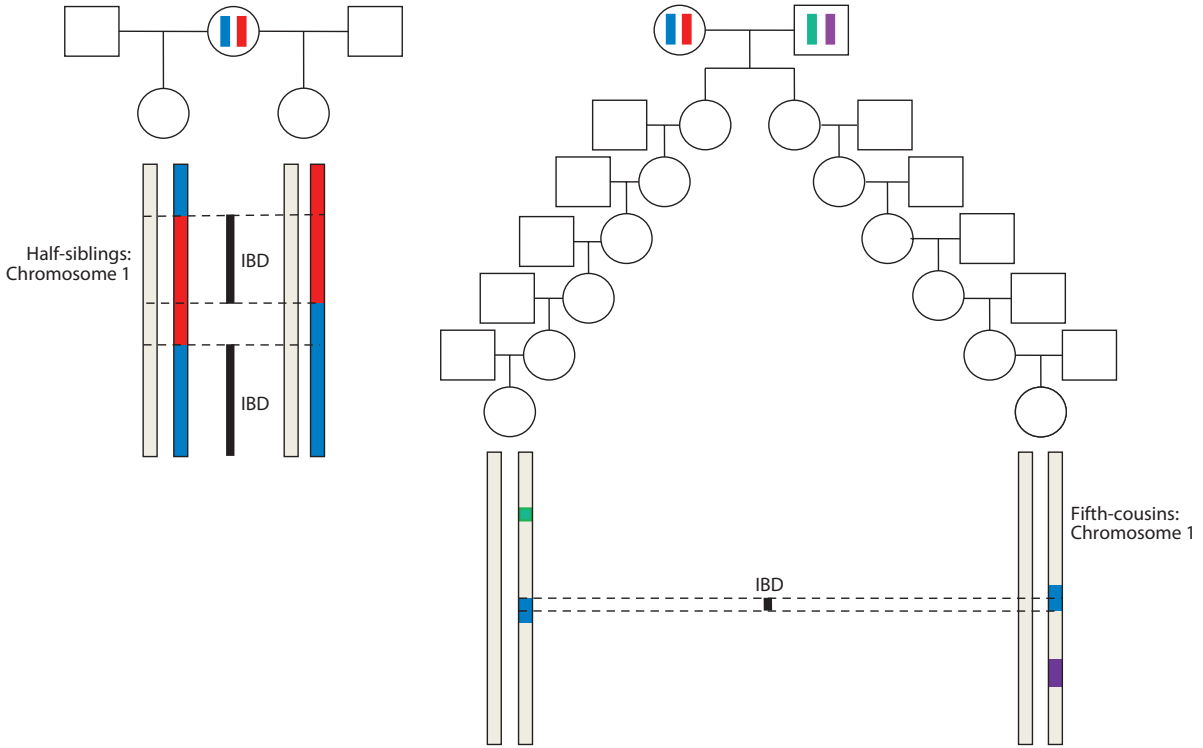
Identity by descent

Consanguinité

entre deux individus reliés, on copte sur l'identité de haplotypes par descente



Browning & Browning (2012)

Segments IBD

centiMorgan : distance sur laquelle 0.01 recombinaisons occurrent en une génération

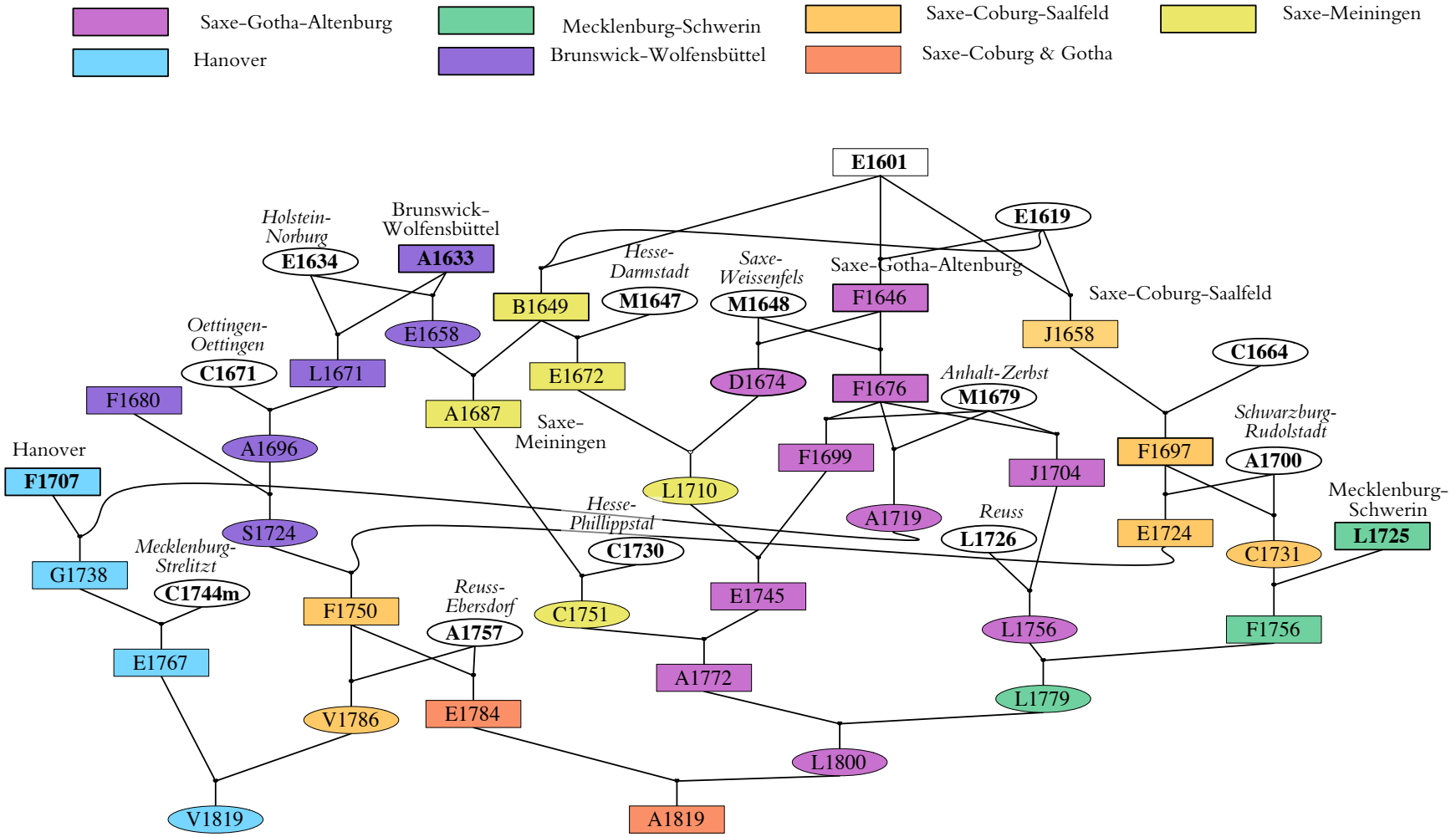
génome humain : $10^6\,\mathsf{pb} \approx 1\,\mathsf{cM}$

distance de m générations : proportion de génome partagée $1/2^{m-1}$
longueur de segment : $100/m\,\mathsf{cM}$

$\Rightarrow$ cousins de 5e degré : partagé $\frac{3000\,\mathsf{Mb}}{2^{11}} \approx 1.5\,\mathsf{cM}$ en moyenne venant de chaque grand^{5}parent, de longueur $8.3\,\mathsf{cM}$ en moyenne ; aucun segment en commun avec $\geq 65\%$ probabilité :

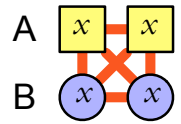
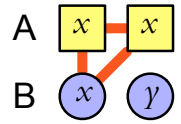
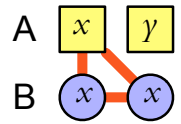
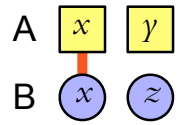
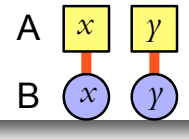
$$\frac{2 \cdot 1.5}{3.5} = 0.35$$

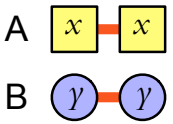
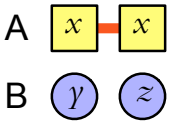
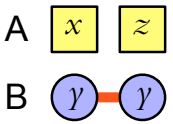
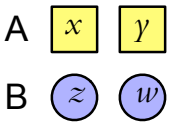
Reine Victoria et Prince Albert



Csuros (2014)

Mode d'identité (Jacquard)

IBD mode	A's genotype	B's genotype	identity coefficient
	x/x	x/x	Δ_1
	x/x	x/y	Δ_3
	x/y	x/x	Δ_5
	x/y	x/z	Δ_8
	x/y	x/y	Δ_7

IBD mode	A's genotype	B's genotype	identity coefficient
	x/x	y/y	Δ_2
	x/x	y/z	Δ_4
	x/z	y/y	Δ_6
	x/y	z/w	Δ_9

IBD → IBS

(identity-by-descent et identity-by-state)

Table 2 | **Joint genotypic probabilities**

	Genotypes	Genotypic state	Number of shared alleles	General	Non-inbred
1	A_iA_i, A_iA_i	Hom/hom	2	$\Delta_1 P_i + (\Delta_2 + \Delta_3 + \Delta_5 + \Delta_7)P_i^2 + (\Delta_4 + \Delta_6 + \Delta_8)P_i^3 + \Delta_9 P_i^4$	$k_2 P_i^2 + k_1 P_i^3 + k_0 P_i^4$
2	A_iA_i, A_jA_j	Hom/hom	0	$\Delta_2 P_i P_j + \Delta_4 P_i P_j^2 + \Delta_6 P_i^2 P_j + \Delta_9 P_i^2 P_j^2$	$k_0 P_i^2 P_j^2$
3	A_iA_i, A_iA_j	Hom/het	1	$\Delta_3 P_i P_j + (2\Delta_4 + \Delta_8)P_i^2 P_j + 2\Delta_9 P_i^3 P_j$	$k_1 P_i^2 P_j + 2k_0 P_i^3 P_j$
4	A_iA_i, A_jA_m	Hom/het	0	$2\Delta_4 P_i P_j P_m + 2\Delta_9 P_i^2 P_j P_m$	$2k_0 P_i^2 P_j P_m$
5	A_iA_j, A_iA_j	Het/het	2	$2\Delta_7 P_i P_j + \Delta_8 P_i P_j (P_i + P_j) + 4\Delta_9 P_i^2 P_j^2$	$2k_2 P_i P_j + k_1 P_i P_j (P_i + P_j) + 4k_0 P_i^2 P_j^2$
6	A_iA_j, A_iA_m	Het/het	1	$\Delta_8 P_i P_j P_m + 4\Delta_9 P_i^2 P_j P_m$	$k_1 P_i P_j P_m + 4k_0 P_i^2 P_j P_m$
7	A_iA_j, A_mA_l	Het/het	0	$4\Delta_9 P_i P_j P_m P_l$	$4k_0 P_i P_j P_m P_l$

The table shows seven distinct patterns of genotypes that are possible for two unordered individuals, and the probabilities of these pairs of genotypes in general, or assuming no inbreeding. Two genotypes could be homozygous (hom) for the same or different alleles (rows 1 and 2), one could be homozygous and the other heterozygous (het) with one or zero shared alleles with the homozygote (rows 3 and 4), or both individuals could be heterozygous with two, one or zero shared alleles (rows 5–7). There are nine pairs of genotypes if the ordering of individuals is important (not shown), as the genotypes in rows 3 and 4 (one homozygote and one heterozygote) each have two orders. k_i, the probability of sharing i number of alleles that are identical-by-descent (where $i = 0$–2; see also FIG. 1); P, allele frequency; Δ_{1-9}, Jacquard coefficients, which are measures of identity-by-descent status (BOX 1; FIG. 1).

→ probabilité d'émission pour HMM avec états Δ_i

Weir & al. (2012)

Détecter les segments IBD

Problèmes

I. On a 2 individus — on veut identifier le niveau exact de parenté :
kcoeffs, CARROT

II. Identifier des liens de parenté dans une population : on a n génomes (diploides) — on veut identifier les paires d'individus avec segments IBD :
fastIBD, GERMLINE, ...

Quantification

coefficient de consanguinité γ (*inbreeding coefficient*) :
probabilité d'IBD entre les 2 allèles du même individu

modèle iid (allèle A avec p, a avec $q = 1-p$) + consanguinité : proba de génotypes non-ordonnés

$$\phi\Big(\mathsf{Aa}\Big) = 2(1-\gamma)pq$$
$$\phi\Big(\mathsf{AA}\Big) = p^2\left(1+\gamma\frac{q}{p}\right)$$
$$\phi\Big(\mathsf{aa}\Big) = q^2\left(1+\gamma\frac{p}{q}\right)$$

avec $\gamma = 0$: coefficients de Cotterman pour fréquences de $\mathsf{IBD0}$, $\mathsf{IBD1}$, $\mathsf{IBD2}$

Méthode de Lee

homozygotes discordantes ($\mathcal{D}$) : $(\mathsf{AA}, \mathsf{aa})$ ou $(\mathsf{aa}, \mathsf{AA})$
hétérozygotes concordantes ($\mathcal{C}$) : $(\mathsf{Aa}, \mathsf{Aa})$

probabilités sans IBD : $d = \mathbb{P}\mathcal{D} = 2p^2q^2$, $c = \mathbb{P}\mathcal{C} = 4p^2q^2$

compter seulement les sites $\mathcal{C}$ ou $\mathcal{D}$: $X_1, X_2, \ldots, X_n \in \{0, 1\}$ où $X_i = 1$ dénote des hétérozygotes concordantes

définir le compte $N_{\mathcal{C}} = \sum_{i=1}^{n} X_i$; on a

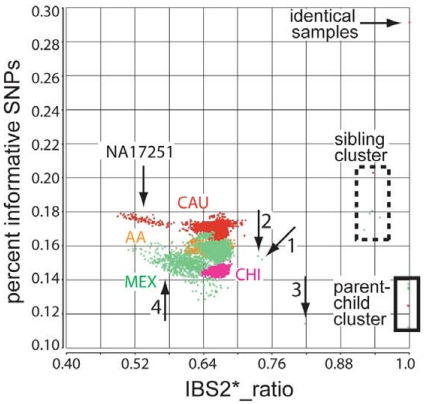
$$\mathbb{E}N_{\mathcal{C}} = \frac{2n}{3}; \qquad \mathsf{Var}N_{\mathcal{C}} = \frac{2n}{9}$$

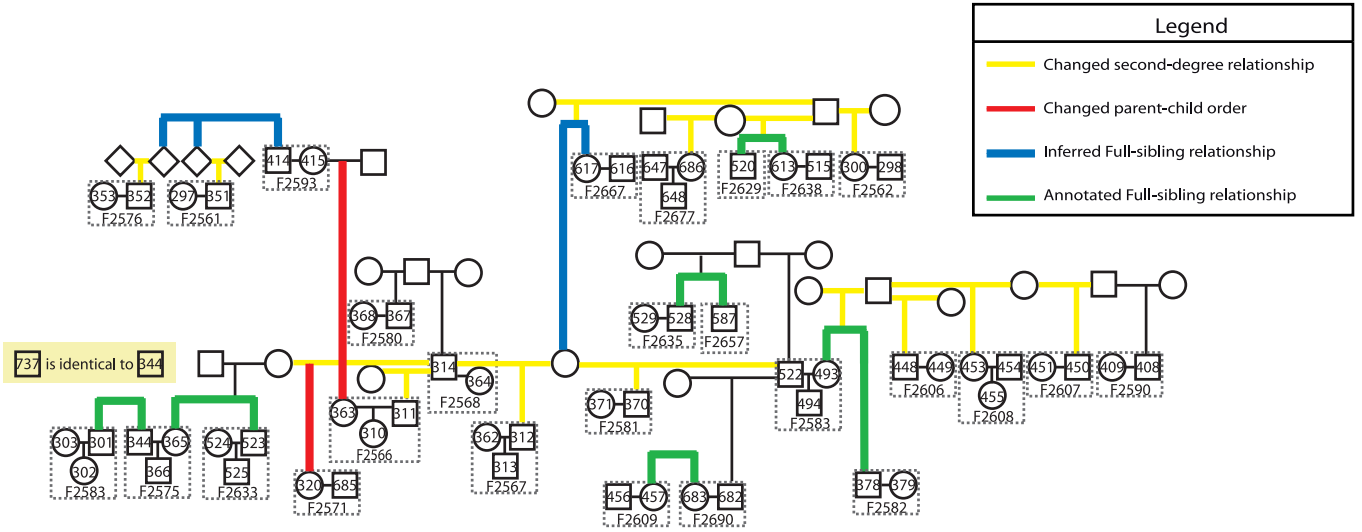
$\Rightarrow$ test d'hypothèse : $\frac{N_{\mathcal{C}}}{n} \sim \mathcal{N}\left(2/3, \frac{\sqrt{2}}{3n}\right)$ si indépendents
si IBD, alors $\mathbb{E}N_{\mathcal{C}} > 2n/3$;
si populations différentes (aucun IBD, p différentes), $\mathbb{E}N_{\mathcal{C}} < 2n/3$

Lee *Annals of Human Genetics*, 67 :618 (2003)

kcoeff

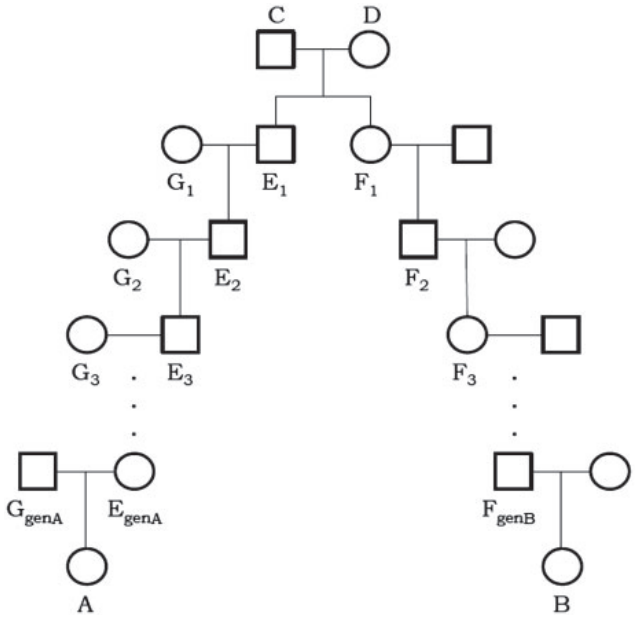
distribution jointe de IBS2*_ratio $= \frac{N_{\mathcal{C}}}{N_{\mathcal{C}}+N_{\mathcal{D}}}$ et $\frac{N_{\mathcal{C}}+N_{\mathcal{D}}}{N_{\mathcal{C}}+N_{\mathcal{D}}+N_{autres}}$





Stevens & al *PLoS Genetics*, 7 :e1002287 (2011)

Segments à fine échelle



HMM avec état [$\emptyset$ consanguinité]
comme vecteurs de 8 indicateurs :
$m_C \in \{0,1\}$: E_1 et F_1 héritent le même haplotype du parent C ; m_D
$m_E \in \{0,1\}$: E_2 reçoit l'allèle de C ; m_F (reçoit de D)
$d_E = 0$ si $E_1 \rightarrow E_2 \cdots \rightarrow E_k$
p_A : phase de A/E_{gen_A}

Transitions : par coordonnées, indépendamment !

Kyriazopoulou-Panagiotopoulou & al *ISMB* 2011

CARROT

Variable	$Pr(0 \to 0)$	$Pr(1 \to 0)$
m_C	$\theta_k^2 + (1-\theta_k)^2$	$2\theta_k(1-\theta_k)$
m_{E_1}	$1-\theta_k$	θ_k
d_A	$(1-\theta_k)^{\text{gen}_A - 1}$	$\left(\sum_{n=1}^{\text{gen}_A - 1} \binom{\text{gen}_A - 1}{n} \theta_k^n (1-\theta_k)^{\text{gen}_A - 1 - n}\right) / \left(2^{\text{gen}_A - 1} - 1\right)$
p_A	$1-\omega$	ω

Degree	Relationship	(mrcas, gen_A, gen_B)
1	Full siblings	(2, 0, 0)
	Parent–child	(1, −1, 0)
2	Half siblings	(1, 0, 0)
	Aunt-niece	(2, 0, 1)
	Avuncular	(2, 0, 1) or (2, 1, 0)
	Grandparent–grandchild	(1, −1, 1)
3	First cousins	(2, 1, 1)
	Great grandparent–grandchild	(1, −1, 2)
	Great aunt–niece	(2, 0, 2)
	Half aunt–niece	(1, 0, 1)
4	Half first cousins	(1, 1, 1)

- ⋆ émissions selon qualité des bases de sequençage
- ⋆ classification de segments (fenêtre) par probabilité postérieure d'IBD

Rel.	2,0,2	2,1,1	2,2,0	1,−1,2	1,0,1	1,1,0	1,2,−1
2,0,2	**88**	–	–	5	5	2	–
2,1,1	–	**84**	1	–	7	8	–
2,2,0	–	1	**88**	–	2	4	5
1,−1,2	2	–	–	**96**	1	1	–
1,0,1	–	11	–	–	**68**	21	–
1,1,0	–	8	–	–	24	**68**	–
1,2,−1	–	–	3	–	1	1	**95**

The value at row i and column j is the percentage of pairs of relationship i that were predicted to be of relationship j. (2, 0, 2): great aunt–niece; (2, 1, 1): first cousins; (2, 2, 0): great niece–aunt; (1, −1, 2): great grandparent–grandchild, (1, 0, 1): half aunt–niece; (1, 1, 0): great niece–aunt; (1, 2, −1): great grandchild–grandparent.

Kyriazopoulou-Panagiotopoulou & al *ISMB* 2011

Segments IBD dans une population

(méthode de GERMLINE : segments longues)

On a des haplotypes [phases !] : matrice binaire $\mathbf{H}[1..2n][1..s]$ avec haplotypes sur s sites dans n génomes

segment IBD : $\mathbf{H}[i][j..j'] = \mathbf{H}[i'][j..j']$

But : identifier les segments IBD les plus longs

1. groupage de haplotypes identiques (tableau de hachage) dans bloc

2. fusion de paires dans blocs consécutifs ; maintenir debut de segment pour la paire

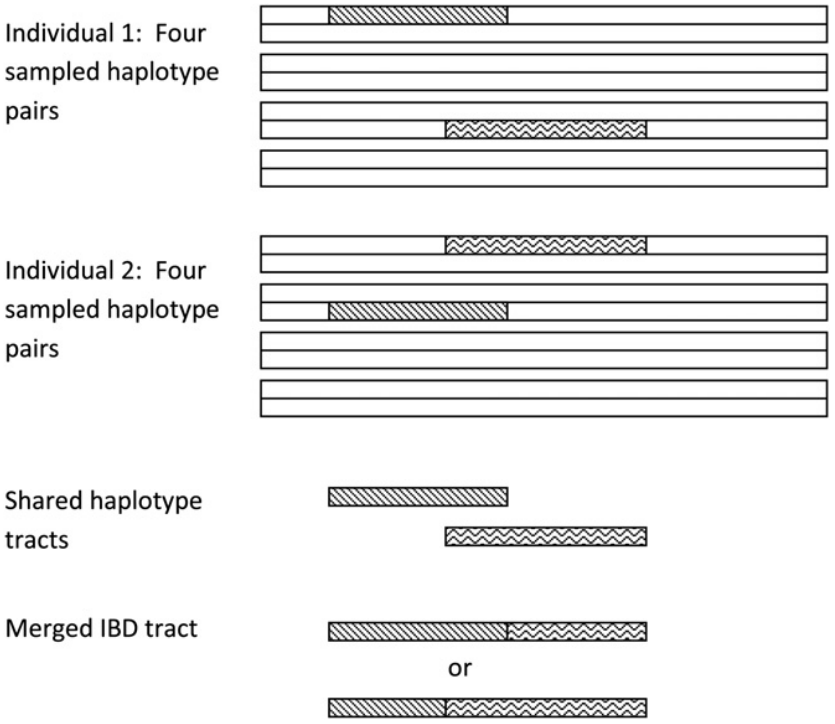
paires de haplotypes (i, i') avec bloc $(k-1)$ IBD : soit extension à bloc k (avec erreurs permises), soit tester longueur (dépasse L_{min} ?)

Gusev & al *Genome Research* 19 :318 (2009)

Segments IBD dans une population 2

Beagle et fastIBD : modèle LD (*linkage disequilibrium*) pour haplotypes

HMM : états = allèle 0/1 dans le haplotype ; transitions par fréquence dans les haplotypes ; émissions incluent erreur de séquençage



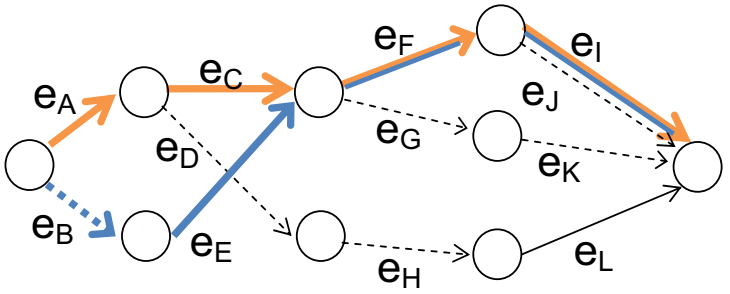


Figure 1. Example of an LD Model on Four SNPs
SNP 1 is represented by edges e_A and e_B; SNP 2 by edges e_C, e_D, e_E; SNP 3 by edges e_F, e_G, e_H; and SNP 4 by edges e_I, e_J, e_K, e_L. For each SNP, allele 1 is represented by a solid line, whereas allele 2 is represented by a dashed line. Haplotype H1 (1 1 1 1) follows the orange path (e_A, e_C, e_F, e_I), and haplotye H2 (2 1 1 1) follows the blue path (e_B, e_E, e_F, e_I).

Browning & Browning *American Journal of Human Genetics* (2010,2011)

Groupage de haplotypes localisés

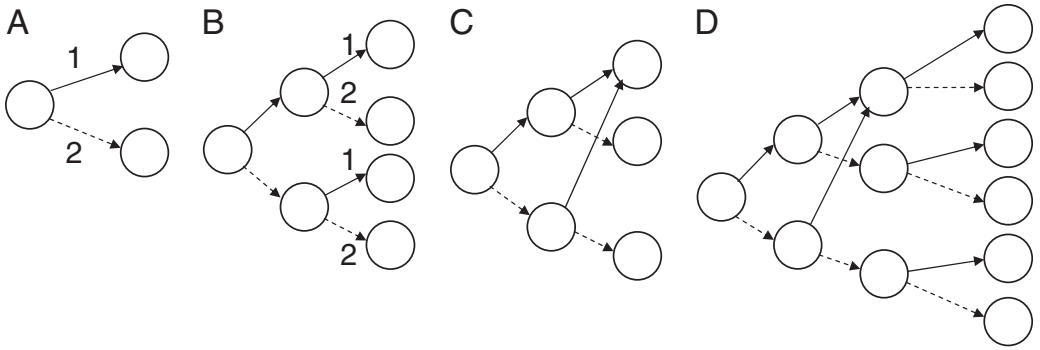


Fig. 1. Initial steps in procedure for creating localized haplotype clusters. (A) Starting with all haplotypes in a single node, the node is split by the alleles at the first marker (a diallelic marker is shown here, however the method can be used with multiallelic markers). The resulting nodes may be merged if the merging criterion is met (for this example the nodes are not merged). (B) The nodes are split by the alleles at the second marker. The four nodes represent the four possible haplotypes when considering the first two markers. (C) Pairs of nodes may be merged. In this example, the first and third nodes are merged. The resulting merged node corresponds to all haplotypes with allele 1 at the second marker. (D) The three nodes are split by the third marker. The process continues by considering merging the resulting nodes, then splitting by the fourth marker and so on.

(fusionner les haplotypes les plus proches selon quelques critères)

Browning & Browning *Genetic Epidemiology* 31 :365 (2007)

Détection de segments IBD

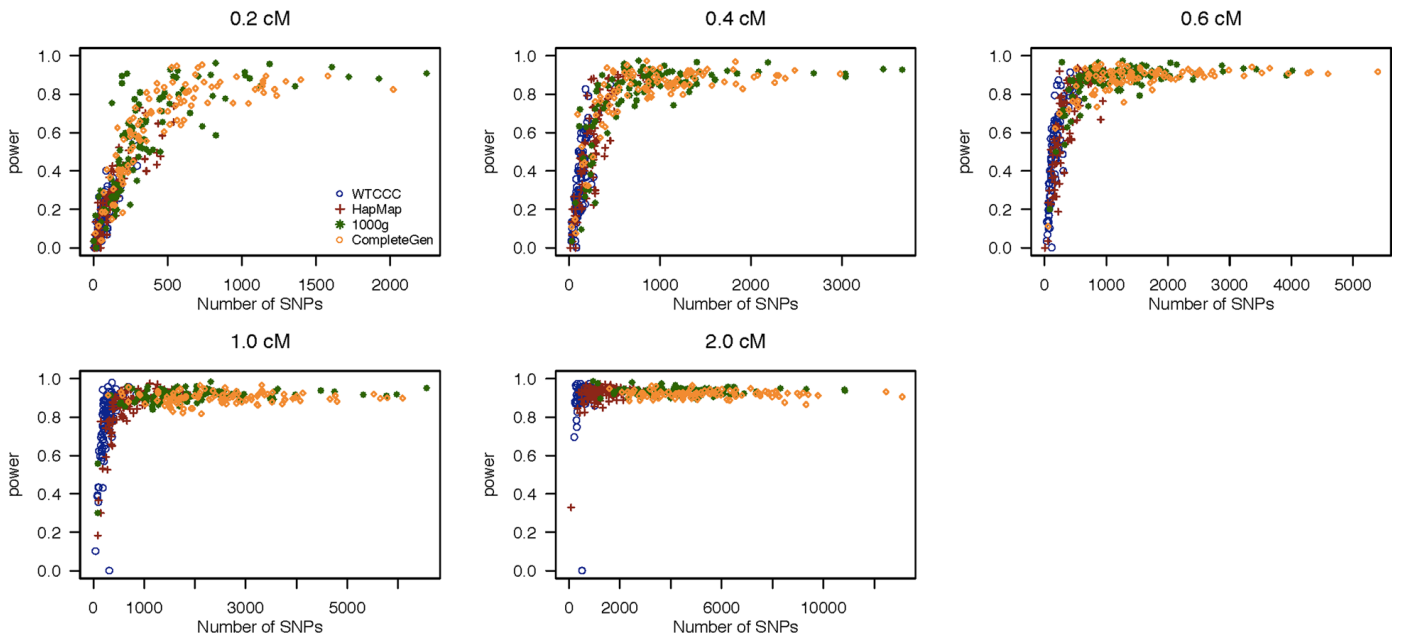
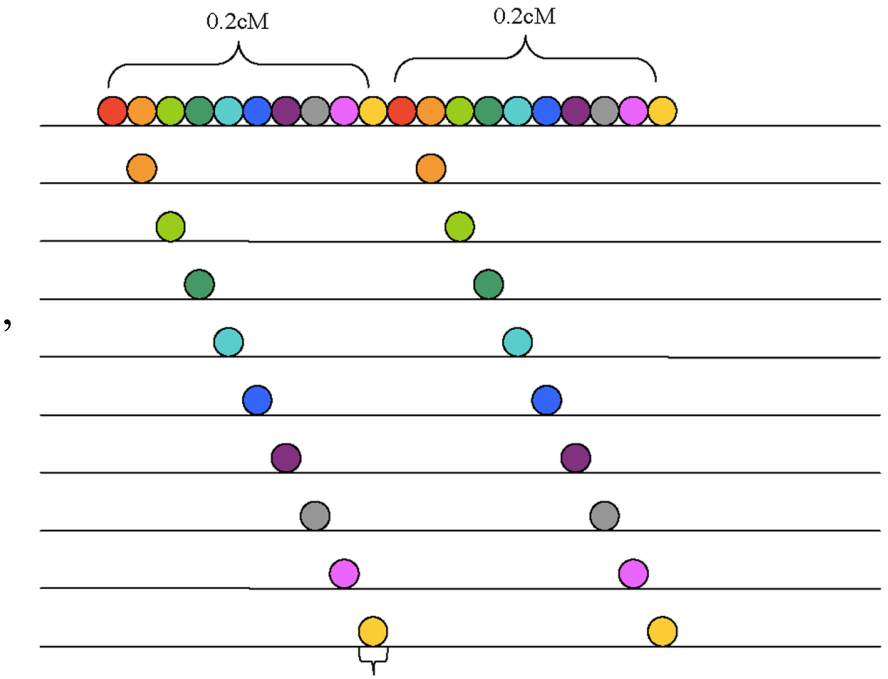


Figure 1 Empirical power of fastIBD. Empirical power of fastIBD to detect an IBD segment as a function of the number of SNPs within a segment in the simulation study. Each plot presents different lengths of IBD segments examined. The power of each dataset is represented by different colored circles and plotted against the number of SNPs contained within a given region.

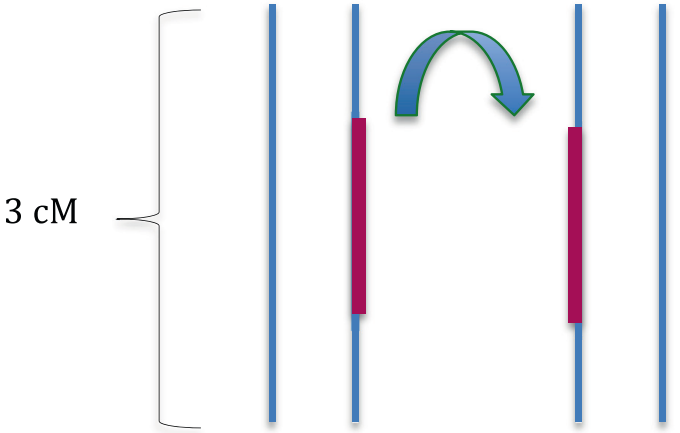
WTCCC : puce avec SNPs éparses ; Hapmap : puce dense ;
1000G : Illumina ; Complete Genomics

Su & al *BCM Bioinformatics* 13 :121 (2012)

Fins points de simulation



«Vrai» haplotypes artificiels après mixage (pas de rélations non-détectées)



IBD par copiage au hasard

Su & al *BCM Bioinformatics* 13 :121 (2012)

Fausses positives. . .

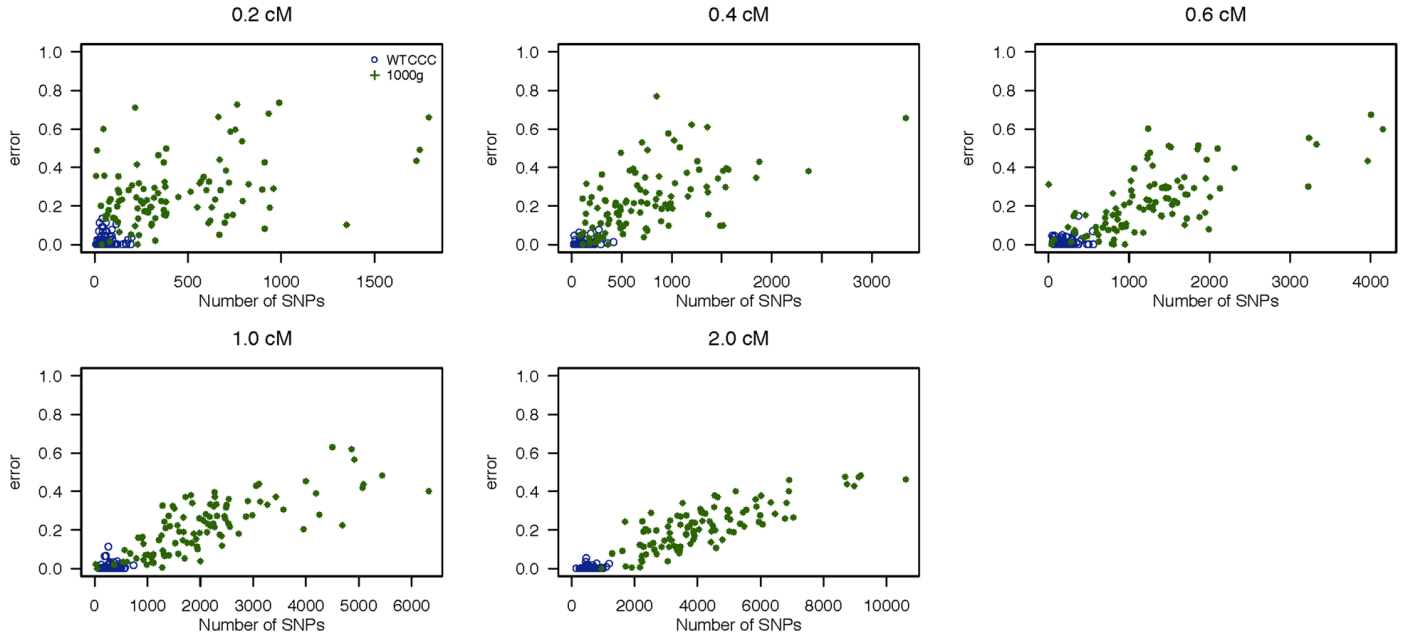
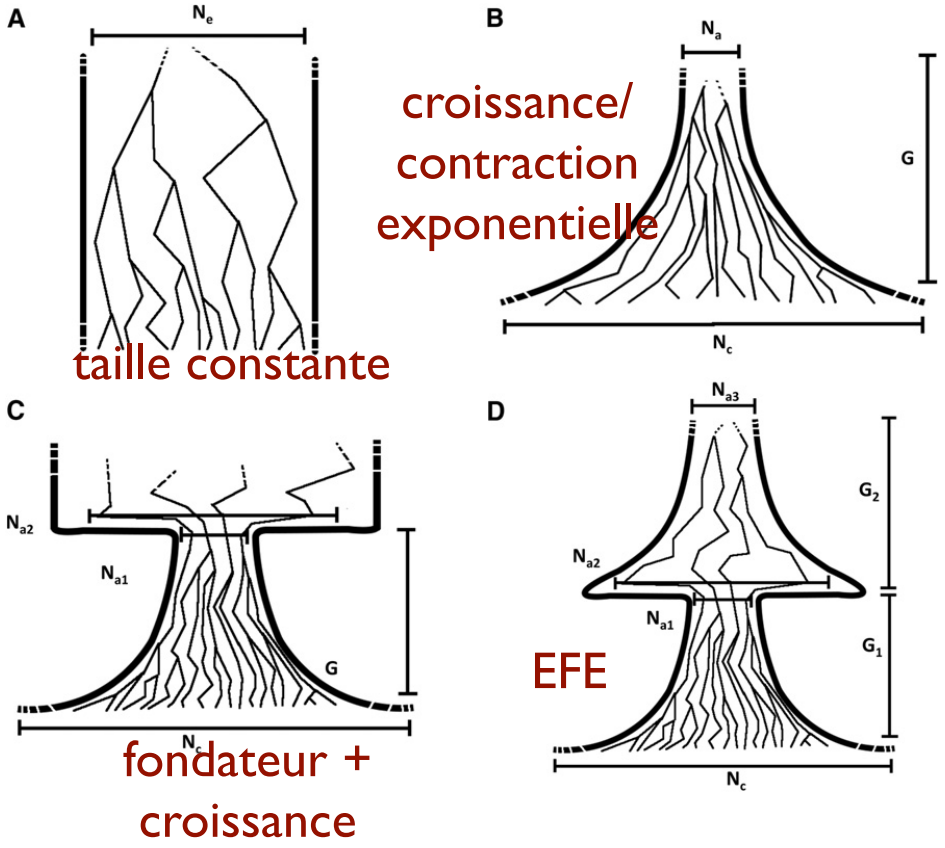


Figure 4 Empirical false positive rate of Germline. Empirical false positive rate of Germline to detect an IBD segment as a function of the number of SNPs within a segment in the simulation study. Each plot presents different lengths of IBD segments examined. The error rate of each dataset is represented by different colored circles and plotted against the number of SNPs contained within a given region.

Su & al *BCM Bioinformatics* 13 :121 (2012)

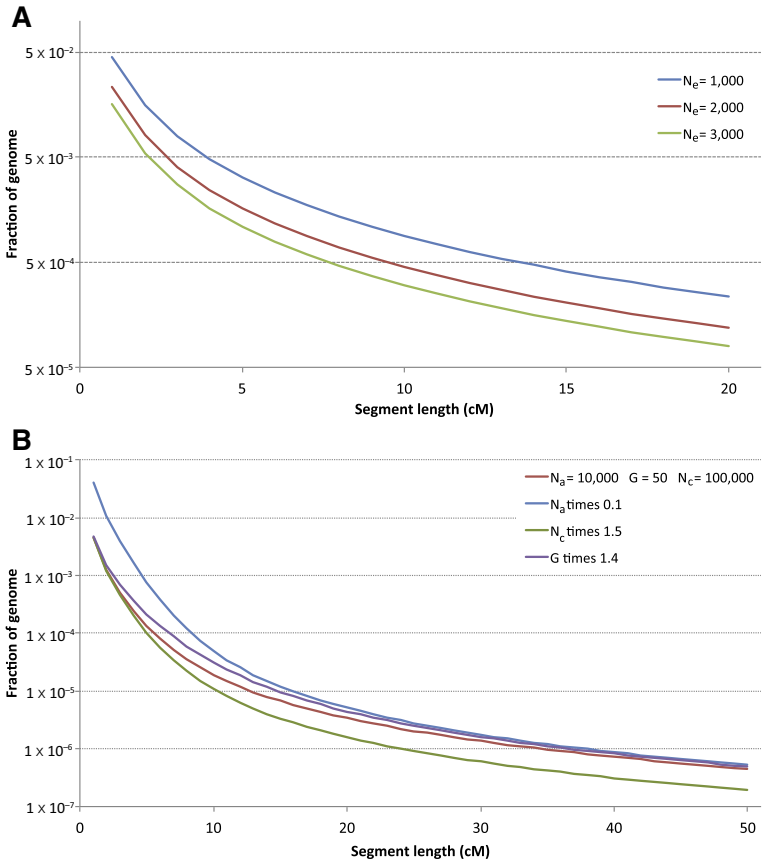
Distribution de segments IBD dans la population



la distribution de segments IBD dépend de l'histoire de la population

Palamara & al *American Journal of Human Genetics* 91 :809 (2012)

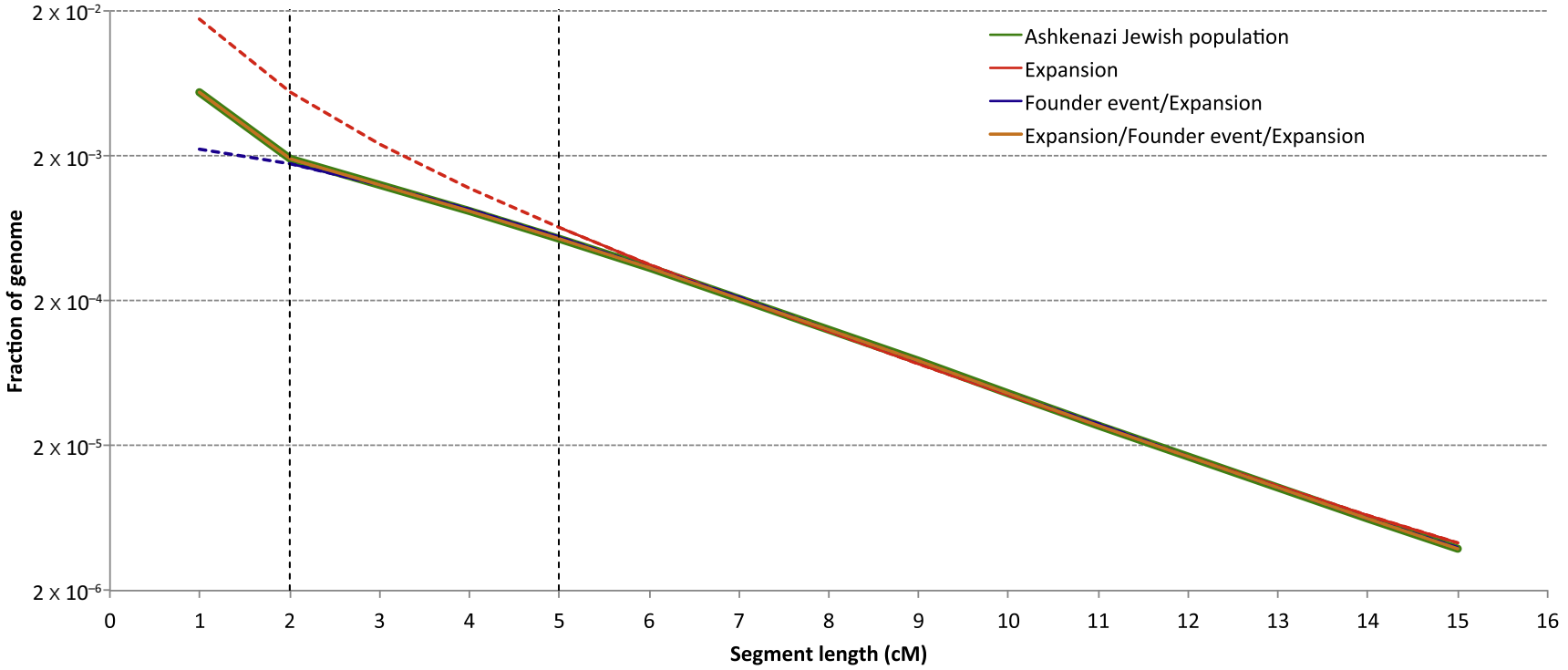
Segments IBD et paramètres



(taille constante, croissance exponentielle)

Palamara & al *American Journal of Human Genetics* 91 :809 (2012)

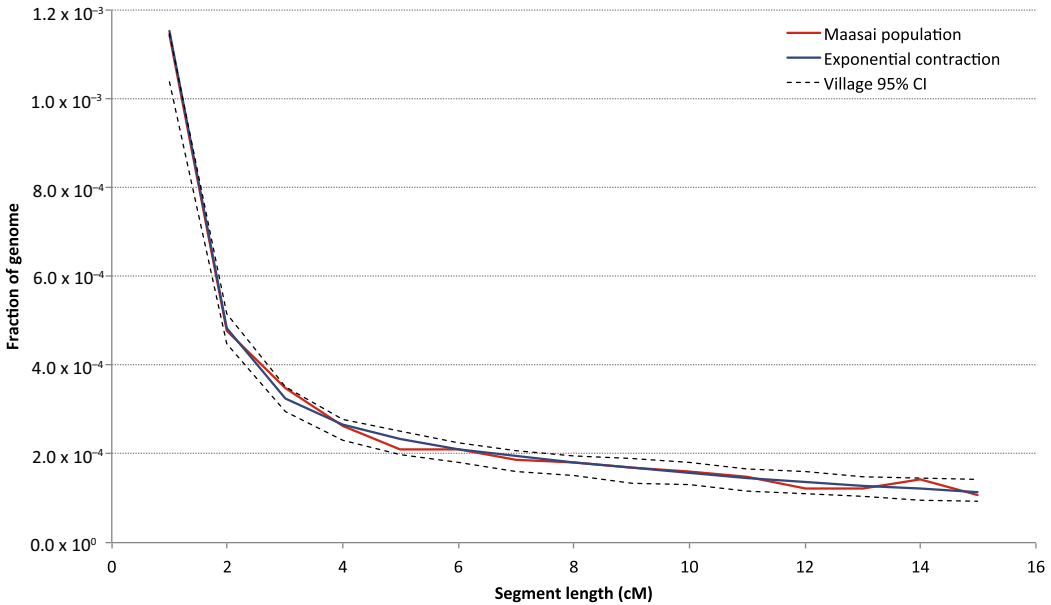
Démographie des Ashkenazim



meilleur fit pour données d'Ashkenazim ($n = 500$ génotypés, $\ell = 750\,\mathsf{k}$ sites) : EFE avec $N_c = 2300$ (-200 générations), ↗ 45 k (-34 générations) ↘ 270 ↗ 4.3 M

Palamara & al *American Journal of Human Genetics* 91 :809 (2012)

Démographie des Maasaï

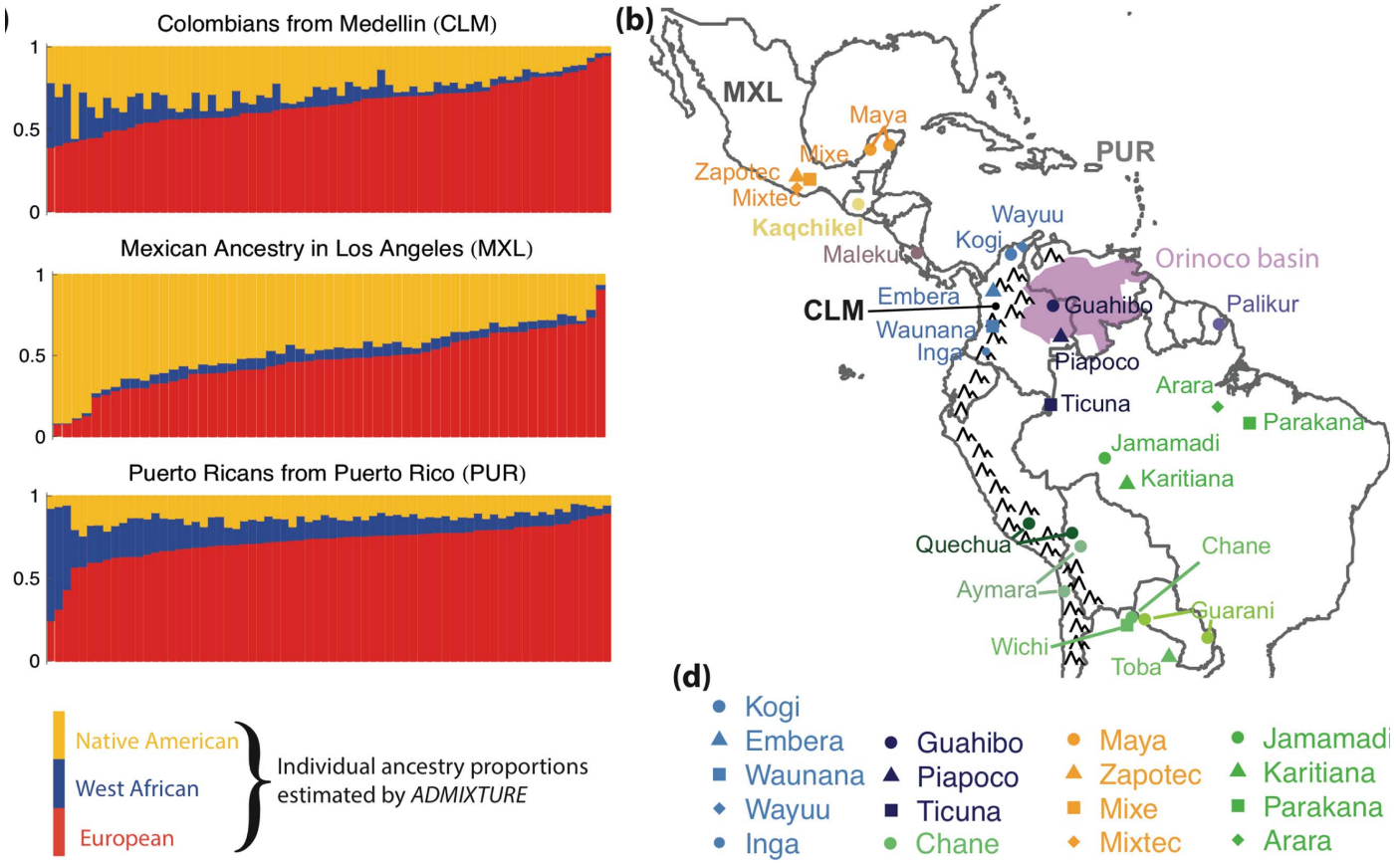


meilleur fit pour données de Maasaï ($n = 78$ génotypés, $\ell = 1.5\,\mathsf{M}$ sites) : EFE avec $N_c = 23500$ (-23 générations) , $\nearrow 500$

explication : villages avec migration de taux bas ($\Rightarrow$ ancêtres communs sont plus anciens)

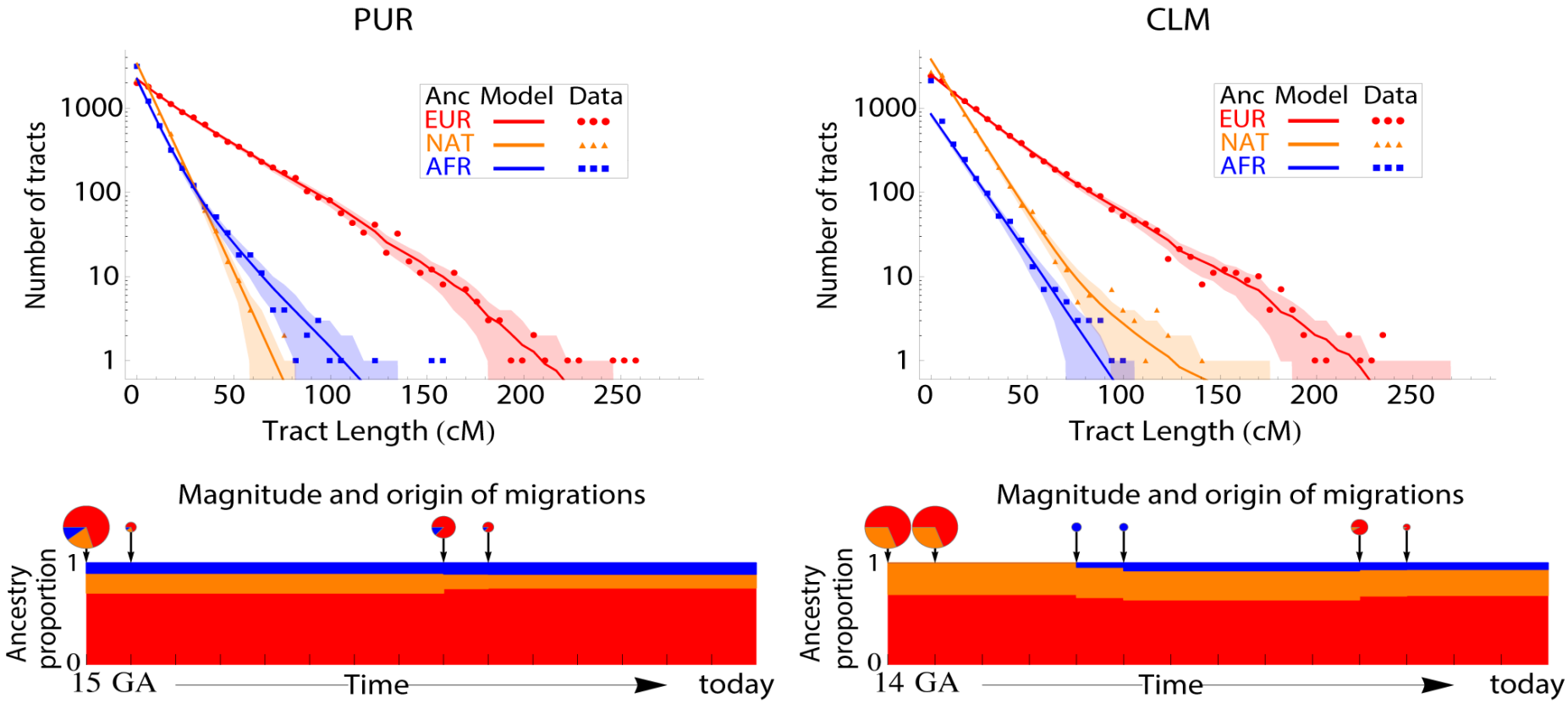
Palamara & al *American Journal of Human Genetics* 91 :809 (2012)

Population des Amériques



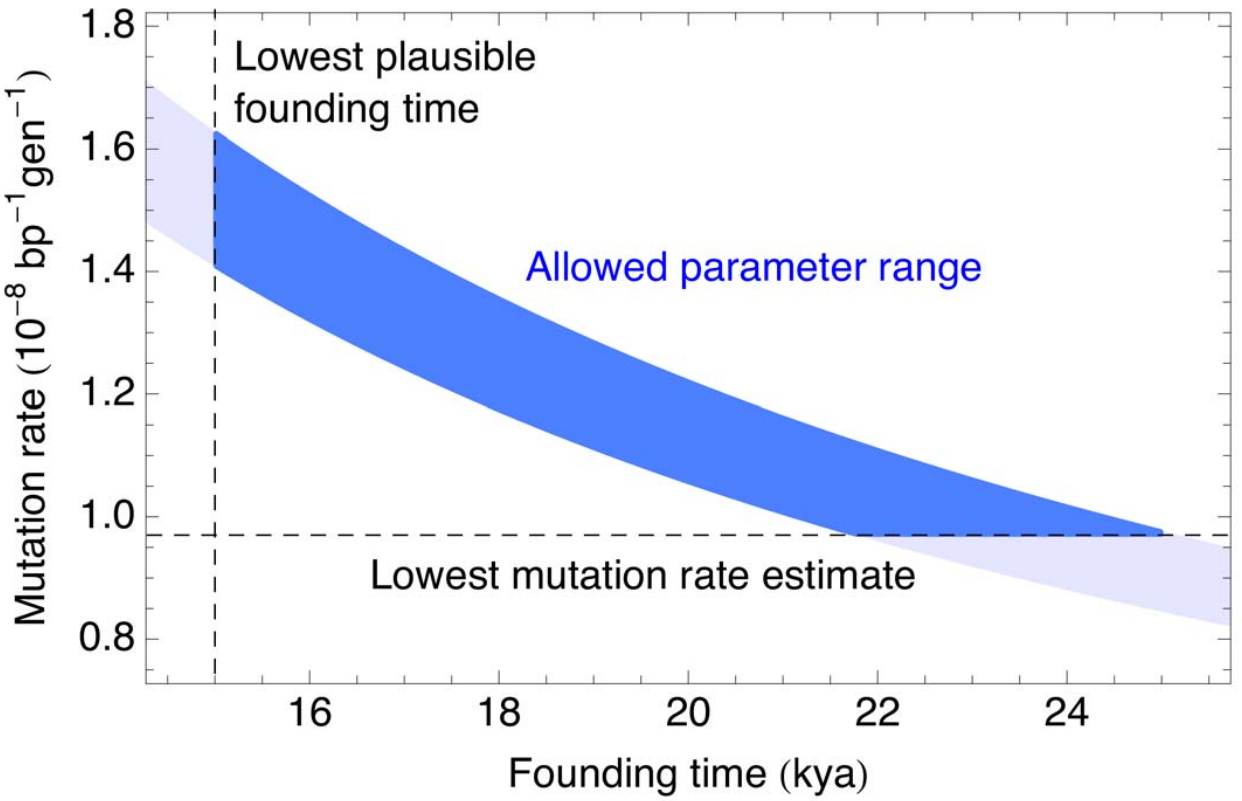
Gravel & al *PLoS Genetics* 9 :e1004023 (2013)

Fondateurs et diversion



Gravel & al *PLoS Genetics* 9 :e1004023 (2013)

Population des Amériques 2



arrivée : 16000 ans, MXL 12200 ans, PUR/CLM 11700 ans, migrations

Gravel & al *PLoS Genetics* 9 :e1004023 (2013)