Freie Universität Berlin

MAX-PLANCK-GESELLSCHAFT

AlgoBio WS 16/17
Protein-DNA Interaktionen
ChiP-Seq Datenanalyse

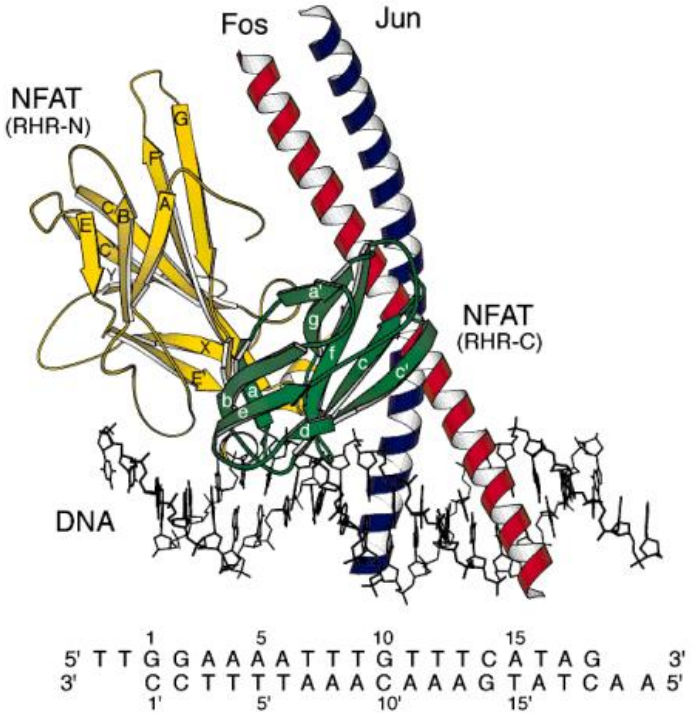
Annalisa Marsico

6.02.2017

Protein-DNA Interaktionen

- Häufig binden sich Proteine an DNA, um ihre biologische Funktion zu regulieren. Transkriptionsfaktoren (TF) beeinflussen die Expression eines Gens.

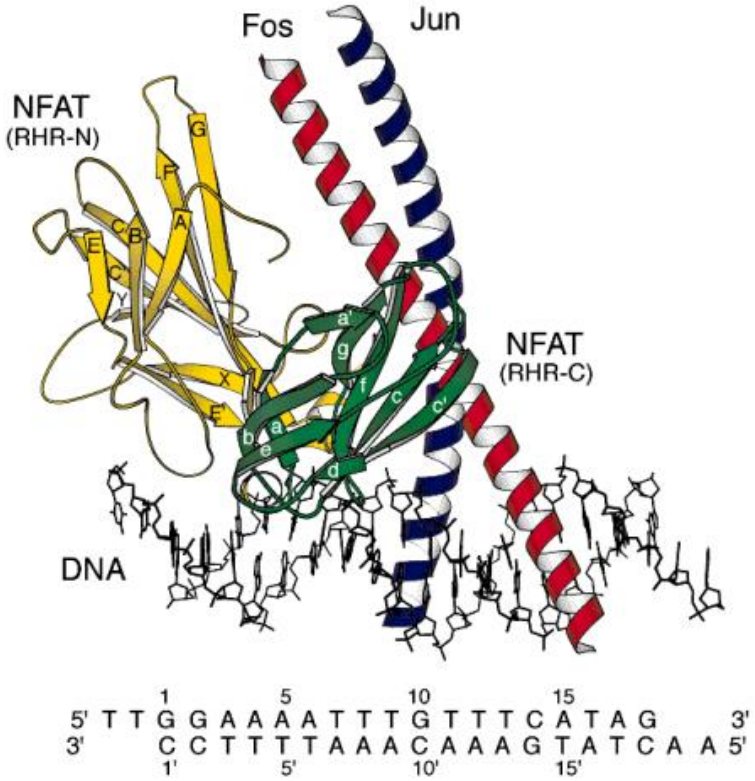
- ChIP-Seq (ChIP + Sequenzierung) ist eine Technik, um Bindungsstellen von TFs auf der DNA zu detektieren (aber nicht nur!).



NFAT-AP1-DNA complex
[Harrison, Nature 1998]

Protein-DNA Interaktionen

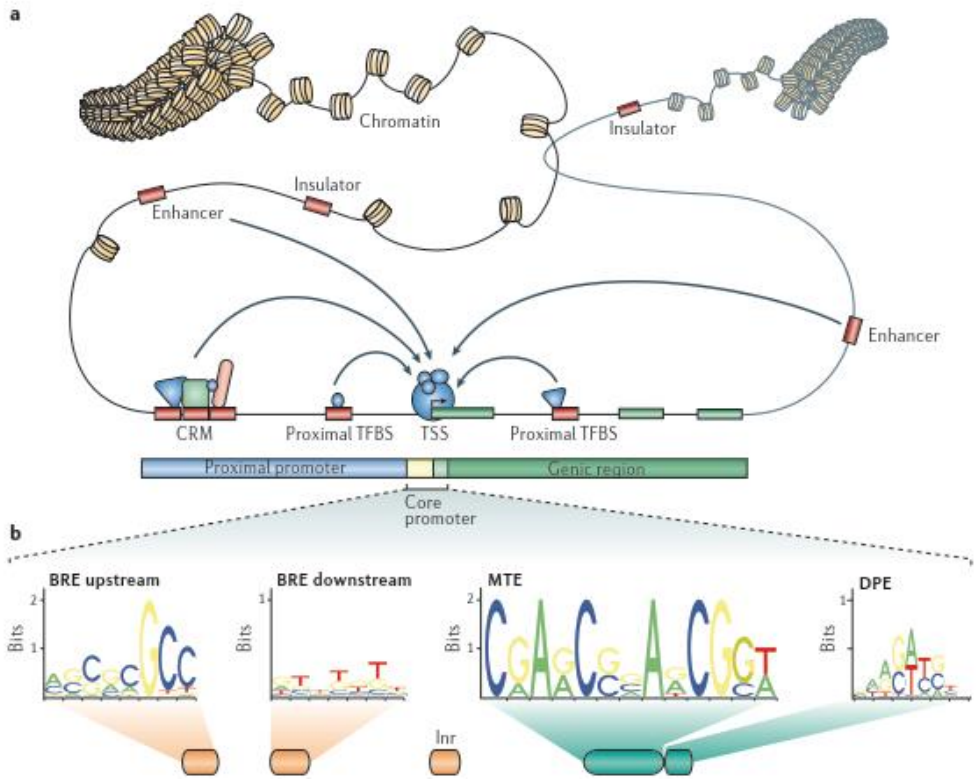
- DNA-bindende Proteine haben eine DNA-bindende Domäne, die eine "Affinität" für einzelne oder doppelsträngige DNA und für ein bestimmtes "Motiv“ hat.



NFAT-AP1-DNA complex
[Harrison, Nature 1998]

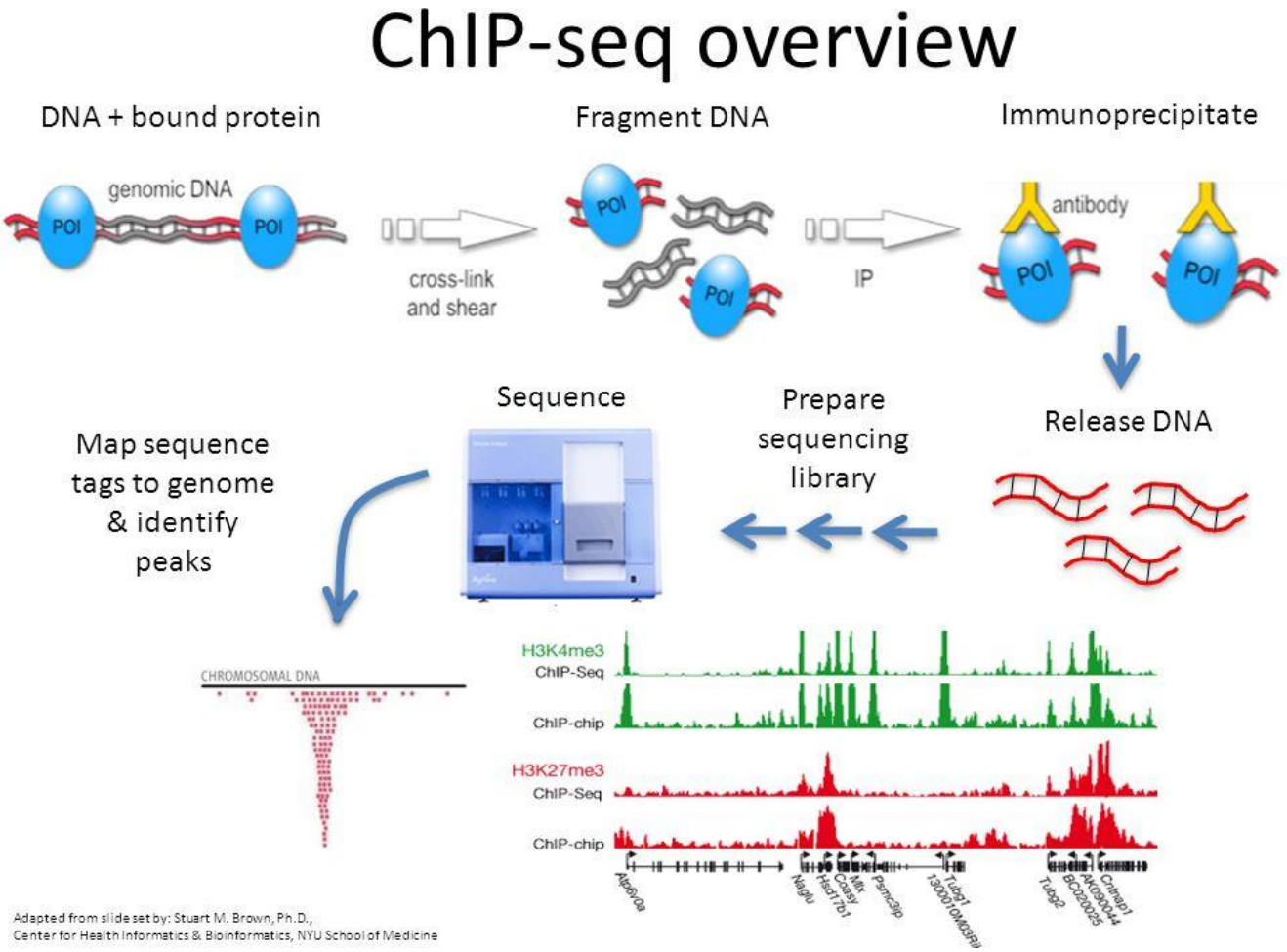
Chromatin Umwelt

- Nicht nur ein DNA-Motiv, sondern auch der Chromatin-Zustand beeinflusst die TF-Bindung und die Genregulation



Um Genregulation und genregulatorische Netzwerke zu verstehen, wollen wir alle Stellen im Genom kennen, an die Transkriptionsfaktoren unter verschiedenen Bedingungen binden.

Schritte eines ChIP-Seq-Experiments



Ziel: Anreichern für DNA-Fragmente, die an ein spezifisches Protein (TF) gebunden sind

Unterschied zu RNA-seq

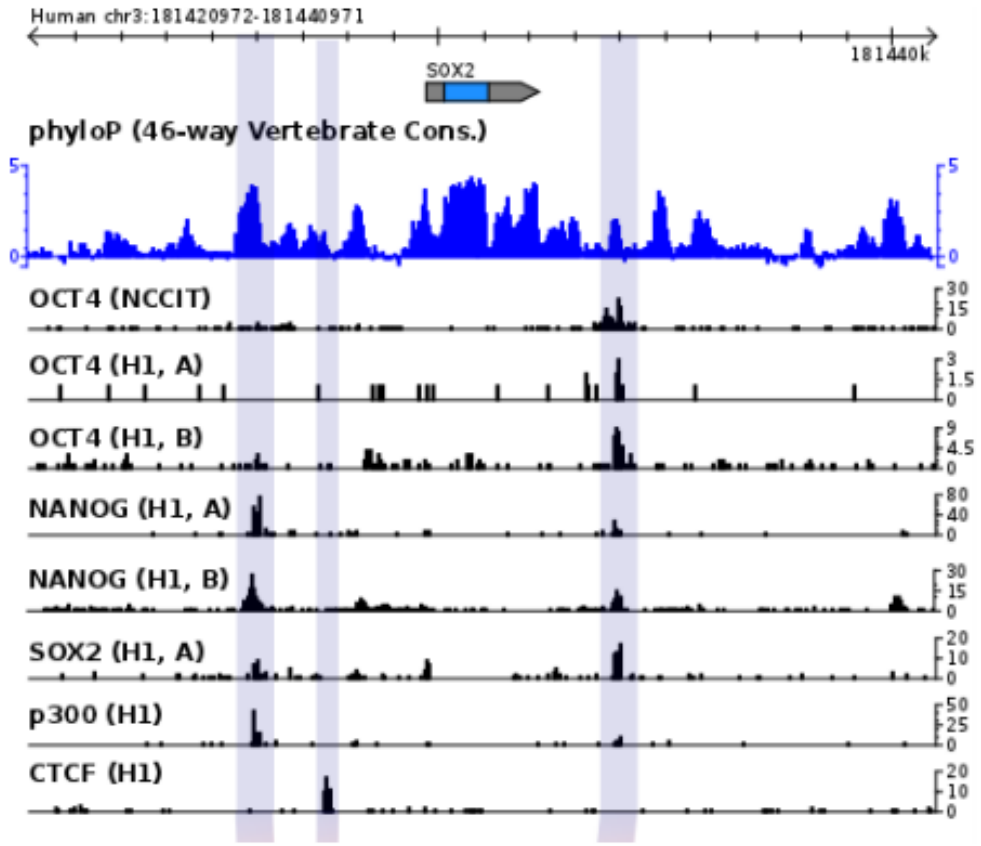
ChIP-Seq

- Geführt von einem ChIP
- Angereichert für DNA-Regionen, die an einen TF gebunden sind (Antikörper)
- **DNA Sequenzierung**
- Reads bedecken Bindungsstellen an den Gen-Promotoren
- Hauptziel: Genomweit Peaks finden (welche Gene werden durch eine bestimmten TF reguliert?)

RNA-Seq

- Kein ChIP
- Angereichert für eine bestimmte RNA-Klasse (z. B. PolyA)
- **RNA Sequenzierung**
- Reads bedecken RNA-Transkripte (z. B. verschiedene Isoformen)
- Hauptziel: Genexpression (wie viele Kopien eines Gens in einem bestimmten Gewebe?)

Dynamische Ansicht: Spezifische Messungen für einen Zelltyp



Schritte der ChiP-seq-Analyse

- Genom Alignment (z. B. Bowtie-Software, schneller Mapper)
  - Erlaubt eine kleine Anzahl von Fehlpaarungen aufgrund von Sequenzierungsfehlern, SNPs etc.

- **„Peak Calling“**
  - Identifizierung von Reads-angereicherten Regionen
  - Signifikanz der Peaks
  - Differentialpeaks?

- **Downstream Analyse**
  - Motivfindung in Peaks

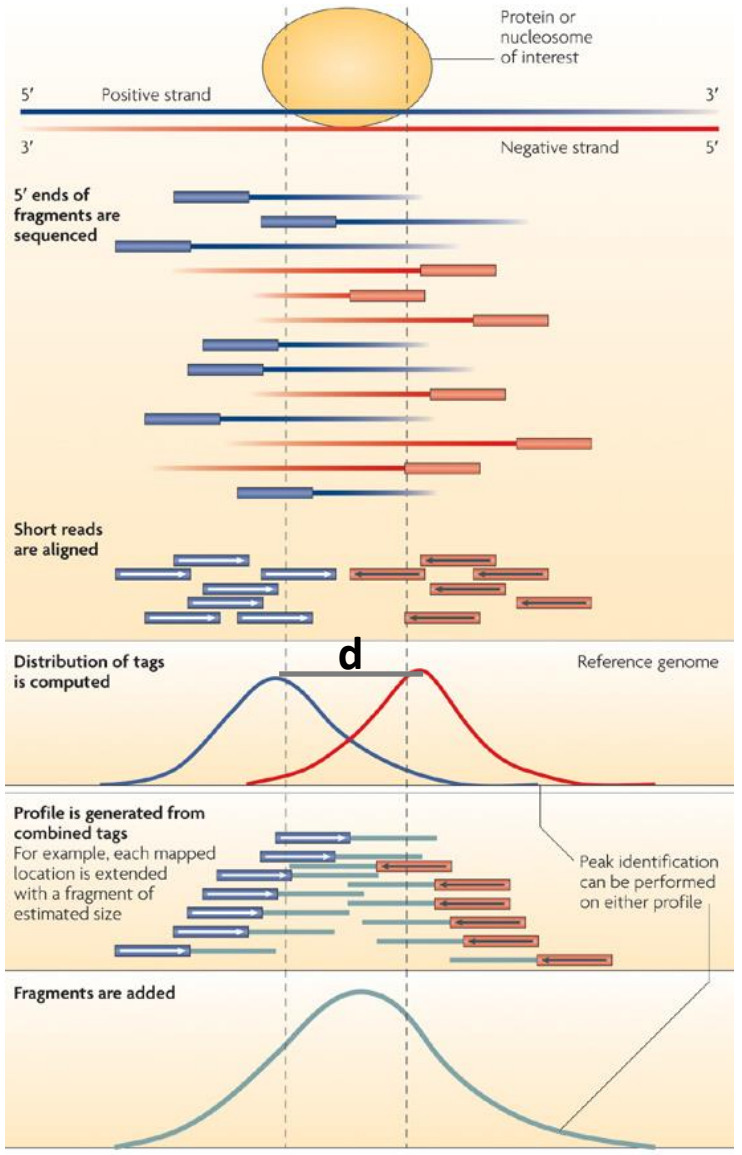
Peak Detektion

Scannen der genomischen DNA und suchen nach angereicherten Regionen mit einem Fenster-Ansatz

Strand-spezifische Muster können beobachtet und verwendet werden, um die Peaks zu lokalisieren

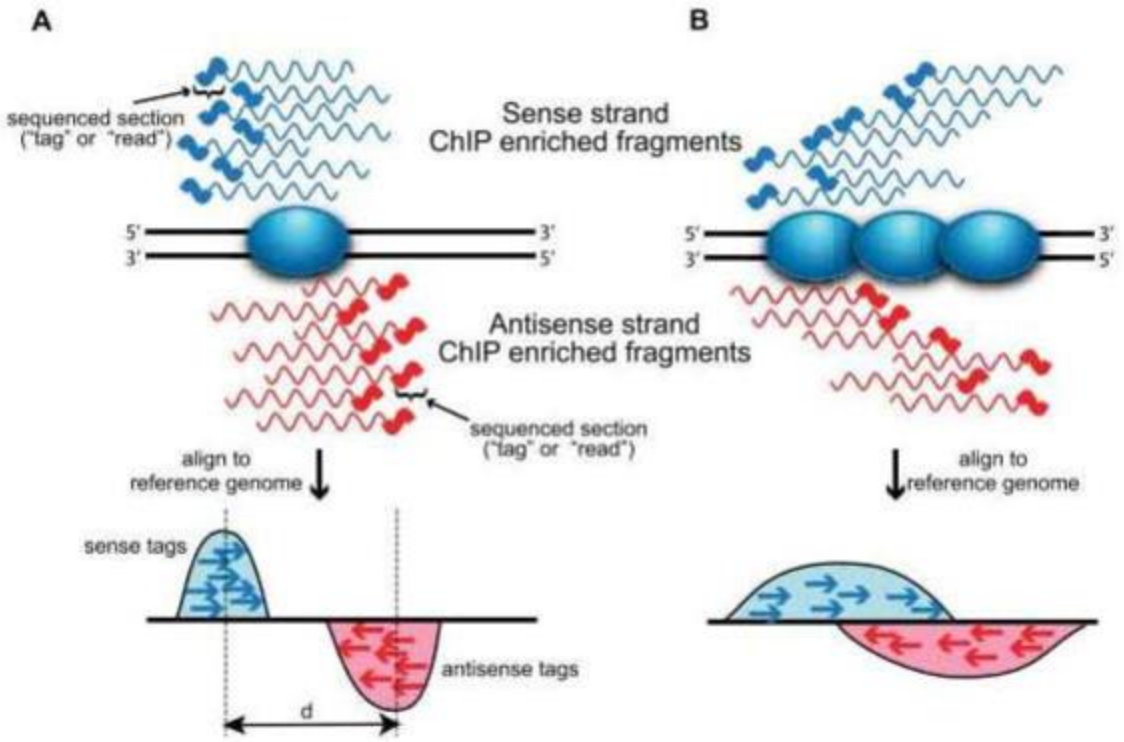
(1) Erweitern die Reads auf die geschätzte Fragmentlänge

(2) Verschieben Reads auf die Mitte der beiden Gipfel (d/2)



Nature Reviews | Genetics

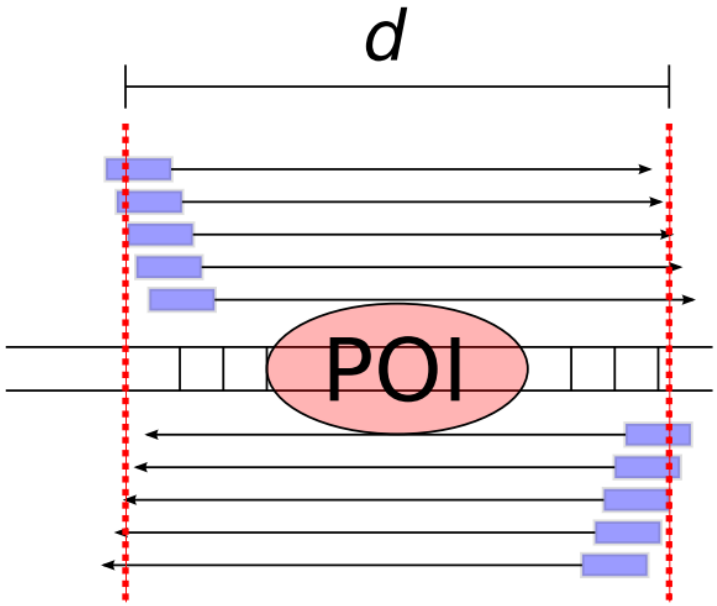
Peak Detektion



In anderen Ansätzen wird Bi-Modalität verwendet, bevor Peaks aufgerufen werden, um unwahrscheinliche Peaks zu filtern. Die Verteilung auf die beiden Stränge muss einander ähneln und der Abstand zwischen den Peaks muss in der Nähe der erwarteten Fragmentgröße liegen.

Warum „Bi-Modalität“?

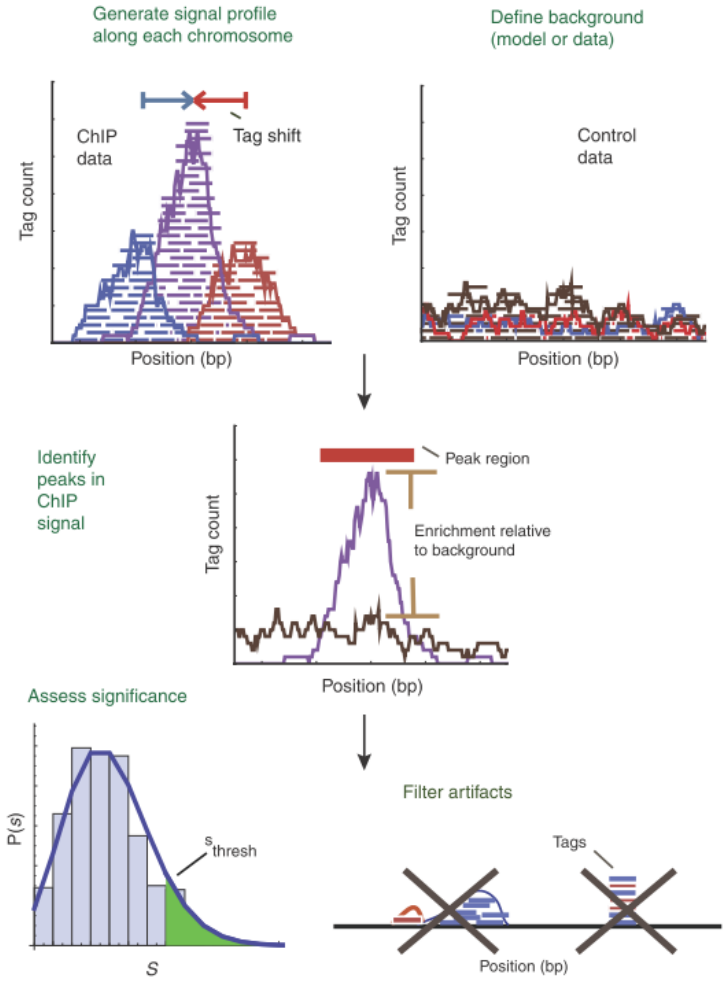
Warum entspricht die Trennung zwischen den Peaks (d) der durchschnittlichen sequenzierten Fragmentlänge?



Die blaue Box zeigt die Region des Fragments, das tatsächlich sequenziert wird (oft 36bp). Das gesamte Fragment ist länger, wobei die genaue Größe von dem experimentellen Fragmentierungsprotokoll abhängt. Normalerweise liegt das Protein (POI) in der Mitte des Fragments , so dass der durchschnittliche Abstand zwischen den Reads der durchschnittlichen Fragmentlänge entspricht.

Handhabung des Hintergrunds

Wie wird der Reads Abstand (oder read-shift) *d* bestimmt? Er ist entweder benutzerdefiniert oder abgeschätzt durch hochwertige Peaks, d.h. diejenigen mit sehr großer Anreicherung im Verhältnis zum Hintergrund.

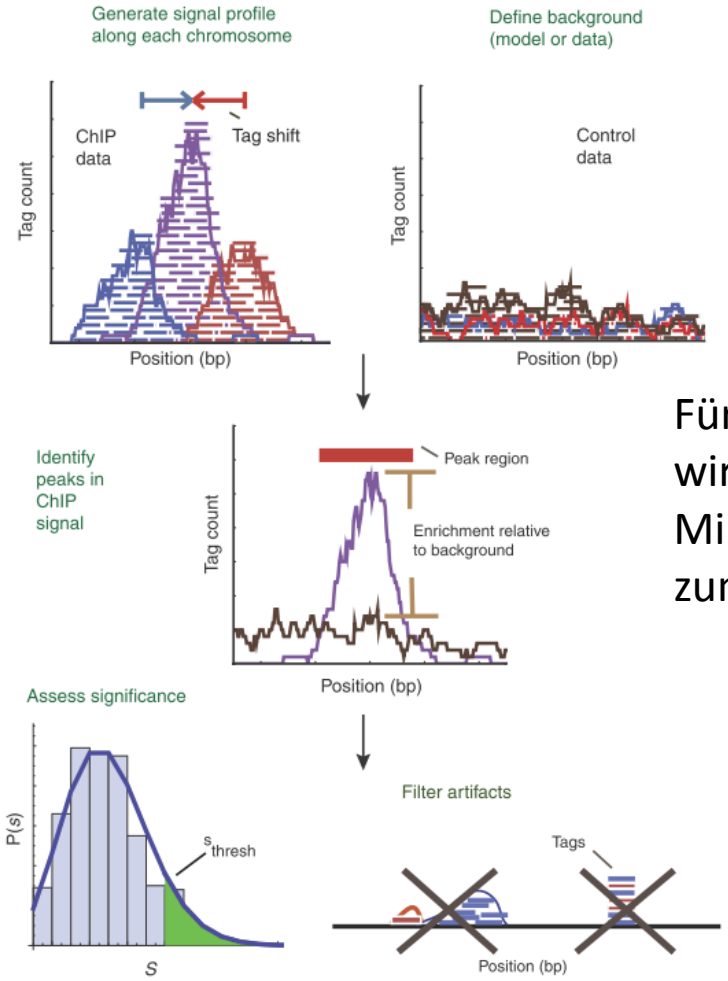


Was ist der Hintergrund?
'Noise' Verteilung von Reads.
Er ist mit einer Poisson-Verteilung modelliert.

Pepke et al., Nature Methods 2009

Handhabung des Hintergrunds

Wie wird der Reads Abstand (oder read-shift) *d* bestimmt? Er ist entweder benutzerdefiniert oder abgeschätzt durch hochwertige Peaks, d.h. diejenigen mit sehr großer Anreicherung im Verhältnis zum Hintergrund.

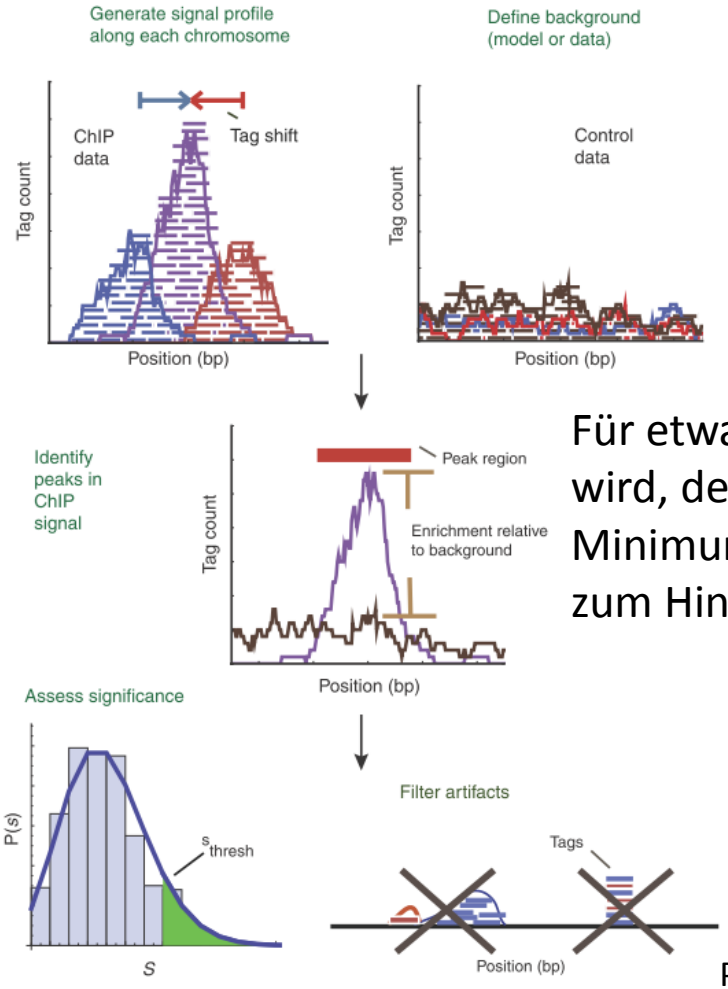


Für etwas, das als Peak bezeichnet wird, definiert man eine Minimumanreicherung relativ zum Hintergrund

Pepke et al., Nature Methods 2009

Handhabung des Hintergrunds

Wie wird der Reads Abstand (oder read-shift) *d* bestimmt? Er ist entweder benutzerdefiniert oder abgeschätzt durch hochwertige Peaks, d.h. diejenigen mit sehr großer Anreicherung im Verhältnis zum Hintergrund.



Für etwas, das als Peak bezeichnet wird, definiert man eine Minimumanreicherung relativ zum Hintergrund

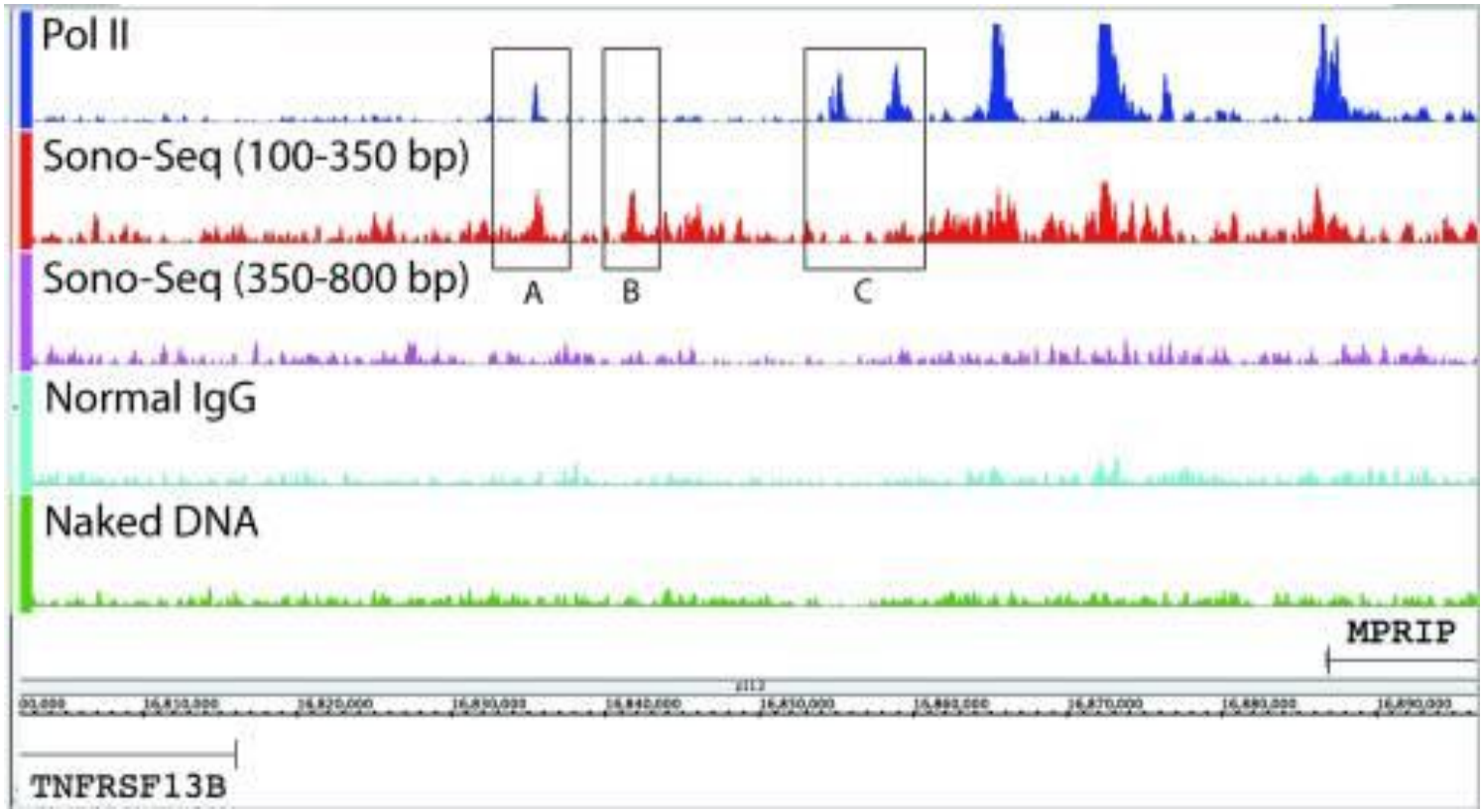
Oder eine Signifikanz (p-Wert). Was ist die Wahrscheinlichkeit N Reads zufällig zu beobachten?

Pepke et al., Nature Methods 2009

Signifikanz der Peaks

- Jeder Peak hat einen zugehörigen p-Wert
- Korrigieren für mehrere Tests
- E.g. FDR-korrigierte p-Werte (BH Korrektur)
- Oder einen empirischen FDR definieren

Die Notwendigkeit der Kontrollen

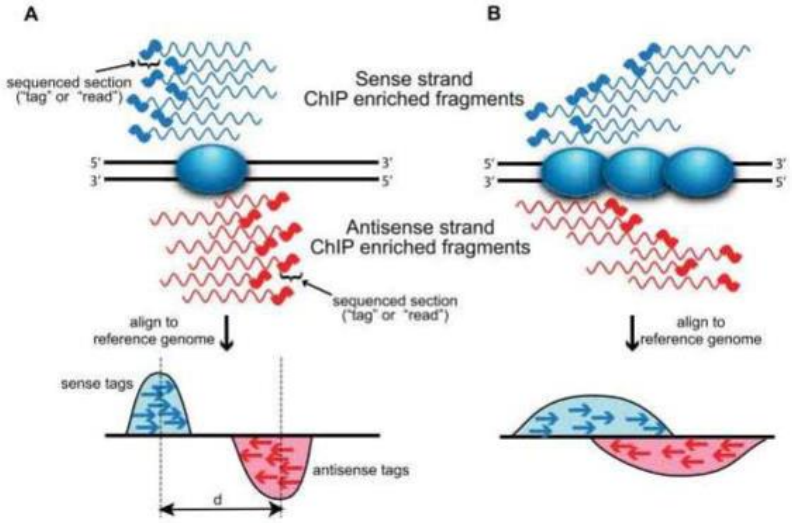


Auerbach et al., PNAS 2009

Model-based Analysis of ChIP-Seq data (MACS)

Es ist einer der am häufigsten verwendeten Peak-Finder.
MACS führt eine fortgeschrittene Art der Modellierung der Fragmentgröße ein.

ChIP-DNA-Fragmente werden gleichermaßen wahrscheinlich von beiden Enden sequenziert. Die „read density“ um eine echte Bindungsstelle sollte ein bimodales Anreicherungsmuster aufweisen



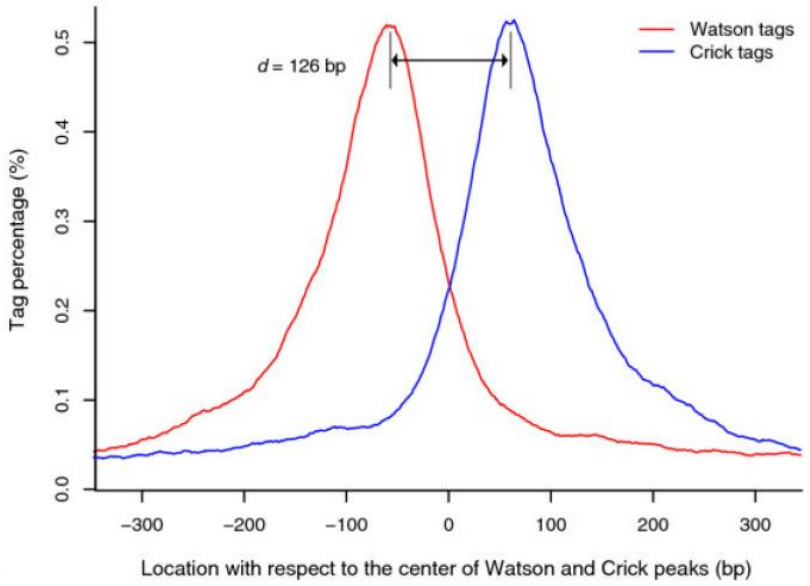
Reads werden häufig in Richtung der 3'-Richtung verschoben / ausgestreckt, um die genaue Protein-DNA-Interaktionsstelle besser darzustellen. Die Größe der Verschiebung ist dem Experimentator jedoch oft unbekannt.

MACS: shift Größe

Algorithm 1 Estimate Fragment Size

1: Slide a window of $2 \times \text{bandwidth}^3$ across genome
2: Identify regions of moderate enrichment (mfold: 10-30 fold)
3: **for each** peak i of 1000 randomly chosen enriched regions **do**
4: separate reads into + and - strand
5: Calculate mode of + and - summit
6: $d_i \leftarrow |\text{mode}_+ - \text{mode}_-|$
7: **end for**
8: $d \leftarrow \text{average}_i(d_i)$

MACS: shift Größe



Sobald *d* geschätzt worden ist, werden alle Reads um *d/2* zu ihrem 3'-Ende, in Richtung zur Mitte des Gesamtpeaks verschoben.

Ein statistischer Test wird dann verwendet, um signifikante Peaks zu bestimmen. Wie? (An der Tafel)

ChIP-Seq: Hintergrund bias

Lokale Eigenschaften des Genoms können zu einer Bias in der Anzahl der mapped Reads führen.

- Chromatin Zustand (e.g. Euchromatin-Fragmente einfacher als Heterochromatin)
- GC Inhalt
- ChIP-Seq-Experimente enthalten oft eine Kontrolle

MACS: Peak Calling

Aufgrund dieses Bias verwendet MACS anstelle eines einheitlichen λ_{BG} das aus dem gesamten Genom geschätzt wird, einen dynamischen Parameter, λ_{local} , der für jeden Kandidatenpeak definiert ist:

$$\lambda_{local} = \max(\lambda_{BG}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k})$$

λ_{BG} wird über das gesamte Genom berechnet, und λ_{1kb} , λ_{5kb} , λ_{10kb} werden aus den 1 kb, 5 kb or 10 kb Fenstern berechnet, das an der Peak-Stelle in der Kontrollprobe zentriert.

MACS: Peak Calling

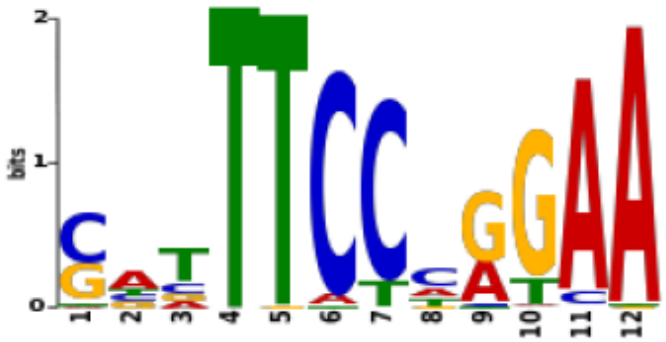
- Kandidatenpeaks mit p-Werten unterhalb eines benutzerdefinierten cutoff (default 10-5) werden als signifikant (Poisson-Verteilung) bezeichnet.

- Das Verhältnis zwischen dem ChIP-Seq-Tag-Zählwert und λ_{local} wird als die Faltenanreicherung angegeben.

Downstream Analyse

Der Satz von Peaks und deren Lage bilden die Grundlage für die biologische Interpretation der Aktion des untersuchenden Transkriptionsfaktors.

TAT1-Motiv aus FASTA-Sequenzen der Peaks, e.g. mit der RSAT Software.



Eine weitere funktionelle Analyse: GO-Anreicherung von Genen mit signifikanten Peaks (Chi-Quadrat-Test)

Referenzen

- *Metazoan: emerging characteristics and insights into transcription*. Lenhard et al., Nature Review Genetics 2012

- *ChIP-seq advantages and challenges of a maturing technology*. Peter J Park, Nature Review Genetics, 2009

- *Computation for ChIP-seq and RNA-seq studies*. Pepke et al., Nature Methods 2009

- *Model-based Analysis of ChIP-Seq (MACS).* Zhang et al.. Genome Biology 2010