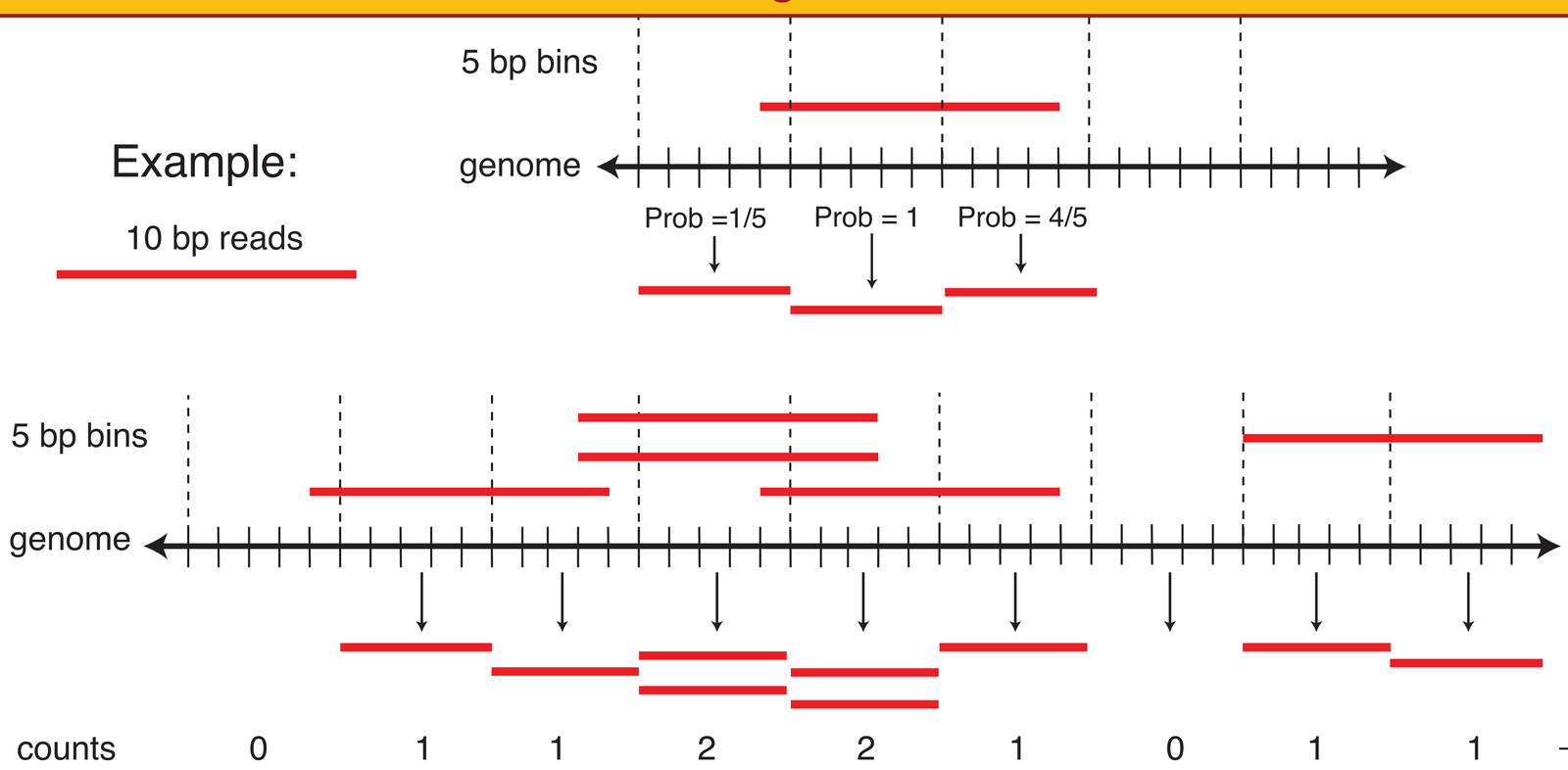
Predicting genomic coverage by probabilistic binning and extrapolation

Timothy Daley[1] and Andrew Smith[2]

[1]Department of Mathematics and [2]Department of Molecular and Computation Biology, University of Southern California

USC University of Southern California

Probabilistic binning of reads:



$n_i = \#$ bins with count i

$= \sum_j 1(\text{ bin } j \text{ has } i \text{ reads covering})$

Assume general Lander-Waterman model

$L =$ Expected # total bins

$=$ genome coverage / bin size

$\mu =$ Poisson rate of a random bin

$E\, n_i =$ expected # bins with count i

$= L \int_0^\infty e^{-\lambda} \lambda^i / i!\, d\mu(\lambda)$

$\longrightarrow n_1 = 5,\ n_2 = 2$

Are bins uncovered due to random chance or because no read overlaps? Can't distinguish true zeros from missing zeros

Genomic coverage ≈ bin size • # covered bins:

Extrapolating coverage $\longleftrightarrow$ Extrapolating covered bins

Daley & Smith *Nature Methods* (2011):

- Predicting number of distinct reads from additional sequencing.
- Treat distinct bins like distinct reads.
- Apply Good & Toulmin's non-parametric empirical Bayes solution (*Biometrika* 1959) and rational function approximation

$$\begin{aligned}
\Delta(t) &= \#\text{ new covered bins from sequencing an additional } tN \text{ reads} \\
&= E\big(\text{bins uncovered in additional experiment}\big) - E\big(\text{initial \# uncovered bins}\big) \\
&= L \int_0^\infty e^{-\lambda}\big(1 - e^{-(t-1)\lambda}\big)\, d\mu(\lambda) \\
&= \sum_{i=1}^{\infty} (-1)^{i+1} (t-1)^i E n_i \ \approx\ \frac{p_0 + p_1(t-1) + \cdots p_P (t-1)^P}{1 + q_1(t-1) + \cdots q_Q (t-1)^Q}
\end{aligned}$$

Results: extrapolating using 2% of the data

Data: 2 Single Cell libraries comparing library construction methods (Zong *et al.* Science 2012). MALBAC (Zong *et al.* Science 2012) & MDA (Dean *et al.* Gen. Res. 2001)

Library	MALBAC	MDA
SRA	SRX202978	SRX204160
Total reads	563.7M	534.8M
Reads mapping to female chroms (bwa)	449.6M	482.1M

Downsample 2% of reads with Picard: