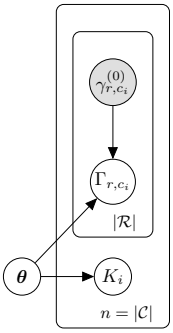
Cell Systems, Volume *7*

Supplemental Information

Statistical Binning for Barcoded Reads Improves Downstream Analyses

Atiya Shajii, Ibrahim Numanagić, Christopher Whelan, and Bonnie Berger

Variable	Description
$\mathcal{C}$	set of all clouds in connected component
$\mathcal{R}$	set of all reads mapping to some cloud in $\mathcal{C}$
$\boldsymbol{\theta} = (\theta_{c_1}, \ldots, \theta_{c_n})$	vector of cloud weights
K_i	number of reads generated by cloud c_i
Γ_{r,c_i}	event that read r truly originates from cloud c_i
γ_{r,c_i}	$\Pr(\Gamma_{rc_i}\|\theta)$
$\gamma^{(0)}_{r,c_i}$	$\Pr(\Gamma_{rc_i})$ (prior based on edit distance, mate, etc.)



$$\boldsymbol{\theta} = (\theta_{c_1}, \ldots, \theta_{c_n}) \sim \mathrm{Dir}(\mathbf{1})$$
$$K_i \mid \boldsymbol{\theta} \sim \mathrm{Cloud}(\theta_{c_i})$$
$$\Gamma_{r,c_i} \sim \mathrm{Ber}(\gamma^{(0)}_{r,c_i})$$
$$\Gamma_{r,c_i} \mid \boldsymbol{\theta} \sim \mathrm{Ber}(\gamma_{r,c_i})$$

Figure S1: Graphical representation of EMA's latent variable model involved in barcoded read alignment, as described in STAR Methods. Related to Figure 1. $\boldsymbol{\theta}$ denotes the vector of cloud weights; K_i denotes the number of reads generated by cloud $c_i \in \mathcal{C}$; Γ_{r,c_i} denotes whether read $r \in \mathcal{R}$ maps to cloud c_i, and $\gamma^{(0)}_{r,c_i}$ is a prior on this event based on barcode-oblivious information like edit distance, mate alignment, etc.

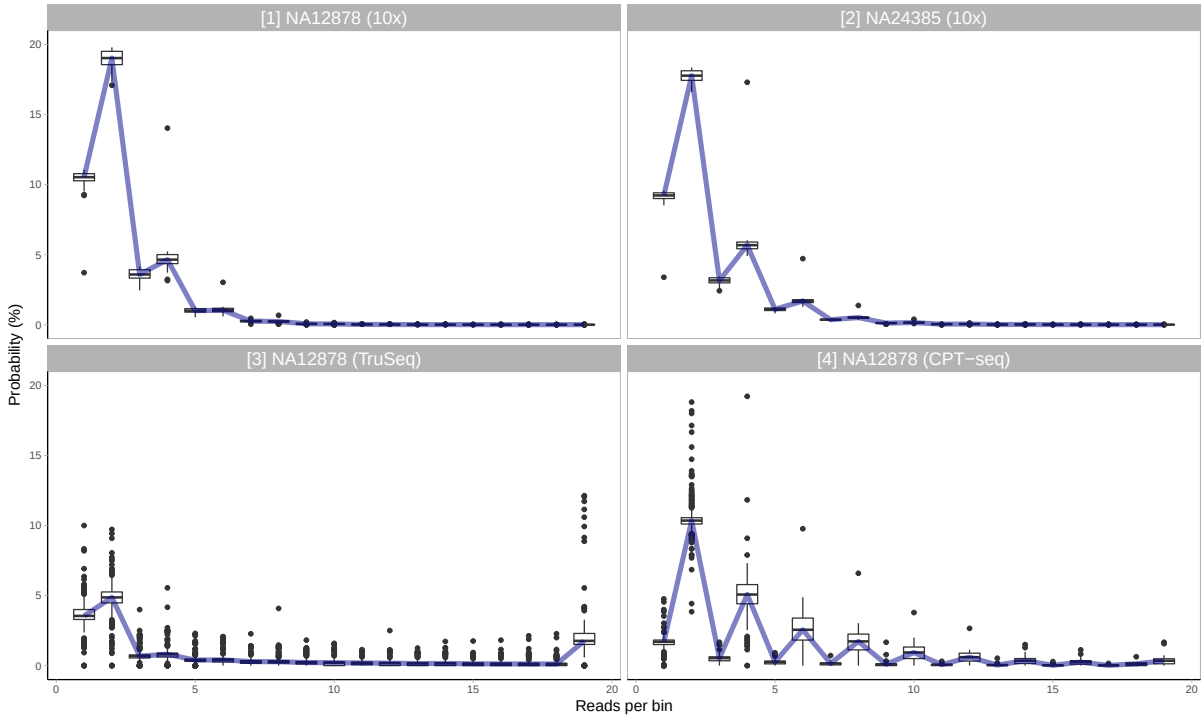


Figure S2: Distribution of the number of reads in a 1kb window within a cloud. Related to Figure 1. The first row shows the distribution of two 10x data samples, while the bottom row shows TruSeq SLR's and CPT-seq's distributions. We only consider the clouds in which no reads have multiple alignments within the cloud. The box plots correspond to different bin offsets within the cloud.

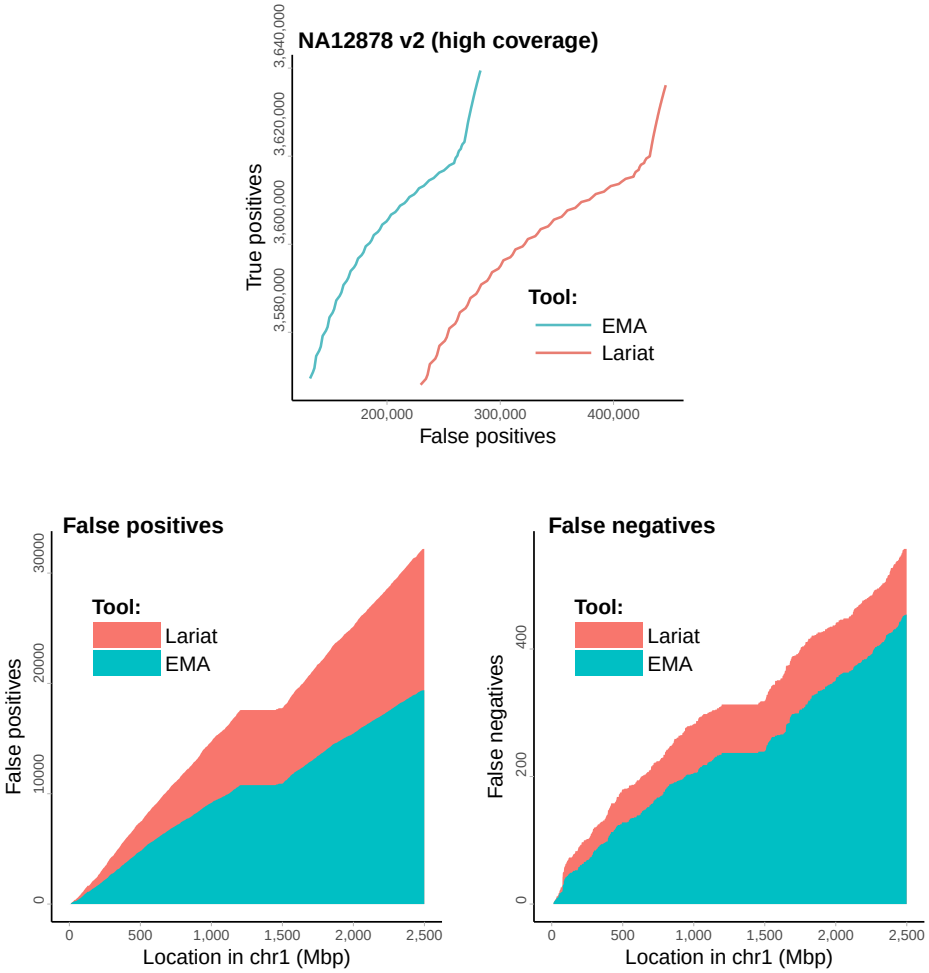


Figure S3: Genotyping accuracy for EMA as compared to Lariat on a high-coverage NA12878 10x dataset. Related to Figure 2. The top plot shows true positives as a function of false positives, and the bottom two plots are cumulative histograms of true and false positives throughout chromosome 1. We note that EMA's improvement is even more substantial with higher coverage.