Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

DEVELOPMENT UNDER EXTREME CONDITIONS: FORENSIC BIOINFORMATICS IN THE WAKE OF THE WORLD TRADE CENTER DISASTER

HOWARD D. CASH, JONATHAN W. HOYLE, AMY J. SUTTON
Gene Codes Forensics, 775 Technology Drive, Suite 100A,
Ann Arbor, MI 48108, USA

The terrorist attacks of September 11, 2001 resulted in death and devastation in three locations, and extraordinary efforts have been exerted to identify the remains of all victims. As mass fatalities go, this one has been unusual at a policy level because the goal has been not merely to identify remains for every decedent, but to identify every bit of remains found so that even small pieces of tissue can be returned to families for burial. While the human impact at the Pentagon and Shanksville, PA was horrific, the World Trade Center site presented a particularly complex challenge for forensic DNA matching and data handling. A complete and definitive list of all those killed is still elusive, and human remains were crushed and co-mingled by the falling towers. Software tools had never been considered for a problem of this scale and scope. New data handling systems had to be created under extreme software development conditions characterized by incomplete requirements specifications, chaotically changing feature priorities, truly impossible deadlines and rapidly rolling production releases. Partly because of the company's experience with mtDNA tools built for the Armed Forces DNA Identification Lab starting in 1997, the New York City Office of Chief Medical Examiner [OCME] contacted Gene Codes Corporation in late September as existing data-handling tools began to fail. We began work on the project in mid-October, 2001. Our approach to the problem included:

- *Extreme Programming* [XP] methodology for functional software development,
- On-site time and motion analysis at the OCME for user interface design,
- Evidentiary references between STR, SNP and mtDNA analysis results, and
- Separate data Quality Control [QC] and software Quality Assurance [QA] initiatives.

A substantial software suite was developed called *M-FISys*, an acronym for *Mass-Fatality Identification System*.

1 Background

The New York City Office of Chief Medical Examiner [OCME] faced an unprecedented problem after September 11, 2001. Early estimates suggested that over 10,000 people had been killed in the attacks at the World Trade Center (this number would be revised down to 2,801 by the first anniversary of the disaster). Forensic identification of remains would be complicated by the substantial fragmenting of remains (with individuals recovered in as many as 200 pieces[a]) and the heat and moisture associated with the fighting of jet fuel fires that burned for months. Certainly some fatalities would be identified by classical methods such as viewed remains, fingerprints, dental records and personal property. Because of the crushing and co-mingling of remains, personal belongings found at "Ground Zero"

[a] See also Chicago Tribune, *To scientists, ID project is a sacred trust,* S. Swanson, Sep11, 02 and Los Angeles Times, *Probing the DNA of Death,* R. Hotz, Oct9, 02

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

have been considered highly suspect as identifying evidence.[b] The vast majority of individual remains would have to be identified by DNA matching, and the associated data handling is the subject of this paper. Because of the public interest that we all share in this case, attention will be given to the circumstantial pressures that impacted the software development process.

Functionality had to be added to the M-FISys program [pronounced like *emphasis*] on a very fast schedule without sacrificing software quality and testability. In fact, a new release of M-FISys has been delivered to the OCME almost every week since mid-December, 2001 (the 38th iteration is being released at the time of this writing). The need for stringent software quality control is self-evident and the moral magnitude of the task plus the knowledge of the catastrophic impact of any false identification was a constant weight on the shoulders of the developers. Extreme Programming [XP] [15,16,17,18,19] has been an invaluable methodology under the existent pressures. Since almost all studies of software correctness indicate that code reviews are one of the most valuable QA tools, XP programmers work in pairs, with one engineer constantly reviewing the work of the other. Both unit tests and acceptance tests must be written before functionality can be added. In order for new work to be checked back into the source code control system, not only must it pass these tests, but it must pass all other tests that have been written against all functions since the project began. An obvious benefit of this discipline is that it makes it difficult to unknowingly "break" a part of the program while fixing a seemingly unrelated bug. It can also add a cost any time there is a major architectural refactoring because many of the legacy tests will have to be re-written to accommodate the new architecture.

Under the XP methodology, functionality is added each week through observation and direct negotiation with users in the field. One or two staffers have been on-site daily at the OCME since the project began, and the chief designer has traveled back and forth between the user location in Manhattan and the developer site in Ann Arbor each week. Weekly reviews identify opportunities for process improvement. Productivity of the engineering is carefully monitored to the point that one can literally choose any week since the project began, recover the source code as it was during that period, identify what new functionality was scheduled to be added, which engineers were available, initial time estimates for tasks and review the actual programming velocity of the entire team for that iteration.

---

[b] To illustrate the complexity of the investigation, it is worthwhile to note that a number of remains would be identified by tracing serial number on implants such as cardiac pacemakers and artificial joints, contacting the manufacturer (sometimes overseas), tracing the hospital to which the item had been sold, and then cross-referencing all implantation surgeries against the list of missing persons from the Trade Center. See also "The Ethical Aftermath of Sept 11: Body Identification Issues of Remarriage and Inheritance" from *The International Conference on Judaism and Contemporary Medicine,* Nov 15-17, 2002, sponsored by The National Institute of Judaism and Medicine.

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

An important principal of our methodology has been to rigorously resist the urge to propose functionality to the OCME. This is very different from taking the luxury of adding our own "helpful" tools as we would in a commercial design and development project. Our process analysts work through proposals based on the immediate needs and priorities of the forensic biologists, and while we take responsibility for architecture and initial interface design, we do not allow the engineering team to pursue its own vision for the program's functionality. It takes effort for skilled and creative engineers to stick to this discipline, but it is critical to keeping up our development velocity. Outside contractors and advisors have become frustrated and occasionally openly hostile when pet proposals have not made it onto the development list because OCME staff have evaluated those proposals as adding insufficient value. Certainly the emotional content of this development project is a major component of the managerial complexity.

2 Methods of Identification

The most widely used forms of DNA-based human identification involve Short Tandem Repeat [STR] analysis at 13-15 nuclear loci. A second procedure involves sequencing the hypervariable regions of the mitochondrial genome. A relatively new forensic procedure, pioneered by Orchid Biosciences, is based on Single Nucleotide Polymorphisms [SNPs] with well-characterized inheritance and frequency patterns. All of these techniques had to be combined in the M-FISys program.

The theoretical basis of STR analysis in human identification can be traced back to the work of Alec Jeffreys [25]. In modern applications, 13-15 unlinked STR loci are sized, resulting in a "profile."[c]

Table 1: Sample STR data

D3S1358	vWA	FGA	D8S1179	D21S11	D18S51	D5S818
15/16	16/17	26/28	14	30/32.2	15/16	12/13

D13S317	D7S820	D16S539	THO1	TPOX	CSF1PO
12	12	9/12	9.3	6/9	10/11

To orient the reader, the above profile can be read as follows: at the FGA locus, this person inherited 26 repeat-elements from one parent and 28 from the other. The D8S1179 locus is homozygous for 14 repeats. Based on experimentally observed and published frequencies of each repeat value in various ethnic populations, this profile would be expected to be seen no more than one in 10^{17} individuals.

[c] In addition to being chosen for genetic distance and independence, the core loci have been selected to be medically uninformative, thereby protecting the medical privacy of individuals in a forensic or criminal investigation.

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

The likelihood value for a given STR profile is the product of the likelihoods of each of the 13 STR loci. As likelihood values are inverses of frequencies, we can justify a simple multiplication since the value at each locus is an independent event, unrelated to the other loci.

The benefit of this is that we can have comparative likelihoods across partial profiles. If a policy decision is made to set a threshold likelihood in order to report an identification, for example a likelihood of no less than 10^{10}, an incomplete STR profile's likelihood may still surpass this value if the available alleles are rare enough.

The likelihood value for a given profile will differ depending on ethnic population; however, access to that information is not always available or reliable where badly damaged remains are involved. We therefore assume the "worst case scenario" by examining the likelihood values across four races (Asian, Black, Caucasian & Hispanic) and take the lowest value as the final likelihood. By using this lowest value, we err on the side of preventing false positives.

To determine the likelihood of a particular locus, we take the inverse of its frequency. This frequency can be first approximated by ignoring population structure and using the Hardy-Weinberg proportions:

p^2	for homozygous alleles, where p = probability of the allele	(1)
$2pq$	for heterozygous alleles, where p,q = probabilities of each allele	

However, since all humans are related, variance of allele proportions is factored in with the inbreeding coefficient θ (theta), with:

$p^2 + p(1-p)\theta$	for homozygous alleles	(2)
$2pq(1-\theta)$	for heterozygous alleles[d]	

As stated in *Interpreting DNA Evidence* by West & Weir, "$\theta = 0.03$ is a conservative upper bound on the values appropriate for human populations (National Research Council 1996)."[1] For the WTC project, we use the value of 0.03 for θ.

The probabilities p,q are determined empirically themselves, based upon population. (The values we used are in the STR13.xls spreadsheet provided by Dr. Howard Baum, Deputy Director of Forensic Biology at the New York City OCME. [12])

Let S be any STR profile containing A_k loci, k<=13. Furthermore let p_k be the probability frequency of the high allele of A_k, and (if heterozygous), let q_k be the probability frequency of the low allele of A_k. Then the Likelihood of S, L(S), is defined as:

(3)

[d] The Medical Examiner's Office in New York chose to use the equation involving θ only for the homozygous alleles, so that the Likelihood values were always more conservative.

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

$$L(S) = \prod_{k \in Alleles} \frac{1}{P(A_k)}$$

where

$$P(A) = \begin{cases} p^2 + p(1-p)\theta \;\; a \;\; \text{hom}\, ozygous \\ 2pq \;\; a \;\; heterozygous \\ 1 \;\; a \;\; non-existent \end{cases} \tag{4}$$

for each k and θ = 0.03.

3 Implementation and Design Decisions

We will first address the issue of matching profiles that are complete (all 13 loci plus a gender locus) or nearly so. Limit this discussion to those cases where profiles have a likelihood of no less than 10^{10}.[e] [13] Even at this level of match stringency, the first task was to address the information glut of having 25, 50 or 100+ individual samples that had identical profiles because they were fragments of the same person.

For several weeks after the disaster, attempts to match remains used a software package called CoDIS (Combined DNA Index System). CoDIS was designed as a criminal investigation tool to compare a single DNA profile to a database of profiles from crime scenes and from convicted felons.

CoDIS, the result of a federally funded software development project, is a system widely and successfully used in criminal investigation but never intended for mass disasters. It is complicated to use under ideal circumstances, requiring special training for operators. The World Trade Center disaster created a situation so incompatible with the original engineering intention of CoDIS that early attempts to apply it to the World Trade Center disaster were described by the forensic scientists as a disaster in itself. The program generated up to 4,000 printed pages of internally consistent matches (matches between samples and between samples and exemplars). Teams of forensic scientists would literally spend days pouring over CoDIS output with highlighter and ballpoint pens in an attempt to sift through the

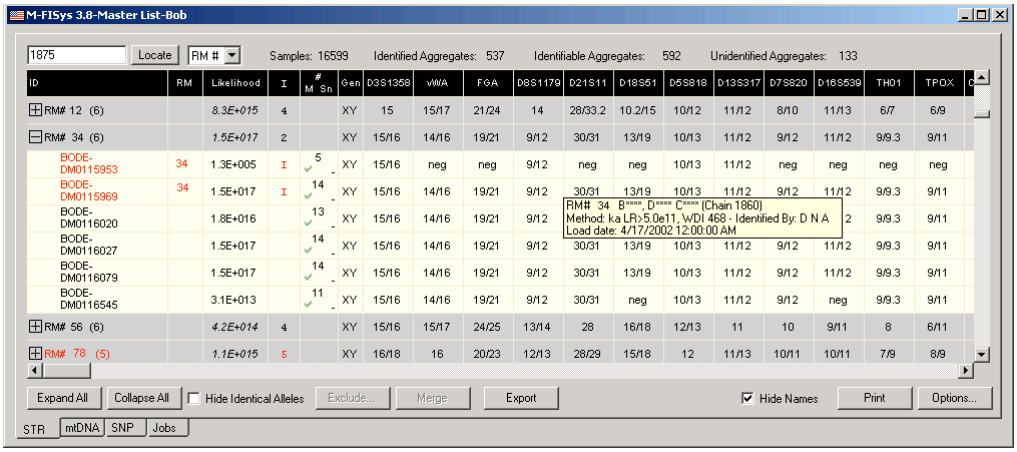
[e] As suggested earlier, the threshold of 10^{10} was a policy decision. In a population of 5,000 victims (the estimate of fatalities at the time of the discussion), this would reduce the likelihood of even a single misidentification based on a fortuitously shared STR profile to less than 1 in 1,000,000. See also "World Trade Center DNA Identifications: The Administrative Review Process," Mike Hennessey, Gene Codes Forensics, Inc; *13th International Symposium on Human Identification*. http://www.promega.com/geneticidproc/ussymp13proc/contents/hennessey.pdf

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

information. The usability issue of reducing the size of this haystack was clearly a first priority.

The first working version of M-FISys, delivered on December 13, had the immediate advantage of collapsing multiple samples with the same profile into an expandable "aggregate" sample. This alone was such an improvement over the CoDIS-based methodology that 55 new ID's were recognized the first day the program was installed. In Figure 1, the line labeled RM #34 (6) indicates that there are six internally consistent profiles in the collapsed set.[f]

Fig. 1. The Master List displays aggregated STR profiles in a searchable and collapsible format. Identifying numbers have been changed.



The software presents data to staff in the DNA Identification Unit at the OCME in three categories: 1) DNA profiles from victim remains, 2) ante-mortem DNA profiles of personal references from missing persons (e.g., a pre-existing pathology sample, a toothbrush or a lipstick) and 3) profiles from buccal swabs taken from first-order relatives. The first attempt is to match a victim sample to a personal reference exemplar.

Unlike a criminal investigation application where a single sample might be compared to the database, the WTC recovery required repeated all-against-all comparisons as additional samples were recovered and analyzed in the lab. Furthermore, as compromised samples were reanalyzed with hopes of extracting additional information, "Virtual Profiles" needed to be created, showing all accumulated values while keeping track of exactly which laboratory attempt had generated which data points. As data is added and operator decisions are made about

[f] RM stands for Reported Missing. Each presumed victim has an RM number. These numbers have been changed and names obscured for purposes of this paper.

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

assignment of data, a history (necessarily including free-text annotations) is kept for each sample.

The possibility of fortuitous (coincidental) matches increases as match stringencies are lowered. For samples with low partial profiles (likelihoods in the 10^3 – 10^4 range) one might expect several possible direct matches. This is particularly true if one further allows for the possibility of allelic drop out (loss of one of the two alleles in a locus because of the damaged condition of the sample – allowing an experimentally homozygous 12 to match a 12/16 heterozygous reference). M-FISys helps a forensic scientist to resolve these ambiguous matches by a process we call *iterative pruning*. The operator can confirm or exclude possible matches for any ambiguous sample, annotating reasons along the way. In a hypothetical example, a right hand with degraded DNA can be excluded from a potential ID on the grounds that a full profile for a right hand has already been reported. As more information is accumulated, more matches can be excluded or confirmed. mtDNA or SNP data may help to confirm or refute a potential match. Finally, experimental errors can also be corrected and annotated as part of an exquisitely meticulous Administrative Review and Quality Control process that is beyond the scope of this paper.

When direct matches to ante-mortem exemplars are not possible, or if the chain of custody for the exemplar cannot be reliably traced[g], kinship analysis is the next source of evidence. Many ethical, legal and social issues are brought into sharp relief at this stage of the process. For instance, the State of New York temporarily suspended Informed Consent rules for the case of collecting buccal swabs from family members. This was not, as some mistakenly inferred, motivated by a need to collect family references more quickly. Rather it was a compassionate ruling allowing for samples to be taken (in many cases from young children) without forcing the contributor to engage in a discussion about the need for DNA samples. There are those who might suggest that the legal suspension of Informed Consent regulations gives tacit permission to participating scientists to perform population genetics research using the database of collected victim and kin samples. This would be a gross and egregious violation of the dignity, privacy and civil rights (under other laws in the State of New York)[h] of the victims and their families. A less extreme issue is the decision on what to do if *false paternity* is ever detected. It seems only reasonable and compassionate to withhold any such findings

---

[g] Verifying the origin of a reference sample has proven to be an enormous challenge in this process. Because of the chaos and emotional stress of the first weeks following 9-11, many contributed samples were accepted with incomplete or even incorrect paperwork, As a result, all identifications are based on no less than two match modalities. See also "World Trade Center DNA Identifications: The Administrative Review Process," Mike Hennessey, Gene Codes Forensics, Inc; *13th International Symposium on Human Identification.*

[h] New York Civil Rights Law Article 79-1 requires informed consent for separate genetic research, and approval by an Institutional Review Board.

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

permanently: What could be more devastating for a father than to learn that not only had a child been lost to a senseless act of terrorism, but that the child had never been his?

As part of the WTC recovery and in collaboration with Dr. Charles Brenner and Dr. Benoit LeClaire, kinship analysis can be tactically approached in two different ways. Using Brenner's approach, family pedigrees are constructed based on the reported relationships to the victim and the set of victim profiles is scanned for promising candidates to fill in the missing (victim's) position in the pedigree. The data can then be revised to allow for incorrectly reported relationships and matches can be further refined. Dr. LeClair takes the conservative approach of assuming that any reported relationship could be wrong (this turns out to be true with unfortunate frequency due to errors in the original collection of data). Each kinship sample is compared to victim samples and tested for likelihood at various relationships such as parent-child, sibling or half-sibling. These are sorted by most-likely match. In an accurate match for a particular victim sample, the correct familial samples will float to the top of the list.

The mathematics of kinship analysis is well established.[1,2,4] Scores are reported as a hypothesis-based "likelihood ratio," calculated as the likelihood that the relationship between the victim sample and the kin sample is as hypothesized (e.g., full sibling in the Hispanic population) divided by the likelihood of seeing the same number of matching alleles in two unrelated persons in the same population.

Similar to the Likelihood in direct STR matches, Kinship Likelihood is the product of Kinship Locus Likelihood across all existing alleles (partial profiles are possible here as well). What is different is that two profiles are being compared instead of just one. In the case of partial profiles, only those loci extant in both profiles can be compared. As in the case of direct STR Likelihood, Kinship Likelihoods were calculated across the four aforementioned races, the lowest value taken as "a worst case scenario".

The Kinship Locus Likelihood takes as input a relationship between the two profiles, which is one of: Parent/Child, Full Sibling, Half Sibling. Formulae exist for First Cousin relationships as well, but it was felt to be less useful for the World Trade Center project, so was never used (swabs from cousins were not taken); however, because it was coded into M-FISys, the formula will be shown here as well. Let p be the frequency probability of the high allele value, q the frequency probability of the low allele value. Then the Kinship Locus Likelihood is defined the following way [15]:

$$(a_2p_2 + a_1p_1 + a_0p_0) / p_2 \tag{5}$$

where a_i's are proportions based on relationship:

$$\text{Full Sibling:} \quad a_2 = 1/4,\ a_1 = 1/2,\ a_0 = 1/4 \tag{6}$$

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

Half Sibling:	$a_2 = 1/2,\ a_1 = 1/2,\ a_0 = 0$
Parent-Child:	$a_2 = 0,\ a_1 = 1,\ a_0 = 0$
First Cousin:	$a_2 = 3/4,\ a_1 = 1/4,\ a_0 = 0$

and the pi's are defined as follows:

(7)

p_2	$= p^2$	if the victim is homozygous and matches one of the relative's alleles
	$= 2pq$	otherwise
p_1	$= 0$	if relative and victim alleles share no common value
	$= p$	if the relative is homozygous and the victim shares its (homozygous or heterozygous) low value
	$= q$	if the relative is homozygous and the victim shares its (heterozygous) high value
	$= p/2$	if the relative is heterozygous and shares exactly one value with the victim's (homozygous or heterozygous) low allele
	$= q/2$	if the relative is heterozygous and shares exactly one value with the victim's heterozygous high allele
	$= (p+q)/2$	if relative and victim are identical and heterozygous
p_0	$= 1$	if the relative and victim alleles are identical
	$= 0$	otherwise

The unfortunate truth is that some samples are so severely burnt that full STR profiles will not be available for either direct or kinship analysis. When samples are badly compromised through harsh environmental degradation, many of the loci may be blank and the likelihood of observation can drop to levels where one would expect many instances of shared "partial profiles" in a population of 2,801 victims. At this point, the options for making a positive ID are a) to re-extract and retest DNA with hopes of collecting information at additional loci, or b) to use other techniques that might be more effective on highly degraded samples, even if they are not as discriminating. One of these alternative techniques is mitochondrial DNA analysis.

Mitochondrial typing involves direct sequencing of the highly variable regions of the genome adjacent to the origin of replication. The standard in forensic communities is to report results not in the form of the sequence itself, but in a compact format that only shows the difference between the experimental results and an established and internationally recognized standard known as *The Anderson Sequence*[i].[20] If the sequence being typed is identical to the Anderson reference, the difference report will be null. Point mutations are described as a base position, plus the base that differs from the reference sequence. Deletions are represented as a "D"

[i] The Anderson reference [Nature 1981 Apr 9;290(5806):457-65] is a published representative mitochondrial sequence but is not necessarily the most common. An alternative reference sequence called The Modified Cambridge Reference Sequence [Nature Genetics 1999 23: 147] is gaining in popularity.

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

character (not to be confused with the IUB code for "A, G or T, but not C"). We call this difference list the "delta representation." A typical mitotype might look like this:

```
16093:     C
16224:     D
16311:     C
195:       C
263:       G
315.1:     C
```

Note that the base positions cross the origin of the 16569-base genome. The first lines indicate that this sample has a C at position 16093 where a T is usually found, and a deletion of the base at position 16224. To maintain the integrity of numbering, an insert is indicated as a decimal point position on the base that the insert follows. In this case, "315.1: C" indicated that there is a C insert after position 315 in the reference sequence.

Although standards have been promulgated, there is still the occasional disagreement over nomenclature, particularly as it applies to inserts. For instance, if the reference sequence includes the sequence "TTT" starting at position 16091, and the sample under study has four T's rather than three, it can logically be represented as any of the following:

```
16090.1:   T
16091.1:   T
16092.1:   T
16093.1:   T
```

The sequence used in mitotyping is not in a coding region and, since the inserted T has no biological significance, it can be referenced as an insert in any of several places. Two of the collaborating labs were not in precise agreement on the standard notation for this insert, even though both were producing accurate mtDNA profiles. Inflexible and sometimes emotional positioning (almost exclusively from less experienced outside advisors, rather than members of the labs themselves) over the "correctness" of one naming representation or another was an example of one of the greatest non-technical challenges of this project. We defined this to be a non-issue for the current project, since our solution was to reverse-translate the difference list to generate a non-ambiguous original sequence and compare those sequences themselves.[j]

mtDNA is abundant and very hardy material that can survive intact under conditions where nuclear DNA degrades. However, the variation in the hypervariable d-loop of the mitochondrial genome is not nearly as discriminating as a 13 locus STR profile. In fact, over 5% of the Caucasian population share the same, common

---

[j] There was a concern that the differences in nomenclature could obscure matches between reported mitotypes. A provable alternative that avoided that problem did not allow us to avoid a significant expenditure of resources in the debate over which nomenclature was more "correct." We hope that this distraction can be avoided in future disasters.

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

mitotype. Mitotyping is most often used to exclude a match but by itself would not normally contain enough information to confirm an identification with a high degree of confidence. Fortunately, we can take advantage of the fact that mitotypes are independent of the STR profiles and use them as additional information to supplement a partial STR match.

Mitochondrial DNA is inherited only through the maternal line. Therefore, many of the samples available to the OCME are useless for mitotype matching. Fathers will not contribute mtDNA to their offspring, and children of male victims (roughly 2/3 of the victims were male) do not have relevant mtDNA for identification. However, as an Operations Research issue, it was found to be more efficient and less expensive to mitotype all kin swabs than to pull out and re-array only those samples that were part of a matrilinear line to a victim.

A proposal was made that mitotype frequency statistics be generated from roughly 5,200 personal effect and kinship samples available. With just a quick review of the database as of September 16, 2002, it was found that at least 24% of the 2,801 victims had no maternally-related kin samples available to the process. This, plus the enormous amount of effort that would be required to confirm the validity of all kin swabs and personal effects, argues that a frequency database from the general population might be preferable to one that could be reasonably and reliably created specifically for WTC victim population.

At the time of this writing SNP data has not been applied to the identification process. The ability to collect data from very short sequences makes this an exciting technology which offers great hope for collecting identifiable information from badly degraded samples. SNPs occur within the human genome on average every 100 to 300 bases and are stable from an evolutionary standpoint, making them easier to follow in populations as they do not change much from generation to generation. A panel of 70 SNPs has been characterized by the GeneScreen division of Orchid Biosciences, all from nuclear DNA. Roughly 2 out of every 3 SNPs involves replacing a C with a T, and it is among these biallelic loci that the 70 forensic SNPs are chosen, specifically ones in which C and T are equally likely.

For kinship analysis, it is important to ensure that these loci are genetically independent of each other and not linked to nuclear STR loci used in identifying the same sample. In an ideal world, all loci would be no less than 50MB or 50cM apart. With 70 SNPs plus up to 15 STR loci in a genome that is only 3.5 billion bases long, this is clearly not possible.

In a study by Dr. Ranajit Chakraborty of the Center for Genome Information, the estimated allele frequencies for the current panel ranged between 15.5% to 81.3%, although the great majority were nearly equal (an average heterozygosity of 46% across the three population groups Caucasian, Black and Hispanic). Fortunately, his study showed the allelic dependence being relatively small, 5.71%, comparable to the 5% expected by chance alone. He further concluded that despite

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

the lack of theoretically independent loci, an analysis of 713 genotypes from 3 different populations support the efficacy of using this 70 SNP panel for identification purposes. With a complete profile of 70 SNPs at probability 1/2, the likelihood of match would be 2^{70}, or approximately 10^{21}. Even with a revised probability average of 46%, the likelihood is 10^{18}.

The authors would reject the proposal that this panel of SNPs should be replaced with one that is more evenly spaced. Because of the limitation of the genome size, linkage in a 70 locus panel is a problem intractable in the laboratory but addressable in software. M-FISys can take an arbitrarily large panel of SNP and STR loci and, given a threshold likelihood value, select those loci that create an unlinked set of STRs and SNPs to reach that threshold in a valid match (assuming the data is available). The remaining loci, of course, would still be reviewed and any directly conflicting data highlighted to indicate experimental error or other anomalies that should be resolved by a forensic biologist before a positive ID is reported.

Combining STR, mtDNA and SNP data is as much a human-computer interaction issue as it is a computational one. Keep in mind that the goal is not for the software to make identifications; as a matter of law, neither the developers nor the software have such authority. Rather it is to present data to a qualified and authorized forensic biologist to make it easier for that person to certify an identification. Instead of developing an interface that combined all STR, mtDNA and SNP data in a single view, M-FISys is broken up into STR-*centric*, mito-*centric* and SNP-*centric* views of the data, with indications that other data supports or contradicts a proposed identification.

Fig. 2. STR-centric view of the Master List displays details of STR results as well as indicators of supporting or conflicting experimental evidence using mitotyping or SNP analysis. Identifying numbers have been changed.

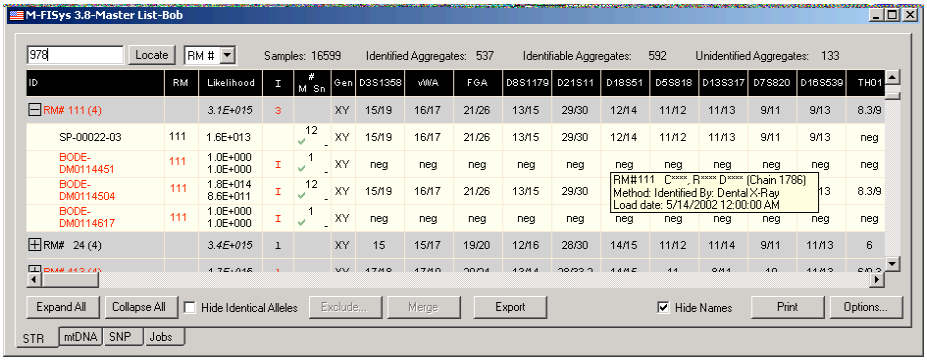


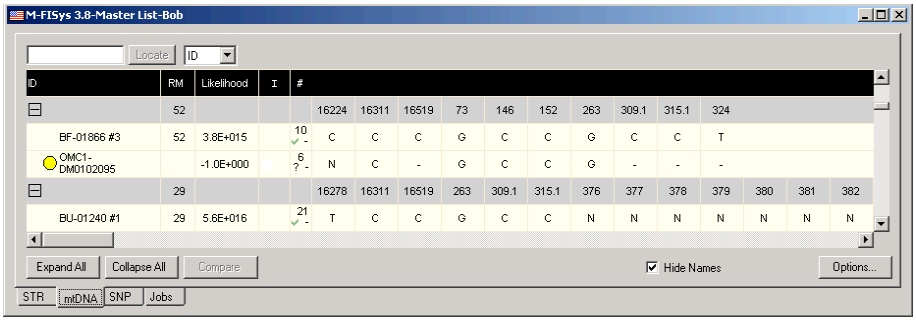
Figure 2 shows the STR-centric view of the master comparison list. The first column shows the names of samples including personal effects (starting with "SP" because they were typed at the State Police lab in Albany) and disaster samples that,

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

following an optional prefix, are coded as DM01nnnnn, where the final 5 digits are the order in which the remains were discovered and checked in at the morgue.[k] The fifth column contains three values. The number at the top is the number of loci (including the gender locus) that yielded results. There is a character on the lower left and the lower right of that column that indicates consistency between the STR match and mito data, and SNP data, respectively. The possible values for mito and SNP data are a green check mark (✓: the data is consistent with the STR results), a red X (X: the data is inconsistent or contradictory), a question mark (?: the data is partially or ambiguously consistent) or a dash (— : this analysis has not yet been performed). Following this is the complete STR profile of the sample.

Similarly, in the mito-centric view of the same data (Figure 3), the operator can see the complete mitotype, plus an indication of whether any STR or SNP data supports or contradicts a mito-based match. This approach allows us to add an arbitrarily large number of different modalities for identification. Individual forensic biologists may be charged with reviewing a particular type of data within their specialty, but are immediately alerted if contradictory data needs to be reviewed.

Fig. 3. The mito-centric view of the Master List displays mitotypes as well as indicators of supporting or refuting STR and SNP data. Identifying numbers have been changed.



The last area of functionality we will review is quality analysis of incoming data. An enormous amount of energy at the OCME is expended to avoid false identifications. While every effort is made to make the process run as quickly and efficiently as possible, the idea of having to tell a family that they have buried the wrong remains is horrifying. Because remains were crushed, fragmented and co-mingled, it is important not only to make sure that the correct remains have been typed, but that no additional remains are being released when a funeral director accepts custody of human tissue from Memorial Park.

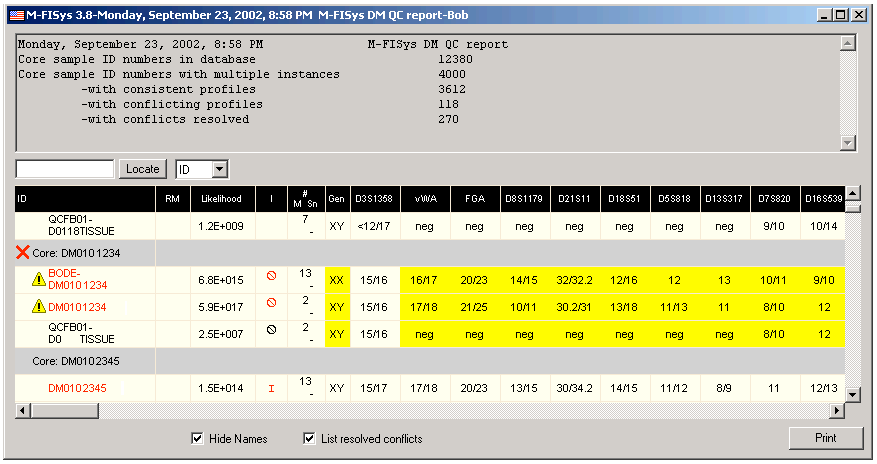
---

[k] The header "DM01" stands for "Disaster: Manhattan, 2001." There are also "DQ01" samples, referring to the "Disaster: Queens" when flight 587 lost part of its tail and crashed in November of the same year.

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

For that reason, the first person to review remains when they arrive at the morgue is not a Medical Examiner, but a Forensic Anthropologist. This person's job is to be as certain as possible that a) the recovered remains are human, and b) that only one person is included in a collection. The policy is that if remains are not physically connected by tissue or sinew, that they should be separated into individual pieces and typed independently.

Fig. 4. Quality Control Report displays experimental results with inconsistent STR profiles. Identifying numbers have been changed.



An example of the kind of QC that M-FISys performs is shown in Figure 4. Sample DM0101234 (not the real number) has been typed twice. In one case, the tissue was typed by Myriad Genetics. But bone from the same sample was sent to Bode Technology Group, a well regarded forensics lab in Virginia with a highly developed capacity for extracting DNA from bone.[1] If you look at the values for each locus, several of them (in yellow) are in conflict. To a forensic biologist, these profiles clearly represent two different people. In fact, they are different genders.

There are several possible resolutions to this conflict. It could be that, upon examining the shipping manifests, it is found that a sample has been mislabeled. This is almost never a consideration, but it still must be an option for resolving the problem within the program. Smaller discrepancies can be the result of experimental error or allelic drop out. In this case, a review of the remains showed that bone from one person was embedded in the muscle tissue of another. Certainly a bizarre circumstance in most medical examiner's offices, but a sadly familiar possibility at

[1] Bone is difficult for several reasons. One issue is that the dust created when a sample is extracted must not contaminate the next sample from the same bench. More significantly, Calcium and other components are PCR inhibitors. Bode Technology Group has a core competency in overcoming these obstacles that has been invaluable in the WTC project.

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

the World Trade Center. The program will prevent these samples from being identified and released until this conflict of data for a single "victim sample" has been resolved.

Another of the several data-integrity tests in M-FISys compares the profiles of personal effects to the swabs taken from family members. Families of missing persons are able to call the "DNA Hotline" at the OCME and find out if enough DNA has been collected for an identification. For instance, a mother whose husband did not come home after the attacks might bring in the presumed victim's toothbrush as well as the oldest child of the victim to give a buccal swab. However, she might ask that the OCME determine if there is enough DNA available from the personal effect to make an identification without additional kin samples because she does not want to further traumatize her younger children by asking them to provide swabs to help identify their father. A call to the DNA hotline might indicate that the toothbrush gave a full profile and that further samples are not required. But what if the wrong toothbrush was brought in? If the toothbrush matches the DNA for the wife or the oldest child, the full profile is not valid and a swab from additional children might be key to making a final identification.

4 Summary

When we were brought onto this project several weeks after the September 11 attacks, we developed five goals and two major deliverables. The goals were:

1. Identify individual remains,
2. Reunify partial remains so that they can be returned to families,
3. Collect and warehouse meta-data for administrative review of reference samples,
4. Track samples among collaborating labs, and
5. Create an information management system to report metrics and make problem resolution proposals to supervisors at the OCME.

Not all aspects of the engineering effort have been discussed in this paper, but we hope a broad overview of both the process and the challenges has been conveyed.

The first deliverable has been to create a mass-fatality identification and recovery system for the OCME, creating installed value to address specific data-handling problem on the fastest possible schedule without running the risk of false identifications. This has been a seven-day-a-week job for over eleven months as of this writing, and all of the participants in the project feel that it is the most important thing they will do in their professional lives. The second deliverable will be to create a generalized and portable version of this tool that can be deployed anywhere in the world where there is a massive human tragedy, be it natural or man made.

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

Acknowledgments

Thanks to the many contributors including the primary process analysis, development and support team: Tracy Beeson, Debra Cash, Peter Fattori, Lucy Hadden, Mike Hennessey, Jeff Ingalls, Anna Khizhnyak, Tom Kubit, Anna Korn, Simon Mercer, Judy Nolan, Gregory Poth, David Relyea, Matt Smith, John Snyder, Francis Sullivan and Bill Wake. Other truly major contributors have included staff at Gene Codes Corporation, Gene Codes Forensics, the New York State Police, AFDIL and the OCME too numerous to name. Special thanks to Chief Medical Examiner Dr. Charles Hirsch who I would follow into war, George Carmody for his always generous technical assistance on short notice, Charles Brenner, Benoit LeClaire, and to the many inspiring families we have met along the way who lost loved ones on September 11, with particular love and admiration for the Cartier, Hannon, Ielpi, Lynch and Weiner families.

References

1. I. Evett et al, *Interpreting DNA Evidence* (Sinauer Associates, 1998).
2. National Research Council, *The Evaluation of Forensic DNA Evidence* (National Academy Press, 1996).
3. National Research Council, *DNA Technology in Forensic Science* (National Academy Press, 1992).
4. J. Ballantyne et al, *DNA Technology and Forensic Science* (Cold Spring Harbor Laboratory Press, 1989).
5. L. Kirby, *DNA Fingerprinting* (Stockton Press, 1990).
6. R. Saferstein, *Forensic Science Handbook -Volume 1* (2nd Edition, Prentice Hall, 1982, 2002).
7. N. Rudin et al, *Forensic DNA Analysis* (2nd Edition, CRC Press, 2002).
8. J. Butler, *Forensic DNA Typing* (Academic Press, 2001).
9. National Institute of Justice, *The Future of Forensic DNA Testing* (2000).
10. Federal Bureau of Investigations, *State DNA Database Statutes* (1999).
11. Federal Bureau of Investigations, *National DNA Index Systems (NDIS) Procedures Manual* (1999).
12. H. Baum, *PCR Statistics: STR13.xls* (personal communication, Office of Chief Medical Examiner, 6/14/99).
13. M. Hennessey, "World Trade Center DNA Identifications: The Administrative Review Process," *13th International Symposium on Human Identification*
14. G. Carmody, *KinTest - CODIS 13 Core Loci* (personal communication, 2001).
15. W. Wake, *Extreme Programming Explored* (Addison-Wesley, 2002)
16. K. Beck and M. Fowler, *Planning Extreme Programming* (Addison-Wesley, 2001)

Pac Symp Biocomput. 2003;:638-53.
PMID: 12603064

17. J. Newkirk and R. Martin, *Extreme Programming in Practice* (Addison-Wesley, 2001)
18. K. Beck, *Extreme Programming Explained* (Addison-Wesley, 2000)
19. R. Jeffries et al, *Extreme Programming Installed* (Addison-Wesley, 2001)
20. S. Anderson et al, *Sequence and Organization of the Human Mitochondrial Genome* (Nature 290: 457-465, 1981).
21. P. Awadalla et al, *Linkage Disequilibrium and Recombination in Hominid Mitochondrial DNA* (Science 286: 2524-2525, 1999).
22. D. Balding and P. Donnelly, *Inferring Identity from DNA Profile Evidence* (Proc. Natl. Acad. Sci. USA 92: 11741-11745, 1995).
23. B. Budowle et al, *Population Data on the Thirteen CODIS Core Short Tandem Repeat Loci in African-American, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians* (J. Forensic Sci. 44: 1277-1286, 1999).
24. A. Jeffreys et al, *Hypervariable Minisattellite Regions in Human DNA* (Nature 314: 67-72, 1985).
25. A. Jeffreys et al, *Individual Specific "Fingerprints" of Human DNA* (Nature 316: 75-79, 1985).
26. H. Cash. "Informatics and Forensics: DNA Data Handling in the World Trade Center Disaster;" Cambridge Healthtech Institute Conference: *DNA Forensics*, June 24-28, 2002
27. RM Andrews et al, *Reanalysis and revision of the Cambridge reference sequence of human mitochondrial DNA* (Nature Genetics 23: 147. 1999).