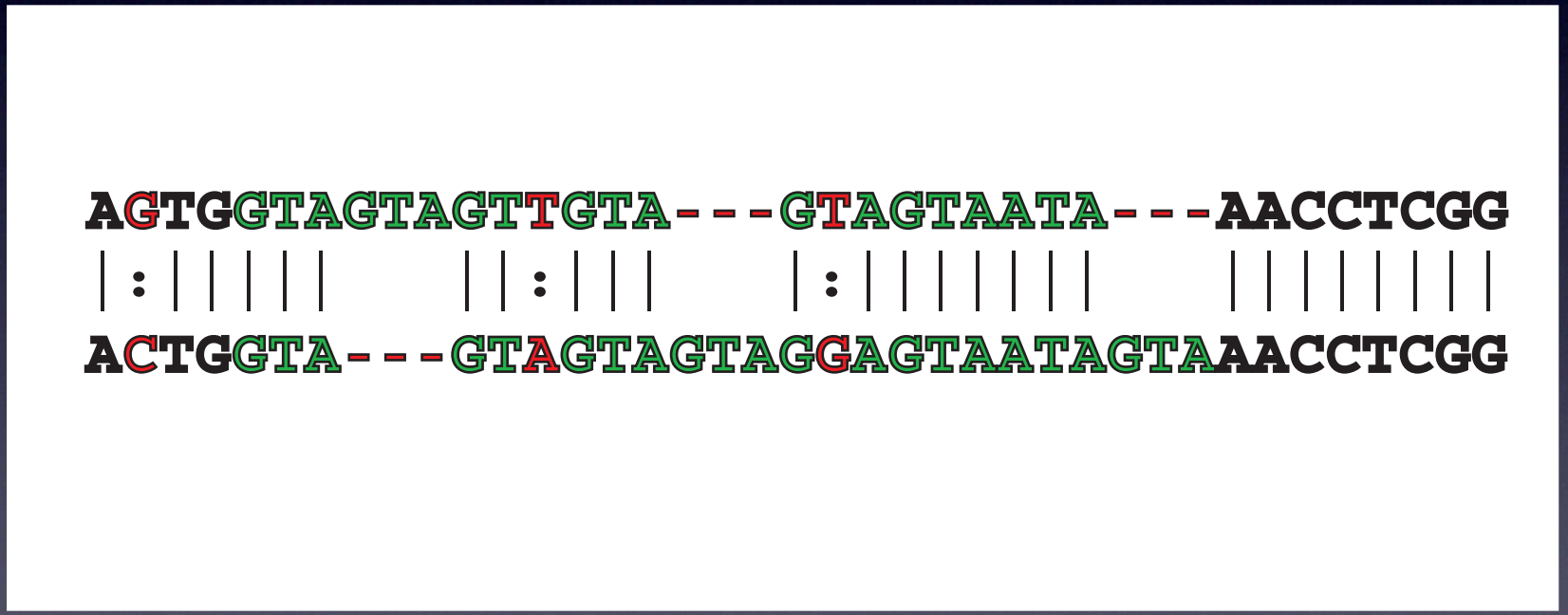
Genome Evolution by Repeat Duplication

Dominic Grün
Institute of Theoretical Physics, University of Cologne

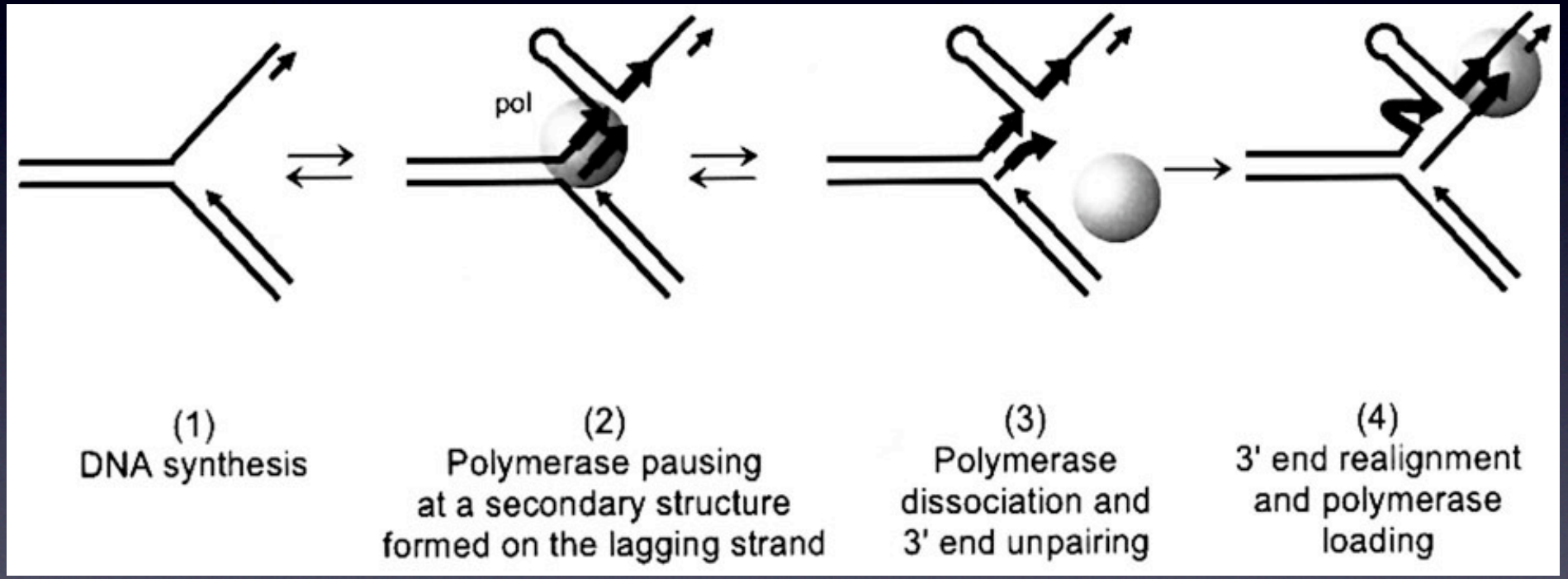
Are base substitutions the leading small scale evolutionary process?

Microsatellites abundant in higher eukaryotes
and
explain a large fraction of small gaps

Sinha, S. & Siggia, E. D. Mol. Biol. Evol. (2005)



Evolution by replication slippage



Viguera, E., Canceill, D. & Ehrlich, S.D. The EMBO Journal (2001)

replication slippage rates >> base substitution rates >> background indel rates

local accumulation of gaps; polymorphic

Inference of correct alignment (phylogeny) difficult

Why were microsatellites studies?

- used as genetic markers
- regulation of gene activity
- involved in cancer and genetic disorders
- DNA metabolism
- Li et al., Molecular Ecology (2002)

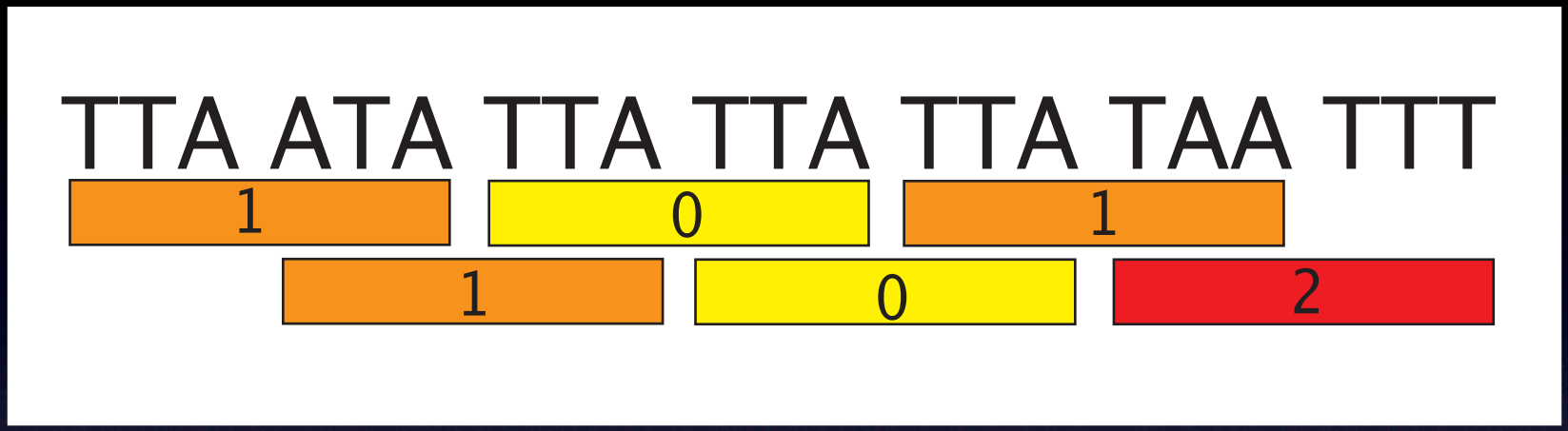
Why do we study microsatellites?

Supply of novel sequence $\Rightarrow$ quantitative model

Grün, D. , Rajewsky, N. & Lässig, M. (2007), in prep.

A minimal model for repeat evolution

A composite object:



Doublets are the dynamical units.

Doublets of Hamming distance $i=0,1,2,...$ falls into error class $i=0,1,2,...$
with occupation number $n_0=2$, $n_1=3$, $n_2=1$, ...

Composition distribution $P(n_0, n_1)$?

Elementary processes for microsatellite evolution:

initiation of novel microsatellites at rate γ_i

CGCTCTTATAAGTCAA ⇨ CGCTCTTATTAGTCAA

forward slippage at rate γ_+

CGCTCTTATTAGTCAA ⇨ CGCTCTTATTATTAGTCAA

backward slippage at rate γ_-

CGCTCTTATTATTAGT ⇨ CGCTCTTATTAGTC

base substitution at rate μ

CGCTCTTATTAGTCAA ⇨ CGCTCTTATGAGTCAA inactive

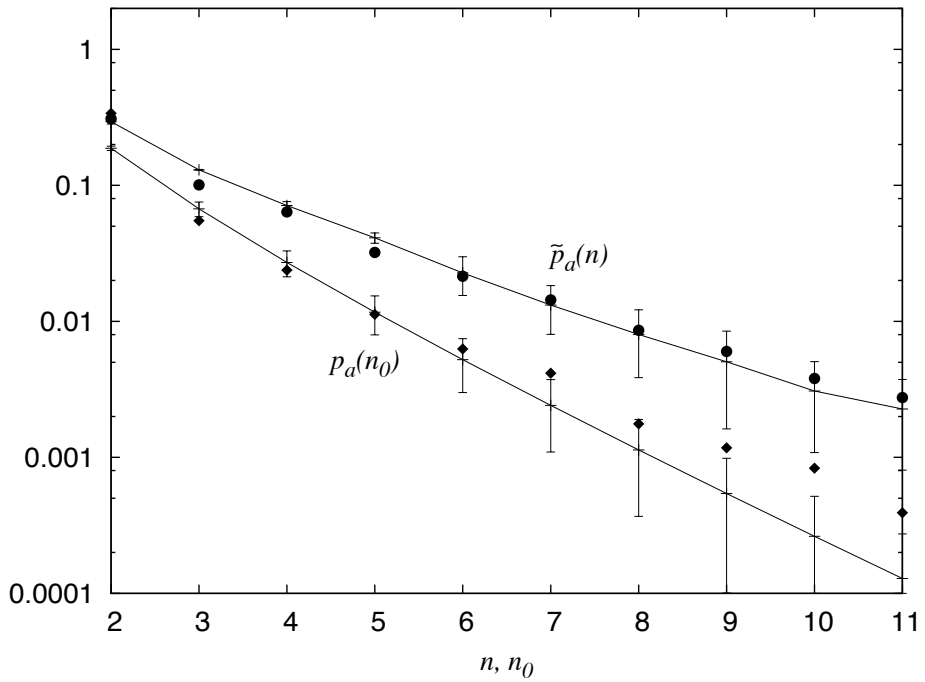
repeat statistics and rate inference from single species data

Microsatellites in *D. melanogaster*:

- identify microsatellites for distinct classes of genomic sequence using Tandem Repeats Finder (Benson, G. Nucleic Acids Res. (1999))

How does the composition distribution look like?

analytical expression for the stationary configuration probability distribution:

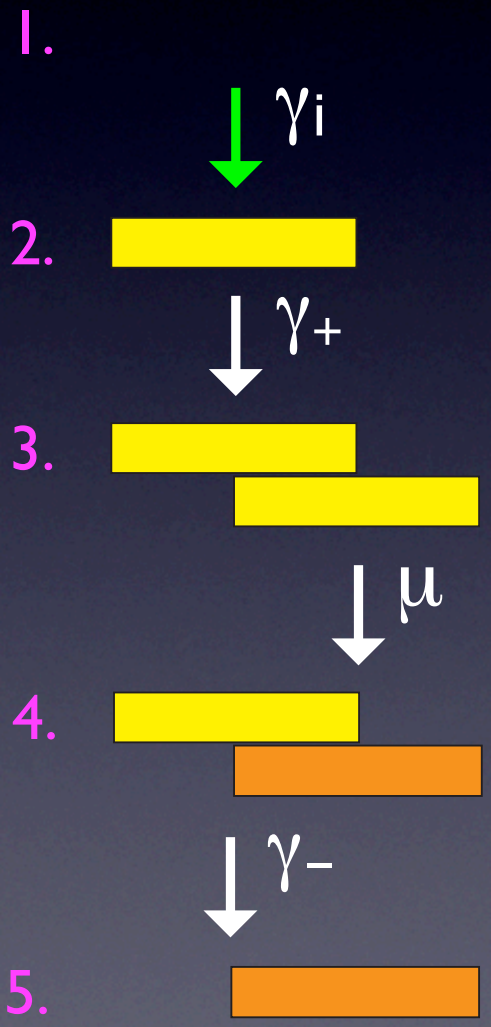


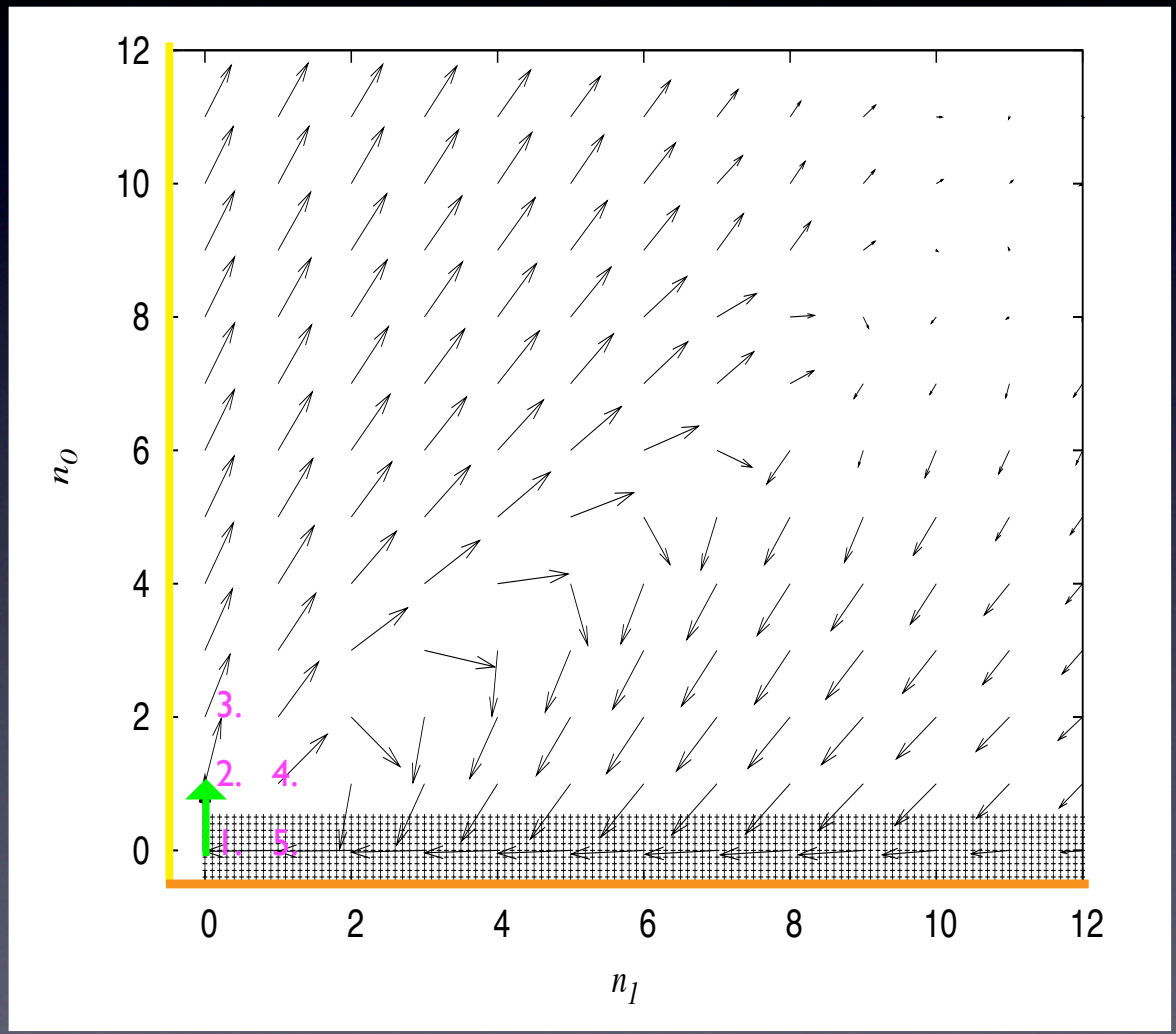
$$p_a(n_0) = \frac{C}{n_0}\,\mathrm{e}^{-\alpha n_0}$$

$$\tilde{p}_a(n) = \frac{C_0}{n}\,\mathrm{e}^{-\tilde{\alpha} n} \qquad n \equiv n_0 + n_1$$

Grün, D. , Rajewsky, N. & Lässig, M. (2007), in prep.

Life cycle of a microsatellite:





Inferred evolutionary rates for trinucleotide repeats in *D. melanogaster*:

- Slippage rates: $\gamma_+/\mu = 4.6 \pm 1.2$ $\gamma_-/\mu = 2.7 \pm 1.2$
- Initiation rate: $\gamma_i/\mu = 0.1 \pm 0.01$
- Coverage: $\lambda = 1.8 \pm 0.1\%$
- Constraint: $\mu/\mu_0 = 0.94 \pm 0.16$ $(\mu_b/\mu_0 = 0.56 \pm 0.02)$

Dynamical Hierarchy:

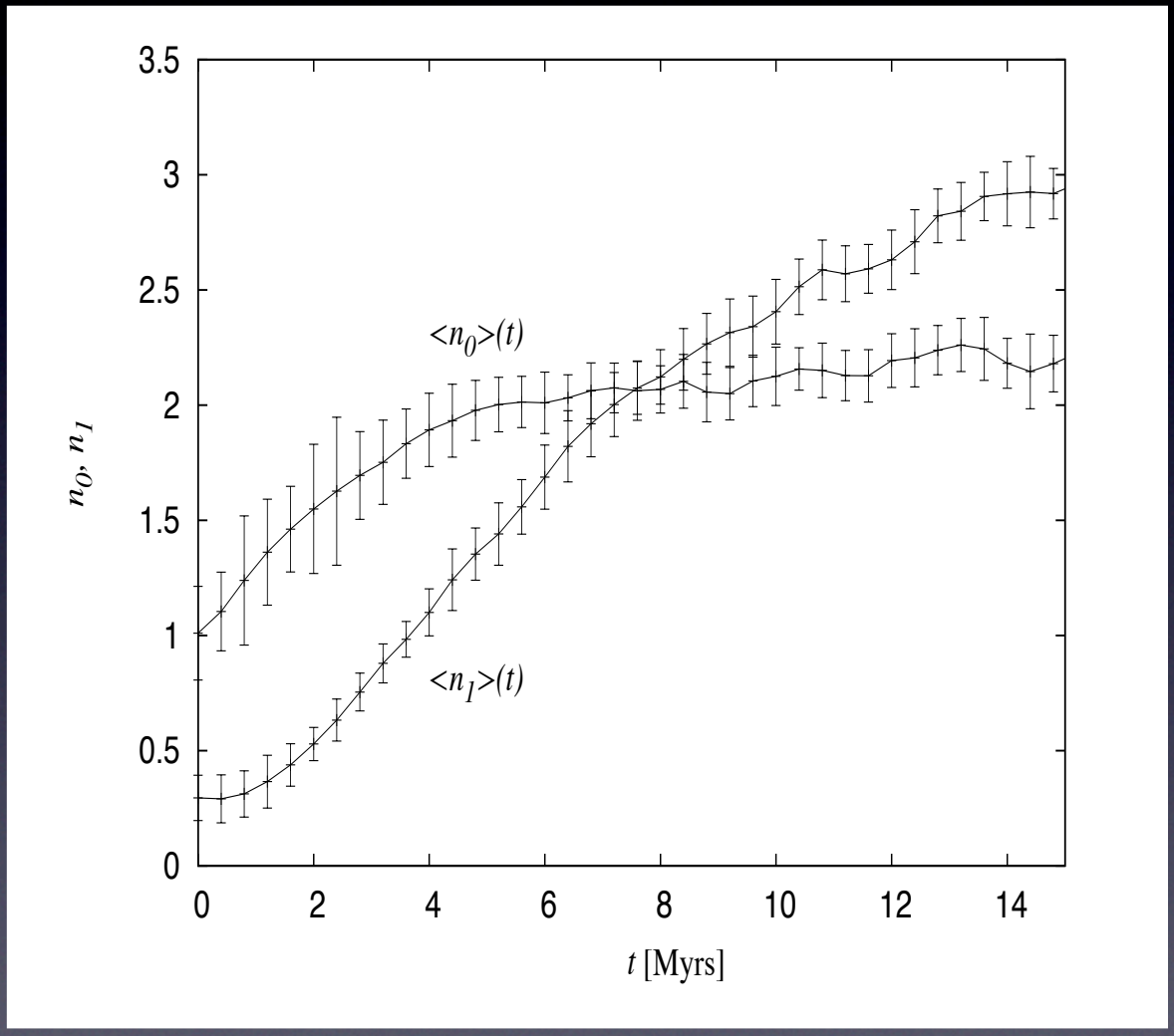
$$\gamma_+ >> \mu >> \gamma_i$$

Microsatellites are rare events and only survive due to random fluctuations

$\gamma_+ > \gamma_-$ but $\gamma_+ < \gamma_- + 2\mu l$

permanent turnover of repeat elements, emergence at rate γ_i

Composition has predictive power for microsatellite age



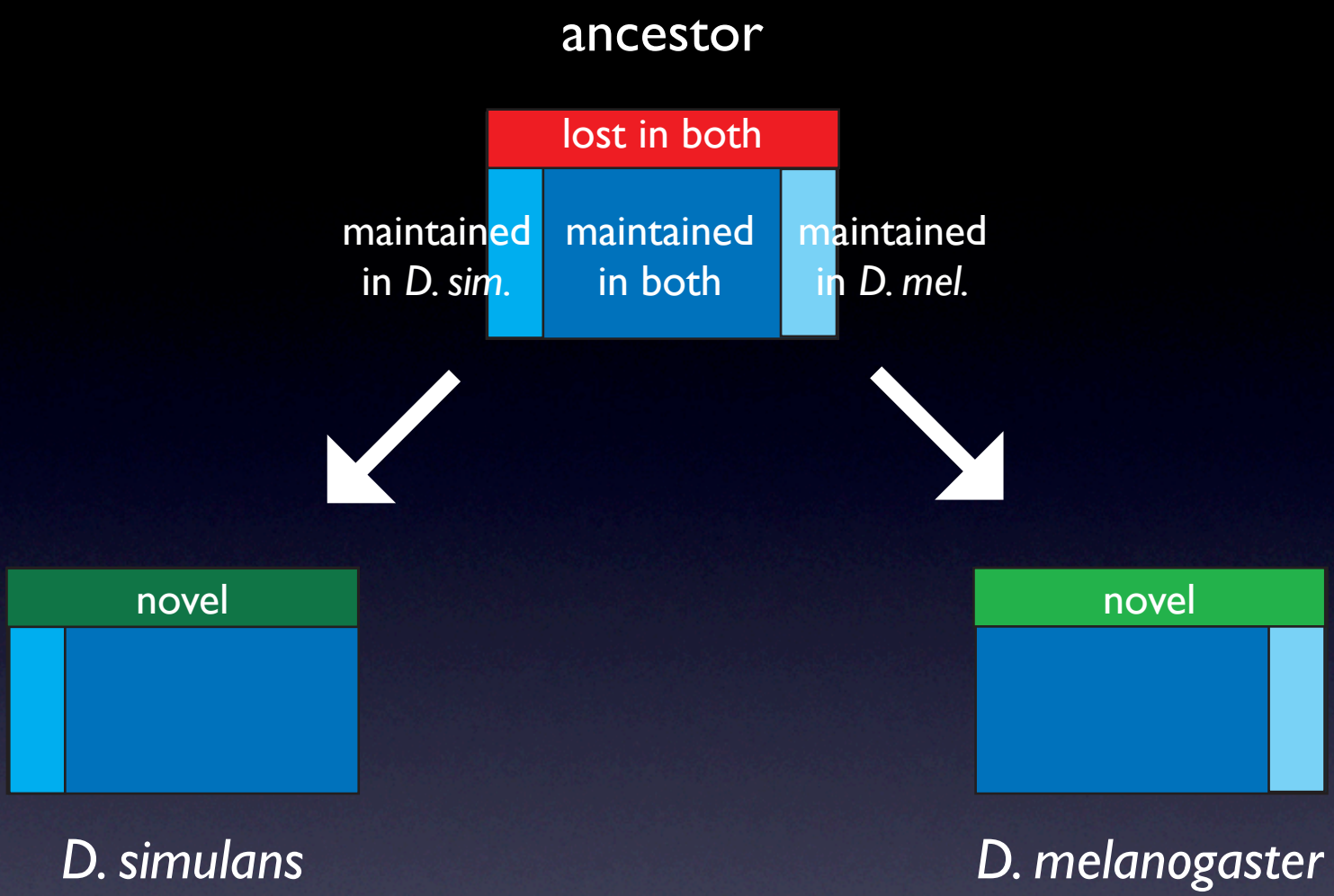
...average age: 3 Myrs

cross species analysis

Sequence turnover by repeat duplication in *D. melanogaster*

test model predictions by cross species comparison between *D. melanogaster* and *D. simulans* (2 Myrs)

How much novel sequence has emerged from microsatellites?



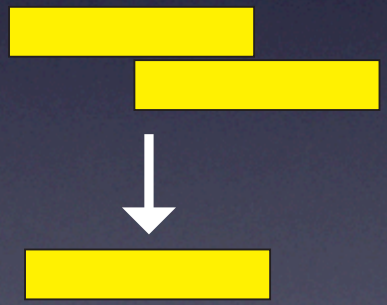
25% of the repeats in *D. melanogaster/D. simulans* are species specific.
60% of the species-specific repeats are novel.

Contributions to sequence turnover:

Sequence emergence rate:

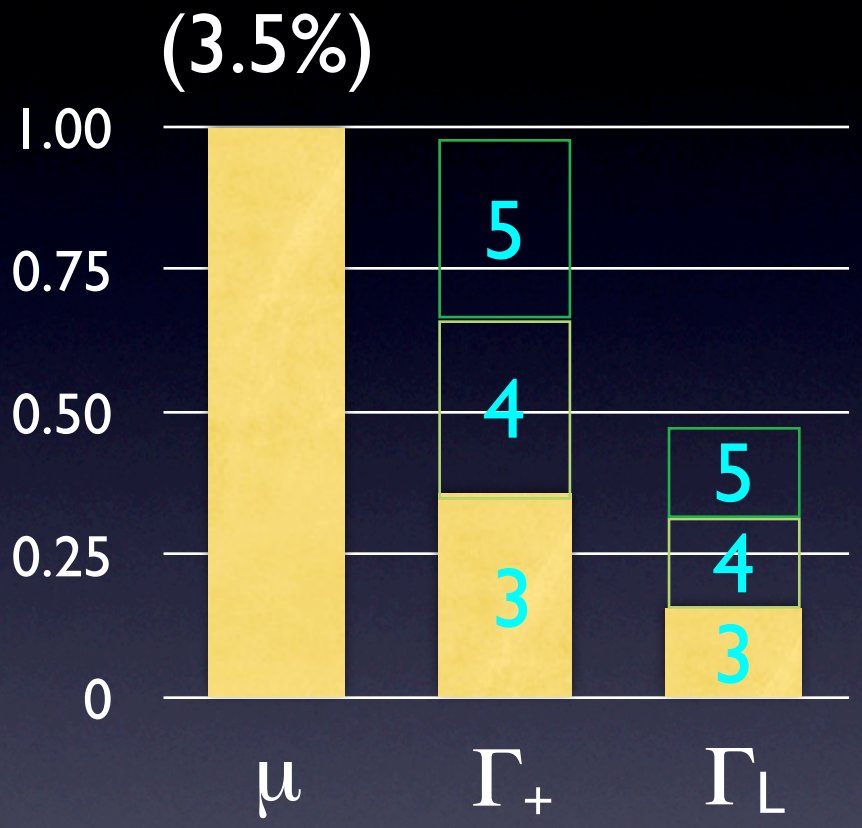
$\Gamma_+ =$  $+$

Net growth rate:

$\Gamma_L = \Gamma_+ -$ 

Why do we look at these two variables?
Selection can set in to maintain favorable sequence

Relative sequence turnover for *trinucleotide repeats*



repeat motif length

Microsatellites contribute to sequence turnover to a similar extent like base substitution.

Grün, D. , Rajewsky, N. & Lässig, M. (2007), in prep.

Conclusions and future work

- microsatellites are genomic fluctuations
- composition allows prediction of age
- in *D. melanogaster*, microsatellites
  are rare
  mutate fast
  contribute to sequence turnover comparably to base substitutions
- need for progressive alignment tool to improve gap alignments and infer the correct phylogeny

Acknowledgements:

Michael Lässig
Nikolaus Rajewsky

SFB 680 "Molecular Basis of Evolutionary Innovations"
joint project with the Institute of Genetics, University of Cologne