1. Stock mixture analysis

Estimation Problems

1. Stock proportions composing the mixture
2. Source identification of the mixture individuals

Solution via mixture modeling

1. Stock proportions estimated directly by frequentist or Bayesian methods
2. Stock proportions treated as prior probabilities for individual source identification with Bayes' rule or theorem

2. Multilocus genotypes

•Individual's multilocus genotype, **X**, for L loci with J_h alleles at locus h is denoted

$$\mathbf{X} = \left(x_{hj} = 0, 1, \text{or } 2\right), h = 1,\ldots,L;\ j = 1,\ldots,J_h$$

•Corresponding allele (relative) frequencies in stock i are denoted

$$\mathbf{Q}_i = \left(q_{ihj}\right), i = 1,\ldots c;\ h = 1,\ldots,L;\ j = 1,\ldots,J_h$$

•Relative frequency of genotype **X** in stock i is

$$f(X;Q_i) = \prod_{h=1}^{L} 2^{1-\delta_h(\mathbf{X})} \prod_{j=1}^{J_h} q_{ihj}^{x_{hj}}, \qquad \delta_h(\mathbf{X}) = \begin{cases} 1 & \max_j \{x_{hj}\} = 2 \\ 0 & \max_j \{x_{hj}\} = 1 \end{cases}$$

3. Unknowns to be estimated for mixture analysis

Stock proportions:

$$\mathbf{p} = (p_1, \ldots, p_c)' \qquad 0 \le p_i \le 1, \text{ and } \sum_{i=1}^{c} p_i = 1$$

Allele (relative) frequencies for c stocks at L loci:

$$\mathbf{Q} = (\mathbf{Q}_1, \ldots, \mathbf{Q}_c)$$

$$\mathbf{Q}_i = \begin{pmatrix} \mathbf{q}_{i1} \\ \vdots \\ \mathbf{q}_{iL} \end{pmatrix}, \qquad \mathbf{q}_{ih} = (q_{ih1}, \ldots, q_{ihJ_h})', \quad 0 \le q_{ihj} \le 1, \text{ and } \sum_{j=1}^{J_h} q_{ihj} = 1$$

$i = 1, \ldots, c.$

4. Baseline allele frequencies at a locus and non-ignorable "sampling zeros" problem

•Low frequency alleles are likely not to be detected in some stock baseline samples.
•Unless a "zero" baseline allele count is an actual "zero", non-ignorable bias in mixture analysis can result

Possible solutions:

1. Pool alleles
    - No basis for categories
    - Likely loss of information
2. Bayesian modeling to prevent "zeros"

5. Bayesian baseline analysis for informative mixture prior of allele frequencies at a locus

Dirichlet prior + allele counts = Dirichlet posterior

Pella-Masuda prior and Pseudo-Bayes fit

- Dirichlet prior with mean equal to central or regional baseline values
- Dirichlet posterior with mean = objectively weighted average of observed and baseline center

Rannala-Mountain prior

- Dirichlet prior with arbitrary mean of equal frequencies
- Dirichlet posterior with mean = arbitrary weighted average of observed and equal frequencies

6. Finite genetic mixture model for c stocks and a sample of individuals $m = 1,\ldots,M$

$$f(X_m;Q_i) = \prod_{h=1}^{L} 2^{1-\delta_h(X_m)} \prod_{j=1}^{J_h} q_{ihj}^{x_{mhj}}$$

Probability of drawing a mixture individual with genotype $\mathbf{X}_m$:

$$\Pr\{X_m; p, Q\} = \sum_{i=1}^{c} p_i \cdot f(X_m;Q_i),$$

where $0 \le p_i \le 1$ for $i = 1,\ldots,c$, and $\sum_{i=1}^{c} p_i = 1$.

Likelihood function of $\mathbf{p}$ and $\mathbf{Q}$:

$$\Pr\{X_1,\ldots,X_M; p, Q) \propto \prod_{m=1}^{M} \sum_{i=1}^{c} p_i \cdot f(X_m;Q_i)$$

7. Estimation methods of stock proportions

- Most frequently used
- Conditional maximum likelihood with bootstrap resampling for precision
- Bayesian posterior distribution via Markov chain Monte Carlo sampling
- Others
- Classification-based method
- Unconditional maximum likelihood
- Constrained least squares

8. Conditional maximum likelihood

- Assumes allele frequencies in stocks ($\mathbf{Q}_i$) are known from baseline samples
- Stock proportions $\mathbf{p} = (p_1,\ldots,p_c)$ are estimated
- Algorithms for computing estimate of **p** that maximizes the mixture model likelihood function includes EM, conjugate gradient, and iteratively re-weighted least squares (IRLS)
- Precision from bootstrap resampling of baseline and mixture samples

9. EM algorithm for *c* stock mixture

Initial guess: $\mathbf{p}^{(0)} = \left(1/c \ldots, 1/c\right)'$

Iterate the set of c equations for $t = 1,2,\ldots$

$$p_i^{(t)} = \frac{1}{M} \cdot \sum_{m=1}^{M} \left[\frac{p_i^{(t-1)} f(X_m; Q_i)}{\sum_{s=1}^{c} p_s^{(t-1)} f(X_m; Q_s)} \right] = \frac{1}{M} \sum_{m=1}^{M} \Pr^{(t-1)}(i \mid X_m)$$

$i = 1, \ldots, c$

At convergence, each individual has been partitioned into c source fractions, or posterior source probabilities, and the estimated p_i equals the average (over individuals) of the source fractions for the i-th stock.

10. Bayesian estimation of stock composition, **p**, baseline allele frequencies, **Q**, and individual posterior source probabilities, $P(i|\mathbf{X}_m)$

- Unit Dirichlet (low information ~ single fish) mixture prior for **p** will be updated with assigned mixture fish
- Informative baseline posterior becomes the mixture prior for **Q** and will be updated with alleles of assigned mixture fish
- Markov chain Monte Carlo samples are generated for **p** and **Q**
- Advantages over CML:
  1. Uses mixture information to update **Q**
  2. More flexible in attacking non-standard problems
  3. Direct probability statements about **p**

11. MCMC sampling of posterior distribution

1. Draw stock identities of mixture individuals at random with probabilities equal to current values at iteration *t* of posterior source probabilities of genotypes

$$\Pr(i \mid X_m) = \frac{p_i^{(t)} f(\mathbf{X}_m \mid \mathbf{Q}_i^{(t)})}{\sum_{s=1}^{c} p_s^{(t)} f(\mathbf{X}_m \mid \mathbf{Q}_s^{(t)})}, \quad i = 1, \ldots, c$$

2. Draw $\mathbf{p}^{(t+1)}$ and $\mathbf{Q}^{(t+1)}$ from their posterior Dirichlet distributions given the baseline samples and updated with the current assignments of mixture individuals from step 1.
3. Return to step 1 until specified sample size is achieved.

17. Software

- Individual assignments without mixture modeling
  - GeneClass (Cornuet and Piry of INRA Centre of Montpellier, France
  - WHICHRUN (Banks and Eichert, Oregon State U., or U. California Davis)
- Mixture modeling
  - SPAM (ADF&G Genetics Lab) for CML
  - BAYES (Auke Bay site) for Bayesian analysis