Methods of Parentage Analysis in Natural Populations

Using molecular markers, estimates of genetic maternity or paternity can be achieved by excluding as parents all adults whose genotypes are incompatible with offspring under consideration

Parentage

Important tool for:

•Breeding systems – Difference between observed and actual, mate choice

•Estimating reproductive success of males and females and the traits of successful males and females (e.g., body size, social status)

•Managers
  –Breeding assessments for captive individuals

• Identification of individuals dispersing into the population

•Population estimates – Mark recapture estimates using genetic signatures of known family groups.

•Identifying dead-beat dads (a human issue)

Methods

- Exclusion
  - Earliest and conceptually simplest technique
- Categorical and fractional likelihood
  - Complete exclusion is not possible
  - Assigns progeny to non-excluded parents based on likelihood scores derived from their genotypes
- Genotype Reconstruction
  - Uses multilocus genotypes of parents and offspring to reconstruct the genotypes of unknown parents contributing gametes to a progeny array for which one parent is known a priori

Exclusion

•Based on Mendelian rules of inheritance

•Uses incompatibilities between parents and offspring to reject particular parent offspring hypotheses.

Female Genotype: A/A
Offspring Genotype: A/B

Excluded: A/C
Not Excluded: A/B

*Powerful when there are few candidate parents and highly polymorphic genetic markers available.

•Impractical if the pool of candidate parents becomes large

-Due to the large number of loci needed to yield a single non-excluded parent.

•Many exclusion programs can allow the user to specify the number of mismatches necessary for an exclusion to be considered valid, making the method more robust to the difficulties imposed by mutations or scoring errors.

Categorical Allocation

- Categorical allocation uses likelihood-based approaches to select the most likely parent from a pool of non-excluded parents.
- This method involves calculating a logarithm of the likelihood ratio (LOD score) by:
  - Determining the likelihood of an individual (or pair of individuals) being the parent (or parents) of a given offspring divided by the likelihood of these individuals being unrelated.
  - Offspring are assigned to the parent (or parental pair) with the highest LOD score.
- LOD = 0 or negative - offspring are unassigned.
- Contrary to strict exclusion methods, likelihood-based allocation methods usually allow for some degree of transmission errors due to genotype misreading or mutation.

Fractional allocation

- The fractional allocation method assigns some fraction, between 0 and 1, of each offspring to all non-excluded candidate parents.
- The portion of an offspring allocated to a particular candidate parent is proportional to its likelihood of parenting the offspring compared to all other non-excluded candidate parents.
- Single parent and parent pair likelihoods are calculated in the same way as in the categorical allocation method.
- Assumes genotypes are known from all parents in the population and that one parent is known for the offspring under consideration.
- The fraction of offspring (O=k) awarded to a candidate male j (MP=j) conditional on female I (FP=i) is denoted by F_{ij}:

$$\hat{F}_{ij} = \sum_{k} X_{ik} P(MP = j \mid FP = i, O = k)$$

LOD=ln(L_1/L_2)
Natural log of the ratio of 2 likelihoods
ln(1)=0

Male	LOD	F_{ij}
1	0.5	0.8
2	0.4	0.2

Parental Reconstruction

- Uses the multilocus genotypes of parents and offspring to reconstruct the genotypes of unknown parents contributing gametes to a progeny array for which one parent is known a priori.
- Existing techniques reconstruct the minimum number of parental genotypes necessary to explain the data set.
- For the case in which the mother is known, all possible paternal genotypes consistent with at least one progeny in the data set are tested in combination to determine which minimum set of paternal genotypes can explain the entire progeny array.
- Extremely computationally intensive using algorithms
  - Especially for progeny arrays with more than six fathers.

Genotypes in offspring array

(A) / A (A) / C

(C) / C (C) / D

Male alleles (unknown)

Female alleles (known)

	A	B	C	D	E
A*	X			X	
B					
C*	X		X	X	
D					
E					

Prior to analysis

- Collection of data for parentage analysis is just as important as the management of the compiled data set.

- Ideal situation - large groups of offspring are collected from known mated pairs of adults
  - Molecular techniques needed only to verify the truth.

- Estimating parentage is still accurate if offspring can be collected in family groups with their mothers and a complete sample of males from the population is obtained.

- As we lose sample size the likelihood that the missing samples contain the true parental genotype increases, along with our ability to correctly assign offspring.

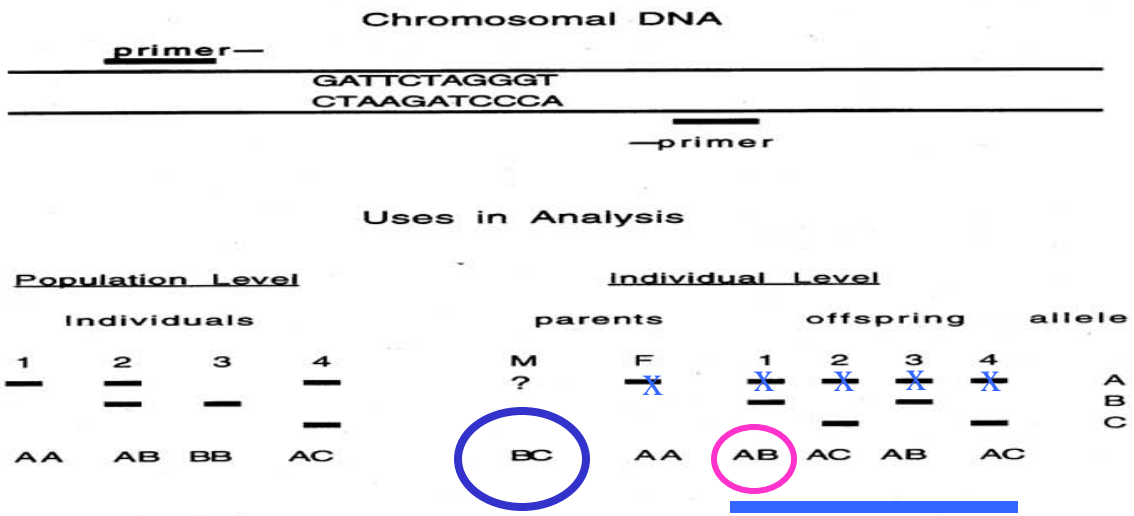
- Jones and Arden (2003) emphasized the importance of knowing the constraints of your particular study. The proportion of adults that can be sampled, the techniques and markers to be used, and how the analysis will proceed is critical in the design of an experimentally or hypothesis driven research design.

Markers?

- Statistical power is increased as a function of

  (a) the number of loci used

  (b) allelic diversity and heterozygosity
- Analyses assume Hardy-Weinberg (i.e., you can infer population genotype frequencies and the expectation of observing a genotype at random in the population from the frequency of alleles)
- Microsatellites are the most powerful for biological systems.
- # of loci used will depend upon exclusion probabilities.
- Analytical techniques can apply to any dominant marker
  - Amplified Fragment Length Polymorphisms (AFLP's)
  - Uni-parentally inherited cytoplasmic (e.g., mitochondrial) markers.



No BC male in pool of candidates

Are offspring data more consistent with 1 or >1 male parents?

Qualitative Inferences

assume female F is the true mother

If offspring array were produced by a single male we may predict the male's genotype (BC).

If males 1-4 are the only potential males in the population and we assume a single male parent, this is evidence of gene flow

If we only had samples of one offspring (e.g. 1) then our ability to infer the male(s) genotype is less likely

Given the progeny array there is some probability of multiple paternity

Genetic Determination of Parentage

Situation – an offspring or progeny array of unknown parentage is found and we wish to assign parents

- one parent (usually the female) is known
- neither parent is known

Analyses become more computationally difficult when fewer data are available

- assignment of one vs two parents
- numbers of progeny in progeny array
- number of loci available (high or low)
- characteristics of loci (allelic diversity and distribution of allele frequencies)
  with k alleles there are $(k^2+k)/2$ possible genotypes
- background information and setting in which offspring and parents are placed

Analyses become more computationally difficult depending on whether analyses are based on *exclusions* or in cases where all putative parents (or parental pairs) can't be excluded, when parentage is *assigned on the basis of probabilities*

Uses of paternity analysis

1. In absence of observational data on movements, analyses provide a measure of the distances males (or their gametes) moved.
2. Actual rates of selfing and outcrossing can be obtained (plants) as can relative to inbred vs outbred matings
3. Genetic relatedness of progeny from a singe female can be determined – proportion of progeny that are full ½ sibs
4. Number of paternal individuals which have fertilized a single female can be determined
5. Paternity studies provide insights into sperm competition and sperm storage
6. Relative reproductive contributions of males as a function of phenotype or other ecological correlates can be determined
7. Differential survival and fertility of offspring from specific events can be followed
8. In absence of known pedigrees, analysis aid design of breeding programs

Evaluation of Statistical Power

- Probability of exclusion
  - Exclusion= mis-match [putative parent not possible given genotypes of offspring and parent(s)]
  - Probability of exclusion and probability of assignment: probability of finding a specific genotype in the population
    - 1. number of loci assayed
    - 2. degree of polymorphism
    - 3. allele frequency distribution
    - 4. number of potential parents
    - 5. number of progeny

- Inclusion
  - Inclusion= no mis-matches
  - However, programs account for the possibility of mutation or error and for the fact that you have not sampled all possible parents

Estimating the likelihood of multiple paternity or maternity

Based on exclusions – for example, in a clutch the most alleles you can have at a single Mendelian locus is 4 (e.g., both parents heterozygous for different alleles).

Based on probabilities of concurrent paternity – even when multiple paternity is not observed on the basis of presence of “foreign” alleles, there is often a non-zero probability that the genotypes of the progeny array are consistent with multiple parentage. This should be tested against the probability of single parentage.

Statistical power increases as a function of the number of loci, allelic diversity, and number of offspring in the clutch

Descriptive Statistics– Heterozygosity

- Heterozygosity
  - Expected heterozygosity

    $H_e = \frac{n}{n-1}(1 - \Sigma p_i^2)$

    Where n is the number of individuals used to determine the allele frequencies and p_i is the frequency p of the ith allele

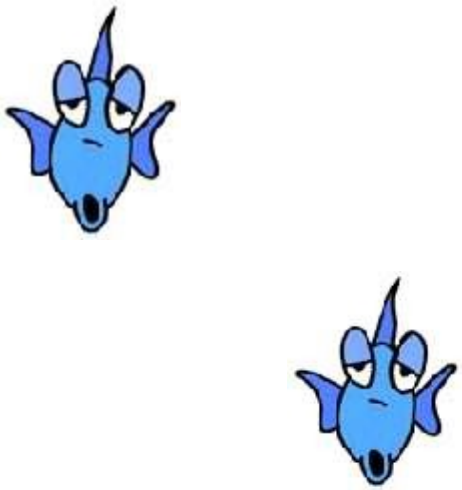
  - Observed heterozygosity

    $H_o = N_{AB}/N$

    Where AB represents a heterozygous genotype (i.e. A and B are alleles)

Descriptive Statistics

- Hardy-Weinberg Equilibrium



Genotype	**AA**	**AB**	**BB**
Genotypic Frequencies	p^2	2pq	q^2
Expected	Np^2	N2pq	Nq^2
Observed	N_{AA}	N_{AB}	N_{BB}

Estimating the likelihood of parentage in absence of exclusions (after Meagher and Thompson, 1986)

Consider an ordered triplet of genotypes (g_B, g_C, g_D) at a single autosomal locus for 3 individuals (B, C, D). We are interested in identifying triplets consisting of an offspring (B) and the maternal (C) and paternal (D) parents. The statistical properties of triplets under different relational situations are:

(UU) B, C, and D are unrelated and thus the triplet contains neither parent of B
(QU) C is the parent of B but D is unrelated and the triplet contains 1 parent
(QQ) C and D are the parents of B and thus the triplet contains both parents

Estimating the likelihood of paternity given non-exclusion (con't)

The probabilities of these triplets will be denoted as $P(g_B,g_C,g_D|R)$ where the relationship R is one of the 3 previous possibilities (UU, QU, QQ)

$P(g_B,g_C,g_D|UU) = P(g_B)*P(g_C)*P(g_D)$

$P(g_B,g_C,g_D|QU) = P(\text{offspring } g_B|\text{parent } g_D)*P(g_C)*P(g_D)$

or $T(g_B|g_D,\text{--})*P(g_C)*P(g_D)$

$P(g_B,g_C,g_D|QQ) = P(\text{offspring } g_B|\text{parents } g_C,g_D)*P(g_C)*P(g_D)$

or $T(g_B|g_C,g_D)*P(g_C)*P(g_D)$

Which relationship is more likely given the data [P(R|data)] – use LOD

$P(g_i)$ is the expected frequency of the i^{th} genotype (under Hardy-Weinberg) and T denotes the transmission probabilities from putative parents to offspring

Assignment Statistics– LOD Scores

- Likelihood
  - T(gB|gC,gD)*P(gC)*P(gD)
  - Where T is the probability of allele transmission of parents (C and D) to Offspring B; and gB, gC, and gD are the genotypes of offspring individual B and candidate parents C and D

- Likelihood ratio

$$\mathbf{L(H1,H2|D) = \frac{P(D|H1)}{P(D|H2)}}$$

  - Where H1 is the hypothesis that the candidate parental pair is the true parental pair and H2 is the hypothesis that another candidate parental pair is the true parental pair and D denotes the data in the form of offspring and parental genotypes

- LOD scores– (Logarithm of Odds) used in instances where there is more than one possible relationship in order to demonstrate which is more likely
  - $\text{LOD} = \log_e \frac{P(D|H1)}{P(D|H2)}$

A fish example using real data to show formula are applied

Here, population allele frequencies of the 'A' and 'a' alleles were estimated to be 0.3 and 0.7, respectively so expected frequencies of each genotype can be estimated.

C

Possible mother

D Possible father

AA

$p^2 = 0.09$

aa

$q^2 = 0.49$

?

B

Offspring whose parents we wish to find

Aa

2pq=0.42

Likelihood relationship QQ= T(gB|gC,gD)*P(gC)*P(gD)

= 1 * 1 * 0.09 * 0.49

= 0.0441

Likelihood relationship $P(g_B,g_C,g_D|UU) = P(g_B)*P(g_C)*P(g_D)$

= 0.42 * 0.09 * 0.49

=0.0185

LOD = $\log_e$ (0.0441/0.0185)

So…the probability of adults C and D being the parents are about 3 times more likely than 2 random adults from the population based just on this one locus.