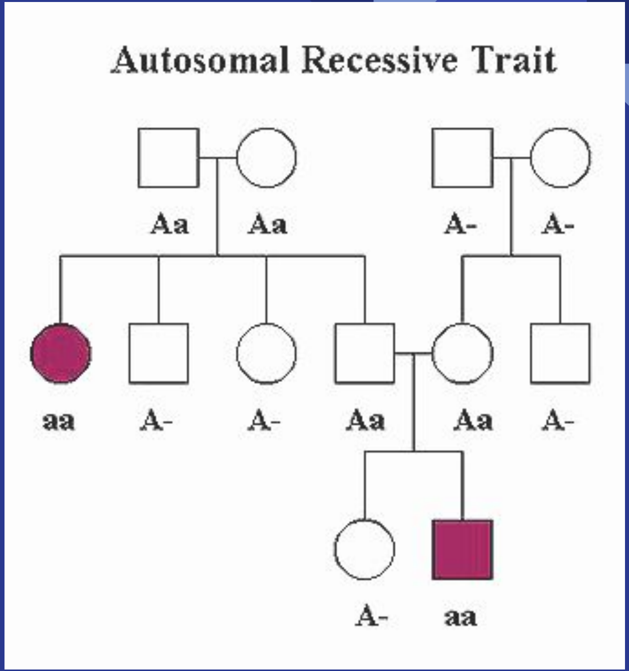
# Estimating Theoretical Allele Frequencies of Cystic Fibrosis Using an EM Algorithm

Rachel Ferina

# Genetics Background

- Genes: functional units of heredity
  - Made up of DNA
- Alleles: versions of a gene
  - Dominant--phenotypically expressed
  - Recessive--only phenotypically expressed when dominant alleles aren't present
  - 1 allele inherited from each parent



# Allele Frequencies

- Frequency of Allele A: $\frac{\text{Number of copies of allele A in population}}{\text{Total number of copies of gene in population}}$
- Change in allele frequencies over several generations indicates evolution in a population
- Applications in population genetics
  - Genetic diversity and gene pool richness
  - Genetic association with diseases, estimating number of individuals in a population susceptible to disease or drug resistance

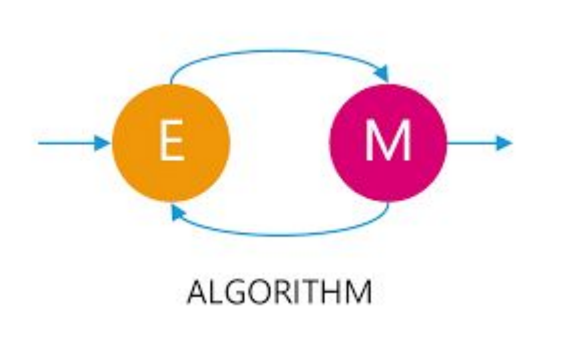
# Cystic Fibrosis

- Genetic disease resulting in excess production of thick mucus
  - Affects the lungs and digestive system
  - Often results in shorter lifespan
- Inherited recessively in CFTR gene

| | **A** | **a** |
|---|---|---|
| **A** | AA | Aa |
| **a** | Aa | aa |

# Expectation Maximization (EM) Algorithm

- Useful for calculations with incomplete data
- E Step: Expectation
  - Compute expected genotype based on observed phenotype
- M Step: Maximization
  - Determines maximum for parameters
- Iterate until convergence of likelihood
  - Slow convergence
- Use it to find maximum likelihood estimate (MLE)



# Estimating Theoretical Allele Frequencies of Cystic Fibrosis

- Random sample of 200 subjects
- 12 subjects diagnosed with Cystic Fibrosis
- Goal: Estimate population allele frequencies for A and a --$p_A$ and $p_a$
    - Missing data: genotypes

| Cystic Fibrosis | Unaffected |
|---|---|
| aa | AA<br>Aa |

- $t_0$ = aa, $t_1$ = Aa, $t_2$ = AA

# Log Likelihood

- Use Hardy-Weinberg equilibrium allele frequencies
  - $p^2 + 2pq + q^2 = 1$
- Complete Log Likelihood = $n_{AA}\log(p^2) + n_{Aa}\log(2pq) + n_{aa}\log(q^2)$
- Incomplete Log Likelihood = $n_A\log(p^2+2pq) + n_a\log(q^2)$
- While loop to carry out EM algorithm with incomplete log likelihood
  - Current likelihood - previous likelihood > 0.0001

# E step

- Calculate current allele frequency estimate with function 2q/(1+q)
    - Probability based on current q estimate
        - Probability of Aa given AA or Aa
- Split up unaffected group into carriers ($t_1$ = Aa) and unaffected homozygous ($t_2$ = AA); example calculation for first iteration
    - Multiply 188 by function 2q/(1+q)
        - $t_1 = 21.283$
    - 188 - previous answer
        - $t_2 = 166.717$

# M step

- Estimates new q with function $(2t_0 + t_1)/2n$
  - Conditional on current allele frequency estimates
  - Updates $p_a$ via gene counting
- Alternate with E step until convergence of likelihood is reached

# EM Algorithm Results

| Iteration | $t_0$<br>(aa) | $t_1$<br>(Aa) | $t_2$<br>(AA) | New q $(2t_0 + t_1)/2n$<br>$(p_a)$ | 1 - q<br>$(p_A)$ | Log Likelihood |
|---|---|---|---|---|---|---|
| 1 | 12 | 21.2830 | 166.7170 | 0.3733 | 0.6267 | -51.8655 |
| 2 | 12 | 38.2373 | 149.7627 | 0.3155 | 0.6845 | -47.3999 |
| 3 | 12 | 50.6259 | 137.3740 | 0.2855 | 0.7145 | -46.6292 |
| | | | | … | | |
| 13 | 12 | 73.9795 | 114.0205 | 0.2452 | 0.7546 | -45.3936 |
| 14 | 12 | 73.9796 | 114.0204 | 0.2451 | 0.7548 | -45.3935 |
| 15 | 12 | 73.9796 | 114.0204 | 0.2450 | 0.7550 | -45.3935 |

# Cystic Fibrosis Allele Frequencies

- For a theoretical sample of 200 subjects with 12 diagnosed with Cystic Fibrosis:
  - $p_a = 0.245$
  - $p_A = 1 - p_a = 0.755$

| | **A** | **a** |
|---|---|---|
| **A** | AA | Aa |
| **a** | Aa | aa |

# Acknowledgements

Seth Temple--PhD student, Statistics

Elizabeth Thompson--*Statistical Inference from Genetic Data on Pedigrees*

Sharon Browning--Biostatistics Research Professor