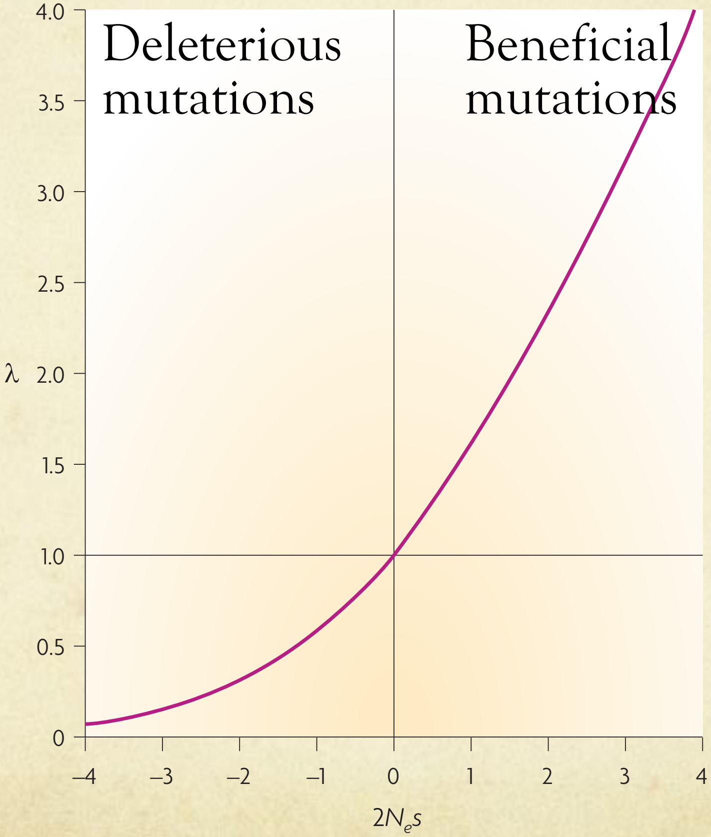
Detecting selection within genomes: recombination clouds the clues

Laurent Duret
Laboratoire de Biométrie et Biologie Evolutive, CNRS, Université Lyon 1

gBGC interferes with selection

- Without gBGC

Fixation probability
(relative to neutral mutations)

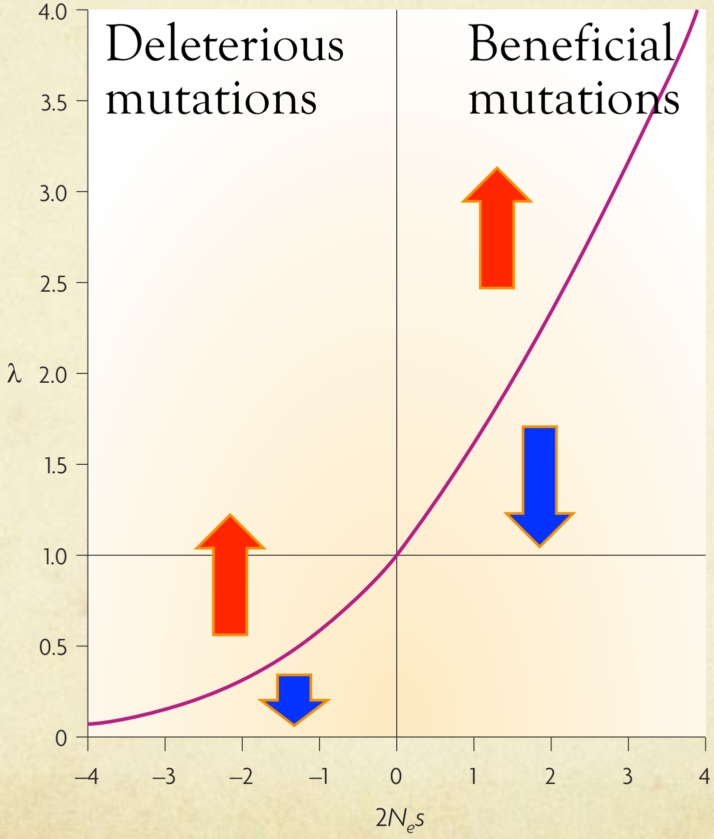


Selection coefficient (population-scaled)

gBGC interferes with selection

- With gBGC

Fixation probability
(relative to neutral mutations)



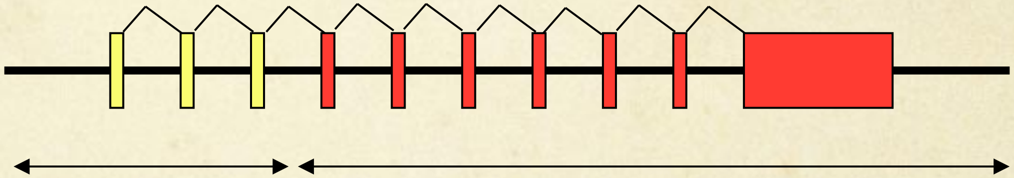
Selection coefficient (population-scaled)

AT→GC mutations

GC→AT mutations

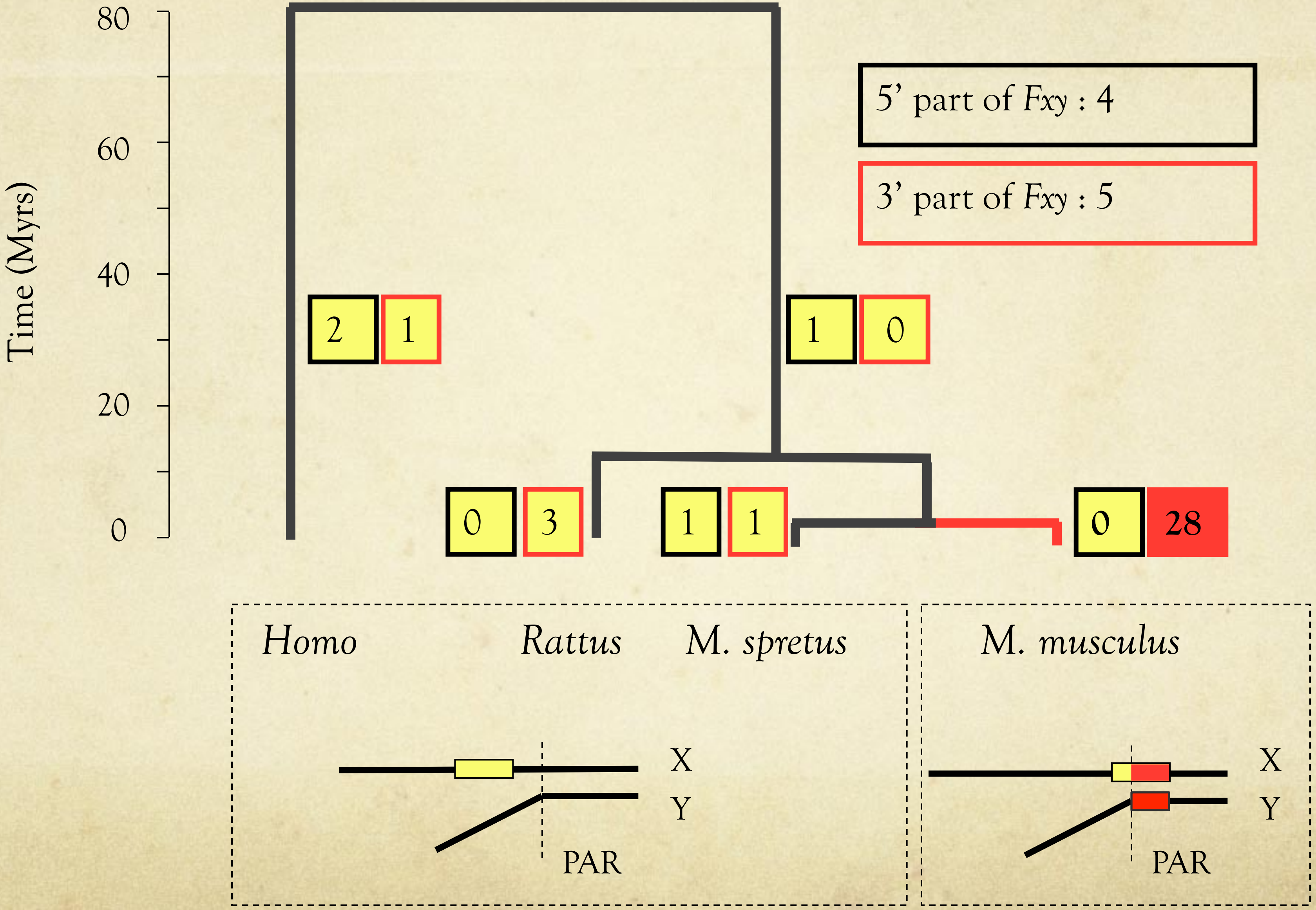
gBGC interferes with natural selection

- *Fxy* gene : translocated in the pseudoautosomal region (PAR) of the X chromosome in *Mus musculus*

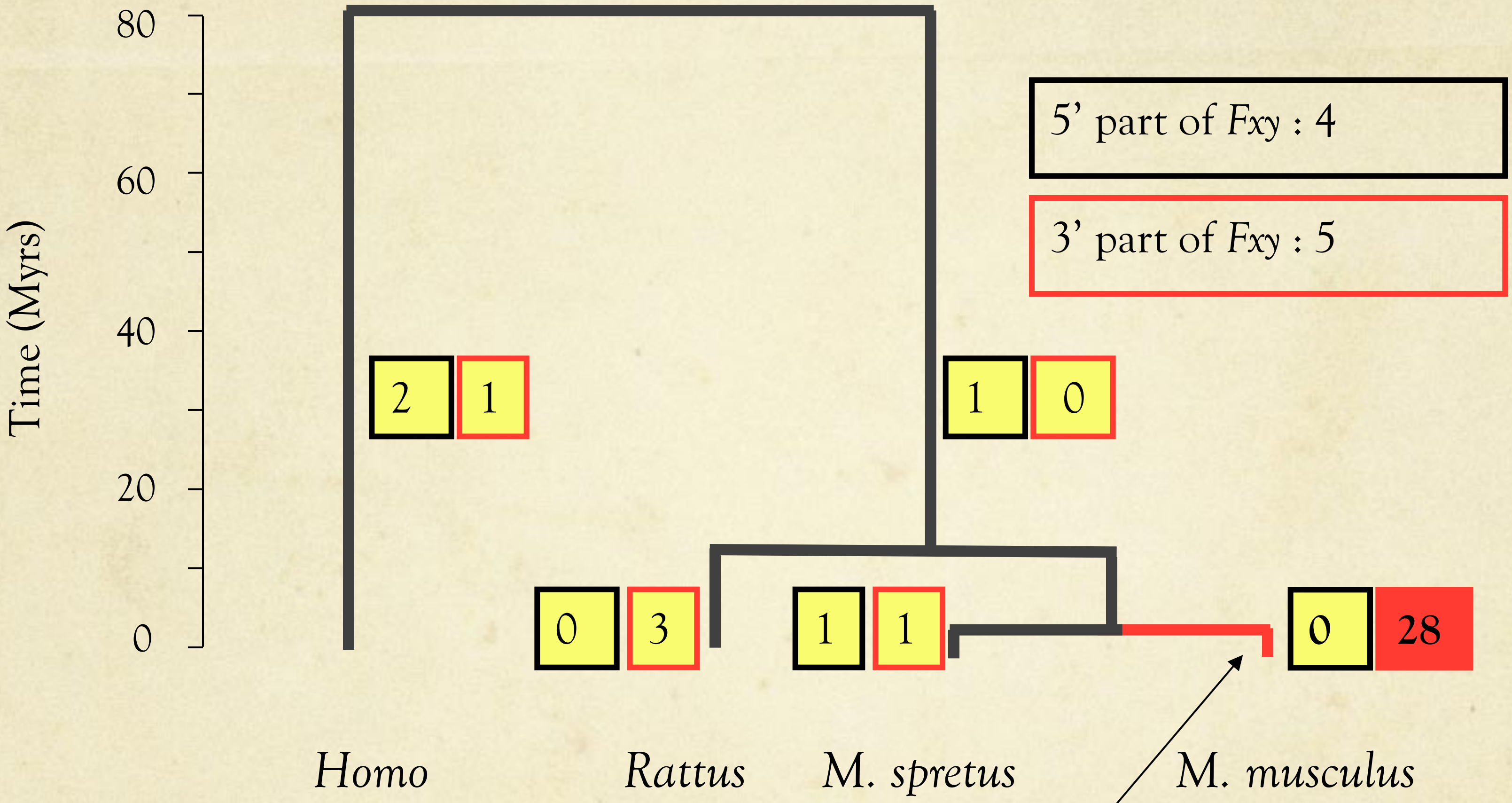


	X specific	PAR
Recombination rate	normal	extreme
GC synonymous sites	normal (55%)	very high (90%)

Amino-acid substitutions in *Fxy*



Amino-acid substitutions in *Fxy*



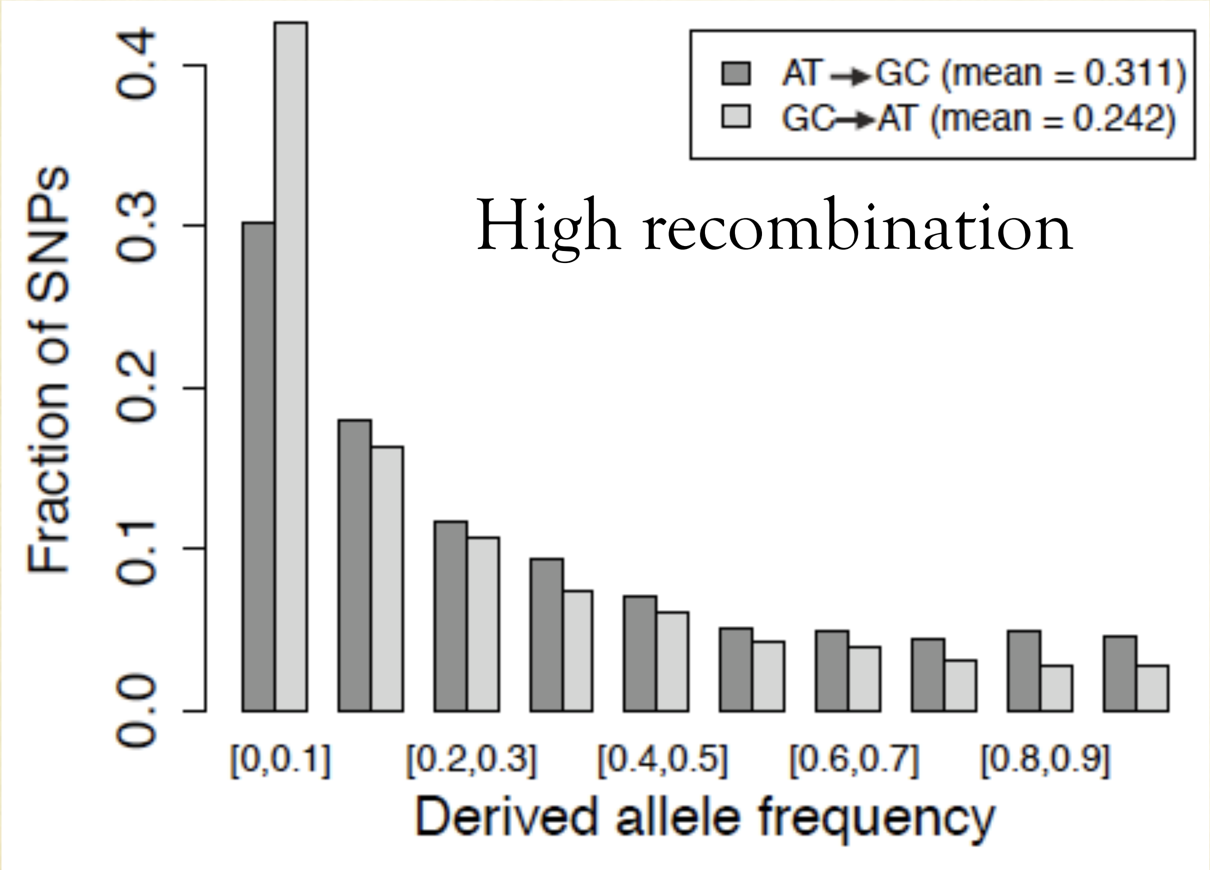
28 non-synonymous substitutions, all AT→GC
Acceleration: x 327 NB: strong negative selection

Is *Fxy* just an exception?

Is gBGC strong enough in other regions of the genome to affect the spreading of deleterious mutations?

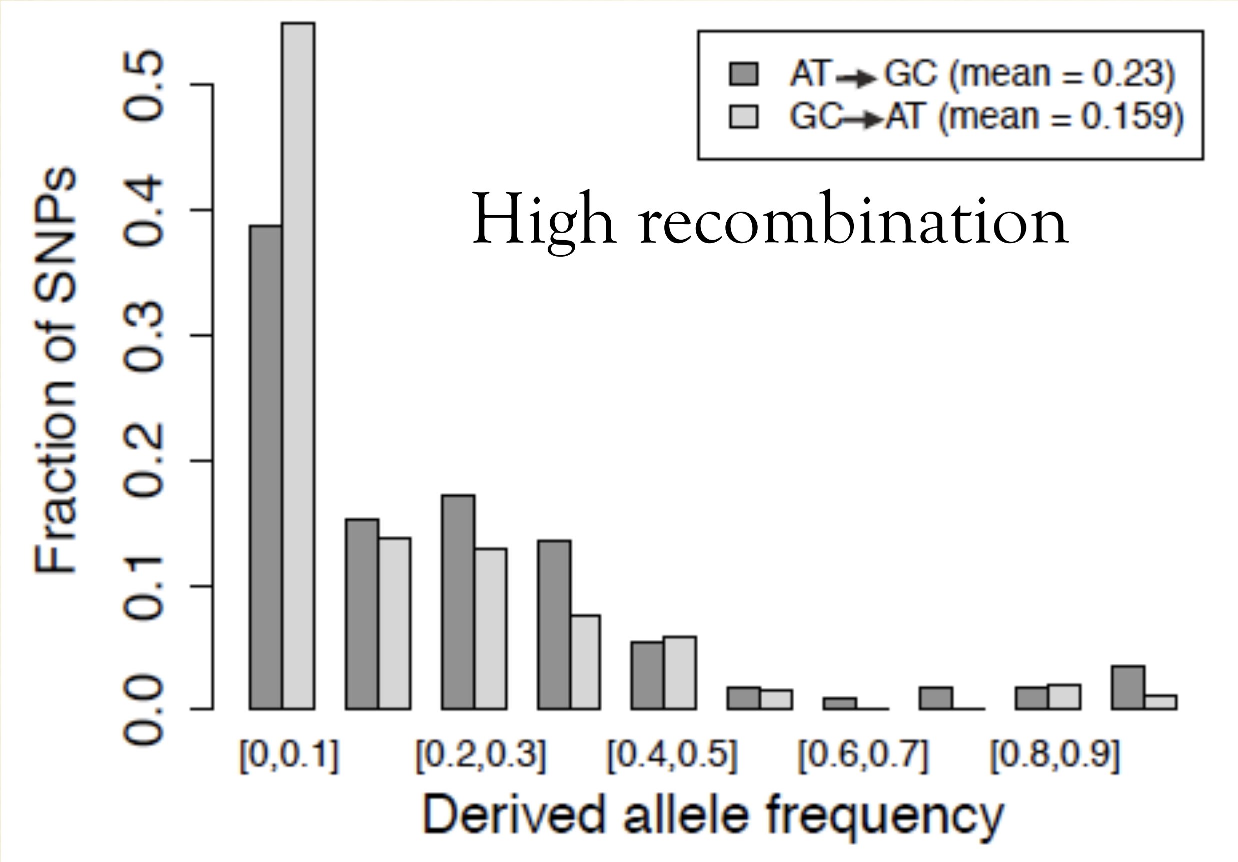
Does gBGC affect the fate of deleterious mutations in extant human populations?

DAF spectrum: non-synonymous SNPs



N=4,975 SNPs, from HapMap (YRI). $p < 10^{-3}$

DAF spectrum: probably damaging non-synonymous SNPs



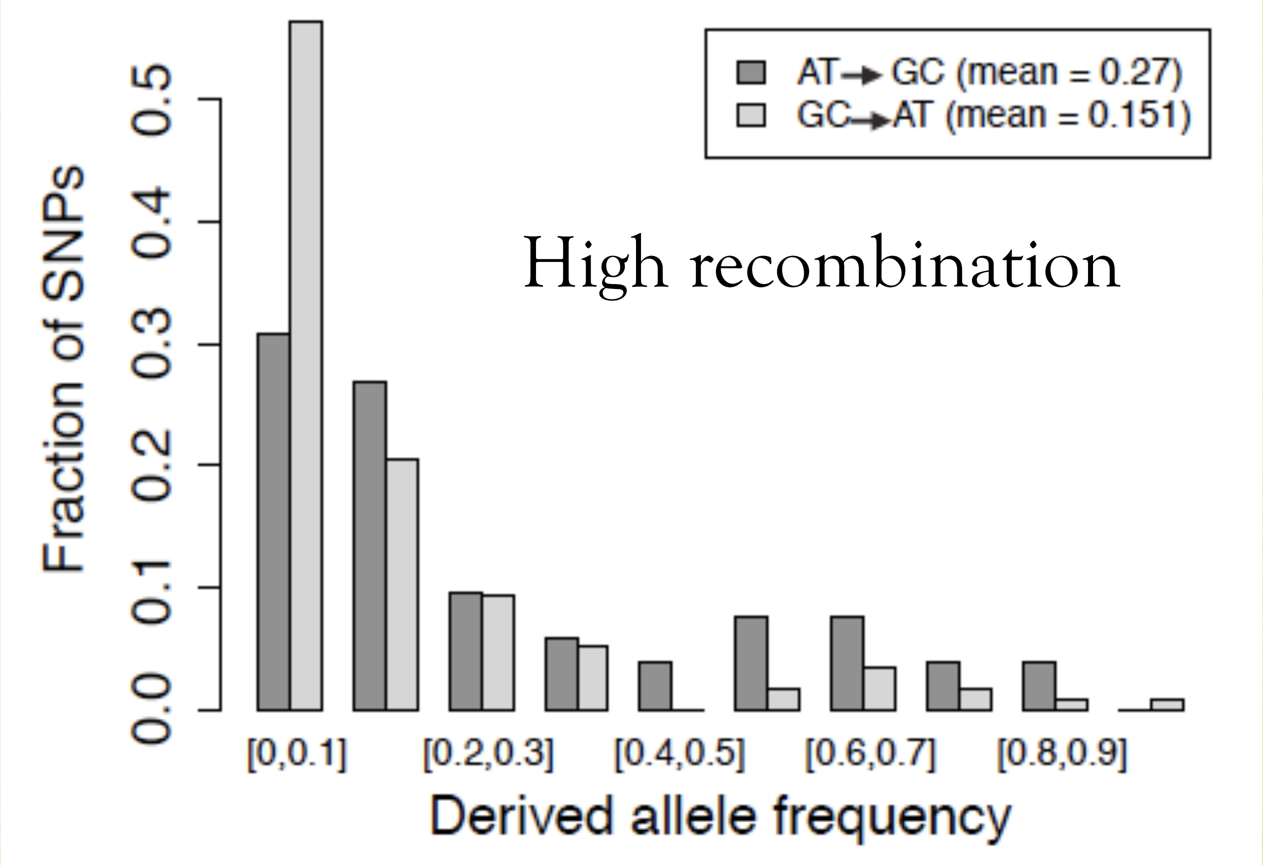
High recombination

Polyphen predictions

N=351 SNPs, from HapMap (YRI). $p = 10^{-3}$

DAF spectrum: mutations involved in genetic diseases

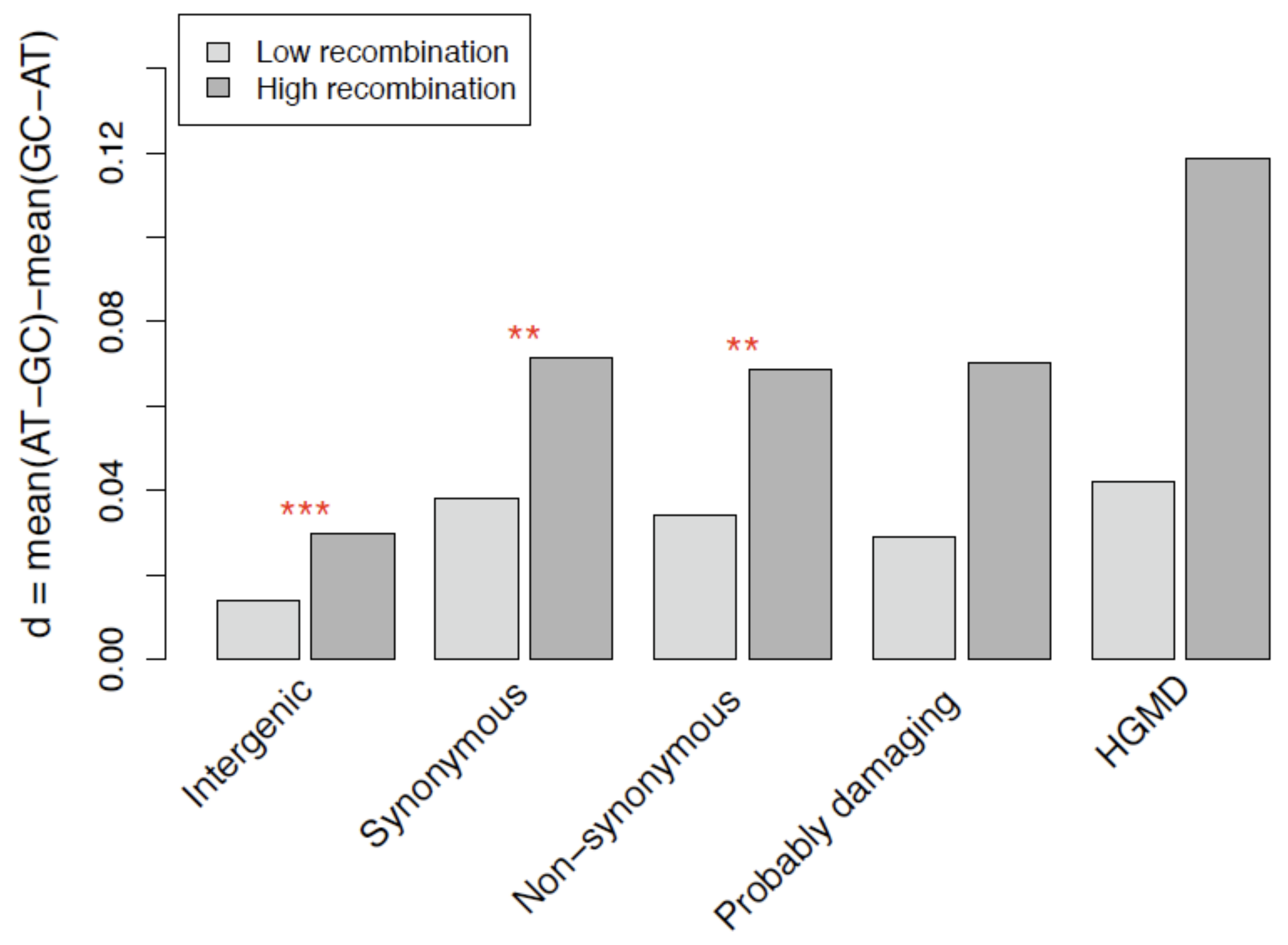
HGMD database



High recombination

N=169 HGMD mutations present in HapMap (YRI). $p < 10^{-3}$

The fixation bias in favor of GC-allele increases with recombination



Summary

- Non-synonymous AT→GC mutations segregate at higher frequency than GC→AT mutations in regions of high recombination
- This pattern is observed for all SNPs, including those that are involved in genetic diseases
- => **gBGC favors the spreading of deleterious AT→GC mutations in human populations**

Recombination hotspots: the Achilles' heel of our genome

- Recombination occurs essentially in hotspots (<2kb)
- gBGC => substitution hotspots in recombination hotspots

(Dreszer et al. 2007, Genome Res.; Duret & Arndt 2008, Plos Genet.)

- gBGC can drive the fixation of deleterious mutations in genes overlapping hotspots

Galtier N. and Duret L. (2007) *Trends Genet*

Galtier N., Duret L., Glemin S., and Ranwez S. (2009) *Trends Genet*

ELSEVIER

TRENDS in Genetics

The Achilles' heel of our genome

The evolution of microbial virulence

Silencing by imprinted non-coding RNAs

Are chromosomal imbalances important in cancer?

Trends Genet. June 2007 Vol. 23 No. 6, pp. xxx–xxx ISSN 0168-9525

The impact of gBGC on selection tests

Tracking natural selection ...

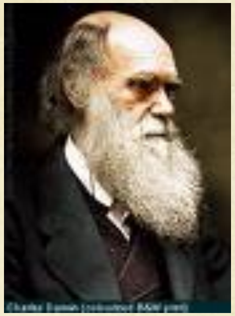
- **Demonstrate the action of selection = reject the predictions of the neutral model**

- Compare substitution rate (K) to mutation rate (u) :
  - Neutral evolution => $K = u$
  - Negative selection => $K < u$
  - Positive selection => $K > u$

Protein-coding genes:
  Non-synonymous substitution rate: dN
  Synonymous substitution rate: $dS \approx u$

Searching for signatures of positive selection within genomes:

What make chimps
different from us ?



Positive selection => accelerated evolution (K > u)

gBGC: a non-adaptive process that looks like selection

- Positive selection => acceleration
- But, gBGC also => acceleration
- gBGC can confound selection tests

Genome scans of positive selection on non-coding functional elements

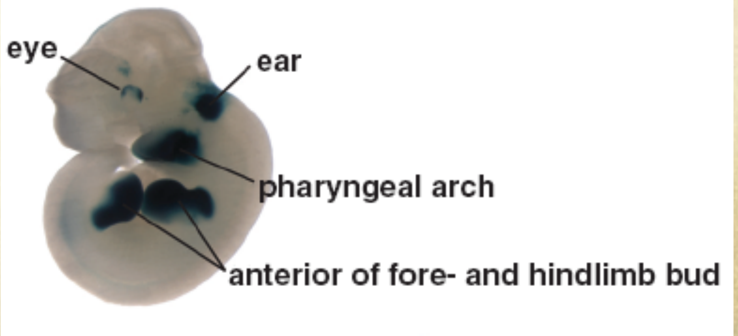
- Regulatory elements: responsible for human-specific adaptations (?)
- Pollard et al. *Nature* (2006), Prabhakar et al. *Science* (2006) : searching for positive selection in non-coding regions
    - Search for conserved non-coding sequences (CNCs) that have significantly accelerated in the human lineage
    - HARs: human-accelerated regions

Positive selection in the human lineage ?

- 49 significant HARs
- HAR1: 120 bp (Pollard et al. 2006 *Nature)*
  - Extreme rate of evolution (18 fixed substitutions in the human lineage, *vs.* 0.7 expected)
  - Part of a non-coding RNA gene
  - Expressed in the brain
  - Involved in the evolution of human-specific brain features ?

Positive selection in the human lineage ?

- HAR2: 546 bp (Prabhakar *et al. 2008 Science)*
  - Extreme rate of evolution (16 fixed substitutions in the human lineage, *vs.* 4 expected)
  - Enhancer activity: drives gene expression in the limb during early development (transgenic mice)
  - Involved in the evolution of human-specific movement capacities (tool use, bipedalism)?



Positive selection ?

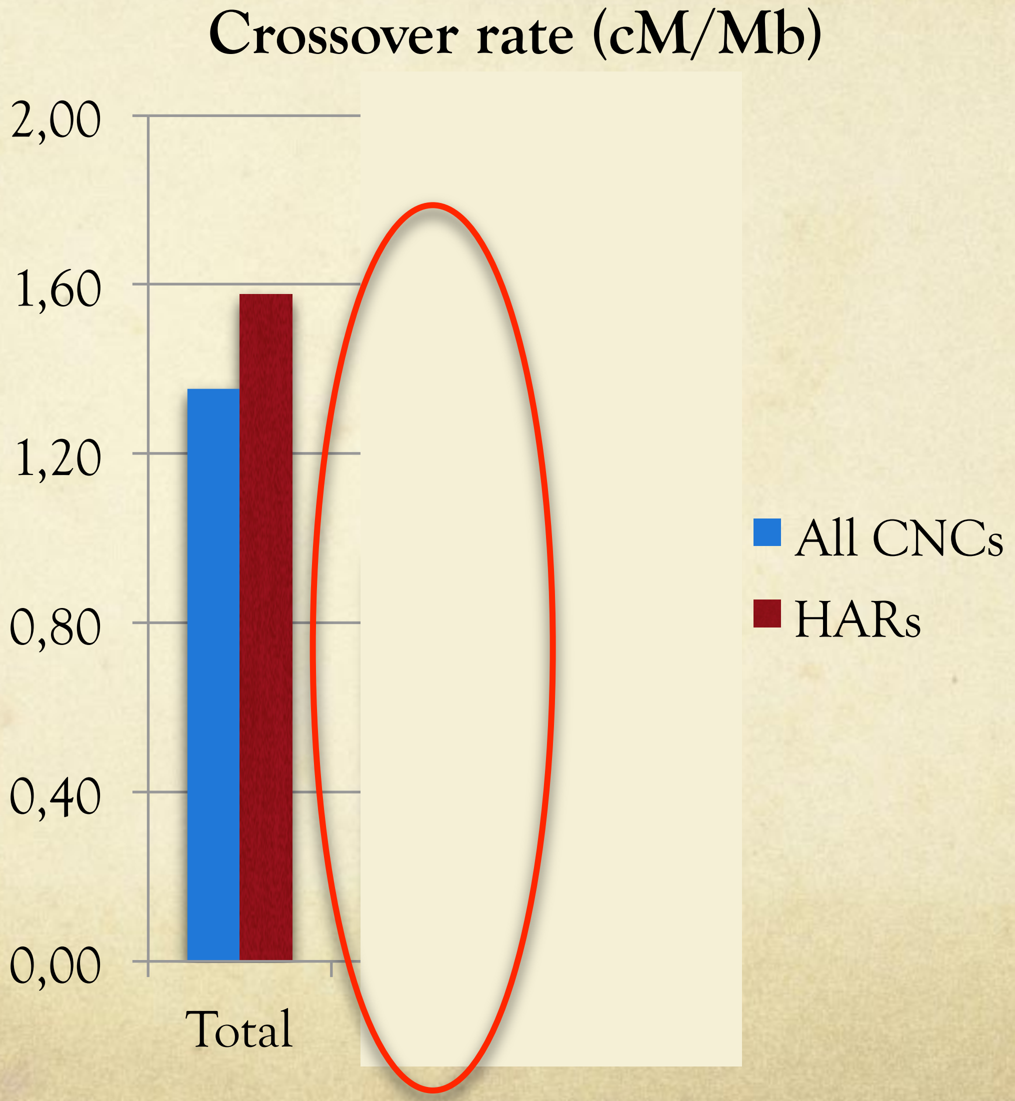
- GC-biased substitution pattern in HARs
  - Proportion of AT→GC changes in HARS = 72%
  - HAR1: the 18 substitutions are all AT→GC changes
  - HAR2: 16 substitutions: 14 AT→GC + 2 CG→GC changes

- Known functional elements (coding or non-coding) are not GC-rich !!
  - GC-content of conserved non-coding sequences (CNCs) = 41%
  - GC-content at 1st and 2nd codon positions = 50%

- HAR1: the accelerated region covers >1 kb, *i.e.* is not restricted to the functional element (120 bp)

HARs are located in regions of high recombination

- N=48 HARs
- Control= 34,829 conserved non-coding sequences (CNCs)



Positive selection or gBGC ?

- All observations are consistent with predictions of the gBGC model
- Null hypothesis: HARs = result of the non-adaptive gBGC process, not positive selection
- HARs = accumulation of (weakly) deleterious mutations driven to fixation by gBGC
- Sumiyama & Saitou (2011): the functional change of HAR2 is due to a loss of function (not a gain)

Duret L. and Galtier N. (2009) *Science* 323:714 [Technical Comment]

Genome scans of positive selection on protein-coding genes

- gBGC affects both synonymous and non-synonymous sites => dN/dS tests expected to be more robust to gBGC than simple acceleration tests

- But… GC-content at synonymous sites (GC3) >> GC-content at 1st and 2nd codon position (GC12)

- => more opportunities for gBGC to drive the fixation of AT→GC mutations at non-synonymous sites

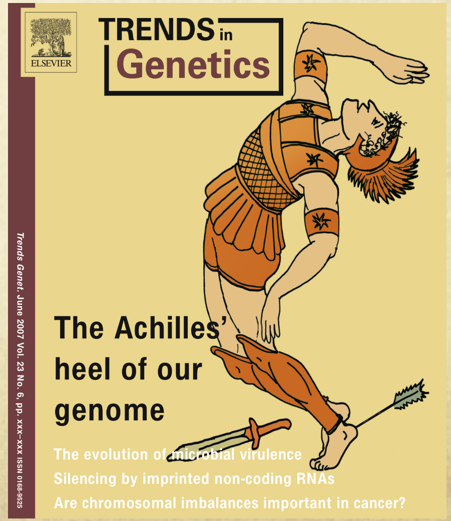
- => gBGC increases the dN/dS ratio and leads to false positive dN/dS tests *(Berglund et al. 2009; Galtier et al. 2009, Ratnakumar et al. 2010)*

How to distinguish positive selection from gBGC?

- Positive selection may favor any type of substitution
- gBGC favors specifically AT→GC substitutions
- Positive selection may affect any locus in the genome
- gBGC occurs in regions of high recombination rate
- Positive selection: affects only a limited number of sites
- gBGC : regional process, affecting all sites (functional or not) located in a recombination hotspots (~1 kb)
- Positive selection: selective sweep
- gBGC: no hitch-hiking (except in the conversion tract: ~1 kb)

Conclusion (1)

gBGC can drive the fixation of deleterious mutations and contribute to the spreading of disease-causing mutations in human populations



ELSEVIER

TRENDS in Genetics

Trends Genet. June 2007 Vol. 23 No. 6, pp. xxx–xxx ISSN 0168-9525

The Achilles' heel of our genome

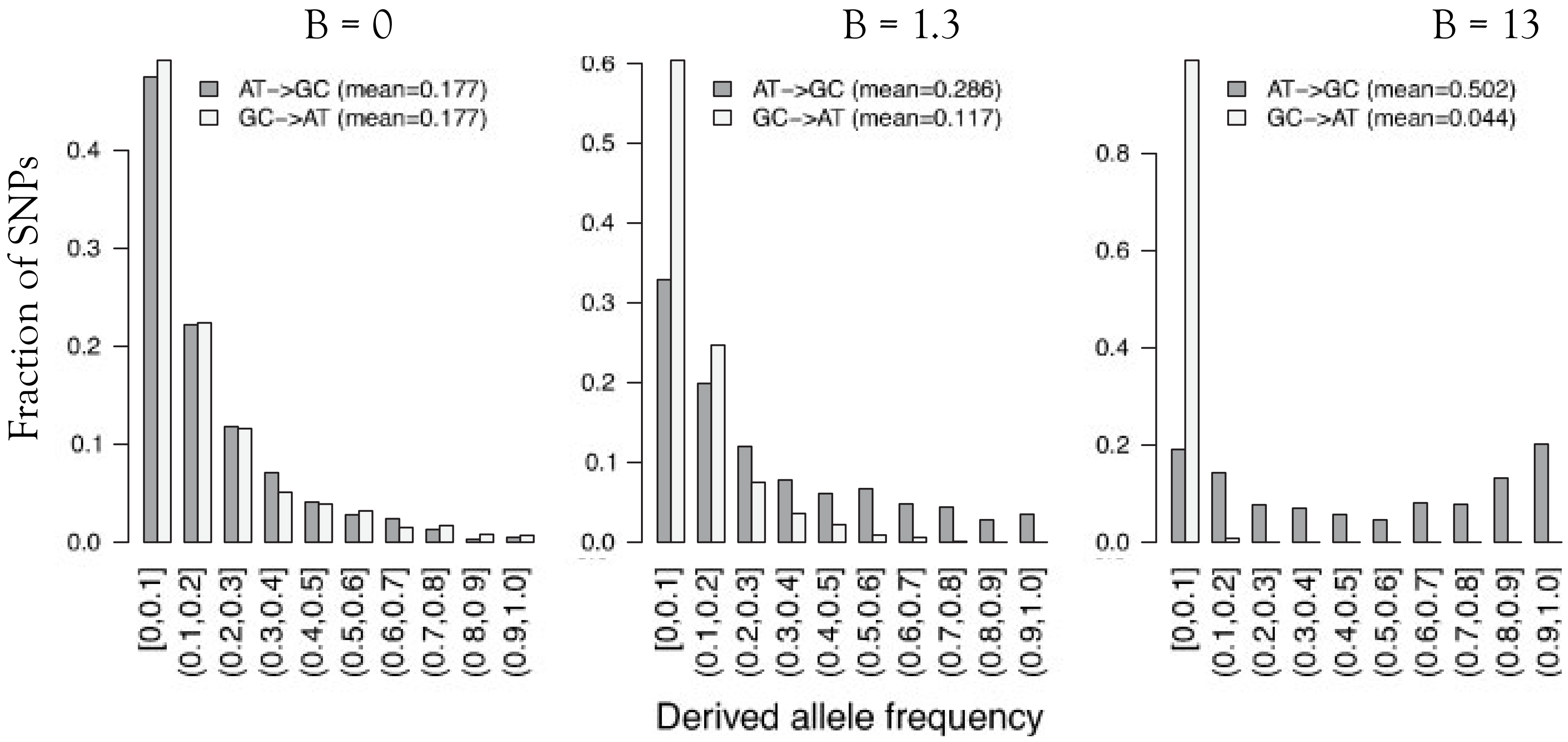
The evolution of microbial virulence
Silencing by imprinted non-coding RNAs
Are chromosomal imbalances important in cancer?

Conclusion (2)

- gBGC can confound selection tests
- Extending the null hypothesis of non-adaptive evolution:
  - Mutation
  - Genetic drift
  - Biased gene conversion

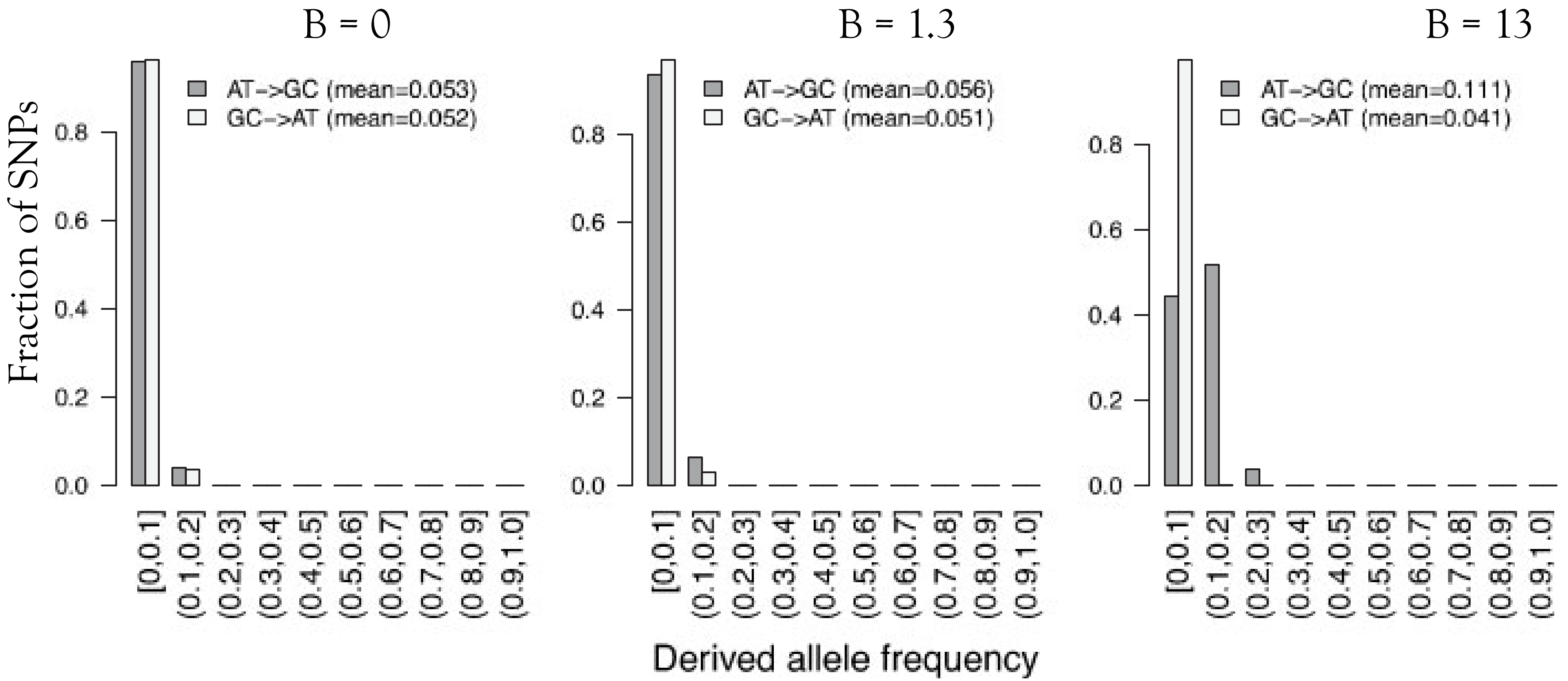
Impact of gBGC on site frequency spectra

- Simulation study:
  - **Neutral sites:** $N_e s = 0$
  - Population-scaled gBGC coefficient: $B = N_e b$



Impact of gBGC on site frequency spectra

- Simulation study:
  - Sites under **strong purifying selection:** $N_e s = -100$
  - Population-scaled gBGC coefficient: $B = N_e b$


B = 0
B = 1.3
B = 13
Fraction of SNPs
AT->GC (mean=0.053)
GC->AT (mean=0.052)
AT->GC (mean=0.056)
GC->AT (mean=0.051)
AT->GC (mean=0.111)
GC->AT (mean=0.041)
[0,0.1]
(0.1,0.2]
(0.2,0.3]
(0.3,0.4]
(0.4,0.5]
(0.5,0.6]
(0.6,0.7]
(0.7,0.8]
(0.8,0.9]
(0.9,1.0]
Derived allele frequency