Evolution of the Ancestral Recombination Graph along the genome in case of selective sweep

Etienne PARDOUX (with Stéphanie LEOCARD)

LATP, Université de Provence, Marseille, France.

Annecy, TAGp10, Oct 22, 2010

The paper on which this talk is based has just appeared as

E. Pardoux, S. Leocard
Evolution of the ancestral recombination graph along the genome in case of selective sweep,
Journal of Mathematical Biology **61**, 1, 819–841, Dec. 2010

Contents

Plan

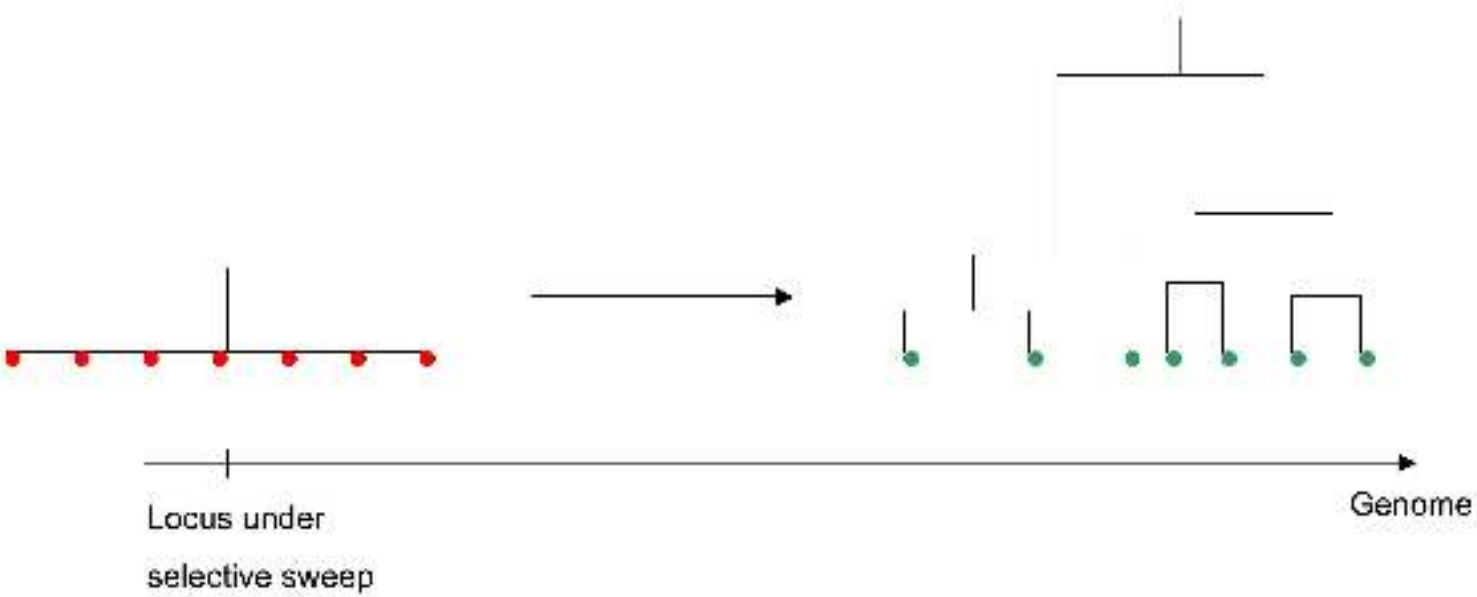
- We consider a population of large size at the end of a **selective sweep**, i. e. a period when a mutation providing a selective advantage which has appeared in a single individual fixate in the population.

- We consider a population of large size at the end of a **selective sweep**, i. e. a period when a mutation providing a selective advantage which has appeared in a single individual fixate in the population.
- At the beginning of the sweep, a single individual carries that mutation; at the end of the sweep, everyboby in the population carries that mutation.

- We consider a population of large size at the end of a **selective sweep**, i. e. a period when a mutation providing a selective advantage which has appeared in a single individual fixate in the population.
- At the beginning of the sweep, a single individual carries that mutation; at the end of the sweep, everyboby in the population carries that mutation.
- For simplicity, we do as if the sweep was instantaneous.

- We consider a population of large size at the end of a **selective sweep**, i. e. a period when a mutation providing a selective advantage which has appeared in a single individual fixate in the population.
- At the beginning of the sweep, a single individual carries that mutation; at the end of the sweep, everyboby in the population carries that mutation.
- For simplicity, we do as if the sweep was instantaneous.
- We are interested in understanding the **genealogy** back in time of a given sample of size n sampled from the population. As we shall see, because of **recombinations**, the genealogy varies dramatically as we move along the genome.

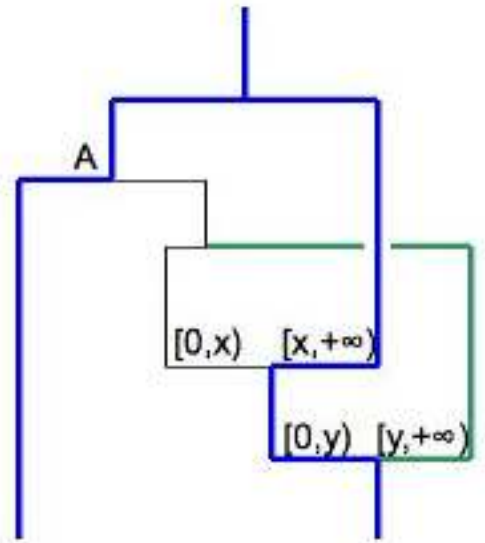
- We consider a population of large size at the end of a **selective sweep**, i. e. a period when a mutation providing a selective advantage which has appeared in a single individual fixate in the population.
- At the beginning of the sweep, a single individual carries that mutation; at the end of the sweep, everyboby in the population carries that mutation.
- For simplicity, we do as if the sweep was instantaneous.
- We are interested in understanding the **genealogy** back in time of a given sample of size n sampled from the population. As we shall see, because of **recombinations**, the genealogy varies dramatically as we move along the genome.
- The genome is here identified with $\mathbb{R}$. The locus where the mutations has appeared is 0, and we shall consider the loci on the right of 0.


Locus under
selective sweep
Genome

Recombinations $\Rightarrow$ if we want to represent jointly the genealogy of a sample at various loci of their genome, we need to replace *coalescence trees* by *ancestral recombination graphs*.

Definition

$ARG(u)$ = graph that sums up the genealogy of the sample, implied by coalescence and recombination events on the portion $[0, u]$.



(In this figure, $0 < x < y$)

WARNING :

- our model of an instantaneous sweep neglects coalescences whithin the sample during the sweep,
- however, we do assume that recombinations happen during the sweep, as a Poisson process of rate 2γ along the genome. There are in fact two independent processes of recombination : recombinations with an individual carrying B, recombinations with an individual carrying b, each one at rate γ.

WARNING :

- our model of an instantaneous sweep neglects coalescences whithin the sample during the sweep,
- however, we do assume that recombinations happen during the sweep, as a Poisson process of rate 2γ along the genome. There are in fact two independent processes of recombination : recombinations with an individual carrying B, recombinations with an individual carrying b, each one at rate γ.

- In other words, we start at the end of the sweep and look backward in time.
  1. During the instantaneous sweep, only recombinations happen.
  2. Before the sweep, during the *neutral period*, both coalescences and recombinations happen. We look backward in time until we find a unique common ancestor of the sample (the so–called MRCA).

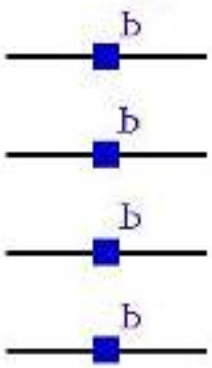
Plan

Haploid population of infinite size.
Genome considered as a single chromosome, identified with $\mathbb{R}$.

Haploid population of infinite size.
Genome considered as a single chromosome, identified with $\mathbb{R}$.

We assume that the gene at locus 0 bear an advantageous mutation :

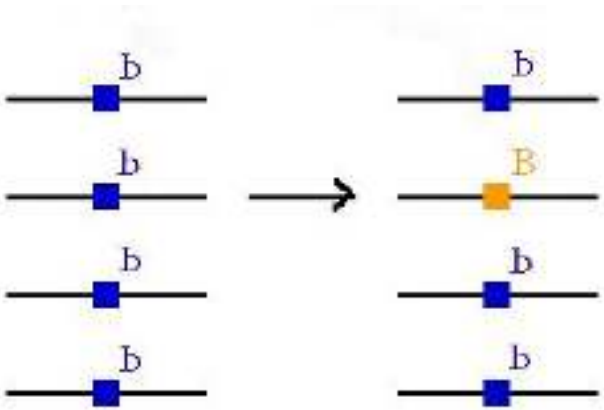
- Before the sweep, everybody carries the wild-type allele b at locus 0.



Haploid population of infinite size.
Genome considered as a single chromosome, identified with $\mathbb{R}$.

We assume that the gene at locus 0 bear an advantageous mutation :

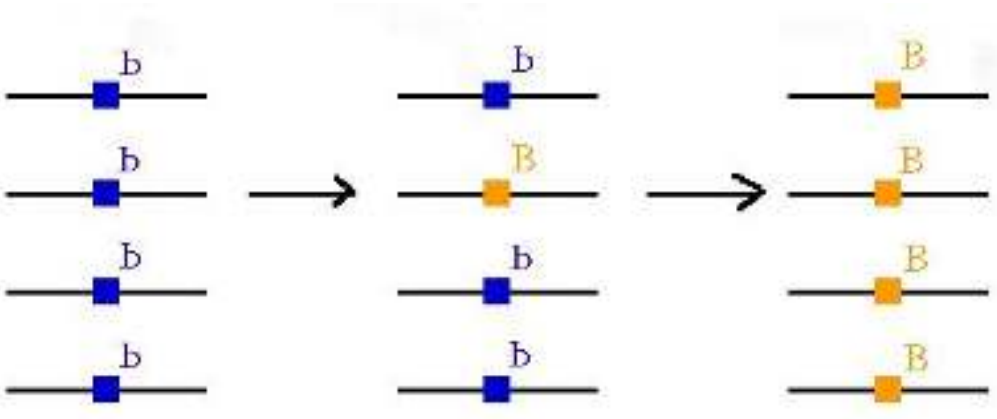
- Before the sweep, everybody carries the wild-type allele b at locus 0.
- An advantageous mutation from b to B happens in a SINGLE individual J. Selective advantage of B over b: α.



Haploid population of infinite size.
Genome considered as a single chromosome, identified with $\mathbb{R}$.

We assume that the gene at locus 0 bear an advantageous mutation :

- Before the sweep, everybody carries the wild-type allele b at locus 0.
- An advantageous mutation from b to B happens in a SINGLE individual J. Selective advantage of B over b: α.
- B quickly spreads in the population until fixation.



Assume $\alpha \to +\infty$.
The duration of the selective sweep tends to 0 as the selective advantage α goes to $+\infty$, so the selective sweep is instantaneous.

Assume $\alpha \to +\infty$.
The duration of the selective sweep tends to 0 as the selective advantage α goes to $+\infty$, so the selective sweep is instantaneous.

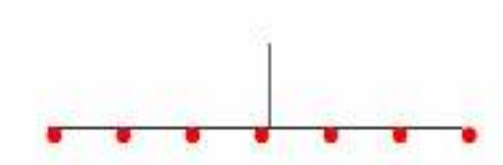
Sample of n individuals at the end of the selective sweep.

Objective: Describe the evolution of the coalescent tree as the distance from the selected site increases (restriction to $[0, +\infty)$).

Question 1: What is the shape of the coalescent tree near and far from the locus under selection?

Question 1: What is the shape of the coalescent tree near and far from the locus under selection?

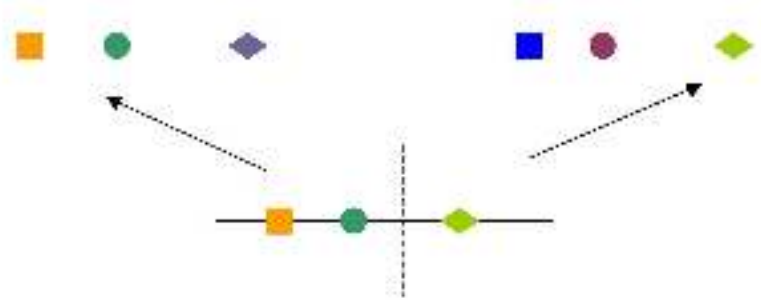
First case: Genealogy at the site under selection (or very close to this site).



Second case: Genealogy far from the site under selection.

Definition

A recombination happens when a portion of the genome is inherited from an individual and the other portion by another individual.



Consider a neutral site at position $x > 0$.

- Without recombination on $[0, x]$ in the sample, same genealogy as the site under selection : comb. (hitchhiking)
- With recombinations on $[0, x]$, no comb anymore.

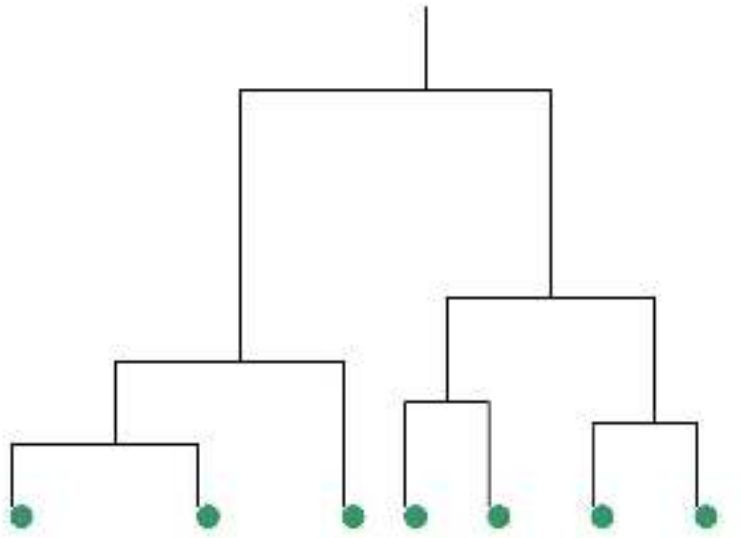
Far from the site under selection (x large):

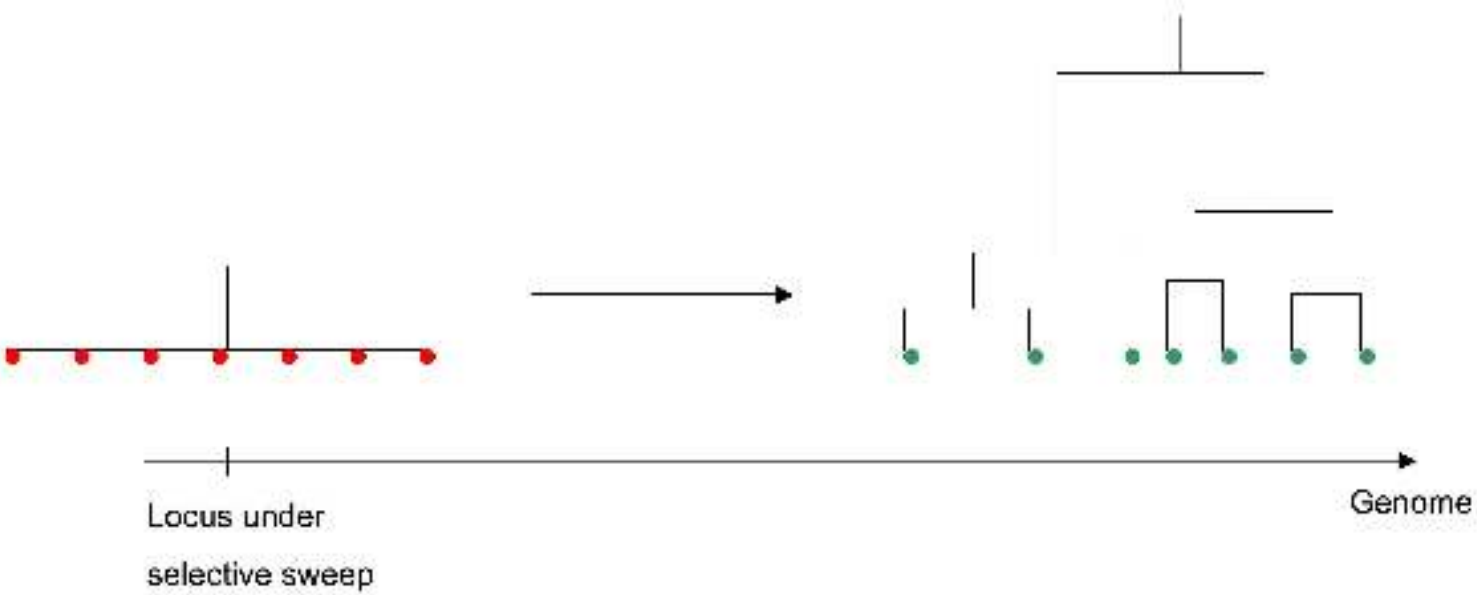
Many recombinations on $[0, x]$
$\Rightarrow$ No hitchhiking anymore
$\Rightarrow$ Evolution under neutrality.

Far from the site under selection (x large):

Many recombinations on $[0, x]$
$\Rightarrow$ No hitchhiking anymore
$\Rightarrow$ Evolution under neutrality.

Genealogy of the n-sample : Kingman's n-coalescent.
Coalescence rate when k lineages: $\binom{k}{2}$.




Locus under
selective sweep
Genome

Question 2: What is the shape of the coalescent tree at various distances from the selected locus?

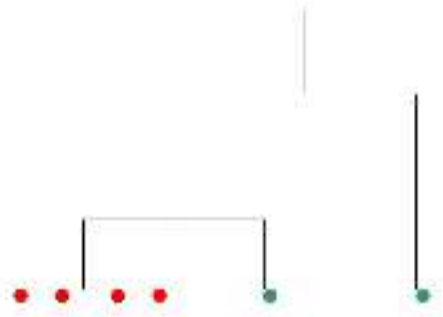
Plan

Let $x > 0$ be a position on the genome.
$R(x) =$ number of alleles at position x in the n-sample, that are inherited from J.

Let $x > 0$ be a position on the genome.
$R(x) =$ number of alleles at position x in the n-sample, that are inherited from J.

Proposition (Shape of the coalescent tree)

The coalescent tree at locus x is a Kingman $(n - R(x) + 1)$-coalescent, where one leaf is a comb with $R(x)$ teeth. The comb gathers the alleles inherited from J.



Question 3: Evolution of the process $x \in \mathbb{R}_+ \to R(x) \in \{0, ..., n\}$*?*

Theorem

The process $x \in \mathbb{R}_+ \to R(x) \in \{0, ..., n\}$ *has the following properties:*

1. $R(0) = n$,
2. *R is a non-homogeneous jump–Markov process, whose jump rates are given as follows:*

$$Q_{k,\ell}(x) = \begin{cases} (1 - \frac{1}{2}\exp(-\gamma x))k \times 2\gamma & \text{if } \ell = k-1, \\ \frac{1}{2}\exp(-\gamma x)(n-k) \times 2\gamma & \text{if } \ell = k+1, \\ 0 & \text{if } \ell \notin \{k-1, k, k+1\}. \end{cases}$$

3. $\exists x(\omega) > 0; R(x) = 0 \ \forall x \geq x(\omega)$ *a.s.*

Theorem

The process $x \in \mathbb{R}_+ \to R(x) \in \{0, ..., n\}$ *has the following properties:*

1. $R(0) = n$,
2. *R is a non-homogeneous jump–Markov process, whose jump rates are given as follows:*

$$Q_{k,\ell}(x) = \begin{cases} (1 - \frac{1}{2}\exp(-\gamma x))k \times 2\gamma & \text{if } \ell = k-1, \\ \frac{1}{2}\exp(-\gamma x)(n-k) \times 2\gamma & \text{if } \ell = k+1, \\ 0 & \text{if } \ell \notin \{k-1, k, k+1\}. \end{cases}$$

3. $\exists x(\omega) > 0; R(x) = 0 \ \forall x \geq x(\omega)$ *a.s.*

Idea for $Q_{k,k+1}(x)$

Question 4: Evolution of the coalescent tree along the genome?

Wiuf and Hein, 1999: The process of the coalescent trees along the genome is not Markovian.

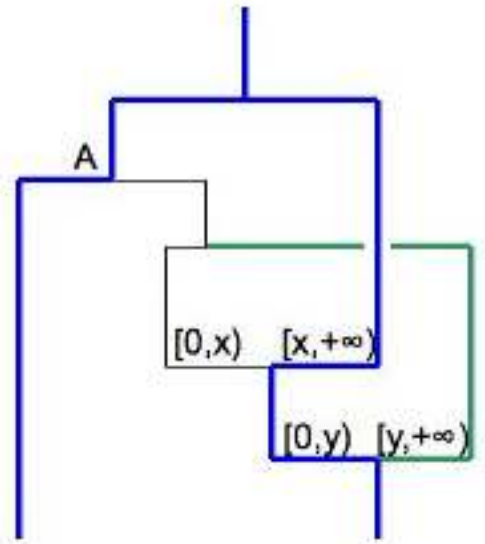
To obtain a Markovian process, we must add all the "past": Ancestral Recombination Graph (ARG).

Wiuf and Hein, 1999: The process of the coalescent trees along the genome is not Markovian.

To obtain a Markovian process, we must add all the "past": Ancestral Recombination Graph (ARG).

Definition

$ARG(u)$ = graph that sums up the genealogy of the sample, implied by coalescence and recombination events on the portion $[0, u]$.



(In the above figure, $0 < x < y$)

Question 4': Evolution of the ARG along the genome?

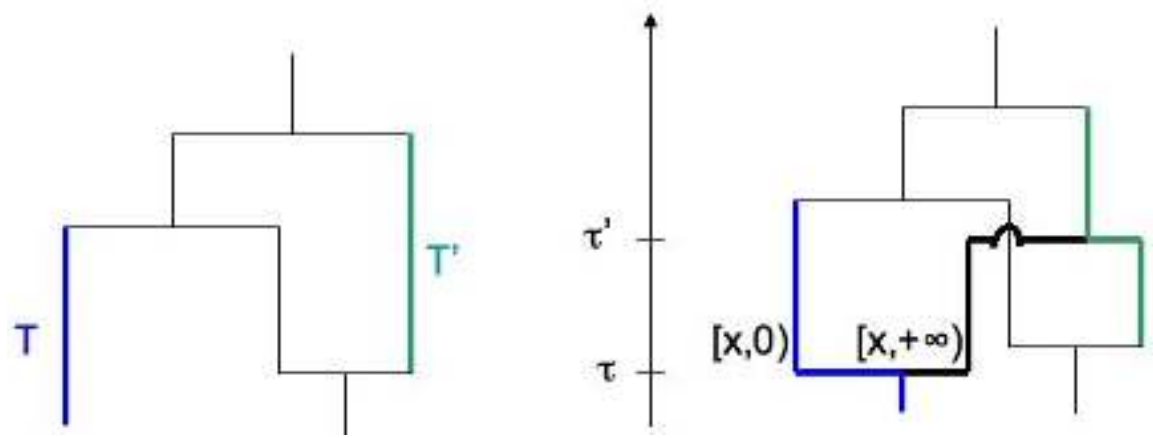
Plan

The ARG is modified:

- when a recombination impacts the ARG during the neutral period that predates the selective sweep (rate $\lambda\times$ total length of the ARG)
- when a recombination occurs during the sweep (rate 2γ)

1) Evolution under neutrality / Recombination during the neutral period:

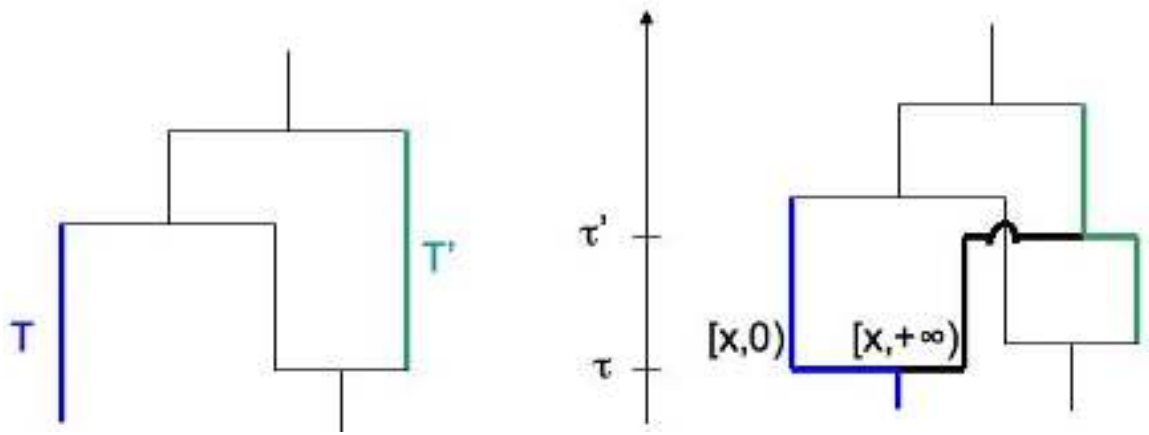
τ: recombination time;
T: branch of the ARG where the recombination happens;
τ': coalescence time of the recombinant lineage;
T': branch of the ARG that coalesces with the recombinant lineage;



$A_t(x)$: number of lineages in ARG(x) at time $t \geq 0$.
$L(x)$: total length of ARG(x)
$H(x)$: height of ARG(x)

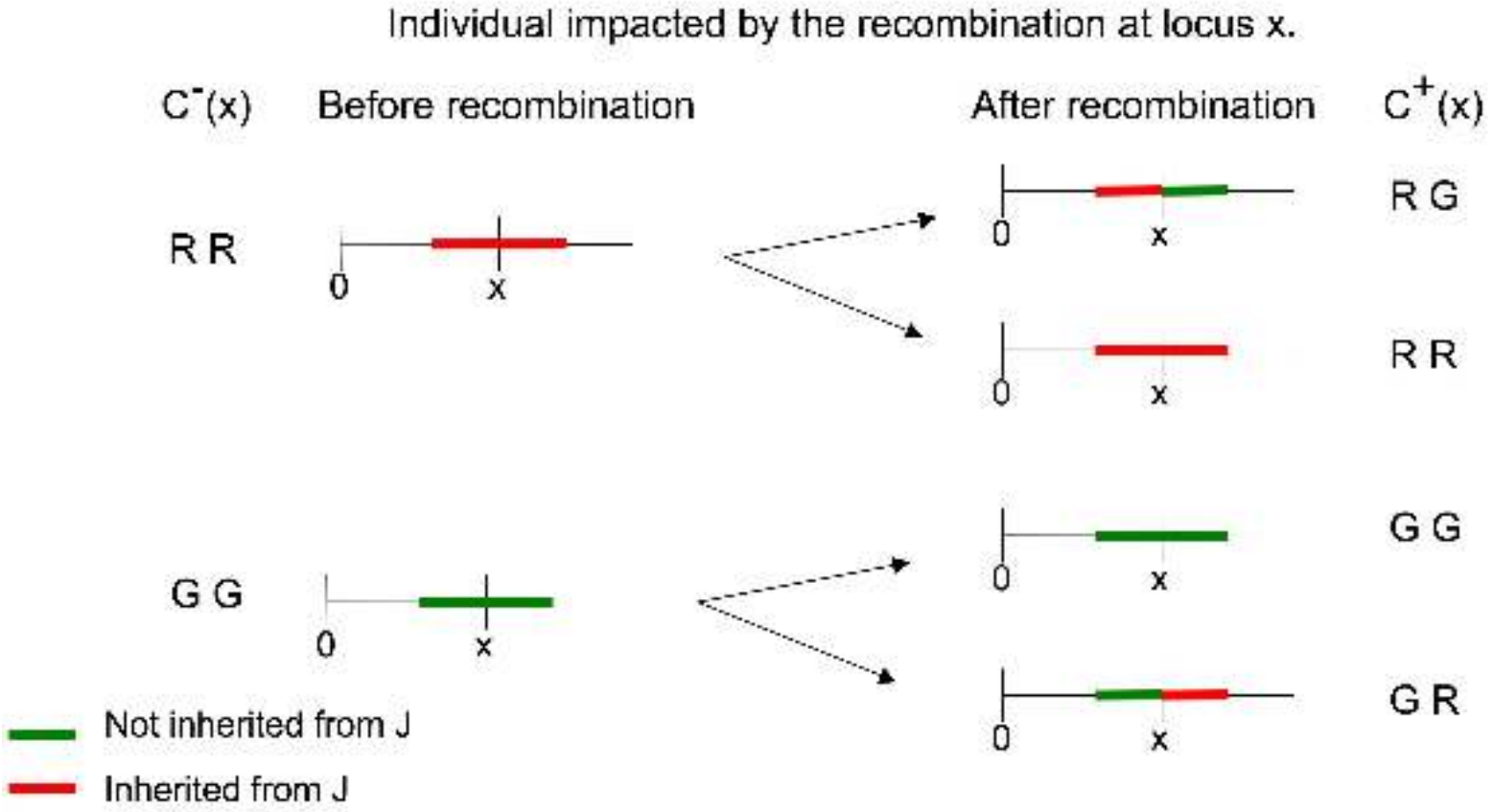
Theorem

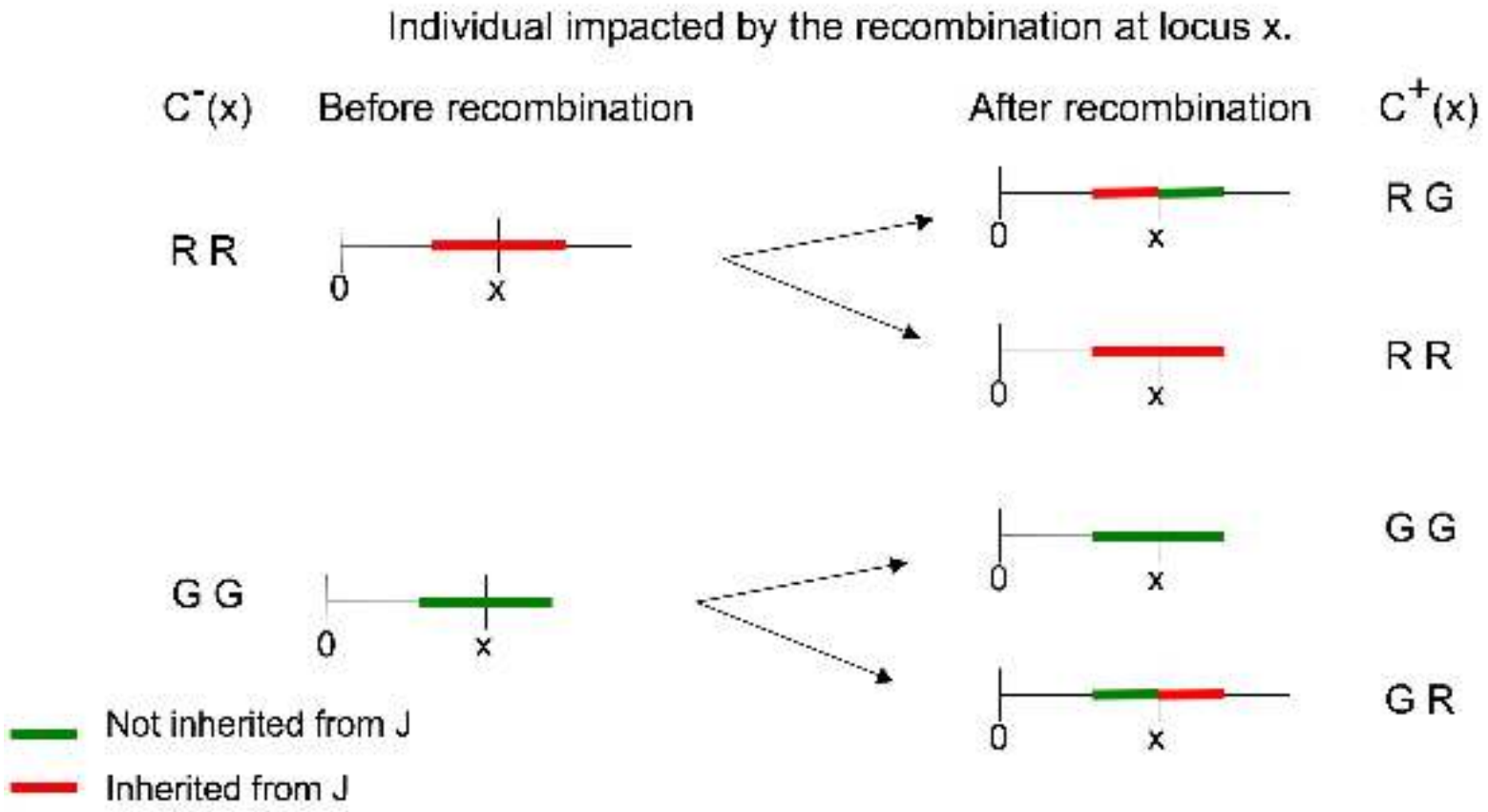
1. *The positions of the recombinations are the jump positions of a Poisson process with intensity* $\lambda \times L(x)$.
2. $\mathbb{P}_{(\tau,\tau',T,T')}(dt, dt', \mathcal{T}, \mathcal{T}') = \dfrac{\exp\left(-\int_t^{t'} A_s(x^-)ds\right)}{\int_0^{H(x^-)} A_w(x^-)dw} 1_{t<t'} dt' dt.$



2) Effect of a recombination during the selective sweep

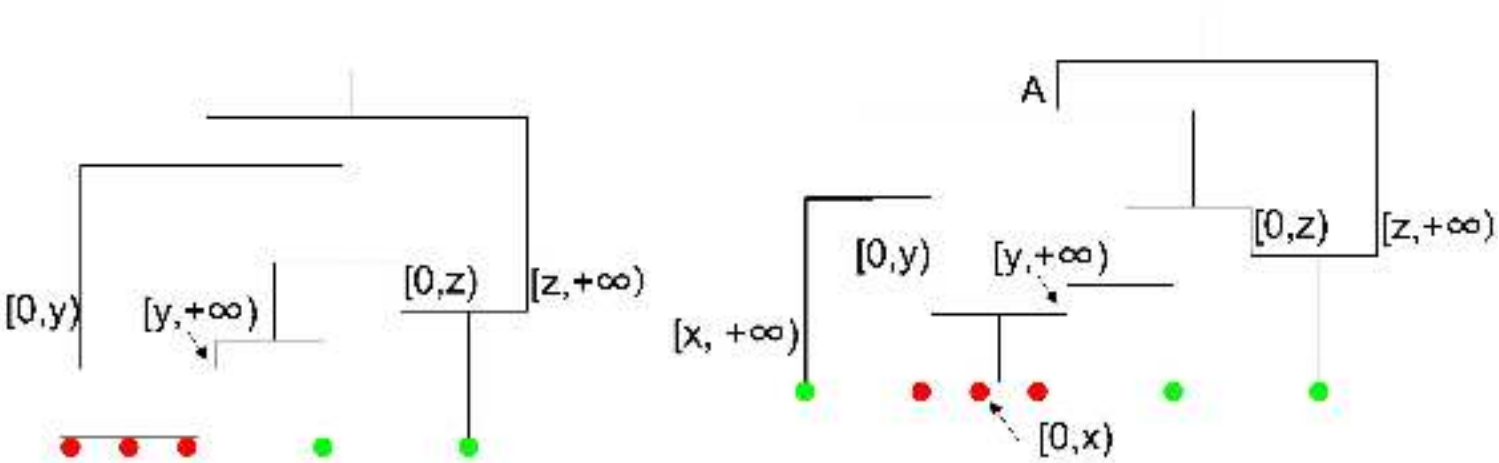
Suppose that a recombination happens at position $x > 0$ during the selective sweep and that $ARG(x^-)$ is given. Then the impact of this recombination is of one of the four types:



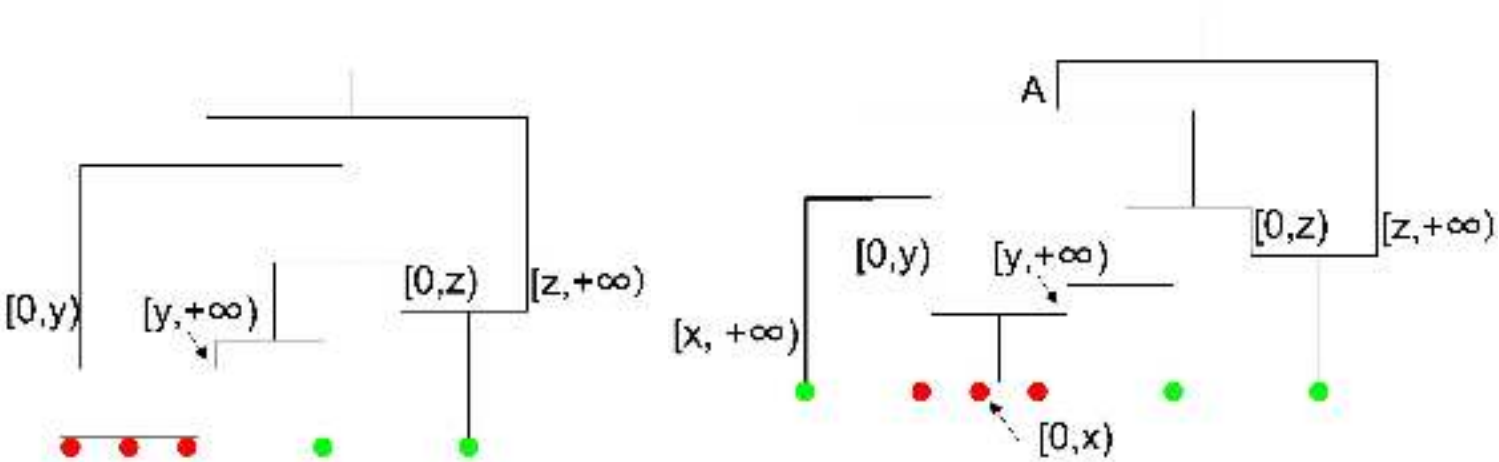


$$\begin{aligned}
& \mathbb{P}(C^+(x) = RG | R(x^-) = k; \text{recomb at } x) \\
= \quad & \mathbb{P}(C^+(x) = RG | C^-(x) = RR; \text{recomb at } x) \mathbb{P}(C^-(x) = RR | R(x^-) = k) \\
= \quad & \left(1 - \frac{1}{2} e^{-\gamma x}\right) \frac{k}{n}
\end{aligned}$$

Modification of the ARG:



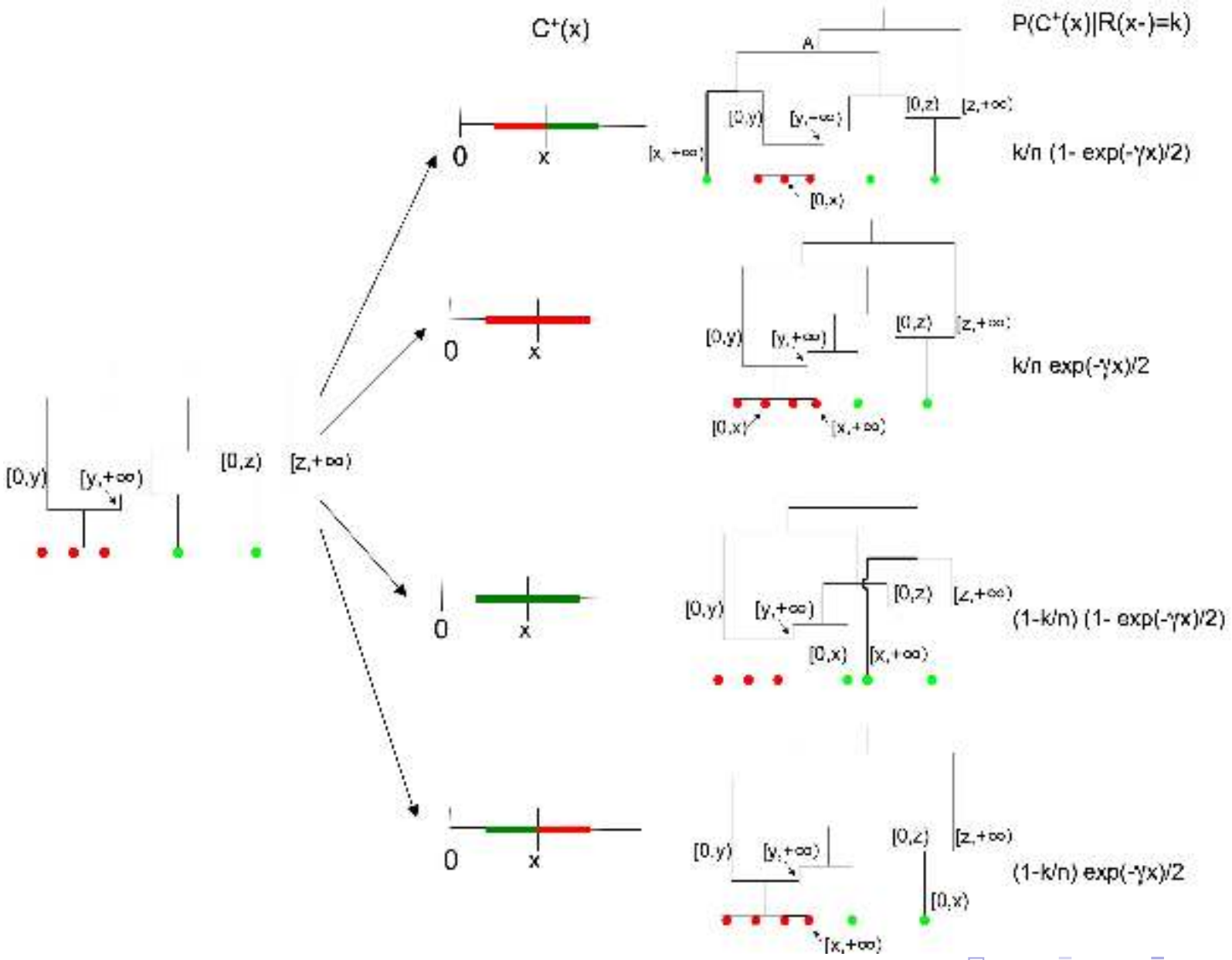
Modification of the ARG:



Let I be the impacted tooth of $\mathcal{P}(x^-)$.

$$\mathbb{P}_{(I,\tau',T')}(\mathcal{I}, dt', \mathcal{T}') = \frac{1}{|\mathcal{P}(x^-)|} \exp\left(-\int_0^{t'} A_s(x^-)ds\right) 1_{t' \in \mathcal{T}'} dt'.$$

$A_t(x)$: number of lineages in ARG(x) at time $t \geq 0$.
Coalescence of the new branch at rate $A_t(x)$.

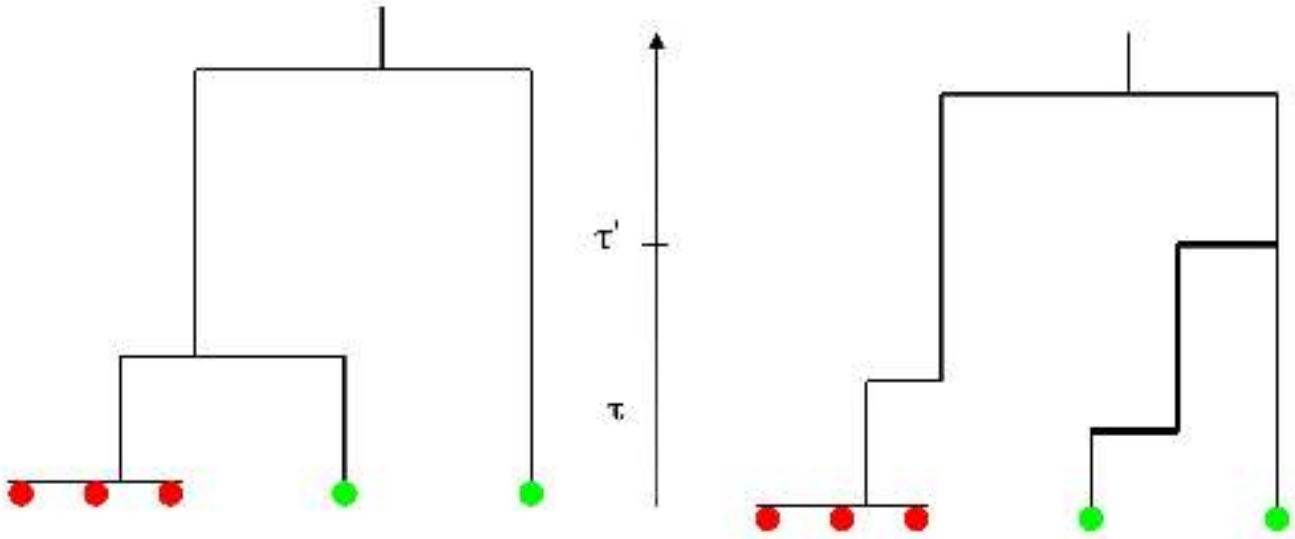

C+(x)
P(C+(x)|R(x-)=k)
A
0
x
[0,y)
[y,+∞)
[0,z)
[z,+∞)
[x,+∞)
[0,x)
k/n (1- exp(-γx)/2)
k/n exp(-γx)/2
(1-k/n) (1- exp(-γx)/2)
(1-k/n) exp(-γx)/2

Theorem (Answer 4': Evolution of the ARG along the genome)

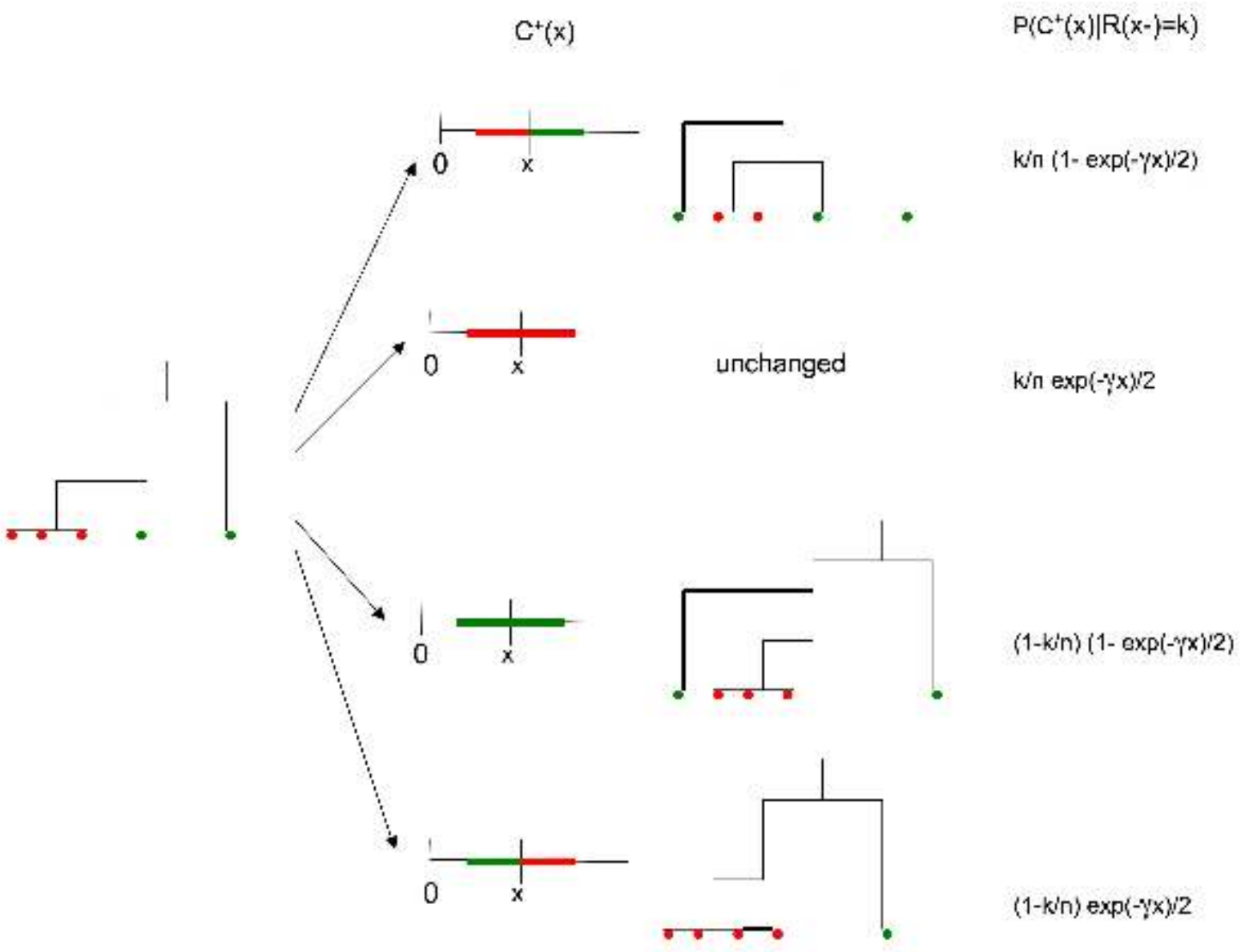
1. *$ARG(0)$ is a comb with n teeth.*
2. *The recombination positions follow a Poisson process with parameter $2n\gamma + \lambda L(x)$, where $L(x)$ is the total length of $ARG(x)$.*
3. *At the position x of a jump:*
   - *With probability $\frac{\lambda L(x^-)}{2n\gamma+\lambda L(x^-)}$, the recombination occurs during the neutral period. The comb is unchanged and the graph is modified as under neutrality.*
   - *With probability $\frac{2n\gamma}{2n\gamma+\lambda L(x^-)}$, the recombination occurs during the selective sweep and the evolution of the comb and the graph is one of the 4 cases presented above.*

Back to question 4 : Evolution of the coalescent tree along the genome?

1) If the recombination occurs during the neutral period, a part of a branch is suppressed and a new one is created and coalesces with another branch of the tree.



2) If the recombination occurs during the selective sweep:



$$\nu(x) := 1 - \frac{R(x^-)}{n}\frac{\exp(-\gamma x)}{2}$$

Theorem (Answer 4: Evolution of the coalescent tree along the genome)

1. *Tree(0) is a comb with n teeth.*
2. *The recombination positions follow a Poisson process with parameter $2n\gamma\nu(x) + \lambda\tilde{L}(x)$, where $\tilde{L}(x)$ is the total length of tree(x).*
3. *At the position x of a jump:*
   - *With probability $\frac{\lambda\tilde{L}(x^-)}{2n\gamma\nu(x)+\lambda\tilde{L}(x^-)}$, the recombination occurs during the neutral period. The comb is unchanged and the tree is modified as under neutrality.*
   - *With probability $\frac{2n\gamma\nu(x)}{2n\gamma\nu(x)+\lambda\tilde{L}(x^-)}$, the recombination occurs during the selective sweep and the evolution of the comb and the tree is one of the 3 cases presented above.*

Current work (with Majid Salamat) :

1. compute the joint law of the various coalescence trees,
2. deduce the joint law of the number of SNIPs on the various segments of the genome between two consecutive recombination locations.

THANK YOU FOR YOUR ATTENTION !