Quiver: modeling consensus accuracy

David H. Alexander
Pacific Biosciences

August 23, 2013

Seeking a model for consensus accuracy

- How do characteristics of chemistry influence consensus accuracy?
  - Merge rate
  - Branch rate
  - Miscall rate
- Predictions for C2, XL, P4, and dyeball chemistries

Previous approaches

- Most obvious approach is binomial sampling model,

$$mathgoeshere$$

- This approach makes wrong assumptions about PacBio
    - Suggests very high consensus accuracy
    - For PacBio, aligning the reads is the challenge, not tabulating bases in columns (miscall rate ˜0.5%, indel rate ˜12-15%)
    - Homopolymer errors are the problem

Our approach: focus on homopolymers

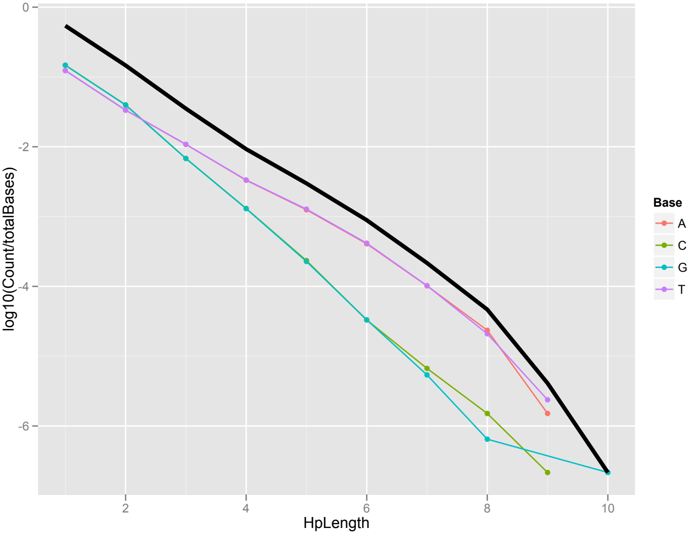


Figure : *E. coli* K12 homopolymer length distribution

Simple model for homopolymer errors

$$Y = X + B - M;$$

$$B \sim \mathrm{Bin}(X, \beta);$$

$$M \sim \mathrm{Bin}(X - 1, \mu);$$

$$B \perp M$$

Y: observed HP length
X: true HP length
B: branches
M: merges
β: branching rate
μ: merging rate

Parameters estimated from EDNA

Chemistry	Branch	Merge	Dark
C2	0.061	0.067	0.026
P4C2	0.056	0.057	0.023
Dyeball.9566.Std	0.029	0.154	0.048
Dyeball.Final	0.035	0.120	0.038

For now, averaging across channels, SNRs

Model (with C2 parameters) seems realistic

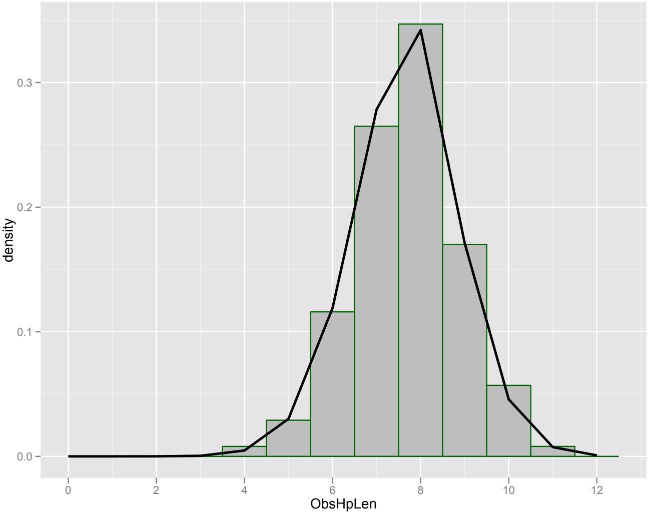
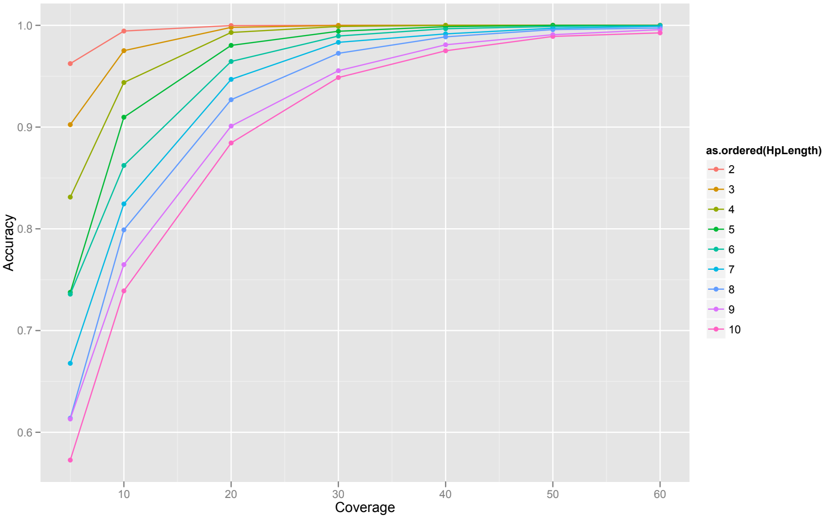
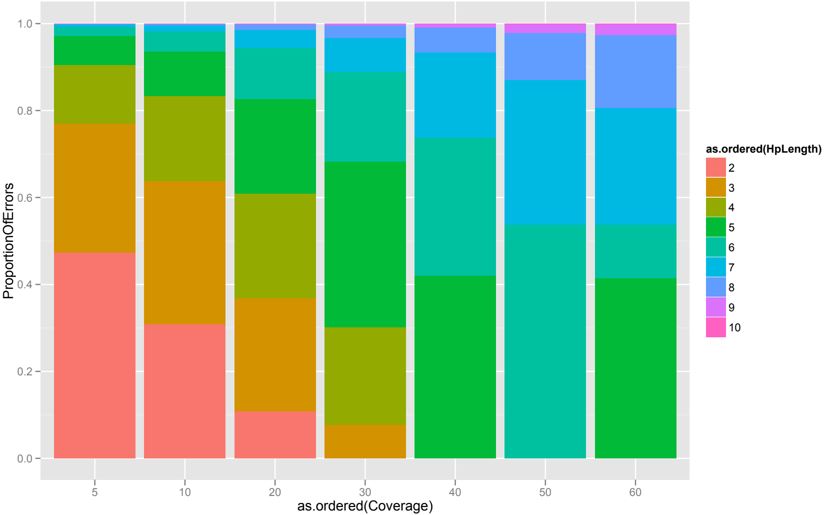


Figure : Monte-Carlo simulated observed HP length distribution

Predicted HP accuracy by length, coverage (C2 params)

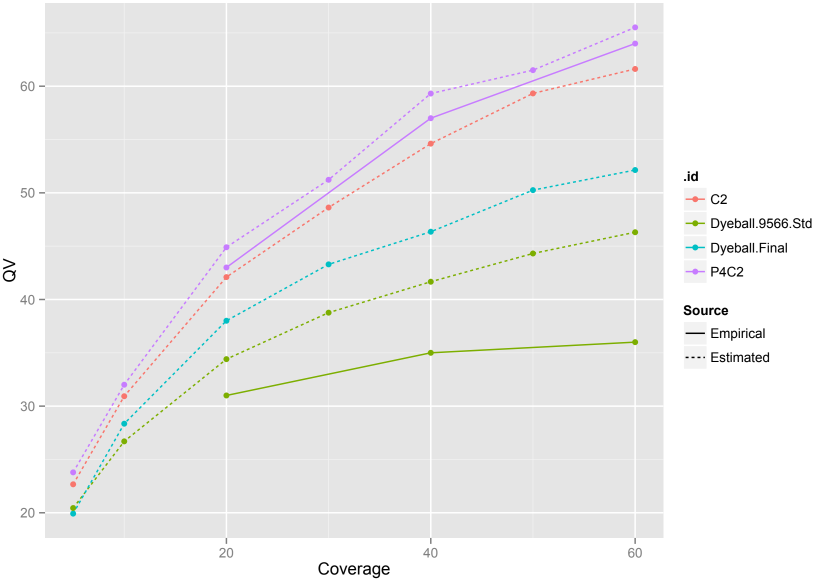


Distribution of homopolymer errors by length (C2 params)



(Based on distribution of HP lengths in *E. coli* K12)

Overall consensus accuracy prediction for *E. coli* K12


QV
Coverage
.id
C2
Dyeball.9566.Std
Dyeball.Final
P4C2
Source
Empirical
Estimated