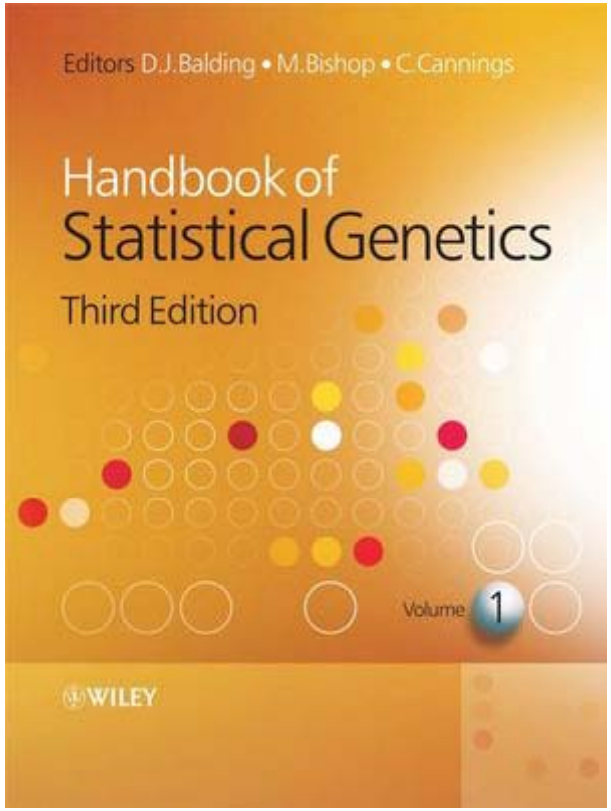
Statistical genetic approaches for estimating population structure with applications to fisheries populations

Toshihide Kitakado[1], Shuichi Kitada[1] and Hirohisa Kishino[2]

[1]Tokyo University of Marine Science and Technology

[2]University of Tokyo

Handbook of Statistical Genetics, 3rd ed. (Balding et al. eds 2007)



FOCUS ON STATISTICAL ANALYSIS
NATURE REVIEWS | GENETICS (2006) REVIEWS

Computer programs for population genetics data analysis: a survival guide

Laurent Excoffier and Gerald Heckel

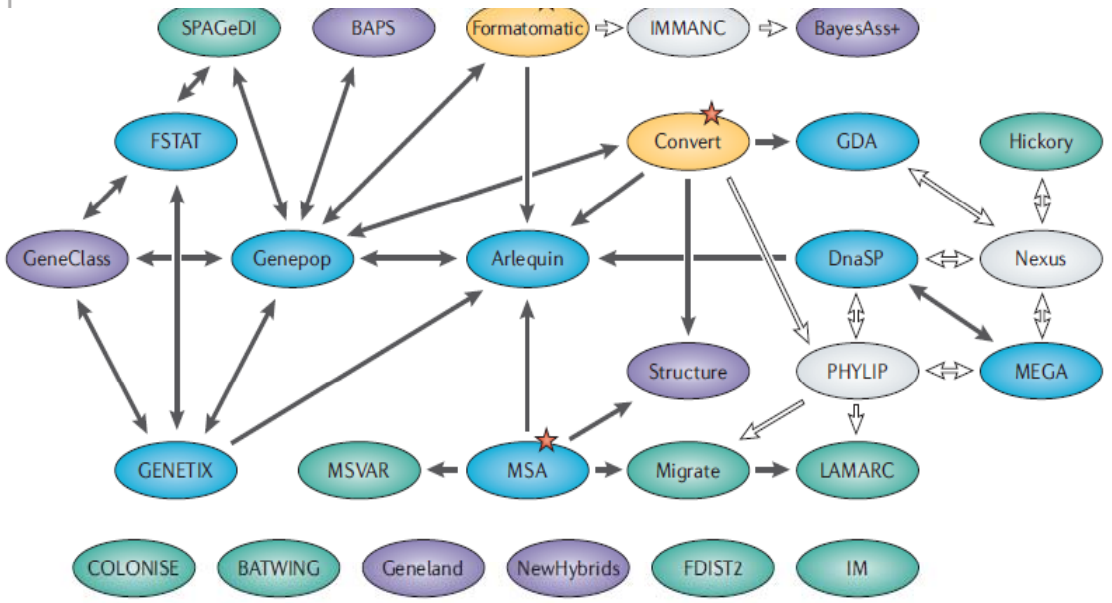
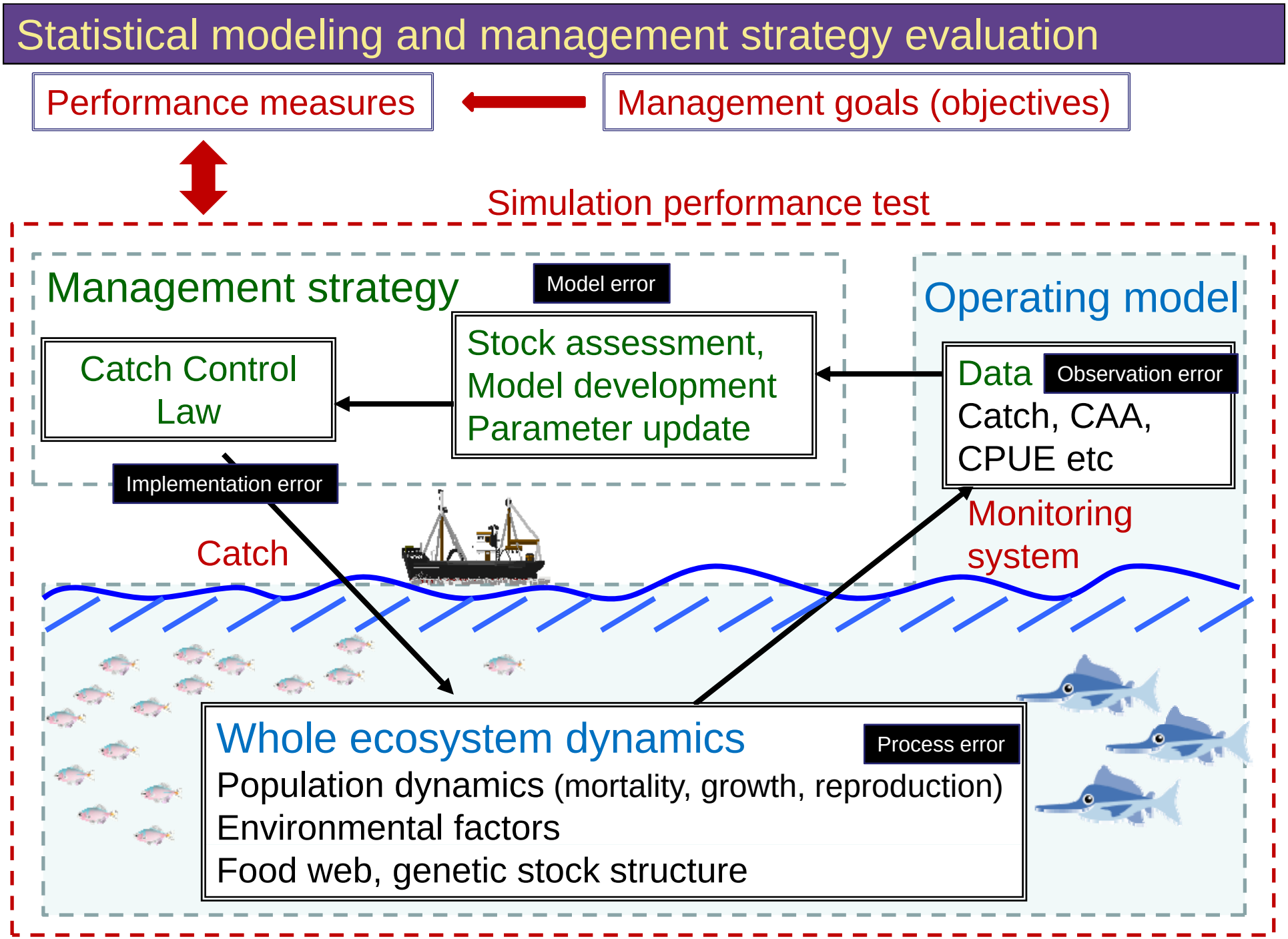


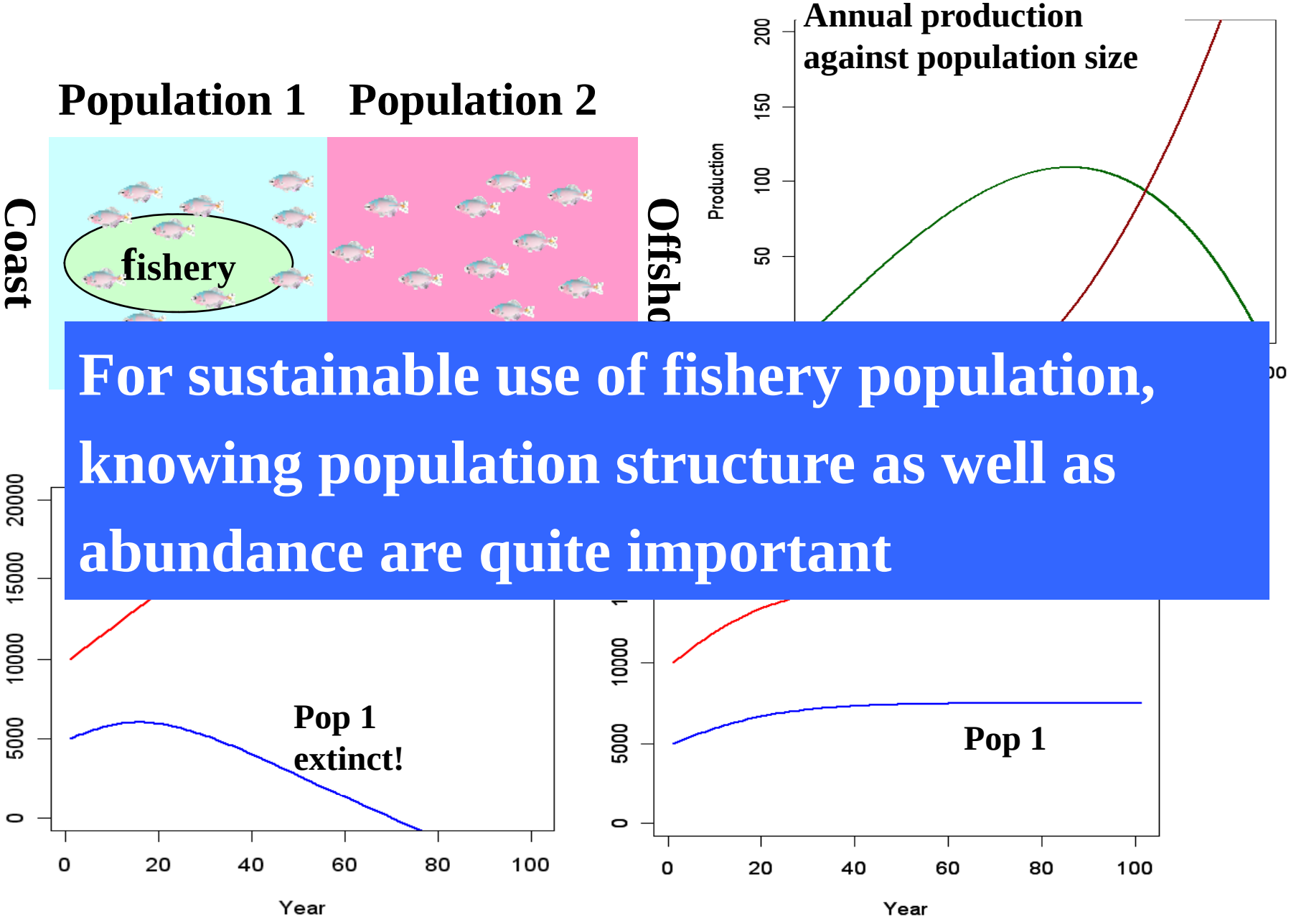
Figure 1 | Flow chart of possible data exchange between different population genetics programs.

1.Why must we consider population structure for fishery resource management ?

Statistical modeling and management strategy evaluation



Fishery management and population structure

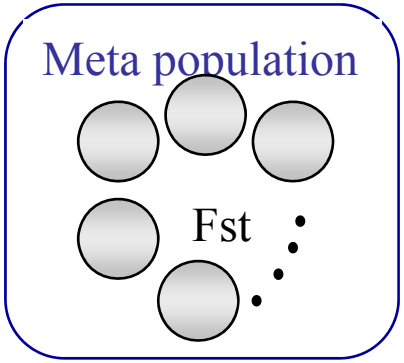


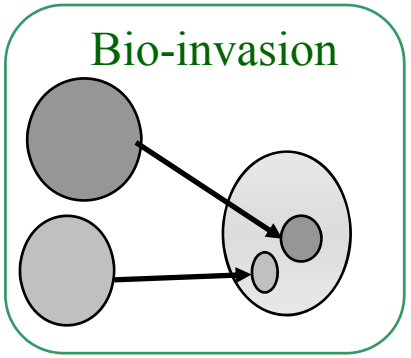
Several types of population structures

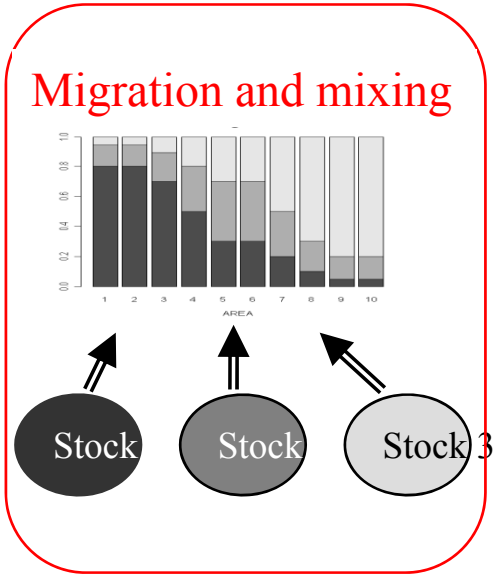
Spatial Population structure

Affect ↓ ↑ Infer

Genetic composition in population





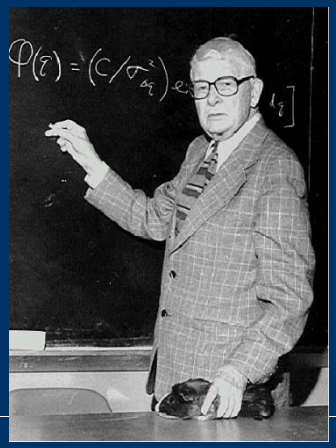


For statistical modeling of population structure

- Hierarchical structure
- Latent variables

Here, examples with population differentiation and mixing are introduced

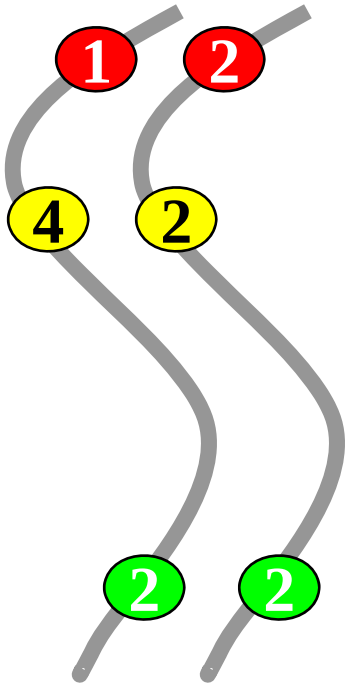
2. Estimation of population differentiation



- Population differentiation
- Likelihood for estimating Fst under a metapopulation
- Empirical Bayes estimation of pairwise Fst

Allele frequencies

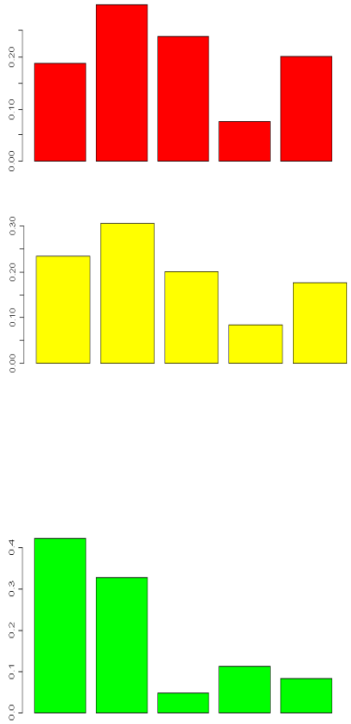
Individual genotype



Locus 1

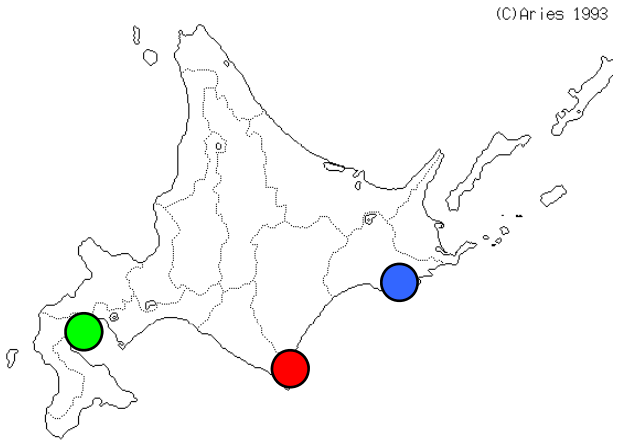
Locus 2

Locus *L*

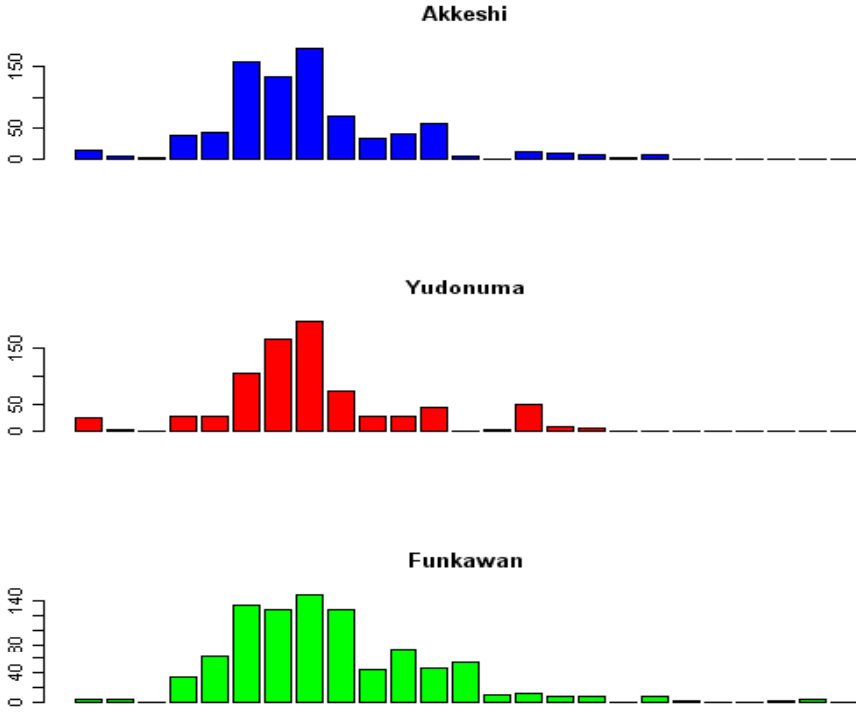


These allele frequencies differ if populations differ

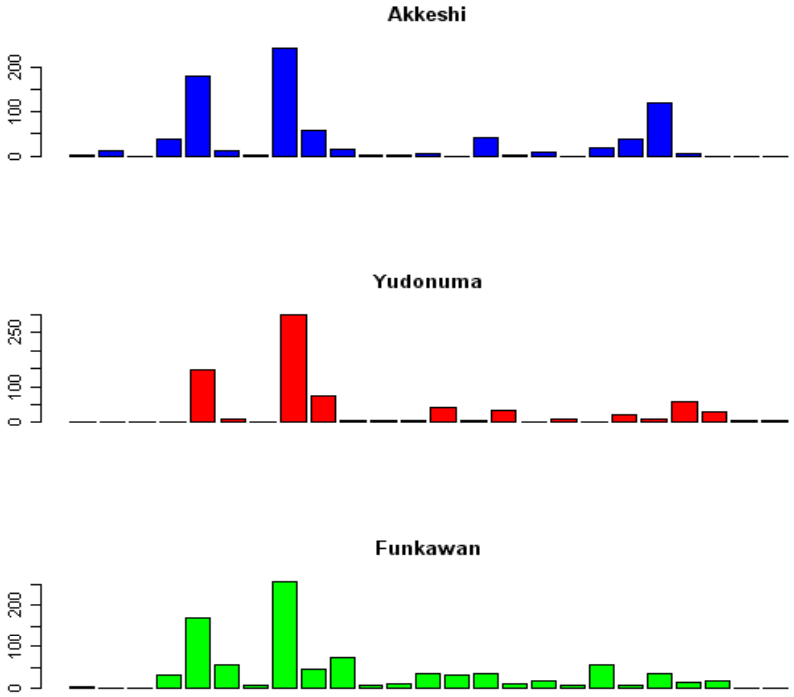
Pacific herring



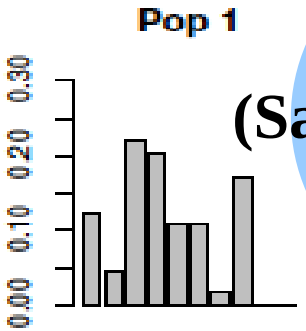
Locus 1



Locus 2

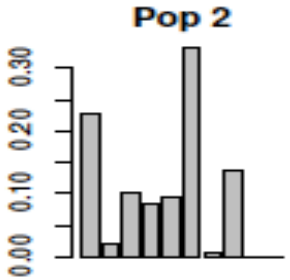


Statistical tests



Locality 1 (Sampling Area 1)

Locality 2 (Sampling Area 2)



$$p_1 = (p_{11}, \ldots, p_{1J})$$

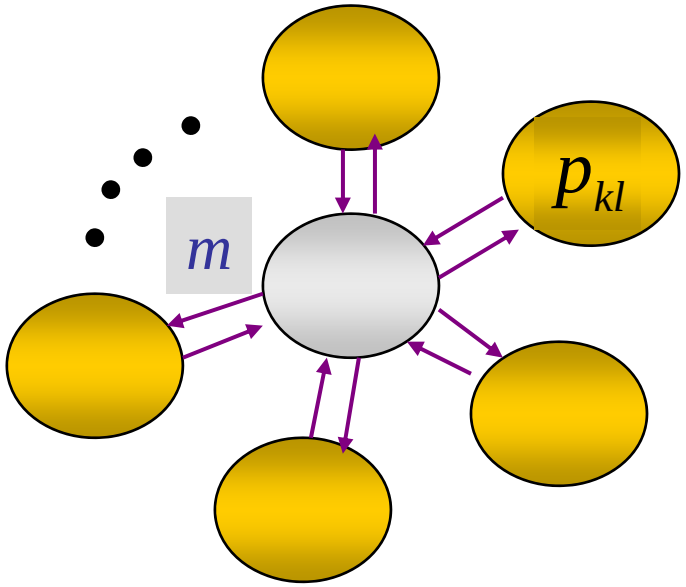
$$p_2 = (p_{21}, \ldots, p_{2J})$$

Testing $H_0 : p_1 = p_2 \quad vs \quad H_1 : p_1 \neq p_2$

- **Hard to say that the two are same!**
- **Need information on gene flow**

Metapopulation model

Migration-drift balance in metapopulation



$$E[\Delta p_t] = -m\ p_t + m\beta$$

$$Var[\Delta p_t] = \frac{p_t(1-p_t)}{2N_e}$$

Equilibrium distribution of allele frequencies

$$p_{kl} = (p_{kl1},\ldots,\ p_{klJ}) \sim D(\theta\beta_{l1},\ldots,\ \theta\beta_{lJ})$$

$$Var(p_{klj}) = \frac{1}{1+\theta}\beta_{lj}(1-\beta_{lj})$$

$$F_{ST} = \frac{1}{1+\theta} = \frac{1}{1+4N_e m}$$

Wright (1969)
Rannala and Hartigan (1995)

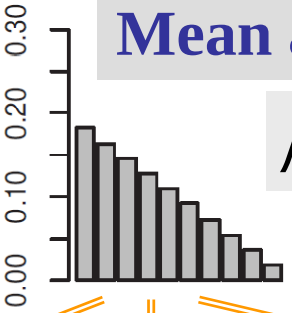
$$p_k \sim Dirichlet\ (\theta\beta_1, ..., \theta\beta_J)$$

$$Var\,(p_{kj}) = \frac{1}{1+\theta}\beta_j(1-\beta_j)$$

$$F_{ST} = \frac{1}{1+\theta} = \frac{1}{1+4N_e m}$$

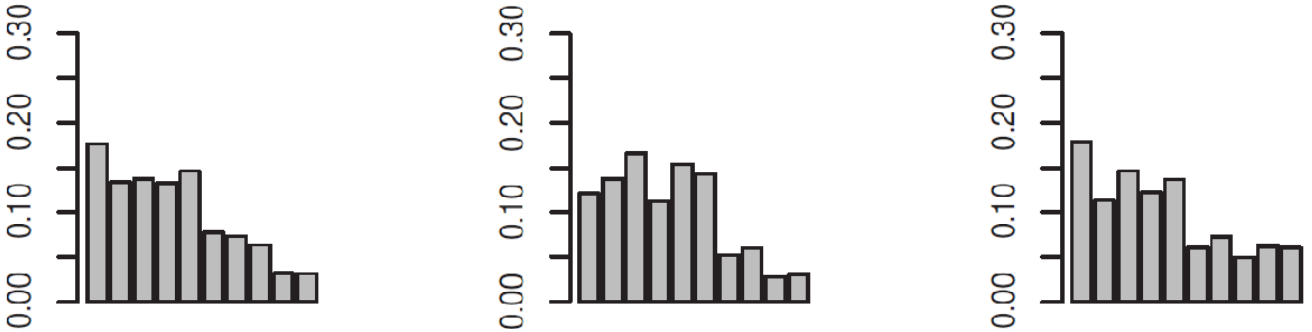
Mean allele frequencies at a locus

$$\beta = (\beta_1, ..., \beta_J)$$

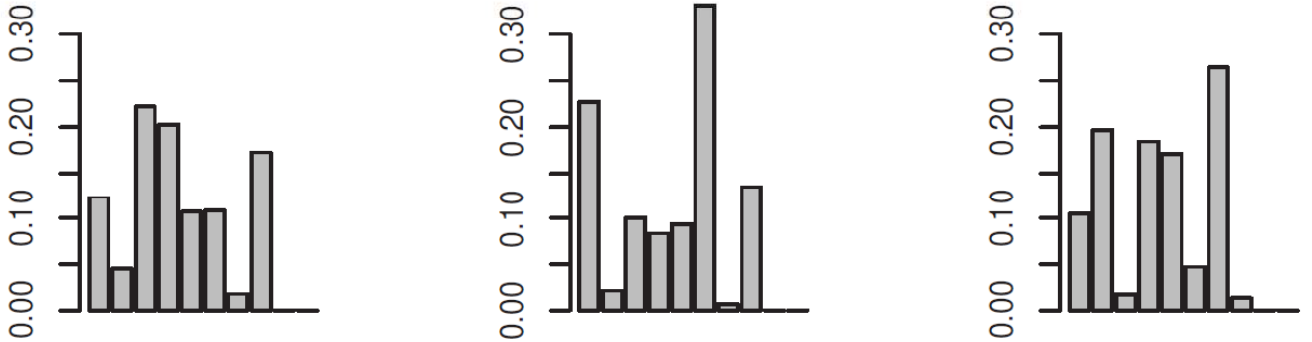


$p_1 = (p_{11}, ..., p_{1J})$ $p_2 = (p_{21}, ..., p_{2J})$ $p_3 = (p_{31}, ..., p_{3J})$

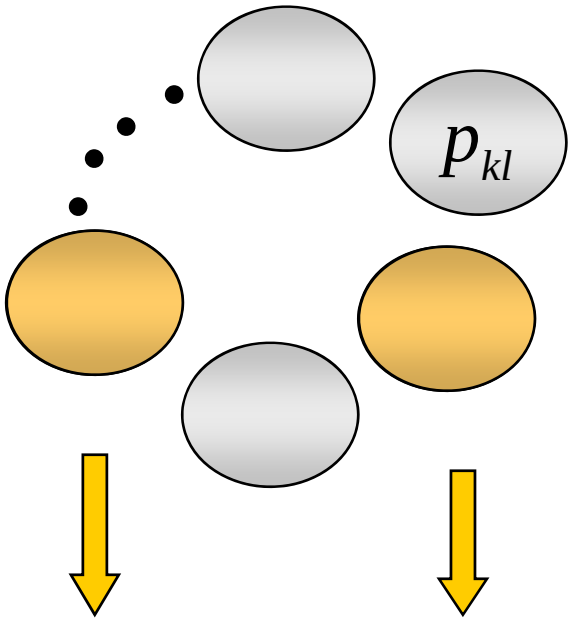
$F_{ST} = 0.01$
$(\theta = 99)$



$F_{ST} = 0.1$
$(\theta = 9)$



Sampling from a metapopulation



Sampling of localities
Sampling of allele counts

Distribution of allele frequencies

$$p_{kl} = (p_{kl1},\ldots,\ p_{klJ}) \sim D(\theta\beta_{l1},\ldots,\ \theta\beta_{lJ})$$

Variance of allele frequencies

$$Var\,(p_{klj}) = \frac{1}{1+\theta}\,\beta_{lj}(1-\beta_{lj})$$

$$F_{ST} = \frac{1}{1+\theta}$$

$$n_{kl} = (n_{kl1},\ldots,\ n_{klJ}) \mid p_{kl} \sim Multi\,(N_k;\ p_{kl1},\ldots,\ p_{klJ})$$

$$n_{kl} \sim DM(N_k;\theta\beta_1,\ldots,\theta\beta_J)$$

Probability distribution of allele counts given true allele frequencies

Multinomial $$f(n_{kl}|p_{kl}) = \frac{N_k!}{\prod_{j=1}^{J_l} n_{klj}!} \prod_{j=1}^{J_l} p_{klj}^{n_{klj}}$$

True allele frequencies

Dirichlet $$f(p_{kl};\theta,\beta_l) = \frac{\Gamma(\theta)}{\prod_{j=1}^{J_l}\Gamma(\theta\beta_{lj})} \prod_{j=1}^{J_l} p_{klj}^{\theta\beta_{lj}-1}$$

Marginal distribution of allele counts

Dirichlet-Multinon

$$\begin{aligned} f(n_{kl};\theta,\beta_l) &= \int\cdots\int f(n_{kl}|p_{kl}) f(p_{kl};\theta,\beta_l) dp_{kl} \\ &= \frac{N_k!}{\prod_{j=1}^{J_l} n_{klj}!} \cdot \frac{\Gamma(\theta)}{\Gamma(N_k+\theta)} \cdot \prod_{j=1}^{J_l} \frac{\Gamma(n_{klj}+\theta\beta_{lj})}{\Gamma(\theta\beta_{lj})}. \end{aligned}$$

$$\begin{aligned} L(\theta,\beta) &= \prod_{k=1}^{K}\prod_{l=1}^{L} f(n_{kl};\theta,\beta_l) \\ &= \prod_{k=1}^{K}\prod_{l=1}^{L} \frac{N_k!}{\prod_{j=1}^{J_l} n_{klj}!} \cdot \frac{\Gamma(\theta)}{\Gamma(N_k+\theta)} \cdot \prod_{j=1}^{J_l} \frac{\Gamma(n_{klj}+\theta\beta_{lj})}{\Gamma(\theta\beta_{lj})}. \end{aligned}$$

ML estimate

Neyman-Scott problem

	Pop 1	Pop 2	⋯	Pop K
Locus 1	$\beta_1 = (\beta_{11},\ldots,\beta_{1J_1})$			
Locus 2	$\beta_2 = (\beta_{21},\ldots,\beta_{2J_2})$			
⋯	⋯			
Locus L	$\beta_L = (\beta_{L1},\ldots,\beta_{LJ_L})$			

Non-consistency of ML estimation of θ if K is small :
Typical problems in ML method in the presence of many nuisance parameters

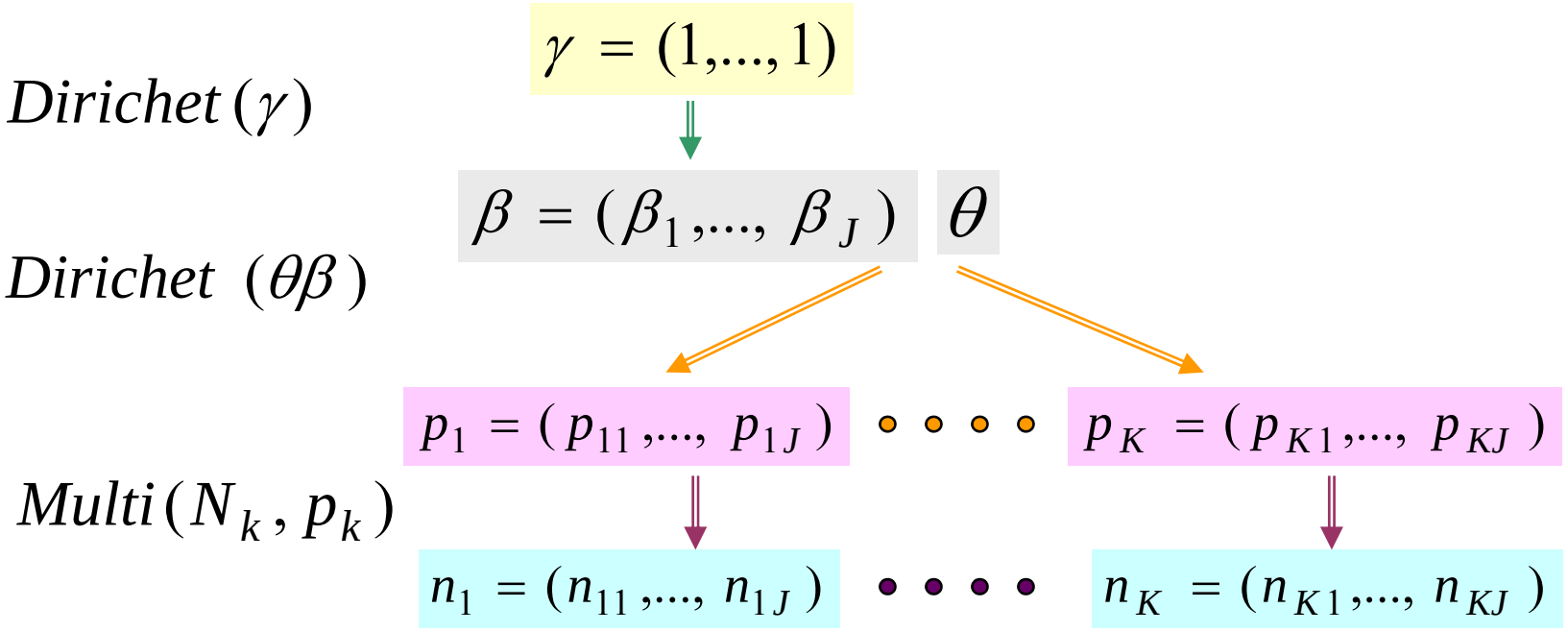
Separation of likelihood is impossible for DM case

Integrated likelihood (Kitakado et al 2006)

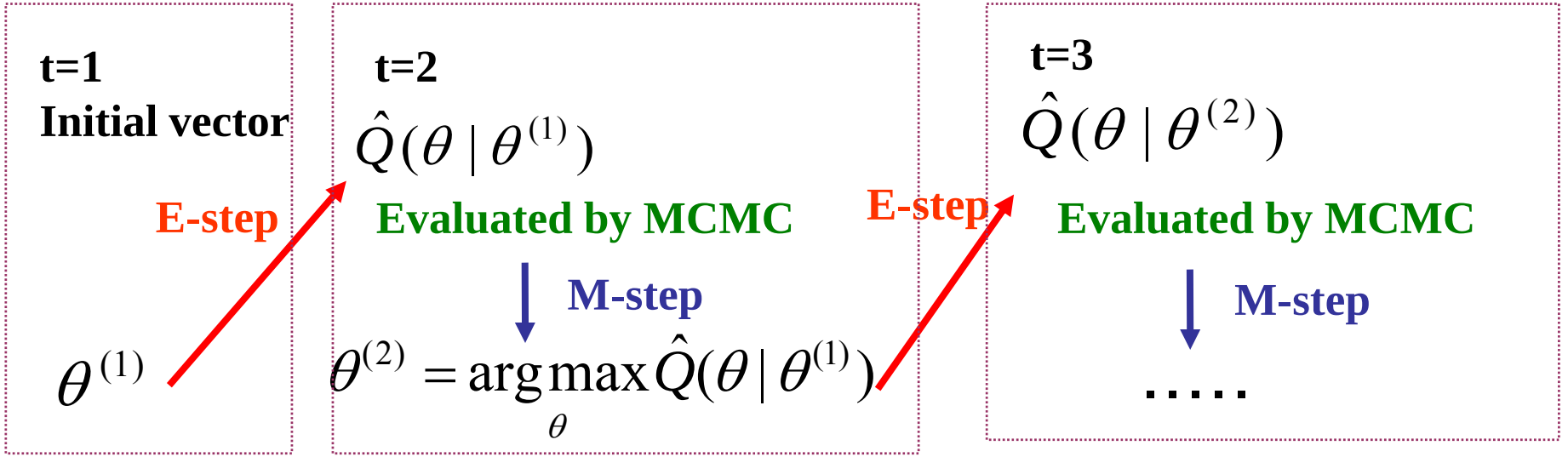
Integrated-likelihood

$$L_I(\theta) = \int_D L(\theta, \beta) d\beta$$
$$= \prod_{l=1}^{L} \left\{ \int \cdots \int \prod_{k=1}^{K} \frac{N_k!}{\prod_{j=1}^{J_l} n_{klj}!} \cdot \frac{\Gamma(\theta)}{\Gamma(N_k + \theta)} \cdot \prod_{j=1}^{J_l} \frac{\Gamma(n_{klj} + \theta\beta_{lj})}{\Gamma(\theta\beta_{lj})} d\beta_l \right\},$$

No closed formula of the integrated likelihood
Direct maximization is impossible



MCEM algorithm



MCEM algorithm:

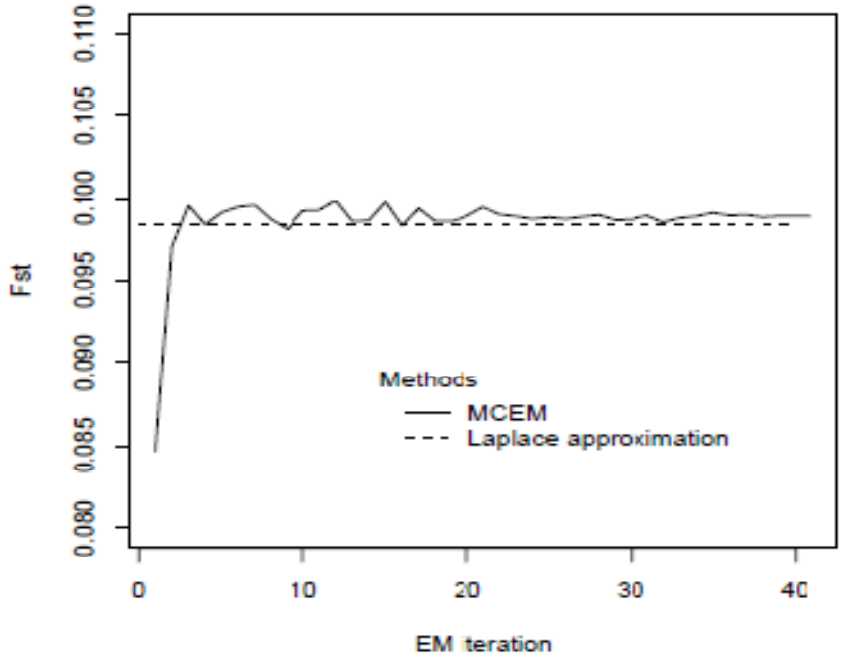
- **Convergence MLE**
- **Iterate MCMC sampling at each step**
- **Slow convergence** ⇒ **Need another algorithm with faster computation**

⇒A Laplace approximation

Laplace approximation

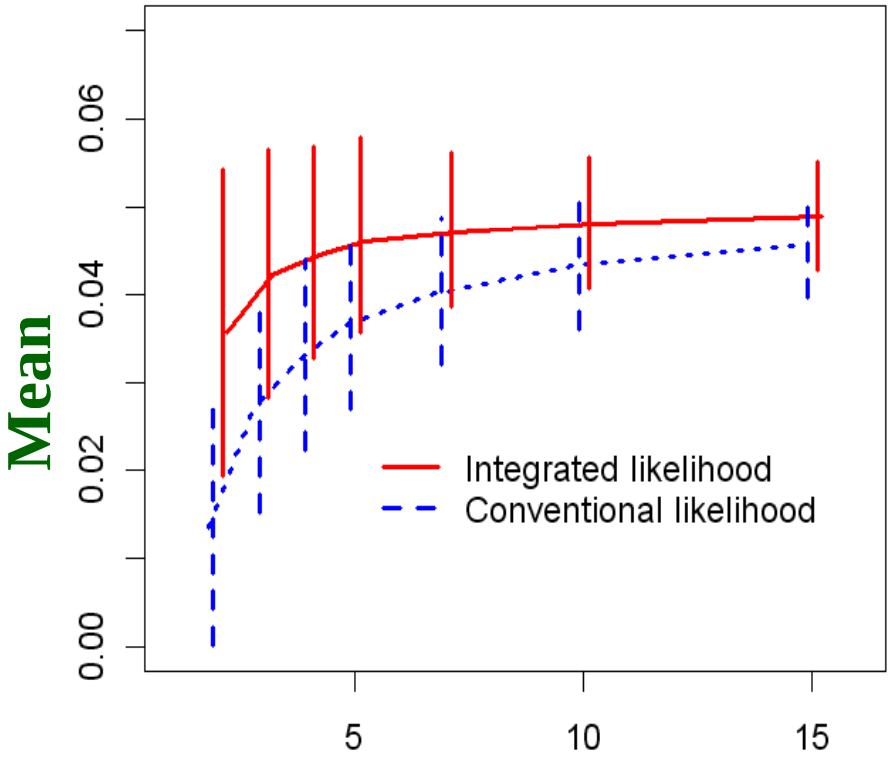
$$L_I(\theta) = \prod_{l=1}^{L} \int f(n_{1l}, \ldots, n_{Kl} | \theta, \beta_l) d\beta_l$$

$$\begin{aligned} L_I^{(l)}(\theta) &\approx \int \exp\left\{ \log f(n_l | \theta, \hat{\beta}_l(\theta)) - \frac{1}{2}(\beta_l - \hat{\beta}_l(\theta))' H(\theta)(\beta_l - \hat{\beta}_l(\theta)) \right\} d\beta_l \\ &= \det\{H(\theta)\}^{-1/2} f(n_l | \theta, \hat{\beta}_l(\theta)), \end{aligned}$$

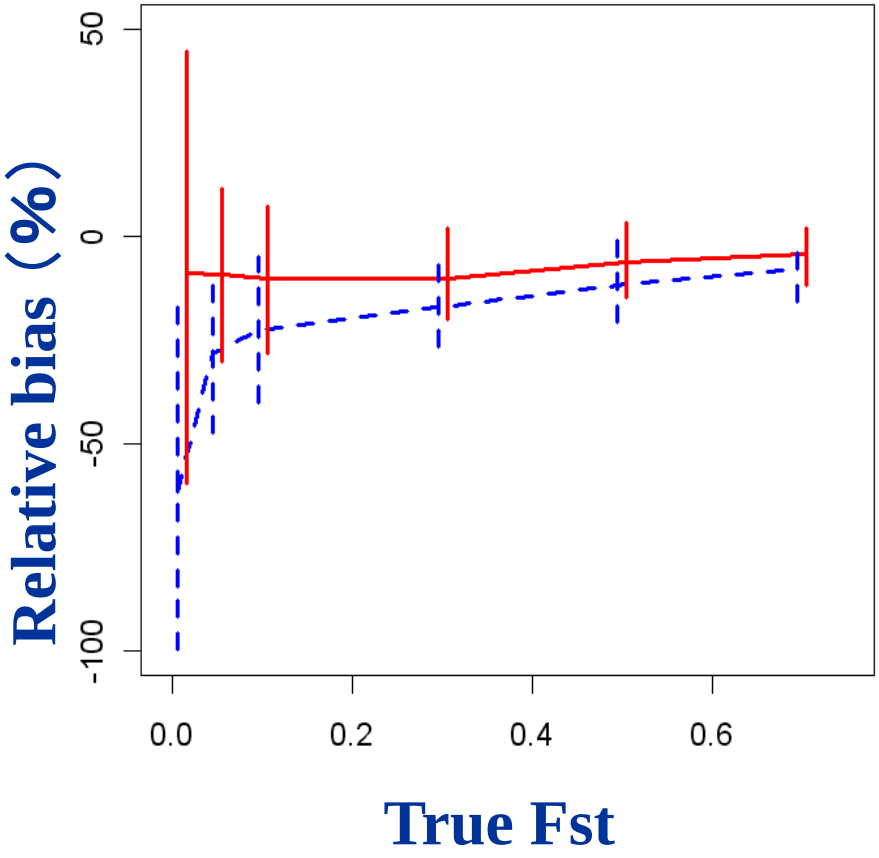


ADMB-RE
(Skaug and Fournier, 2006)

Comparison between conventional ML and IL methods



（True Fst=0.05）



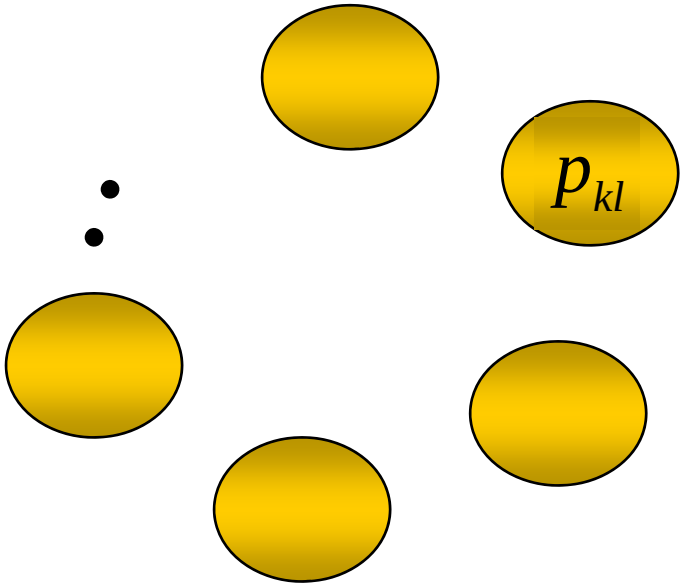
（#subpopulations=5）

Estimation result

Case	Method	θ	F_{ST}
Pacific herring	ML	156.6 (13.4)	0.0063 (0.00054)
	PL	167.9 (13.8)	0.0059 (0.00049)
	IL	91.7 (9.2)	0.0108 (0.00109)
African elephants	ML	1.86 (0.486)	0.350 (0.109)
	PL	1.97 (0.503)	0.337 (0.103)
	IL	1.67 (0.433)	0.374 (0.114)
Channel Island foxes	ML	0.425 (0.210)	0.702 (0.178)
	PL	0.421 (0.206)	0.704 (0.175)
	IL	0.403 (0.198)	0.713 (0.170)

As shown in simulation studies, a large difference between IL and ML(PL) was observed in case of small Fst (like in fish and birds)

Pairwise Fst



Hierarchical models and global Fst improve the estimation performance of pairwise Fst

Empirical Bayes estimation of pairwise Fst

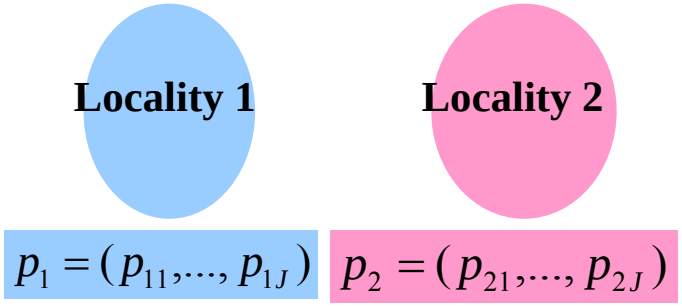
Kitada, Kitakado and Kishino (2007)

$$n_{kl} = (n_{kl1},\ldots,n_{klJ}) \mid p_{kl} \sim Multi(N_k; p_{kl1},\ldots, p_{klJ_l})$$

$$p_{kl} = (p_{kl1},\ldots, p_{klJ}) \sim D(\theta\beta_{l1},\ldots,\theta\beta_{lJ})$$

$$\Rightarrow \quad p_{kl} \mid n_{kl} \sim D(\hat{\theta}\hat{\beta}_{l1} + n_{kl1},\ \ldots,\ \hat{\theta}\hat{\beta}_{lJ_l} + n_{klJ_l})$$

$$F_{St}^{P} = \frac{H_T - H_S}{H_T} \qquad H_T = 1 - \sum \bar{p}_j^2 \qquad H_S = 1 - \frac{1}{2}\sum\sum p_{kj}^2$$



- **Empirical Bayes estimator, simulated posterior distribution, shrinks to the global Fst and has better estimation performance**
- **The nominal estimator has less accuracy**

Improvement of the estimation of pairwise Fst by EB methods

$K=5$,
$N_k=30$

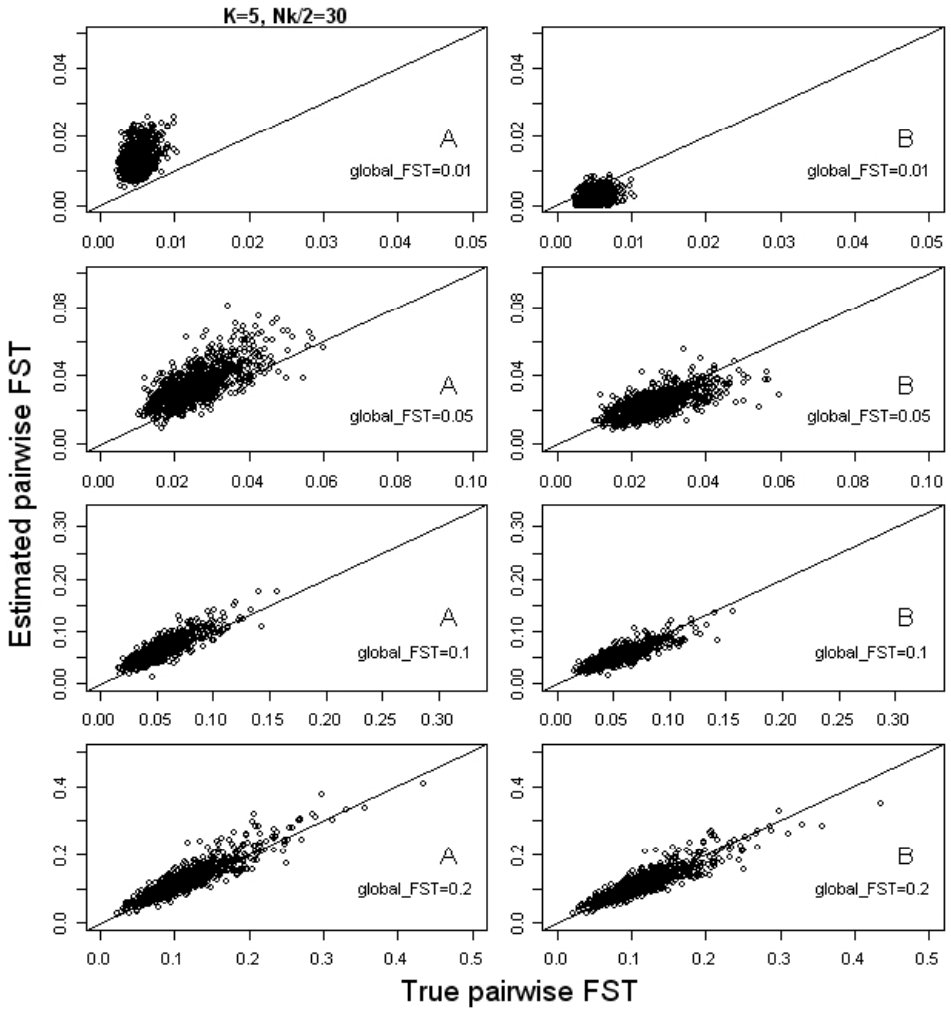
Conventional **EB**

Global Fst =0.01

Global Fst =0.05

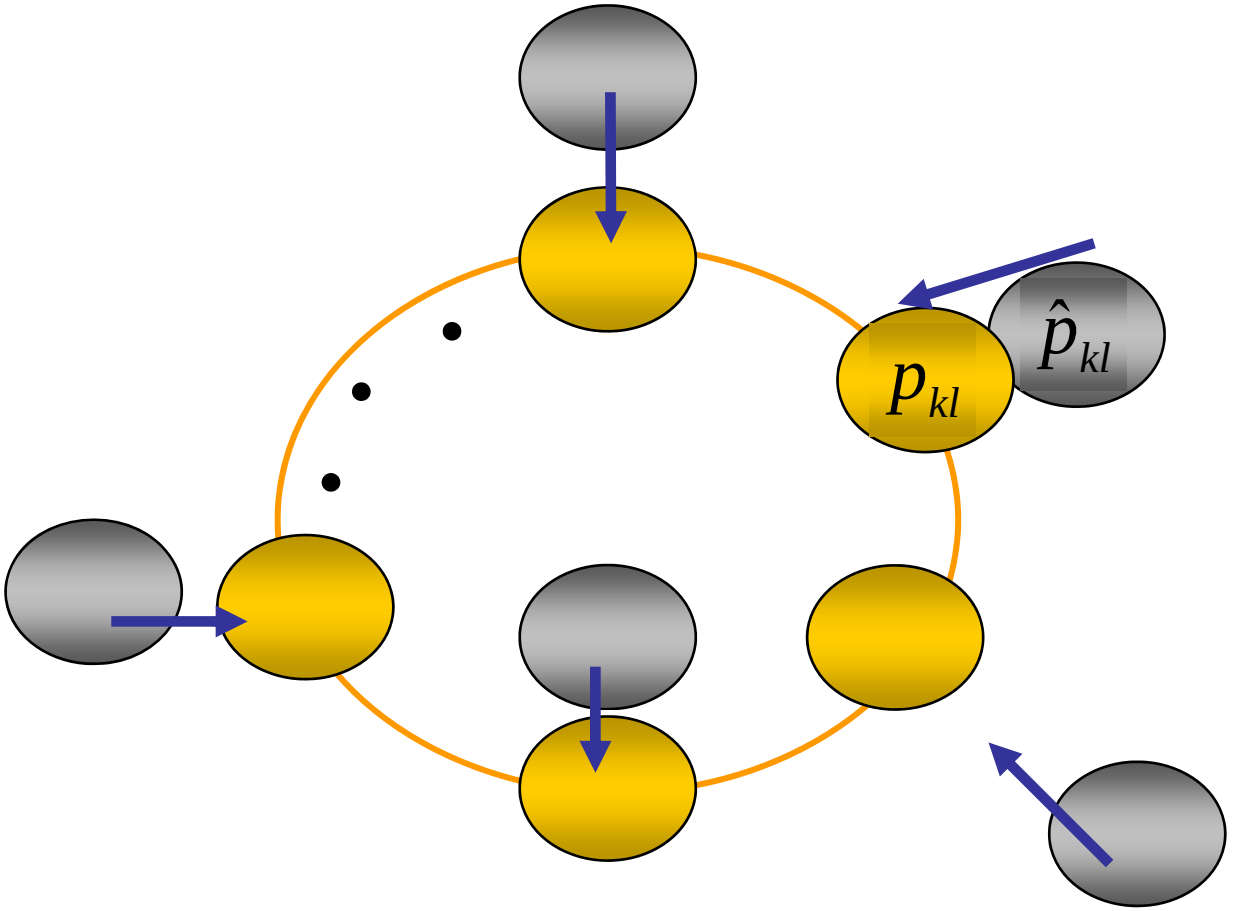
Global Fst =0.1

Global Fst =0.2



Shrink to the mean allele frequencies, which makes the pairwise estimates stable, less biased and variance

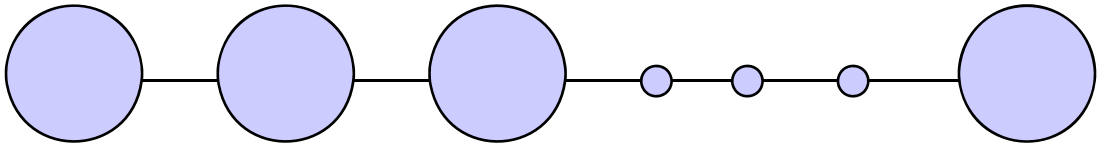
Shrinkage effect in empirical Bayes



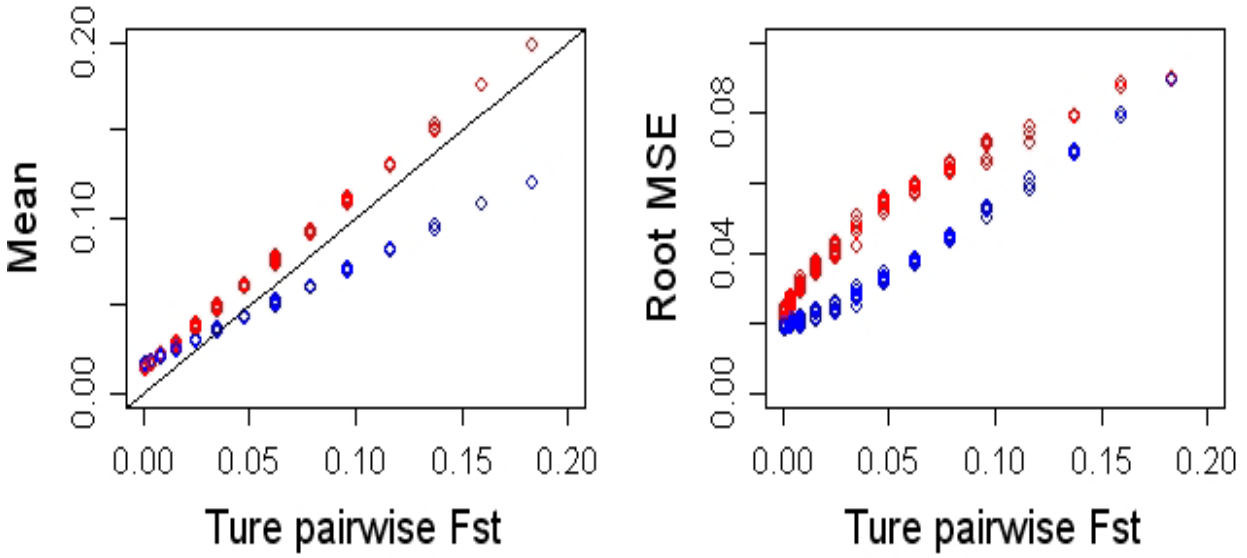
Shrinkage to the mean allele frequencies can make their estimates stable. This is effective especially when the number of sampling localities is large while sample size from each locality is small

Robustness of empirical Bayes for estimation of pairwise Fst

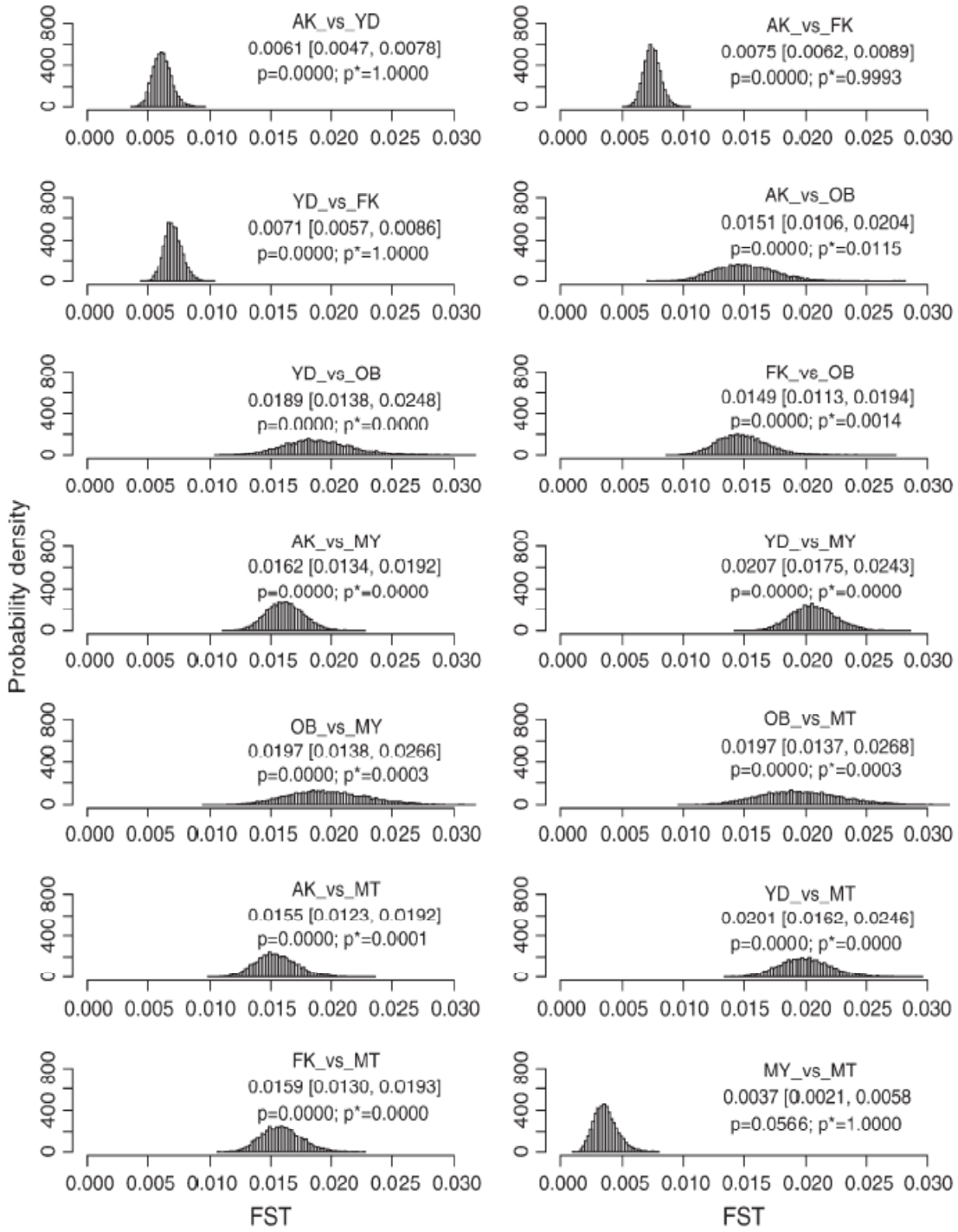
- Stepping Stone Model
- 15 subpopulations
- FST=0.001 between two adjacent subpopulations



Nei's GST and EB



Metapopulation assumption works well as a working model to get a better estimation performance



Posterior distributions of pairwise FST (herring)

From Kitada et al. (2007)

Software List

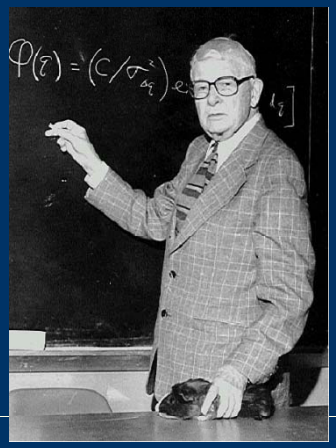
- *POPDIF* : Empirical Bayese estimation of pairwise Fst between subpopulations
- *POPMIX* : Maximum likelihood estimation of mixing proportions from composite genotype data

POPDIF : Empirical Bayese estimation of pairwise Fst between subpopulations

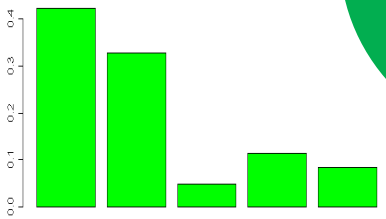
Publication	Kitada, S., T. Kitakado and H. Kishino (2007) Empirical Bayes inference of pairwise FST and its distribution in the genome. Genetics 177, 861-873. (PDF)
	Kitakado, T., Kitada, S., H. Kishino and H. J. Skaug (2006) An Integrated-Likelihood Method for Estimating Genetic Differentiation Between Populations. Genetics 173, 2073-2082. (PDF)
	Kitada, S. and H. Kishino (2004) Simultaneous detection of linkage disequilibrium and genetic diffrentiation of subdivided populations. Genetics 167, 2003-2013. (PDF)
Summary	*POPDIF* estimates locus-specific global Fst and the rate of gene flow by maximum/integrated likelihood method from genotype data of K geographical samples. On the basis of the global Fst estimate, posterior distributions for all sets of the population pairwise Fst are simulated. Allele data with Genepop format or allele (haplotype) frequencies are available.
Author	Toshihide Kitakado, Shuichi Kitada and Hirohisa Kishino
Download	Download program and manual.(in preparation)

http://www2.kaiyodai.ac.jp/~kitada/Conservation/index_eng.html

3. Estimation of population mixture



Population differentiation -> mixture



Locality 1
(Sampling Area 1)

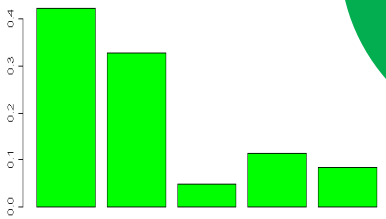
Locality 2
(Sampling Area 2)

$$p_1 = (p_{11}, \ldots, p_{1J})$$

$$p_2 = (p_{21}, \ldots, p_{2J})$$

Population differentiation -> mixture

Baseline 1



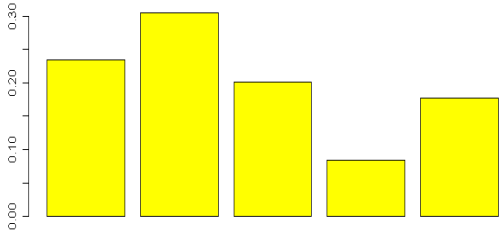
Locality 1
(Sampling Area 1)

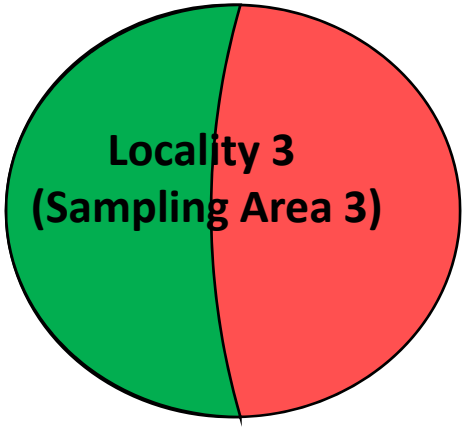
Locality 2
(Sampling Area 2)

Baseline 2

Mixing prop
ω : 1−ω

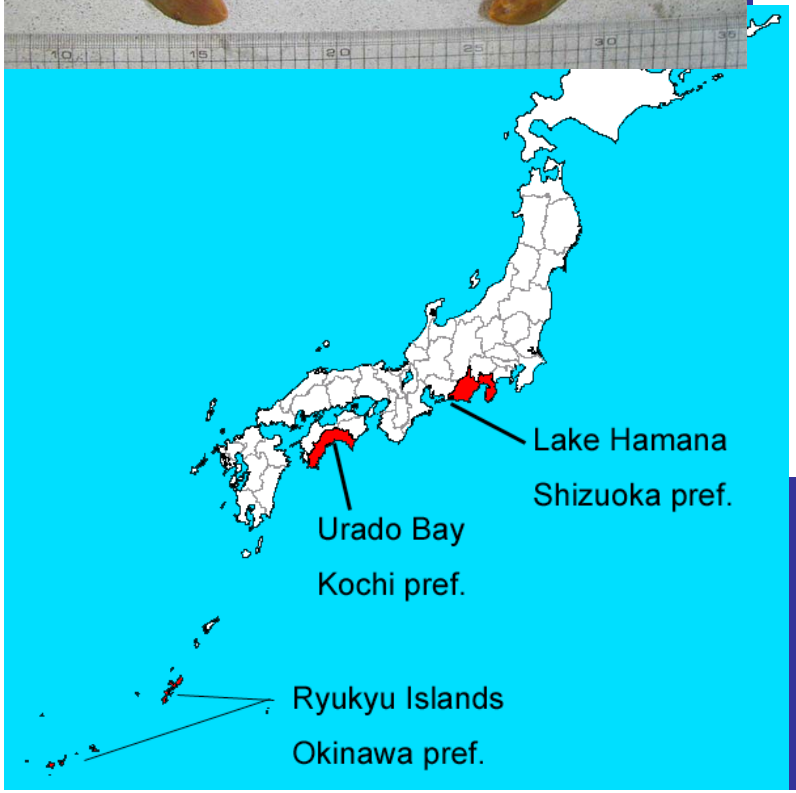
Mixed stock

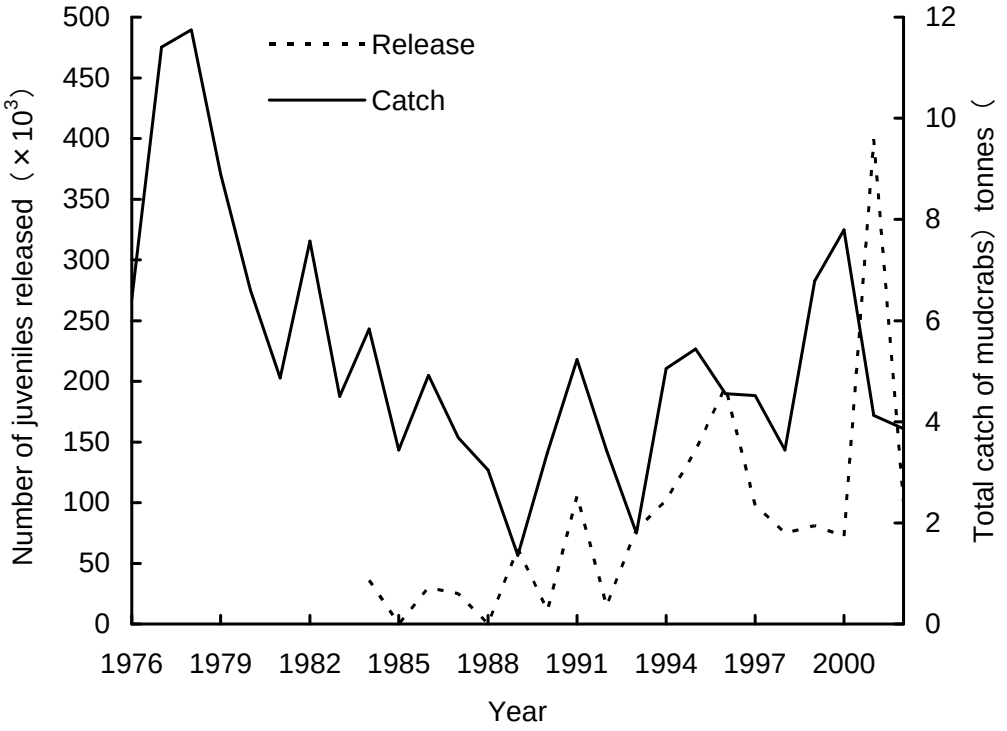




If allele counts from the two baselines and mixed are observable, then possible to estimate the mixing proportion

Enhancement program for mud crab in Japan





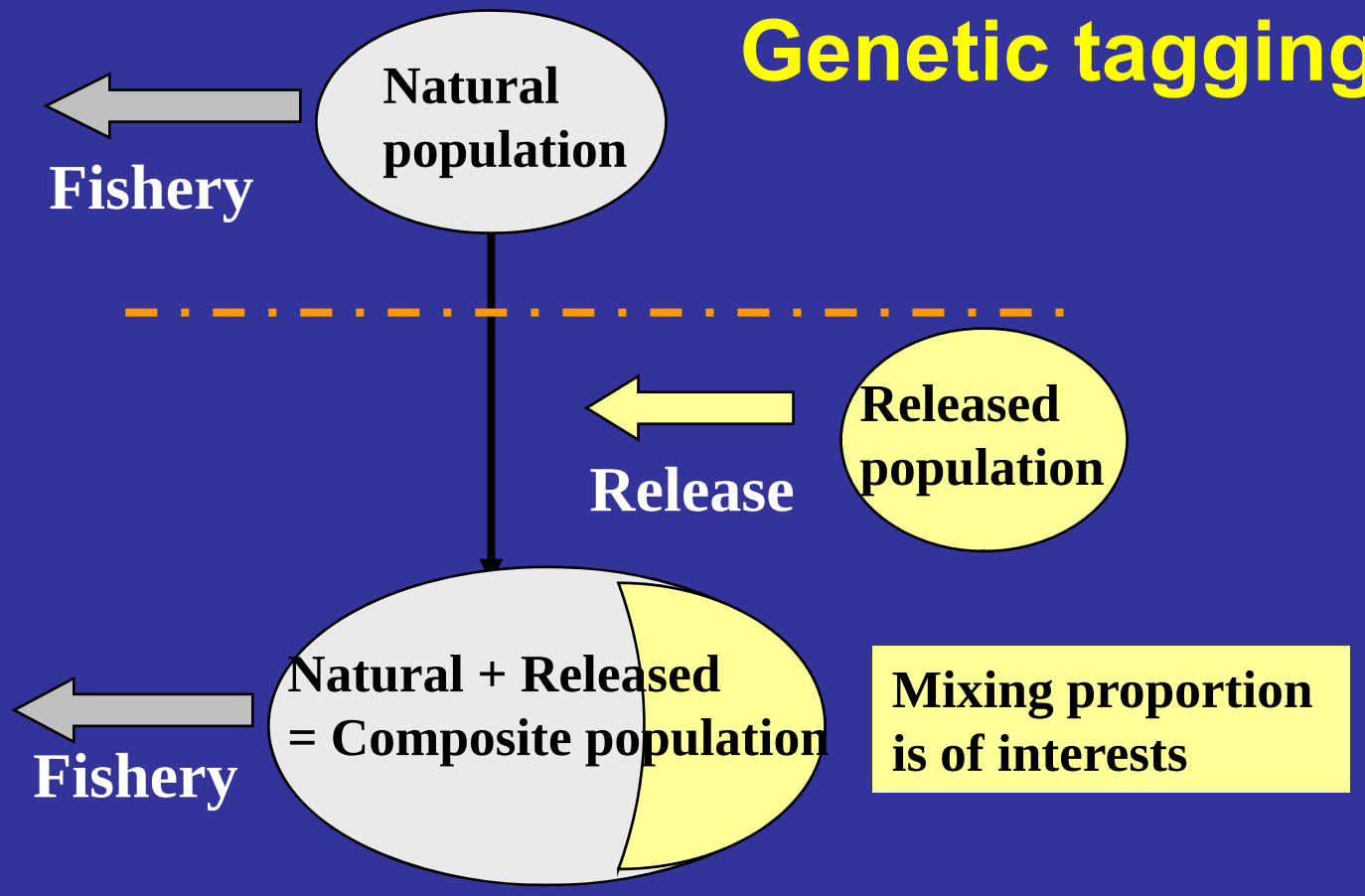
From Obata et al.(2006)

Effectiveness of Stock Enhancement Programs

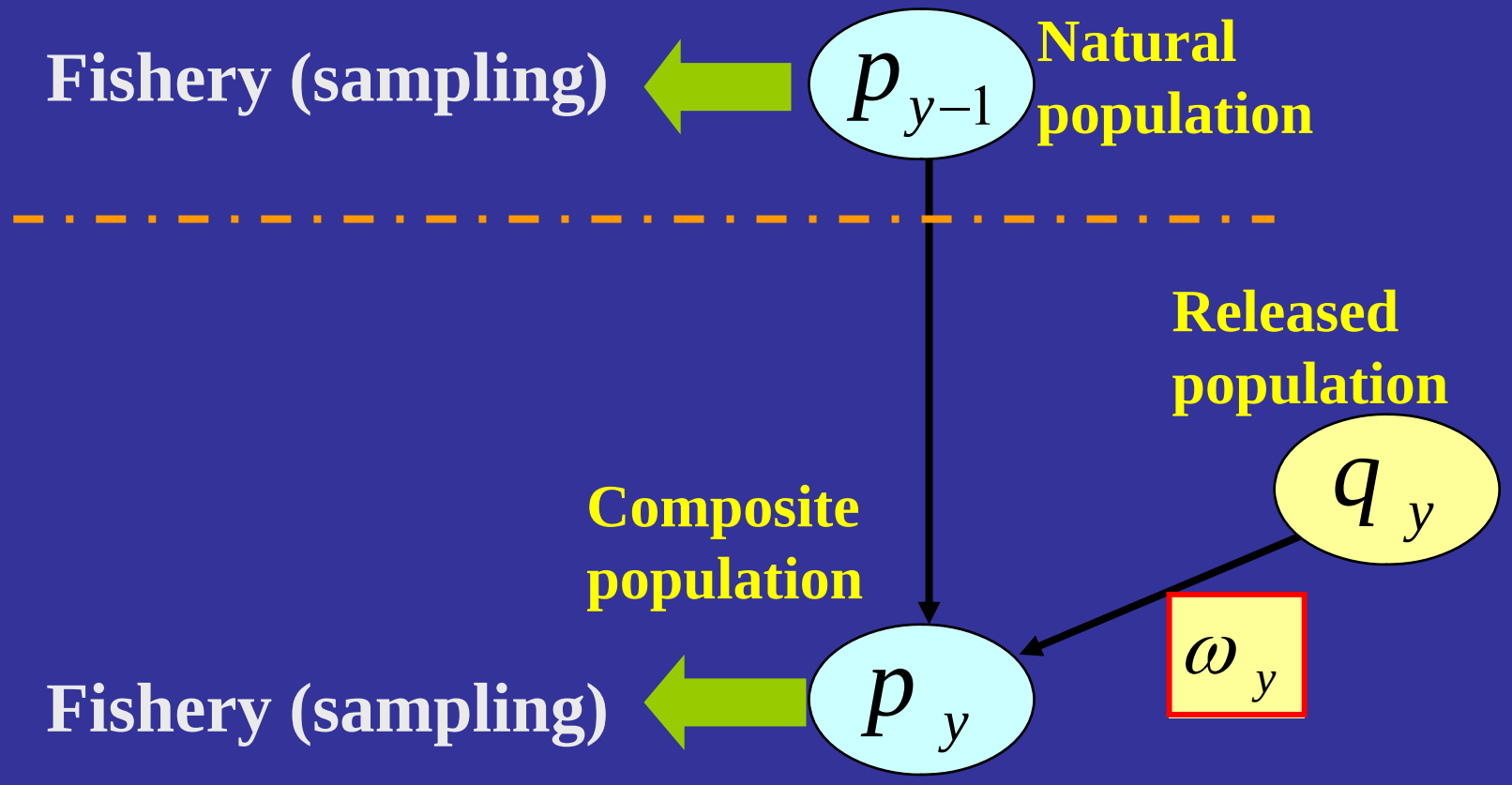
- **Recovery rate**
  Too small recovery rates means ineffectiveness of programs

- **Mixing proportion**
  Too large contribution of released juveniles to a natural population raises concern reducing its effective population size (e.g., Ryman and Laikre, 1991)

Estimation of mixing proportions

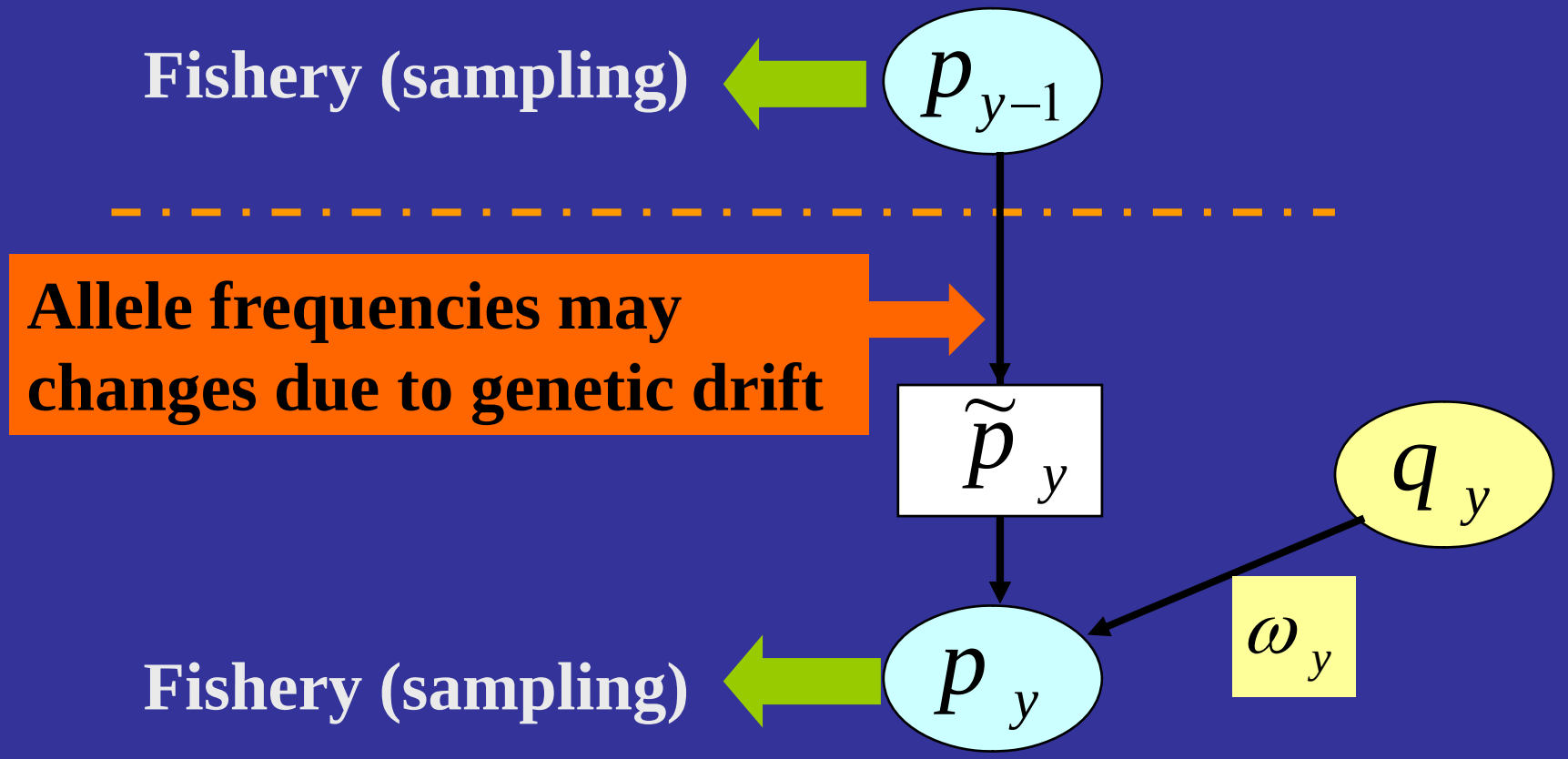


A schematic diagram for the estimation of mixing proportion



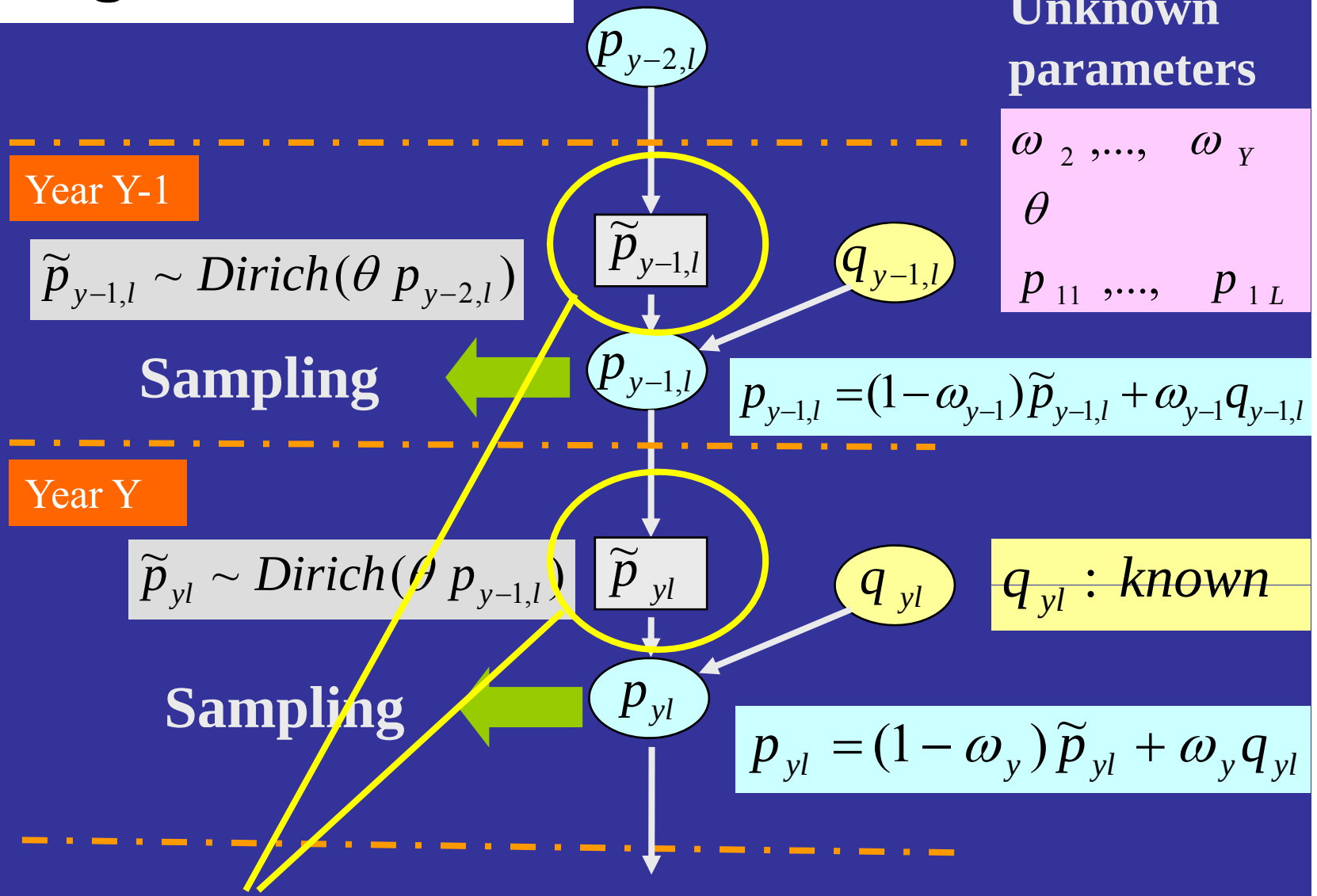
$$p_y = (1 - \omega_y)\, p_{y-1} + \omega_y q_y$$

A schematic diagram with genetic drift



$$p_y = (1-\omega_y)\tilde{p}_y + \omega_y q_y$$

Full diagram of model



Unobservable latent variables makes model complicated

Constructing a full likelihood

Individual genotype

$$x_{yil}^{(1)} = (x_{yil1}^{(1)}, ..., x_{yilJ_l}^{(1)}) \quad maternal\,allele$$

$$x_{yil}^{(1)} = (x_{yil1}^{(1)}, ..., x_{yilJ_l}^{(1)}) \quad paternal\,allele$$

Joint probability of observation and latent variables

$$L_{complete}(p_1, \tilde{p}_2, ..., \tilde{p}_Y, \theta, \omega_2, ..., \omega_Y)$$

$$= L_1(p_1, \tilde{p}_2, ..., \tilde{p}_Y, \omega_2, ..., \omega_Y \mid Data) \cdot L_2(p_1, \theta, \omega_2, ..., \omega_Y \mid \tilde{p}_2, ..., \tilde{p}_Y)$$

Marginal likelihood for observation (Integrated likelihood)

$$L_{obs}(\theta, \omega_2, ..., \omega_Y)$$

No closed form

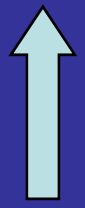
$$= \int \cdots \int L_{complete}(p_1, \tilde{p}_2, ..., \tilde{p}_Y, \theta, \omega_2, ..., \omega_Y) \prod_{l=1}^{L} dp_{1l} d\tilde{p}_{2l} \cdots d\tilde{p}_{Yl}$$

Estimation with latent variables

Several approaches have been developed

- Consider as a full Bayesian model
- Monte Carlo EM algorithm
- Importance sampling

- **Laplace approximation (analytical)**



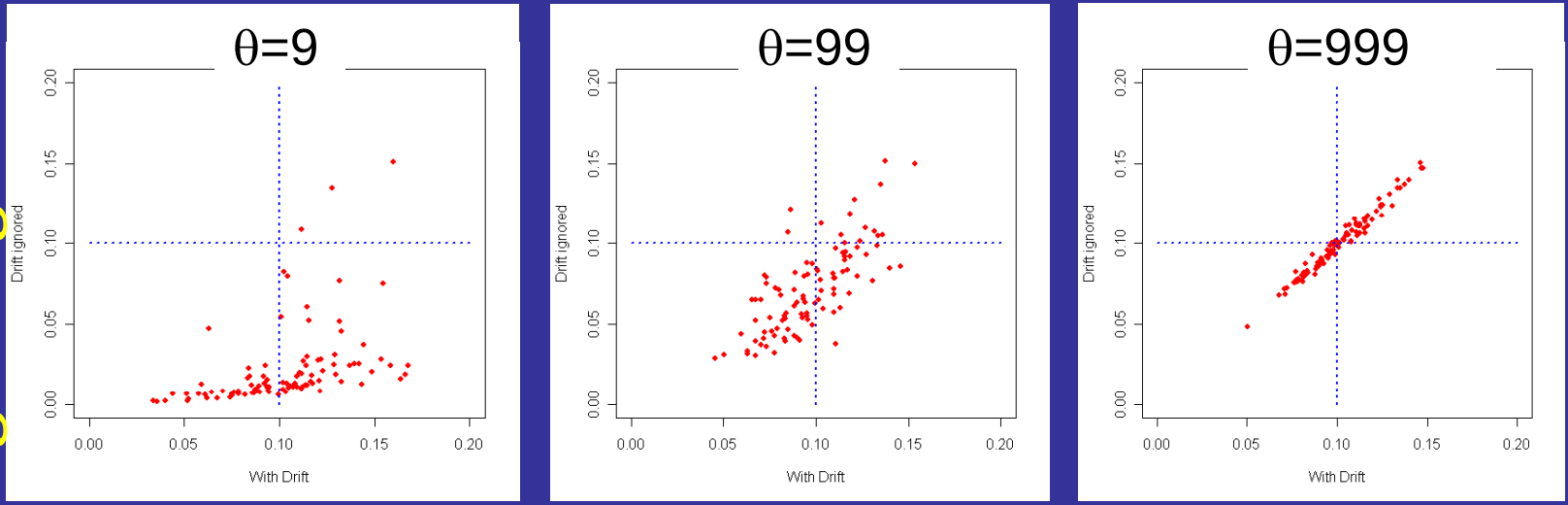
Markov chain Monte Carlo (MCMC) is typically utilized.

Computationally intensive

Simulation

Scatter plots of estimates in 3rd year

Ignoring drift



Considering genetic drift

- Estimates with considering genetic drift distributed around the true value of mixing proportion
- Model with ignoring genetic drift caused severe underestimation when the drift was large

Application to mud crab data

Observed haplotype frequencies from fishing ground

	Haplotype									
Year	1	2	3	4	5	6	7	8	9	Others
1996	79	23	13	16	10	3	1	3	0	1
1997	90	65	16	13	11	2	5	2	0	9
1998	57	37	22	11	7	0	1	2	0	6
1999	159	124	39	15	11	3	5	5	5	37
2000	225	147	47	32	21	2	4	5	4	34
2001	71	32	8	7	5	0	0	1	11	23
2002	105	50	30	17	0	5	0	0	5	35

Haplotype frequencies of released populations

	Haplotype									
Year	1	2	3	4	5	6	7	6	7	Others
1996	0	0	0	0	0	0	0	0	0	0
1997	0	98,300	0	0	0	0	0	0	0	0
1998	0	0	64,000	0	0	0	11,000	0	0	0
1999	0	0	0	81,000	0	0	0	0	0	0
2000	0	0	0	0	72,000	0	0	0	0	0
2001	0	0	0	0	0	0	0	0	149,000	0
2002	0	0	0	0	0	0	0	0	0	0

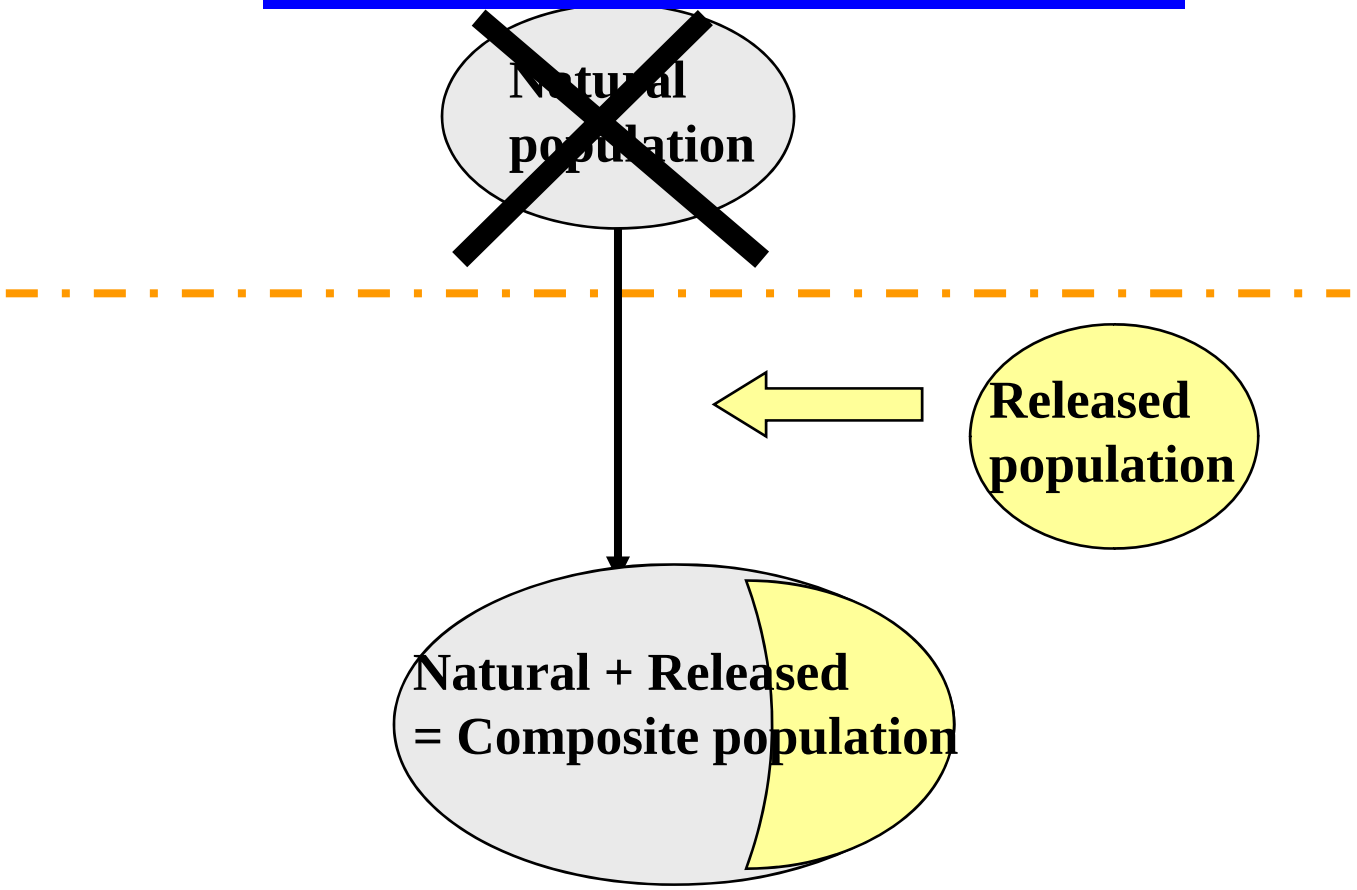
Estimation results

Year	Mixing Proportion	
1997	0.152	(0.023)
1998	0.025	(0.021)
1999	0.000	(0.000)
2000	0.015	(0.007)
2001	0.039	(0.009)

- Relatively small level of genetic drift was observed
- Due to this, only a slight impact was given on the estimation of mixing proportions in this example
- Low level of mixing proportions except for 1997

Mixing proportions released (again)

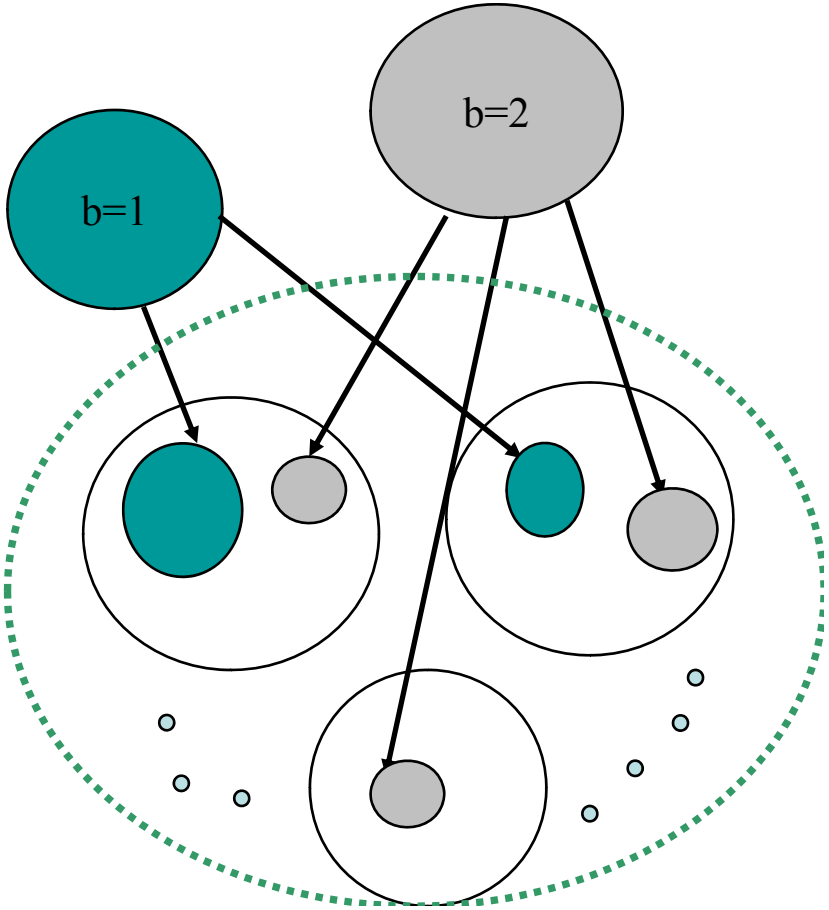
Sometimes no information is available before releasing

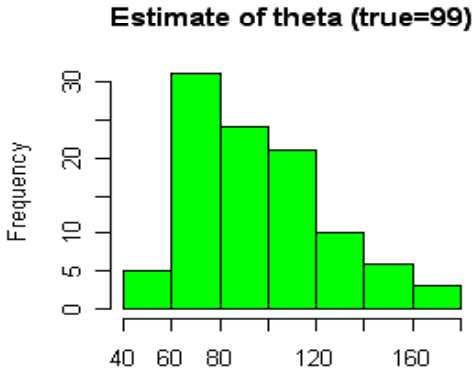


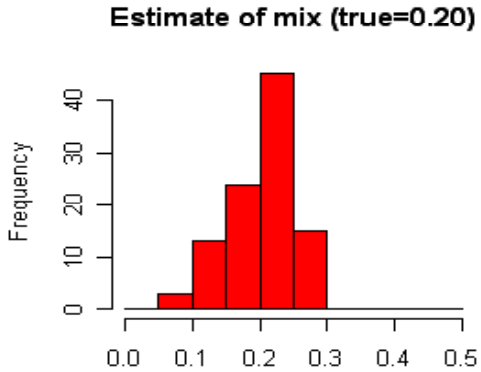
An integrated likelihood function and results

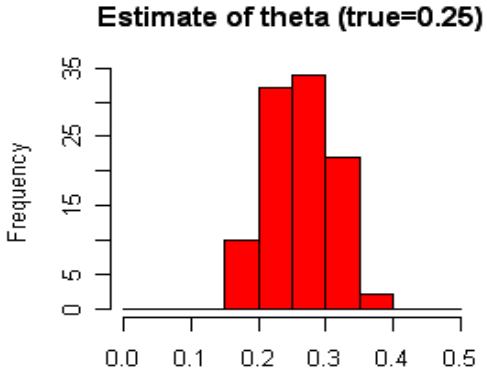
$$L(\boldsymbol{\phi}, \boldsymbol{p}, \boldsymbol{q}) = \left[\prod_{a=1}^{A}\prod_{i=1}^{N_a}\left\{\sum_{k=0}^{B}\phi_{ak}\prod_{l=1}^{L}\prod_{h=1}^{2} f(y_{ail}^{(h)}|z_{ai}=k)\right\}\right]\cdot\left[\prod_{b=1}^{B}\prod_{l=1}^{L}\left\{\frac{(2N_b)!}{\prod_{j=1}^{J_l} n_{blj}!}\prod_{j=1}^{J_l} q_{blj}^{n_{blj}}\right\}\right]$$

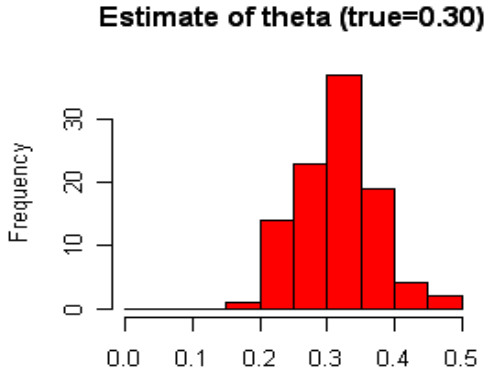
$$L_I(\boldsymbol{\phi}, \theta) = \int_{D_{\boldsymbol{p},\boldsymbol{q},\boldsymbol{\beta}}} L(\boldsymbol{\phi}, \boldsymbol{p}, \boldsymbol{q}) f(\boldsymbol{p}; \theta, \boldsymbol{\beta}) d\boldsymbol{p} d\boldsymbol{q} d\boldsymbol{\beta}$$











4. Concluding remarks

Concluding remarks

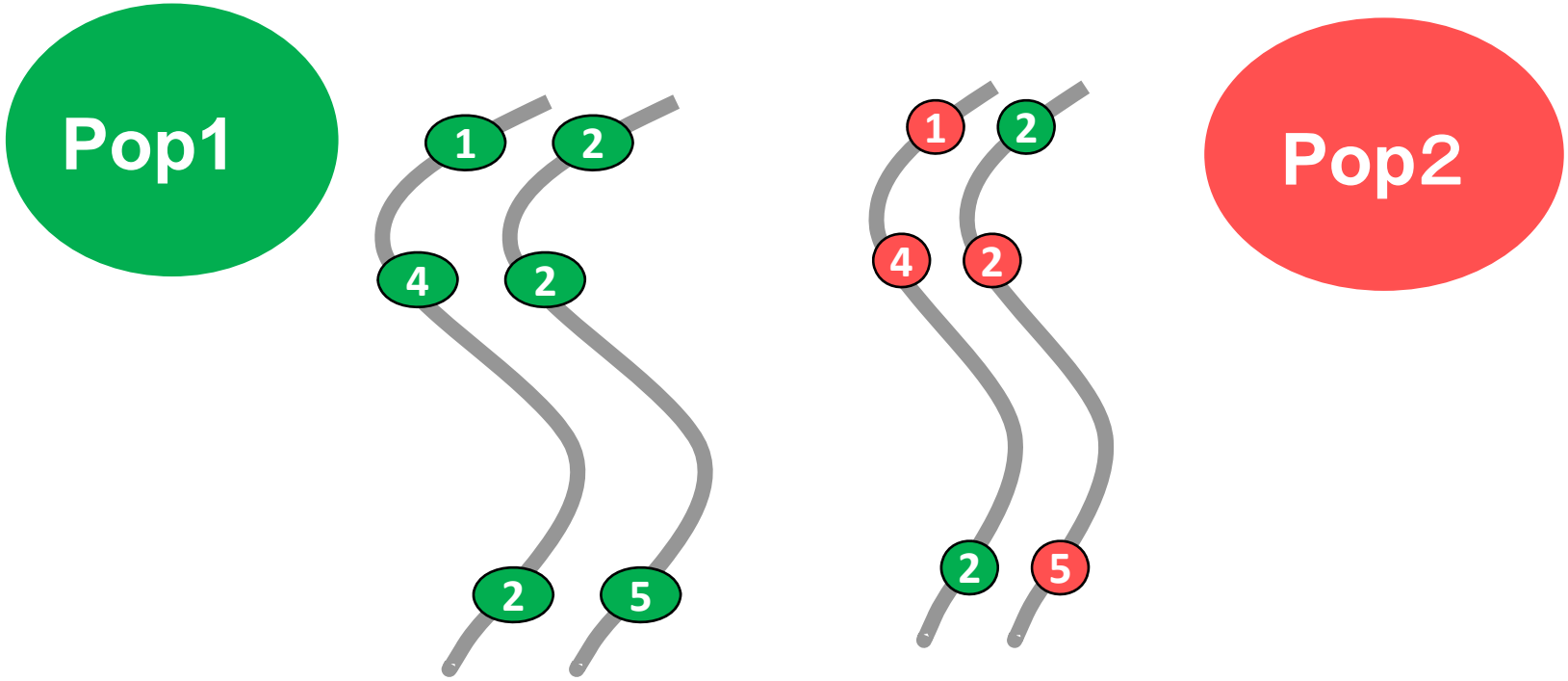
- As in analyses for human population, use of SNPs with a large number of sites may be common in science for wildlife

Concluding remarks

- As in analyses for human population, use of SNPs with a large number of sites may be common in science for wildlife

- Quite often, individual assignment methods without assuming baseline populations are applied, but the reliability of method especially in the estimation of the number of populations is still open to question

Mixture and admixture

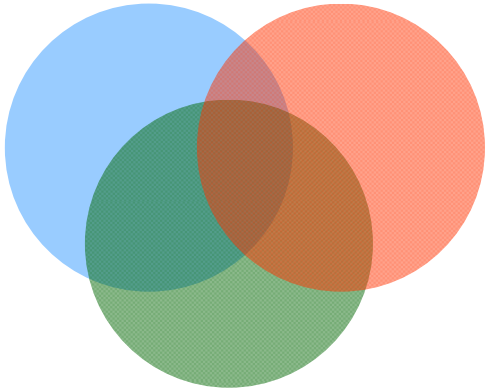


Mixture
The target is origin of an individual
Assignment probability =prob that the individual comes from a population

Admixture
The target is origin of alleles in an individual
Assignment probability =proportion of alleles coming from an ancestral population

Individual assignment

Pritchard et al(2000), Falush et al(2003)



$$y_{i1}^{(1)} = (0,0,1,0,0), \quad y_{i1}^{(2)} = (0,1,0,0,0),$$
$$\cdots \qquad\qquad \cdots$$
$$y_{iL}^{(1)} = (0,0,0,0,1,0,0), \quad y_{iL}^{(2)} = (1,0,0,0,0,0,0).$$

Z_i : latent variable representing individual's origin

$$Y_{il}^{(m)} \mid Z_i = z_i,\ p_{z_i l} \sim Multi(1; p_{z_i l})$$

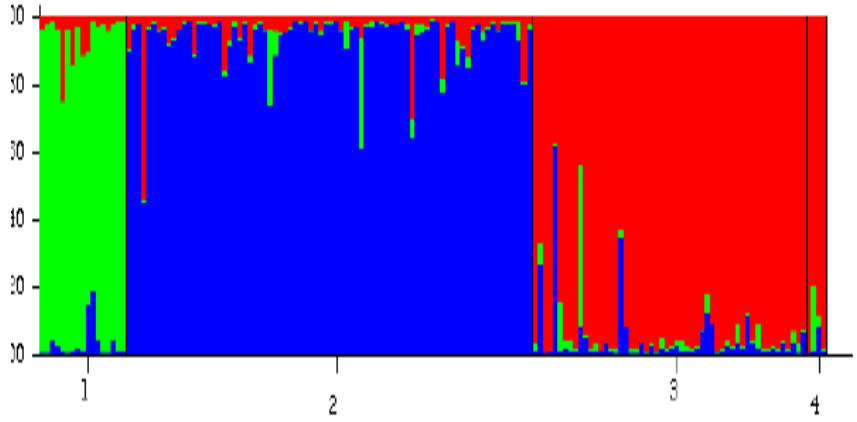
$$P(Z_i = k) = \frac{1}{K}$$

$$p_{kl} \sim Dirichlet(\lambda_1, \ldots, \lambda_{J_l})$$

Hyperparameter $\lambda_j = 1 \ (j = 1, \ldots, J)$

(As noninformative prior)

Posterior $p(Z_i = k \mid Y)$




Thank you very much
for your kind attention!