DNA Insertions and Deletions in the Human Genome

Philipp W. Messer

Max Planck Institute
for Molecular Genetics

Genetic Variation

```
CGACAATAGCGCTCTTACTACGTGTATCG
|||||||:|||||   |||||||||:||||
CGACAATGGCGCT---ACTACGTGCATCG
```

1. Nucleotide mutations
2. Genomic rearrangements
3. **DNA insertions / deletions (indels)**

Outline

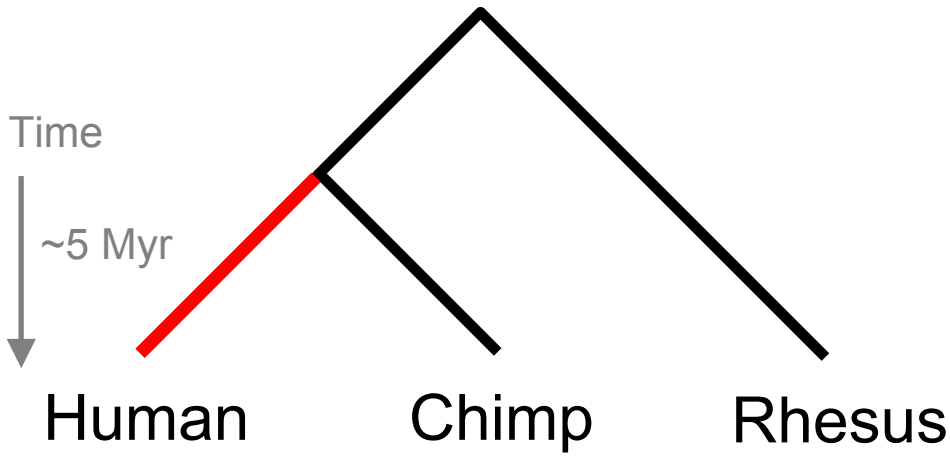
1. Origin and characteristics of indels
2. Indels in protein-coding regions
3. Duplications and genomic correlations
4. Duplications and alignment scores

Part 1

Origin and Characteristics of Indels in the Human Genome

[Messer and Arndt, *Mol Biol Evol* 2007]

Identifying Insertions/Deletions

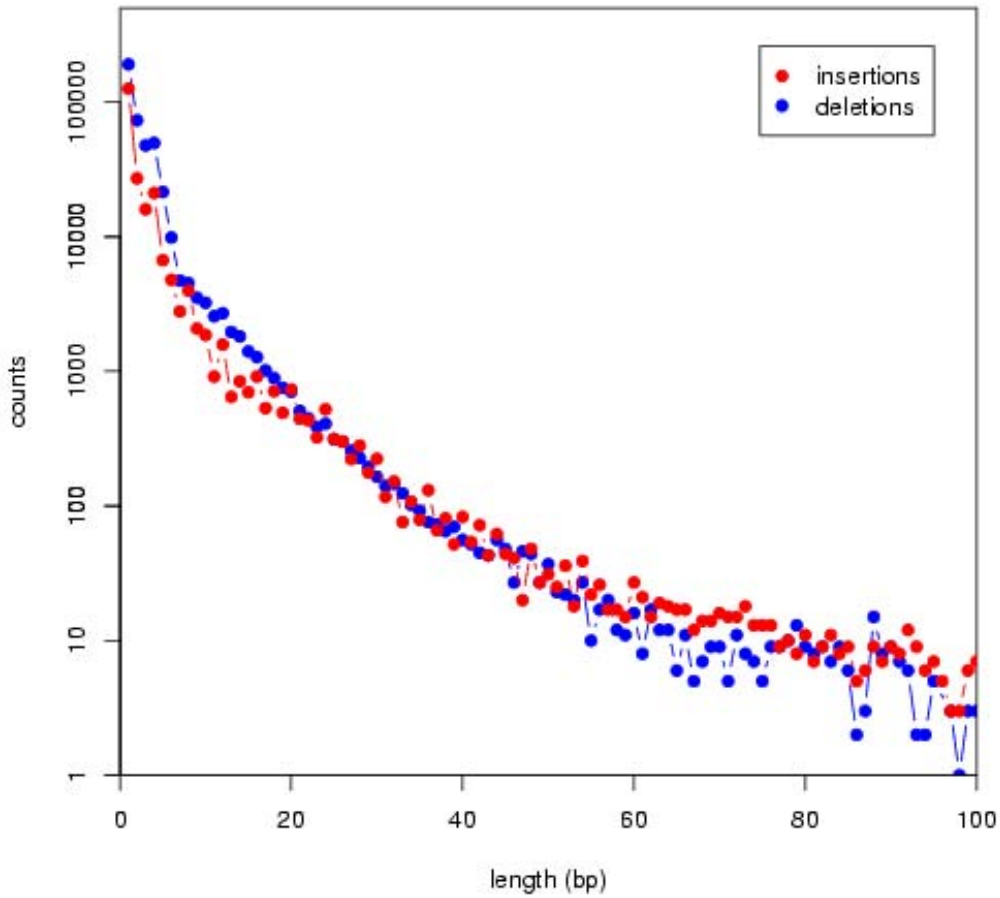


High-quality flanks

```
H:  ..ATACCTCGTACAATAGCGCTGGTACAGATA..
C:  ..ATACCTCGTACAAT--CGCTGGTACAGGTA..
R:  ..ATACCTCGTACAAT--CGCTGCTACAGATA..
```

Insertions can be distinguished from deletions (parsimony)

Insertion/Deletion Statistics



Insertions:

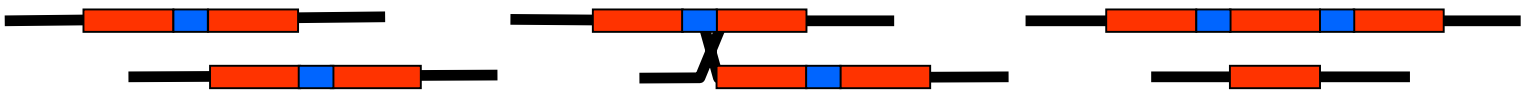
- ~250 000 events
- 15% SSR

Deletions:

- ~450 000 events
- 5% SSR

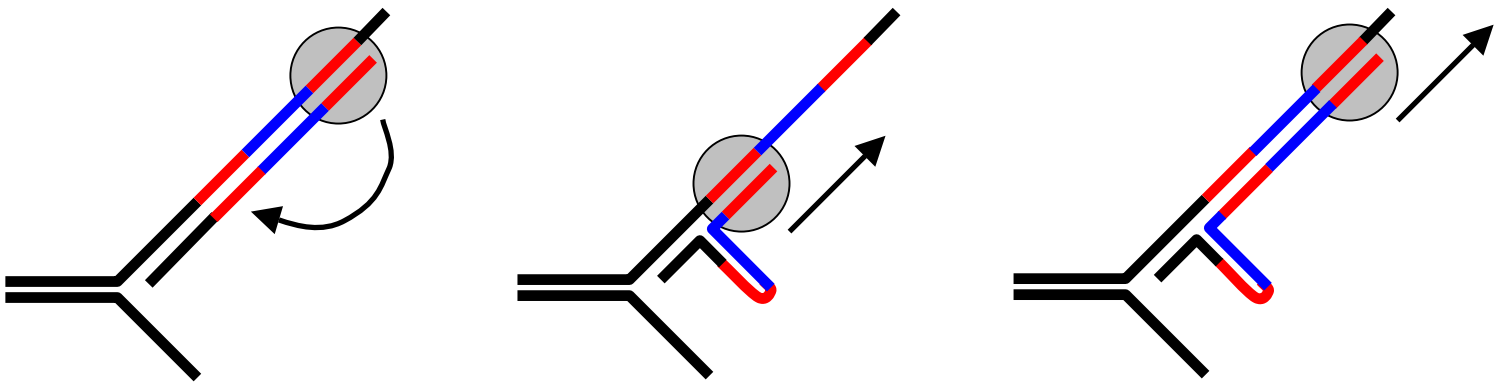
Molecular Mechanisms

1. Unequal Crossing Over (UCO)



nonhomologous recombination

2. Replication Slippage (RS)

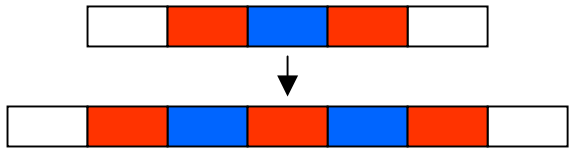


slipped-strand mispairing

Indel Signatures for UCO and RS

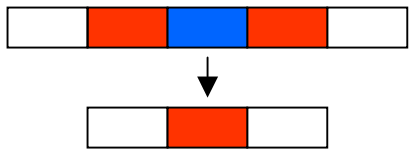
Insertions:

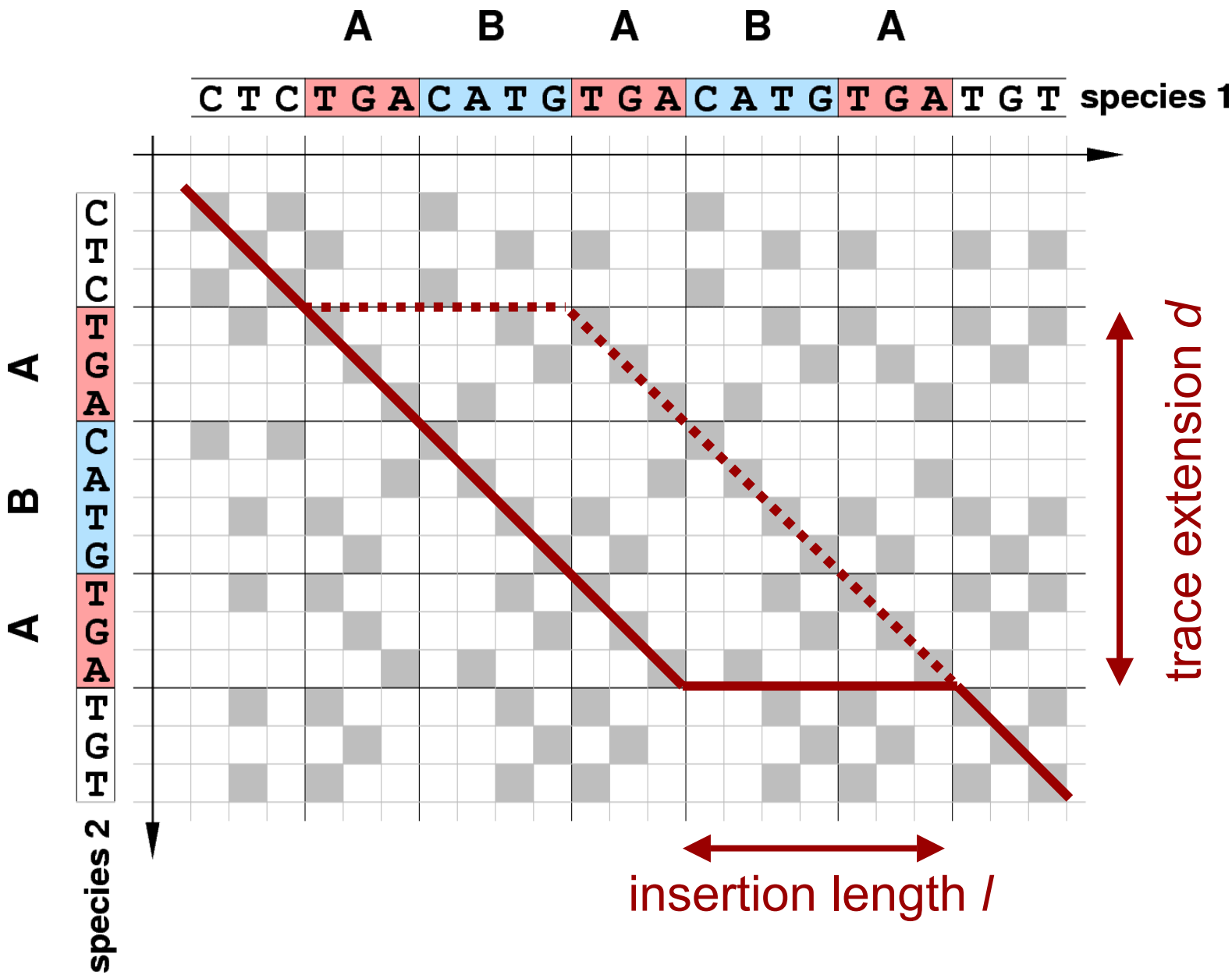
Tandem duplications of preexisting duplicates



Deletions:

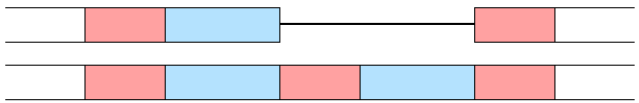
Remove one copy of preexisting duplicates




A B A B A
C T C T G A C A T G T G A C A T G T G A T G T
species 1
A B A
C T C T G A C A T G T G A
species 2
trace extension d
insertion length l

Indel Trace Extensions

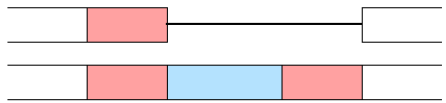
a)



d>l

UCO, RS (insertion)

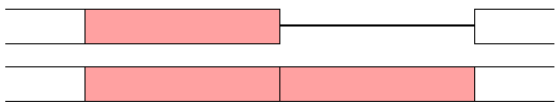
b)



0<d<l

UCO, RS (deletion)

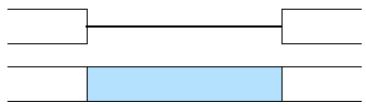
c)



d=l

Tandem duplication

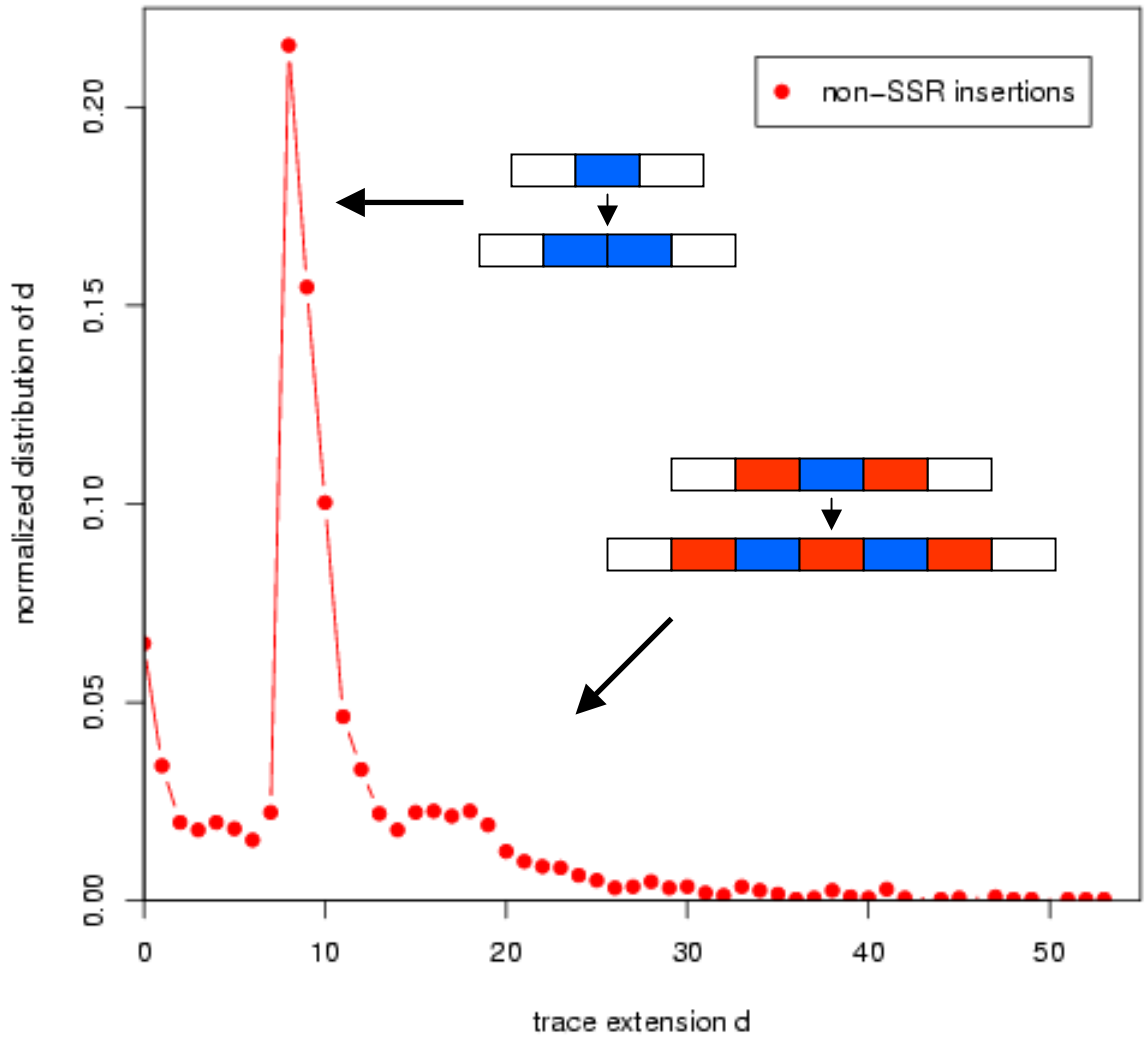
d)



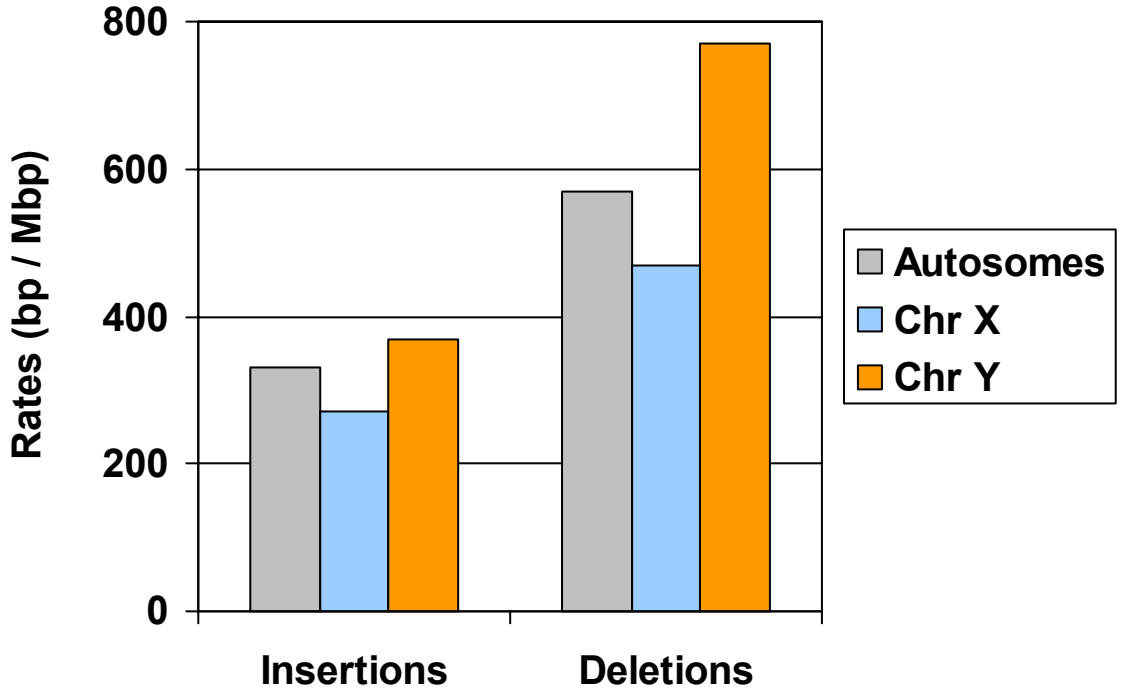
d=0

Random indel

Measured Trace Extensions (*l*=8 bp)



Chromosomal Rate Differences

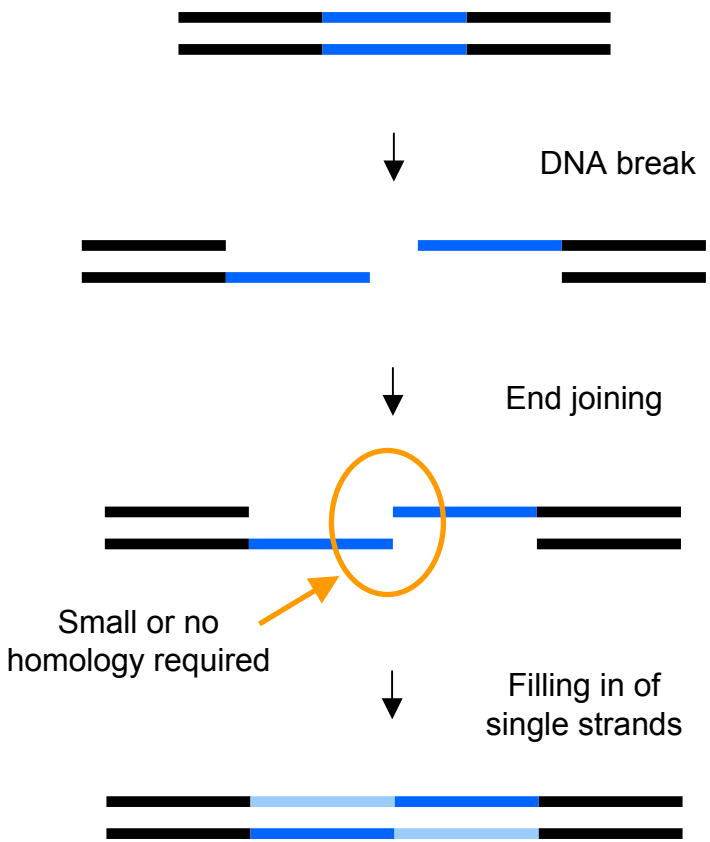


1. Indels occur preferentially in the male germline
2. Indels are not recombination-mediated

Indel Characteristics

1. The majority of insertions are tandem duplications
2. Long preexisting duplicates are often missing
3. Indels occur preferentially in the male germline
4. Indels are not recombination-mediated

Nonhomologous End Joining

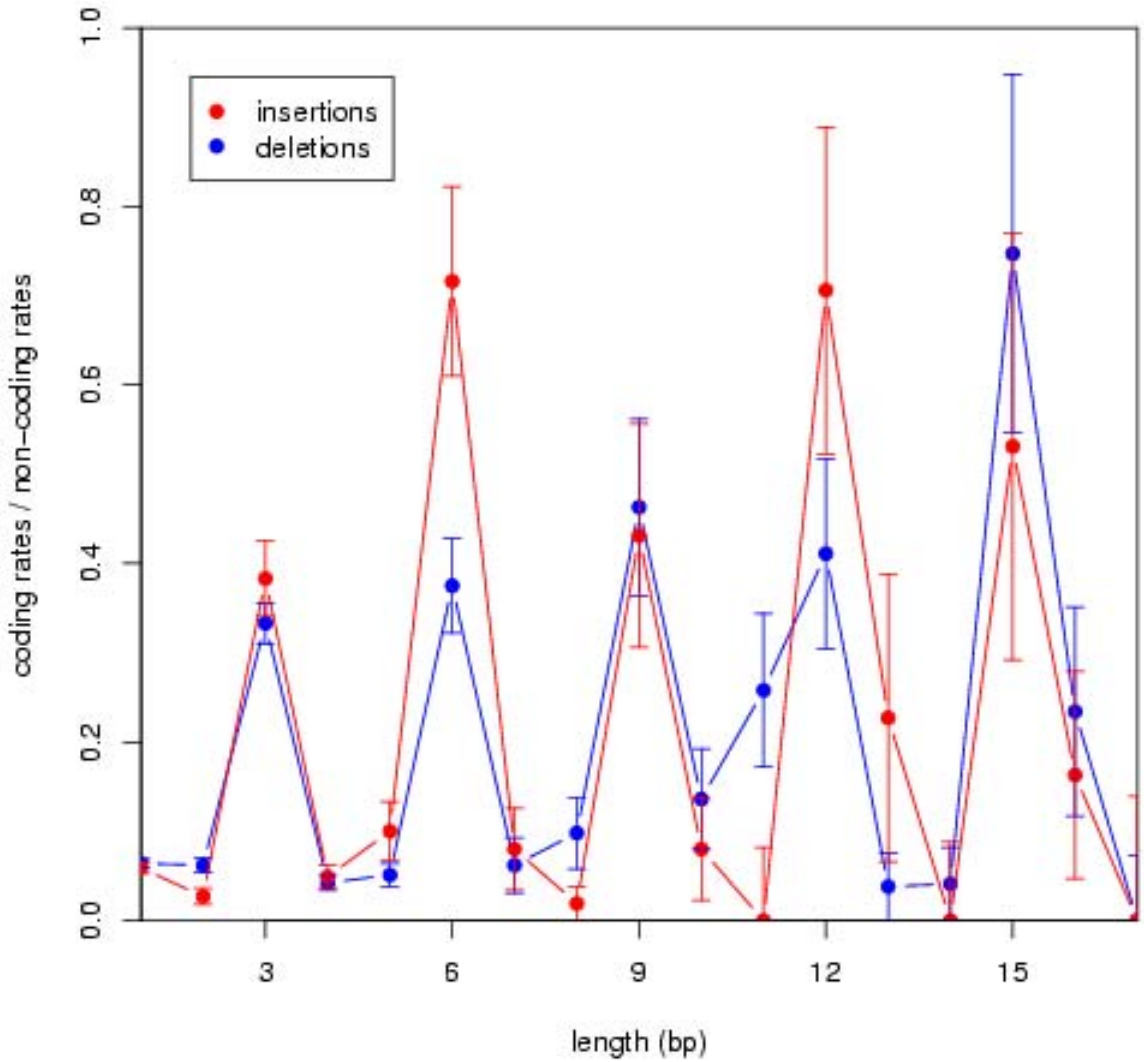


Part 2

Indels in Protein-coding Regions of the Human Genome

[Chaux, Messer, Arndt, *BMC Evol Biol* (submitted)]

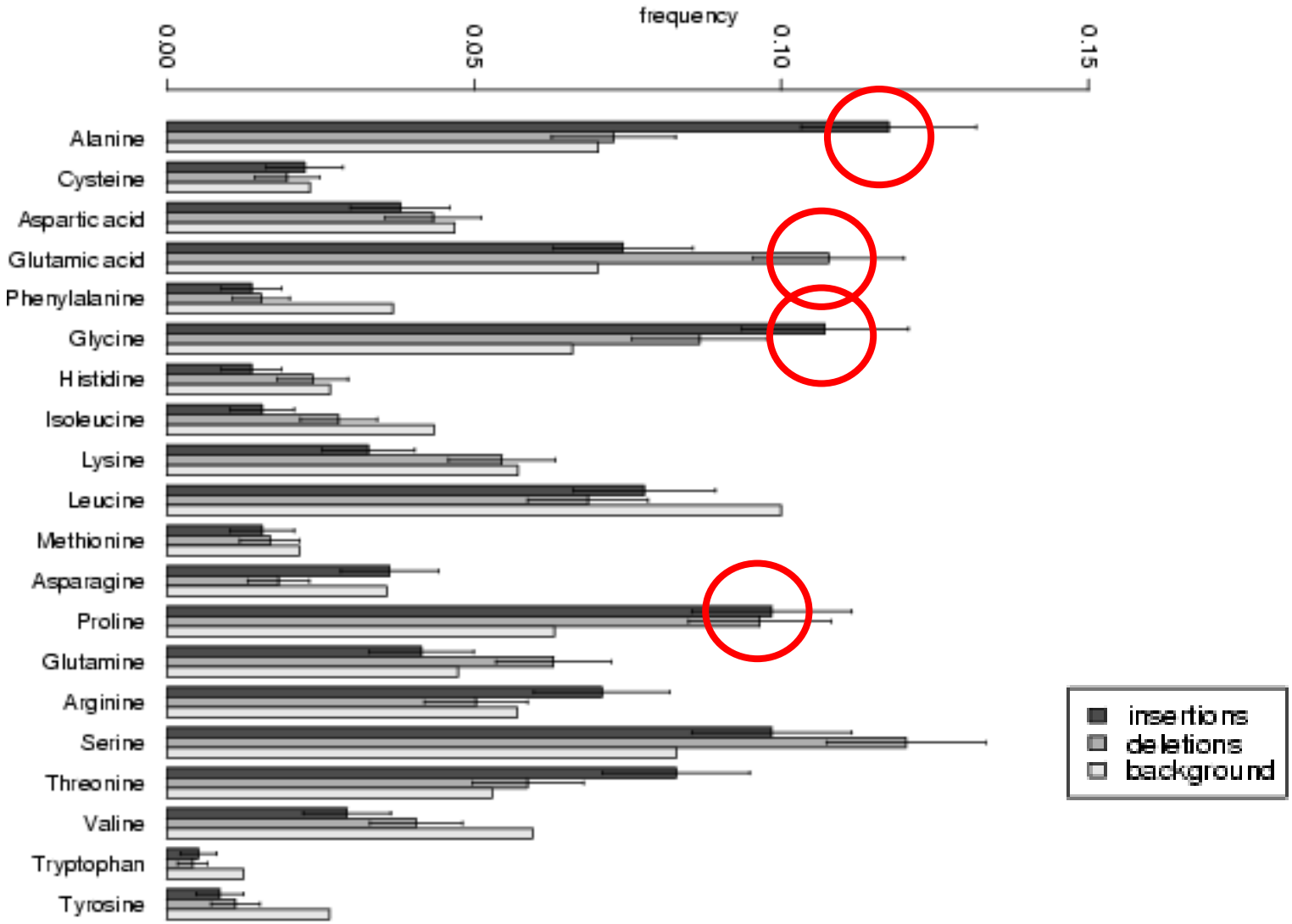
Indel Rates in Coding Regions



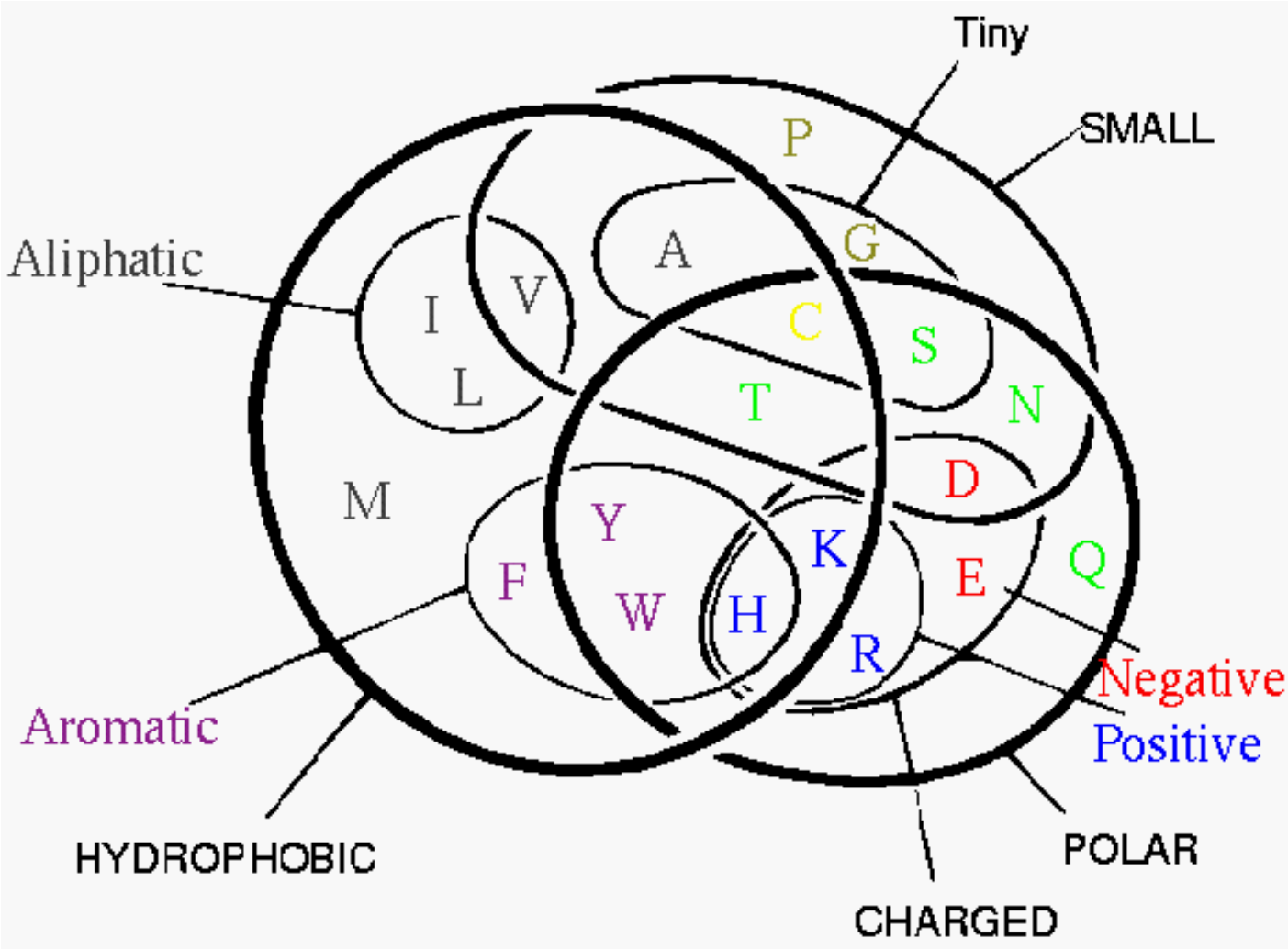
Genetic Code

	U	C	A	G	
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

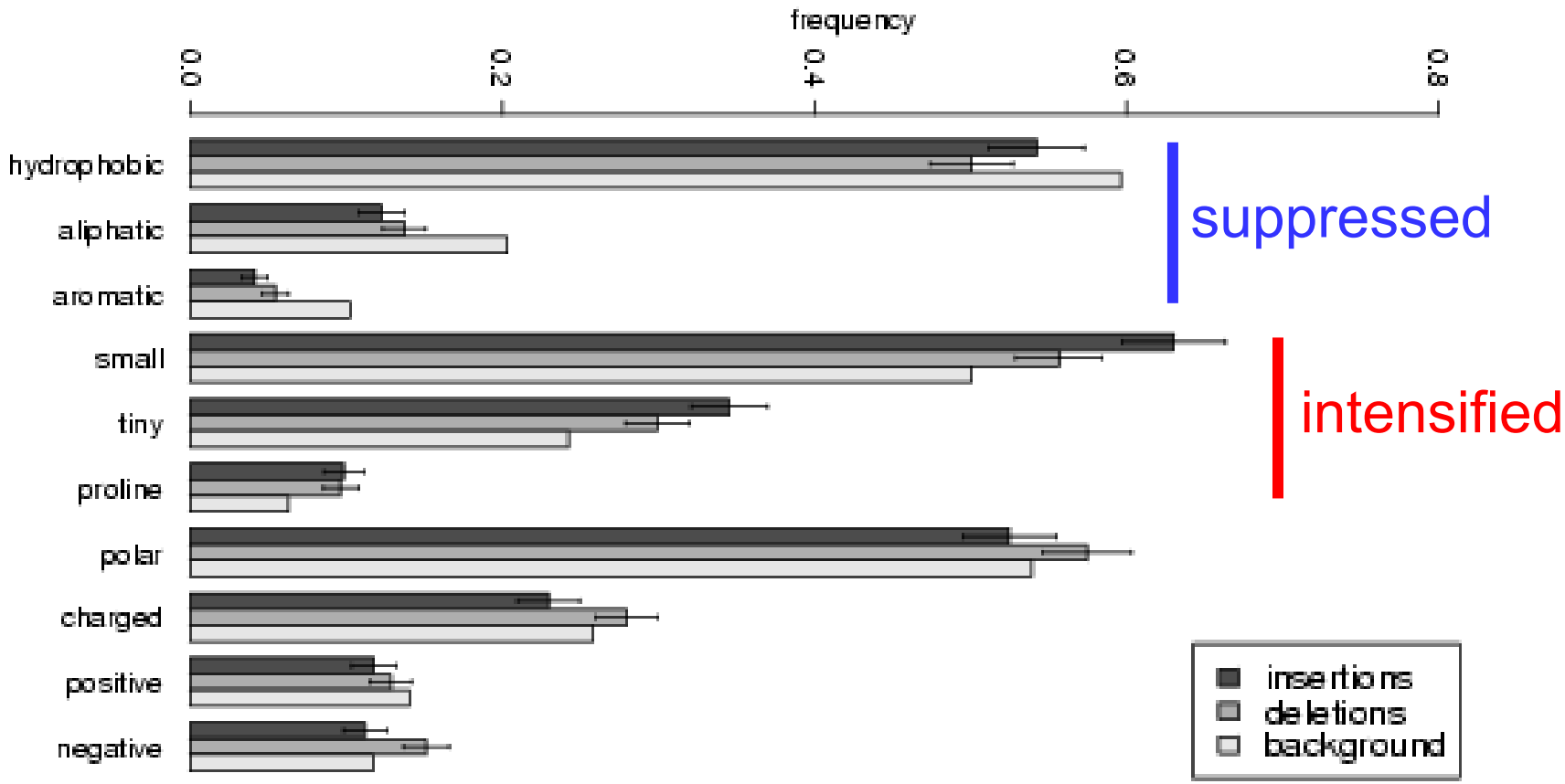
Inserted/Deleted Amino Acids



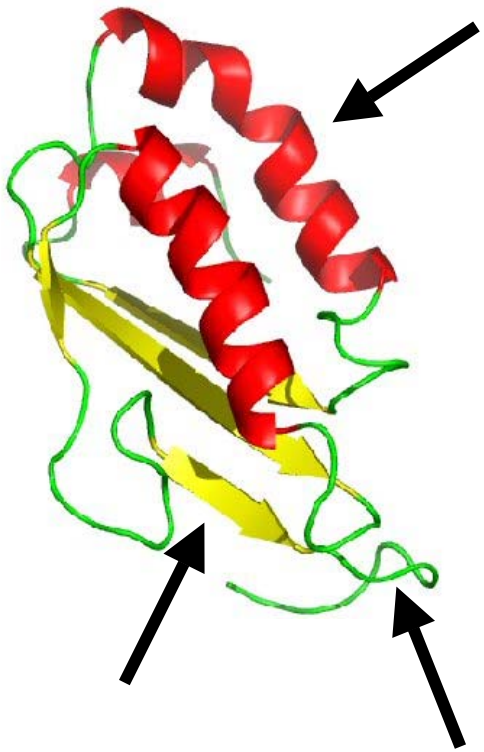
Physico-chemical Properties

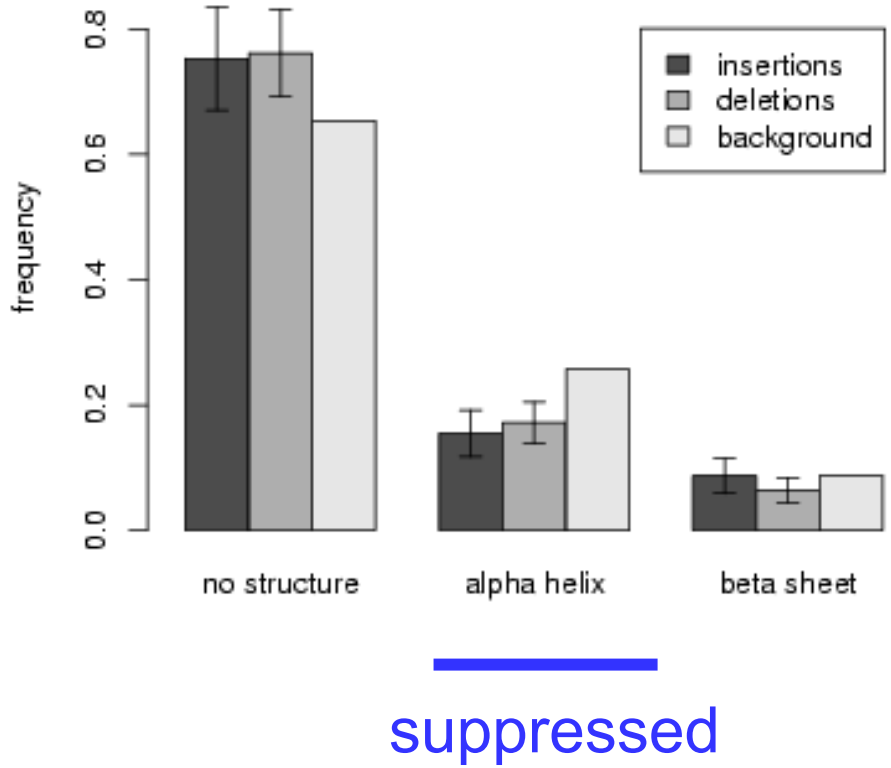


Physico-chemical Properties



Protein Secondary Structure





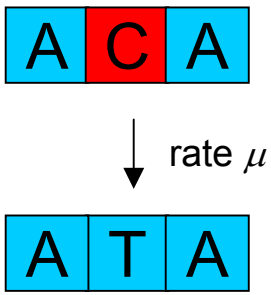
Part 3

Tandem Duplications and Genomic Correlations

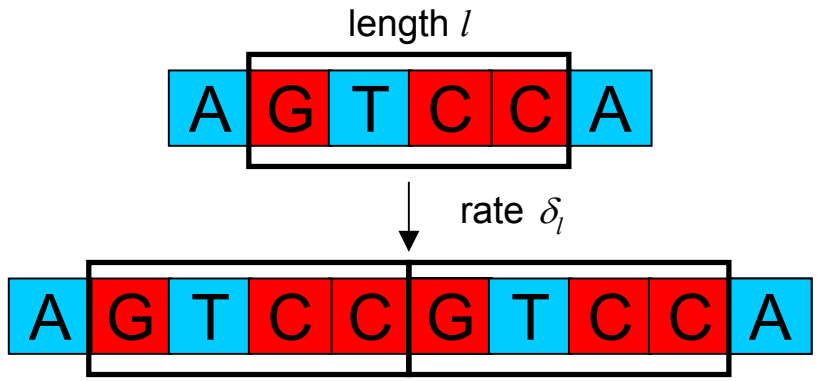
[Messer, Arndt, Lässig, *Phys Rev Lett* 2005]
[Messer, Lässig, Arndt, *J Stat Mech* 2005]
[Messer, Arndt, *Nucleic Acid Res* 2006]

Sequence Evolution Model

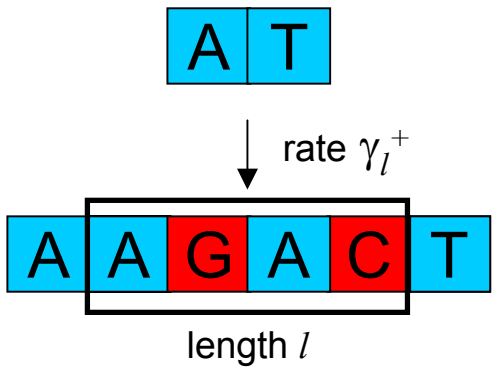
Mutation



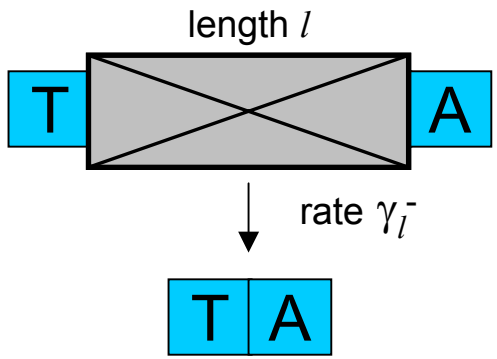
Segmental Duplication



Random Insertion



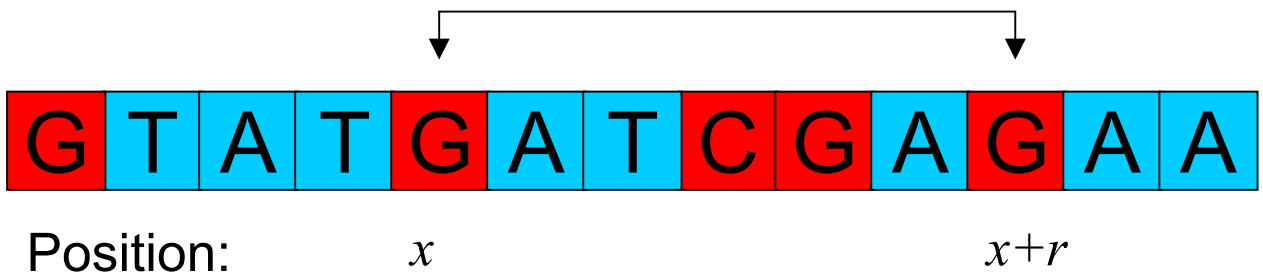
Segmental Deletion



The Correlation Function

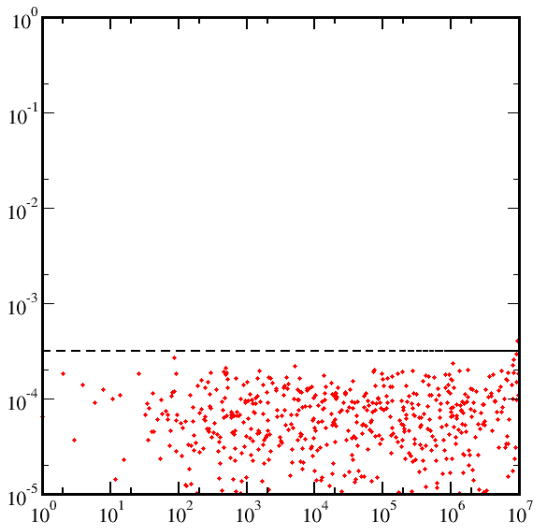
Measures likelihood of finding two G/C base pairs separated by a distance *r* along the genome

$$C(r) = P_{\mathrm{G/C}}(x, x+r) - P^2_{\mathrm{G/C}}(x)$$



Types of Correlation Behavior

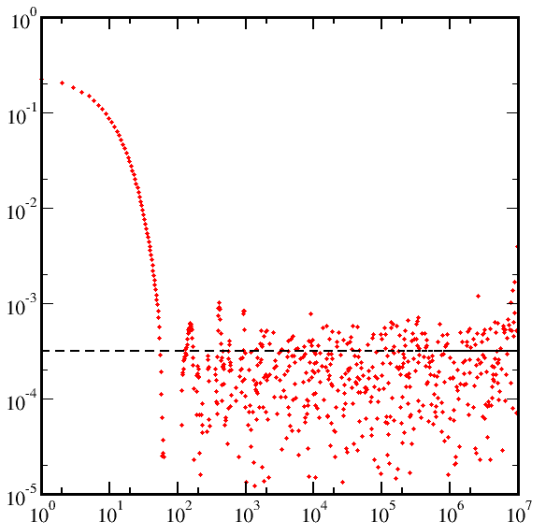
Random sequence



$$C(r) \approx 0$$

- Noise fluctuations

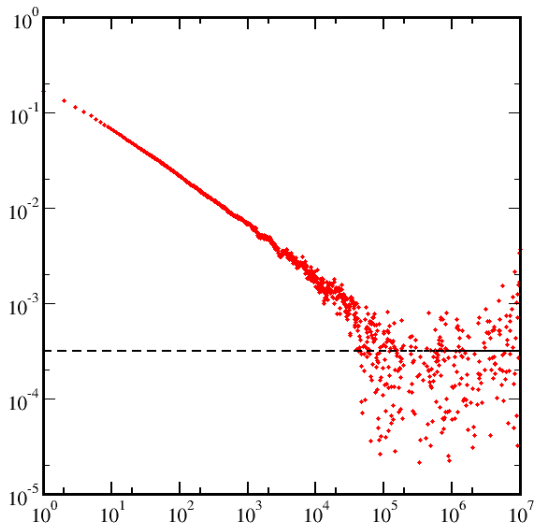
Local correlations



$$C(r) \propto \exp\left(-r / r_0\right)$$

- Characteristic scale
- Generated e.g by Markov-processes

Long-range correlations



$$C(r) \propto r^{-\alpha}$$

- Scale free, fractal
- Generated by a non-trivial dynamical model

Calculation of *C(r)* for Our Model

Approach: continuous time Master Equation formalism

⇩

Exact equation for the dynamics of *C(r)*

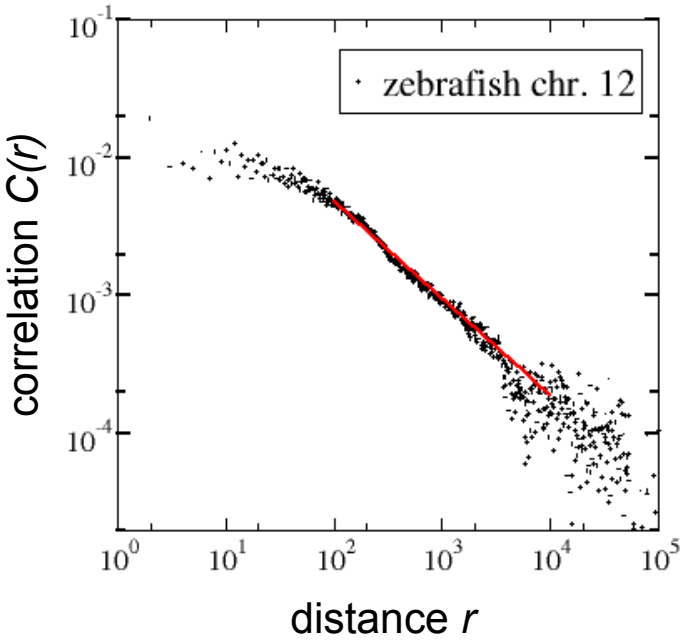
⇩

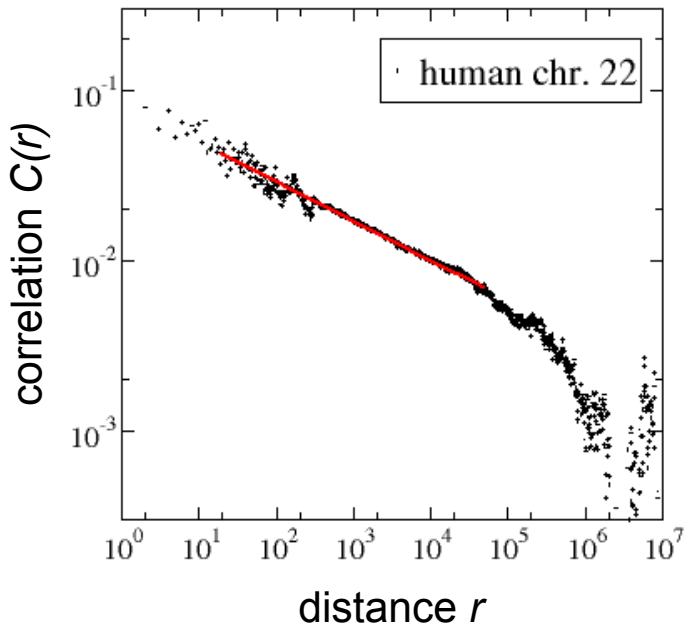
Stationary solution in a continuum limit:

$$C(r) \propto r^{-\alpha} \quad \text{with} \quad \alpha = \frac{4\mu_{\text{eff}}}{\lambda}$$

Correlations in Genomic DNA

$$C(r) \propto r^{-\alpha}$$

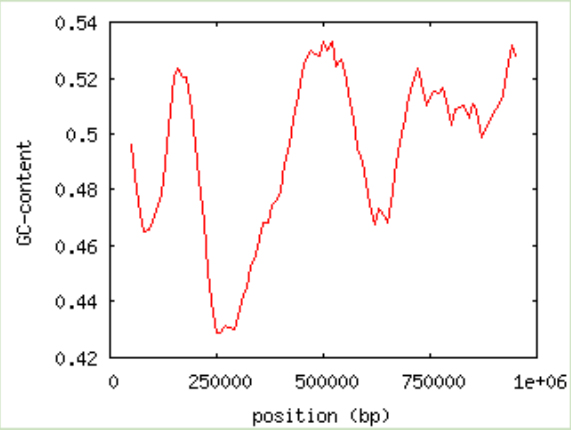


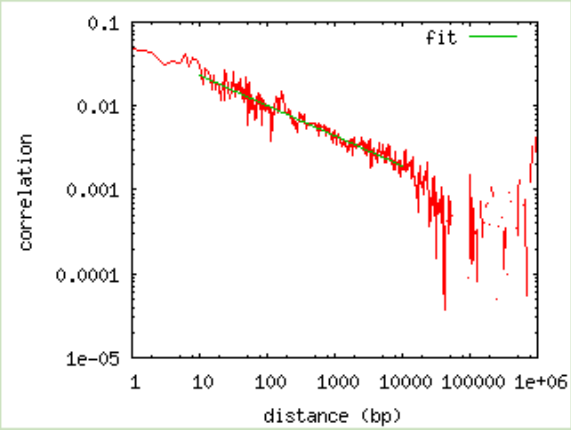


CorGen measuring and generating long-range correlations for DNA sequence analysis

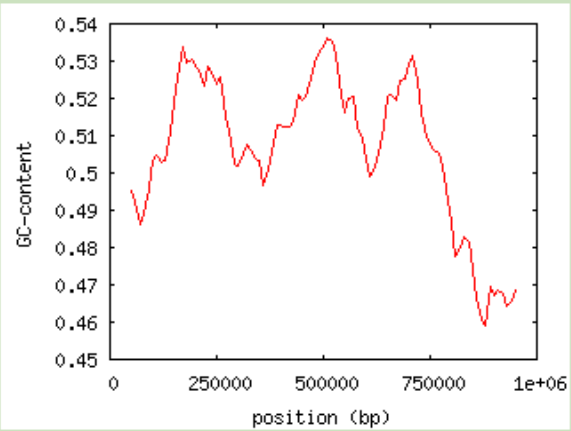
Your uploaded sequence was **1000000 bp long** and has a **GC content of 0.497**. A power-law has been fitted to the correlation function in the range 10-10000. The **decay exponent is 0.359** and the **amplitude(at distance 10 bp) is 0.02340**.

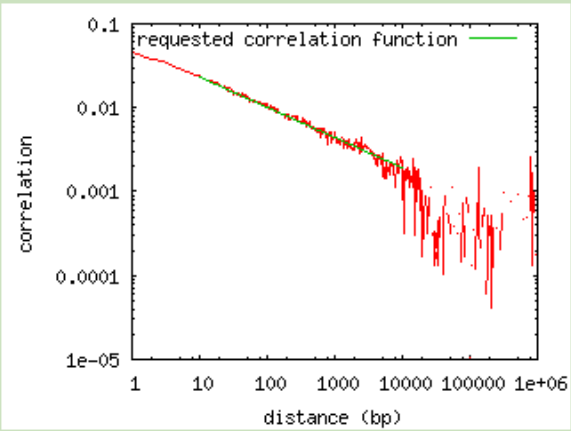
GC profile and the correlation function of the submitted sequence:





A sequence with the same correlation parameters has been generated (and can be downloaded here). Its GC profile and the correlation function are shown below:





You can get an independently sampled sequence here.

It is also possible to retrieve independent samples using non-interactive network clients, e.g. using:
" wget -q -O - 'http://corgen.molgen.mpg.de/cgi-bin/corgen.cgi?seqonly=1&len=1000000&gc=0.497&alpha=-0.35932&dist=10&c=0.02340' ".

Part 4

Tandem Duplications and Alignment Score Statistics

[Messer, Bundschuh, Vingron, Arndt, *RECOMB 2006*]
[Messer, Bundschuh, Vingron, Arndt, *JCB* 2007]

Alignment Score Statistics

Sequence alignment

⇩ Significance

P-values of scores

⇩ Requires DNA null model

Standard iid model

⇩ Problem

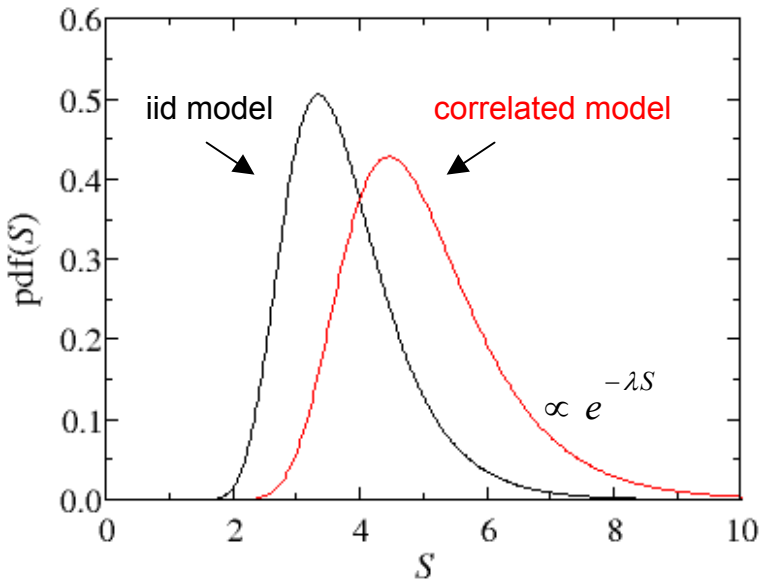
Correlations in DNA

⇩ Incorporate into null model

P-values change

Seq1:	ACCTAGTGCTA
Seq2:	ATCTAGTGATA

Score distribution in the null model

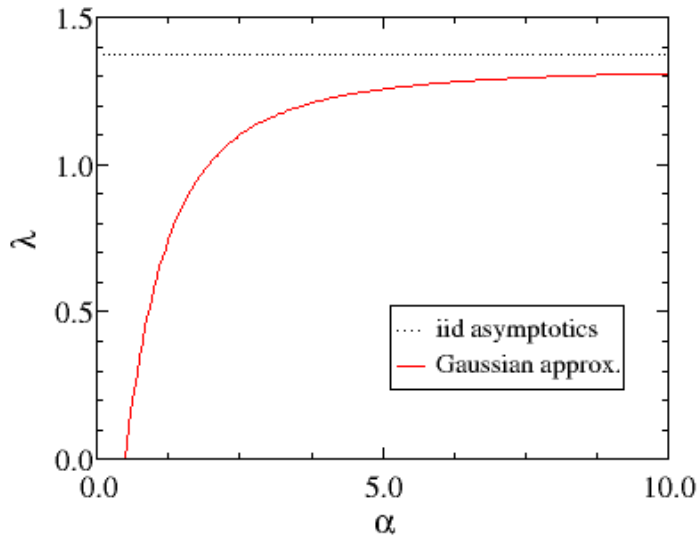


Gaussian Approximation

Analytic approach to calculate alignment score statistics for null models with LRC sequences, i.e. $C(r) \propto r^{-\alpha}$

$$\lambda = \frac{-2\langle s \rangle}{\sigma^2 + c\zeta(2\alpha)}$$

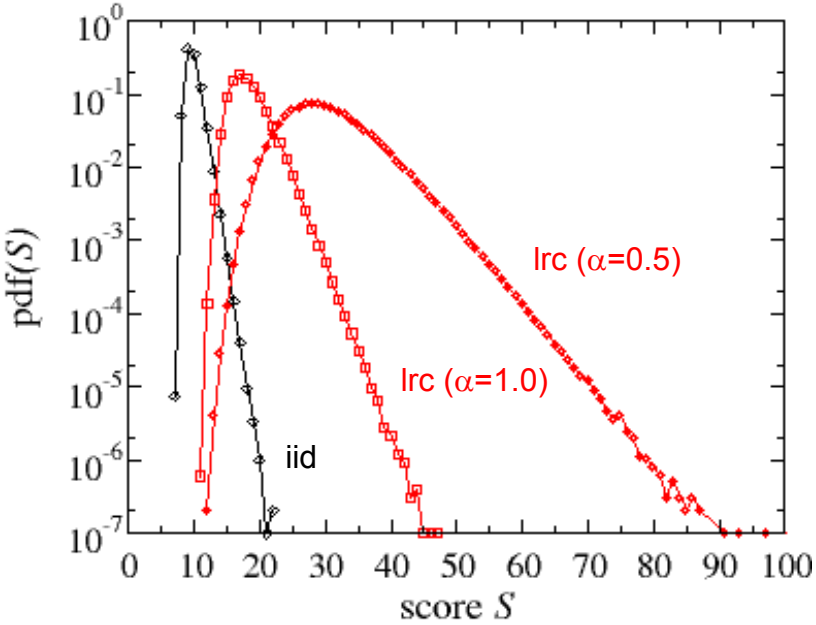
vanishes for iid sequences



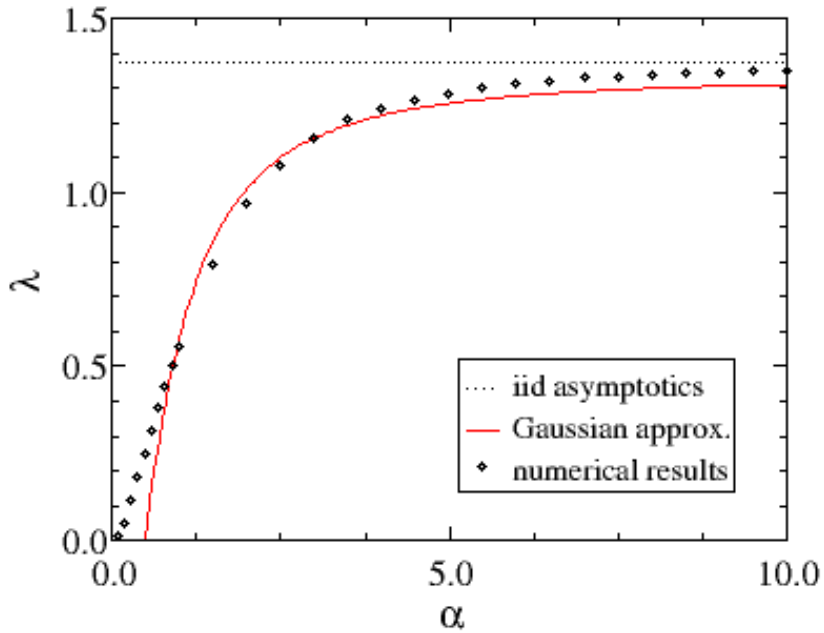
LRC's increases probability of finding high alignment scores by chance

Numerical Verification

Score distribution



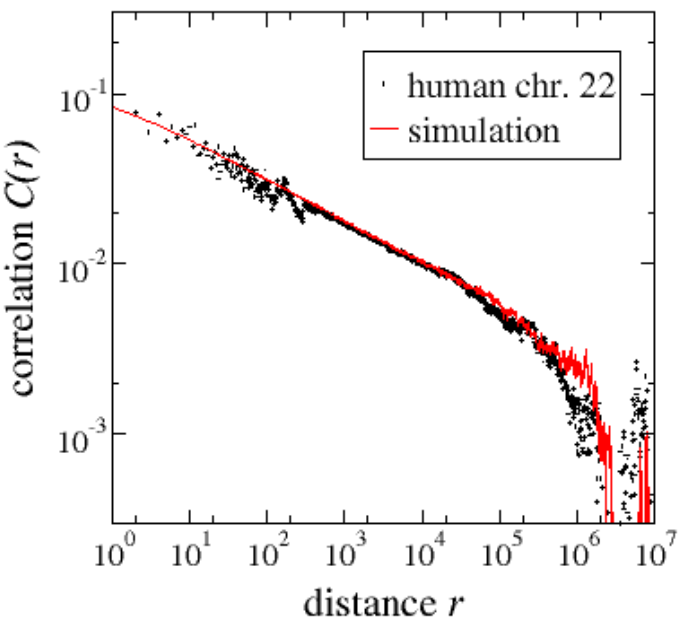
Decay parameter λ

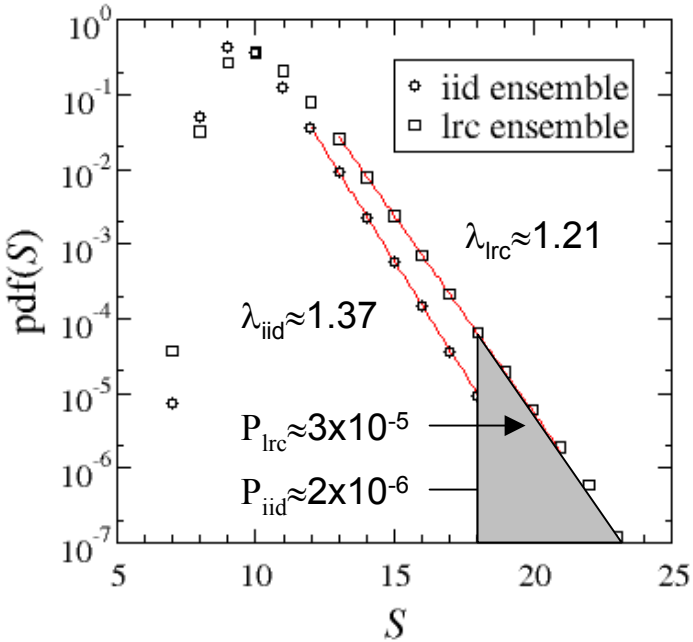


Gaussian approximation captures qualitative behavior

Biological Significance

Alignment of random sequences with same correlation parameters as human chr. 22





Difference in λ is approx.16 %, p-value increase > 10 x

Summary

1. The majority of short DNA insertions are tandem duplications

2. Amino acid insertions/deletions are less deleterious than substitutions

3. Tandem duplications cause long-range correlations in genomic base composition

4. These correlations have profound impact on the statistics of sequence alignment scores

Acknowledgments

Peter Arndt
Nicole de la Chaux
Paz Polak
Federico Squartini

Ralf Bundschuh
Michael Lässig
Martin Vingron

Max Planck Institute
for Molecular Genetics

International
Max Planck Research School
for Computational Biology
and Scientific Computing