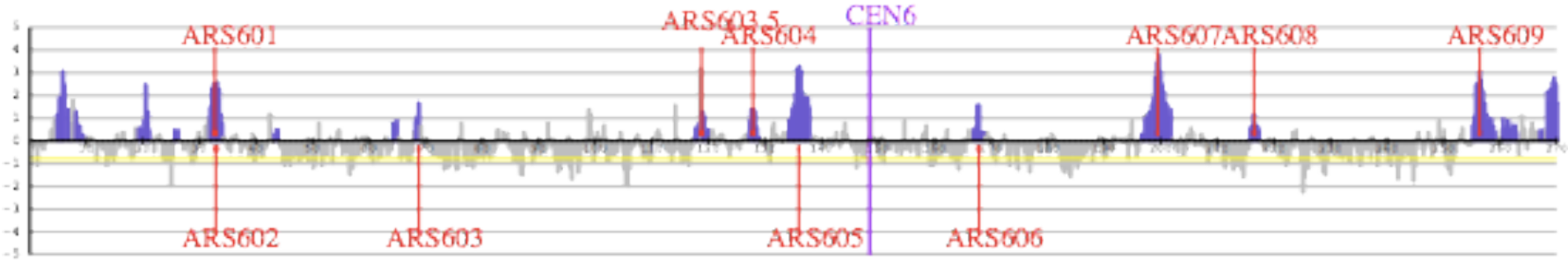


アクセスできる領域も増える。定量性はq-PCRのレベルではある。
q-PCRによる検証に意味はあるのか？KOによる検証の方が有意義。

Katou

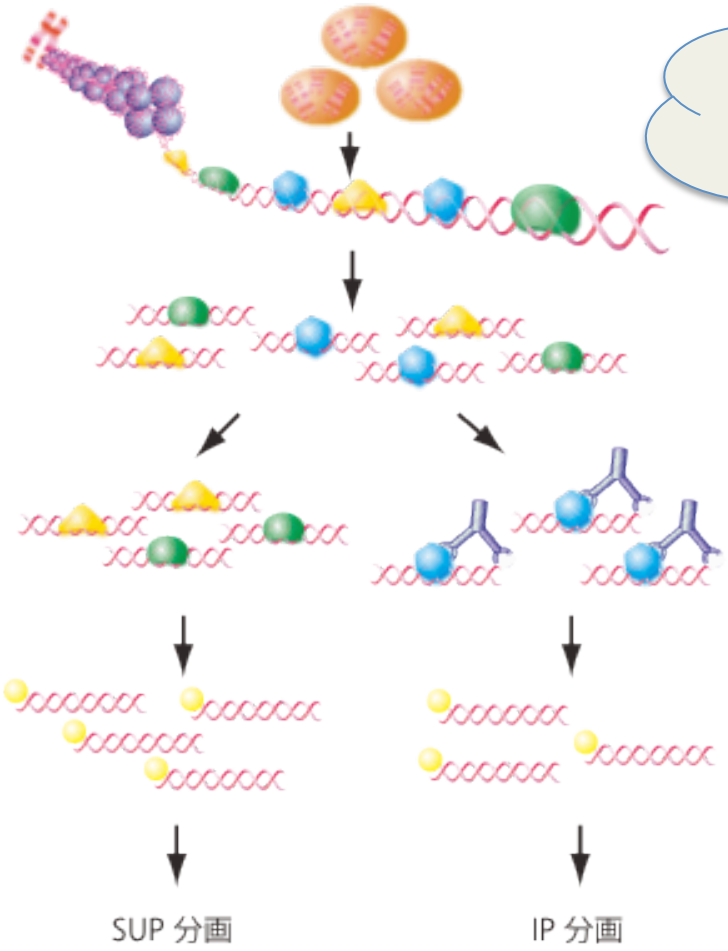
ChIP-chipとChIP-seqの比較



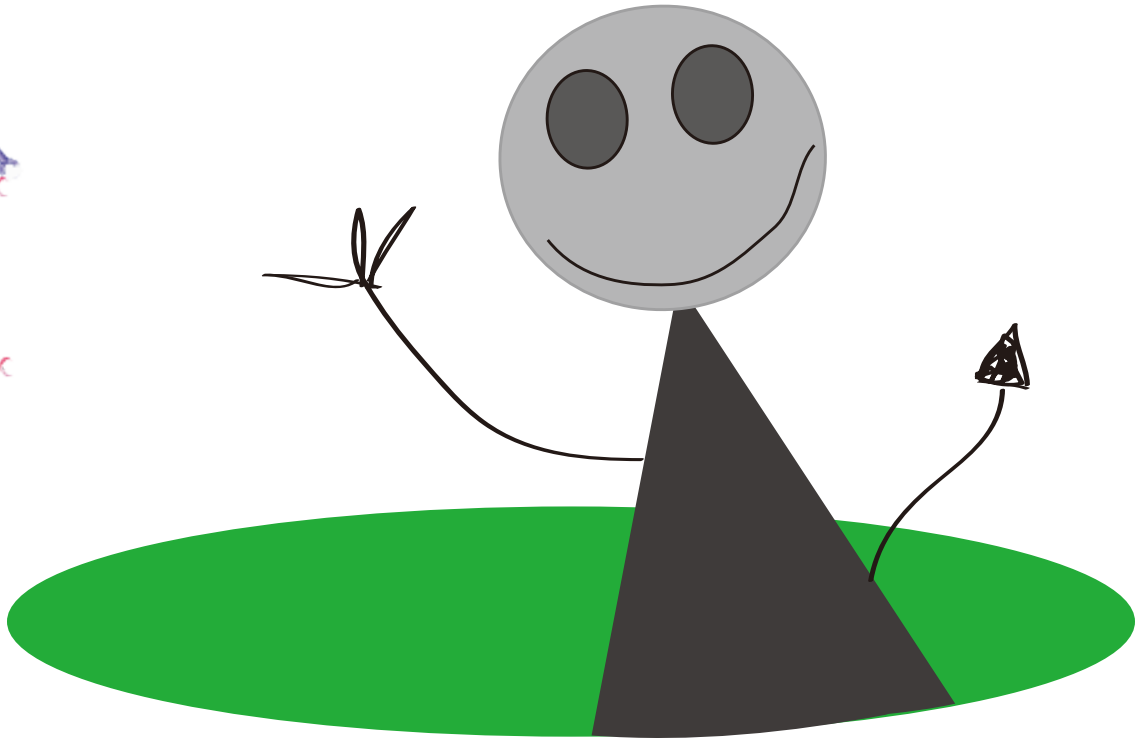


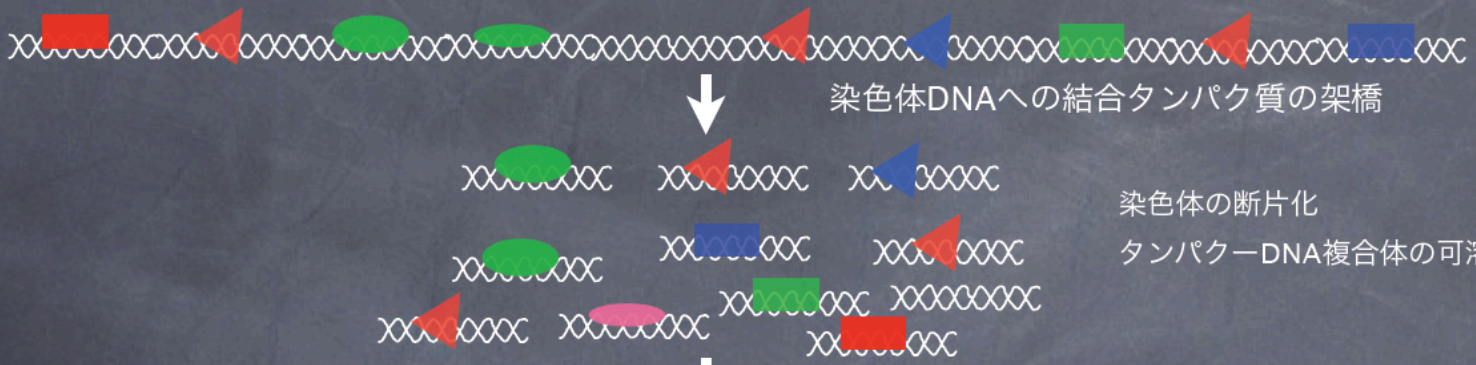
Katou

ChIP-seq:実験のデザインと落とし穴



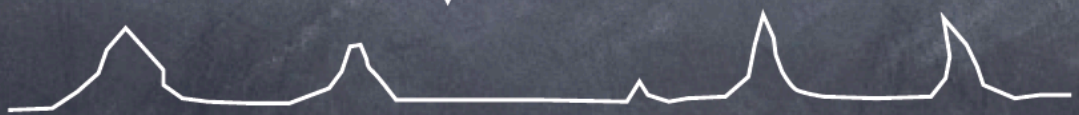
ウエット編





染色体免疫沈降(ChIP)

DNA配列の決定(Sequence)



蛋白結合プロファイル

ゲノムへの
マッピング

全ゲノムレベルでタンパク結合位置解明のためのChIP-seq法
タンパクがいつどこでどのように働いているかを知る

細胞固定

固定

DNA 断片化

免疫沈降

リバースクロスリンク
精製

ライブラリ作成

細胞数

ヒト細胞	酵母
1-2x10^7 ～ cells	5x10^8 ～ cells

この数で何も出なければ諦めましょうかという数です。
逆に最低必要な細胞数は10の6乗ぐらいです。
10の4乗では使える抗体は限られていると思います。

固定時間

ヒト細胞	酵母
～10min.	～30min. 時には ON

DNA 断片化（クロマチンの可溶化）

固定

DNA 断片化

免疫沈降

リバースクロスリンク
精製

ライブラリ作成

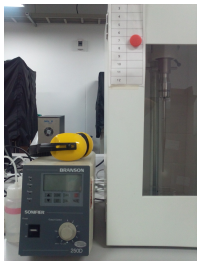
細胞破砕(酵母)

ON:60sec
OFF:60sec
2500rpm

x20cycles @4C

安井器械社製
マルチビーズショッカー

DNA断片化



Branson社製
Sonifier250D/
マイクロチップ使用

ヒト細胞	出芽酵母
Power1.5/12sec./8回	Power1.5/15sec./7回

基本は超音波、ただし、酵素による切断も。検討し、最適な方法を選ぶ。Nucleaseを使うとデータがnoisyになる傾向がある。

免疫沈降

	ヒト細胞	出芽酵母
抗体量	～ 10µg	～ 20µg
プロティンビーズ量	200µL	80µL
免沈時間	～ON	5hrs.～ON

リバースクロスリンクと精製

リバースクロスリンクの時間

ヒト細胞	出芽酵母
6hrs～	8hrs～

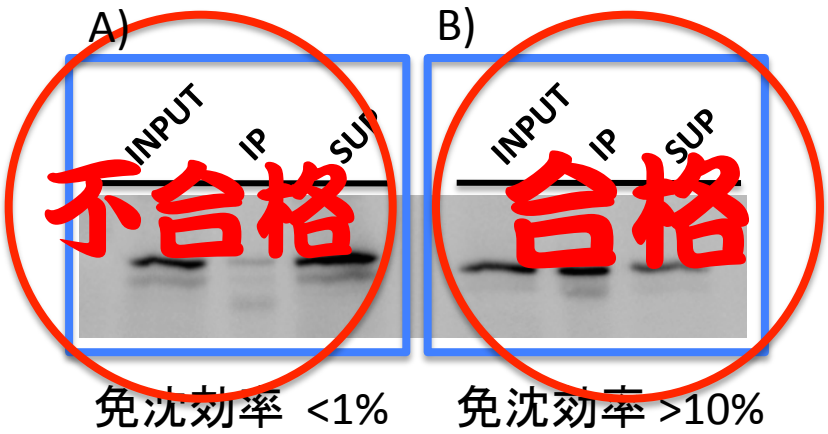
DNA精製(ProteinaseK、Rnase処理後)

QIAquick PCR Purification Kit　(QIAGEN)を使用

サンプルのチェック

結合する部位がわかっているのならば、**定量PCR**による確認が
一番「確か」だと思いますが、、、sampleがもったいない場合や候補が解らないときは、

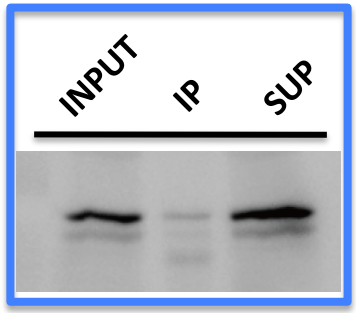
最低でも**ウエスタンブロッティングによる免沈効率の確認**は必要

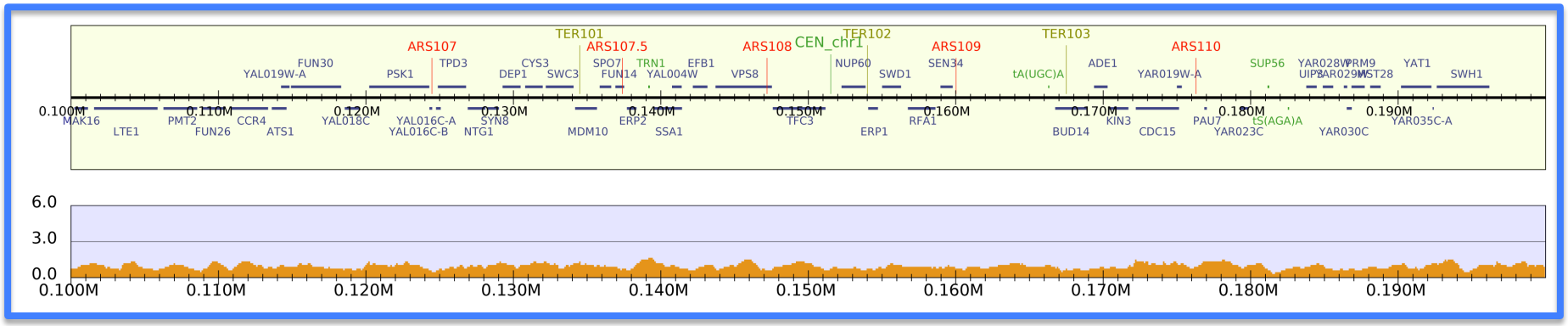


最近では投稿にあたり、このデータを求めるジャーナルもあります。
とにかく出ればいい、という発想は危険。
10%落ちていてもq-PCRで調べると0.1%程度の濃縮率なんてことはよくある。
抗体がChIPに使えるか否かの一つの指標は免疫抗体染色に使えるか否か（いずれも固定したサンプルを解析するため）。

ウエスタンによるチェック

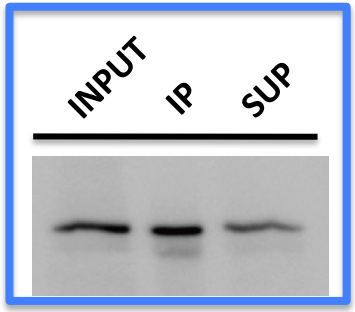
A)

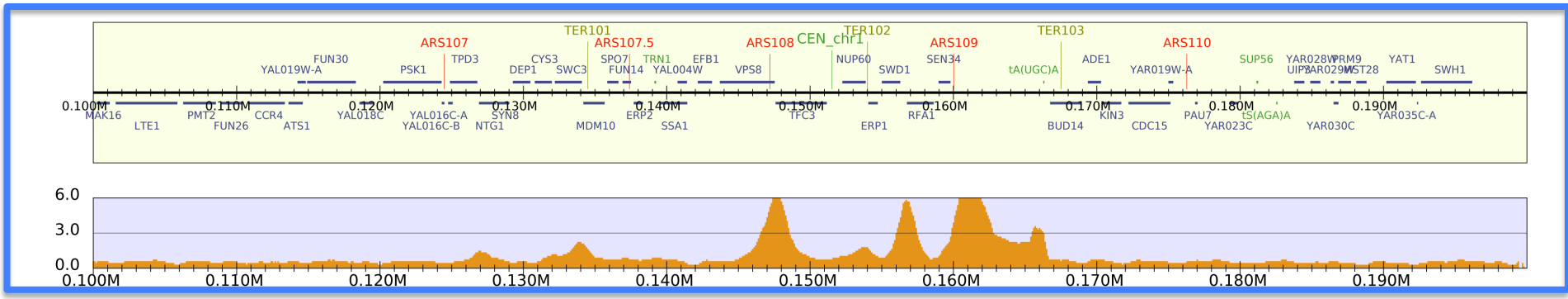




ウエスタンによるチェック

B)





ライブラリ作成

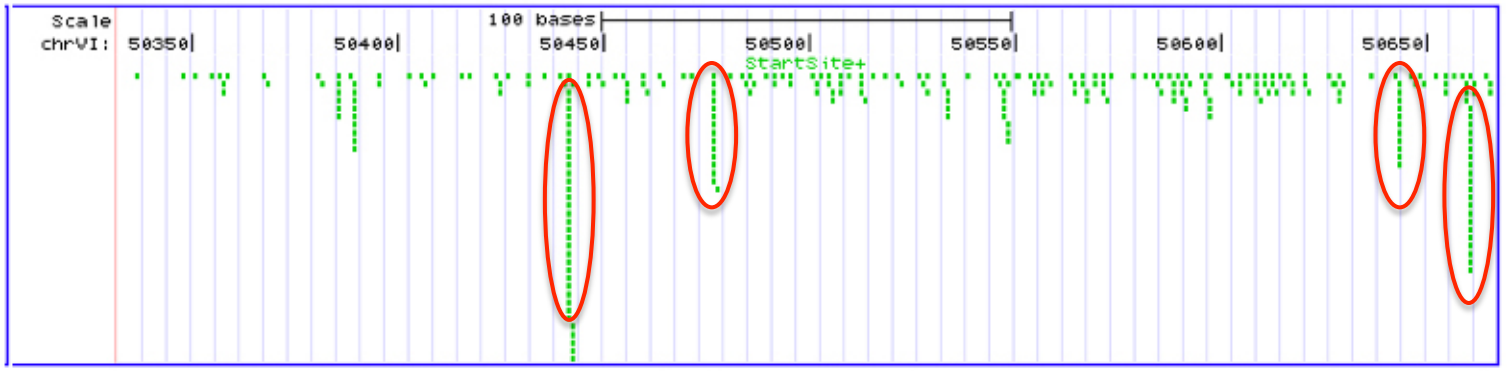
NEBNext ChIP-Seq Library Prep Master Mix Set for Illumina
(BioLabs社NEB #E6240S)

Multiplexing Sample Prep Oligo Kit
(illumina社 #PE-400-1001)

BioLabs社からOligoが提供されていなかった為、
複数の会社の試薬を使用しています。
今はBiolabs社からもOligoが提供されていますし、
イルミナ社からChIP-seq SamplePrep kitも提供されています。

リードの開始点の分布をみると、、、、(酵母の場合)

Genome DNA sample



100 bp

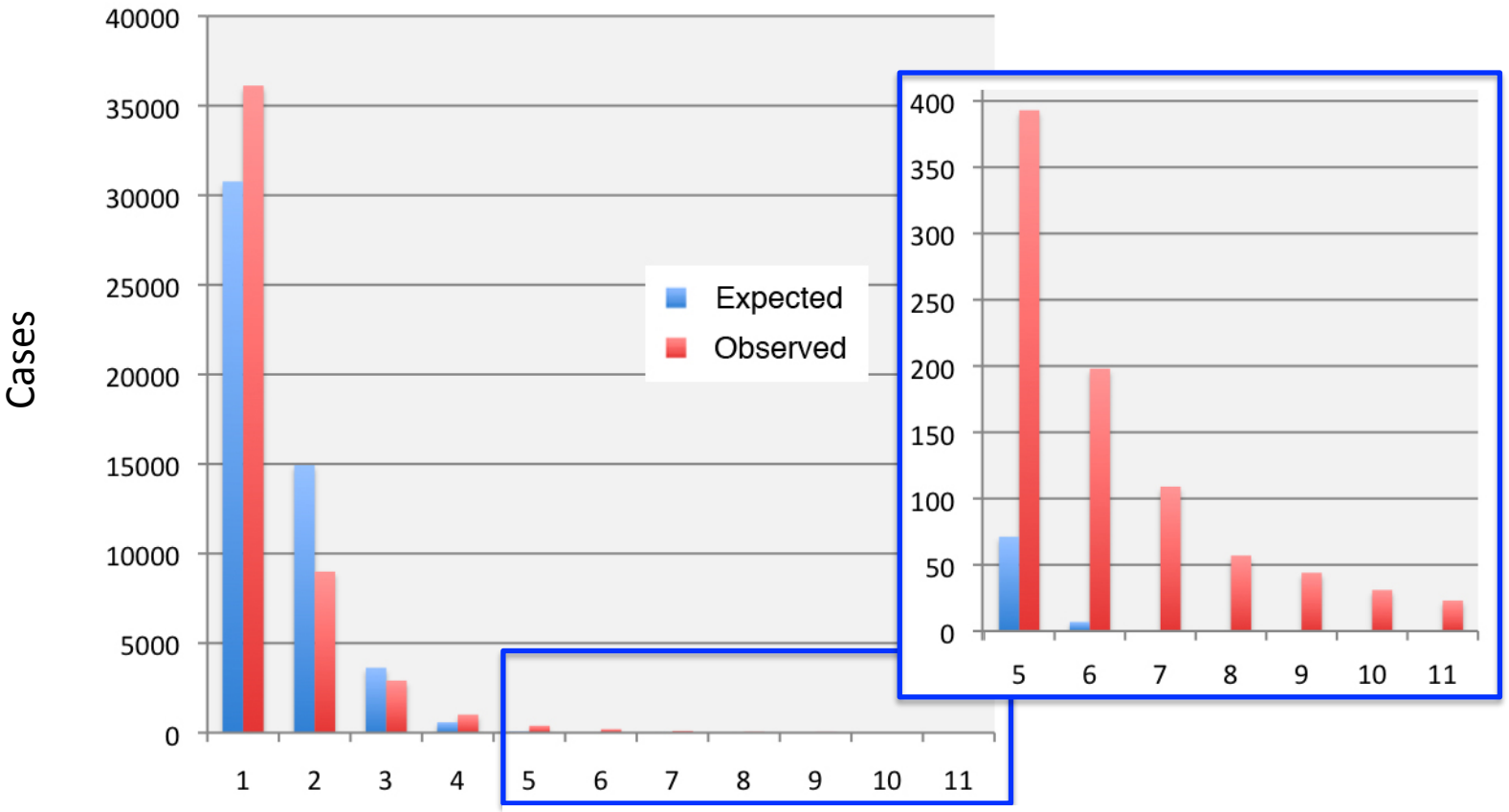
5-10リードが同じスタートサイトを持つ

ライブラリを作成する際のPCRバイアス？？

PCRバイアスによるアーティファクト

同じ開始点を有するリードの数は理論値と逆転している事例がある。

酵母六番染色体の50kbの領域についてみて見ると（24000リード）

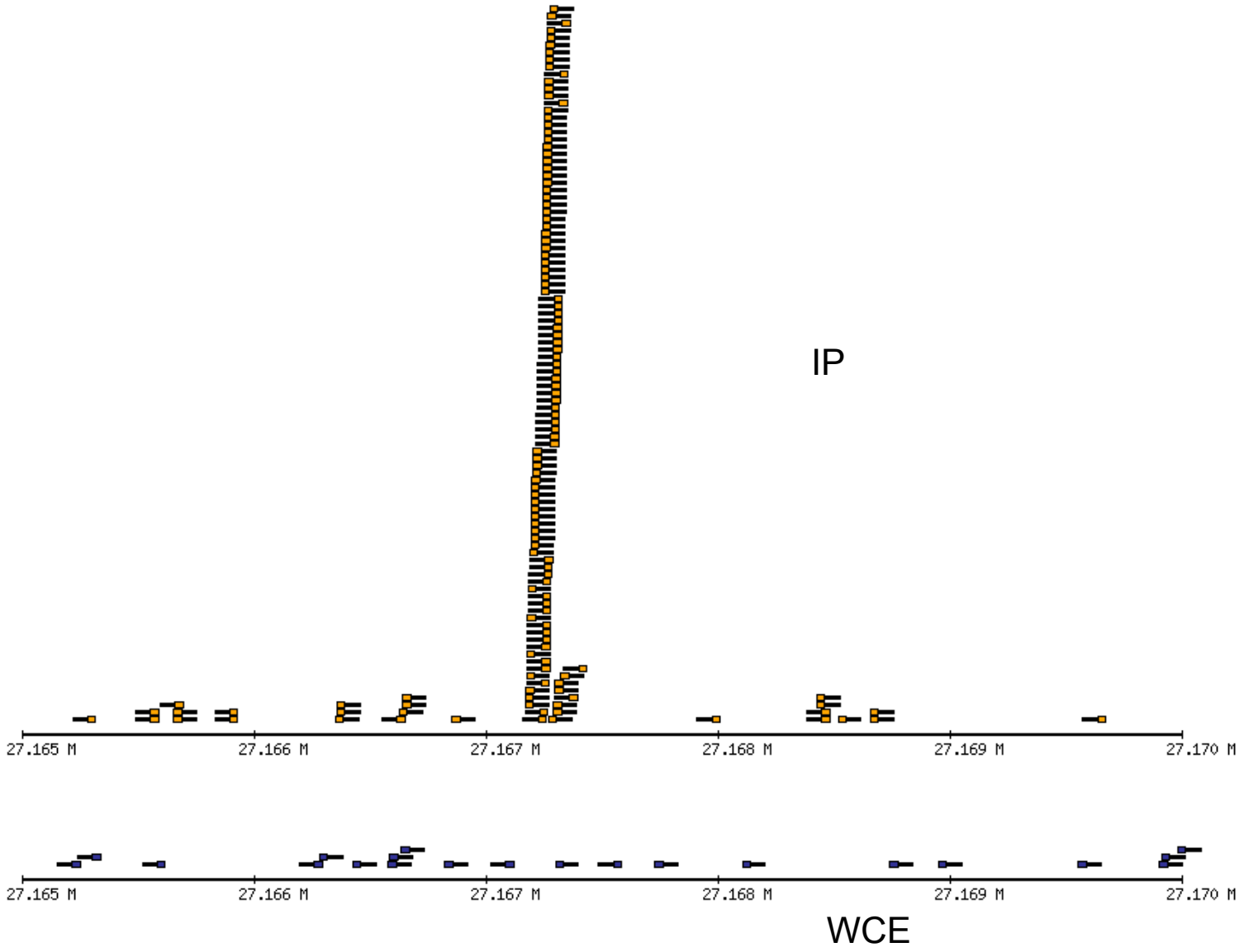


同じ開始点を

~ 1% の開始点に 5 リード以上落ちている

理論値を超えるのでPCRバイアスの可能性もある

結論としてなるべくライブラリを作成する際のPCRは抑える


IP
27.165 M
27.166 M
27.167 M
27.168 M
27.169 M
27.170 M
27.165 M
27.166 M
27.167 M
27.168 M
27.169 M
27.170 M
WCE

ヒトの場合（もっと深刻？）

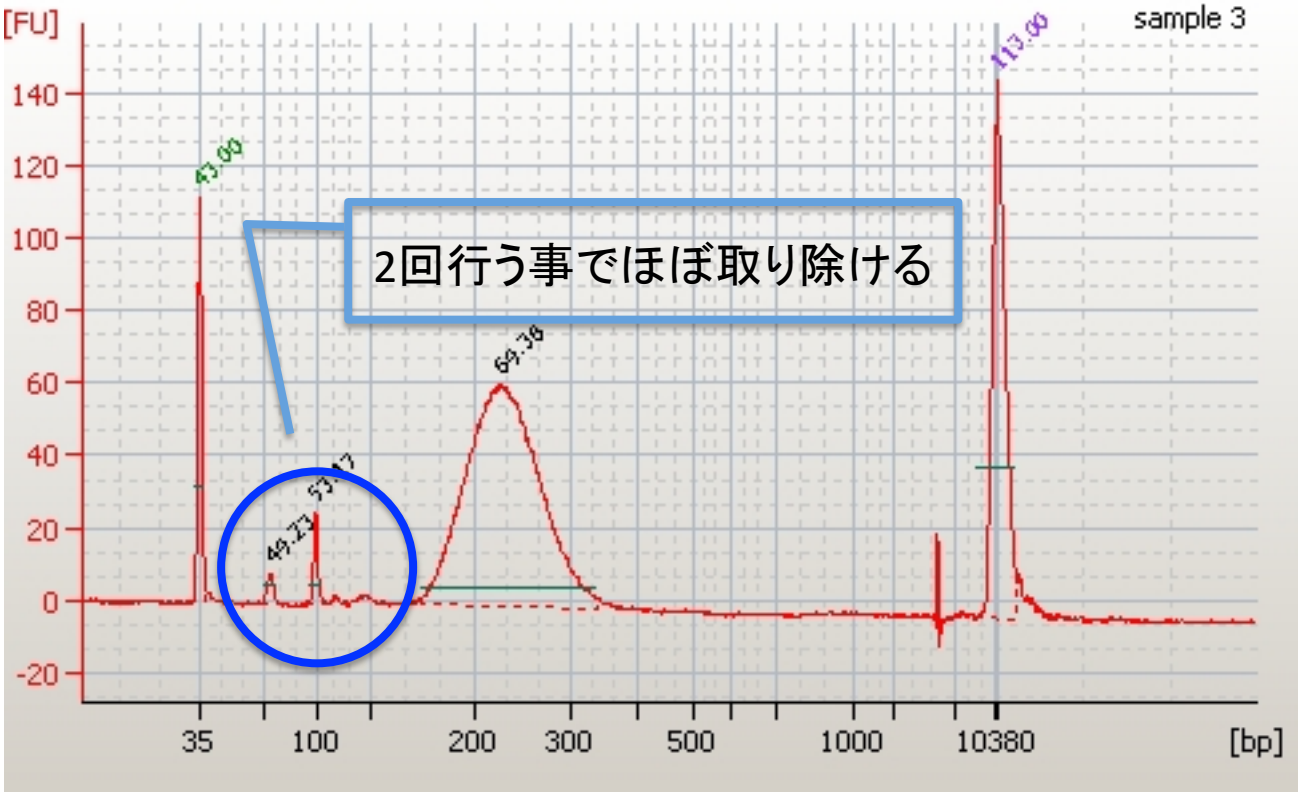
ライブラリ作成のTips

INPUTなどで高濃度の場合は、QubitもしくはNanodropで測定し
50ng-100ng使用

AMpureを使用して精製、またサイズセレクトを行っている
詳しくはBiolabs NEBNext DNA Library Prep Master Mix Set for illumina
(初期量が1μg以上ある場合のプロトコール)の**version3.0**のプロトコールを参照
#酵素量は、ChIP-seqプロトコール通り使用

ライブラリ作成のTips

各箇所でのAMpureでの精製は2回ずつ行う
特にenrich後は2回行った方が、primer dimerを取り除ける



ライブラリ作成のTips

ChIP-seqライブラリー作成時にCovarisを使用する必要があるか

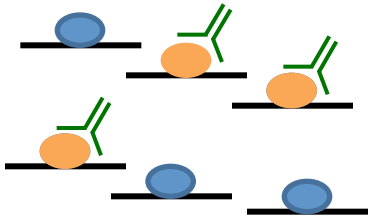
した方がいいと思います。
理由は、covarisにより更に、断片化できて、シークエンスに回せるDNA量が確保できるからです。
ヒト等高等真核生物の場合は繰り返し配列によるPCRバイアスの問題等が必ずついて回るので、出来るだけ均一に短くした方がいいと思います。
またDNAの長さによっては、AMpureではなく、切り出しの方がいい場合もあります。

解像度が下がるかもと考えるかもしれませんが、読めなきゃ意味ないですからね。

解析編

ChIP-seq法の流れ

ChIP（クロマチン免疫沈降）

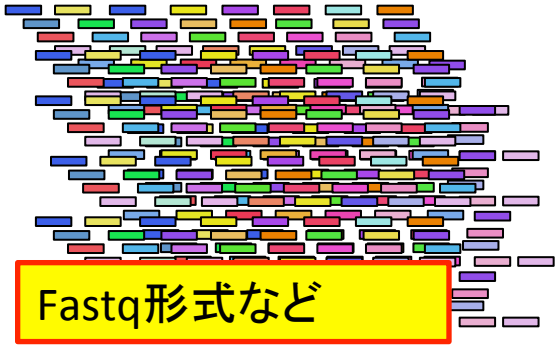


ゲノム断片（150bp～300bp）

Sequencing

数億個の断片配列 （リード）
（36bp～100bp）

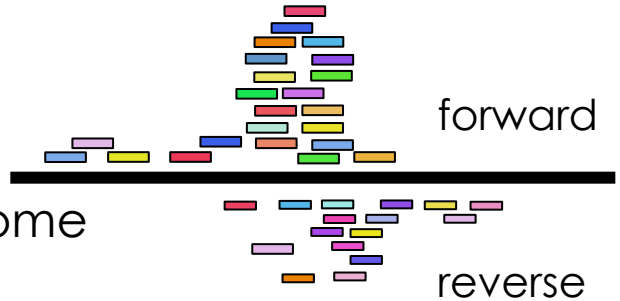
ヒト：5千万　酵母：1500万



Fastq形式など

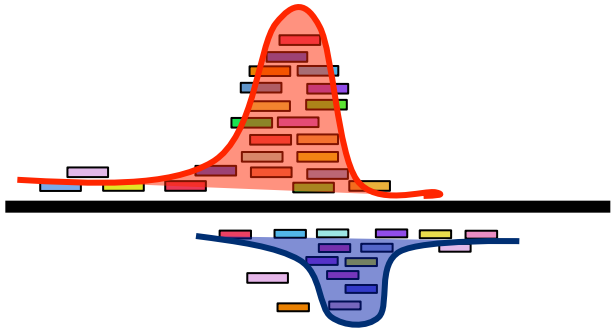
フラグメント、ペアーエンド

Genome mapping



bam形式など

Peak-calling



bed形式など

解析に必要なもの

- 解析用PC
  - データの量が多くなければ、それほど高級なPCは必要ない
- バックアップ用HDD（外付けHDDなど）
  - ある日突然PCが壊れてデータが全て消えた、とならないように、、、
  - 500GB～2TB程度

OS (Operating System)

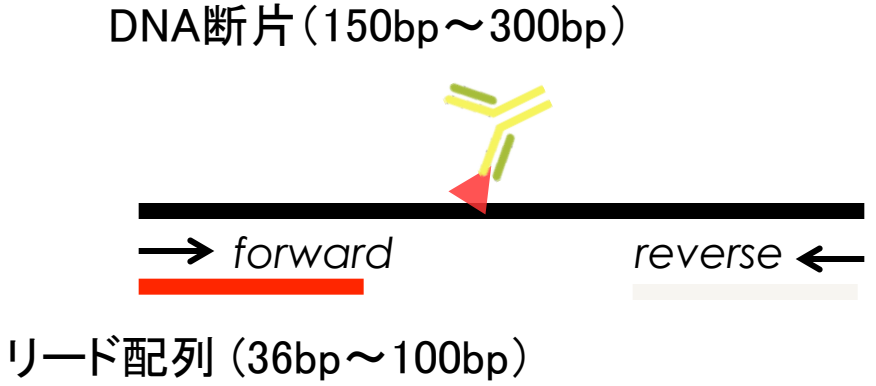
- Linuxが最も自由度が高い
  - Linux用の無償プログラムが多く開発されている
  - ファイル形式の変換も容易
  - 解析に必要な時間が短め
  - エラーのログが明確
- Macでもターミナルを使えば可能
- Windowsの場合
  - Cygwinを使う
  - 有料の統合解析用ソフトを使う

プログラミング技術は必要か

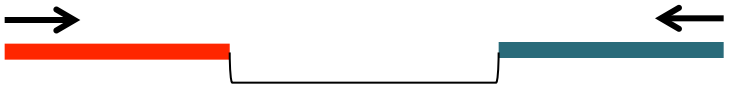
- Perl, pythonなど：データの整形や統合など
  - （例：遺伝子のIDとアノテーション）
- R：統計解析、グラフ描画に強い
  - OSに依存せずプログラムを使える
- 必要に迫られてからでも良い
  - エクセルだけでも何とかなったりする
  - 誰かにお願いする
- 最も必要なのはネット検索スキル
  - Linuxソフトウェアのインストール、実行
  - エラー解決能力

シークエンシング

- Single-end
  - 片端を読む
  - 1断片あたり1リード



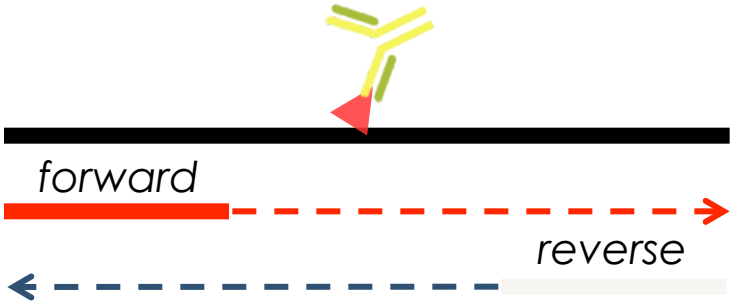
- Paired-end
  - 両端を読む
  - 1断片あたり2リード



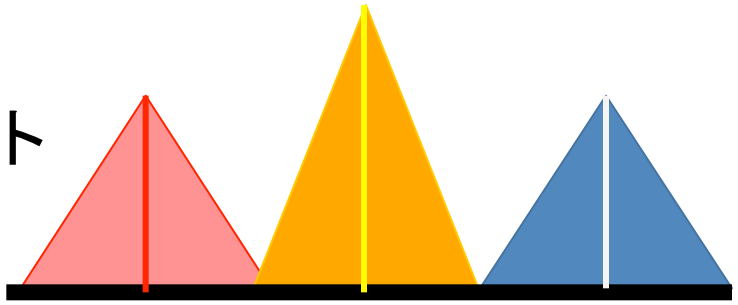
- 断片配列の端を読むので、実際のタンパク結合部位とずれが生じる
- どれくらいのリードを読むか？多ければ多いほどよい。

マップ領域の補正

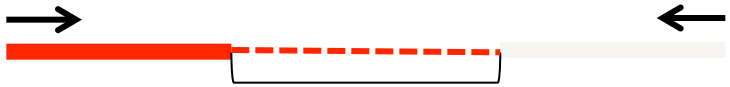
- Single-end
  1. 推定断片長まで伸長



  2. Forwardとreverseを別にカウント

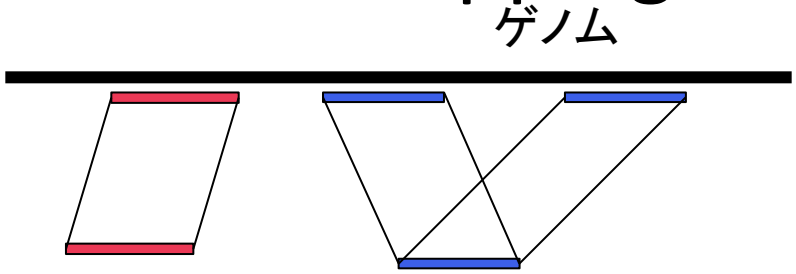


- Paired-end
  –両側がゲノムにマップされた
  リードのみを利用する



Paired-endは繰り返し配列を丹念に解析したい時等、特殊な用途でほとんどはSingleでいい

Genome mapping



ゲノムのただ1箇所に張り付く
unique read

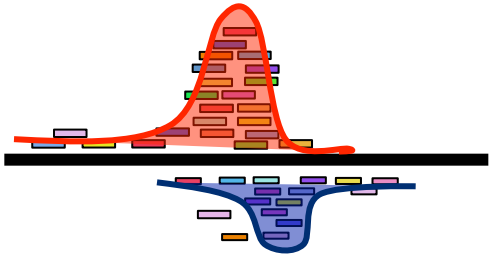
ゲノムの複数箇所に張り付く
multi read

- ツール： Eland, BWA, Bowtie, maq, …
  - 読み取りエラー, SNPなどに起因するミスマッチを考慮
  - Indelは通常考慮しない
- ほとんどの論文はunique readのみを考慮している
  - 多くのケースはこれで十分
- 反復配列領域などを調べたい場合はmulti readも考慮すべき

酵母の場合、ヒトの場合

- 酵母 (ゲノム長16Mbp)（標準～１５００万リード）
  - リードの深さ（depth）は十分 （～60）
  - セントロメア領域も可
  - マップ率: ～90%, ゲノムカバー率: ～99%
- ヒト (ゲノム長3Gbp)（標準～５千万リード）
  - リードの深さ（depth）は十分でない （～1.0）
  - セントロメア領域など、未解読の領域が存在する
  - ゲノム配列の約半分が反復配列領域
  - マップ率: ～70%, ゲノムカバー率: ～70%
  - でも、一度は10億リードぐらい読んでみたい。。。

Peak-calling

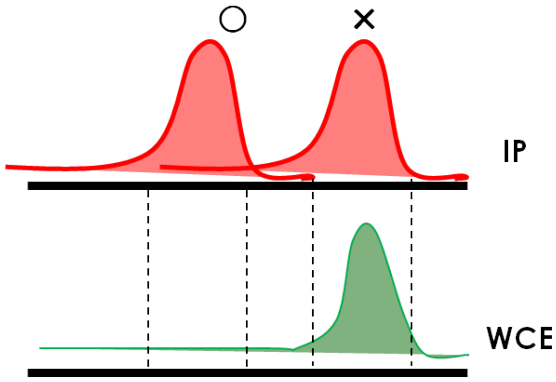


- マッピングファイルから、ChIP (IP) リードが濃縮している場所（山）を「ピーク」として抽出
  - E.x. [Johnson *et al.*, Science, 2007]

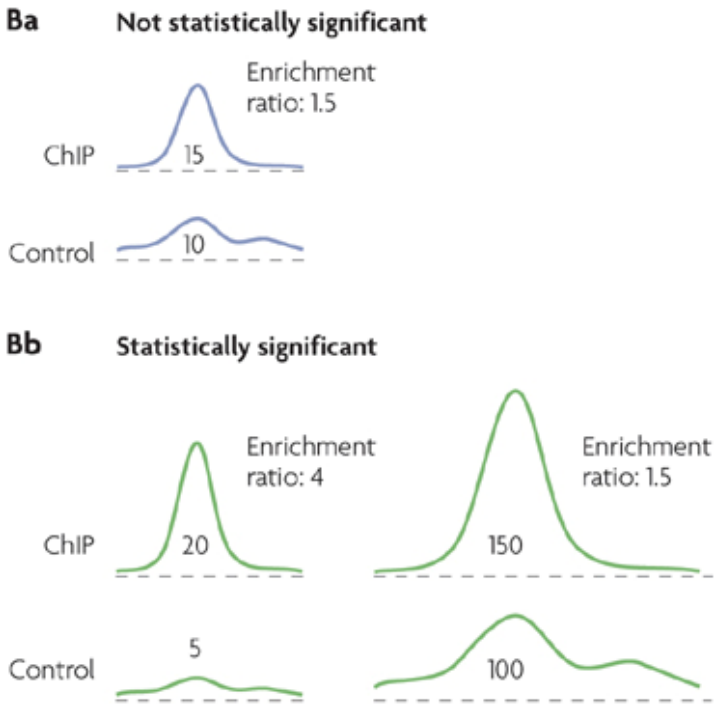
【問題点】

- ChIPから得られたリード集合にも必ず一定量のバックグラウンドが含まれる
  - ほとんどバックグラウンドであることもしばしば
- 結合領域でない偽陽性のピークが多く検出される
  - 反復配列, PCR bias, Open chromatinなど

Peak-calling



- ChIP (IP)と Control (WCE)を比較
  - IP/WCEで有意に濃縮している領域をピークとして抽出
  - 酵母だとこれでも十分
- ヒトの場合、深さが足りないこともあり、偽陽性のピークが生まれやすい
- 更なる統計処理が要る
- 既存のピークコール手法は主にヒト対象

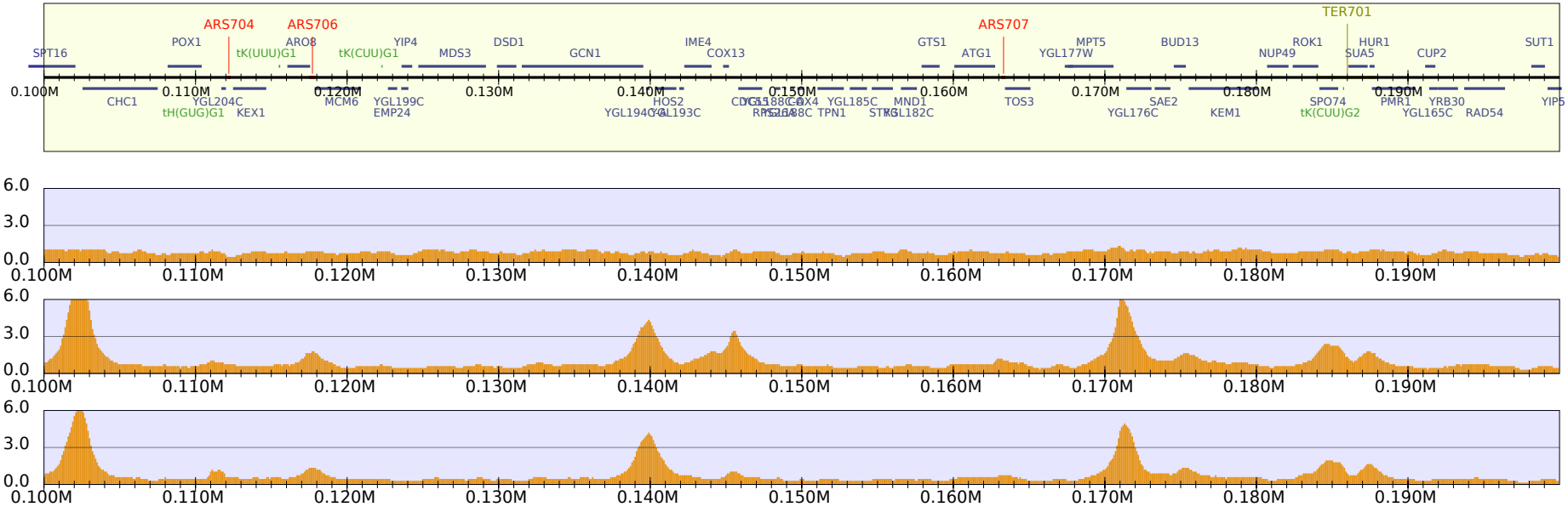


[Park, *Nature Review Genetics,* 2009]

可視化の例(出芽酵母)

[Tanaka et al., Current Biology, 2011]

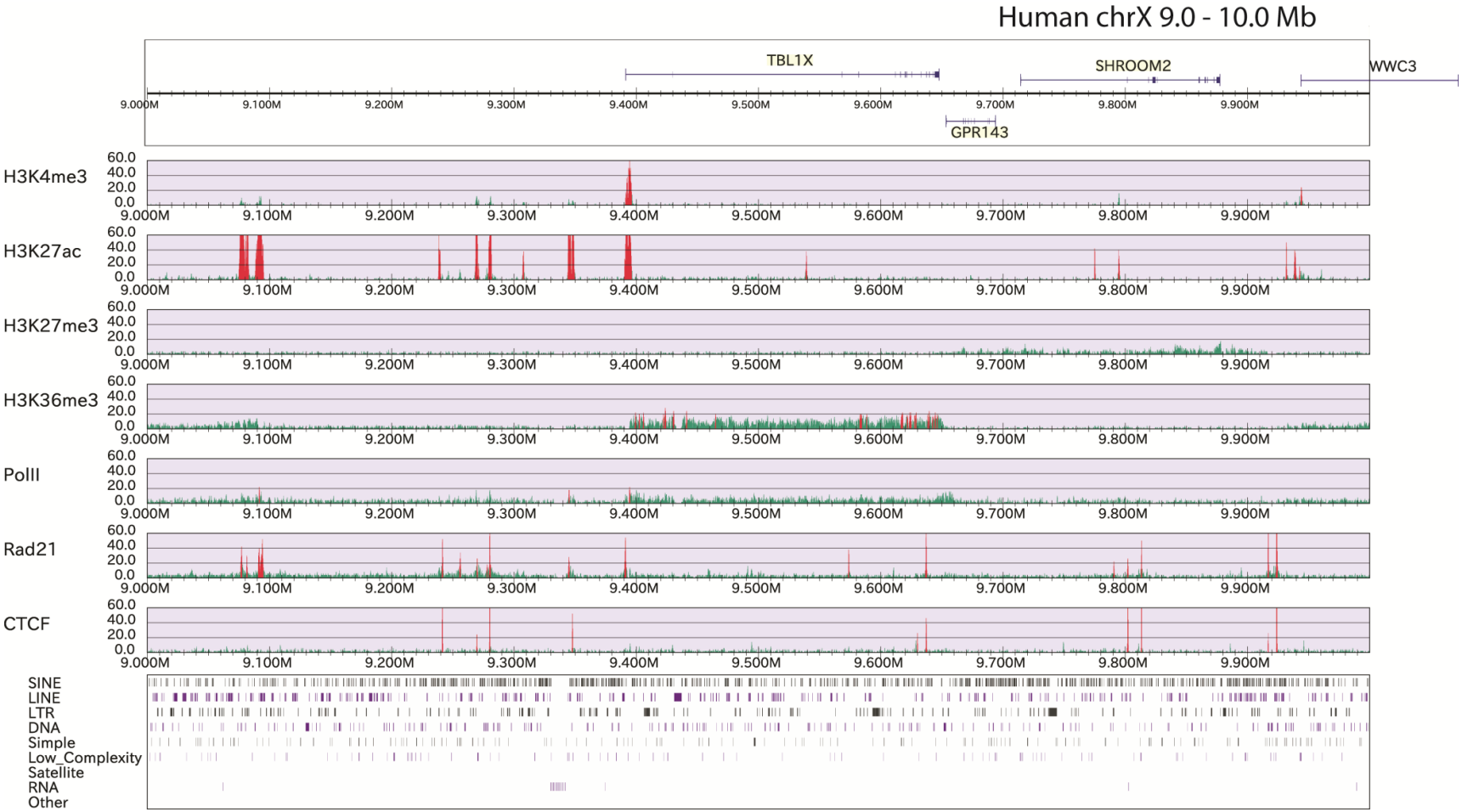
酵母の場合はX60ぐらい読むことになるのでそもそもピークコールと言う発想はいらないかも



酵母の場合はX60ぐらい読むことになるのでそもそもピークコールと言う発想はいらないかも

そのままプロファイルとして理解すべき

可視化の例(ヒト)



タンパクの結合モードに応じた結合領域の定義が必要

既存のピークコール手法

[Wilbanks and Facciotti, PLoS ONE, 2010]

Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

X* = Windows-only GUI or cross-platform command line interface

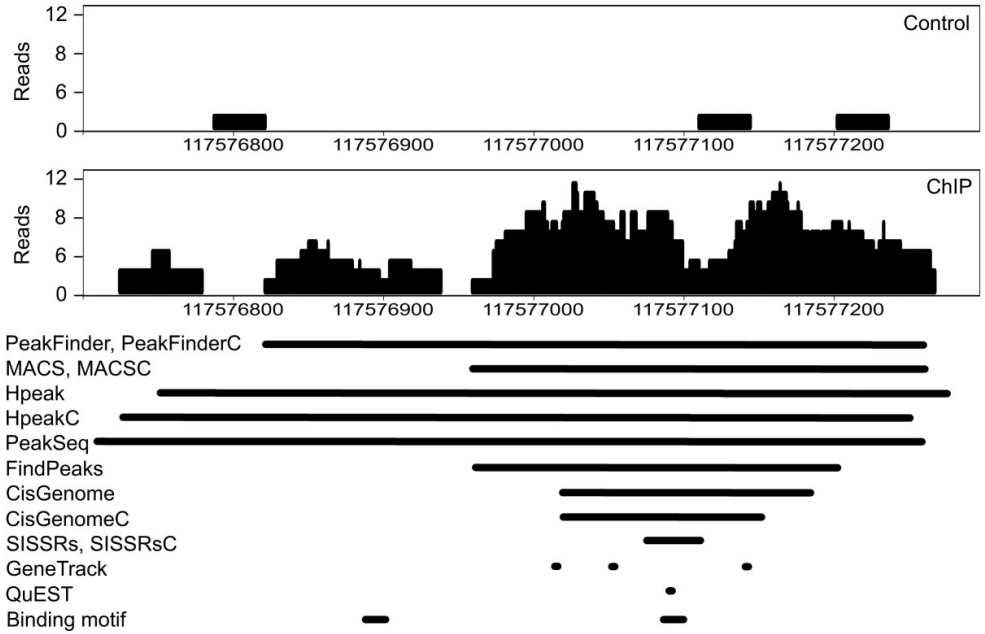
X** = optional if sufficient data is available to split control data

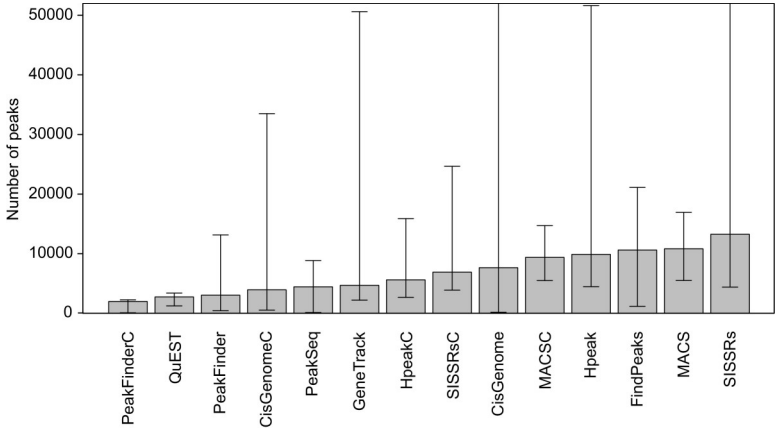
X' = method exludes putative duplicated regions, no treatment of deletions

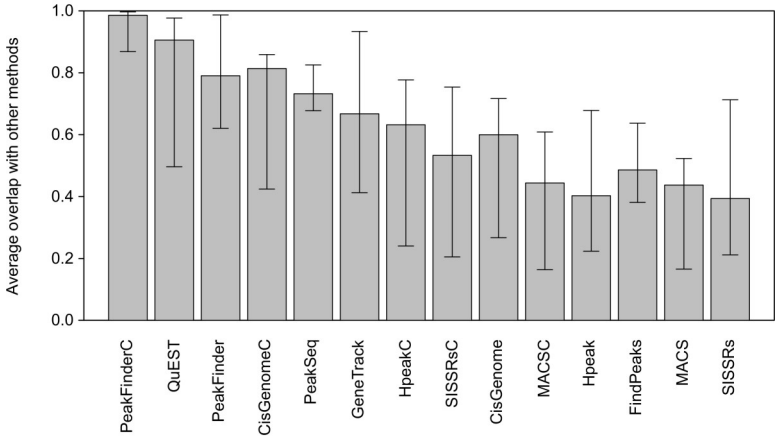
どの手法を用いるのが良いのか？

- 各プログラムで得られるピークはさまざま
- ピークの数や幅はプログラム次第

[Laajala et al. BMC Genomics, 2009]







どの手法を用いるのが良いのか？

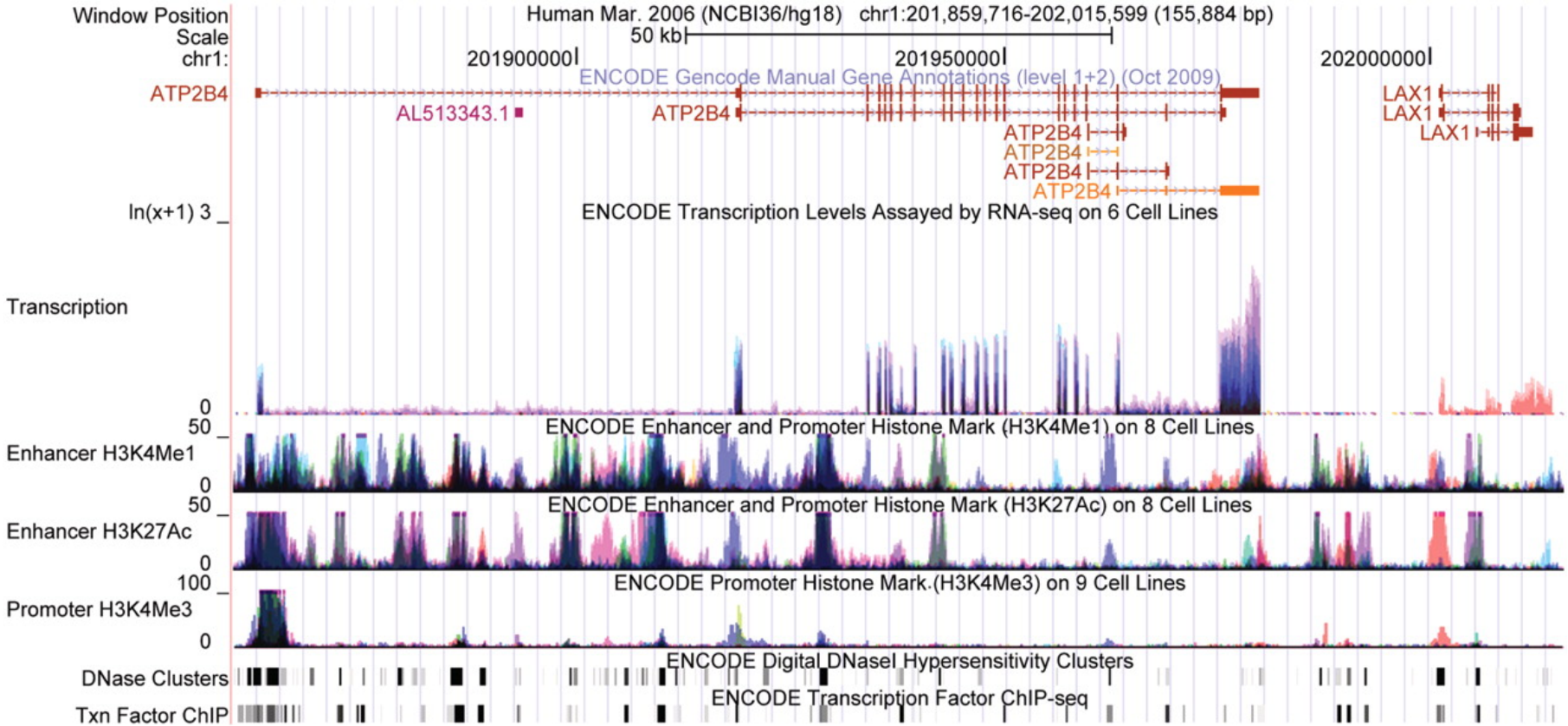
- 強くはっきりしたピークはどの手法でも得られる
  - サンプルのS/N比が良ければ、どの手法でも大丈夫
  - 小さなピークを感度良く検出することは難しい
- 最適なパラメータは生物種や目的に依存する
  - 感度＞特異度 （とにかく候補を見つけたい）
  - 感度＜特異度 （信頼性を高い候補だけを見たい）
- 入力・出力フォーマットもさまざま
- 結局、結果とにらめっこしながらパラメータを変えつつ試行錯誤するしかない。
- ただ、あまりお気楽にほいほい変えることが出来るものでもない。試行錯誤には時間がかかる。

ピークコール以降の解析

- ピークがどのように分布しているか可視化したい。
- 得られたピークが近傍に現れる遺伝子の機能を知りたい。
- ピークが表れる領域に特異的な配列（モチーフ）は存在するか？
- 他のどのような転写因子と共結合しているか？

UCSC genome browser

- 自分のデータをアップロードして既知アノテーションと比較可能



[Raney et al. Nucl. Acids Res. 2011]

NCBI Reference Sequence (RefSeq)との比較

- http://www.ncbi.nlm.nih.gov/RefSeq/

	A	B	C	
1	<binding site> all: 1093, genic: 899, promoter: 621, downstream: 39, intergenic: 296			
2	<gene>			
3	name	chromosome	type	description
4	C5orf34	5	protein coding	Uncharacterized protein C5orf34 [Source:UniProtKB/Swiss-Prot;Acc:Q96MH7]
5	MRPL13	8	protein coding	mitochondrial ribosomal protein L13 [Source:HGNC Symbol;Acc:14278]
6	CBY1	22	protein coding	chibby homolog 1 (Drosophila) [Source:HGNC Symbol;Acc:1307]
7	MRPL54	19	protein coding	mitochondrial ribosomal protein L54 [Source:HGNC Symbol;Acc:16685]
8	GINS3	16	protein coding	GINS complex subunit 3 (Psf3 homolog) [Source:HGNC Symbol;Acc:25851]
9	SNRPA	19	protein coding	small nuclear ribonucleoprotein polypeptide A [Source:HGNC Symbol;Acc:11151]
10	AIDA	1	protein coding	axin interactor, dorsalization associated [Source:HGNC Symbol;Acc:25761]
11	NEDD8	14	protein coding	neural precursor cell expressed, developmentally down-regulated 8 [Source:HGNC Symbol;Acc:7732]
12	CCT4	2	protein coding	chaperonin containing TCP1, subunit 4 (delta) [Source:HGNC Symbol;Acc:1617]
13	CCNB1IP1	14	protein coding	cyclin B1 interacting protein 1 [Source:HGNC Symbol;Acc:19437]
14	KCNH6	17	protein coding	potassium voltage-gated channel, subfamily H (eag-related), member 6 [Source:HGNC Symbol;Acc:18862]
15	ERLIN1	10	protein coding	ER lipid raft associated 1 [Source:HGNC Symbol;Acc:16947]
16	RPRD2	1	protein coding	regulation of nuclear pre-mRNA domain containing 2 [Source:HGNC Symbol;Acc:29039]
17	ZNF546	19	protein coding	zinc finger protein 546 [Source:HGNC Symbol;Acc:28671]
18	RPL7L1	6	protein coding	ribosomal protein L7-like 1 [Source:HGNC Symbol;Acc:21370]
19	GGPS1	1	protein coding	geranylgeranyl diphosphate synthase 1 [Source:HGNC Symbol;Acc:4249]
20	NR3C1	5	protein coding	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor) [Source:HGNC Symbol;Acc:7978]
21	PIK3C3	18	protein coding	phosphoinositide-3-kinase, class 3 [Source:HGNC Symbol;Acc:8974]
22	CCT5	5	protein coding	chaperonin containing TCP1, subunit 5 (epsilon) [Source:HGNC Symbol;Acc:1618]
23	NQO2	6	protein coding	NAD(P)H dehydrogenase, quinone 2 [Source:HGNC Symbol;Acc:7856]
24	SEC24C	10	protein coding	SEC24 family, member C (S. cerevisiae) [Source:HGNC Symbol;Acc:10705]

GREAT (http://great.stanford.edu/public/html/)

- ピーク周辺の遺伝子の機能で有意に表れるものを抽出

GREAT predicts functions of cis-regulatory regions.

Many coding genes are well annotated with their biological functions. Non-coding regions typically lack such annotation. GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. Thus, it is particularly useful in studying cis functions of sets of non-coding genomic regions. Cis-regulatory regions can be identified via both experimental methods (e.g. ChIP-seq) and by computational methods (e.g. comparative genomics). For more see our Nature Biotech Paper.

News

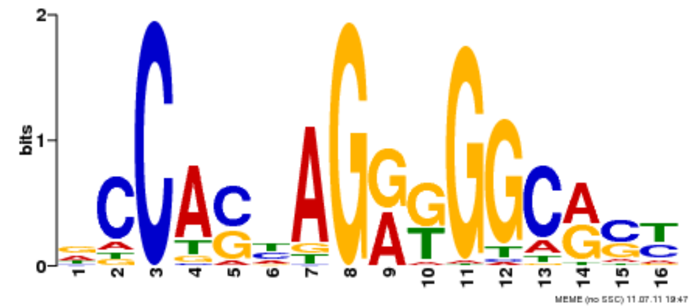
- Apr 3, 2012: GREAT version 2.0 adds new annotations to human and mouse ontologies and visualization tools for data exploration.
- Feb 18, 2012: The GREAT forums are released, allowing increased user-to-user interaction

More news items...

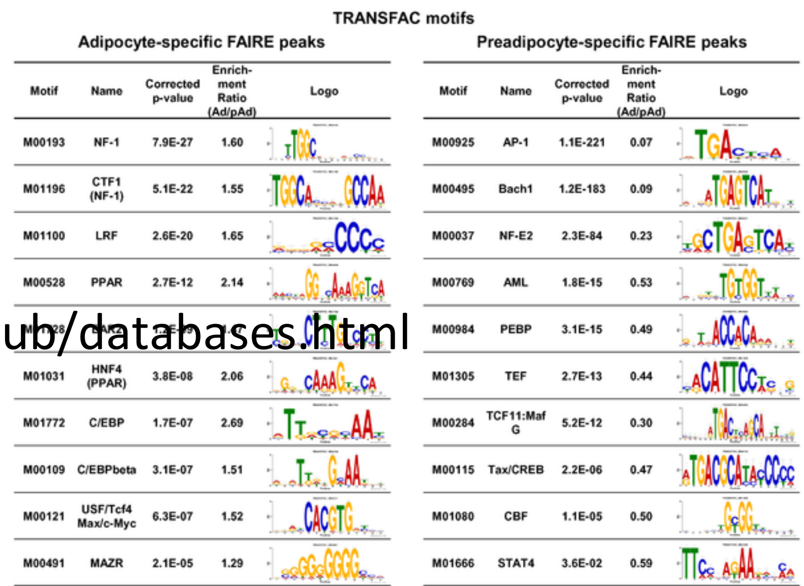
モチーフを使った解析

(例: CTCF motif)

- モチーフ抽出
  - MEMEなど
  - http://meme.sdsc.edu/meme/



- モチーフデータベース
  - Transfac (有料)
  - http://www.gene-regulation.com/pub/databases.html



TRANSFAC motifs

Adipocyte-specific FAIRE peaks

Motif	Name	Corrected p-value	Enrichment Ratio (Ad/pAd)	Logo
M00193	NF-1	7.9E-27	1.60	
M01196	CTF1 (NF-1)	5.1E-22	1.55	
M01100	LRF	2.6E-20	1.65	
M00528	PPAR	2.7E-12	2.14	
[illegible]	[illegible]	[illegible]	[illegible]	
M01031	HNF4 (PPAR)	3.8E-08	2.06	
M01772	C/EBP	1.7E-07	2.69	
M00109	C/EBPbeta	3.1E-07	1.51	
M00121	USF/Tcf4 Max/c-Myc	6.3E-07	1.52	
M00491	MAZR	2.1E-05	1.29	

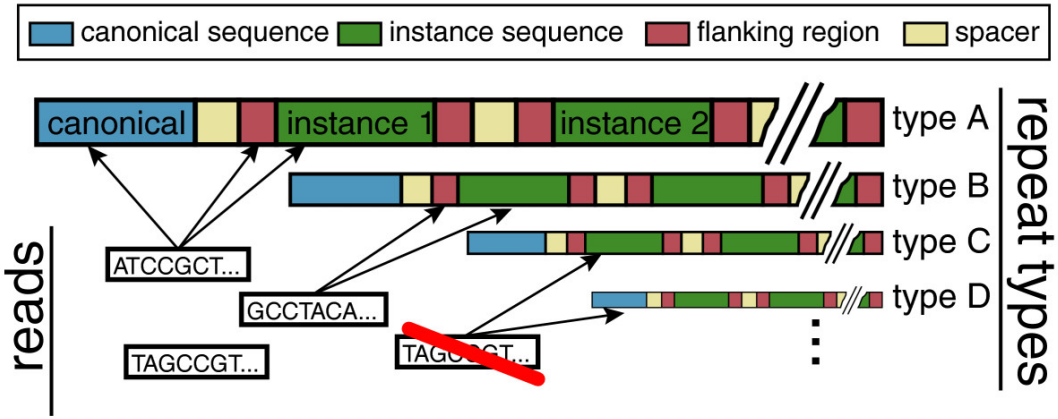
Preadipocyte-specific FAIRE peaks

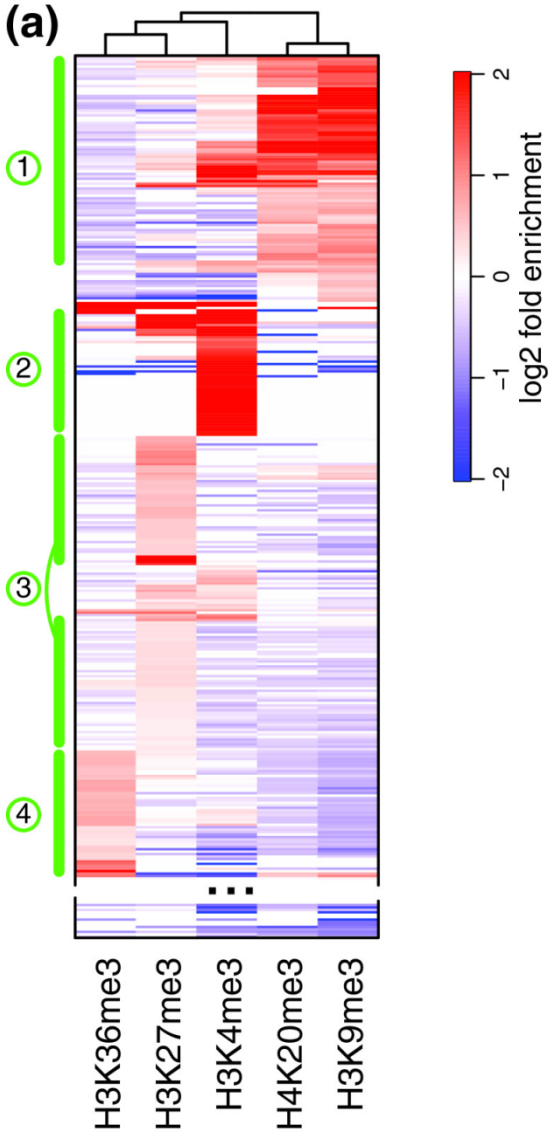
Motif	Name	Corrected p-value	Enrichment Ratio (Ad/pAd)	Logo
M00925	AP-1	1.1E-221	0.07	
M00495	Bach1	1.2E-183	0.09	
M00037	NF-E2	2.3E-84	0.23	
M00769	AML	1.8E-15	0.53	
M00984	PEBP	3.1E-15	0.49	
M01305	TEF	2.7E-13	0.44	
M00284	TCF11:MafG	5.2E-12	0.30	
M00115	Tax/CREB	2.2E-06	0.47	
M01080	CBF	1.1E-05	0.50	
M01666	STAT4	3.6E-02	0.59	

[Waki et al., PLoS Genetics, 2011]

反復配列領域の解析

[Day et al., Genome Biology, 2010]





- multi readsを用いる
- 既知のリピート領域にマップされるリードを統計的に解析

課題 (1)

- 得られたピークの質（パラメータの正しさ）をどうやって評価するか
  - モチーフがあるタンパクの場合、見つかったピーク中にモチーフがあるかどうか
  - 転写因子ならば、遺伝子周辺に結合しているか
  - 共結合するタンパク（同じタンパク複合体のサブユニット同士など）が共結合しているか
  - qPCRはほとんどの場合同じになります。
- ポジティブコントロールがあるのが理想

課題 (2)

- 精度は元となるサンプルの質に強く依存する
  - IPとWCEのリード数が2倍以上違うとうまく正規化できなくなる
  - 元々のDNA量も問題
  - 抗体の質

ChIP-seqデータのデータベース

- Gene Expression Omnibus (GEO)
  - http://www.ncbi.nlm.nih.gov/geo/
  - マッピングデータ、wigデータ、ピークリストなどを格納
- The Sequence Read Archive (SRA)
  - http://www.ncbi.nlm.nih.gov/Traces/sra/
  - 生リードデータを格納