KINSHIP ANALYSIS AND HUMAN IDENTIFICATION IN MASS DISASTERS: THE USE OF MDKAP FOR THE WORLD TRADE CENTER TRAGEDY

Benoît Leclair[1], Steve Niezgoda[2], George R. Carmody[3] and Robert C. Shaler[4]
[1]*Myriad Genetic Laboratories, Salt Lake City, UT.*
[2]*Technical Contractor, National Institute of Justice, U.S. Department of Justice, Washington D.C.*
[3]*Dept. of Biology, Carleton University, Ottawa, Canada.*
[4]*Forensic Biology, Office of the Chief Medical Examiner, New York City, NY.*

Abstract

The World Trade Center (WTC) attack on Sept. 11 2001 resulted in what is the largest and most complex crime scene in U.S. history to date. The task of recovering and identifying the missing from the smoldering tower debris presented a daunting challenge. Recovered human remains were found to have sustained massive fragmentation and decay, which made DNA typing difficult. The task of matching genotypes through kinship was rendered complex by the very large number of genotypes derived from the next-of-kin and from the victims' remains (including a significant proportion of partial genotypes). Several kinship analysis software programs were adapted to assist the identification effort.

A kinship analysis program, originally developed and used in the aftermath of the Swissair Flight 111 Air Disaster in 1998, was entirely rebuilt for use in the identification of the missing of the WTC tragedy. The program, MDKAP (Mass Disaster Kinship Analysis Program), collapses very large collections of complete / partial STR genotypes derived from the remains down to a restricted number of consensus genotypes believed to reflect the many different victims. Each victim consensus genotype is screened against the data set of genotypes from next-of-kin and personal effects for direct matches to personal effects, for evidence of genetic relatedness to kin through the calculation of Kinship Indices, or for successful production of a parentage trio with any two members of the next-of-kin cohort. Kinship Indices are calculated through a segregated score approach, and by pair-wise likelihood ratio calculations through a linkage to the KinTest software. Matches of victims to parentage trios are obtained by a two-step algorithm that tests all possible parentage trio combinations for any given victim. A positive match to a parentage trio is further cross-checked by the software against other kinship scoring results between the victim and other members of the same family to confirm that the purported family structure including the genotype of the victim is consistent with Mendelian inheritance rules. As computing algorithms are not driven by family information provided with the submitted personal effects or reference samples from next-of-kin, no match leading to an identification can be lost to a mislabeling event or error in family history collection.

The performance and usefulness of DNA typing and bioinformatics tools in mass fatality incidents had been demonstrated before but this report demonstrates the applicability and suitability of this technology for events of much greater scale.

Introduction

The unique ability of DNA typing to derive identity information from any type of tissue makes of DNA typing a valuable victim identification tool for mass fatality incidents (MFI) where severe body fragmentation is encountered (1). The Swissair Flight 111 air disaster was the first high body fragmentation MFI where STR-based DNA typing data was derived from nearly all recovered remains and where kinship analysis data provided identification leads for nearly all victims (1). The very large number of genotype comparisons required to carry out kinship analysis on the large data sets generated by such MFIs makes the use of bioinformatics tools a necessity. Such a tool was successfully developed for the Swissair disaster.

With the level of devastation encountered at the site of the World Trade Center disaster, it was again anticipated that DNA typing data would prove pivotal for the identification of a large number of victims.

The much larger number of victims for WTC (≈ 2,800), when compared to most MFIs involving airliners, the associated large number of relatives (6,000+) and personal effects (5,000+), as well as the anticipated significant decay of remains to be recovered added considerable complexity to the design of any kinship analysis approach to be developed. Major software development efforts were initiated to provide kinship analysis capability to the Office of the Chief Medical Examiner (OCME) of New York, which office was responsible for the identification of recovered remains of WTC site.

Materials and Methods

MDKAP is a re-written version of the kinship analysis software used for the Swissair victim identification initiative (original and re-written code by B. Leclair). The basic algorithm was redesigned to drastically increase processing speed, and the software was supplemented with several additional processing steps to address the requirements dictated by the added complexity of the WTC situation. MDKAP compares each victim's genotype to all genotypes derived from samples collected from next-of-kin, as well as those derived from biological material recovered from personal effects purported to having belonged to the victims. There is no assumption as to the accuracy of the collection information of the submitted reference samples as it pertains to the reported biological relationship of a next-of-kin or the ownership of a personal effect. As such, associations between genotypes are inferred solely on genetic evidence, the accessory sample identification information being verified for consistency only in later steps of the process.

The remains' genotypic data is collapsed by the software to generate groups of remains sharing the same partial / complete genotype. The computing process to which each unique genotype identified within the remains data set is subjected to is divided into four steps:

- the first step consists in identifying, within the interrogated data set of next-of-kin and personal effects, those genotypes that either show a perfect match (i.e. personal effect, identical twin) to the query (remains), or share at least one allele per tested locus as expected from parent:offspring relationships (herein referred to as the Parent:Offspring Criteria, or POC); likelihood ratio data for each pair-wise comparison is provided for parent-child, full sibling and half sibling putative relationships through linkage to the KinTest tool (written by G. Carmody);
- the second step consists in identifying, among the entries that have emerged from step #1 as being significant, those who produce a conclusive parentage trio, that is, all alleles from the purported offspring genotype are accounted for with the two purported parents genotypes;
- third, the sample identification information associated with the samples that have produced a conclusive parentage trio is checked for concordance between the reported biological relationships and those suggested by the successful parentage scenario;
- finally, the reported scores from all other available family member for the victim are checked for scoring consistency within the purported family pedigree.

Figures 1 and 2 show a typical scoring report for MDKAP.

The software will be described in more detail elsewhere.

Results and Discussion

The DNA identification process can be considered as a two step process: 1) the matching of two genotypes, either complete match as expected between the genotypes derived from a personal effect and a remains, or partial as expected between genotypes of a victim and his relatives, and 2) a probability calculation on the reported match. The latter step can be performed through the use of software or manually, but few tools exist to efficiently perform the former. MDKAP was designed to perform the first step in the process. DNAView (2) was also modified to perform a similar function.

Several factors were to impact on the software development rationale of MDKAP:

- first, some inaccuracies were detected in reported biological relationships (usually detected as reported relationships inconsistent with the amelogenin results; same gender inaccuracies can only be detected

during pedigree analysis) for a significant number of next-of-kin, making the sample accessory information not reliable in 100% of cases;
- second, the Swissair experience had demonstrated that for 10% of submitted personal effects, derived genotypes were not those of the anticipated victims (1). Hence, for all the convenience personal effects provide in allowing for a direct genotype match, some personal effects were expected to provide a misleading genotype;
- third, in addition to the massive fragmentation inflicted on the bodies of the victims during the collapse of the Towers, considerable decay was sustained by most recovered remains as a results of decomposition related to the length of the recovery operation, high temperature resulting from fires from within the mountain of tower debris at Ground Zero, and humidity from the water used to put out the fires. This damage translated into large numbers of partial STR DNA typing genotypes;
- fourth, the number of samples proved to be very large, leading to huge numbers of pair-wise genotype comparisons;
- fifth, the unavailability of members of parentage trio in a fair number of situations, precluding the use of parentage trio analysis in those situations.

Ideally, if one could secure the genotypes from personal effects, mother, father, siblings, spouse and several offspring, then identification by DNA typing would be a straightforward process. However, this ideal scenario was encountered in a minority of situations (< 4%), and the availability of at least one complete parentage trio (ascendants or descendants) was encountered in only 50% of situations at WTC. For any given victim, any number of reference genotypes may be unavailable: a parent may be pre-deceased, or the victim may not have any siblings or offspring. Additionally, within the group of reference samples available for a given victim, the usefulness of each reference sample varies with the expected allele sharing of its genotype with that of the victim: personal effects will yield a direct match; mother and father, as well as offspring and spouse will account, when considered as trios, for every allele in a victim's genotype; a mother, father or offspring, taken individually, will share at least one allele per tested locus (POC) on pair-wise comparisons; most siblings will share a higher than average number of alleles, but can nonetheless share few.

Although as many relatives as possible should be collected as personal references in any MFI, there are particular combinations of relatives that are more useful than others in the establishment of identification leads. A direct match between the genotype derived from remains and the one derived from biological traces recovered on a personal effect is the easiest way to derive an identification lead with DNA typing. If possible, a confirmation of ownership of the tested personal effect with consistent kinship analysis data with purported relatives should be sought. In the absence of personal effect holding enough trace biological material, the best alternative is to confirm a parentage trio between the victim and his/her parents, and/or between the victim and his/her spouse and offspring. As it is frequent for the parent of a victim to be pre-deceased, in the absence of offspring, the use of one parent and siblings of the victim might provide strong enough evidence of kinship. A group of relatives including siblings only might prove useful if numbers of shared alleles are high. However, as shown during the Swissair incident (1), situations where siblings are sharing large numbers of alleles are common, so the clustering of numerous siblings from one family against a queried remain might incorrectly suggest kinship when, in fact, the matches are fortuitous.

With MDKAP, pair-wise comparisons between queried and interrogated genotypes were used to derive:
- the number of loci at which at least one allele was found to match (SM; displayed as "SM" / "number of loci with allelic data"),
- the number of loci at which both alleles were found to match (DM),
- the presence of matching rare alleles (V) and,
- for situations where POC is missed by a single locus, whether any two compared alleles at the non-matching locus were ± 1 core repeat away from each other, suggestive of a potential core repeat slip mutation (M).

Entries were sorted on four nested levels (SM, DM, V and M scores, consecutively) which produced a ranking favoring personal effects and next-of-kin of a queried victim, especially next-of-kin involved in a parent:offspring relationship with the victim. This ranking was consistently supported by likelihood ratio values calculated with the KinTest tool. The 45 best scoring entries out of the 11,000+ in the next-of-kin and personal effect data set appear on the full screen scoring report for each queried victim (only a

portion of the report is shown Figures 1 and 2). The SM parameter was found to be an effective kinship measurement tool applicable to every pedigree situation. In the case of parentage trios, each individual member of the trio would be found within the restricted number of genotypes meeting POC, at which point, finding which combination of the POC+ genotypes produce a positive parentage trio is facilitated. Individuals with a possible core repeat slip mutation were included in the POC+ subset. For each next-of-kin entry featured in a scoring report, the SM score of every other member of the same family is listed. Upon encountering a successful parentage trio, the software verifies the SM scores from each other family member to ascertain that all other relatives who should meet POC (parents, offspring) actually do. In any other case not involving a parentage trio, any group of related relatives meeting POC or displaying large numbers of shared alleles with the genotype of the victim would be considered a substantial identification lead.

Figures 1-2 provide an example of a strong identification lead generated by MDKAP. The software highlights in red all entries pointing to an identification lead, and posts (in the Flag column) the results of any additional testing done on each highlighted entry. Three personal effects (prefixed with “SP”, “S2” and “H2H”) show either perfect or near perfect matches to the remains pointing to a single reported missing number RM# (all submitted reference samples related to a given victim are given one unique RM#). One biological son (“BS” prefix), two biological daughters (“BD” prefix) are shown to meet POC (13 SM out of 13 tested loci). The genotype of a spouse (“PR-72356 #01”) with the same RM# as those of the personal effects is shown to produce with the genotype of the victim a consistent parentage trio with all three of her offspring, in that all alleles of each offspring were accounted for in the purported parent’s genotypes. The software indicates this with the three flags in the spouse’s flag box. Each offspring has a matching flag pointing back to their mother. The parent:child likelihood ratios support these conclusions as well. The spouse has a low SM score, as expected from a non-genetically related individual, and should appear much lower in this ranking but this sample’s ranking was altered by the positive parentage trio results and the sample was automatically moved next to its offspring. Another sample related to this family, a brother in this family (“BU-72356 #01”) is highlighted because two or more members of this family scored in the top 45 scores out of 10,899 entries of this data set.

As for the Swissair disaster, half of the victim identification scenarios in WTC do not benefit from the availability of at least one complete parentage trio (at time of writing, the WTC victim identification initiative is still in progress). Identification leads can be drawn from a variety of combinations of available relatives. Not all of them will meet or exceed the necessary random match probability threshold to produce an identification, but the lead often points investigators in the right direction and prompts them to search for additional relatives or personal effects, or get additional data (mtDNA, SNPs). Several examples of these situations will be published elsewhere.

In summary, MDKAP provided the OCME with a bioinformatics tool that performs the following:
- collapses the remains data set, and extracts a list of unique victims’ genotypes;
- makes perfect matches to personal effect;
- identifies next-of-kin related to a victim;
- make matches through parentage trios;
- confirms consistency of family pedigrees;
- excludes nearly all fortuitous hits on the basis of the SM scoring performance of the entire family of the relative that triggered the fortuitous hit;
- makes matches despite the know existence of errors in sample names and reported relationships;
- flags samples with inconsistent reported relationships;
- displays matches in their context, so that “close-calls” are brought to the attention of the data reviewer.

Despite the complexity of the WTC situation, maximal use of the information contained in the genotypic data set was achieved through the use of such tools. Again, DNA typing proved pivotal for the identification of a large number of victims.

Acknowledgments

The authors gratefully acknowledge the members of the WTC Kinship and Data Analysis Panel (KADAP) for their helpful suggestions throughout the course of this work, and the National Institute of Justice for creating and supporting the KADAP initiative.

References:

1) Leclair, B., C.J. Frégeau, K.L. Bowen, Borys, S.B., Elliott J. and R.M. Fourney, 1999. “STR DNA typing and human identification in mass disasters: extending kinship analysis capabilities“. Proceedings of the Eighteenth International Congress of the International Society for Forensic Haemogenetics, San Francisco.

2) DNAView program by Charles Brenner, www.dna-view.com

D3S1358 1	D3S1358 2	vWA 1	vWA 2	FGA 1	FGA 2	AMEL 1	AMEL 2	D8S1179 1	D8S1179 2	D21S11 1	D21S11 2	D18S51 1	D18S51 2	D5S818 1	D5S818 2	D13S317 1	D13S317 2	D7S820 1	D7S820 2	THO1 1	THO1 2	TPOX 1	TPOX 2	CSF1PO 1	CSF1PO 2	D16S539 1	D16S539 2	PentaD 1	PentaD 2	PentaE 1	PentaE 2	Ass. Qs:			
16	16	16	16	18	20	X	Y	13	16	30	30.2	11	19	10	10	8	11	12	12	9.3	9.3	8	11	10	10	12	14					Kinship Index			
																																S	D	V	M
16	16	16	16	18	20	X	Y	13	16	30	30.2	11	19	10	10	8	11	12	12	9.3	9.3	8	11	10	10	12	14					13/13	13	2	
16	16	16	16	18	20	X	Y	13	16	30	30.2	11	19	10	10	8	11	12	12	9.3	9.3	8	11	10	10	12	14					13/13	13	2	
		16	16	18	20	X	Y			30	30.2	11	11	10	10	8	11	12	12	9.3	9.3					12	14	12	13			9/9	8	2	
16	17	16	16	20	23	X	Y	15	16	30	30.2	14	19	10	10	11	11	11	12	7	9.3	8	8	10	12	12	12					13/13	3		
16	17	16	17	18	23	X	X	13	15	30	30	11	14	10	11	8	11	12	12	9	9.3	9	11	10	11	12	14					13/13	3	2	
16	17	16	17	20	23	X	X	13	15	30	33.2	11	14	10	10	11	11	11	12	9	9.3	8	9	10	12	12	12					13/13	1	1	
17	17	16	17	23	24	X	X	15	15	30	33.2	13	14	10	11	11	12	11	12	7	9	8	9	11	12	10	12					7/13			
16	19	14	16	21	21	X	Y	13	13	29	30	11	17	11	13	10	11	10	12	9.3	9.3	8	11	10	12	10	12					11/13	2	1	
14	16	16	18	20	22	X	Y	8	13	30	30	13	14	9	10	11	11	12	12	6	9.3	8	11	10	11	11	13					11/13	2		
14	16	16	19	20	24	X	Y	14	16	29	30	16	16	10	11	11	12	12	13	6	9.3	8	11	11	12	12	12					11/13	1		
13	16	14	17	20	25	X	Y	13	14	30	33.2	15	19	11	12	9	11	10	12	8	9.3	8	11	10	12	12	12					11/13	1		
16	18	16	16	20	25	X	Y	13	14	30	33.2	17	19	12	13	11	11	10	12	8	9	8	8	10	12	12	13					11/13	1		
16	18	16	17	19	20	X	X	13	15	29	30	12	14	12	12	8	12	10	12	9	9.3	8	11	10	12	12	13					11/13	1		
13	16	15	16	20	22	X	Y	13	14	29	30	16	18	10	13	8	11	10	12	6	9.3	8	8	11	12	11	12					11/13	1		
15	17	16	17	18	18	X	X	13	14	29	30	12	13	10	13	11	12	10	12	7	9.3	11	11	10	11	11	12					11/13		1	
14	16	16	17	20	22	X	Y	13	13	30	32.2	18	19	11	13	10	11	10	12	7	9.3	8	9	10	14	11	13					11/13			
16	18	15	17	20	25	X	X	14	14	29	30	11	17	10	13	8	11	8	8	9.3	9.3	8	11	10	10	12	12					10/13	4	1	
16	16	16	16	20	21	X	X	13	13	29	30	15	17	10	11	11	12	12	12	7	9.3	8	11	11	13	8	11					10/13	4		
15	16	14	17	18	20	X	Y	10	16	30	34.2	12	17	12	12	8	11	11	12	8	9.3	8	11	10	11	10	12					10/13	3	1	
17	18	16	16	20	21	X	Y	12	13	29	31	11	12	10	12	10	11	10	12	9.3	9.3	8	11	12	12	10	12					10/13	3	1	
14	16	16	18	18	24	X	Y	15	16	30	30.2	14	16	12	13	8	9	12	12	9.3	10	8	9	11	12	12	14					10/13	3	1	
16	16	16	17	20	21.2	X	X	13	14	29	30	16	18	11	12	11	12	9	9	9.3	9.3	8	11	10	11	11	12					10/13	3		
16	17	17	17	18	20	X	Y	13	15	28	30	16	16	9	12	11	11	10	12	8	9.3	8	11	10	11	8	12					10/13	2	1	
16	17	16	17	24	25	X	Y	13	15	30	31.2	11	16	9	13	11	12	10	10	9.3	9.3	8	11	10	13	11	12					10/13	2	1	
14	16	16	17	18	20	X	Y	12	13	29	32.2	15	18	11	13	8	11	9	12	6	9.3	8	9	10	11	11	12					10/13	2	1	
16	17	16	16	20	27	X	X	13	15	30	30	11	15	11	13	10	11	11	12	9	9	8	10	10	10	11	11					10/13	2	1	
14	15	16	16	20	22	X	X	13	14	30	30	11	14	11	11	11	11	10	11	6	9.3	9	11	10	10	12	13					10/13	2	1	
14	15	16	16	20	22	X	X	13	14	30	30	11	14	11	11	11	11	10	11	6	9.3	9	11	10	10	12	13					10/13	2	1	
16	16	16	17	18	23	X	Y	13	15	30	32.2	12	15	11	12	8	11	11	12	6	6	11	11	10	12	10	12					10/13	2	1	

Figure 1: Typical score report from MDKAP.
Zoomed in view of left portion of a typical score report. All identifiers and numbers on this figure have been altered to protect the privacy of the victims and their families. The right portion of this screen is shown in Figure 2.
The queried genotype appears at the top, matching data set genotypes appear below. Matches are shown with gray background, matching variant alleles are shown with a black background. Samples are scored and sorted according to SM, DM, V and M scores (see text).

Ass. Qs:									
Kinship Index				DM0615660		Displayed if >100			2391V / 10889R
S	D	V	M	Sample ID / Kin Ranking	RM#	PChild	FSib	HSib	Flag
13/13	13	2		SP-00176-17	72356	2.E+12	5.E+17	2.E+09	Perfect match.
13/13	13	2		S2-00176-17	72356	2.E+12	5.E+17	2.E+09	Perfect match.
9/9	8	2		H2H-00176-17	72356	2.E+10	7.E+13	7.E+07	Examine match.
13/13	3			BS-72356 #03; R10 D13 D13 U10	72356	7.E+06	1.E+06	5.E+04	F2RM1R72356
13/13	3	2		BD-72356 #05; R7 D13 S13 U10	72356	2.E+08	4.E+07	9.E+05	F2RM0R72356
13/13	1	1		BD-72356 #02; R7 D13 S13 U10	72356	2.E+05	4.E+03	3.E+03	F2RM1R72356
7/13				PR-72356 #01; D13 D13 S13 U10	72356				F1DM1R72356, F1DM0R72356, F1SM1R72356
11/13	2	1		BS-02927 #3; M10 F3 D9	72715	2.E+04		3.E+02	
11/13	2			BF-02701 #5; M7	70369	2.E+04		2.E+02	
11/13	1			BF-00776 #2; M7	70317	3.E+04		4.E+02	
11/13	1			BU-00777 #1; M7 F10 U8 U11	70505	1.E+03			
11/13	1			BU-70505 #05; M7 F10 U8 U11	70505	1.E+03			
11/13	1			BM-00178 #7; F6 U8 U8	71658	5.E+03			
11/13	1			BF-72675 #02; M8 R6 D7 S8 U8	72675	2.E+03			
11/13		1		BM-72057 #03; R9 U9 U7 U7	72057	2.E+04		3.E+02	
11/13				BU-71677 #13; M9 F6 R6 R7 D7 S5 S7 U6 U6	71677	4.E+02			
10/13	4	1		BU-01197 #1	71527	2.E+06	1.E+04	5.E+03	
10/13	4			BM-02927 #9; F3 D9 S11	72715	8.E+04	5.E+02	3.E+02	
10/13	3	1		BS-70251 #03; M6 R9 D9	70251	5.E+04	5.E+02	3.E+02	
10/13	3	1		BS-71777 #03; R8 D8 S8	71777	2.E+05	5.E+02	9.E+02	
10/13	3	1		BU-72356 #06; R7 D13 D13 S13	72356	1.E+07	3.E+05	2.E+04	
10/13	3			BM-70630 #06; R6 D6 S9 U7 U6 U7	70630	2.E+03			
10/13	2	1		BU-00817 #1; M8 F6	70379	5.E+03			
10/13	2	1		BF-70667 #02; M7 U9	70667	6.E+03			
10/13	2	1		BF-00708 #2; U8	70967	3.E+03			
10/13	2	1		BU-71036 #02; M7 U8 U8	71036	7.E+03			
10/13	2	1		BM-00027 #7; M10 R8 U10	71379	1.E+04			
10/13	2	1		BM-71379 #01; M10 R8 U10	71379	1.E+04			
10/13	2	1		BF-01817 #6; M7	72770	1.E+04			

Figure 2: Typical score report from MDKAP.
Zoomed in view of right portion of a typical score report. All identifiers and numbers on this figure have been altered to protect the privacy of the victims and their families. The left portion of this screen is shown in Figure 1.
Samples are scored and sorted according to SM, DM, V and M scores (see text). The prefix legend is the following: BM = Biological Mother, BF = Biological Father, BD = Biological Daughter, BS = Biological Son, BU = Biological Sibling, PR = Undefined Personal Relationship, SP or S2 or H2H = personal effect.
The software highlights entries of interest in red. The kin ranking is listed next to the sample ID: the single letter prefix in this box (i.e. "D") refers to the second letter of the referred sample's name (i.e. "B**D**"), the number that follows is the SM score against the victim of that other relative. For a relative meeting POC ("D" (daughter), "S" (son), "M" (mother) or "F" (father)), this number should be 13. The RM# indicates the reported missing number assigned to the victim for which the reference sample was submitted. The PChild, FSib and HSib columns provide the pair-wise likelihood ratios. The Flag column provides the results of any additional test carried out on the entry. For example, the positive parentage trio flag "F2RM1R72356" for the biological son of this victim indicates that this son is being considered as an offspring (**F2**RM1R72356) in this productive parentage trio scenario, that the other living next-of-kin party to this trio is the spouse (F2**R**M1R72356) of the victim assigned RM# 72356 (F2RM1**R72356**), and that one core repeat mutation was encountered (F2R**M1**R72356) in this trio. Conversely, the mother (PR-72356 #01) carries a "F1SM1R72356" flag, indicating she is considered a parent (**F1**SM1R72356) in this productive scenario flag, and that the other living next-of-kin party in this trio is a son (F1**S**M1R72356) of the victim assigned RM# 72356 (F1SM1**R72356**), and that one core repeat mutation was encountered (F1S**M1**R72356) in this trio. Therefore, any pair of flags can easily be located in a score report. Any

discrepancy between a reported relationship in a trio and that deduced from genetic data would cause for question marks to be added to the end of the flag (i.e. “F1SM1R72356**??**”), drawing the attention of the data reviewer to the anomaly.