Analyzing ChIP-seq data

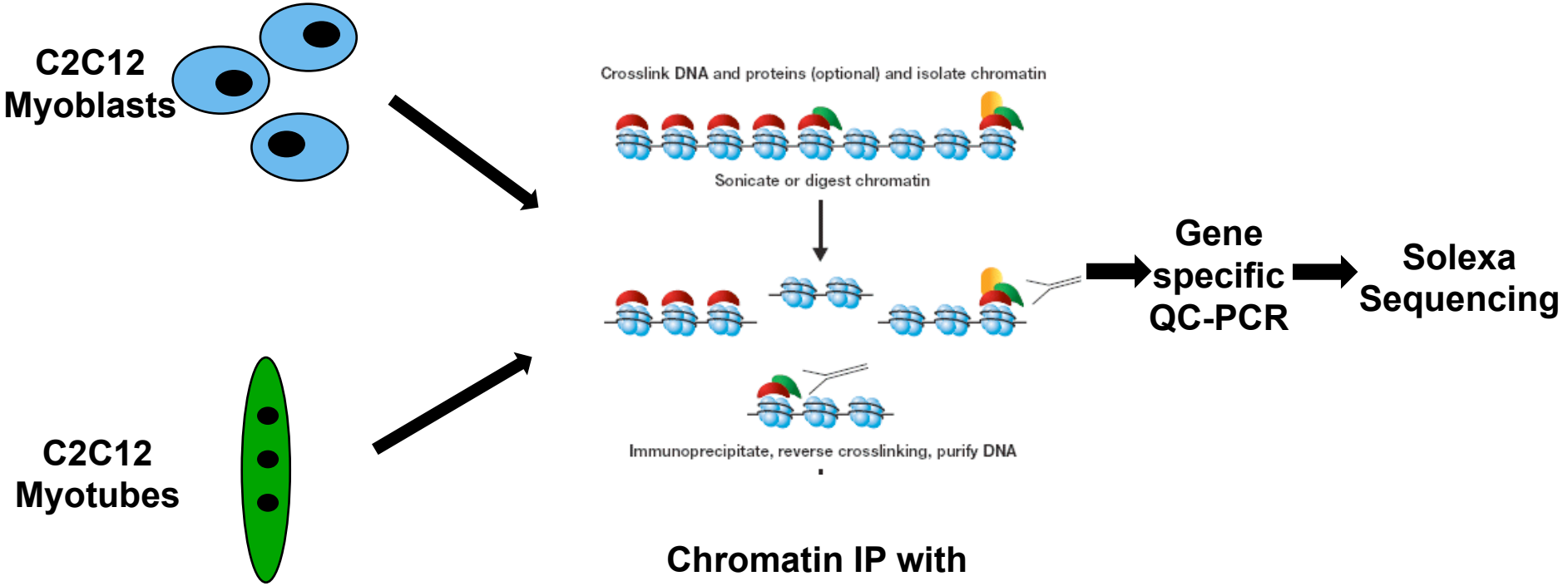
R. Gentleman, D. Sarkar, S. Tapscott, Y. Cao, Z. Yao, M. Lawrence, P. Aboyoun, M. Morgan, L. Ruzzo, J. Davison, H. Pages

FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Biological Motivation

- Chromatin-immunopreciptation followed by sequencing (ChIP-seq) is a powerful tool for:
  - epigenetics
    - histone modifications
    - methylation
  - locating transcription factor (TF) DNA interactions
- HTS technologies have made a number of experiments possible
- my interest is in somewhat complex ones (time-course; multi-factor experiments)

Experimental Design



Computational Challenges

- we are studying MyoD, a member of the bHLH family of TFs, and CTCF
- MYOD bind to an EBOX; CANNTG
  - there are lots of potential binding sites
  - 14 million in mice; 16 million in humans
  - do different members have different sequence specificity
- CTCF: 11 zinc finger protein long binding site
  - Long complex PWM
  - Association with Tes

Computational Challenges

- what role do co-factors play
- experiments with them ko-d or silenced
- time course
- other data
  - methylation
  - Histone modifications

Workflow

- Preprocessing
  - fragment length estimation; finding the most likely binding site
  - estimate background; do you need a control lane? Which peaks represent binding?
  - did we sequence deeply enough?
- tools to perform these tasks are in the chipseq package
- comparison of complex experiments is on going research
- adding genomic context: IRanges/ rtracklayer etc

Observed Data

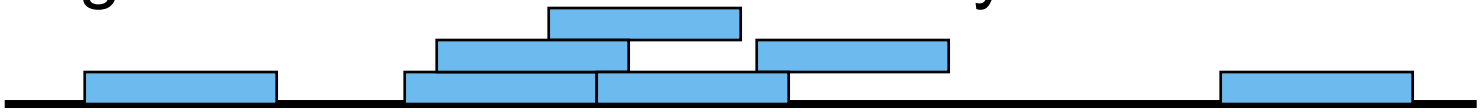
- we exclude (but ultimately won't) reads that map to more than one location
- we exclude reads that map to the same start location and orientation (since in our setting we believe that these are likely due to PCR bias)
- this forces us to think a bit about the *mappable genome:* that part of the genome we could have mapped to
  - so for 36nt reads we want to know how much of the genome is unique

Observed Data

- each fragment contributes a read, of some length (36mers for much of our data), but the real fragment of DNA was likely longer and the protein DNA interaction was somewhere on that longer fragment
  - single end reads: we read a short sequence from one end
  - paired end reads: we get a short sequence from both ends
- XSET: eXtended single-end tags
  - how much should they be extended

Notation

- island: a contiguous section covered by reads
- singleton: an island covered by 1 read



- island size: number of reads in the island
- island depth: maximum number of reads that overlap
- inter-island gap: the number of nt between two islands

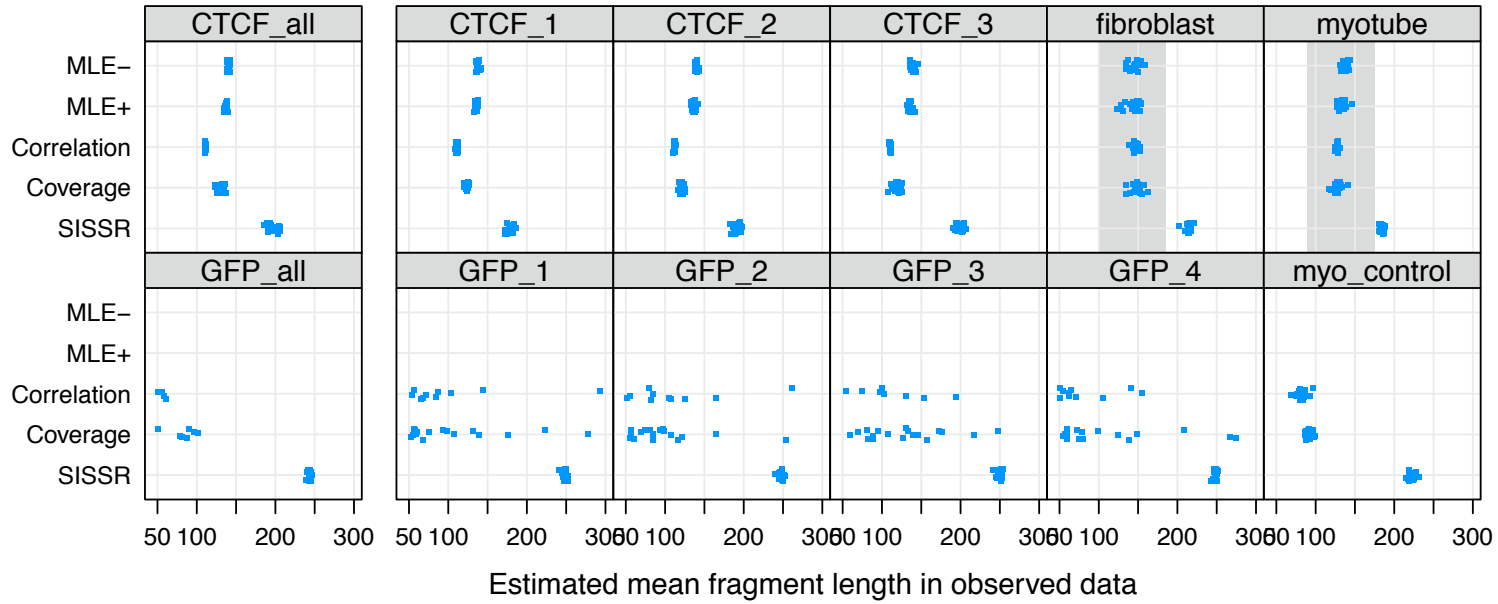
Estimating Fragment length

- there are several methods in the literature for estimating the mean fragment length
  - Kharchenko et al is quite good
  - Jothi et al is quite bad
- our method:
  - choose a lower bound, *w*, for the mean fragment length; extend all reads by *w*
  - shift each negative strand read by an amount *u*
  - compute the total number of bases covered by any read
  - find the value u_{min} of *u* for which the number of bases covered is a minimum
  - estimate the mean length by $w + u_{min}$

Estimating the fragment length

- mean fragment length is not such a good thing
  - something more like the 90%-ile of the distribution is likely to be more useful
  - with the xSet method we want to extend and cover the binding site
- when you have a known TF you can (and probably should) make use of its known PWM to find putatitive binding sites
- then for each read that maps to the genome you can find the nearest potential binding site, and from this we get a set of truncated estimates for L
- and then we can estimate percentiles of that distribution

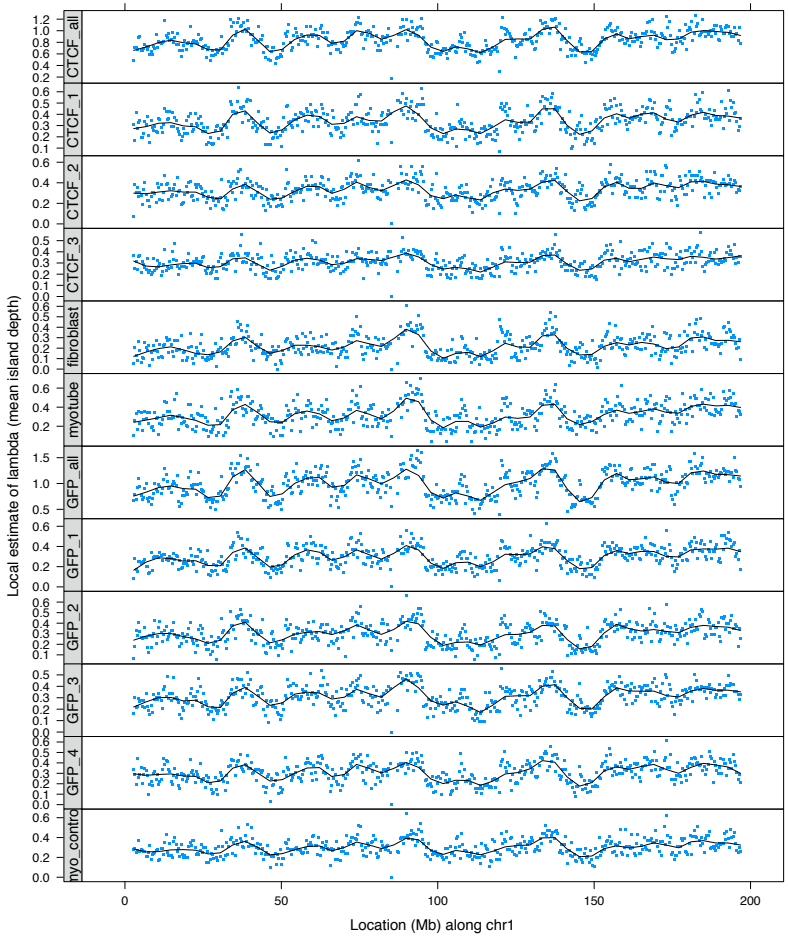
Comparison of Methods



Foreground vs Background

- we observe both reads that correspond to
  - foreground: they represents or some kind of affinity (not necessarily just what we want)
  - background:low density reads from throughout the genome
- we want to separate these two types of signal
  - the background varies within a genome and between individuals
- finding foreground is not the same problem as finding the most likely binding site
  - some peaks cover multiple binding sites
  - some peaks cover no TF binding sites

Background Varies



Null model

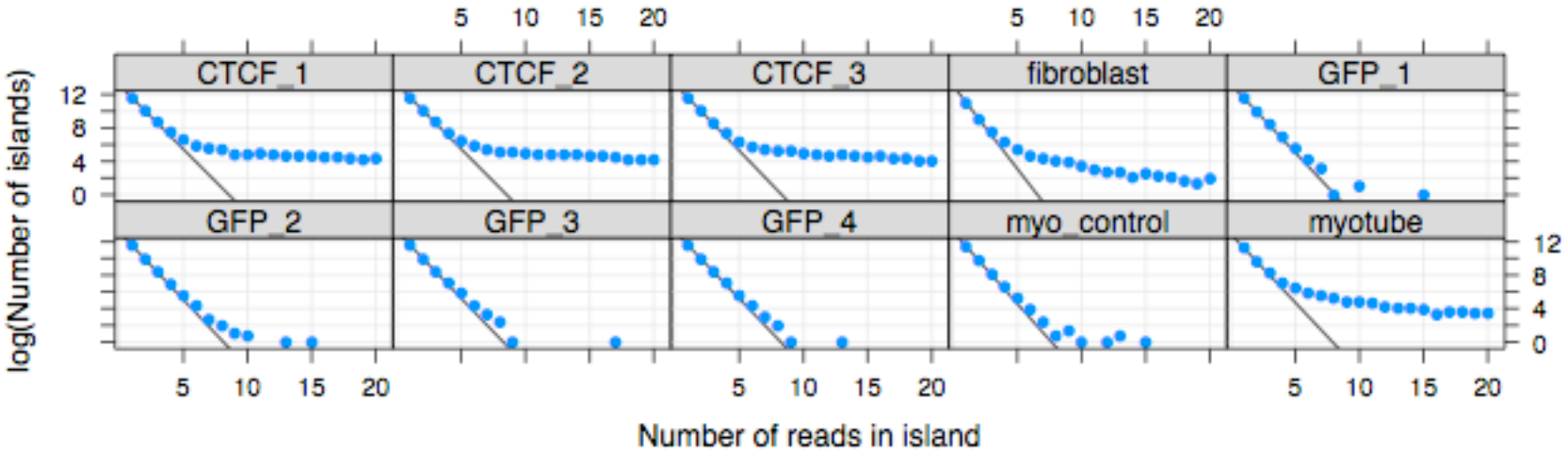
- null model: assume reads are distributed uniformly along the genome (Lander and Waterman, 1988)
- if all XSETs are of length L and let α denote the probability of a new XSET starting at any base
- then we can easily show that the number of reads in an island follows a Geometric distribution $P(N=k) = p^{k-1}(1-p)$

  where $p = 1 - (1- \alpha)^L$
- but we should only use background reads!
- we propose estimating p by using islands of size 1 or 2; and this gives us an estimate of α

Peak Discovery

- given the Poisson model for background, and α, we can develop criteria for peak heights
- we can then select a cut-off based on the probability that a peak of height k is unlikely given the background rate
- for de novo peak detection there are some problems, since the data also determine the peaks
- we did some simulation to show the effect is not so large, and we can use the simple Poisson model

Estimation of the background



- number of reads per island for Chromosome 1 (mouse)
- black line is an estimate of p, using islands with only one or two reads

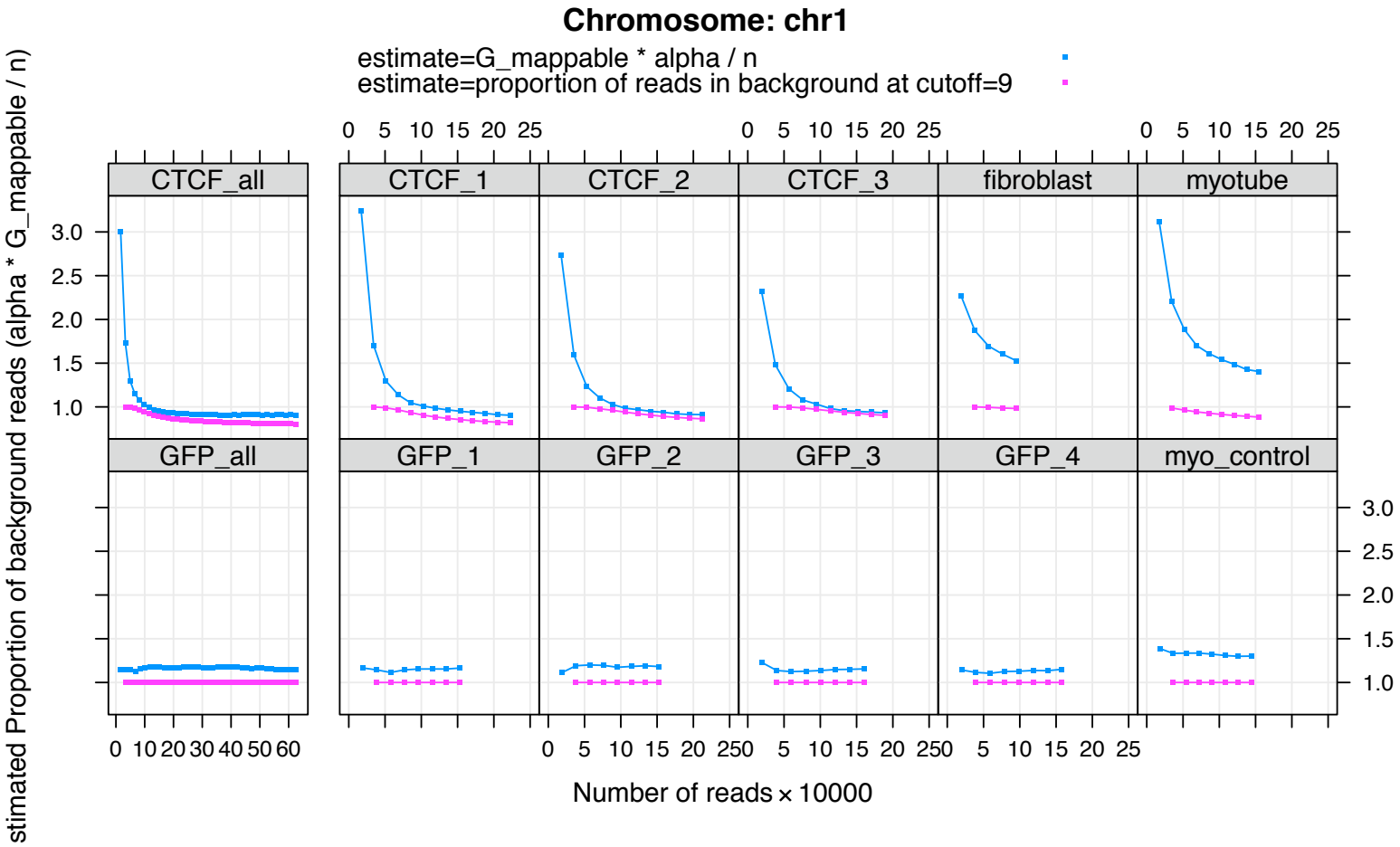
Did we sequence deeply enough?

- we can divide the genome into three categories
  - foreground, background, empty
- foreground is not informative about whether you have sequenced deeply enough
- background is informative

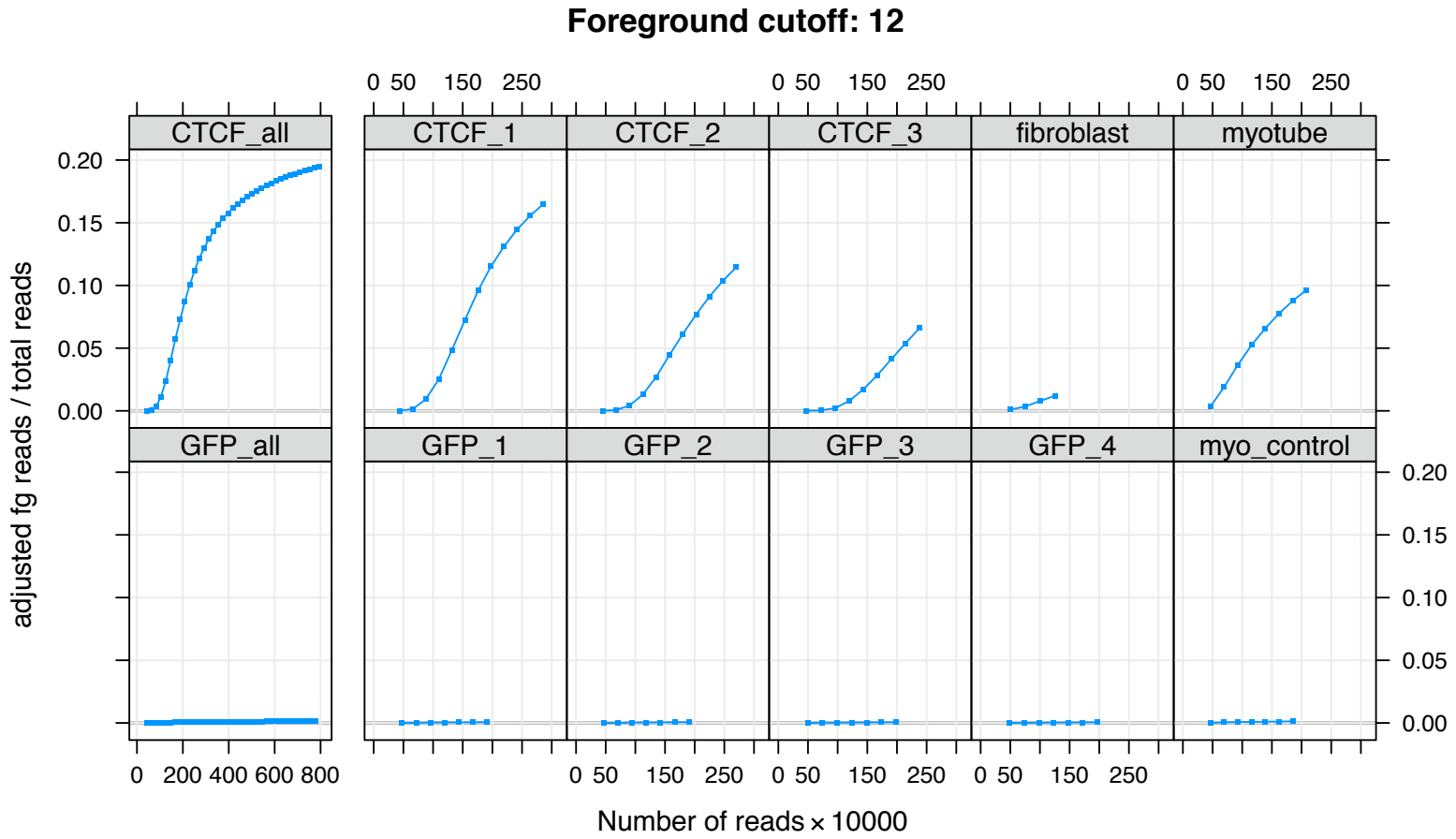
Deep Enough?

- partition the data into k groups
- add each group sequentially, and after it is added compute proportion covered by foreground (peak >= l); background (covered by reads, count < l); empty (not covered)
- for the next group we can estimate the expected number of reads that will cover each of these regions
- if we have undiscovered foreground, then we will see that the number of reads that map to background is larger than expected.

Deep Enough?


Chromosome: chr1
estimate=G_mappable * alpha / n
estimate=proportion of reads in background at cutoff=9
stimated Proportion of background reads (alpha * G_mappable / n)
CTCF_all
CTCF_1
CTCF_2
CTCF_3
fibroblast
myotube
GFP_all
GFP_1
GFP_2
GFP_3
GFP_4
myo_control
Number of reads × 10000

Foreground



FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Where did the TF bind?

- we should get reads from both the + and - strand
- the reads on the - strand should be upstream of the binding site
- those on the + strand should be downstream