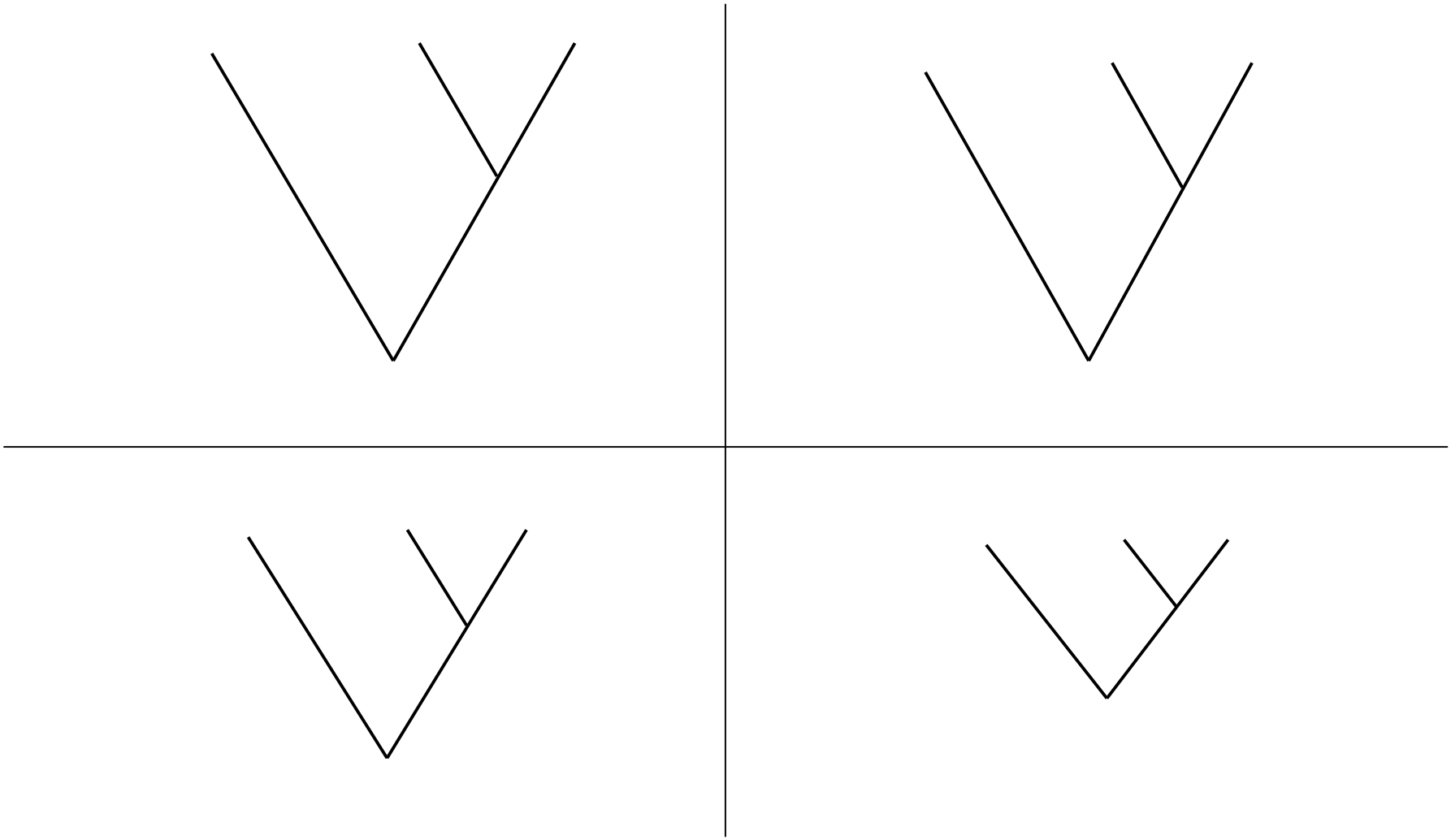
Estimating the contribution of sequence context to nucleotide substitution rate heterogeneity

Helen Lindsay and Gavin A. Huttley

The Gamma Model

- Yang (1993) used a gamma distribution to model rate variation in α- and β-globin genes
- The gamma distribution is often approximated by four equi-probable bins

Gamma rate variation

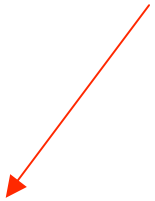


Improvements on the Gamma model

- Allow sites to change rates
- Allow clustering of rates
- Consider other/multiple rate distributions

What causes substitution rate variation?

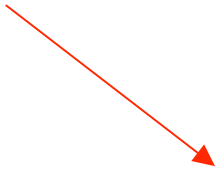
What causes substitution rate variation?



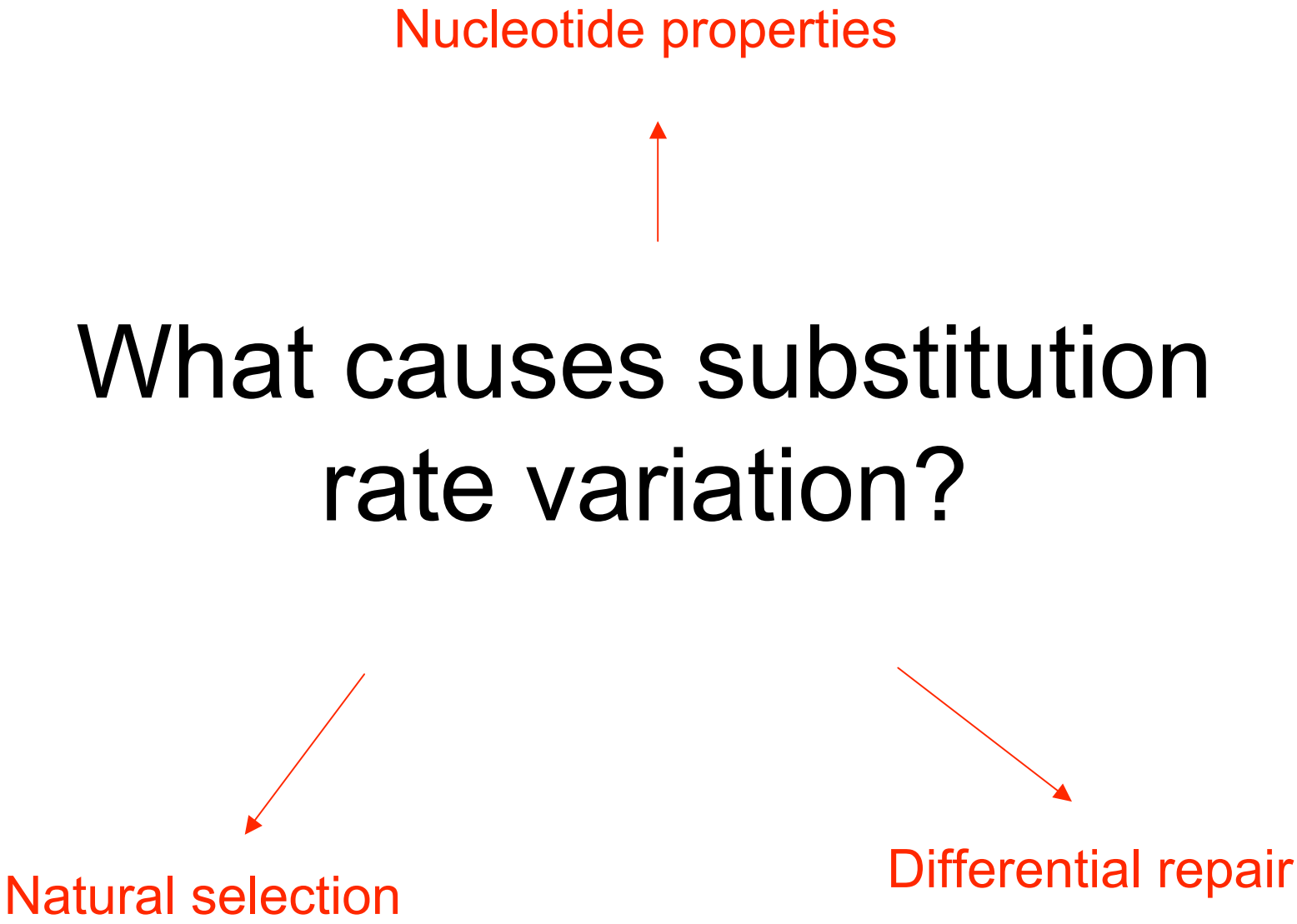
Natural selection

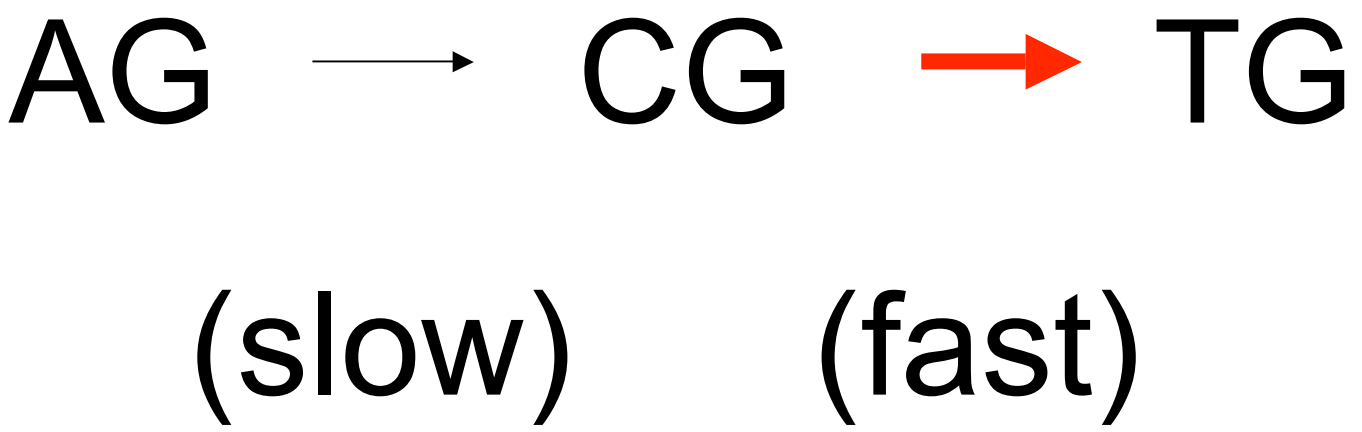
What causes substitution rate variation?

Natural selection



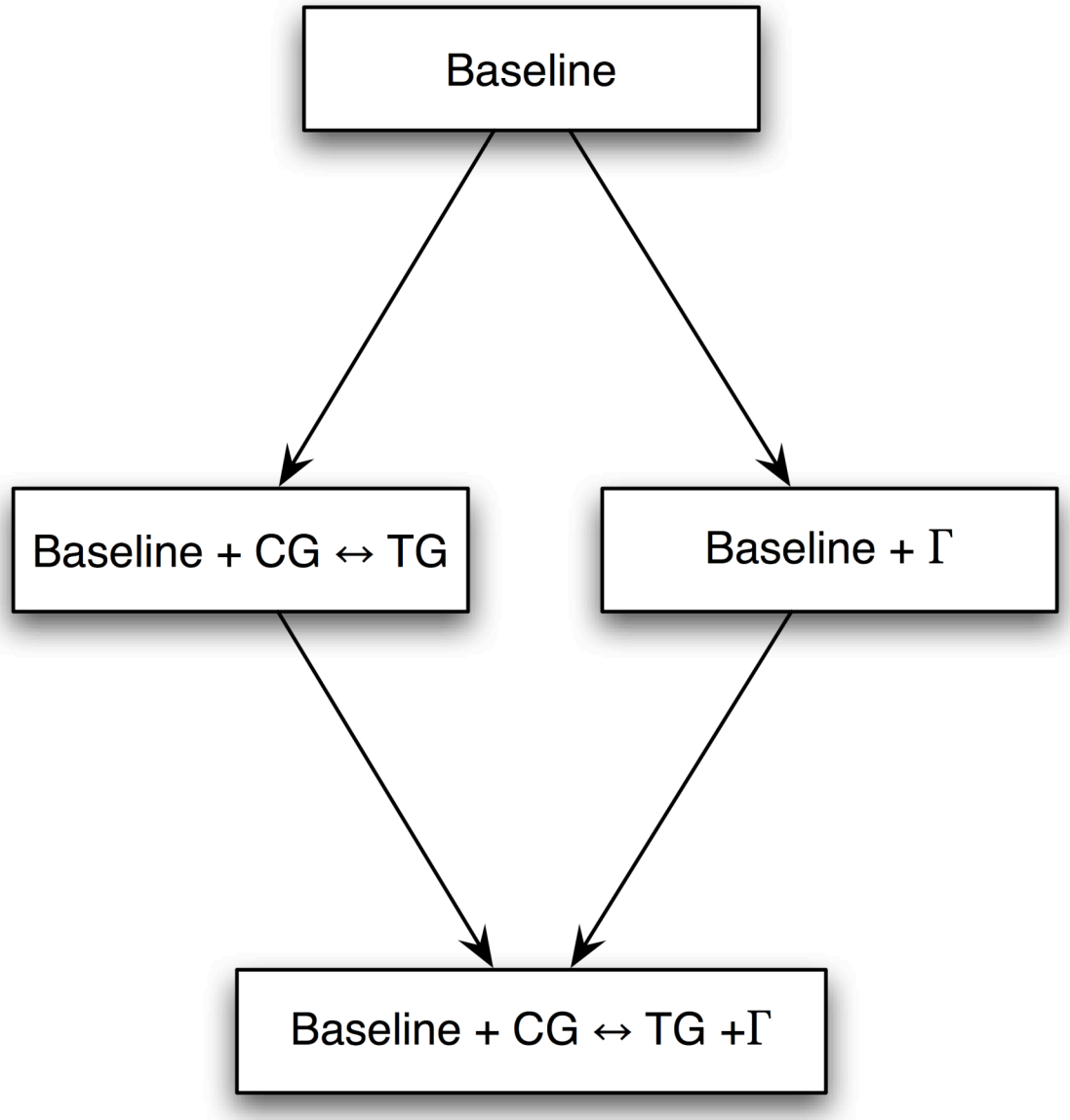
Differential repair


Nucleotide properties
What causes substitution rate variation?
Natural selection
Differential repair


AG
CG
TG
(slow)
(fast)

Data

- 470 alignments, each 50 000 nucleotides long, of introns from human, chimpanzee and macaque one-to-one orthologs.
- Sampled from Ensembl version 49.


Baseline
Baseline + CG ↔ TG
Baseline + Γ
Baseline + CG ↔ TG +Γ

The baseline model

$$q_{i_1 i_2, j_1 j_2} = \begin{cases} 0 & \text{more than one nucleotide difference} \\ \pi_{j_x} r_{A \leftrightarrow C} & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } A \leftrightarrow C \\ \pi_{j_x} r_{A \leftrightarrow G} & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } A \leftrightarrow G \\ \pi_{j_x} r_{A \leftrightarrow T} & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } A \leftrightarrow T \\ \pi_{j_x} r_{C \leftrightarrow G} & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } C \leftrightarrow G \\ \pi_{j_x} r_{C \leftrightarrow T} & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } C \leftrightarrow T \end{cases}$$

where x is the index at which $i_1 i_2$ and $j_1 j_2$ differ

The CpG model

$$q_{i_1 i_2, j_1 j_2} = \begin{cases} 0 & \text{more than one nucleotide difference} \\ \pi_{j_x} r_{A \leftrightarrow C} & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } A \leftrightarrow C \\ \pi_{j_x} r_{A \leftrightarrow G} & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } A \leftrightarrow G \\ \pi_{j_x} r_{A \leftrightarrow T} & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } A \leftrightarrow T \\ \pi_{j_x} r_{C \leftrightarrow G} & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } C \leftrightarrow G \\ \pi_{j_x} r_{C \leftrightarrow T} & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } C \leftrightarrow T, \{i, j\} \neq \{\text{CpG, TpG}\} \\ \pi_{j_x} r_{C \leftrightarrow T} r_{CG \leftrightarrow TG} & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } C \leftrightarrow T, \{i, j\} = \{\text{CpG, TpG}\} \end{cases}$$

where x is the index at which $i_1 i_2$ and $j_1 j_2$ differ

The Gamma Model

$$q_{i_1 i_2, j_1 j_2} = \begin{cases} 0 & \text{more than one nucleotide difference} \\ \pi_{j_x} r_{A \leftrightarrow C} \Gamma_n & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } A \leftrightarrow C \\ \pi_{j_x} r_{A \leftrightarrow G} \Gamma_n & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } A \leftrightarrow G \\ \pi_{j_x} r_{A \leftrightarrow T} \Gamma_n & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } A \leftrightarrow T \\ \pi_{j_x} r_{C \leftrightarrow G} \Gamma_n & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } C \leftrightarrow G \\ \pi_{j_x} r_{C \leftrightarrow T} \Gamma_n & i_1 i_2 \text{ and } j_1 j_2 \text{ differ by } C \leftrightarrow T \end{cases}$$

where x is the index at which $i_1 i_2$ and $j_1 j_2$ differ
and Γ_n is the rate category of site n.

Gamma vs Dinucleotide models

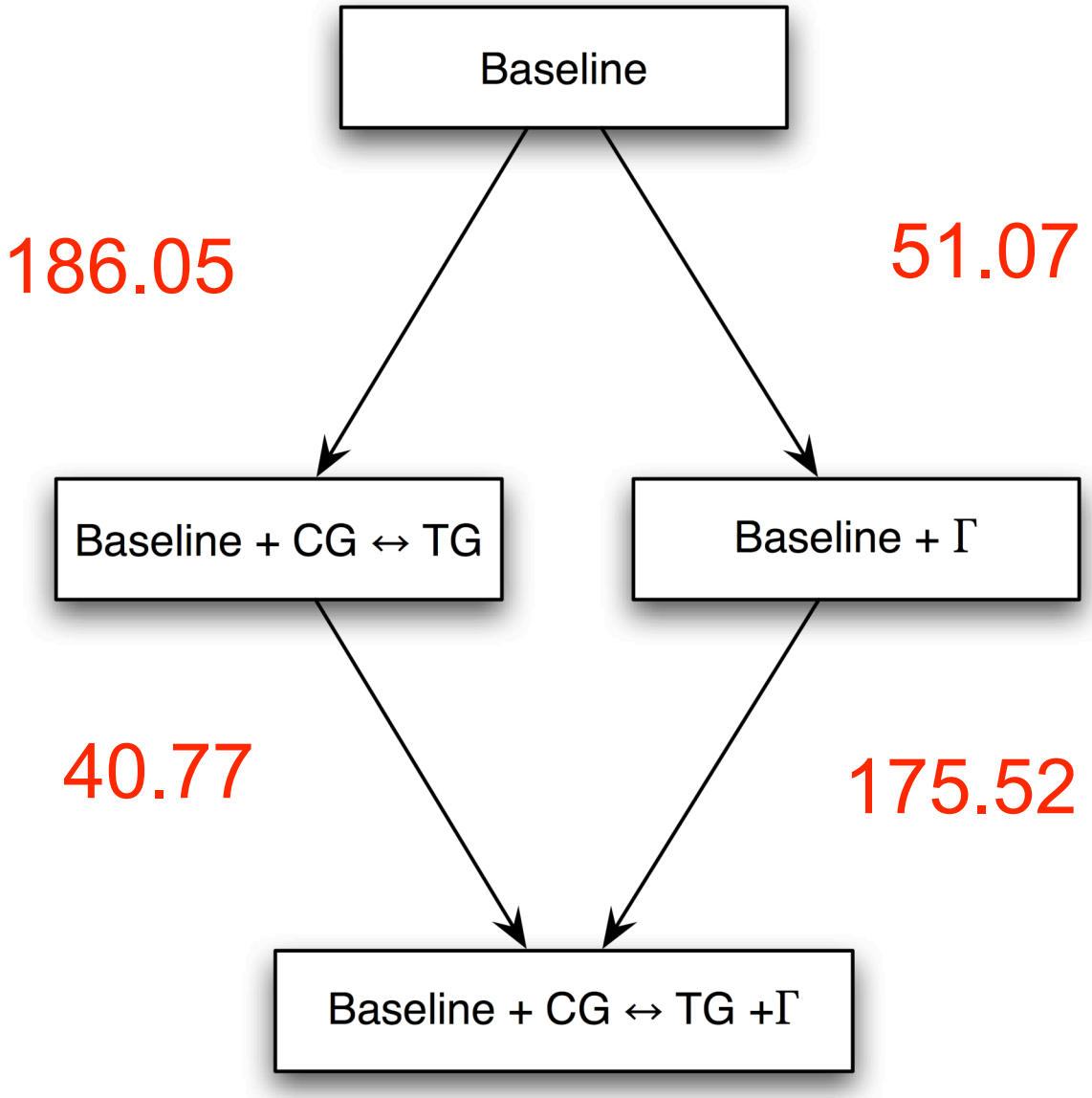
Parameter (P)	**LR** $(\frac{baseline+P}{baseline})$	**LR** $(\frac{baseline+P+\Gamma}{baseline+\Gamma})$	**LR** $(\frac{baseline+P+\Gamma}{baseline+P})$
baseline	-	-	51.07
$TG \leftrightarrow CG$ or $CA \leftrightarrow CG$	186.05	175.52	40.77
$TG \leftrightarrow CG$	89.53	84.79	46.07
$CA \leftrightarrow CG$	84.74	80.08	45.86
$AT \leftrightarrow GT$	28.83	28.01	50.09
$AA \leftrightarrow GA$	25.51	24.66	50.05
$TT \leftrightarrow CT$	23.24	22.48	50.23
$AA \leftrightarrow AG$	19.22	18.31	50.20
$TT \leftrightarrow TC$	15.41	14.63	50.22
$GA \leftrightarrow GG$	14.18	13.53	50.44
$TT \leftrightarrow CC$	10.80	10.43	50.78
$AT \leftrightarrow AC$	9.85	9.33	50.44

Gamma vs Dinucleotide models

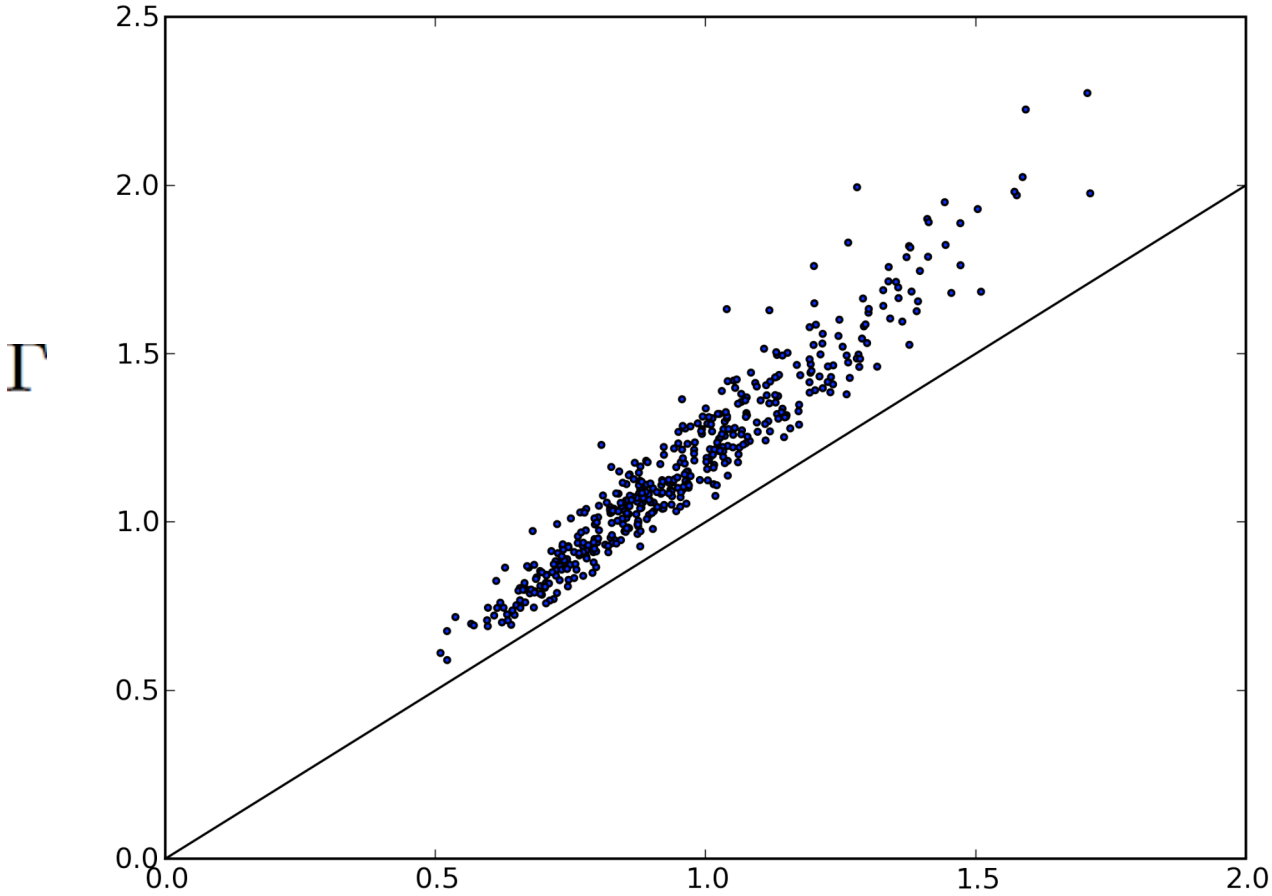
Parameter (P)	**LR** $(\frac{baseline+P}{baseline})$	**LR** $(\frac{baseline+P+\Gamma}{baseline+\Gamma})$	**LR** $(\frac{baseline+P+\Gamma}{baseline+P})$
baseline	-	-	51.07
$TG \leftrightarrow CG$ or $CA \leftrightarrow CG$	186.05	175.52	40.77
$TG \leftrightarrow CG$	89.53	84.79	46.07
$CA \leftrightarrow CG$	84.74	80.08	45.86
$AT \leftrightarrow GT$	28.83	28.01	50.09
$AA \leftrightarrow GA$	25.51	24.66	50.05
$TT \leftrightarrow CT$	23.24	22.48	50.23
$AA \leftrightarrow AG$	19.22	18.31	50.20
$TT \leftrightarrow TC$	15.41	14.63	50.22
$GA \leftrightarrow GG$	14.18	13.53	50.44
$TT \leftrightarrow CC$	10.80	10.43	50.78
$AT \leftrightarrow AC$	9.85	9.33	50.44

Gamma vs Dinucleotide models

Parameter (P)	**LR** $(\frac{baseline+P}{baseline})$	**LR** $(\frac{baseline+P+\Gamma}{baseline+\Gamma})$	**LR** $(\frac{baseline+P+\Gamma}{baseline+P})$
baseline	-	-	51.07
$TG \leftrightarrow CG$ or $CA \leftrightarrow CG$	186.05	175.52	40.77
$TG \leftrightarrow CG$	89.53	84.79	46.07
$CA \leftrightarrow CG$	84.74	80.08	45.86
$AT \leftrightarrow GT$	28.83	28.01	50.09
$AA \leftrightarrow GA$	25.51	24.66	50.05
$TT \leftrightarrow CT$	23.24	22.48	50.23
$AA \leftrightarrow AG$	19.22	18.31	50.20
$TT \leftrightarrow TC$	15.41	14.63	50.22
$GA \leftrightarrow GG$	14.18	13.53	50.44
$TT \leftrightarrow CC$	10.80	10.43	50.78
$AT \leftrightarrow AC$	9.85	9.33	50.44

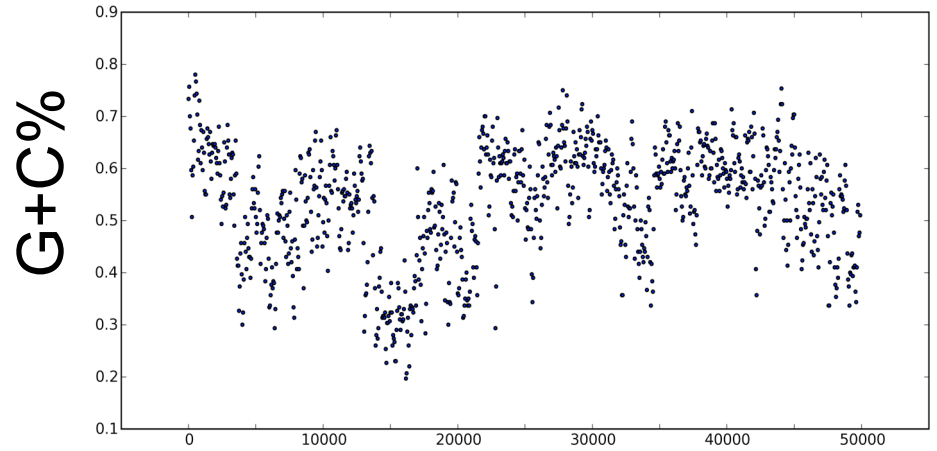

Baseline
186.05
51.07
Baseline + CG ↔ TG
Baseline + Γ
40.77
175.52
Baseline + CG ↔ TG +Γ

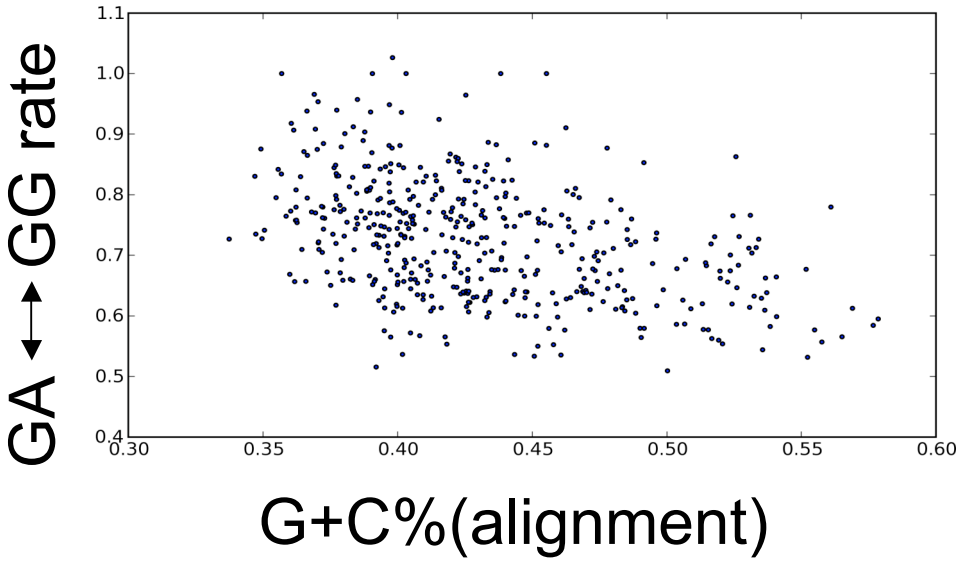
Accounting for CpG substitutions decreases rate variation



$\Gamma + CpG \leftrightarrow TpG$ symmetric

- Independent sites
- Reversible
- Compositional variance





Advantages of dinucleotide models

- Less likelihood computation
- Equivalently parameter-rich
- No assumed distribution of rate variation
- Can incorporate known mutation biases, for example deamination of methylated cytosine.
- Smaller alphabet than amino acids

Acknowledgements

Australian National University

- Gavin Huttley
- Hua Ying

University of Singapore

- Von Bing Yap