

# Computer-Assisted Forensics for Mass Disasters

*Gene Myers*

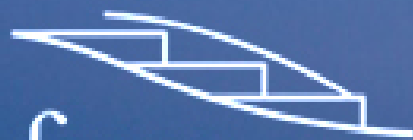
*HHMI Janelia Farm Research Campus*

*Washington, DC*

*[myersg@janelia.hhmi.org](mailto:myersg@janelia.hhmi.org)*

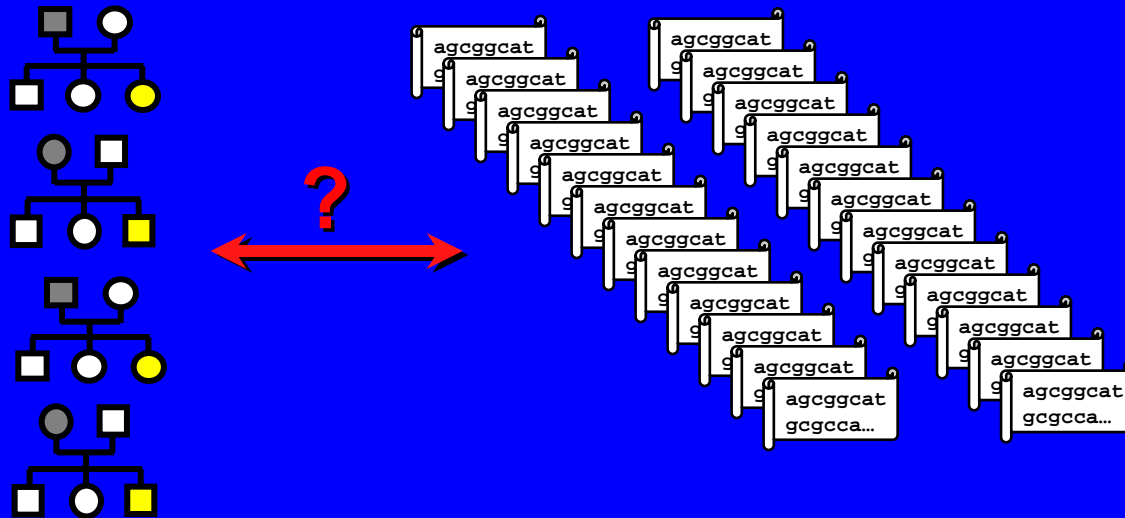
HHMI

janelia farm  
research campus



# The problem

- Given genetic fingerprints of F family pedigrees for alleged victims and genetic fingerprints of S samples found at a disaster site:
  - Who can you confirm died at the site? (legal)
  - Who died at the site that is outside the alleged set? (law enforcement)
  - Cluster the remains for burial. (closure)



# Complications

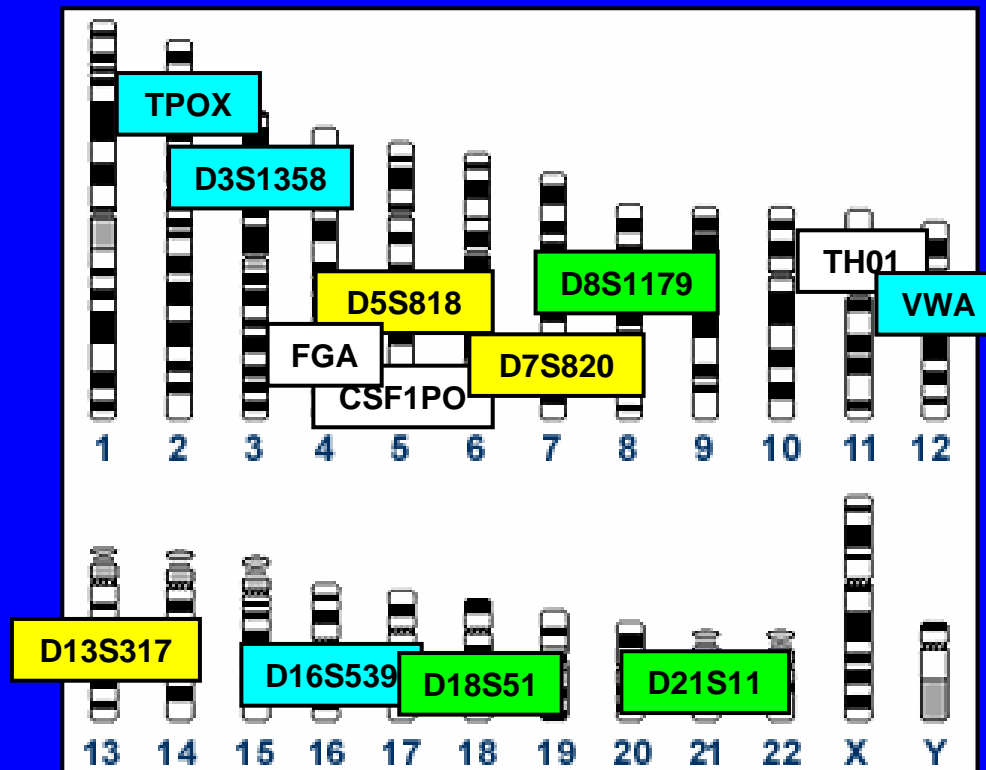
---

- The sample data is often partial due to degraded remains.
- There are experimental errors in the data
- Pedigrees may not be correct
- Direct reference samples may not belong to the victim
- The system is often only partially closed.



# Fingerprints: Nuclear DNA standard.

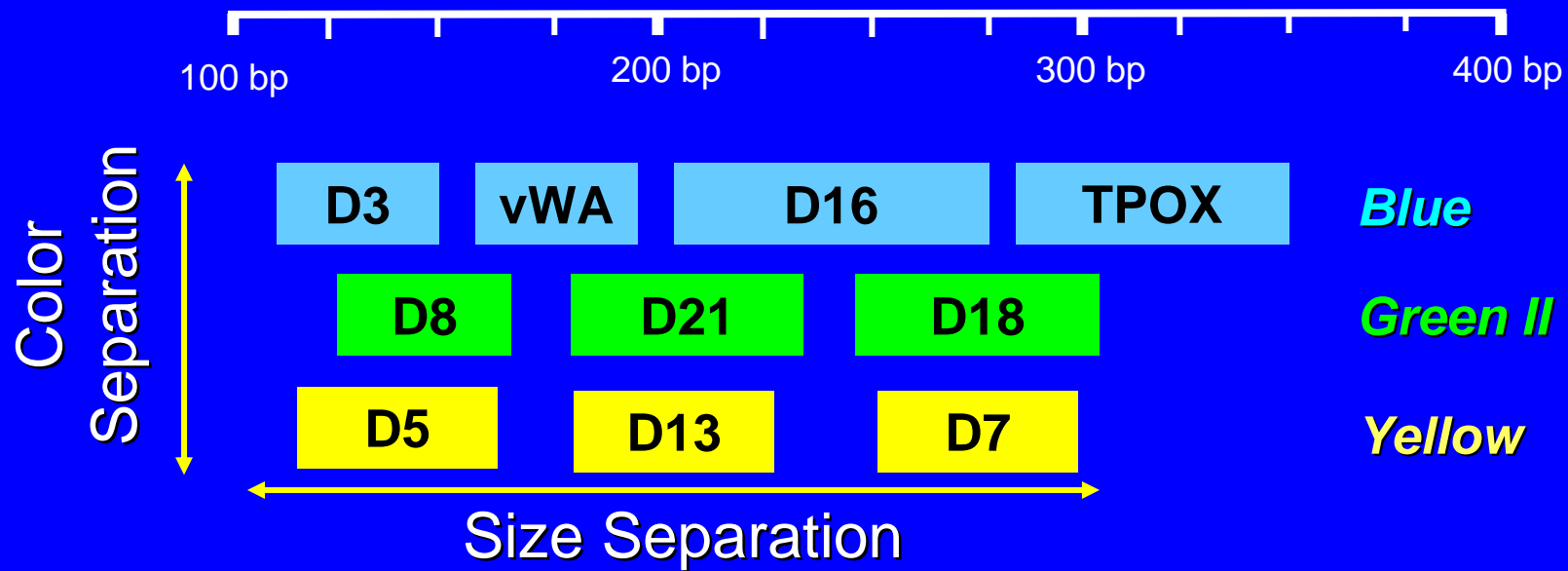
- FBI CODIS (Combined DNA Index System) standard for nuclear DNA utilizes 13 highly-variable tetramer STR sites.



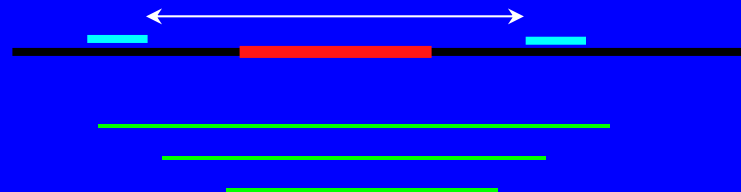
Multiplexed groupings:  
Blue, Green II,  
Yellow, & Cofiler

# Multiplex STR Analysis

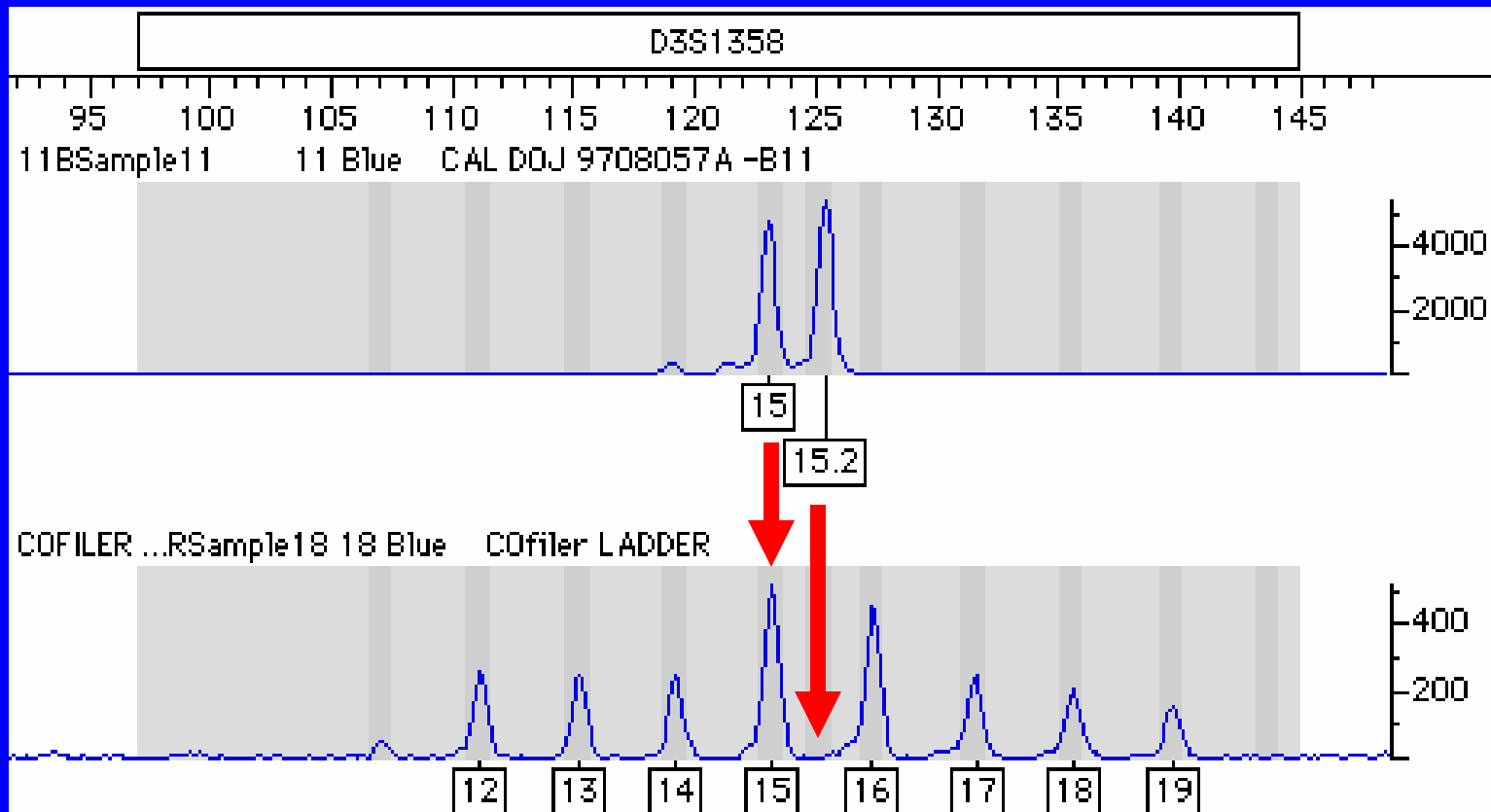
## AmpFISTR® Profiler Plus™



Select PCR primer sites at a distance around site to separate the products in each dye grouping:



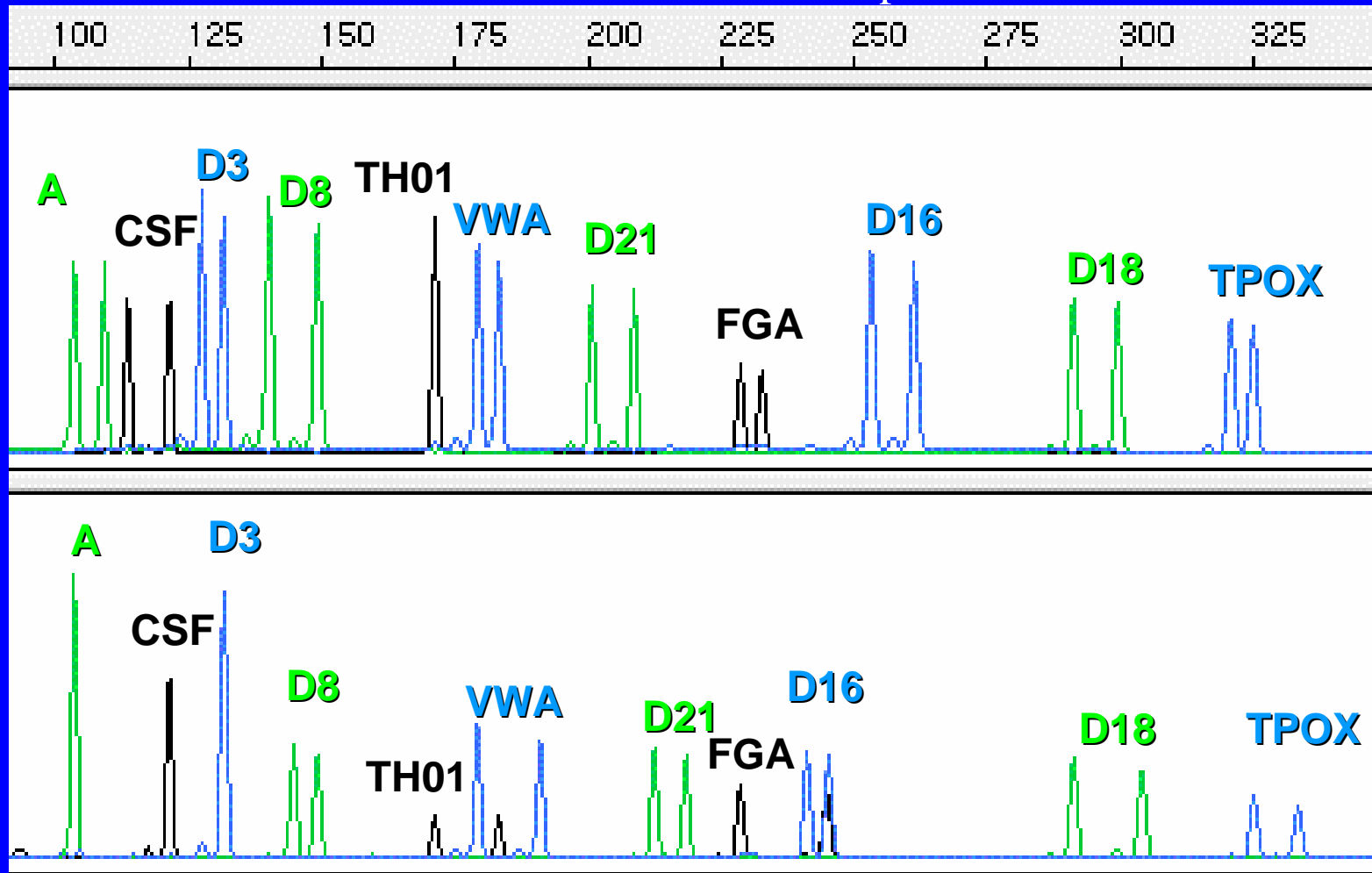
# Measure against Reference Ladder



# Multiplex STR Analysis

AmpFISTR® SGM Plus™ kit

Two different individuals

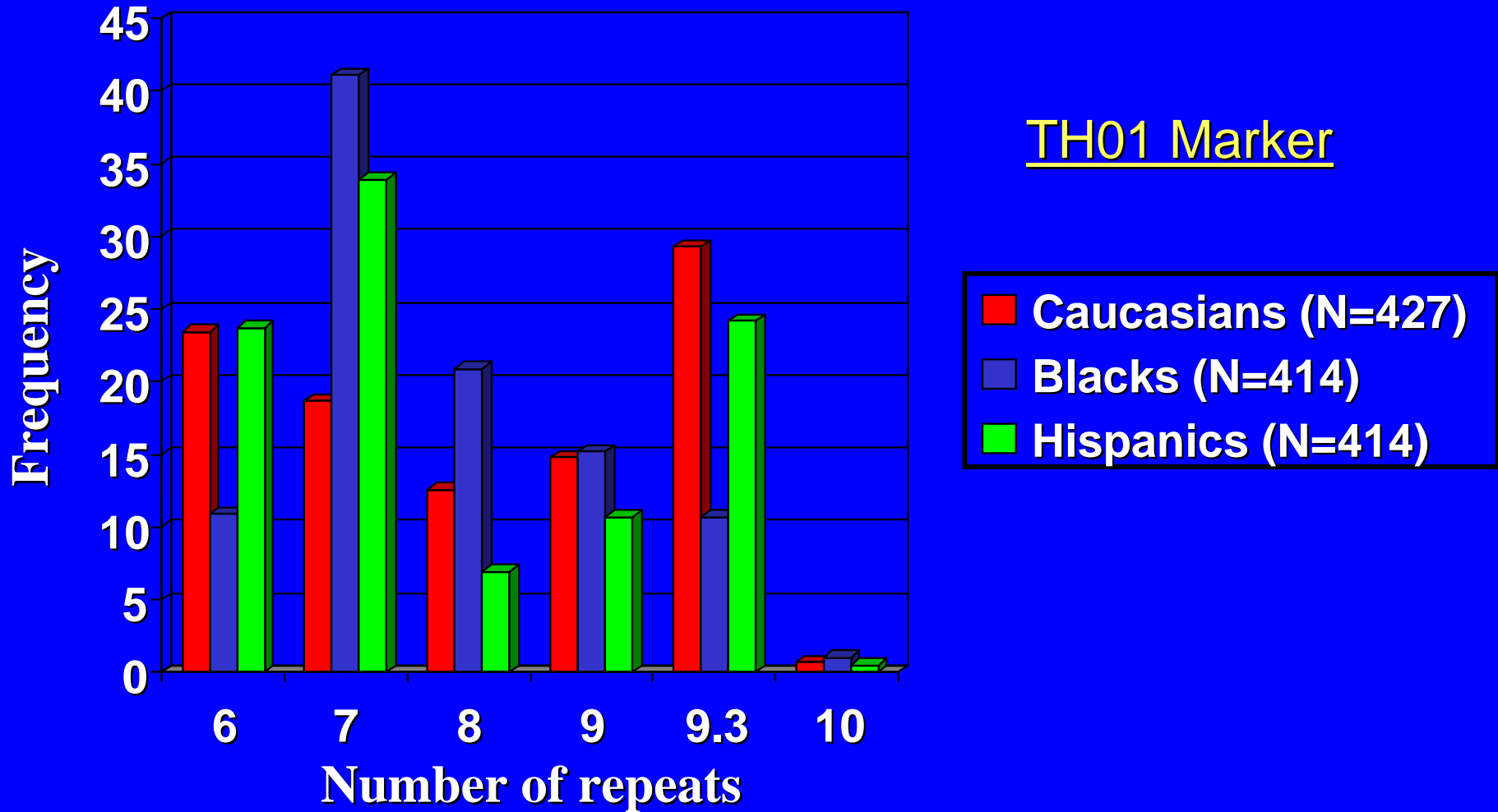


# Example of A CODIS Genotype

CSF1PO	10	12	
TPOX	8	8	
THO1	9	9.3	← 10 copies & a 1bp deletion in the non-STR part
VWA	16	16	
D16S539	8	12	
D7S820	10	12	
D13S317	11	13	
D5S818	11	13	
FGA	22	23	
D8S1179	14	15	
D18S51	14	20	
D21S11	28	35.2	← 35+.2 or 36-.2?
D3S1358	15	18	



# STR Allele Frequencies



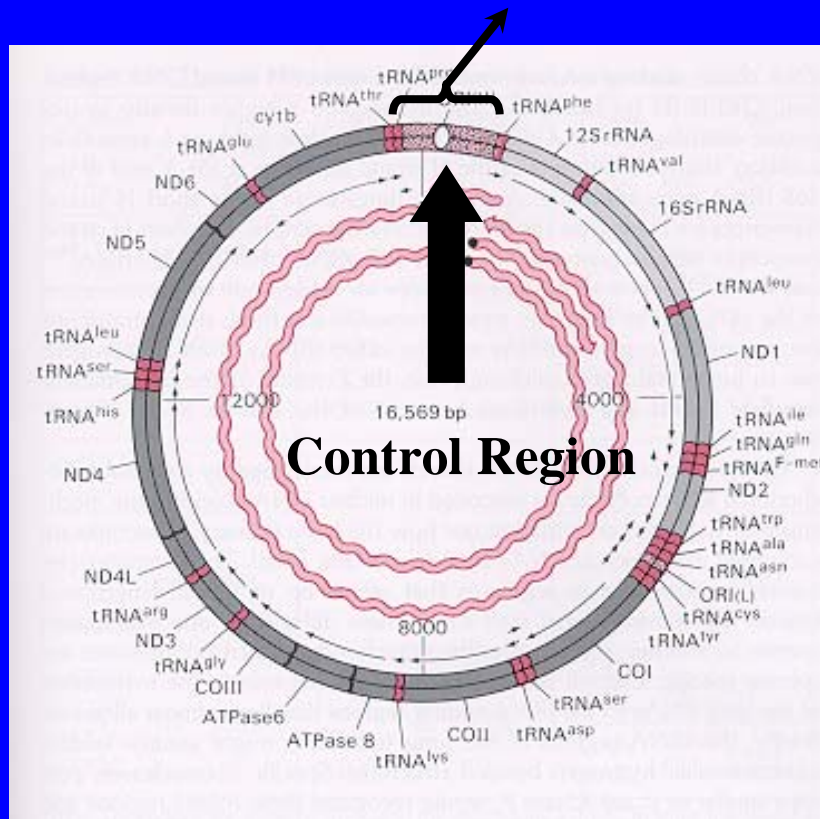
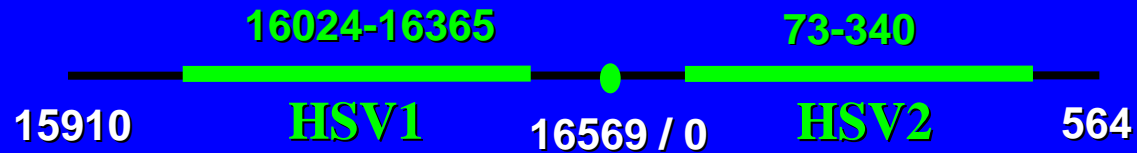
# Fingerprints: Mitochondrial DNA

---

- Mitochondrial DNA, because of its much greater copy # in the cell, is often sequencable when the sample is too degraded to permit nuclear DNA fingerprinting.
- The sequence of two hyper-variable regions of the d-loop control region are becoming a standard.



# Mitochondrial Architecture



# Fingerprints: Mitochondrial DNA

---

- Mitochondrial DNA, because of its much greater copy # in the cell, is often sequencable when the sample is too degraded to permit nuclear DNA fingerprinting.
- The sequence of two hyper-variable regions of the d-loop control region are becoming a standard.
- Maternal inheritance.
- Individuals are often heteroplasmic.



# Royal Pedigree Example

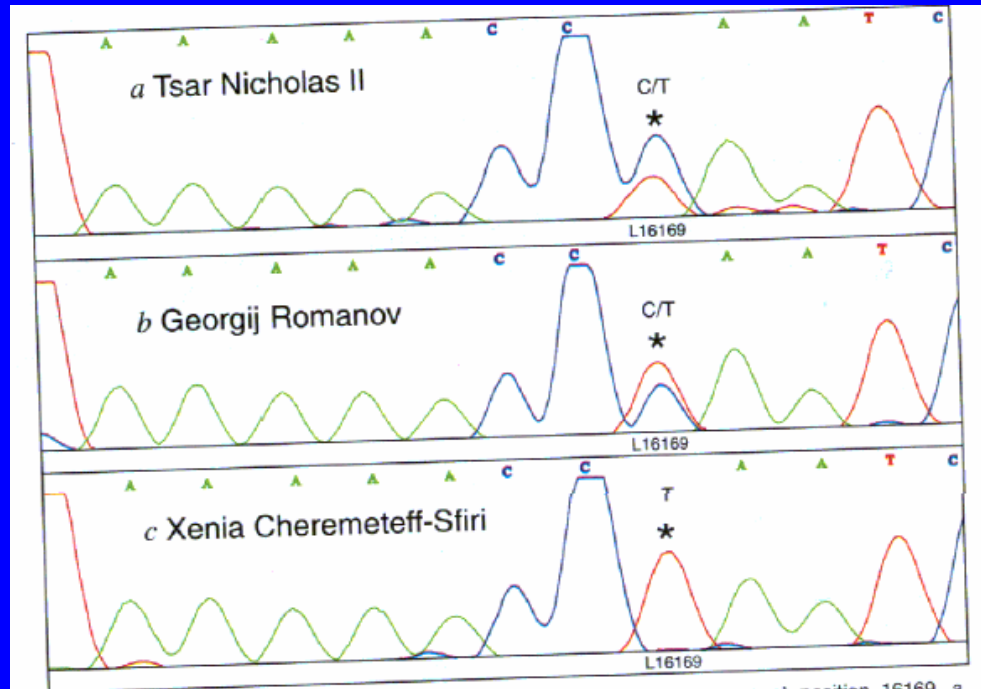
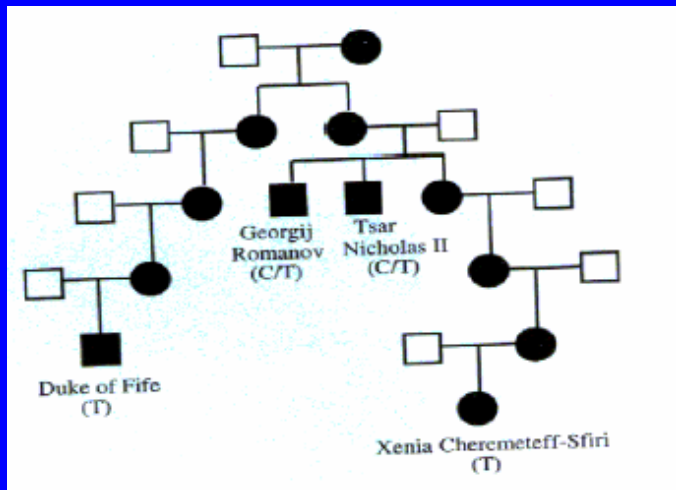


Fig. 2 Automated sequence chromatographs comparing mtDNA sequences at position 16169. *a*, Sequence from bones of putative Tsar Nicholas II, showing heteroplasmy with cytosine predominating thymine; *b*, sequence from bones of Grand Duke Georgij Romanov, showing heteroplasmy with thymine predominating cytosine; *c*, sequence from Countess Xenia Cheremeteff-Sfiri, homoplasmic for thymine.

# Fingerprints: Mitochondrial DNA

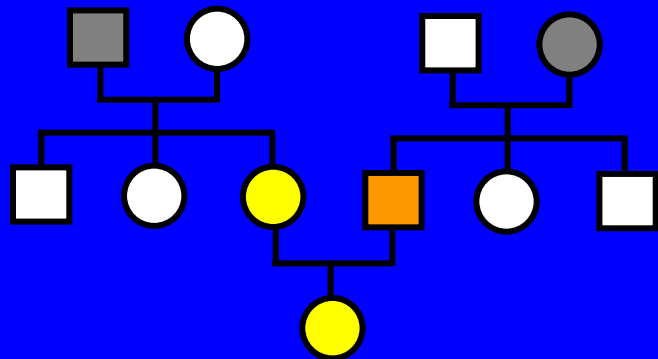
---

- Mitochondrial DNA, because of its much greater copy # in the cell, is often sequencable when the sample is too degraded to permit nuclear DNA fingerprinting.
- The sequence of two hyper-variable regions of the d-loop control region are becoming a standard.
- Maternal inheritance.
- Individuals are often heteroplasmic.
- Some variations are common, e.g. 7% of all Caucasian males have the same d-loop sequences.
- Stored as a  $\Delta$  from the Anderson reference sequence (typically 13 diffs over 608bp).
- Considered good for exclusionary inferences, but not the converse.



# Modelling Pedigrees

- Known (Relative)
- Unknown (Deceased or unsampled relative)
- Variable (Alleged Victim)
- Variable + Alleged samples (e.g. Comb, Toothbrush, etc.)



Parents & Child alleged deceased.

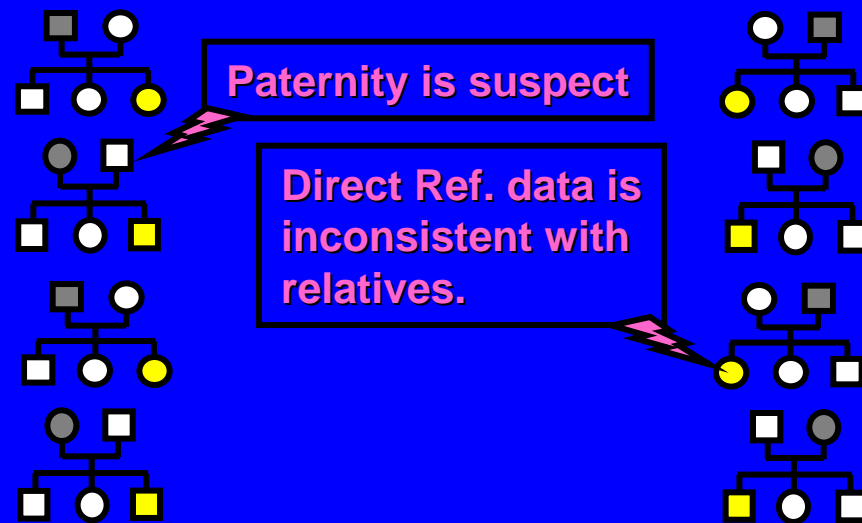
Data on brother, sister, and mother of wife  
& brother, sister, and father of husband  
& sample for husband.

Basic problem:  $\text{Prob}(\text{Pedigree} \mid \text{Fingerprints})$

# Approach: 1. Vet Pedigrees

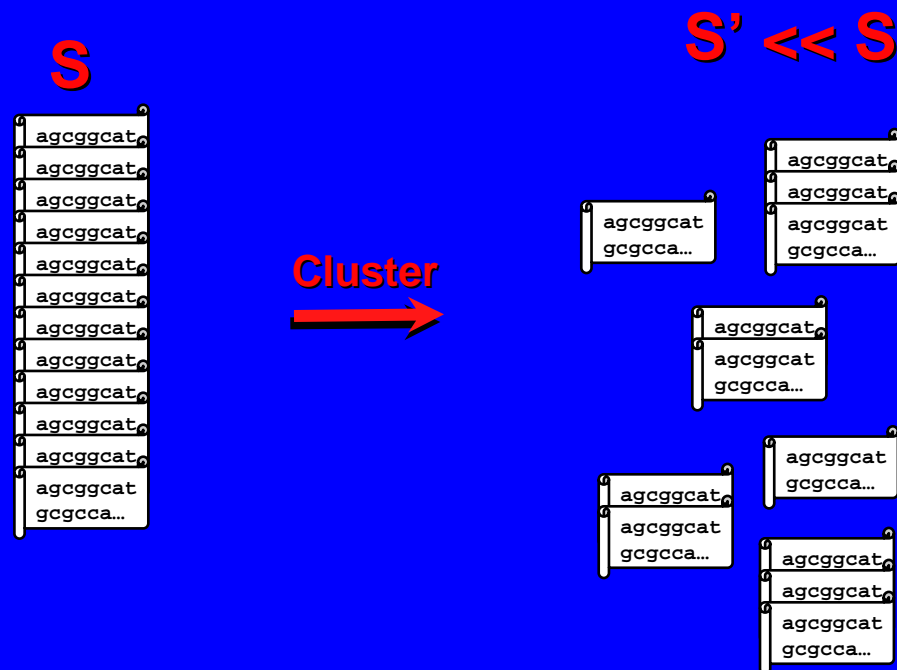
---

- Examine every pedigree and assess its probability ignoring variables (but not direct reference data). Flag all suspect pedigrees for human vetting.



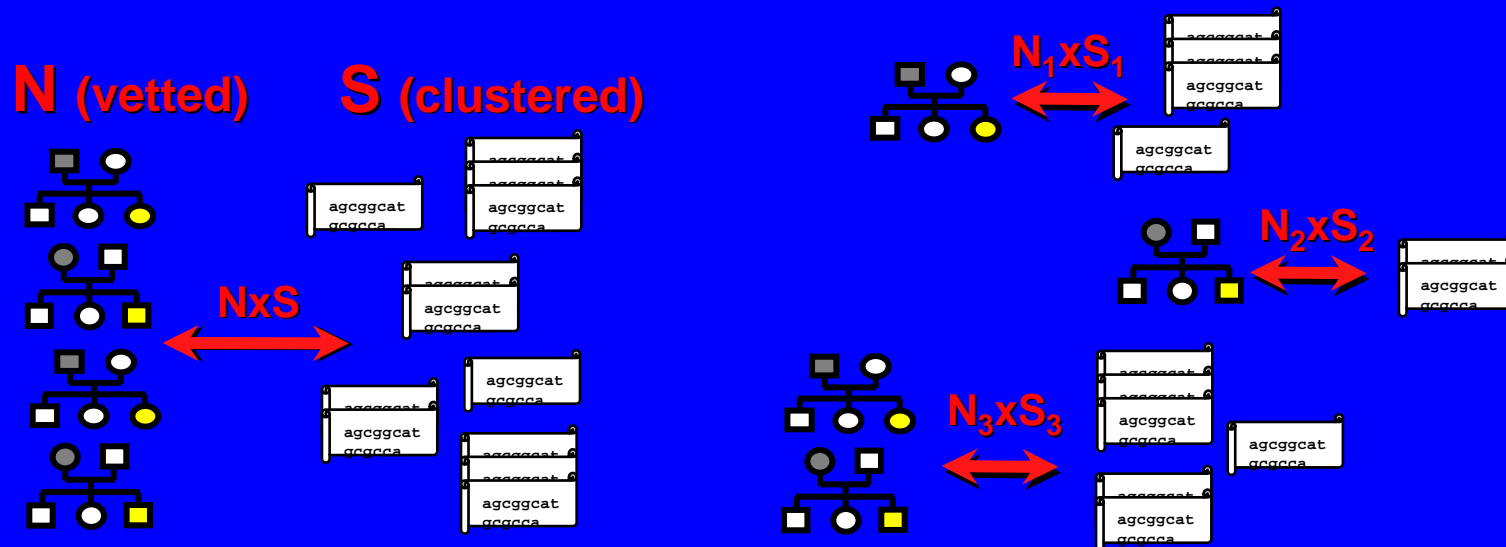
# Approach: 2. Cluster Samples

- Condense the sample fingerprints into distinct genotype clusters. Some partial fingerprints get “assembled” into more complete fingerprints. Can now think about assigning clusters, not individual samples.



# Approach: 3. Prune Pairings

- Given an  $N \times S$  pedigree-to-cluster problem, partition it into a collection  $\{ N_i \times S_i \}$  of  $T \leq N$  smaller subproblems, where with very high confidence no correct assignment has been eliminated. Present the partition to a forensic expert.



- Mitotypes are used to eliminate pairings at both the clustering and matching stages.

# Genotyping Error/Mutation Model

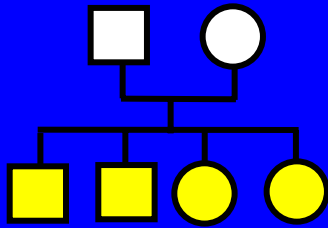
---

- PCR Stutter Error ( $\epsilon_s$ ): Is  $(x,x)$ , but instrument shows  $(x,x \pm 1)$ .
- Threshold Error ( $\epsilon_t$ ): Is  $(x,y)$ , but  $y$  signal below threshold so reported as  $(x,x)$ .
- Calibration Error ( $\epsilon_c$ ): Is  $(x,y)$ , but calibration ladder is off by one so reported as  $(x,y) \pm 1$ .
- Measurement Error ( $\epsilon_m$ ): Is  $(x,y)$ , but technician reads  $(x,y \pm .1)$ .
- Mutation Rate: Parent had allele  $x$  but becomes  $x \pm \Delta$  with probability  $.5\mu(1-\alpha)\alpha^\Delta$  where  $\mu$  is the mutation rate and  $\alpha$  is geometric decay parameter.



# Pedigree Probabilities (Basis)

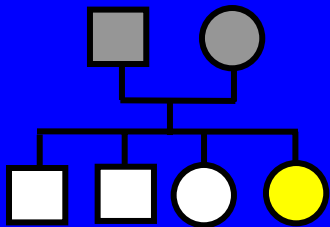
- Probability of allele pairs for  $k$  children from parents with alleles  $F = \{f_1, f_2\}$  and  $M = \{m_1, m_2\}$  depends only on the heterozygosity of each parental pair. **\*\*\* Assuming no error or mutation \*\*\***



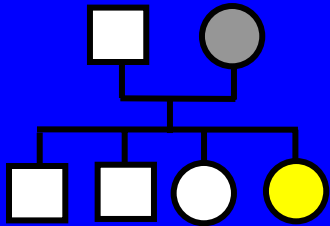
$$\Pr(C_1, C_2, \dots, C_k | F, M) = \begin{cases} 0 & \text{if } \exists C_i \not\subseteq \{f_1, f_2\} \otimes \{m_1, m_2\} \\ \frac{k!}{c_{11}!c_{12}!c_{21}!c_{22}!} (1/4)^k & \text{else if } f_1 \neq f_2 \text{ and } m_1 \neq m_2 \\ \frac{k!}{(c_{11}+c_{12})!(c_{21}+c_{22})!} (1/2)^k & \text{else if } f_1 \neq f_2 \text{ and } m_1 = m_2 \\ \frac{k!}{(c_{11}+c_{21})!(c_{12}+c_{22})!} (1/2)^k & \text{else if } f_1 = f_2 \text{ and } m_1 \neq m_2 \\ 1 & \text{otherwise (i.e. } f_1 = f_2 \text{ and } m_1 = m_2) \end{cases}$$

where  $c_{ab} = |\{C_i | C_i = \{f_a, m_b\}\}|$

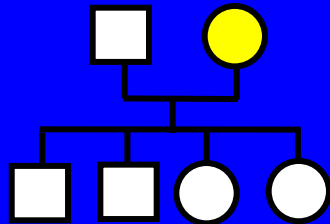
# Pedigree Probabilities (Families)



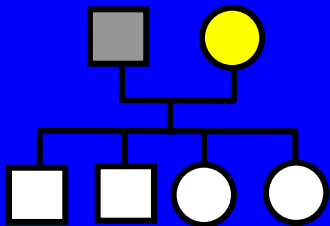
$$\Pr(C \mid C_1, \dots, C_k) = \frac{\sum_{Y,X} \Pr(C, C_1, \dots, C_k \mid Y, X) \Pr(Y) \Pr(X)}{\sum_{Y,X,Z} \Pr(Z, C_1, \dots, C_k \mid Y, X) \Pr(Y) \Pr(X)}$$



$$\Pr(C \mid C_1, \dots, C_k, F) = \frac{\sum_X \Pr(C, C_1, \dots, C_k \mid F, X) \Pr(X)}{\sum_{X,Z} \Pr(Z, C_1, \dots, C_k \mid F, X) \Pr(X)}$$



$$\Pr(M \mid C_1, \dots, C_k, F) = \frac{\Pr(C_1, \dots, C_k \mid F, M) \Pr(M)}{\sum_X \Pr(C_1, \dots, C_k \mid F, X) \Pr(X)}$$



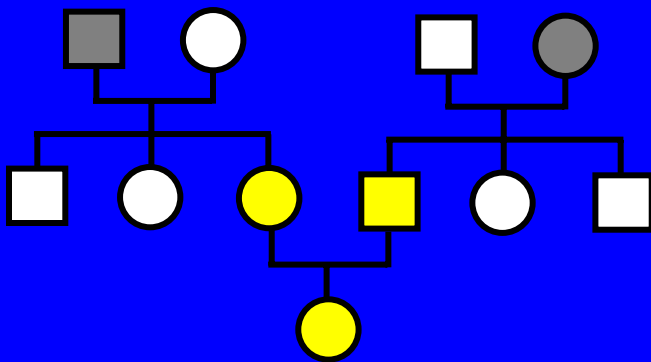
$$\Pr(M \mid C_1, \dots, C_k) = \frac{\sum_Y \Pr(C_1, \dots, C_k \mid Y, M) \Pr(Y) \Pr(M)}{\sum_{Y,X} \Pr(C_1, \dots, C_k \mid Y, X) \Pr(Y) \Pr(X)}$$

# Pedigree Probabilities

$$\Pr(\text{Family}) = \Pr(\text{Children}|\text{F},\text{M}) \Pr(\text{F}) \Pr(\text{M})$$

$$\Pr(\text{Pedigree}) = \sum_{\text{Unknowns } X_1, \dots, X_k} \prod_{\text{Family}} \Pr(\text{Family}(X_1, \dots, X_k))$$

$$\Pr(V | \text{Pedigree}) = \frac{\Pr(\text{Pedigree}+V)}{\Pr(\text{Pedigree})}$$



The order of summing over unknowns or “peeling” the pedigree can have a dramatic effect on complexity.

$$\Pr(\text{F},\text{M},\text{C}|\text{Pedigree}) = \frac{\sum_{X,Y} \Pr(\text{C}|\text{F},\text{M}) \Pr(\text{B}_M, \text{S}_M, \text{M}|\text{X}, \text{M}_M) \Pr(\text{B}_F, \text{S}_F, \text{F}|\text{F}_F, \text{Y})}{\sum_X \Pr(\text{B}_M, \text{S}_M|\text{X}, \text{M}_M) \sum_Y \Pr(\text{B}_F, \text{S}_F|\text{F}_F, \text{Y})}$$

In general: Bayesian Network !

# Pedigree Matches

---

- To vet a pedigree, compute  $\Pr(\text{G-type}|\text{Ped})/\Pr(\text{G-type})$  (including alleged samples) and report those that are above a given threshold
- Let the “score” of matching pedigree  $F$  to cluster  $C$  be  $\Lambda_{FC} = \Pr(C|F)/P(C)$
- If you seek a 1-1 matching between clusters and pedigrees then maximum bipartite matching under  $\log \Lambda_{FC}$  gives the matching with maximum *likelihood ratio*.
- But multiple clusters can match a family, so we use the *posterior likelihood*  $\Lambda_{FC}/(1+\sum_X \Lambda_{XC})$  that  $F$  is matched to  $C$  in a solution to *eliminate* unlikely edges (those below some threshold, e.g.  $10^{-6}$ )

# Clustering

---

- Find all (partial) genotype pairs that match exactly over all but at most  $e$  ( $=2$ ) loci.
- Evaluate matches, say  $S$  vs  $T$ , w.r.t. error model and allele frequencies:  $\Pr(S|T)/\Pr(S)$
- Cluster at specifiable probability threshold  $\Theta_C = O(S)$
- Produce consensus genotype for each cluster.



# Cluster Example

TPOX		CSF1PO		D5S818		D3S1358		D16S539		TH01		D7S820		
8	8	10	10	10	11	15	17	8	10	7	9	8	13	P002b
*	*	10	10	*	*	16	17	*	*	*	*	*	*	P002d
8	8	10	10	10	11	15	17	8	10	7	9	8	13	P002g
8	8	10	10	10	11	16	17	8	10	7	9	8	13	P002f
8	8	*	*	10	11	16	17	*	*	7	9	8	13	P002e
8	8	10	10	10	11	16	17	8	10	7	9	8	13	P002c
8	8	10	10	10	11	16	17	8	10	7	9	8	13	P002a
-----														
8	8	10	10	10	11	16	17	8	10	7	9	8	13	CONSENSUS
D13S317		VWA		D8S1179		D21S11		FGA		D18S51				
11	11	16	18	11	13	30	32.2	21	22	14	15	P002b		
*	*	*	*	*	*	30	32.2	21	22	14	15	P002d		
11	11	16	18	11	13	30	32.2	21	22	14	15	P002g		
11	11	16	18	11	13	30	32.2	21	22	14	15	P002f		
11	11	16	18	11	13	30	32.2	21	22	14	15	P002e		
11	11	16	18	11	13	30	32.2	21	22	14	15	P002c		
11	11	16	18	11	13	30	32.2	21	22	14	15	P002a		
-----														
11	11	16	18	11	13	30	32.2	21	22	14	15	CONSENSUS		



# Clustering: Algorithm

---

- Use  $m=2$  loci (4 #s) as an index for matching.
- Order loci according to specificity  $\sum_a f_a^2 \geq 1/\#a$
- Partition genotypes based on  $2^{13}$  loci **chords** (1-bit per loci in order of specificity, bit is on iff data for loci exists).
- For chords  $c$  and  $d$ , take all 6 pairs of 2 positions of the 4 highest loci in  $c$ & $d$  and compare genotypes in the same 2-loci bucket of an index.
- Filtration efficiency  $\sim m^e / \rho^m$



# Experiments:

---

➤ Pedigree is parents + child throughout.

➤ Typical Scenario:

$\varepsilon_u = 1/10$ ,  $\varepsilon_{m,c} = .001$ ,  $\varepsilon_{t,s} = .004$ ,  $|C| \in [3,7]$ , child is victim

➤ Noisy Scenario:

$\varepsilon_u = 1/3$ ,  $\varepsilon_{m,c} = .002$ ,  $\varepsilon_{t,s} = .008$ ,  $|C|$  in  $[1,9]$ , parent is victim

➤ Sizes: 100, 500, and 1000 families ( $\Rightarrow 5X$  samples)



# Results:

---

- Measure size of reduced matching problem: # of ambiguous clusters, families, and # of unresolved edges as a % of total.

	Unresolved/Total		
	Clusters	Families	Edges
Typical, 100	.01	.06	.0007
Noisy, 100	.12	.10	.0119
Typical, 500	.01	.04	.0004
Typical, 1000	.01	.03	.0002

Always the highly-degraded samples that are ambiguous

---

---

---

# Acknowledgement

---

Eric Xing and Tien-ho Lin (Carnegie Mellon) picked up this work 1/2 way through and took the software and methods to the conclusion presented here and also at:

T. Lin, E.W. Myers, and E.P. Xing,

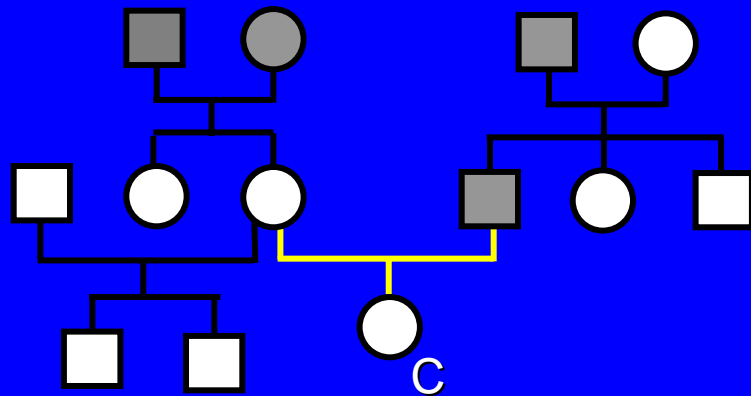
“Interpreting Anonymous DNA Samples From Mass Disasters -- Probabilistic Forensic Inference Using Genetic Markers,”

ISMB 2006

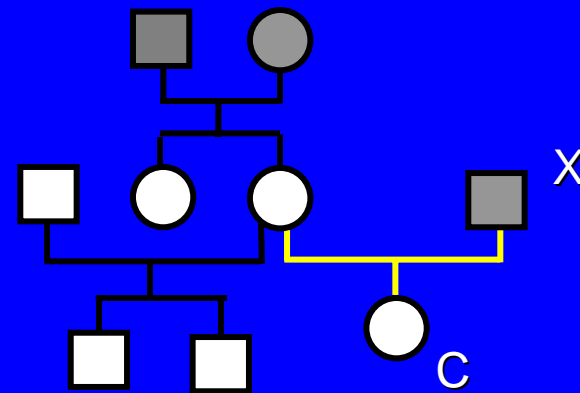


# Paternity Problems

alleged pedigree



null pedigree



$$\frac{\Pr(C|\text{alleged pedigree})}{\max \Pr(C|\text{null pedigree}(X))}$$



# Algorithmic Components

---

- Evaluation of a Pedigree: Peeling algorithm on hypergraph with new error/match model.
- Matching of Mitotypes: String matching on deltas.
- Clustering of Genotypes: New approximate match filter
- With E. Xing (CMU) are building an initial system and display.



# Progress

---

- Algorithm for clustering STR sample data: 20min. on a laptop for ½ million samples. Deals with partial data and finds all overlaps with at most 2 loci mismatches that are above a user-supplied probability level.
- Algorithm for comparing mitotypes that is linear in the amount of data.
- Algorithms to compute the probability of any pedigree given the frequency profiles for the loci and (partial) data on the individuals.
- Contacts with the FBI to take their CODIS system to the next level in handling disaster and missing persons cases.
- Can get WACO, an airplane crash, and have several complex paternity cases.



# Counting & Conditional Probability

- Given  $n$  things, # of ways to pick  $k$  is  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- Given  $n$  things, # of ways to pick  $k_1$ , then  $k_2$  of what's left is  $\binom{n}{k_1, k_2} = \binom{n}{k_1} \binom{n-k_1}{k_2} = \frac{n!}{k_1!k_2!(n-(k_1+k_2))!}$
- Given  $n$  things, the # of partitions into 4 groups of sizes  $k_1, k_2, k_3$ , and  $k_4$  where  $k_1+k_2+k_3+k_4 = n$  is  $\binom{n}{k_1, k_2, k_3, k_4} = \frac{n!}{k_1!k_2!k_3!k_4!}$
- Recall that  $\Pr(A|B) = \Pr(A \cap B) / P(B)$
- Lots of things then follow easily such as:
  - $\Pr(X) = \sum_i \Pr(X|A_i)\Pr(A_i)$  provided  $\sum_i \Pr(A_i) = 1$
  - $\Pr(A|B) = \Pr(B|A)\Pr(A)/\Pr(B)$
  - $\Pr(A|B) = \Pr(B|A)\Pr(A) / \sum_i \Pr(B|A_i)\Pr(A_i)$  (Baye's Rule)