# Estimating allele frequencies in non-model polyploids using high throughput sequencing data

Paul Blischak, Laura Kubatko, Andrea Wolfe

Dept. of EEOB
The Ohio State University

June 1, 2015

# ABBA-BABA statistics

$$\mathbf{D} = \frac{\sum \mathcal{C}_{ABBA} - \mathcal{C}_{BABA}}{\sum \mathcal{C}_{ABBA} + \mathcal{C}_{BABA}}$$

$$f = \frac{\sum \mathcal{C}_{ABBA} - \mathcal{C}_{BABA}}{\sum \mathcal{C}_{ABBA} + \mathcal{C}_{BABA}}$$

# Penstemon attenuatus

The *Penstemon attenuatus* Doug. ex Lindl. species complex ...

# Posterior distribution of allele frequencies

$$P(p_l, g_{li}|R_{li}^b, \epsilon) \propto \prod_l \prod_i P(R_{li}^b|g_{li}, \epsilon)P(g_{li}|p_l)P(p_l). \tag{1}$$

# Notation

| Symbol | Description |
| --- | --- |
| $L$ | The number of loci. |
| $l$ | Index for loci ($l \in \{1, \ldots, L\}$). |
| $N_k$ | The number of individuals sampled from population k. |
| $k$ | Index for populations ($k \in \{P_1, P_{poly}, P_2, O\}$). |
| $i$ | Index for individuals in a population $k$ ($i \in \{1, \ldots, N_k\}$). |
| $N_{lk}$ | The number of individuals sampled at locus $l$ in population $k$. |
| $N_{lk}^a$ | The number of individuals homozygous for A at locus $l$ in population $k$. |
| $N_{lk}^b$ | The number of individuals homozygous for B at locus $l$ in population $k$. |
| $N_{lk}^{ab}$ | The number of heterozygous individuals at locus $l$ in population $k$. |
| $\hat{p}_{lk}$ | Frequency of the derived allele (B) at locus $l$ in population $k$. |
| $P_k$ | The ploidy of individuals in population $k$. |
| $R_{li}$ | The number of reads for individual $i$ at locus $l$. |
| $R_{li}^a$ | The number of reads with allele A for individual $i$ at locus $l$. |
| $R_{li}^b$ | The number of reads with allele B for individual $i$ at locus $l$. |
| $r_{li}^a$ | Proportion of reads with allele A ($R_{li}^a/R_{li}$). |