Infering local admixture for polyploid species

S. Rio[1], F. Gauthier[2], T. Mary-Huard[2,3]

1 Agap (CIRAD), 2 GQE Le Moulon, 3 MIA Paris-Saclay

Biological context

1/ Ancestral populations

Consider K (unobserved) populations back in the past.

Each population has its own genetic characteristics.

2/ Migration event

Populations experienced a single event of genetic material exchange

(i.e. migrations).

3/ Admixture

After the migration event, T generations of random mating.

Genetic material from pops $1, .., k-1, k+1, .., K$ segregates in pop k.

4/ Actual populations

Each population k now includes

- “pure” individuals with full genome inherited from ancestral pop k,

- admixed individuals with genome having a mosaic of genomic regions from the different ancestral pops.

One does not observe the populations, but a set of **unlabeled individuals**.

Objectives

Data

A sample of *n* individuals,

Genotyped at *M* biallelic markers (genomic positions),

Phased genotypes (= access to chromosome copies separately).

Inferring global admixture

One needs to estimate

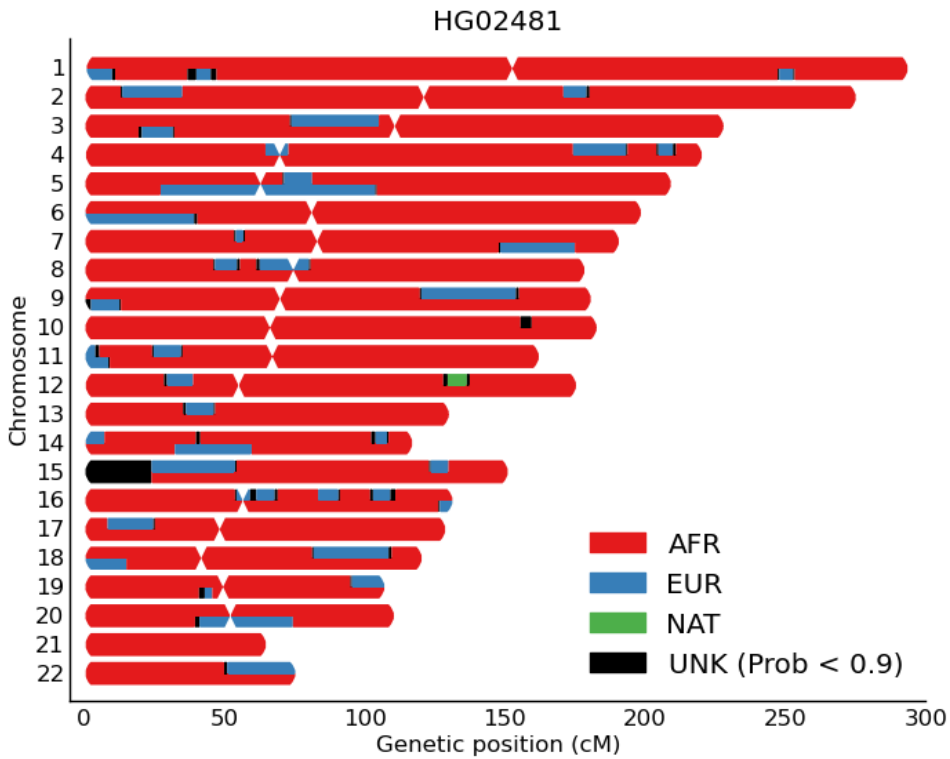
- K, the number of populations,
- $\left(f_{mk}\right)$, the matrix of allelic frequencies,
- $\left(\tau_{ik}\right)$, the ancestry proportions.

$\Rightarrow$ Mixture models

Inferring local admixture

Same + the **haplotypic blocks**

$\Rightarrow$ HMM models



STRUCTURE (Falush et al, 2003)

For individual i and a given chromosome copy, and assuming K is known, define

- $Z_m \in \{1, \ldots, K\}$: ancestry (origin) at position m,
- $G_m \in \{1, \ldots, A_m\}$: allele at position m.

Transition probabilities

$$\mathbb{P}\left(Z_{m+1} = k' \,\middle|\, Z_m = k\right) = \begin{cases} \exp(-\theta d_m) + (1 - \exp(-\theta d_m))\tau_{ik}, & k = k' \\ (1 - \exp(-\theta d_m))\tau_{ik'}, & k \neq k' \end{cases}$$

⋆ d_m: genetic distance between pos. m and $m + 1$ (known),

⋆ θ: recombination rate factor

⋆ τ_{ik}: admixture proportion of ancestral pop. k for ind i.

Emission distribution

$$\mathbb{P}\left(G_m = a \,\middle|\, Z_m = k\right) = f^k_{m,a}$$

⋆ $f^k_{m,a}$: frequency of allele a in ancestral pop. k.

Inference

Through EM algorithm

Haplotypic blocks obtained from posterior probabilities $\mathbb{P}\left(Z_m \,\middle|\, G\right)$.

STRUCTURE (Falush et al, 2003)

For individual i and a given chromosome copy, and assuming K is known, define

- $Z_m \in \{1, \ldots, K\}$: ancestry (origin) at position m,
- $G_m \in \{1, \ldots, A_m\}$: allele at position m.

Transition probabilities

$$\mathbb{P}\left(Z_{m+1} = k' \,\middle|\, Z_m = k\right) = \begin{cases} \exp(-\theta d_m) + (1 - \exp(-\theta d_m))\tau_{ik}, & k = k' \\ (1 - \exp(-\theta d_m))\tau_{ik'}, & k \neq k' \end{cases}$$

⋆ d_m: genetic distance between pos. m and $m + 1$ (known),

⋆ θ: recombination rate factor

⋆ τ_{ik}: admixture proportion of ancestral pop. k for ind i.

Emission distribution

$$\mathbb{P}\left(G_m = a \,\middle|\, Z_m = k\right) = f^k_{m,a}$$

⋆ $f^k_{m,a}$: frequency of allele a in ancestral pop. k.

Inference

Through EM algorithm

Haplotypic blocks obtained from posterior probabilities $\mathbb{P}\left(Z_m \,\middle|\, G\right)$.

Local admixture for polyploid species

Polyploidy

More than 1 (or 2) copy of each chromosome (sugar cane: $P = 12$).

Genotyping through sequencing

Only access to **aggregated** genotypic data at each position.

Ex: $P = 12$, $A_m = 5$

$$G_m = \{3, 2, 2, 4, 1\}$$

What's new ?

Requires to follow ancestry on all copies **simultaneously**.

$$\mathbb{P}\left(Z_{m+1} = c' \mid Z_m = c\right)$$

$\star$ $c \in \mathcal{C}$: ancestry configuration

Ex: $P = 12$, $K = 4 \Rightarrow |\mathcal{C}| = K^P = 16,777,216$

$\Rightarrow$ Work with unordered configurations, i.e. ancestry dosages ($|\mathcal{C}| = 455$).

Quizzzz

Order these species w.r.t their ploidy levels

Crapaud Batura

Ble tendre

Fraisier

Transition matrices (1/2)

One needs to compute for each m

$$T_m[u, u'] = \mathbb{P}\left(Z_{m+1=u'} \ \middle| Z_m = u\right), \quad u, u' \in \mathcal{U}.$$

Step 1: conditioning

Let R_m be the number of recombination events at position m.

One has

$$T_m \quad = \quad \sum_{r=0}^{P} \mathbb{P}\left(R_m = r\right) \times T_{(r)}$$

where $T_{(r)}[u, u'] = \mathbb{P}\left(Z_{m+1} = u' \ \middle| Z_m = u, R_m = r\right)$.

Still...

$\star$ One needs to compute $T_{(r)}[u, u']$

$\star$ $P = 12, K = 8 \Rightarrow |\mathcal{U}| = 50,388$

Transition matrices (2/2)

Step 2: Approximation

$$T_m \quad = \quad \frac{1}{\mathbb{P}\left(R_m \leqslant R_{\max}\right)} \sum_{r=0}^{R_{\max}} \mathbb{P}\left(R_m = r\right) \times T_{(r)}$$

For small values of $R_{\max}$, $T_{(r)}$ is **sparse** !

Step 3: Computing $T_{(r)}\left[u, u'\right]$

Let $u = \left[u_1, \ldots, u_K\right]$, $u' = \left[u'_1, \ldots, u'_K\right]$, one has

$$T_{(1)}\left[u, u'\right] = I_{\left\{u \middle|\middle| u'=0\right\}} \sum_k \frac{u_k}{P} \tau_k \; + \; I_{\left\{u \middle|\middle| u'=1\right\}} \left(\sum_{k \neq k'} I_{\left\{u_k - u'_k = 1\right\}} I_{\left\{u_{k'} - u'_{k'} = -1\right\}} \frac{u_k}{P} \tau_{k'} \right)$$

where $u \middle|\middle| u'$ is the number of visible recombinations between u and u'.

Doable for small values ($R_{\max} \leqslant 2$).

Emission distribution

Denote

⋆ A_m be the number of possible alleles at position m,

⋆ a an allelic configuration at position m,

⋆ a^S a **sorted** allelic configuration compatible with a,

One has

$$\mathbb{P}\left(G_m = a \,\middle|\, Z_m = u\right) \propto \sum_{c \equiv u} \prod_{p=1}^{P} f_{m,a_p^s}^{c_p}$$

Multinomial approximation

$$G_m | Z_m = u \approx \mathcal{M}\left(P, \gamma(m,u,1), \ldots, \gamma(m,u,A_m)\right)$$

where

$$\gamma(m,u,a) = \sum_{k=1}^{K} \frac{u_k}{P} f_{m,a}^k$$

Emission approximation (example)

Let $K = 2$, $P = 2$, $A = 2$.

Z	G	True	Approx
(2,0)	(2,0)	$\left(f_1^1\right)^2$	$\left(f_1^1\right)^2$
(2,0)	(1,1)	$2f_1^1 f_2^1$	$2f_1^1 f_2^1$
(2,0)	(0,2)	$\left(f_2^1\right)^2$	$\left(f_2^1\right)^2$
(1,1)	(2,0)	$f_1^1 f_1^2$	$\left(\frac{1}{2}f_1^1 + \frac{1}{2}f_1^2\right)^2$
(1,1)	(1,1)	$f_1^1 f_2^2 + f_1^2 f_2^1$	$2\left(\frac{1}{2}f_1^1 + \frac{1}{2}f_1^2\right)\left(\frac{1}{2}f_2^1 + \frac{1}{2}f_2^2\right)^2$
(1,1)	(0,2)	$f_2^1 f_2^2$	$\left(\frac{1}{2}f_2^1 + \frac{1}{2}f_2^2\right)^2$
(0,2)	(2,0)	$\left(f_1^2\right)^2$	$\left(f_1^2\right)^2$
(0,2)	(1,1)	$2f_1^2 f_2^2$	$2f_1^2 f_2^2$
(0,2)	(0,2)	$\left(f_2^2\right)^2$	$\left(f_2^2\right)^2$

index m omitted for ease

Simulations

Parameters

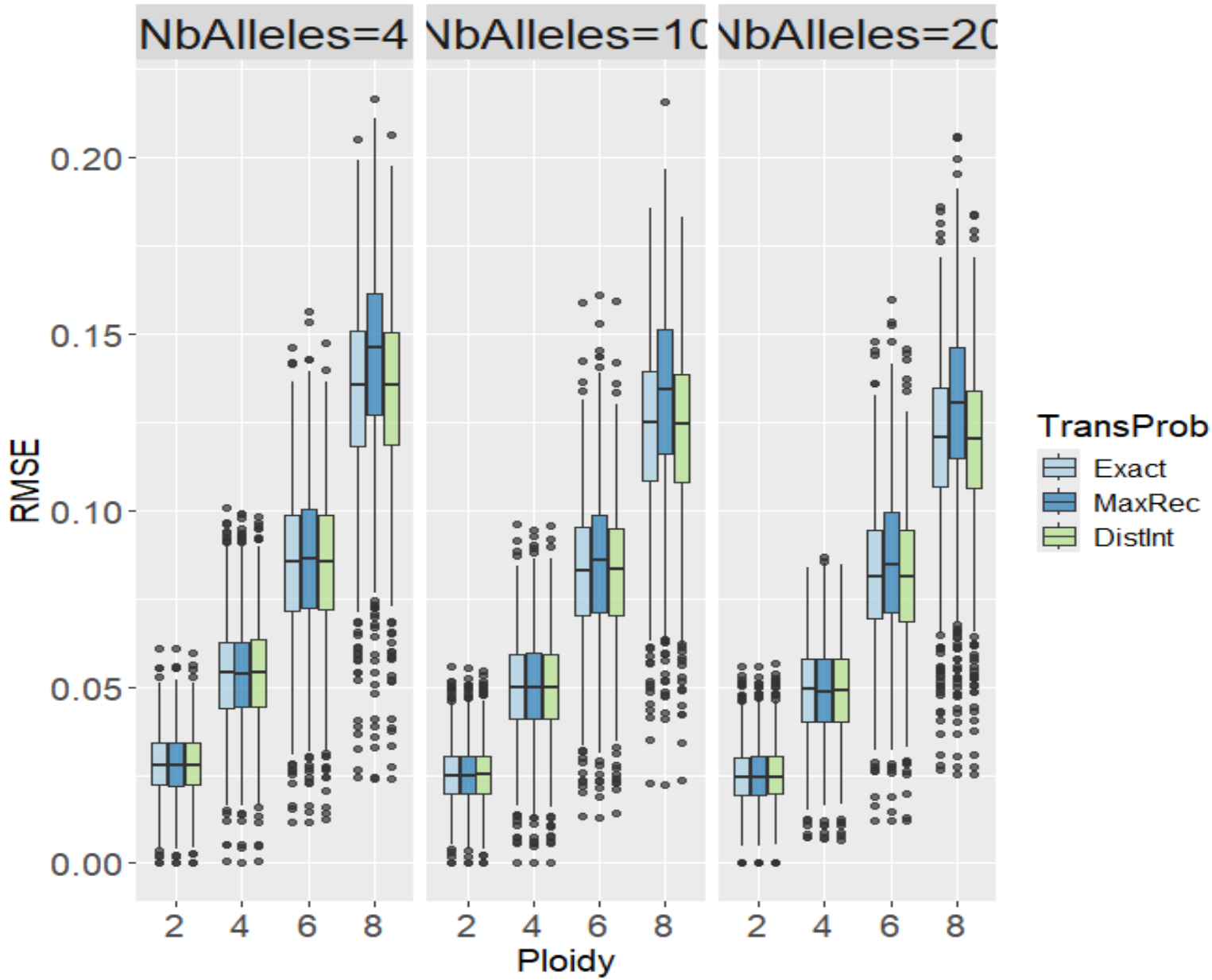
- $P = 2, 4, 6, 8$ *(Ploidy level)*
- $M = 2000$ *(Nb positions)*
- $A = 4, 10, 20$ *(Nb alleles per position)*
- $K = 4$ *(Nb ancestral pops)*

Evaluation

- Generate some chromosomes
- Perform the E step with the true parameters
- Using the different tricks (or not)
- Compare the estimated ancestry dosages with the true ones

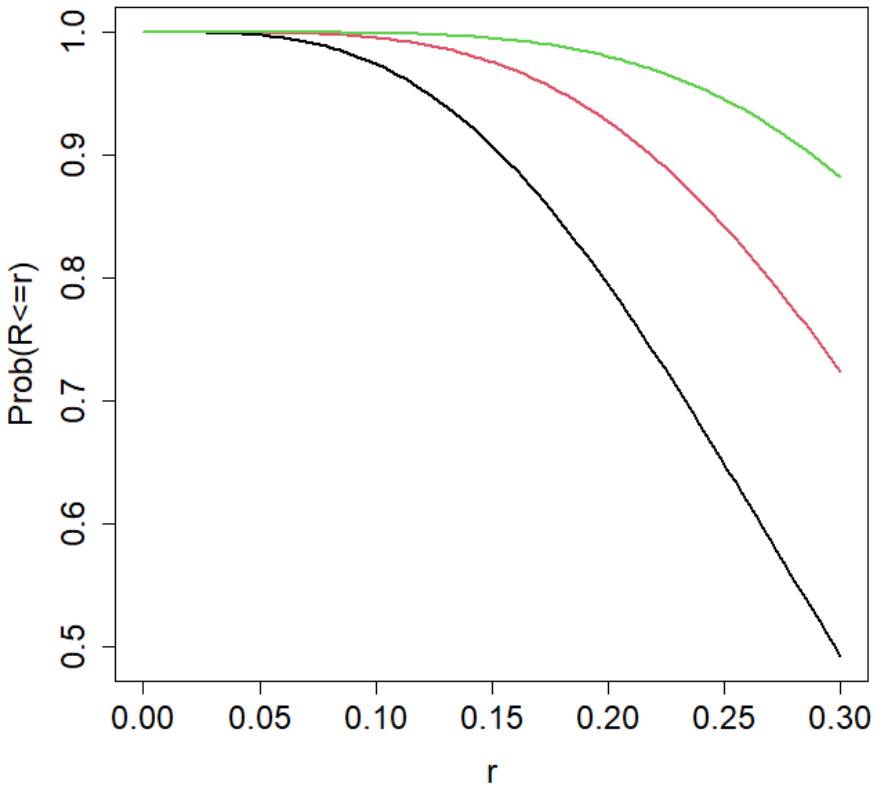
$$RMSE = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \left(u_{mk} - \hat{u}_{mk}\right)^2}$$

Recombination approximation



Genetic distances in sugar cane

Prob. mass vs genetic distance

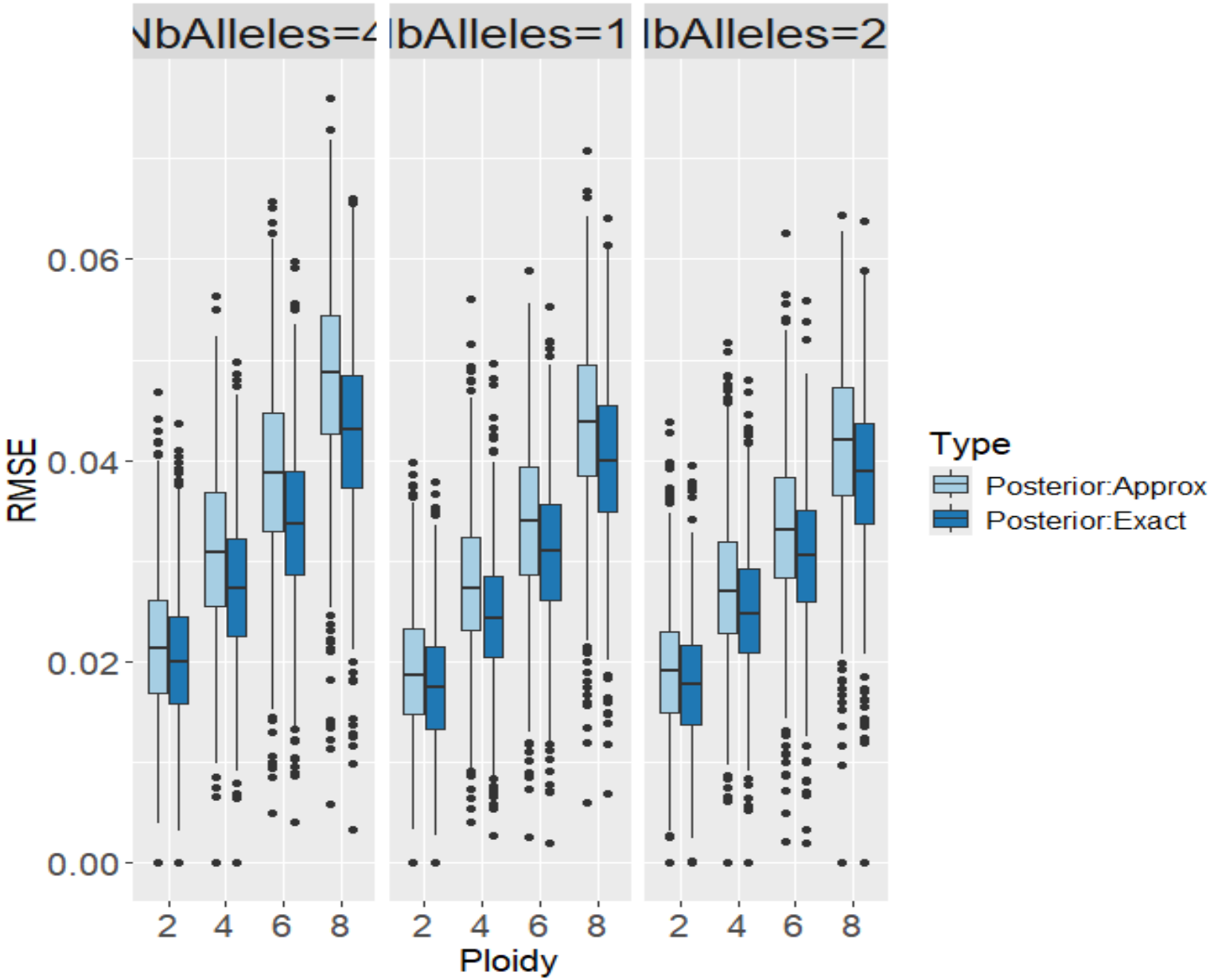


NbMax Recomb: 1, 2, 3

Genetic distance quantiles

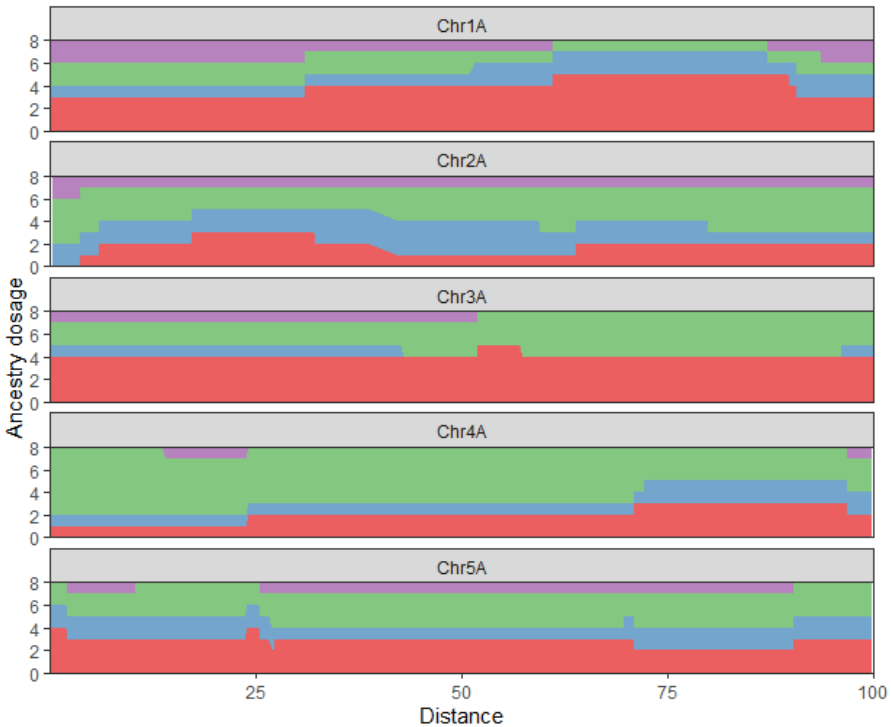
Chrom.	95%	99%	99.9%
1A	0.0320	0.0847	0.2547
2A	0.0389	0.0923	0.2986
3A	0.0488	0.1304	0.3100
4A	0.0609	0.1836	0.6515
5A	0.0475	0.1232	0.3433
6A	0.0746	0.1852	0.7828
7A	0.0614	0.2067	0.5594
8A	0.0872	0.2300	0.6756
9A	0.1183	0.3515	1.4560
10A	0.1456	0.3896	1.0493

Emission approximation

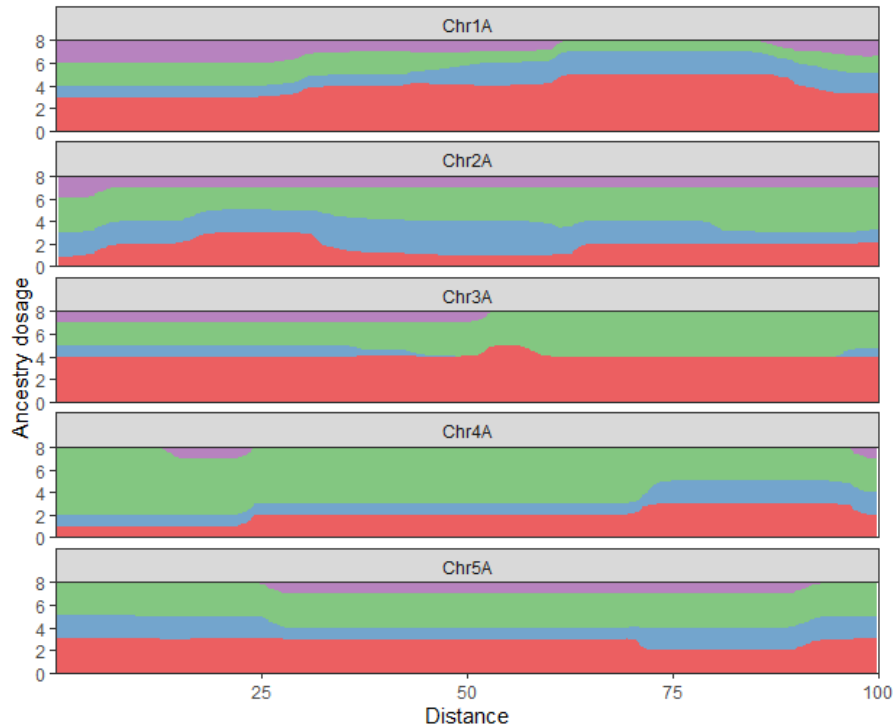


Local ancestry reconstruction

True dosages



Estimated dosages



Conclusions

About local admixture

1/ Some more work for the recombination trick,

2/ Emission approx seems $\approx$ OK for now,

3/ Implementation in R, using Rcpp and parallelization (over chrom and ind).

About global admixture

1/ A proved phased/unphased equivalence,

2/ Implicit E step (i.e. no storage of the posterior matrix),

3/ Implementation similar to local admixture (Rcpp + parallel).