Uncovering Identity By Descent For Varietal Protection in Synthetic Populations

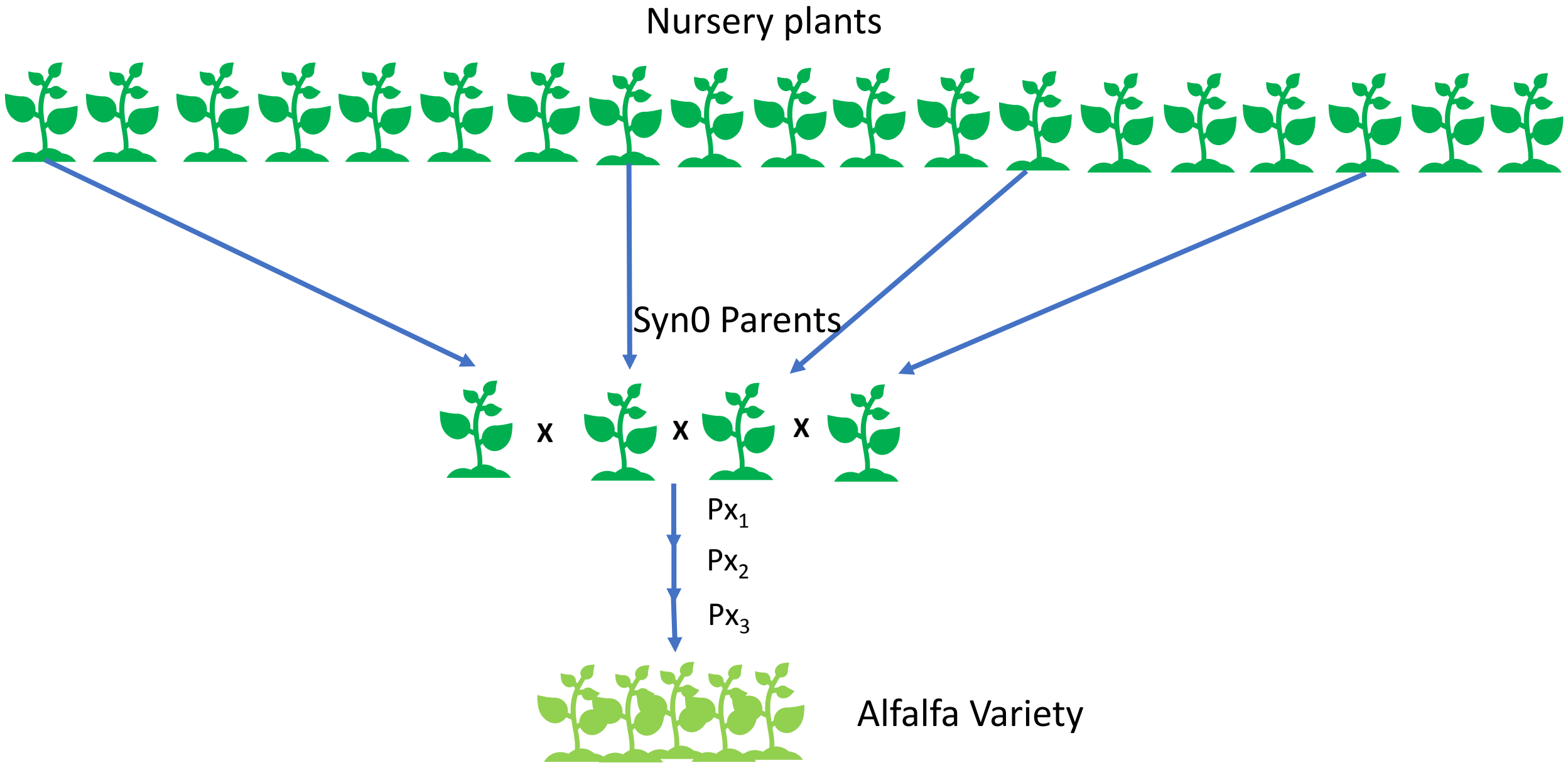
John Cameron

Forage Genetics International

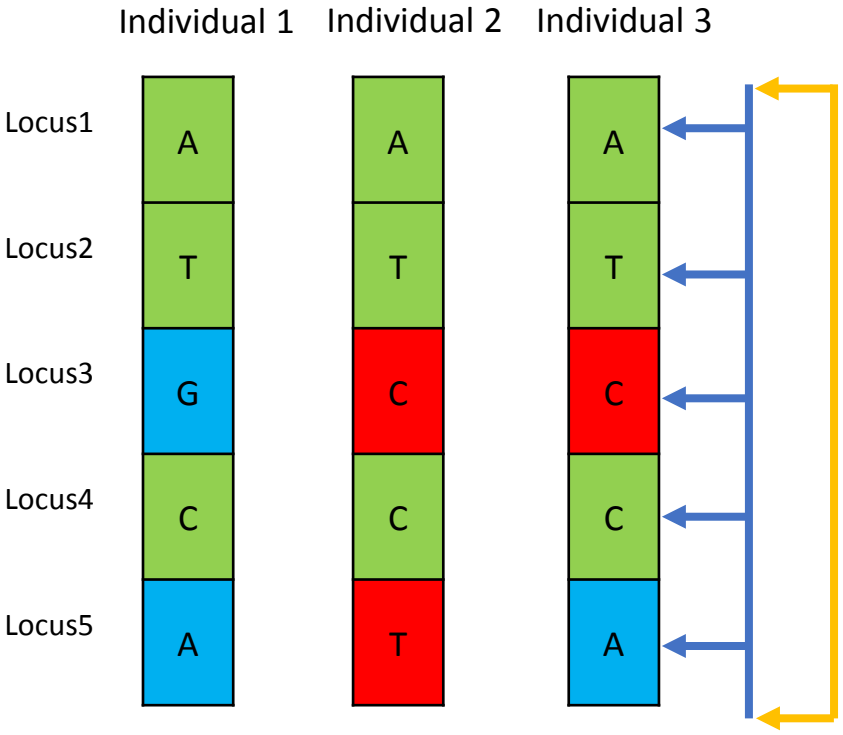
Main concepts

- Haplotypes contain more information about Identity by Descent (IDB) than SNPs due to capturing linkage structures
- Synthetic populations, particularly autotetraploids like alfalfa, cannot be easily “fingerprinted” via phasing entire linkage groups due to nature of synthetic variety breeding
- Microhaplotype markers can be developed using high-throughput targeted-sequencing of small genomic regions (i.e. ≤ 200bp), where more than one known SNP present
- Bulk sampling of variety DNA is a cost-effective way to capture allele frequencies of a variety
- Statistical models can be built using the targeted-sequencing data of synthetic varieties to make inferences about the likelihood of sharing of common progenitors

Alfalfa variety development



IDB: SNPs vs. Microhaplotypes



	Ind 1	Ind 2	Ind 3
Ind 1	1	3/5	4/5
Ind 2	3/5	1	4/5
Ind 3	4/5	4/5	1

	Ind 1	Ind 2	Ind 3
Ind 1	1	0	0
Ind 2	0	1	0
Ind 3	0	0	1

Population allele frequencies

SNPs

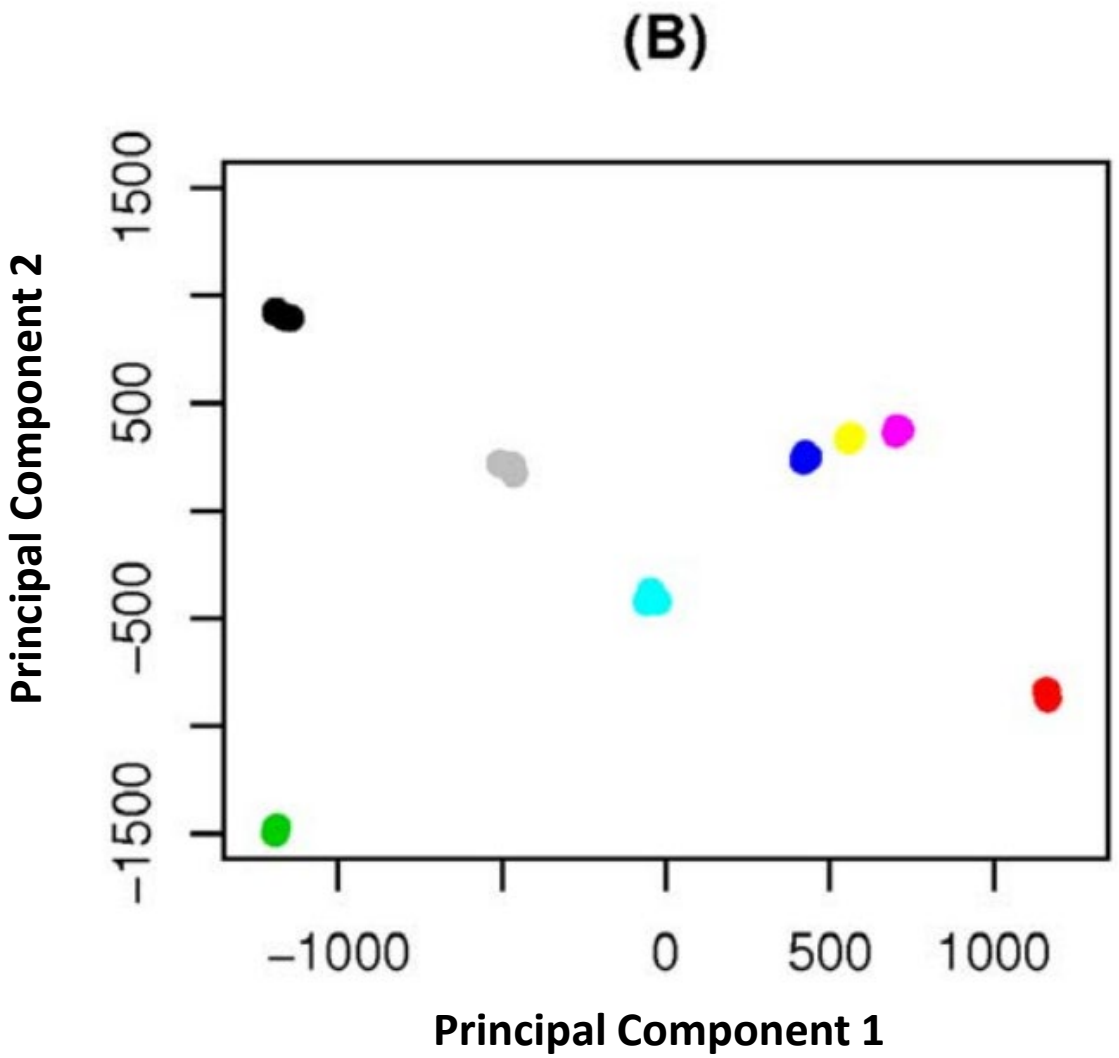
Locus1	1	0
Locus2	1	0
Locus3	0.33	0.66
Locus4	1	0
Locus5	0.66	0.33

Microhaplotypes

	M1	M2	M3
Locus1	0.33	0.33	0.33

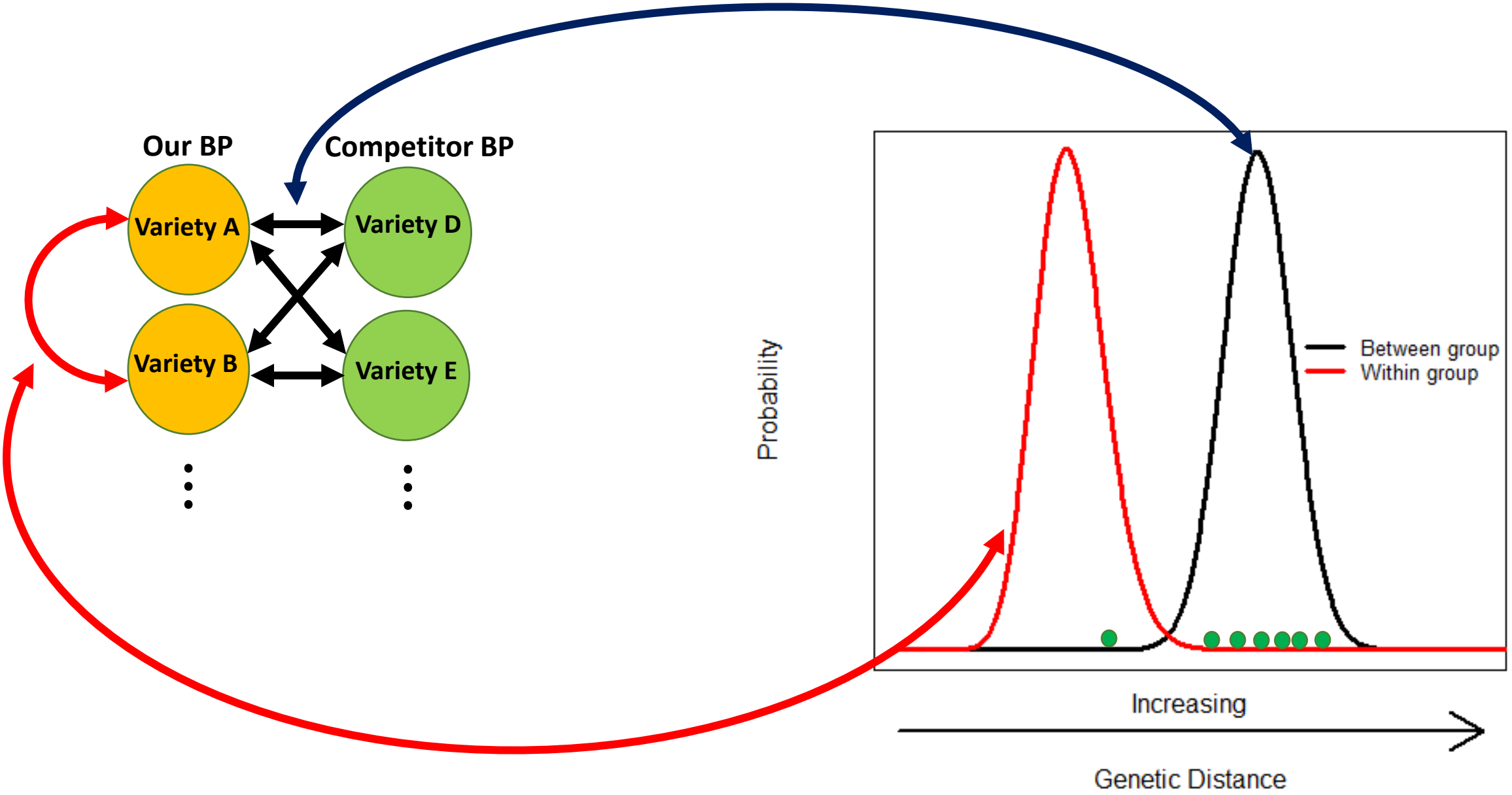
Allele frequency fingerprint

- Rye grass, Bulk harvest, bulk sequence in replicate
- Calculate SNP allele frequencies
  For each replicate
- Calculate Principal Components
- Plot the replicates of varieties in the first 2 or 3 Principal Component (PC) space
- Relative genetic similarity between varieties as (Euclidean) distance

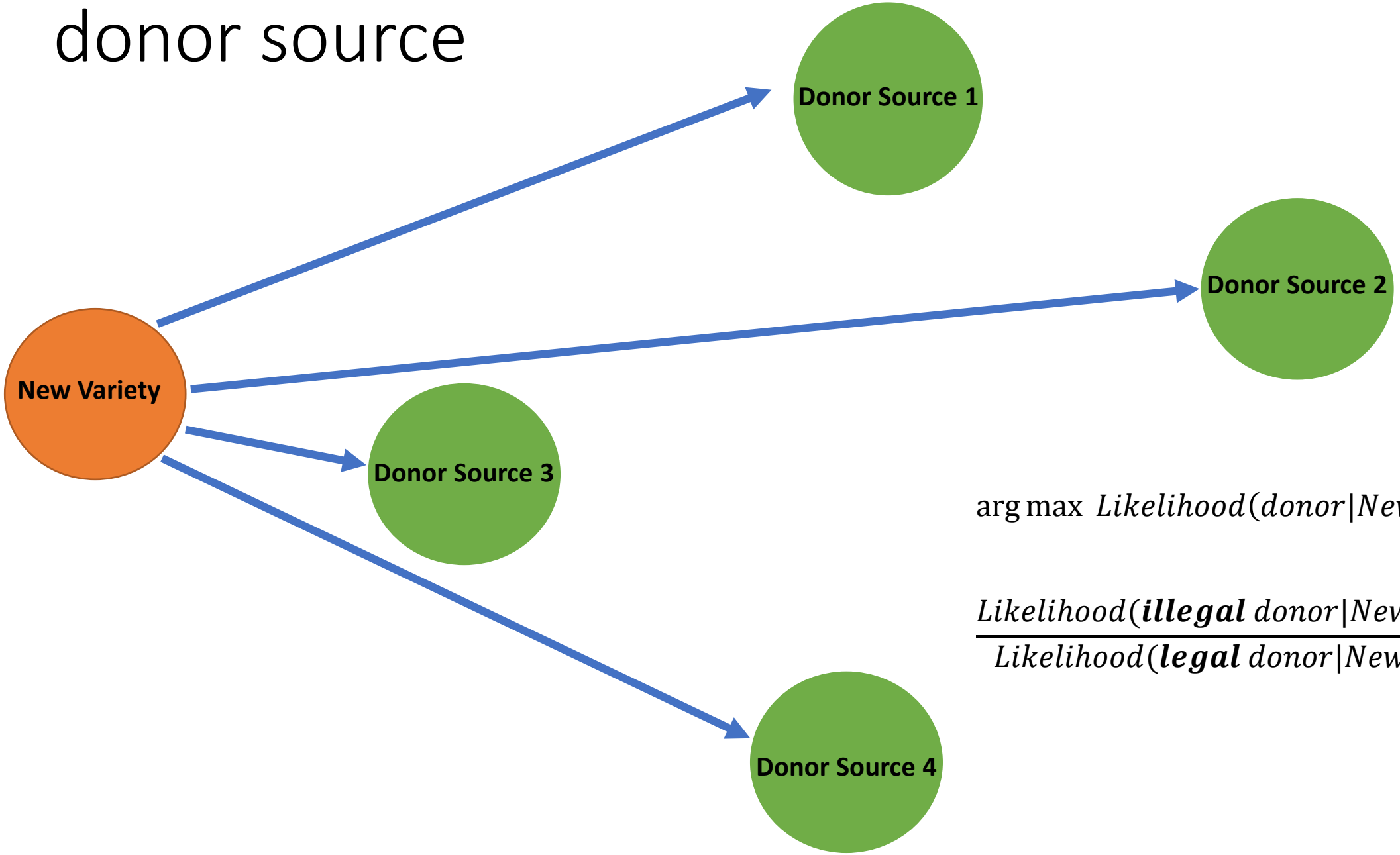


Byrne et al. 2013

Test of relatedness: Parametric



Test of relatedness 2: Maximum likelihood of donor source



$$\arg\max \; Likelihood(donor|New\ Variety)$$

$$\frac{Likelihood(\boldsymbol{illegal}\ donor|New\ Variety)}{Likelihood(\boldsymbol{legal}\ donor|New\ Variey)}$$

Summary

1. Multiallelic haplotype markers contain more information about ancestry (identity by descent) than biallelic SNPs
2. Targeted sequencing can be used to capture adjacent SNPs collectively as microhaplotypes in short genetic intervals (i.e. ≤ 200bp)
    1. Haplotypes can be assembled with SNPs data in autotetraploids, but requires biparental populations, genotyping of individual progenies, and intensive computational methods
    2. Targeted sequencing also enables much more balanced data sets than GBS
3. Targeted sequencing of **bulked DNA** from synthetic varieties allows for cost effective capture of microhaplotype frequencies
4. Logical statistical frameworks can be built using the information obtained from variety sequencing to make inferences about likelihood of IP infraction

Literature

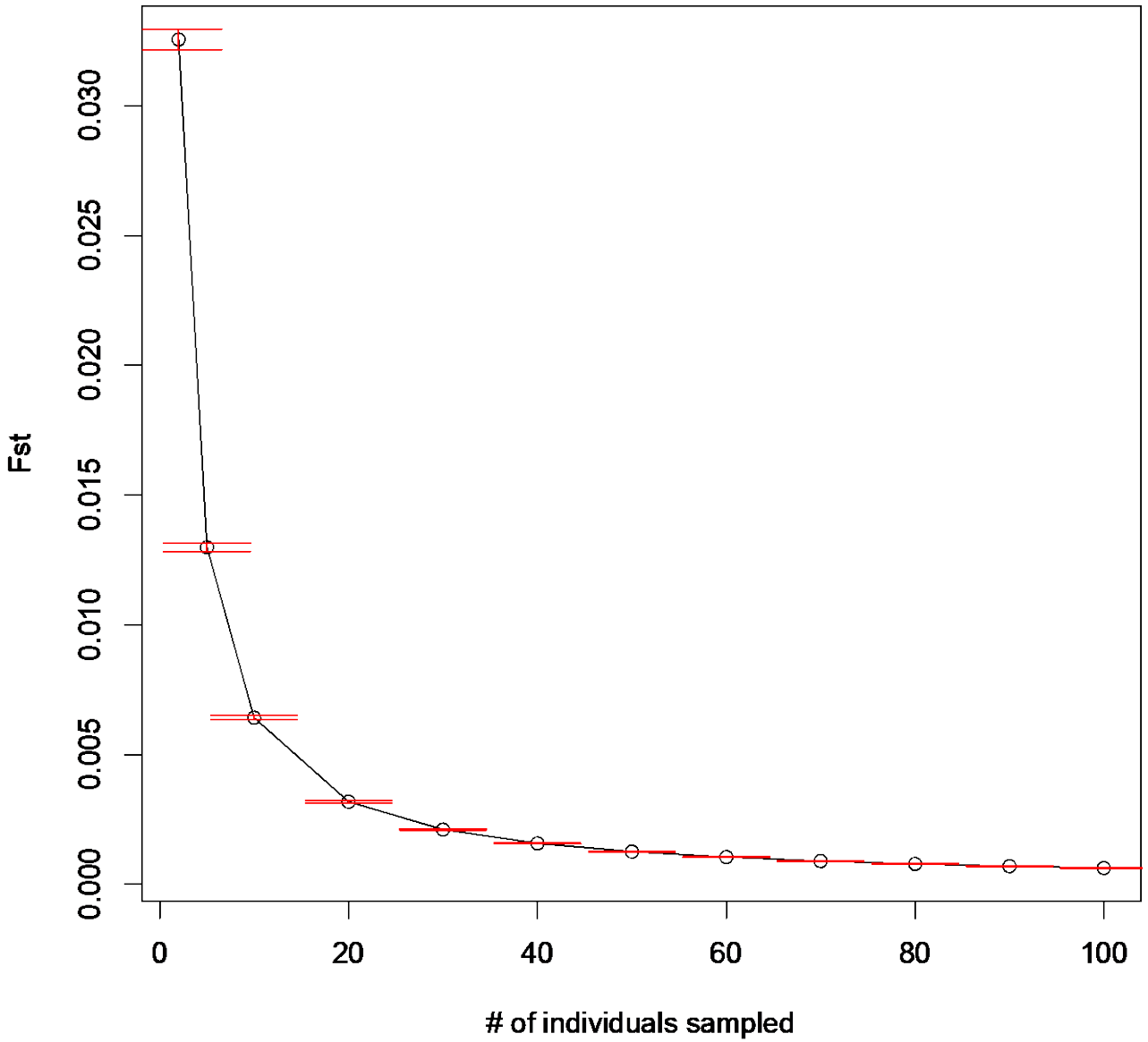
- Byrne, Stephen, et al. "Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing." *PLoS One* 8.3 (2013): e57438.
- Van Geest, Geert, et al. "Micro-haplotyping in polyploids using massively parallel amplicon sequencing." (2020).Forensics papers
- Gattepaille, Lucie M., and Mattias Jakobsson. "Combining markers into haplotypes can improve population structure inference." *Genetics* 190.1 (2012): 159-174.

Thank you

Forage Genetics International

Variety Sampling Strategy

- Accurate estimates of variety-specific allele frequencies, requires many individuals
- Sequencing individuals from each synthetic variety is cost prohibitive and unnecessary
- Sequencing bulks in replicate allows for
    a) Cost-effective capturing of allele frequencies
    b) Estimation of error: sampling, sequencing and handling
- Simulation can be used to determine numbers of individuals to sample in bulk



Final Thoughts

- 2 types of data that can differentiate varieties
  - Unique microhaplotype markers in each variety
  - Microhaplotype frequency makeup of each individual variety
- Sequencing individual plants for all varieties is cost infeasible
  - 96 plant samples * $39/sample * 200 varieties = $748,800
  - 96*$39=$3744 variety
- Sequencing many plants from a single variety as *one* sample will allow us to estimate the allele frequency AND determine unique genetic markers present only in particular varieties,
  - 5 bulk sampled DNA replicates * $39/sample =$200

Microhaplotypes linked to RR gene

- Round Up Ready gene is located @ 12,800,000 bp

Chr	First and last SNP positions in MH marker (bp)	Genotype	freq(RR)-freq(NRR)
Chr6	12999320 - 12999405	GCT	0.90
Chr6	8161619 - 8161738	TGAAATA	0.81
Chr6	18830227 - 18830240	TGG	0.78
Chr6	12082640 - 12082688	GGAC	0.76
Chr6	30433708 - 30433746	TCGG	0.72
Chr6	116479 - 116560	AAAAT	0.72
Chr6	11838413 - 11838433	CT	0.71
Chr6	22522810 - 22522854	GCTA	0.70
Chr6	6143923 - 6143977	GG	0.70
Chr6	22528736 - 22528830	TGCGGTG	0.68

Germplasm Security pipeline

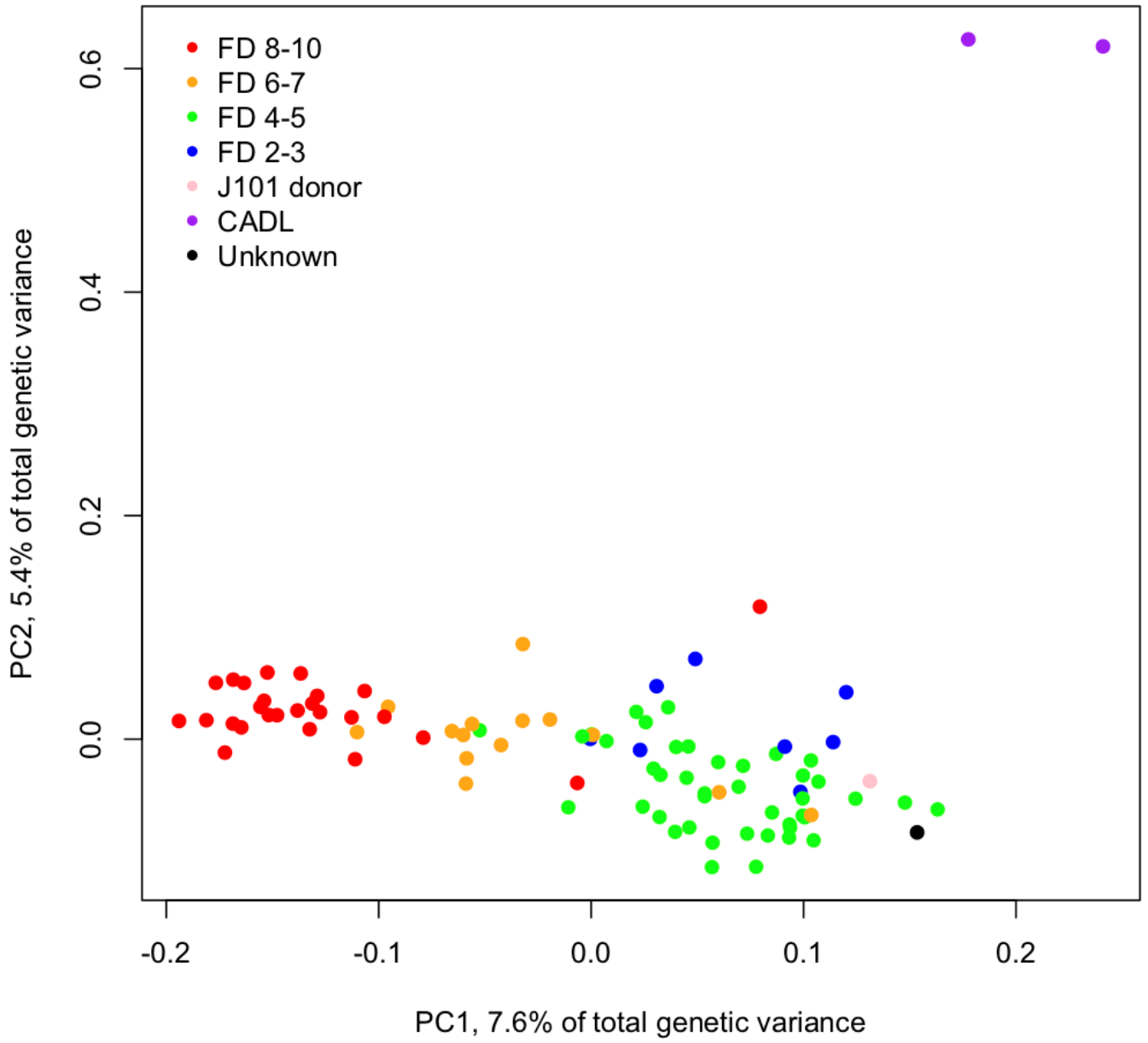
1. Sequence all FGI RR varieties in bulk
2. Determine sets of unique or rare microhaplotypes present in each variety
3. Determine haplotype frequency in region linked to RR gene

4. Sequence all new competitor RR varieties in bulk
5. Determine which FGI variety is most genetically similar to new competitor varieties

6. If a patented variety is most related to new competitor variety, prosecute *or* sequence first *then* prosecute

Germplasm Security: Genetic clustering by dormancy class

Principle Component Analysis



Analytical Methodology: RR example

1. Sample probable allele frequencies from Dirichlet distributions for each variety
2. Calculate genomic contribution of each candidate to target
   - Regression problem with constraints, can be solved via non-linear programming

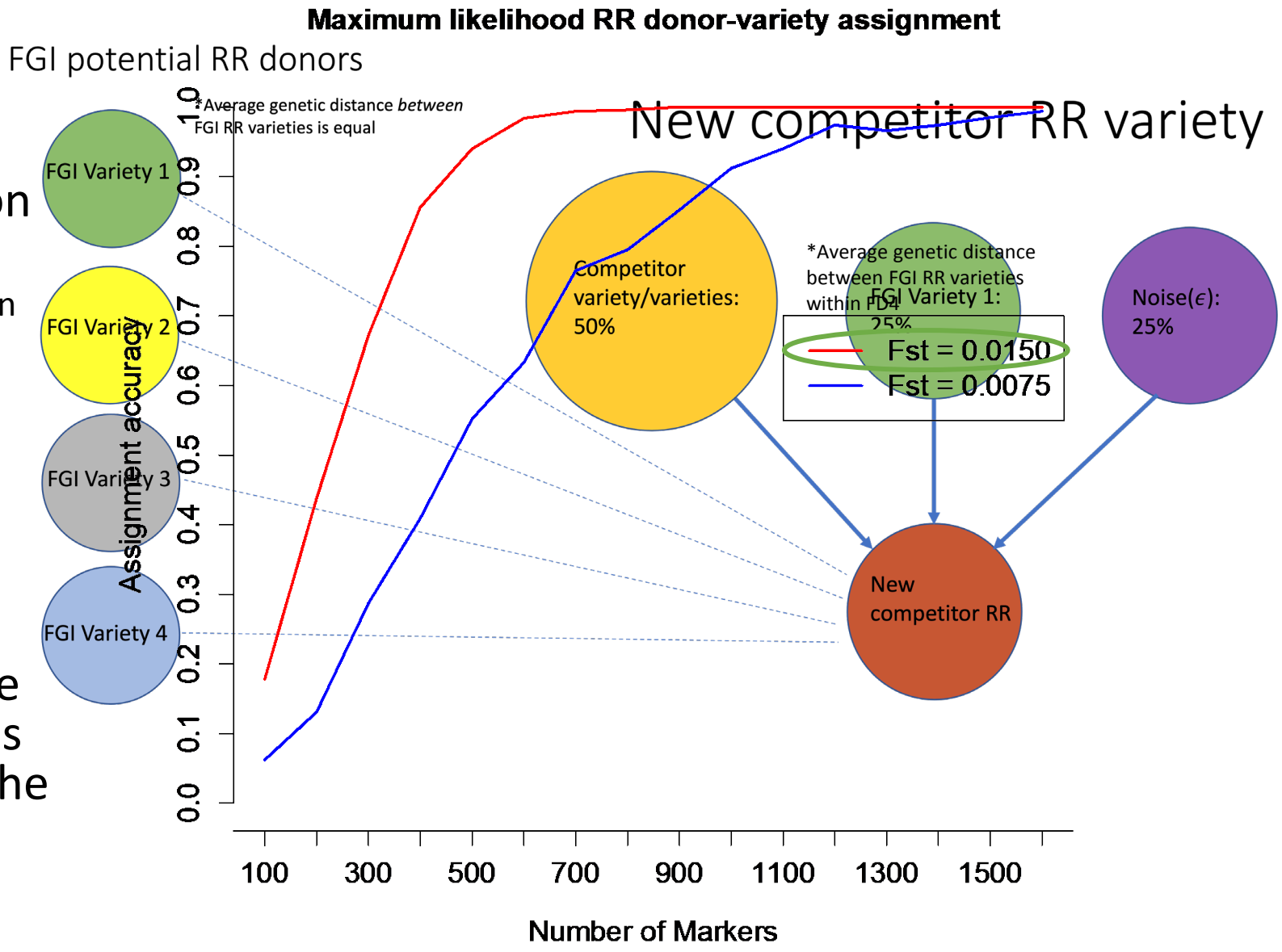
$$\arg\min f(\mathbf{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X} \quad (1)$$

s.t

$$b_i \geq 0\{i \in 1, \ldots, N\} \quad (2)$$

$$\sum_{i}^{N} b_i = 1 \quad (3)$$

3. Repeat many times to estimate the distributions of relatedness for each candidate variety to the target variety



Germplasm Security: Experiment 1

- Development of multiallelic marker panel
  - 96 very diverse varieties, competitor and FGI, 1 plant from each
  - Built bioinformatics pipeline to call microhaplotype markers from targeted sequencing reads (contain more information about ancestry than SNPs)
  - Validated statistical power of ancestry determination from previous simulation experiment, using actual genotypic data
  - Found unique haplotypes in many individual plants
    - Need to sequence many plants from each variety to determine if truly unique

Tests of relatedness: Parametric cont'd

1. Estimate genetic distance between the new variety, and all varieties within the target breeding program
2. Determine probability of a pair of varieties belonging to each distribution
3. Calculate ratio of probabilities of belonging to both distributions