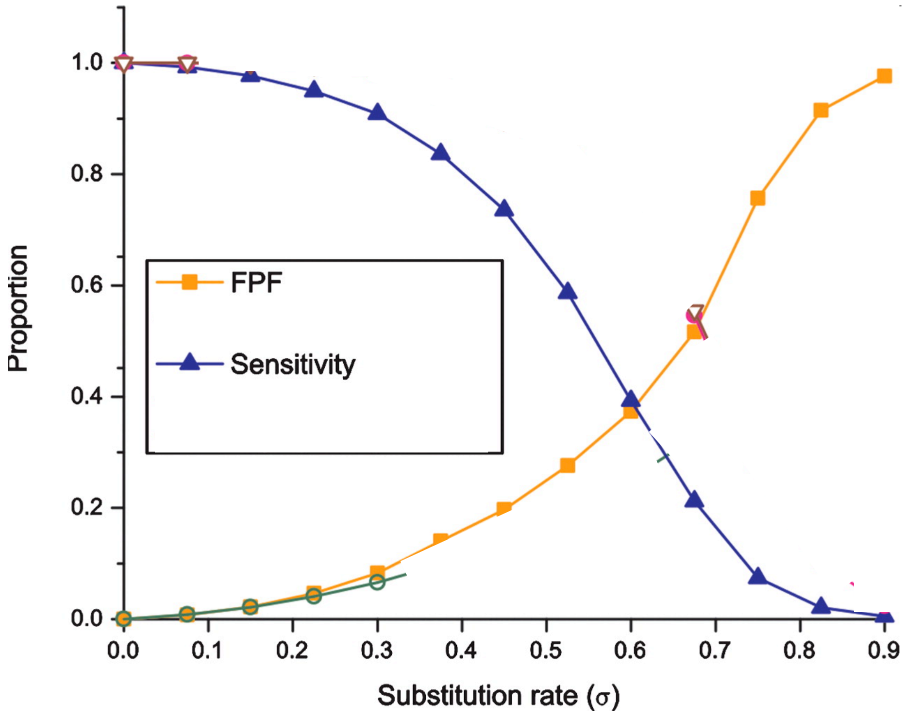
Indel rates and probabilistic alignments

Gerton Lunter

Budapest, June 2008

Alignment accuracy



→ Observed FPF

Simulation:

Jukes-Cantor model

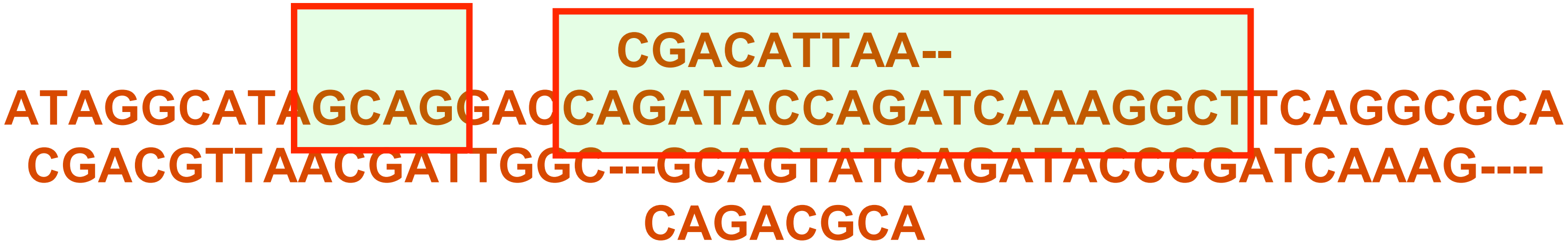
Subs/indel rate = 7.5

Aligned with Viterbi + true model

Neutral model for indels

CGACATTAA--
ATAGGCATAGCAGGACCAGATACCAGATCAAAGGCTTCAGGCGCA
CGACGTTAACGATTGGC---GCAGTATCAGATACCCGATCAAAG----
CAGACGCA

Neutral model for indels



CGACATTAA--
ATAGGCATAGCAGGACCAGATACCAGATCAAAGGCTTCAGGCGCA
CGACGTTAACGATTGGC---GCAGTATCAGATACCCGATCAAAG----
CAGACGCA

- Look at ***inter-gap segments***

  Pr(length = L) ?

Neutral model for indels

CGACATTAA--ATAGGCATAGCAGGACCAGATACCAGATCAAAGGCTTCAGGCGCA
CGACGTTAACGATTGGC---GCAGTATCAGATACCCGATCAAAG----CAGACGCA

i *i+1*

- Look at ***inter-gap segments***

Pr(length = L) ?

Def: $\boldsymbol{p_i}$ = Pr(column ***i+1*** survived | column ***i*** survived)

Assumption: indels are *independent* of each other

Neutral model for indels

CGACATTAA--
ATAGGCATAGCAGGACCAGATACCAGATCAAAGGCTTCAGGCGCA
CGACGTTAACGATTGGC---GCAGTATCAGATACCCGATCAAAG----
CAGACGCA

i $i+1$

- Look at ***inter-gap segments***

$$\mathbf{Pr(\ length = L\ )} \propto p_i\, p_{i+1} \ldots p_{i+L-2}$$

Def: p_i = Pr(column $i+1$ survived | column i survived)

Assumption: indels are *independent* of each other

Assumption: indels occur *uniformly* across the genome

Neutral model for indels

```
CGACATTAA--
ATAGGCATAGCAGGACCAGATACCAGATCAAAGGCTTCAGGCGCA
CGACGTTAACGATTGGC---GCAGTATCAGATACCCGATCAAAG----
CAGACGCA
```

i *i+1*

- Look at ***inter-gap segments***

Pr(length = L) $\propto$ $\boldsymbol{p^L}$

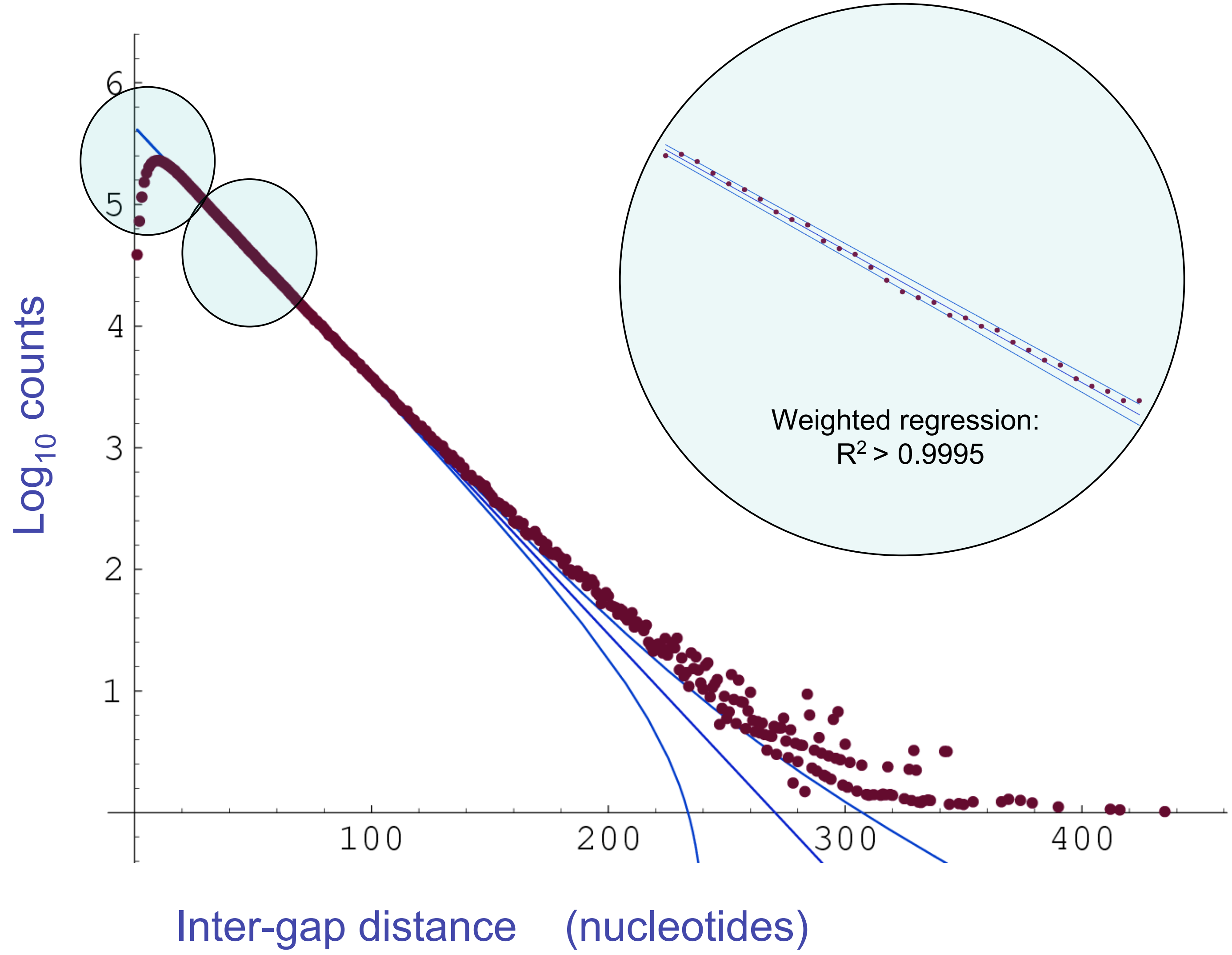
Def: p_i = Pr(column ***i+1*** survived | column ***i*** survived)

Assumption: indels are *independent* of each other

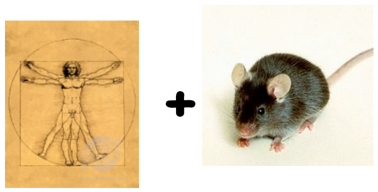
Assumption: indels occur *uniformly* across the genome

Prediction: Inter-gap distances follow a *geometric distribution*

Inter-gap distances in alignments

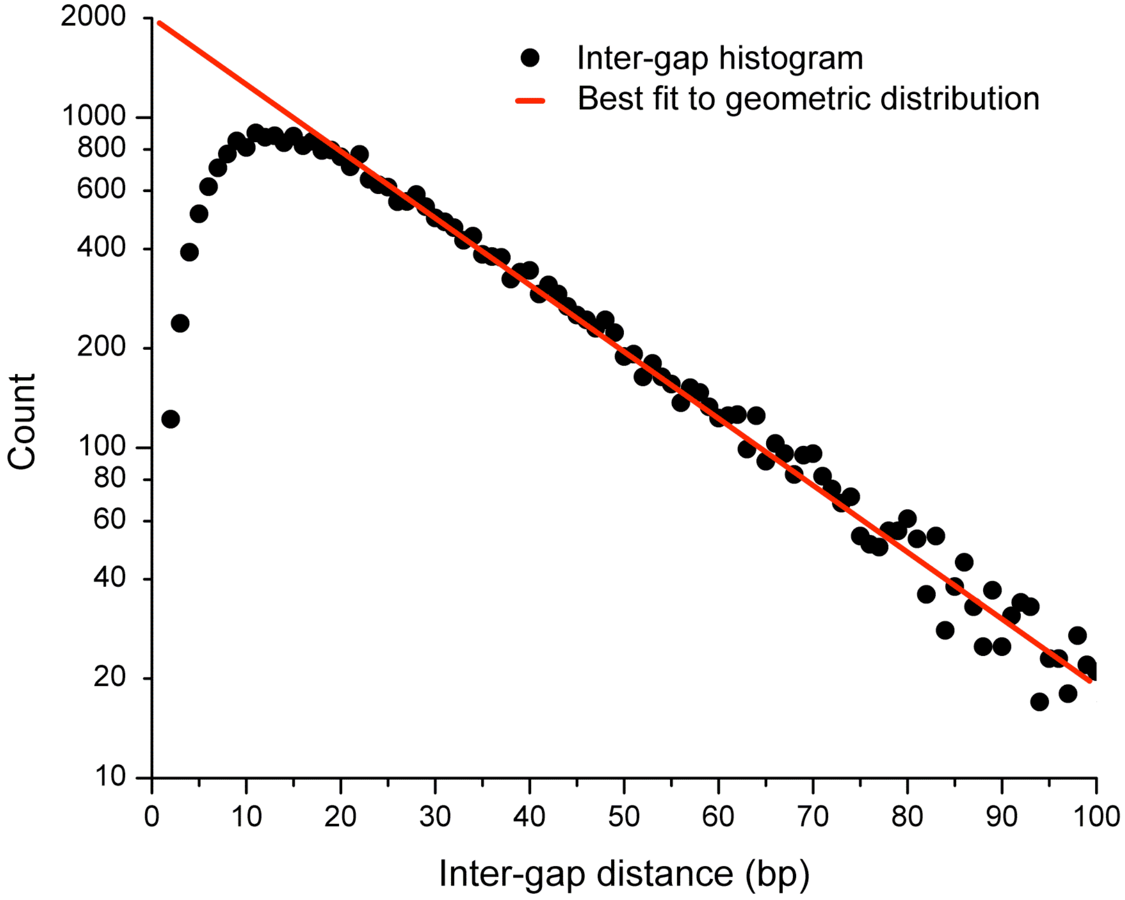


Weighted regression:
$R^2 > 0.9995$

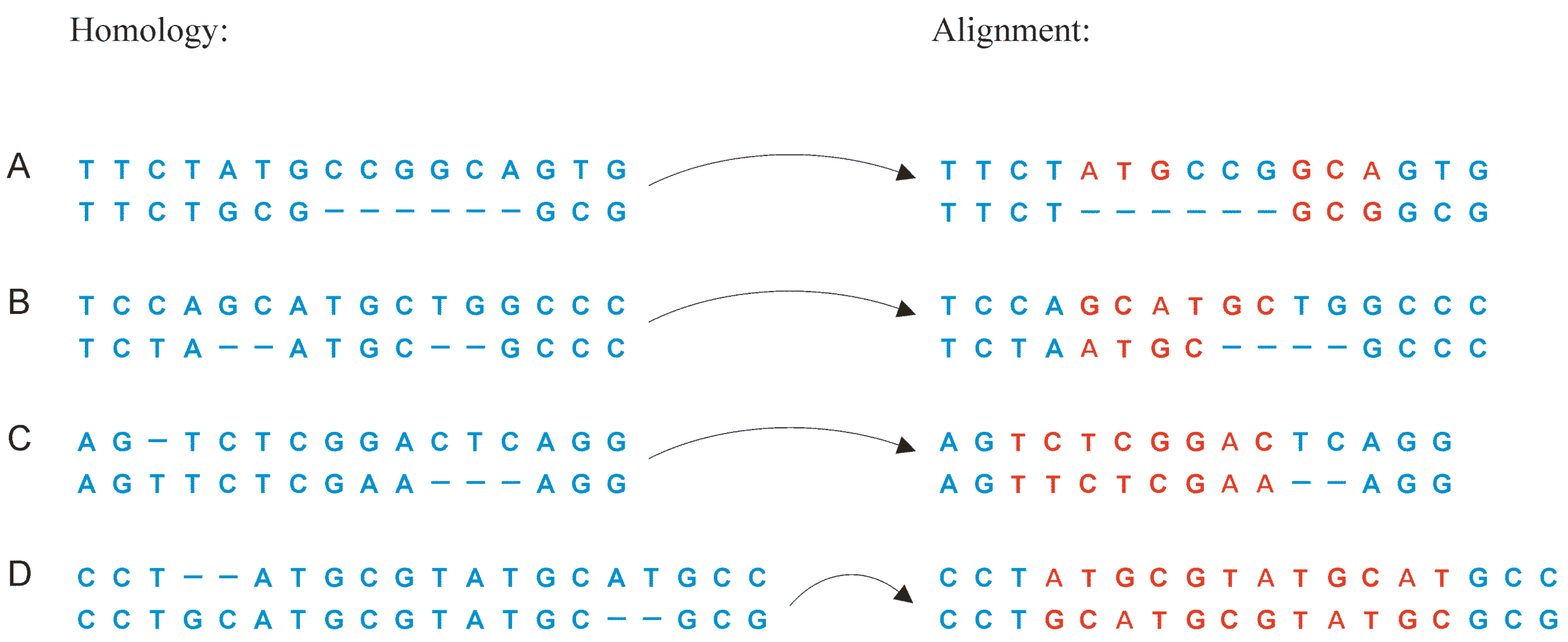


Transposable elements

Inter-gap distances in alignments (simulation)



Biases in alignments



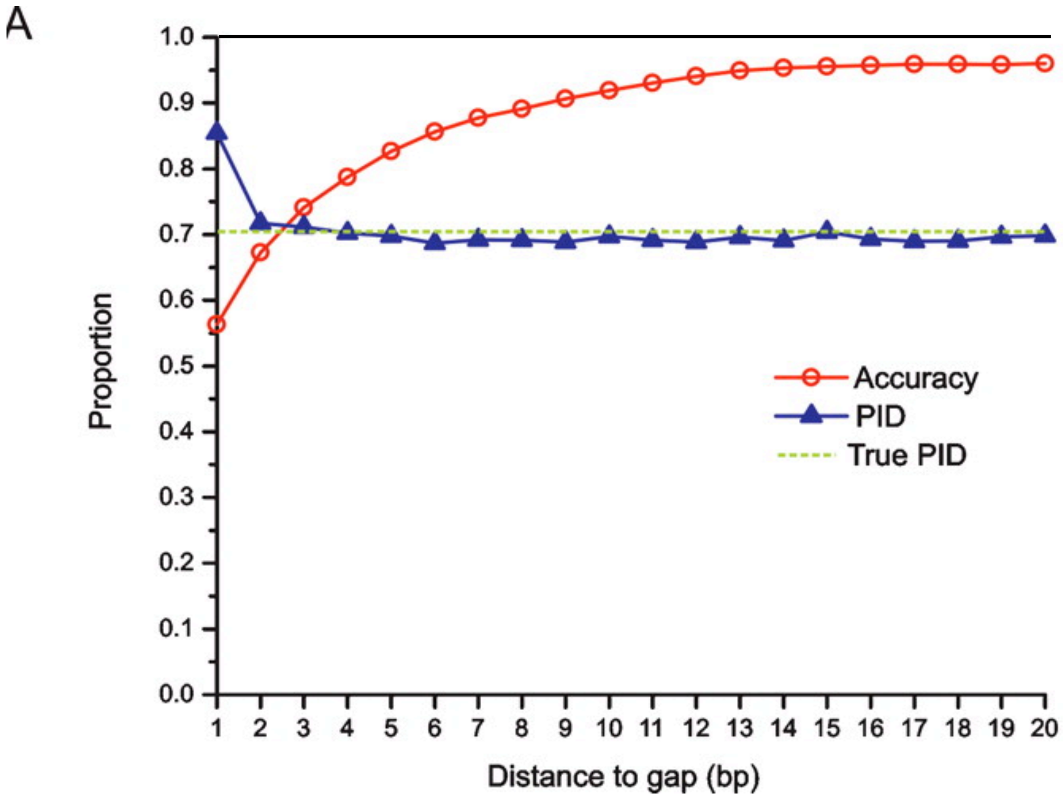
```
      Homology:                                 Alignment:

A     TTCTATGCCGGCAGTG                          TTCTATGCCGGCAGTG
      TTCTGCG------GCG                          TTCT------GCGGCG

B     TCCAGCATGCTGGCCC                          TCCAGCATGCTGGCCC
      TCTA--ATGC--GCCC                          TCTAATGC----GCCC

C     AG-TCTCGGACTCAGG                          AGTCTCGGACTCAGG
      AGTTCTCGAA---AGG                          AGTTCTCGAA--AGG

D     CCT--ATGCGTATGCATGCC                      CCTATGCGTATGCATGCC
      CCTGCATGCGTATGC--GCG                      CCTGCATGCGTATGCGCG
```

A: gap wander (Holmes & Durbin, JCB 5 1998)

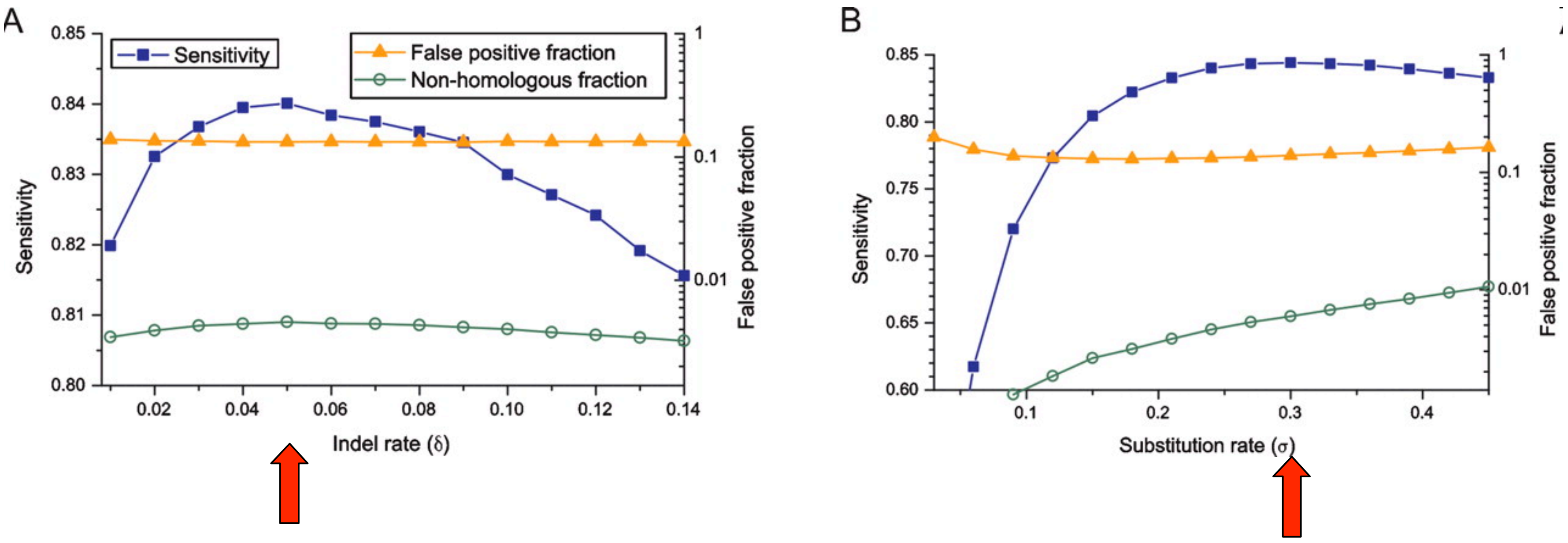
B,C: gap attraction

D: gap annihilation

Biases in alignments

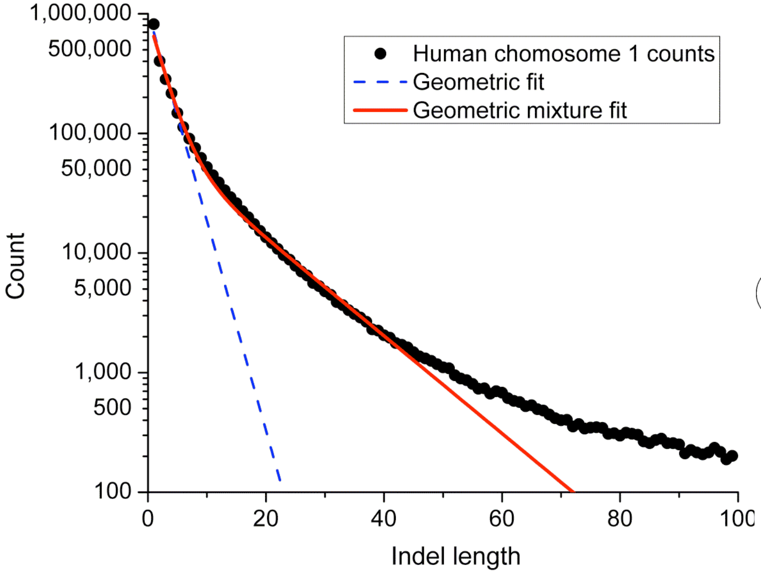


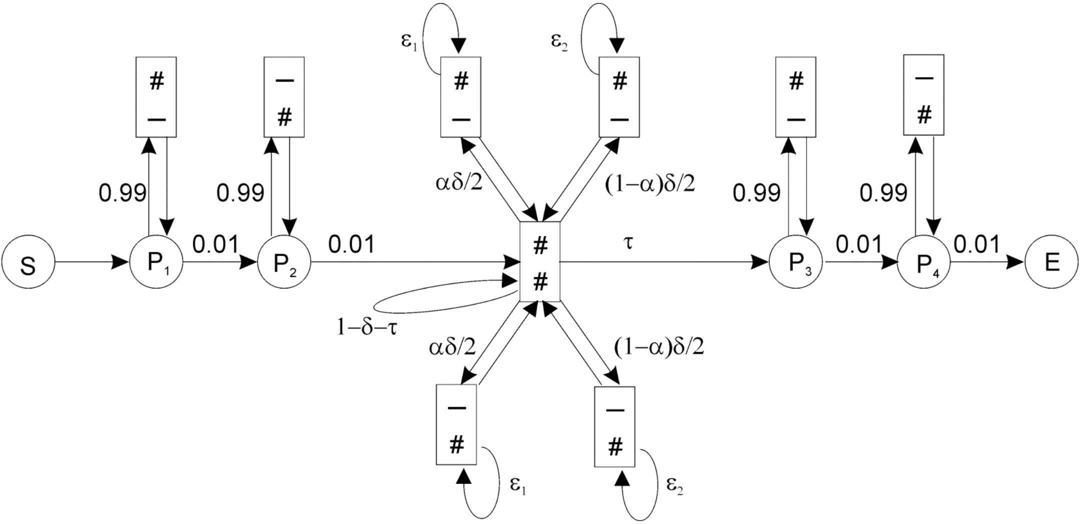
Influence of alignment parameters



- De-tuning of parameters away from “truth” does not improve alignments
- Accuracy of parameters (within ~ factor 2) does not hurt alignments much

Influence of model accuracy



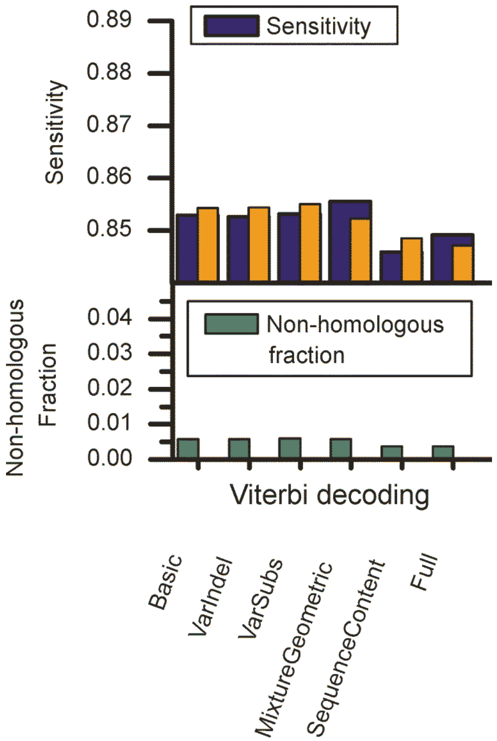


Improved model (for mammalian genomic DNA):

- Better modelling of indel length distribution
- Substitution model & indel rates depend on local GC content
- Additional variation in local substitution rate

Parameters: BlastZ alignments of human and mouse

Influence of model accuracy



Simulation:

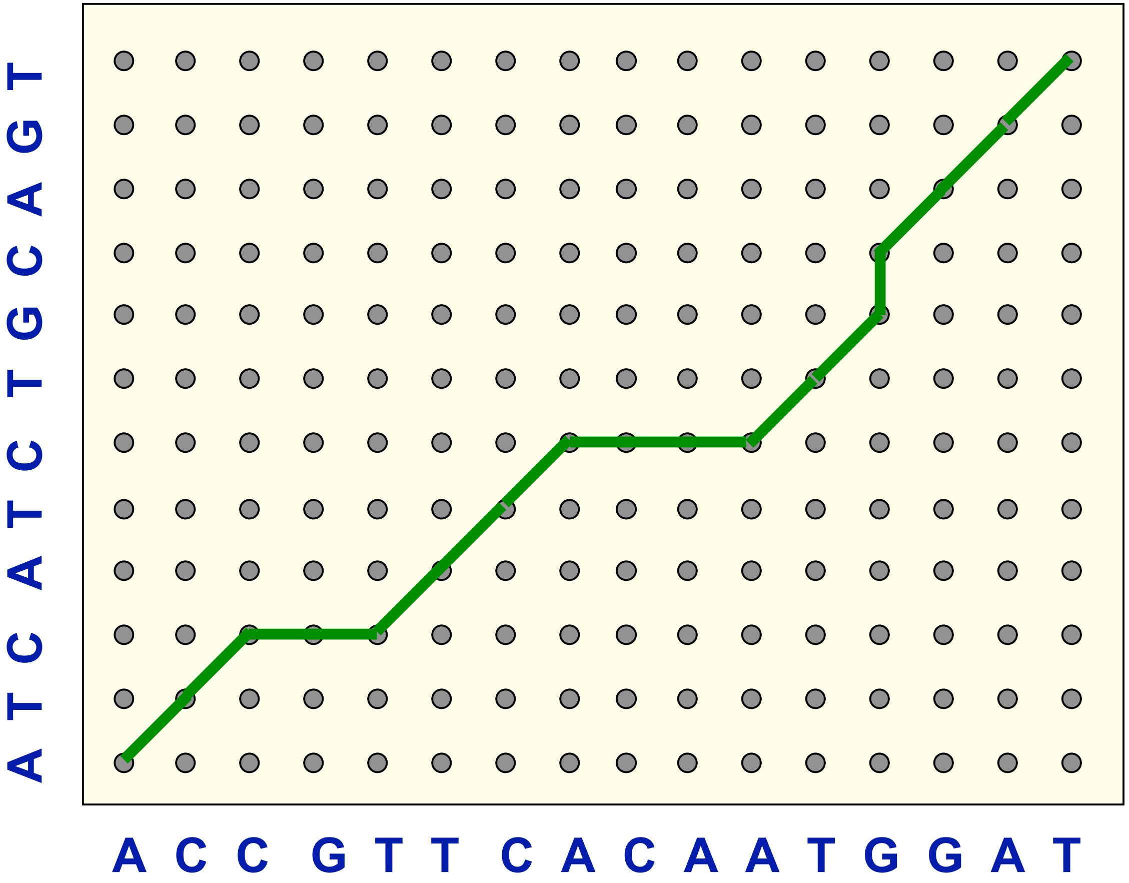
- 20 GC categories
- 10 substitution rate categories
- 100 sequences each = 20.000 sequences
- Each ~800 nt, + 2x100 flanking sequence

Summary so far

- Alignments are biased
  - Accuracy depends on position relative to gap
  - Fewer gaps than indels
- Alignments can be quite inaccurate
  - For 0.5 subs/site, 0.067 indels/site:
    accuracy = 65%, false positives = 15%
- Choice of parameters does not matter much
- Choice of MODEL does not matter much…

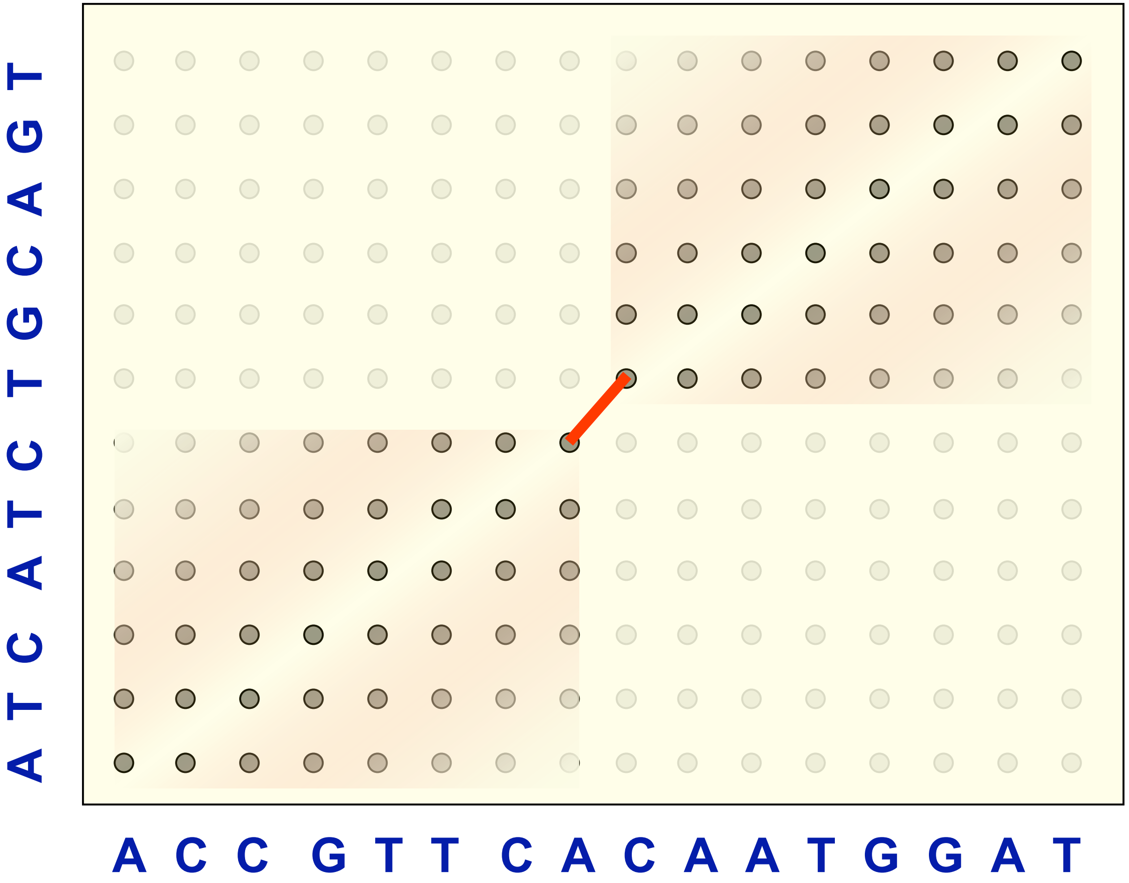
Alignments: Best scoring path

(Needleman-Wunsch, Smith-Waterman, Viterbi)

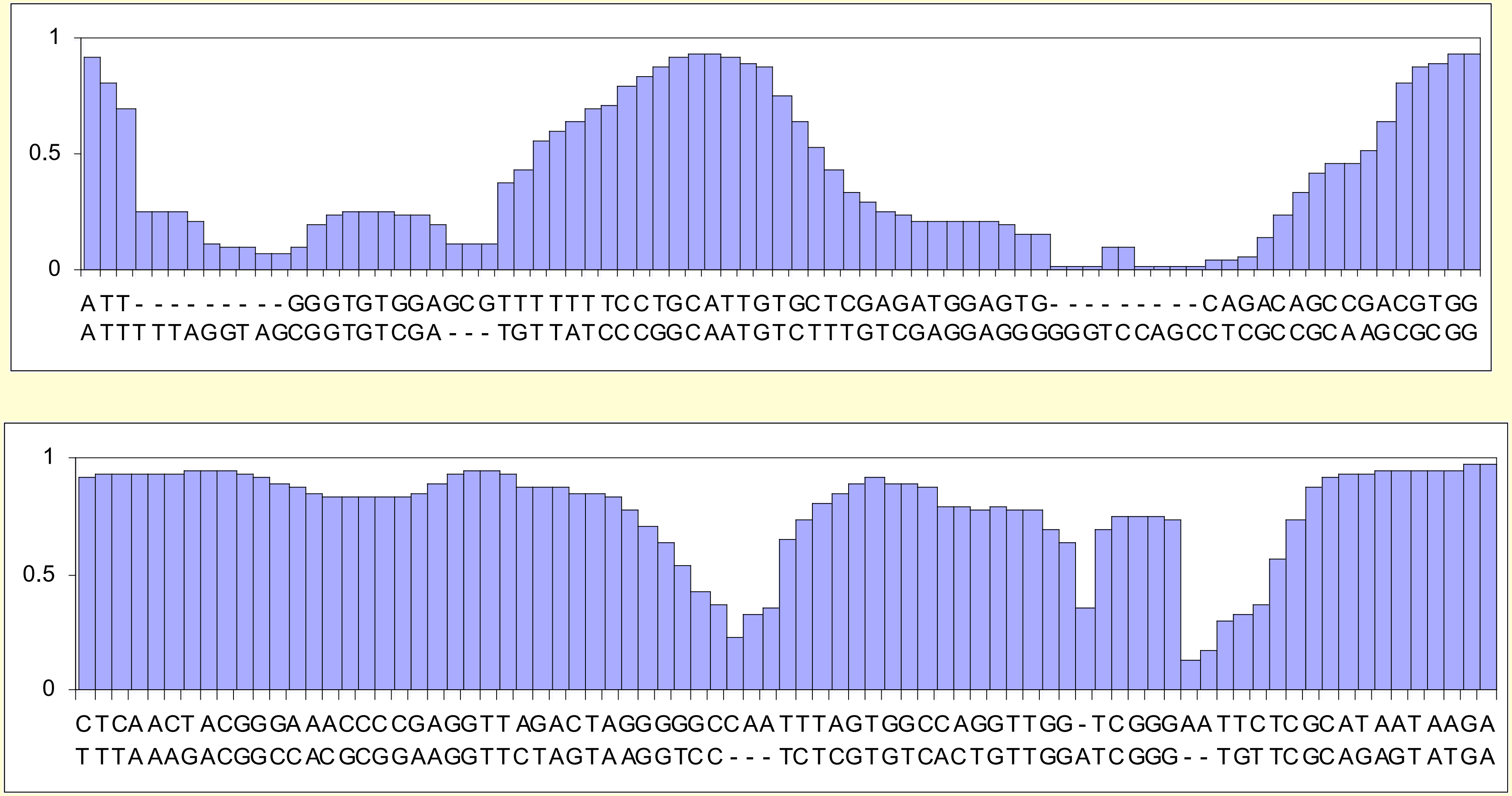


Alignments: Posterior probabilities

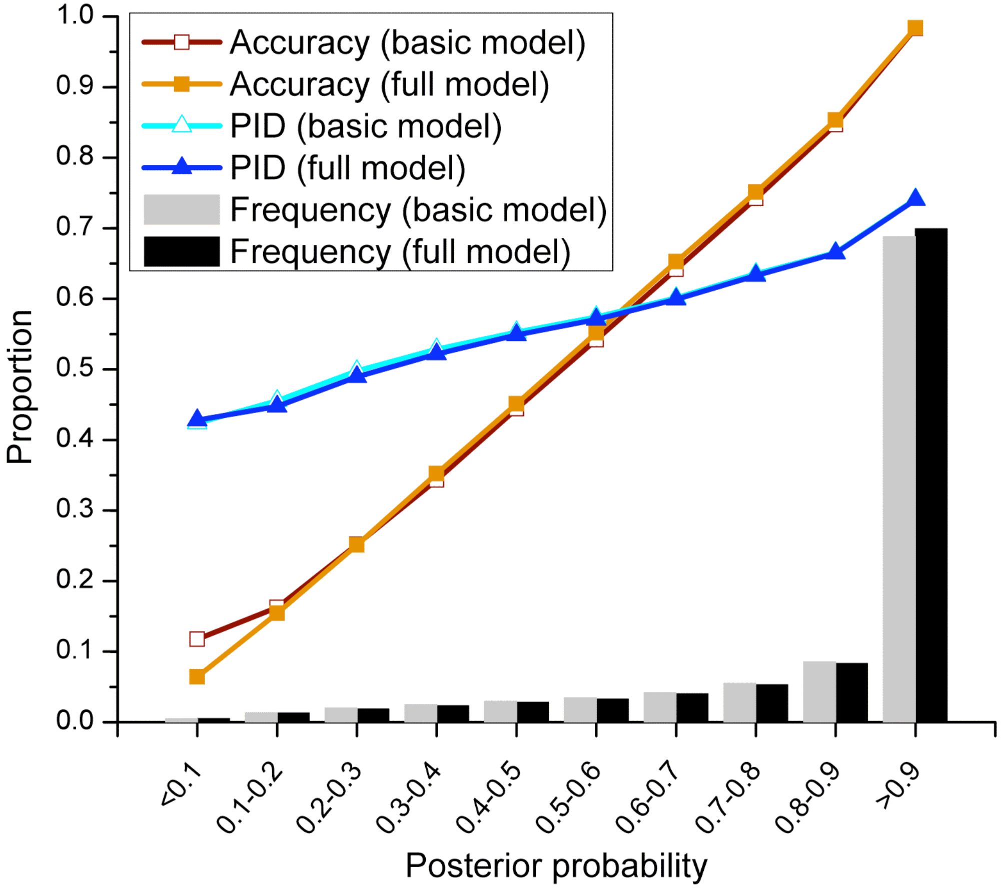
(Durbin, Eddy, Krogh, Mitchison 1998)



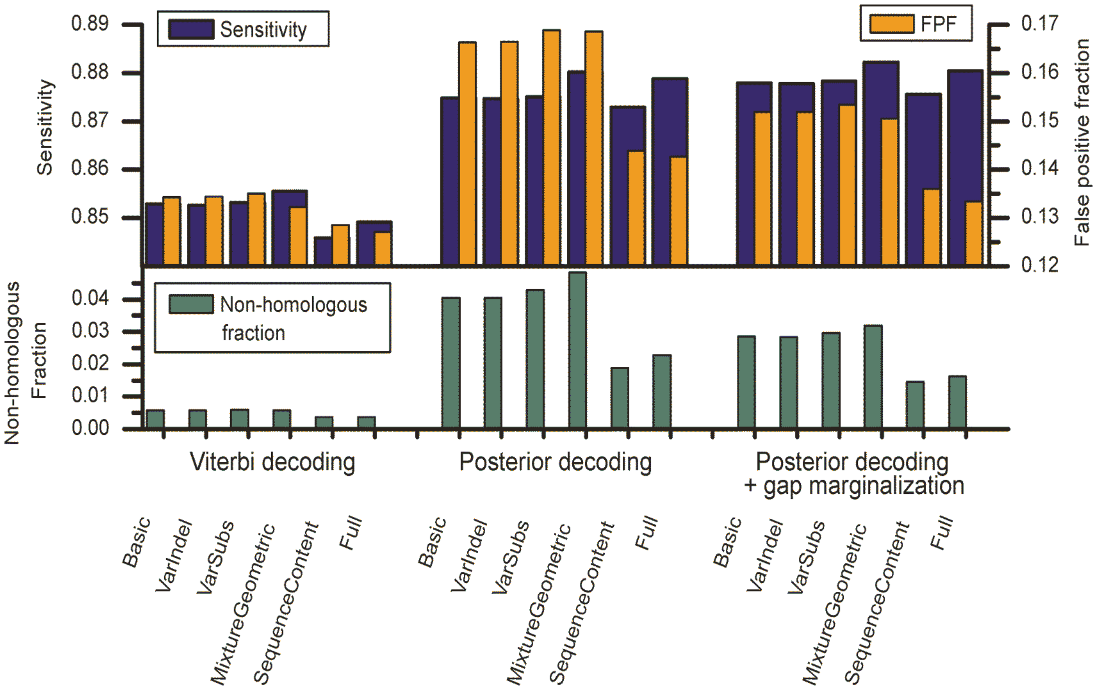
Posterior probabilities

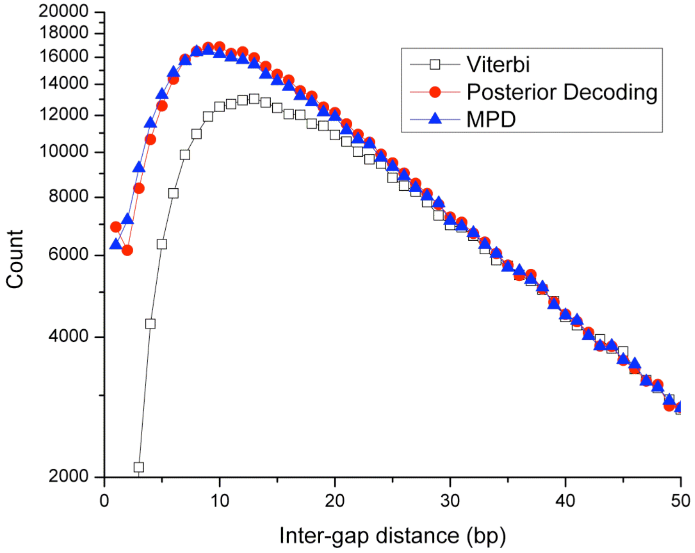


Posteriors: Good predictors of accuracy



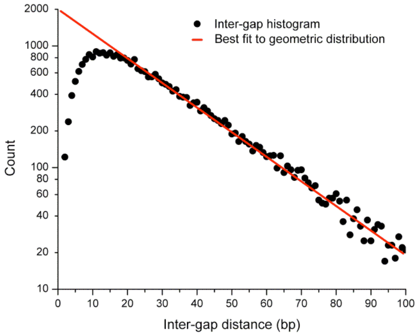
Posterior decoding: better than Max Likelihood



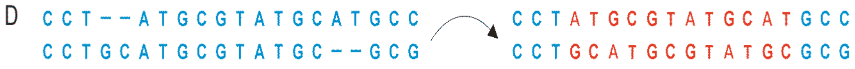


Posteriors & estimating indel rates

The inter-gap histogram slope estimates the indel rate, and is not affected by gap attraction…

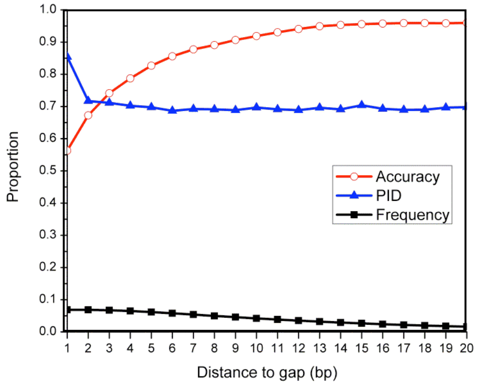


.. but is influenced by gap annihilation…

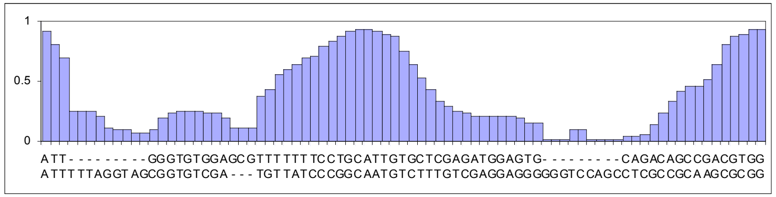


D
C C T - - A T G C G T A T G C A T G C C
C C T G C A T G C G T A T G C - - G C G
C C T A T G C G T A T G C A T G C C
C C T G C A T G C G T A T G C G C G

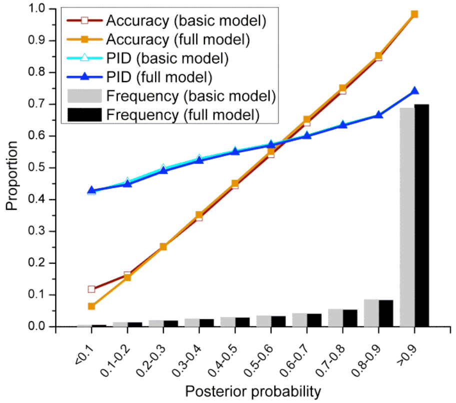
…leading to lower ‘asymptotic accuracy’…



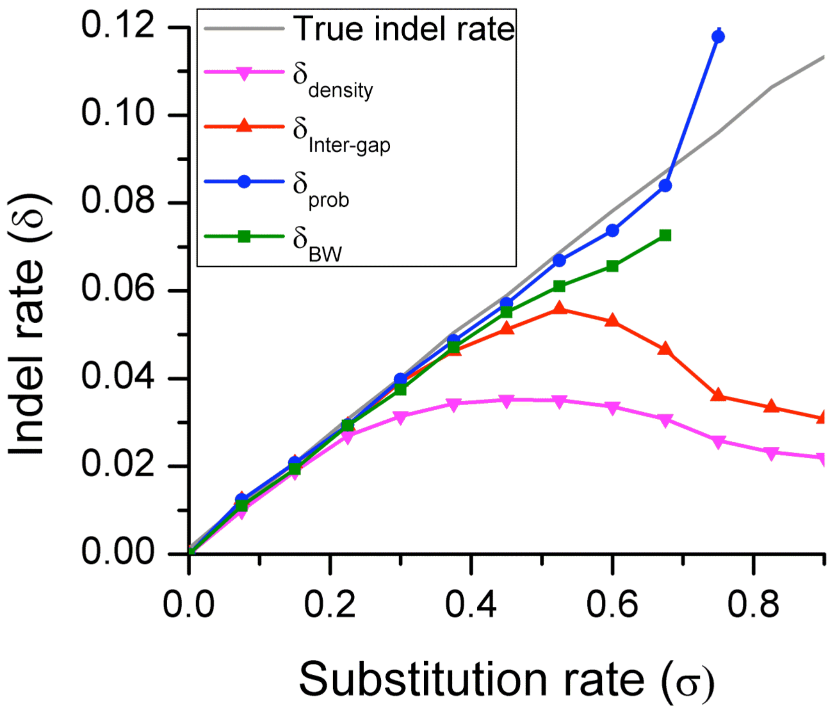
…which cannot be observed – but posteriors can be…



…and they are identical in the mean:

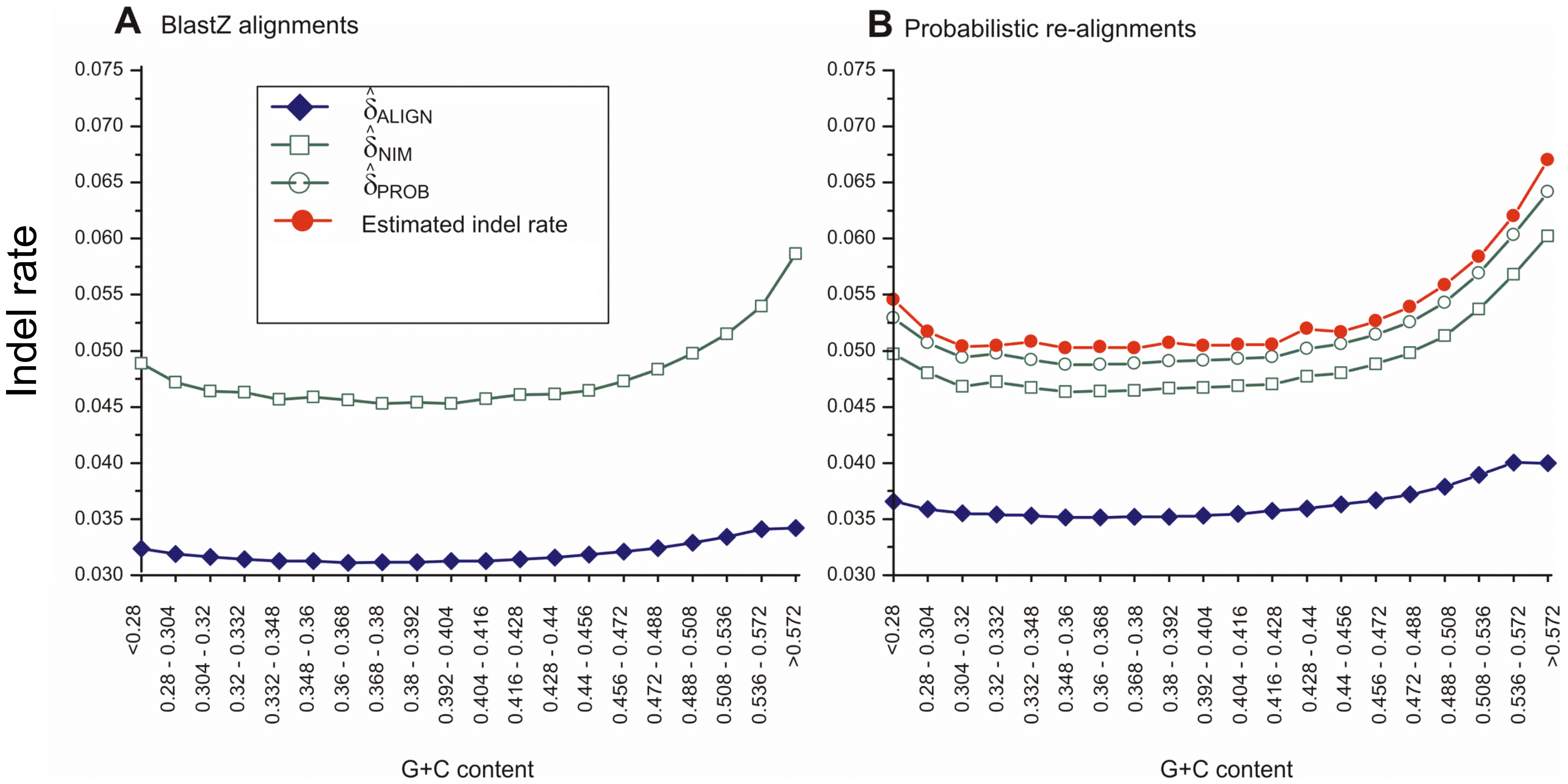


Indel rate estimators

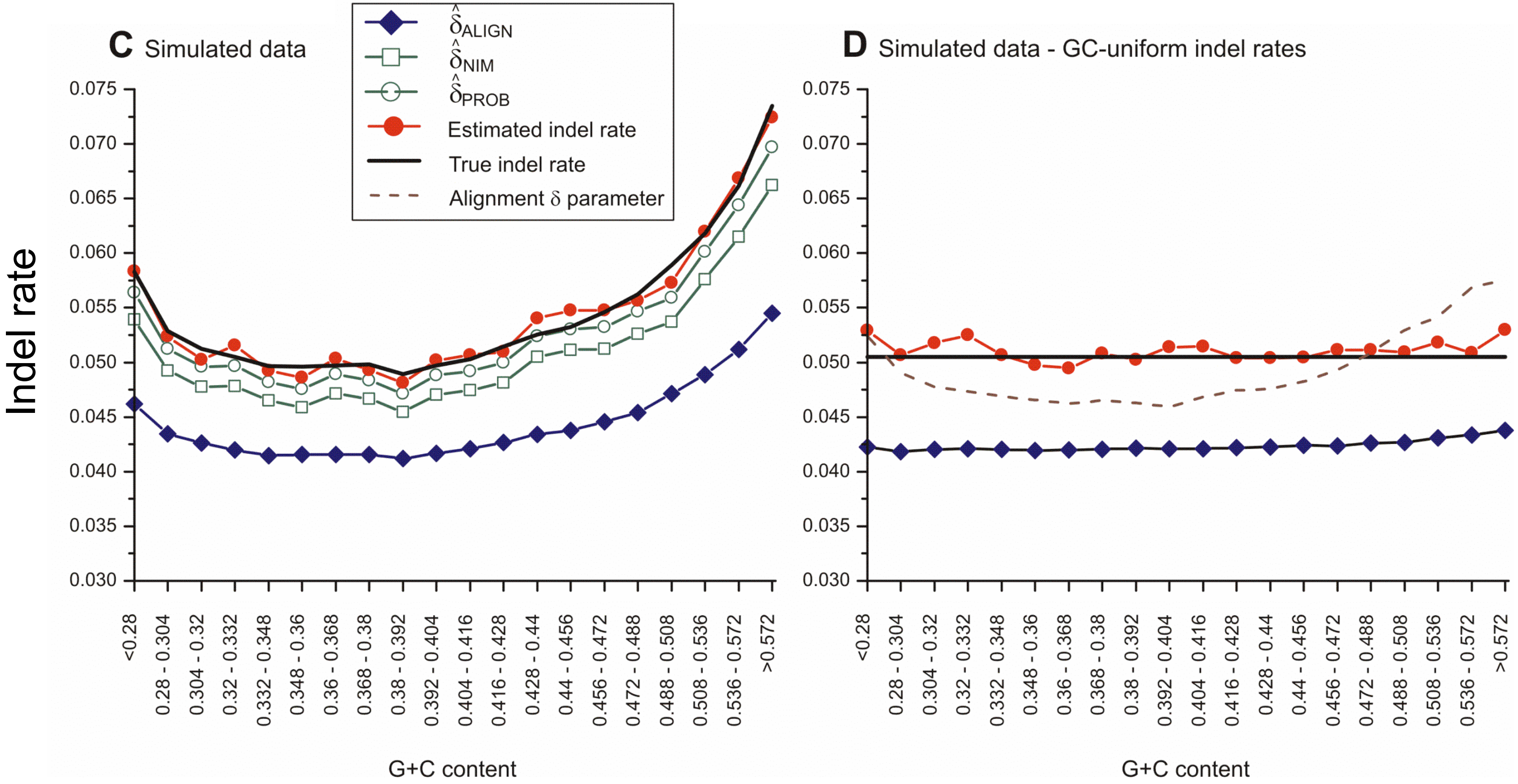


Density: Alignment gaps per site
Inter-gap: Slope of inter-gap histogram
BW: Baum-Welch parameter estimate
Prob: Inter-gap histogram with posterior probability correction

Human-mouse indel rate estimates



Simulations: inferences are accurate

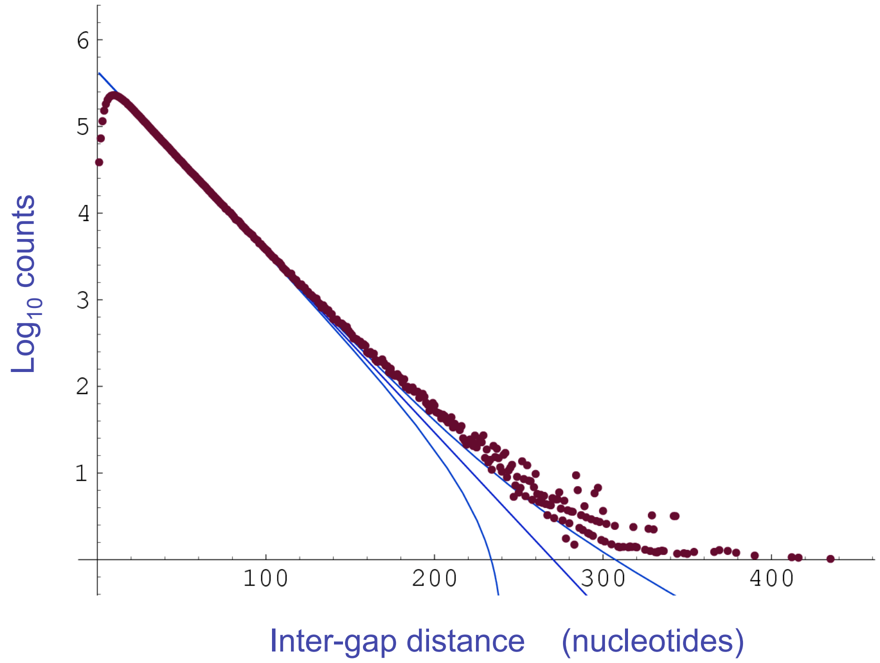


Second summary

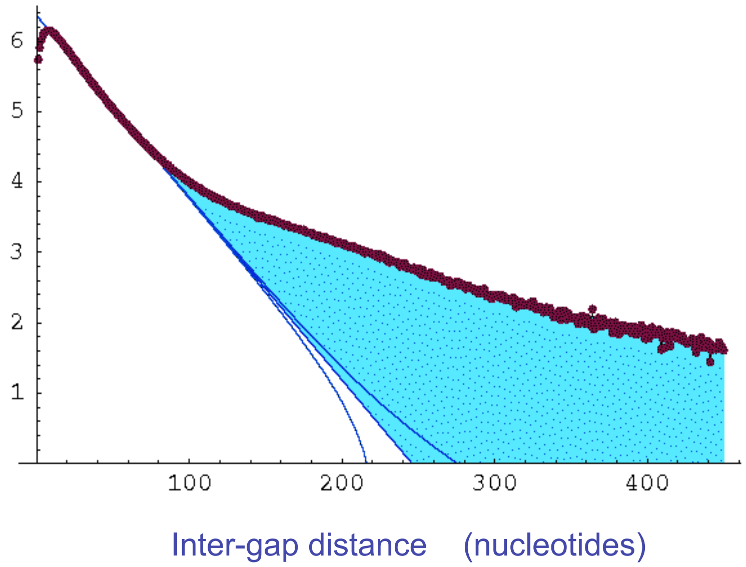
- Alignments are biased, and have errors
- Posterior **accurately predicts** local alignment quality
- Posterior decoding **improves** alignments, **reduces** biases
- With posterior decoding: modelling of indel lengths and sequence content **improves alignments**
- Indel rates (human-mouse) **60-100% higher** than apparent from alignments

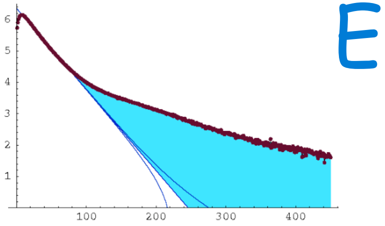
Neutral indel model: Whole genome

Transposable elements:

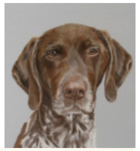


Whole genome:



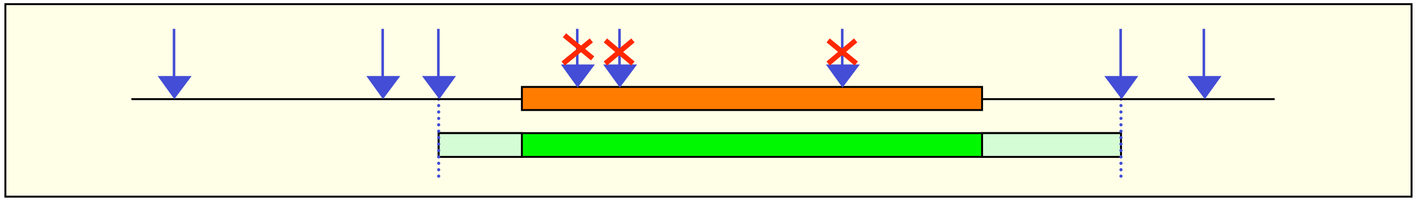


Estimating fraction of sequence under purifying selection



Model:
- Genome is mixture of “conserved” and “neutral” sequence
- “Conserved” sequence accepts no indel mutations
- “Neutral” sequence accepts any indel mutation
- Indels are point events (no spatial extent)

Account for “neutral overhang”:

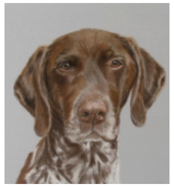


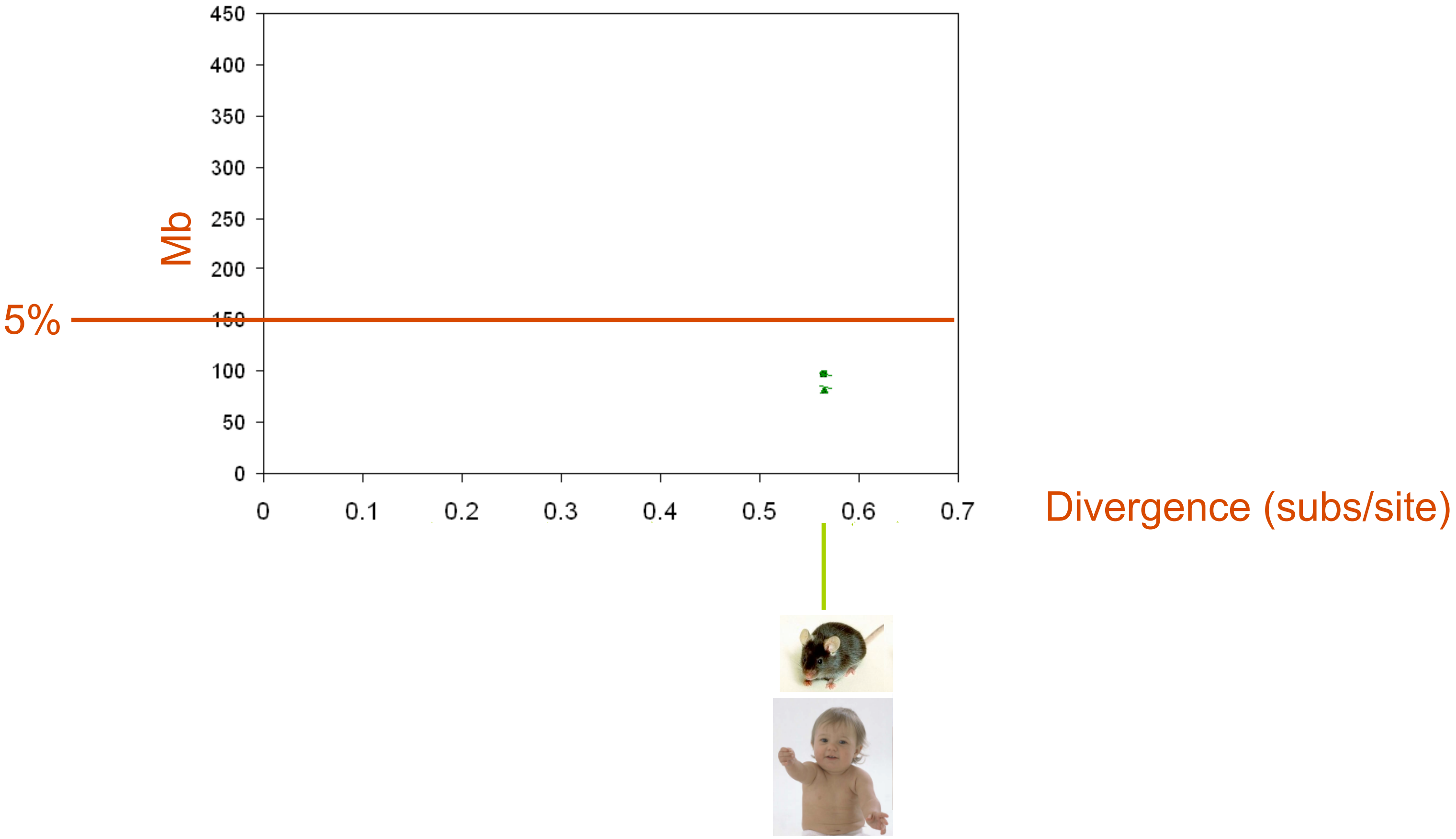
Correction depends on level of clustering of conserved sequence:

- Low clustering: conserved segment is flanked by neutral overhang
  neutral contribution = 2 x average neutral distance between indels
- High clustering: indels “sample” neutral sequence
  neutral contribution = 1 x average neutral distance between indels

Lower bound: **~79 Mb**, or **~2.6 %**
Upper bound: **~100 Mb**, or **~3.25 %**

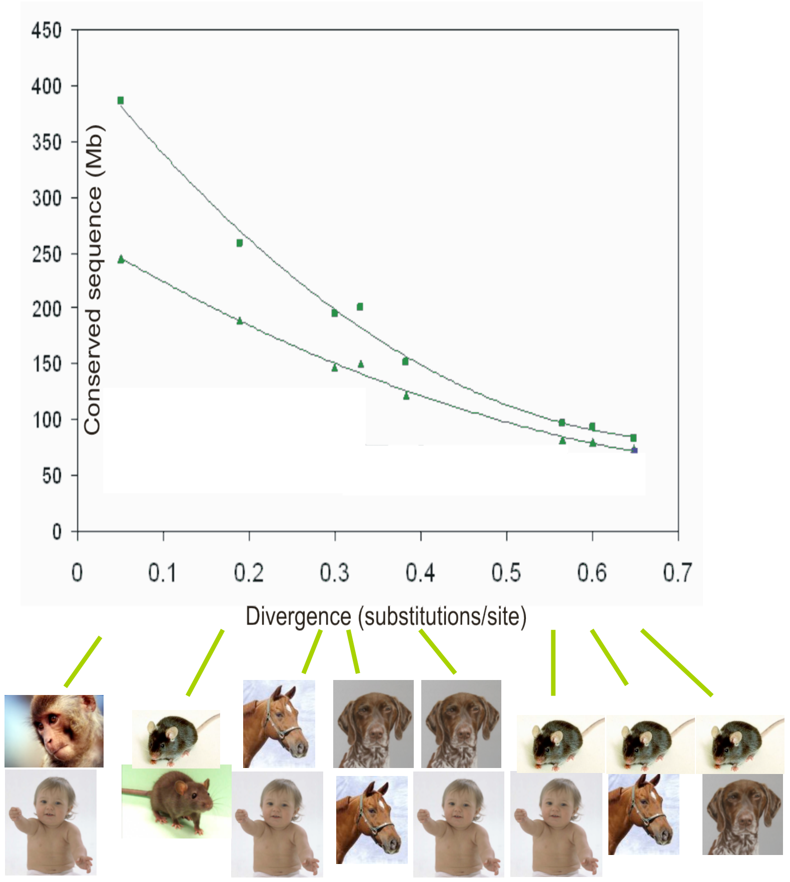
How much of our genome is under purifying selection?

+ +  : 2.56 – 3.25% indel-conserved (79-100 Mb)

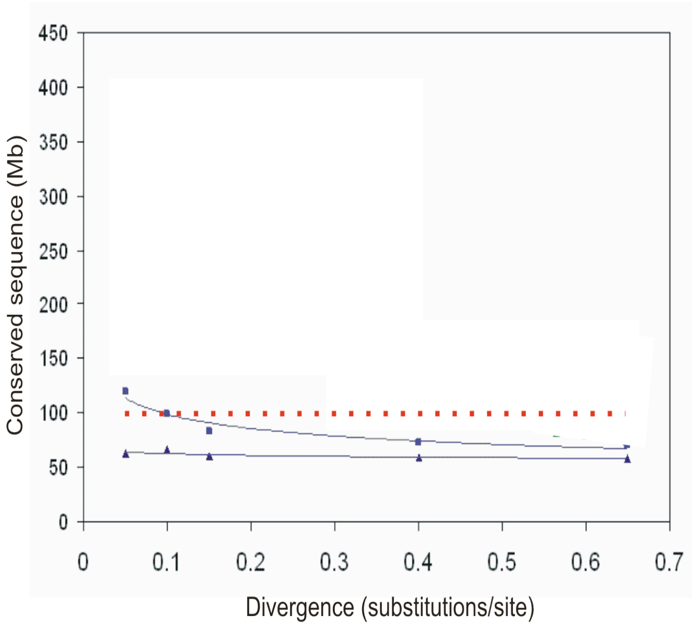


Inferences are not biased by divergence

Inferred from data:



Simulation (100 Mb conserved)



Conclusions

- Alignment is an inference problem; don’t ignore the uncertainties!
- Posterior decoding (heuristic) can be better than Viterbi (exact)
- Indel rates are high. Useful for identifying functional regions, since indels can be more disruptive of function than substitutions.
- Up to 10% of our genome may be functional, and a large proportion is rapidly turning over.