Developing models for genotype uncertainty, inbreeding, and allelic inheritance in non-model polyploids

Paul Blischak[1] :: Laura Kubatko[1,2] :: Andrea Wolfe[1]

Ohio State University

[1]Department of EEOB

[2]Department of Statistics

A brief history

Early theory

The usual suspects:

- **Fisher** – Models for polysomic inheritance of trimorphic style w/ double reduction in *Lythrum salicaria*.
- **Haldane** – Models for determining gamete frequencies w/ partial selfing.
- **Wright** – Models for the distribution of allele frequencies subject to selection, migration and inbreeding.

Empirical motivation

There was a particularly fruitful synergism between the theoretical and empirical work being done for polyploids during the Modern Synthesis.

Polyploid pop-gen

Challenges

The difficulties of making population genetic inferences in polyploids present themselves at two broad levels:

1. **Allelic Dosage Uncertainty**: the inability to fully resolve the genotype of a partially heterozygous polyploid individual for a codominant marker (microsatellite, SNP).

2. **Allelic inheritance**: disomic vs. polysomic vs. heterosomic, selfing, clonal, double reduction, etc.

Allelic Dosage Uncertainty (ADU)

...the inability to fully resolve the genotype of a partially heterozygous polyploid individual for a codominant marker.

- **Partial heterozygote** – when the number of observed alleles at a locus is less than the ploidy level and not equal to 1 (i.e., is homozygous).
- **Ex**: For a locus with observed alleles A, B and C in a tetraploid (4N), the possible genotypes are AABC, ABBC or ABCC.
- The higher the ploidy, the worse the problem.
- Biallelic SNPs will always be partially heterozygous for a polyploid.

Overcoming ADU

For a biallelic locus, such as a SNP, we can use techniques like RAD sequencing to "sample" from the genotype (model with binomial probability distribution*).

Overcoming ADU

For a biallelic locus, such as a SNP, we can use techniques like RAD sequencing to "sample" from the genotype (model with binomial probability distribution*).

$$P(10 \text{ red reads} | k \text{ red alleles}) = \binom{16}{10} \left(\frac{k}{4}\right)^{10} \left(1 - \frac{k}{4}\right)^{6}$$

Overcoming ADU

For a biallelic locus, such as a SNP, we can use techniques like RAD sequencing to "sample" from the genotype (model with binomial probability distribution*).

Let's translate the entire illustration into something more mathematical.

A hierarchical model for polyploids

p_ℓ – allele frequency at locus ℓ.
$g_{i\ell}$ – genotype for individual i at locus ℓ.
$r_{i\ell}$ – sequencing reads for individual i at locus ℓ.

Read count likelihood

Binomial likelihood with error:

$$r_{i\ell} | t_{i\ell}, g_{i\ell}, \epsilon \sim \text{binomial}(p = \mathcal{G}_\epsilon(g_{i\ell}), n = t_{i\ell}),$$

Read count likelihood

Binomial likelihood with error:

$$r_{i\ell}|t_{i\ell}, g_{i\ell}, \epsilon \sim \text{binomial}(p = \mathcal{G}_\epsilon(g_{i\ell}), n = t_{i\ell}),$$

$$P(r_{i\ell}|t_{i\ell}, g_{i\ell}, \epsilon) = \binom{t_{i\ell}}{r_{i\ell}} \mathcal{G}_\epsilon(g_{i\ell})^{r_{i\ell}} (1 - \mathcal{G}_\epsilon(g_{i\ell}))^{(t_{i\ell} - r_{i\ell})}. \quad (1)$$

Read count likelihood

Binomial likelihood with error:

$$r_{i\ell}|t_{i\ell}, g_{i\ell}, \epsilon \sim \text{binomial}(p = \mathcal{G}_\epsilon(g_{i\ell}), n = t_{i\ell}),$$

$$P(r_{i\ell}|t_{i\ell}, g_{i\ell}, \epsilon) = \binom{t_{i\ell}}{r_{i\ell}} \mathcal{G}_\epsilon(g_{i\ell})^{r_{i\ell}} (1 - \mathcal{G}_\epsilon(g_{i\ell}))^{(t_{i\ell} - r_{i\ell})}. \quad (1)$$

Genotype error correction:

$$\mathcal{G}_\epsilon(g_{i\ell}) = \frac{g_{i\ell}}{\psi}(1 - \epsilon) + \left(1 - \frac{g_{i\ell}}{\psi}\right)\epsilon\,. \quad (2)$$

Genotype likelihood

Binomially distributed:

$$g_{i\ell}|p_\ell \sim \text{binomial}(p = p_\ell, n = \psi),$$

Genotype likelihood

Binomially distributed:

$$g_{i\ell}|p_\ell \sim \text{binomial}(p = p_\ell, n = \psi),$$

$$P(g_{i\ell}|p_\ell) = \binom{\psi}{g_{i\ell}} p_\ell^{g_{i\ell}} (1 - p_\ell)^{\psi - g_{i\ell}}. \tag{3}$$

A hierarchical model for polyploids

$$P(\mathbf{p}|R) \propto \sum_{g \in G} P(R|G)P(G|\mathbf{p})P(\mathbf{p}) \tag{4}$$

Simulations
Setup

We used the following settings for the simulations to test our model:

- Tetraploids (4N) and hexaploids (6N).
- Allele frequencies: 0.01, 0.05, 0.1, 0.2, 0.4.
- Sequencing coverage (average # of reads per individual per locus): 5x, 10x, 20x, 50x, 100x.
- Number of individuals sampled: 5, 10, 20, 30.

Results
Heatmaps

x-axis: # of individuals, **y-axis**: coverage, **scale**: error (s.d.)

Results
Heatmaps

x-axis: # of individuals, **y-axis**: coverage, **scale**: error (s.d.)

$$
\begin{aligned}
P(R|T,G,\epsilon) &\sim \text{binomial} \\
P(G|p,\phi) &\sim \text{beta-binomial} \\
P(p) &\sim \text{uniform}[0,1] \\
P(\phi) &\sim \text{uniform}(0,1000]
\end{aligned}
$$

Conclusions

- TAKE HOME: **Don't have to use genotypes as the first line of data**. Using sequencing reads is a viable solution for dealing with ADU.
- **The framework presented here is highly extensible**. Future work includes generalizing to both auto- and allopolyploids, more complex patterns of inheritance.
- **Sampling more individuals appears to be most important**. More individuals over higher sequencing coverage.
- **Need empirical data**. Simulations are nice, but we need to see how a model such as this works for lab-collected data.

Code availability

- **polyfreqs**: an R package for the estimation of allele frequencies in autopolyploids. Available on GitHub – `https://github.com/pblischak/polyfreqs`.
- Manuscript is currently in review, preprint is on bioRχiv – `http://biorxiv.org/content/early/2015/07/02/021907`.
- Data and code for the simulation study and making the figures are on GitHub – `https://github.com/pblischak/polyfreqs-ms-data`.
- Presentation slides are on fig**share**, and the LaTeX source code is also on GitHub – `https://github.com/pblischak/botany2015`.

All these links are in the GitHub repository for this presentation: **pblischak/botany2015**.

#openscience

Acknowledgments

- Aaron Wenzel, and members of the Wolfe and Kubatko labs.
- Ohio Supercomputer Center (Simulations).
- American Society of Plant Taxonomists (Travel Award).
- National Science Foundation (Funding).

Thanks!

Questions?