Structure

Pritchard et al., Genetics 2000

Dakar, February 2011

Clustering in genetics

- We have genetic data from a sample of individuals, each of whom is assumed to have originated from a single unknown population (no admixture).
- We wish to cluster together individuals who are genetically similar.

Structure

- Structure relies on a model-based clustering method for using multilocus genotype data to infer population structure and assign individuals to populations.
- We assume a model in which there are K populations

K=1!

	Locus 1 (f_1)	Locus 2 (f_2)	Locus 3 (f_3)
Indiv 1	0	0	1
Indiv 2	0	1	0
Indiv 3	0	0	0
Indiv 4	0	1	1
Indiv 5	1	1	1
	n_1=1	n_2=3	n_3=3

n=5,L=3

Beta-Binomial model

f_j frequency of 1 for marker j, j=1..L
n_j number of 1 for marker j

Prior

$$f_j \sim \beta(1,1)$$

Modèle

$$n_j \sim \text{Binomial}(n,f_j) \text{ i.i.d.}$$

Posterior

$$f_j \sim \beta(1+n_j,1+n-n_j)$$

K=1

	Locus 1 (f_1)	Locus 2 (f_2)	Locus 3 (f_3)
Indiv 1	0	0	1
Indiv 2	0	1	0
Indiv 3	0	0	0
Indiv 4	0	1	1
Indiv 5	1	1	1
	n_1=1	n_2=3	n_3=3
	$f_1 \sim \beta(1+1,1+5-1)$	$f_2 \sim \beta(1+3,1+5-3)$	$f_3 \sim \beta(1+3,1+5-3)$

n=5,L=3

K=2

		Locus 1 $(f_{10,}f_{11})$	Locus 2 $(f_{20,}f_{21})$	Locus 3 $(f_{30,}f_{31,})$
$Z_1=0$	Indiv 1	0	0	1
$Z_2=1$	Indiv 2	0	1	0
$Z_3=0$	Indiv 3	0	0	0
$Z_4=1$	Indiv 4	0	1	1
$Z_5=1$	Indiv 5	1	1	1
		$n_1=1$	$n_2=3$	$n_3=3$

Param={(f_{10}, f_{20},f_{30}),(f_{11}, f_{21},f_{31}),$(Z_1,….,Z_5)$}

Gibbs sampling step for updating the frequencies (step I)

Given ($Z_1 \ldots Z_5$)
Beta-binomial model within each population

		Locus 1 (f_{10})	Locus 2 (f_{20})	Locus 3 (f_{30})
$Z_1=0$	Indiv 1	0	0	1
$Z_3=0$	Indiv 3	0	0	0
		$n_{10}=0$	$n_{20}=0$	$n_{30}=1$
		$\beta(1+0,1+2)$	$\beta(1+0,1+2)$	$\beta(1+1,1+1)$

Beta-Binomial model
Step I of the algorithm

f_{jk} frequency for marker j in pop k
n_{jk} number of 1 for marker j in pop k

Prior

$$f_{jk} \sim \beta(1,1), j=1..L, k=0..(K-1)$$

Modèle

$$n_{jk} \sim \text{Binomial}(n,f_{jk}) \text{ i.i.d.}$$

Given $(Z_1,\ldots Z_n)$

$$f_{jk} \sim \beta(1+n_{jk},1+n-n_{jk})$$

Gibbs sampling step for updating $(Z_1 \ldots Z_n)$ Step II

Given the allele frequencies

		Locus 1 $(f_{10,} f_{11,})$	Locus 2 $(f_{20,} f_{21})$	Locus 3 $(f_{30,} f_{31})$
$Z_1=1$	Indiv 1	0	0	1

$P(Z_1=1)= c\ (1-f_{11})(1-f_{21})f_{31}$

		Locus 1 $(f_{10,} f_{11})$	Locus 2 $(f_{20,} f_{21})$	Locus 3 $(f_{30,} f_{31})$
$Z_1=0$	Indiv 1	0	0	1

$P(Z_1=0)= c\ (1-f_{10})(1-f_{20})\ f_{30}$

Gibbs sampling algorithm for structure

- *Start with a given* $(Z_1 \ldots Z_n)$

    For (it in 1..num_{it}*)*

    - *Step I*

        Update the allele frequencies (β distributions)

    - *Step II*

        Update the Z_i*'s (use* $P(Z_i=1)$ *and* $P(Z_i=0)$ *)*

- We report the proportion of $(Z_i=1)$ and $(Z_i=0)$ in the chain

R algorithm (Step I)

```
pop1<-sample(0:1,size=n,replace=T)
for (i in 1:nb_it)
{
    ###Count the number of 1 in pop 1 and pop 0
    counts1<-apply(data[pop1==1,],FUN=sum,MARGIN=2)
    counts0<-apply(data[pop1==0,],FUN=sum,MARGIN=2)
    ###Update frequencies (Step I)
    freq1<-rbeta(p,1+counts1,1+sum(pop1==1)-counts1)
    freq0<-rbeta(p,1+counts0,1+sum(pop1==0)-counts0)
    ######..........
}
```

R algorithm (Step II)

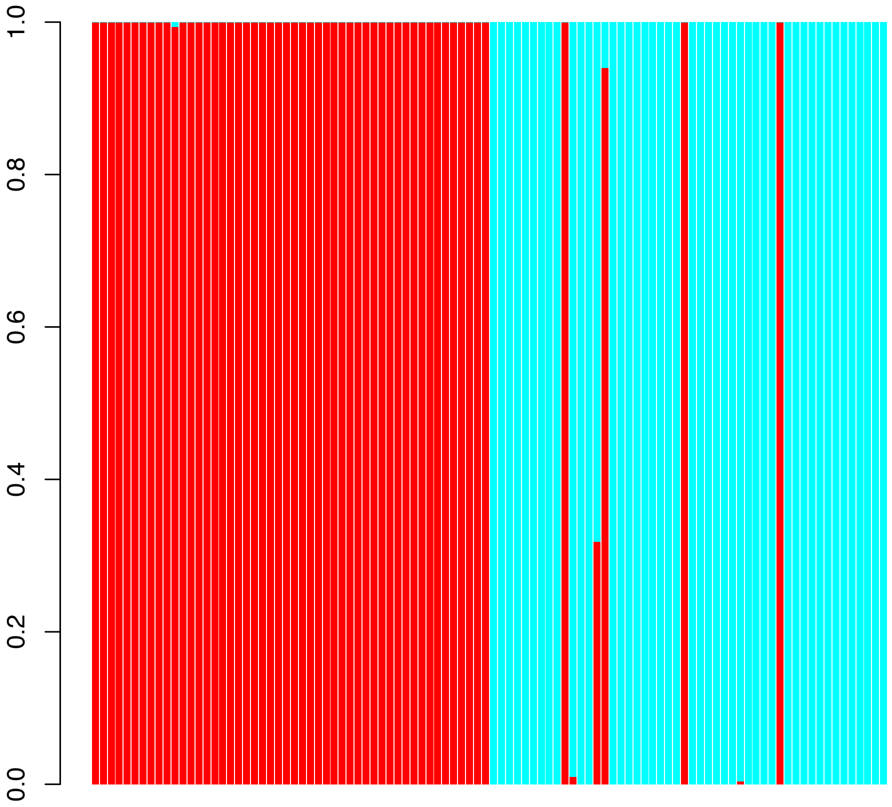
```
for (i in 1:nb_it)
{   ###....
    ###Compute proba to be in pop 1 and pop0
    prob1<-apply(data,FUN=function(x){prod(freq1^(x)*(1-freq1)^
    (1-x))},MARGIN=1)
    prob0<-apply(data,FUN=function(x){prod(freq0^(x)*(1-freq0)^
    (1-x))},MARGIN=1)
    pop1<-rbinom(n=n,size=1,prob=prob1/(prob1+prob0))
}
```

R algorithm (no admixture)

```
n<-dim(data)[1];p<-dim(data)[2];nb_it<-1000;burnin<-500;mem<-NULL;lik<-NULL
pop1<-sample(0:1,size=n,replace=T)
for (i in 1:nb_it)
{       if(i%%100==0)
        cat("Iteration number ",i,"\n")
        ###Count the number of 1 in pop 1 and pop 0
        counts1<-apply(data[pop1==1,],FUN=sum,MARGIN=2)
        counts0<-apply(data[pop1==0,],FUN=sum,MARGIN=2)
        ###Update frequencies (Step I)
        freq1<-rbeta(p,1+counts1,1+sum(pop1==1)-counts1)
        freq0<-rbeta(p,1+counts0,1+sum(pop1==0)-counts0)
        ###Compute proba to be in pop 1 and pop0 (Step II)
        prob1<-apply(data,FUN=function(x){prod(freq1^(x)*(1-freq1)^(1-x))},MARGIN=1)
        prob0<-apply(data,FUN=function(x){prod(freq0^(x)*(1-freq0)^(1-x))},MARGIN=1)
        pop1<-rbinom(n=n,size=1,prob=prob1/(prob1+prob0))
        mem<-rbind(mem,pop1);lik<-c(lik,sum(pop1*log(prob1)+(!pop1)*log(prob0)))}
prob_est<-apply(mem[-(1:burnin),],FUN=function(x){sum(x==1)/length(x)},MARGIN=2)
barplot(rbind(prob_est,1-prob_est),col=rainbow(2),border=NA)
```

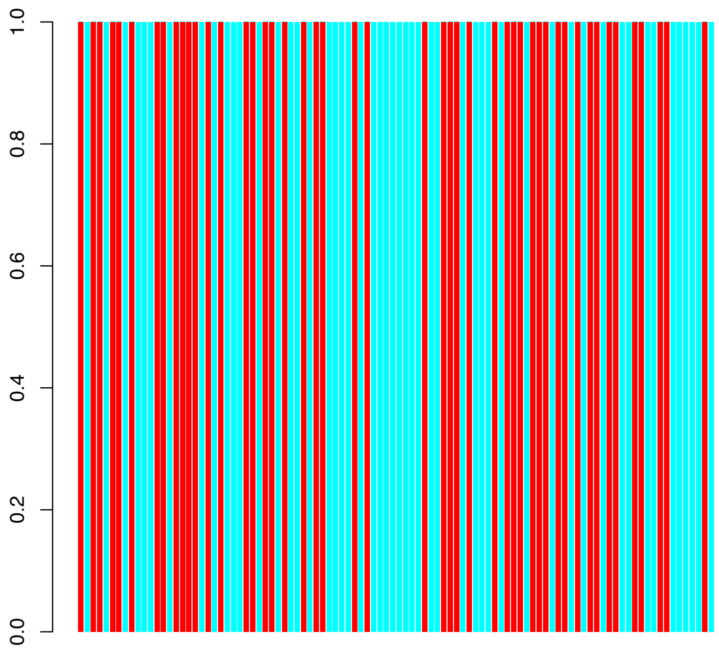
Example 1: true structure

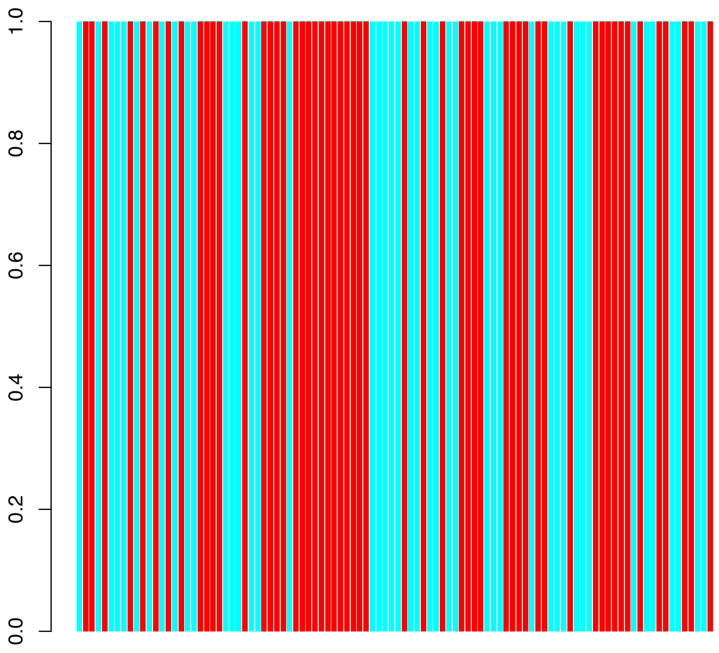
```
###Frequencies of 1: pop 1=60%, pop 0=20%
data<-rbind(matrix(rbinom(50*200,size=1,prob=.
    6),nrow=50,ncol=200),matrix(rbinom(50*200,size=1,prob=.
    2),nrow=50,ncol=200))
###........
barplot(rbind(prob_est,1-prob_est),col=rainbow(2),border=NA)
```



Example 2: no structure

```
data<-matrix(rbinom(100*250,size=1,prob=.5),nrow=100,ncol=250)
```

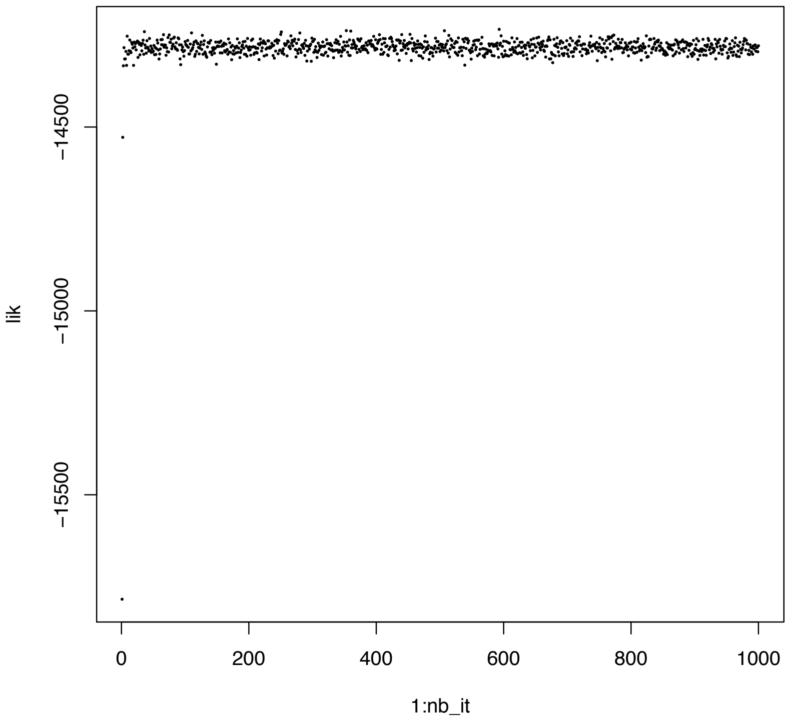




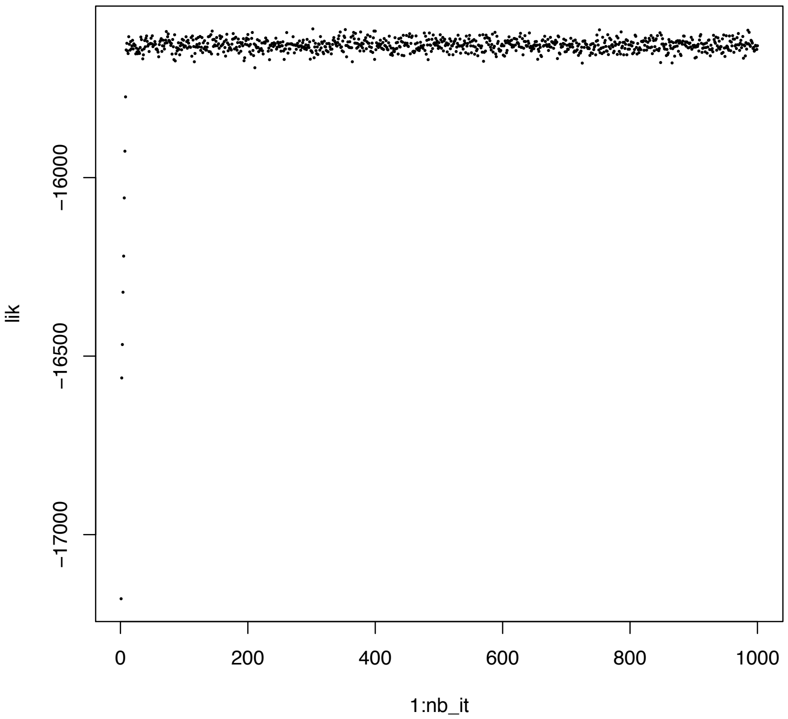
Likelihood

```
plot(1:nb_it,lik,pch=19,cex=.2)
```

Structure



No structure



Model with admixture

		Locus 1 $(f_{10,}f_{11})$	Locus 2 $(f_{20,}f_{21})$	Locus 3 $(f_{30,}f_{31})$
Z_1=(0,0,1)	Indiv 1	0	0	1
	Indiv 2	0	1	0
	Indiv 3	0	0	0
	Indiv 4	0	1	1
	Indiv 5	1	1	1

Individual coefficient of admixture

- Parameter q_i proportion of genes coming from pop 1 in individual i
- Prior

$$q_i \sim \beta(\alpha, \alpha)$$

- #1 dans $Z_i \sim \mathrm{Binom}(L, q_i)$

Model with admixture
Step III of the Gibbs sampling

- Given Z_i and the allele frequencies

$$q_i=\beta(\alpha+(\sharp 1 \text{ in } Z_i),\ \alpha+(\sharp 0 \text{ in } Z_i))$$

		Locus 1 $(f_{11,} f_{10})$	Locus 2 $(f_{21,} f_{20})$	Locus 3 $(f_{31,} f_{30})$
$Z_1=(0,0,1)$	Indiv 1	0	0	1

$$q_i=\beta(\alpha+1,\ \alpha+2)$$

Gibbs sampling algorithm for structure with admixture

- *Start with a given ($Z_1 \ldots Z_n$)*

  For (it in 1..num$_{it}$)

  - *Step I (slightly modified)*

    *Update the allele frequencies (*β distributions)

  - *Step II (slightly modified)*

    Update $Z_1 \ldots Z_n$

  - *Step III*

    Update the admixture proportion $q_1 \ldots q_n$*(*β distributions)

- We report the average of $q_1 \ldots q_n$ *along the chain*

R algorithm (admixture)

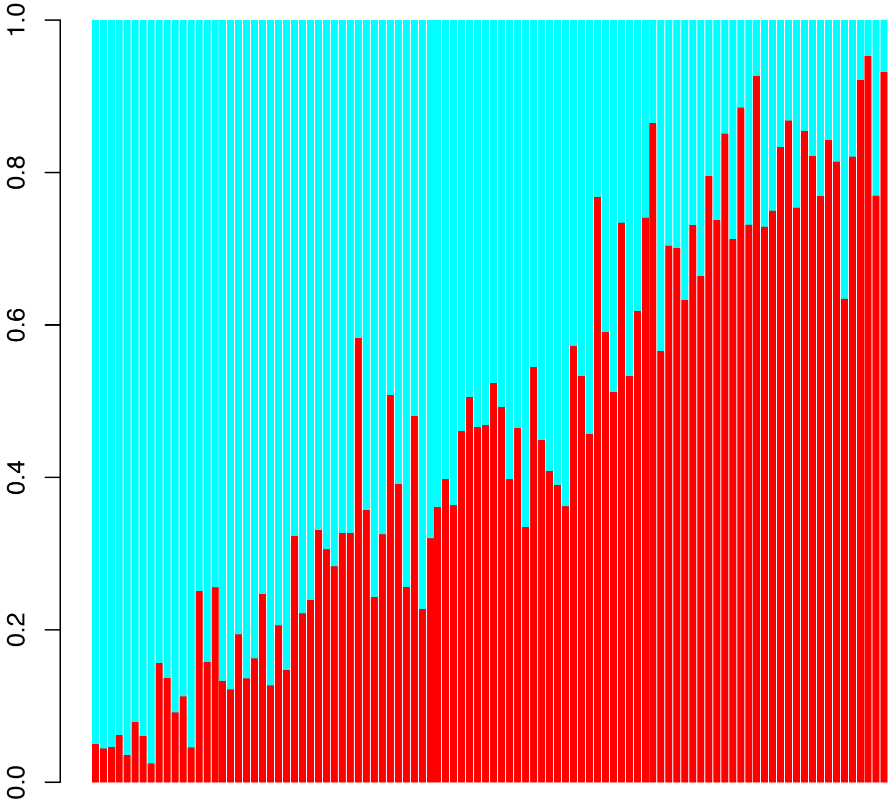
```
n<-dim(data)[1];p<-dim(data)[2];nb_it<-2000;burnin<-1000;mem<-NULL;alpha=1
pop1<-matrix(sample(0:1,size=n*p,replace=T),nrow=n,ncol=p);qadmix<-rep(.5,n);lik<-NULL
for (i in 1:nb_it)
{       if(i%%100==0)
        cat("Iteration number ",i,"\n")
        ###Count the number of 1 in pop 1 and pop 0
        counts1<-sapply(1:p,FUN=function(x){sum(data[pop1[,x]==1,x])})
        counts0<-sapply(1:p,FUN=function(x){sum(data[pop1[,x]==0,x])})
        n1<-apply(pop1,FUN=sum,MARGIN=2)
        ###Update frequencies (Step I)
        freq1<-rbeta(p,1+counts1,1+n1-counts1)
        freq0<-rbeta(p,1+counts0,1+(n-n1)-counts0)
        ###Compute proba to be in pop 1 and pop0 (Step II)
        qaux<-matrix(qadmix,nrow=n,ncol=p)
        prob1<-as.numeric(qaux*freq1^data*(1-freq1)^(1-data))
        prob0<-as.numeric((1-qaux)*freq0^data*(1-freq0)^(1-data))
        pop1<-matrix(rbinom(n=n*p,size=1,prob=prob1/(prob1+prob0)),nrow=n,ncol=p)
        ###Update q (Step III)
        n1<-apply(pop1,FUN=sum,MARGIN=1)
        qadmix<-rbeta(n,alpha+n1,alpha+p-n1)
        mem<-rbind(mem,qadmix);lik<-c(lik,sum(pop1*log(prob1)+(!pop1)*log(prob0)))}
prob_est<-apply(mem[-(1:burnin),],FUN=mean,MARGIN=2)
barplot(rbind(prob_est,1-prob_est),col=rainbow(2),border=NA)
```

Example 1: cline

```
cline<-function(i,n)
{
v1<-rbinom(n,size=1,prob=.6);v2<-rbinom(n,size=1,prob=.2)
choice<-rbinom(n,size=1,prob=i)
return(ifelse(choice, v1, v2))
}
data<-NULL
for(i in 1:100)
data<-rbind(data,cline(i/100,250))
```

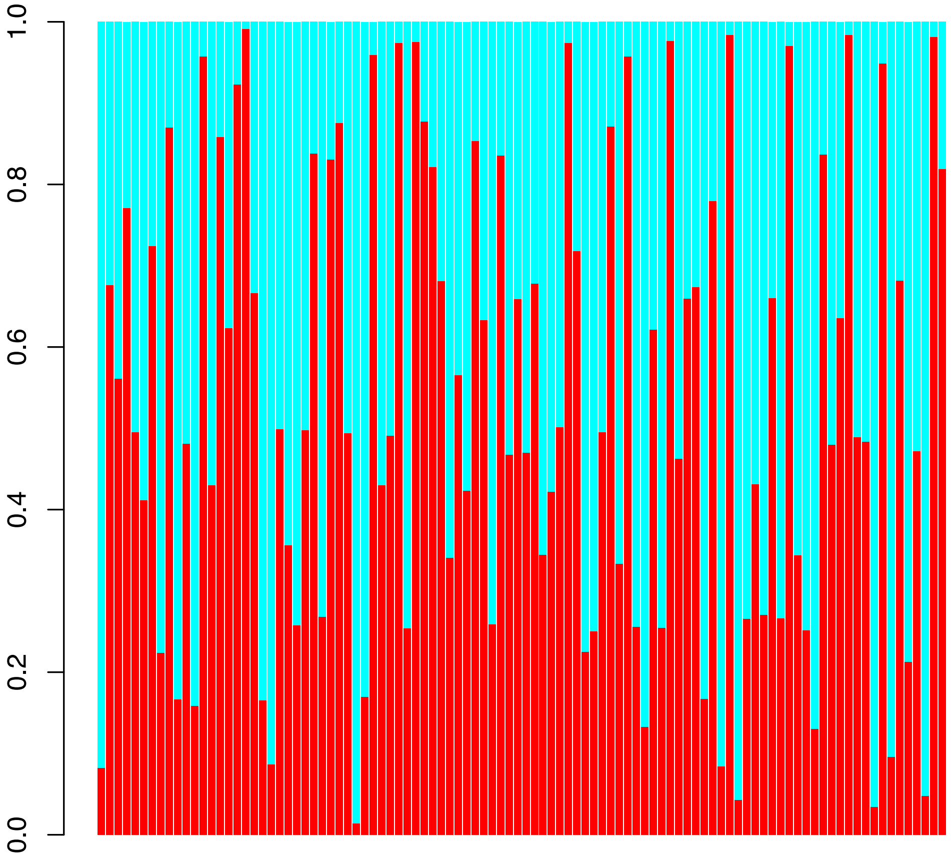
Example 1: cline

```
cline<-function(i,n)
{
v1<-rbinom(n,size=1,prob=.6);v2<-rbinom(n,size=1,prob=.2)
choice<-rbinom(n,size=1,prob=i)
return(ifelse(choice, v1, v2))
}
data<-NULL
for(i in 1:100)
data<-rbind(data,cline(i/100,250))
```



Example 2: no structure

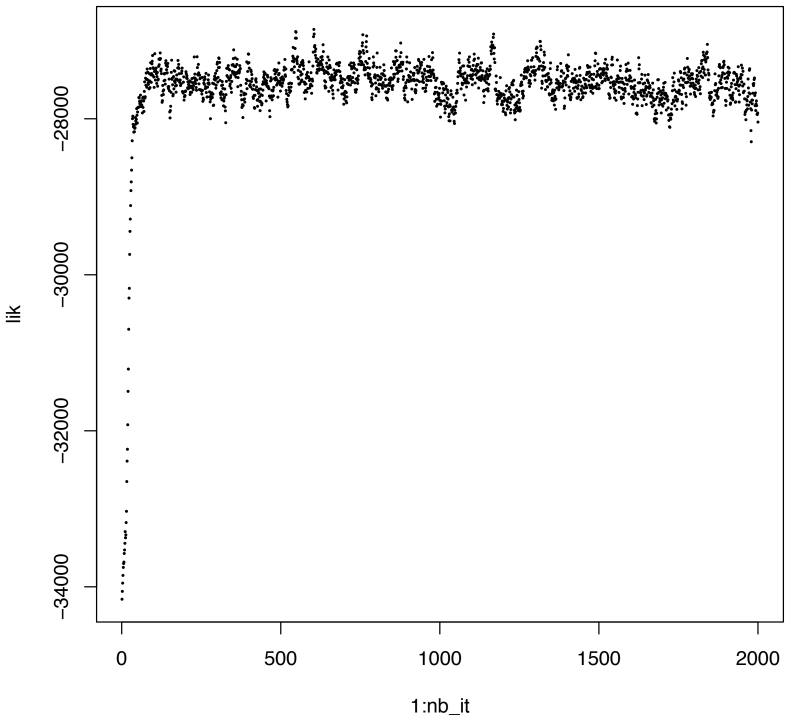
```
data<-matrix(rbinom(1000,size=1,prob=.5),nrow=100,ncol=250)
```



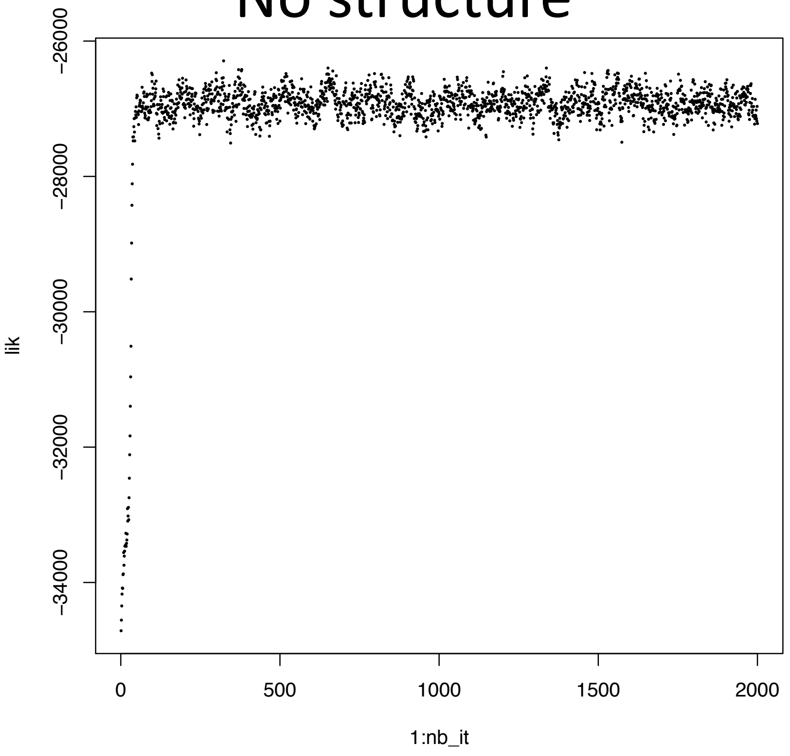
Likelihood

```
plot(1:nb_it,lik,pch=19,cex=.2)
```

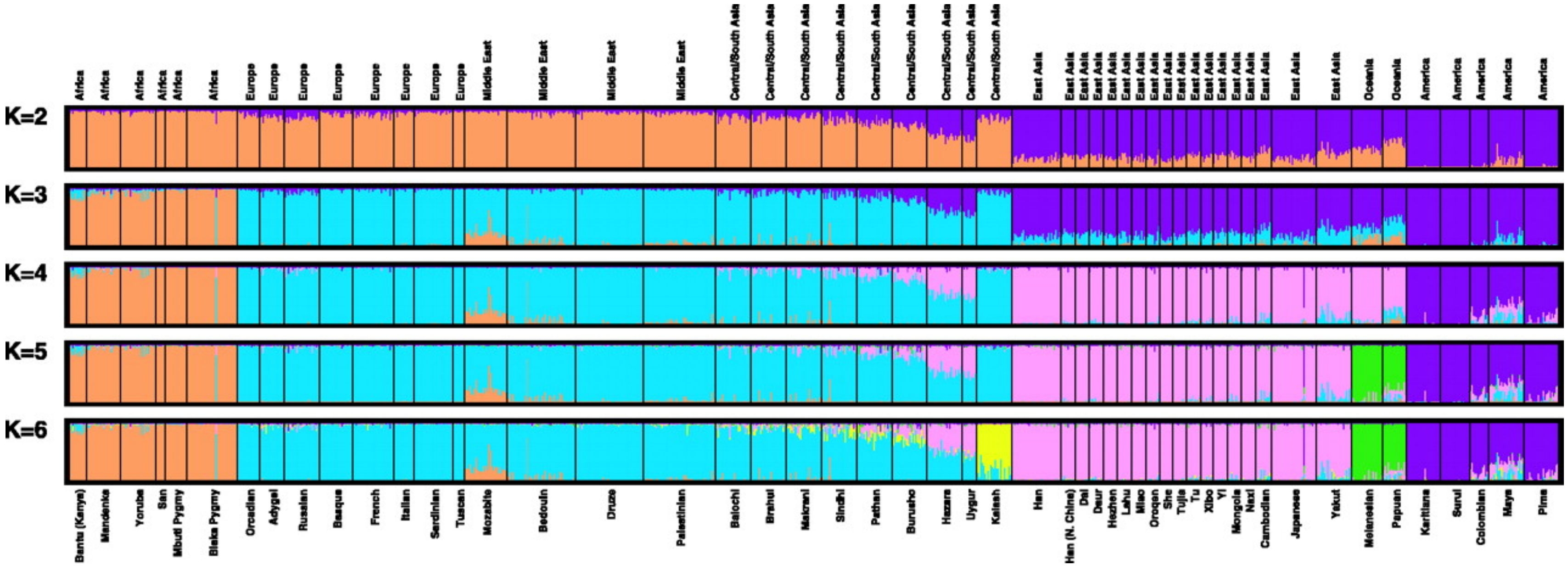
Cline



No structure



Worldwide human population structure

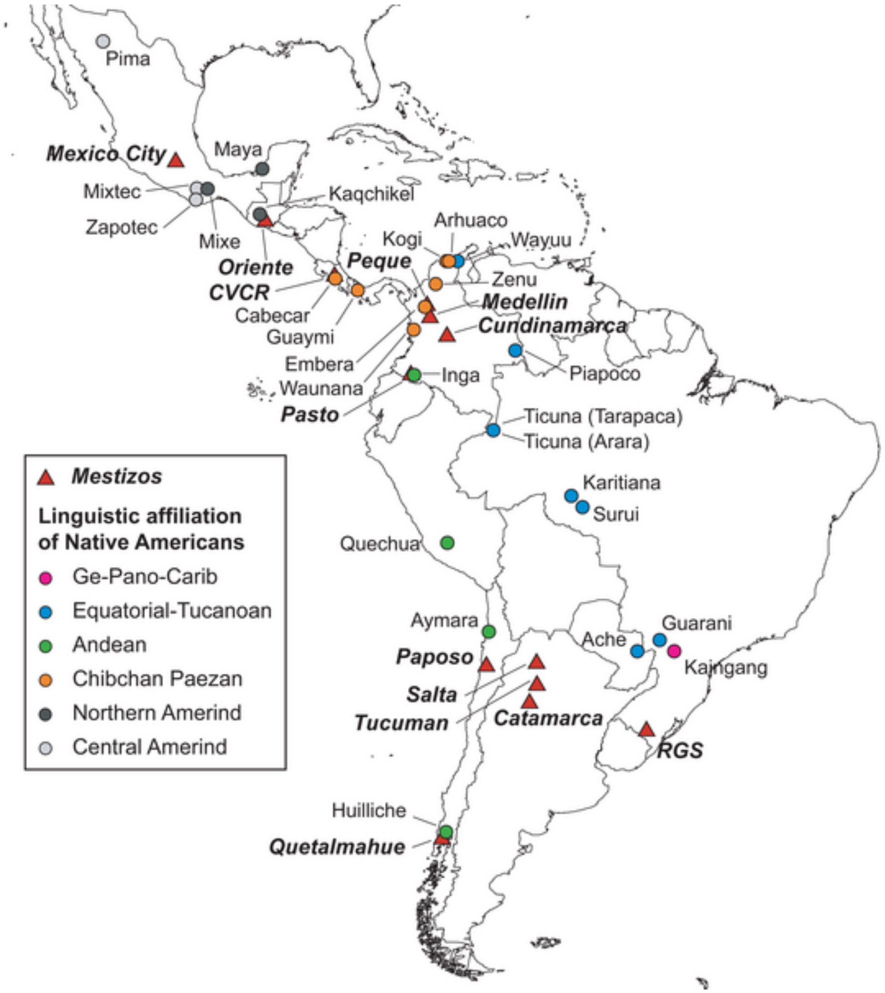


N A Rosenberg et al. Science 2002;298:2381-2385

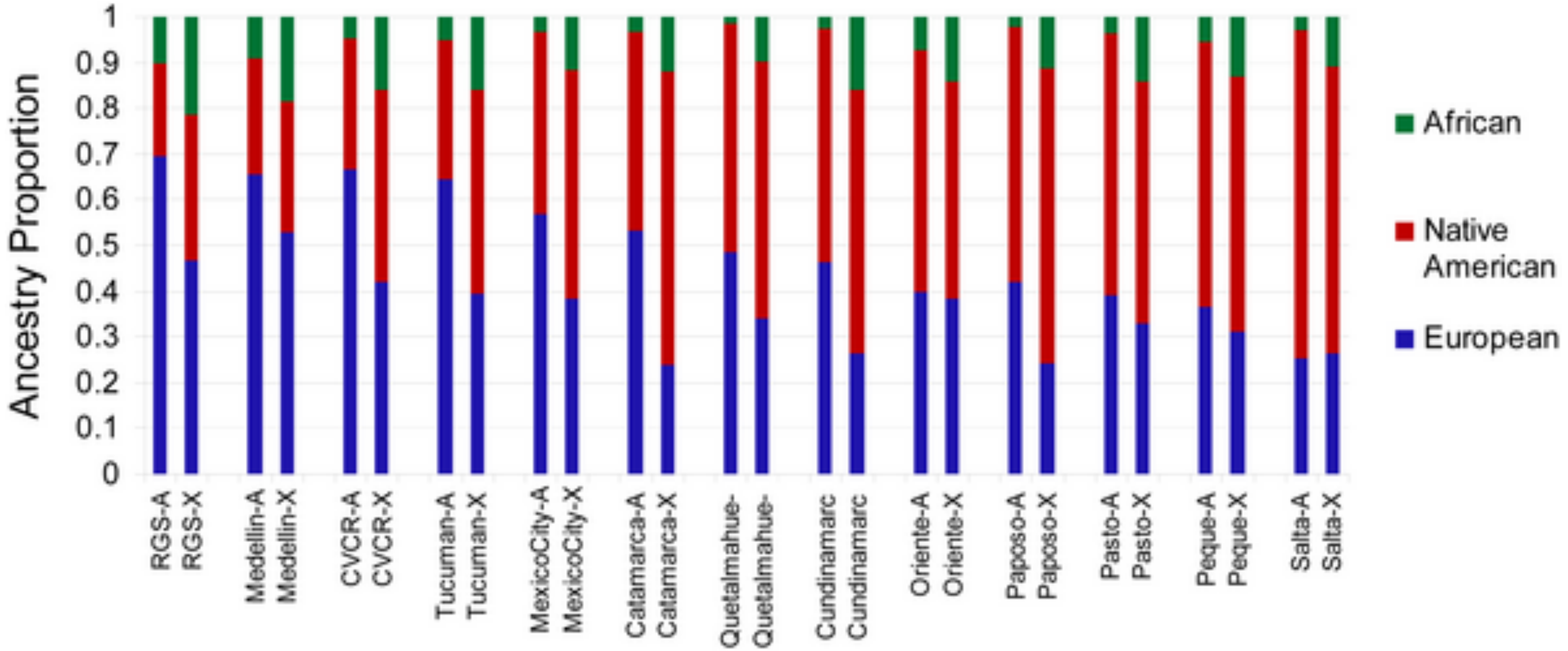
Science
AAAS

Published by AAAS

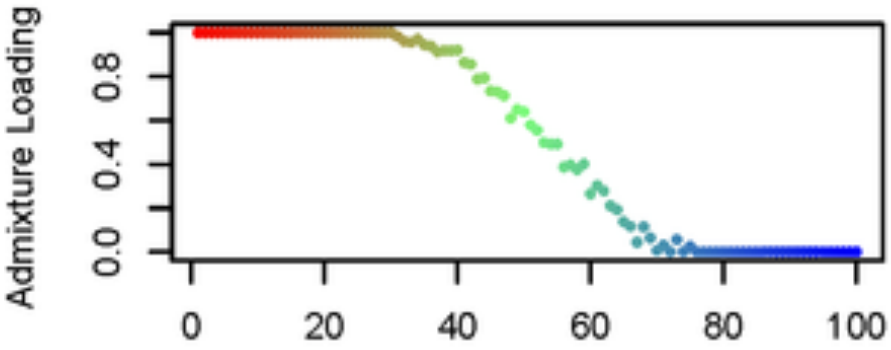
Admixture in Latin American population

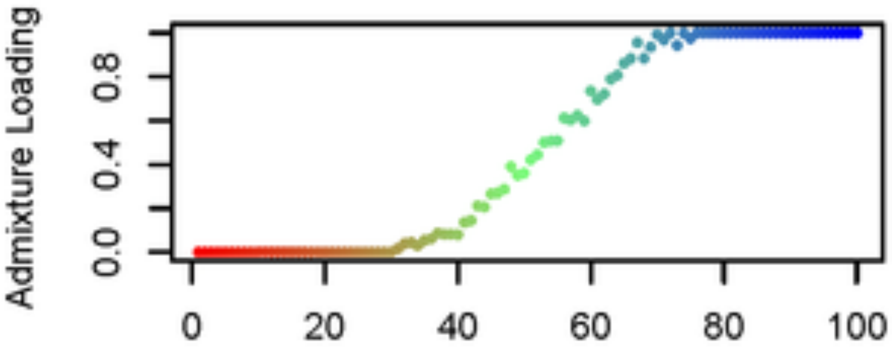


Admixture in Latin American population

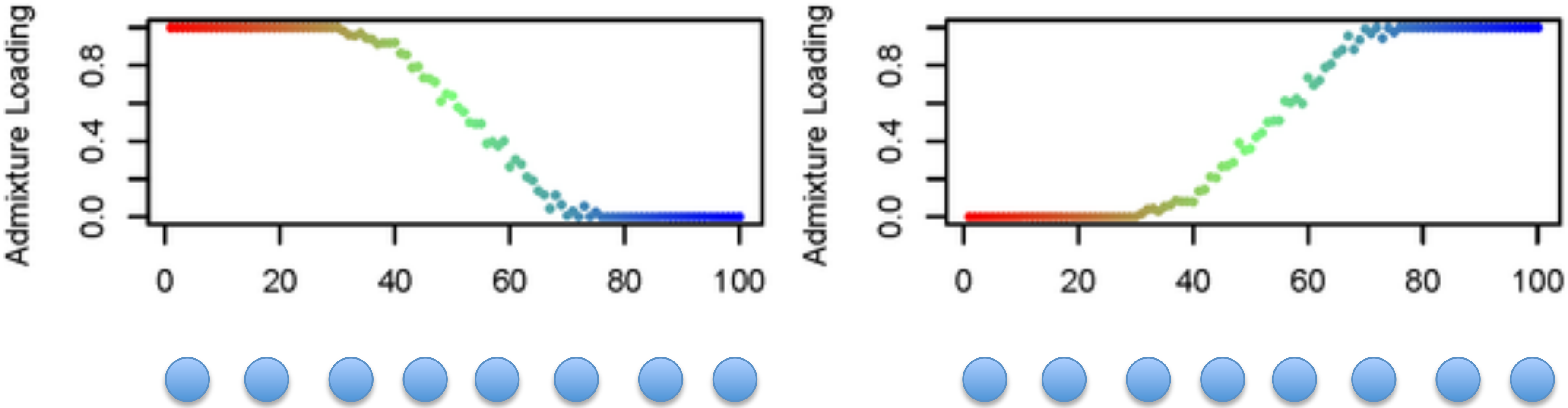


Word of caution





Word of caution


Admixture Loading
Admixture Loading