V. Zvénigorosky, S. A. Fedorova, C. Hollard, A. Gonzalez,
A. N. Alexeev, R. I. Bravina, E. Crubézy, C. Keyser

A NEW APPROACH TO GENETIC KINSHIP TESTING IN YAKUT ARCHAEOLOGY

ABSTRACT

For fifteen years, part of the work of our research team has been focused on the study of parental links between individuals living hundreds or thousands of years ago, whose remains have been found in single graves or large funerary complexes. These studies have been undertaken using methods developed by forensic genetics to identify individuals, mainly based on the genotyping of autosomal STR (Short Tandem Repeats). Issues arose from this work, namely the limits of studying small numbers of subjects, originating from groups of finite sizes where kinships cannot be inferred a priori and for which reference allelic frequencies do not exist. Although ideal human populations are rare when undertaking such studies, the Yakuts of Eastern Siberia constitute a very advantageous model, with large numbers of small pastoral communities and well-preserved archaeological material. The study of kinship in the ancient Yakuts allowed us to highlight the difficulties in analysing genetic data from small ancient human groups and to develop a strategy to improve the accuracy of statistical computations. This work describes this strategy and possible solutions to the study of populations outside of the frame of reference of global meta-populations, due either to isolation, remoteness or antiquity.

Keywords: Ancient DNA – Genetic Kinship – Population Genetics – Short Tandem Repeats

INTRODUCTION

Forensic methods have already reliably identified kinships in ancient human populations, in isolated graves (1,2), as well as funerary complexes (3,4).

Archaeological digs in Yakutia during more than ten years have exposed both single and multiple graves, either isolated or part of larger complexes. The first study applying forensic kinship methods to this data showed that some graves contained the bodies of seemingly closely related individuals (5), often parents and their children or direct siblings. In some cases, however, the occupants of a tomb appeared to be unrelated, or at least not definite close relatives. Others showed ambiguous results, where for example some tests on autosomal STR (Short Tandem Repeat) data would show clear values when studying pairs of subjects and unclear values when studying trios including the same pair. A multiple marker approach made it possible to resolve certain ambiguities, by excluding relationships based on incompatible paternal or maternal lineages, where those were relevant. Some issues however, remained undecided.

In previous studies, autosomal STR had been studied at 15 and 21 loci, while Y-chromosomal STR had been studied at 17 loci. For this study, we present the continued analysis of 15 and 21 autosomal STR loci and a pair of ancient related individuals was analysed at 83 SNPs (Single Nucleotide Polymorphisms), providing new data, independently from already obtained STR genotypes.

This work highlights the two main issues facing the study of kinship in ancient human populations from a statistics standpoint. The first question is that of the efficacy of forensic methods in remote groups, especially in resolving complex or second-degree kinship cases. Although it is advised to use closely related populations as references when a group does not belong to a larger reference population, some human groups are too remote to allow this without great approximation. The second question is the resolution power of new techniques that allow for greater numbers of markers to be analysed. Understanding the scope and reach of kinship testing methods will permit the establishment of tests and standards that supply satisfactory answers that both quantify the quality of kinship calls (going from the qualitative suggestion to the probability of specific kinship) and provide ways to study finer, second-degree kinships, while retaining statistical significance and introducing or identifying the effects of demographic events and population history.

MATERIALS AND METHODS

(1) Samples

Data was available for 128 ancient individuals (5) from four localities in Yakutia, respectively 76 individuals originating from Central Yakutia (the region of Yakutsk), 21 from the basin of the river Villuy (West of Yakutsk), 24 from the region of Verkhoyansk (North of Yakutsk) and 7 from the basin of the river Indigirka (East of Yakutsk).

(2) STR analysis

All 128 ancient individuals had been analysed with regards to 15 STR loci with the AmpFLSTR® Identifiler® Plus kit (Life Technologies™) (5). The SNP (Single Nucleotide Polymorphism) typing protocol is detailed in the Supplementary Information (Materials and Methods 1).

All STR products were run on the 3100 or 3500 genetic analyser (Life Technologies™) and analysed using GeneMapper v. 4.1 (Life Technologies™).

(3) Single Nucleotide Polymorphism (SNP) typing protocol in the Ieralaakh pair

DNA was extracted during the work realized by Keyser et al. 2015 (5).

(a) Library preparation

DNA libraries were constructed using the Ion AmpliSeq™ Library kit 2.0 (Life technologies) and the HID-Ion AmpliSeq™ Identity Panel (Life Technologies). The DNA input was 0.17ng for Ieralaakh 1, 1ng for Ieralaakh 2 and the PCR conditions were those recommended by the manufacturer (Ion AmpliSeq™ Library Preparation for Human Identification Applications Rev. A.0). After partial primer digestion, all libraries were barcoded using the Ion Xpress™ Barcode Adapters kit (Life Technologies) and purified with Agencourt AMPure XP system (Beckman Coulter). Libraries were quantified with the Ion Library Quantitation kit (Life Technologies).

(b) Template preparation and sequencing

Template preparation was done according to the manufacturer's protocol. The emulsion PCR (emPCR) was performed on the Ion OneTouch™

2 Instrument (Life Technologies) using Ion PGM™ Template OT2 200 kit (Life Technologies). Percentage of positive Ion Sphere Particles (ISP) after emPCR was measured with the IonSphere™ Quality Control Kit (Life technologies) on the Qubit® 2.0 fluorometer (Invitrogen). The emPCR products were then enriched on the Ion OneTouch™ Enrichment System (Life Technologies) using Ion PGM™ Enrichment Beads (Life Technologies) and the Ion PGM™ Template OT2 200 kit (Life Technologies). Sequencing was done on the Ion PGM™ system using the Ion PGM™ Sequencing 200 Kit v2 and the Ion 314™ v2 chip according to the manufacturer's protocol.

(c) NGS data analysis

Sequence analyses were performed with Torrent Suite™ 4.2.1 and the HID SNP Genotyper plugin v4.3 (Life Technologies). Integrative Genomics Viewer (IGV) (6) was used to examine each sequence.

(d) Results

For Ieralaakh 1 and Ieralaakh 2 DNA extracts, 72k and 56k reads were obtained respectively from the NGS run. SNP coverage varied from 25x to 1829x (with an average of 427x) for the Ieralaakh 1 DNA extract and from 9x to 4263x (with an average of 457x) for the Ieralaakh 2 DNA extract. For Ieralaakh 1, no SNP position had a sequencing depth under 20x; for Ieralaakh 2, five positions (rs729172, rs993934, rs826472, rs722290, rs12997453), were deleted for the analysis. MAF< 20% = rs1031825 deleted for Ieralaakh 1 and Ieralaakh 2. Finally, strand bias was analysed: rs430046 deleted for the 2 individuals.

Thus, 88 SNPs were successfully typed for Ieralaakh 1 and 83 for Ieralaakh 2.

(4) Computer software and test parameters

Allelic frequencies and diversities, as well as statistical tests were performed using the Genetix (7), MLrelate (8) and Arlequin (9) software.

Likelihood Ratios (LR) were computed on each pair for each relationship category against the likelihood that the individuals were unrelated, using the Familias software (10).

The three metrics that were not dependent on allelic frequencies were measured using the R language (11). Exclusions were accounted for by direct count and Relatedness was computed as a simple proportion of similarity between two genotypes. Identity-by-Descent (IBD) probability was computed using Identity-by-State (IBS) (12,13).

The relationship categories were Parent-Offspring or PO, Full-Sibling or FS, Half-Sibling or HS, Avuncular or AV (Uncle/Aunt-Nephew/Niece), Grandparent-Grandchild or GC, first Cousin or CO, and Unrelated or U. "Unrelated" was the default level of kinship in tests that required one. Thus "LR-PO" is the Ratio of the Likelihood that two individuals are a parent and his child against the Likelihood that they are unrelated.

RESULTS

CONSTITUTION OF REFERENCE POPULATIONS

The first step in constituting unrelated sets of individuals was the elimination of putative parental relationships identified by the MLrelate software, along with those examined in previous studies. Only Parent-Offspring (PO) and Full-Sibling (FS) relationships were considered accurate calls in MLrelate, due to the artefactual nature of the great majority Half-Sibling (HS) calls using this software.

At least one individual was eliminated from the reference frequency set for each putative kinship. However, many individuals were included in more than one relationship, and thus implied the existence of more distant relationships between other individuals. For the purposes of this study, the parental relationships considered grounds for eliminating subjects from the frequency set were Parent-Offspring, Full-Siblings, Half-Siblings, Avuncular, Grandparent-Grandchild and First Cousins.

Based on these criteria, 95 individuals (out of 128) from the ancient Yakut sample were considered unrelated and were included in the computation of allelic frequencies, as well as subsequent statistical tests.

Kinship results and ambiguities

(a) Kinship tests in the "Lepsei Family" using 15 autosomal STR loci

Kinship was analysed within a group of 4 ancient subjects that implied 6 parental relationships whose genealogy had been proposed in a previous study (5).

Because of the way it was built, IBD (Identity-by-Descent) called PO between two infants who showed no exclusions but could not be a parent and his child due to very young age at death. Relatedness, although it could not exclude any relationship conclusively, was only incoherent with other methods in calling a supposed PO kinship as an HS/AV/GC (Half-Sibling, Uncle/Aunt-Nephew/Niece, Grandparent-Grandchild), despite a maternal line fit and the absence of exclusions. Even though LR (Likelihood Ratio) tests gave entirely coherent results, LR values for PO and FS were very similar for the aforementioned infant pair (FS less than five times higher than PO) and distinction between LR-AV and LR-CO was also very weak for two uncle-nephew pairs (respectively less than 4 times higher and less than 2 times higher).

Although some LR values were not significantly distinct from one another and Relatedness and IBD tests each provided one incoherent call, the four methods overall concurred to confirm the proposed genealogy (5) and ambiguities were satisfactorily resolved.

(b) Kinship tests in the "Shamanic Tree Family" using 15 autosomal STR loci

The study of the "Shamanic Tree Family" (Figure 1) revealed kinship ambiguities (Table 1). Five pairs of subjects showed no exclusions, while none showed more that 4 (which is lower than any excluding value at 95% for any relationship category except PO). Relatedness proved a very unstable metric using 15 loci, with at least a third of calls not coherent with one another, as well as PO called in the presence of exclusions and FS called when paternal lineages were incompatible.

IBD called 2 relationships incorrectly (FS with paternal exclusions) and was ambiguous regarding the nature of kinship between three children ST-4, ST-5 and ST-II, identifying the three young boys as brothers, with equivocal calls of FS or HS/AV/GC for the pair ST-4/ST-II.

LR tests also gave problematic results in calling an FS relationship (ST-1/ST-4) where paternal lineages were incompatible and a CO relationship (ST-3/ST-II) where a GC relationship was implied by the study of kinship across generations (ST-3 is the mother of ST-2 who is the mother of ST-II). More significantly, it gave calls for the ST-4, ST-5, ST-II trio that were not logically consistent, making ST-5 the full brother of both ST-4 and ST-II, with ST-II the half-brother of ST-4 and the full brother of ST-5.

With the man ST-1 carrying the Y-chromosomal haplotype the most commonly found in the Yakuts (both modern and ancient) and the three boys ST-4, ST-5 and ST-II being exclusive members of a slightly different haplotype group, it seemed that direct parental or

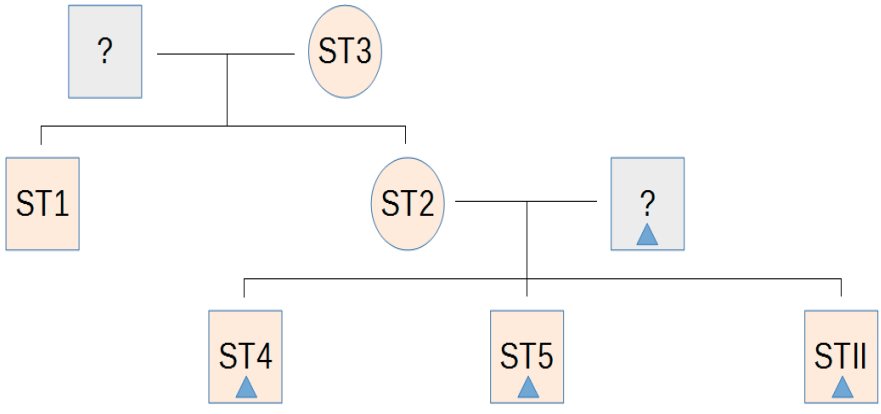

Figure 1. Shamanic Tree genealogy First Hypothesis

fraternal kinship were excluded between the adult and the three children.

(c) Kinship tests in the "Ieralaakh" pair using autosomal STR and autosomal SNP

The "Ieralaakh" pair had been classified as a possible PO relationship on the basis of 15 autosomal STR loci, using LR tests based on allelic frequencies in the ancient dataset, and corrected as an FS relationship using indel typing (5, 16). LR-FS was always higher than LR-PO, even when no exclusions were observed on 15 loci, however, the difference between the two values was less than twofold. Relatedness and IBD tests computed on 15 loci respectively called FS and PO (the present IBD test calls PO in the absence of exclusions).

The study of 21 autosomal STR loci revealed 2 exclusions that eliminated the possibility of a PO relationship (Table 2) across all metrics. Test values were however again computed on the 19 loci that showed no exclusions to identify possible discrepancies. Using 19 loci, LR-PO is higher than LR-FS and only Relatedness calls FS.

All methods analysing autosomal STR converged towards an FS relationship, without however providing clear superiority of an objective metric over another, given that LR must always be understood as a qualitative metric, not a quantitative one.

To improve on these results, 83 autosomal SNP were typed in the "Ieralaakh" pair. Only 2 exclusions were revealed, pointing once again to an FS relationship. LR was not computed in the absence of an ancient Yakut reference set of allelic frequencies but IBD pointed towards FS and more importantly Relatedness pointed towards FS, providing decisive exclusion of all levels of kinship except FS and PO. Computations of the probability of each more distant relationship (HS, AV, GC and CO) compared to the probability of FS (Figure 2) showed that the probability of FS was about 4 times that of HS using 15 STR loci, 20 times using 21 STR loci and 50 times using 83 SNP loci. Although PO is not eliminated by Relatedness (it is eliminated by allelic exclusions), when the mean value for Relatedness of all kinship levels is compared to the value observed in the case of "Ieralaakh" for 83 SNP (Figure 2) it shows that FS is 9 times more likely than PO.

DISCUSSION

(1) "Lepsei", an unambiguous result

The "Lepsei Family" is an example of forensic kinship assessment methods successfully providing a coherent genealogy for a group of ancient individuals. However, this case also highlights two of the main issues that arise from the study of 15 STR loci: an absence of exclusions is sometimes not the sign of a PO relationship but of an FS relationship and Relatedness values, that can be studied in theoretical models outside of a reference population, are poorly efficient on small numbers of markers.

Unknowing the possible level of inbreeding of the population or the validity of mutation rates, the results given by such studies as this one must be understood as indications rather than assessments.

(2) "Shamanic Tree", uncertain second-degree relationships

Many difficulties that are integral to the study of precise kinship in ancient populations arise in the analysis of parental relationships in the Shamanic Tree Family. Relatedness is once more

Table 1

Results of kinship tests in the Shamanic Tree Family using 15 autosomal STR loci

A priori relationship	Subject 1	Subject 2	Exclusions	top Relateness relationship	top IBD relationship	top LR relationship	Y-DNA fit	Mt-DNA fit	Relationship call
ПС	ST2	ST1	2	~FS	FS	FS	i	yes	FS
PP	ST2	ST4	0	FS	PO	PO	i	yes	PO
PP	ST2	ST5	0	HS/AV/GC	PO	PO	i	yes	PO
PP	ST2	STII	0	HS/AV/GC	PO	PO	i	yes	PO
PP	ST3	ST1	0	FS	PO	PO	i	yes	PO
PP	ST3	ST2	0	PO	PO	PO	i	yes	PO
ПС	ST4	ST5	1	FS	FS	FS	yes	yes	FS
ПС	STII	ST4	2	HS/AV/GC	FS/HS/AV/GC	HS/AV/GC	yes	yes	HS/FS
ПС	STII	ST5	4	~FS	FS	FS	yes	yes	HS/FS
ДТП	ST1	ST4	1	~FS	FS	FS	no	yes	AV
ДТП	ST1	ST5	2	HS/AV/GC	FS	HS/AV/GC	no	yes	AV
ДТП	ST1	STII	3	HS/AV/GC	HS/AV/GC	HS/AV/GC	no	yes	AV
ДБВ	ST3	ST4	4	HS/AV/GC	HS/AV/GC	HS/AV/GC	i	yes	GC
ДБВ	ST3	ST5	3	HS/AV/GC	HS/AV/GC	HS/AV/GC	i	yes	GC
ДБВ	ST3	STII	4	CO	HS/AV/GC	CO	i	yes	GC

Table 2

Results of kinship tests in the Ieralaakh pair

A priori relationship	Subjects	Markers	top LR relationship	top IBD relationship	top Relateness relationship	Exclusions	Relationship call	
none	Ieralaakh 1	Ieralaakh 2	15 aSTR	FS	PO	FS	0	FS
			19 aSTR	PO	PO	FS	0	PO
			21 aSTR	FS	FS	FS	2	FS
			83 aSNP	-	FS	FS	2	FS

shown to have poor discriminatory power on only 15 STR loci, while LR tests, that are dependent on allelic frequencies, do not always provide a coherent genealogy. IBD tests vary differently but with the same drawbacks of sometimes not supplying clear cut answers, or even coherent ones.

For a genealogy to be constructed, at least one kinship call must be discarded as false. The three children of ST-2 were either full brothers, or ST-II was the half-brother of the other boys. It must however be posited that, in the second solution (Figure 3), ST-II's father was a close relative of ST-4 and ST-5's father, as they shared the same exclusive paternal line (5). Moreover, to construct a genealogy, all calls between uncle ST-1 and nephew ST-4 must be considered overestimations of kinship (FS is in fact HS/AV/GC), while half the calls between grandmother ST-3 and grandchild ST-II must be considered underestimations of kinship (CO is in fact HS/AV/GC). This genealogy has anthropological implications and cultural studies of the Yakut might shed light on it.

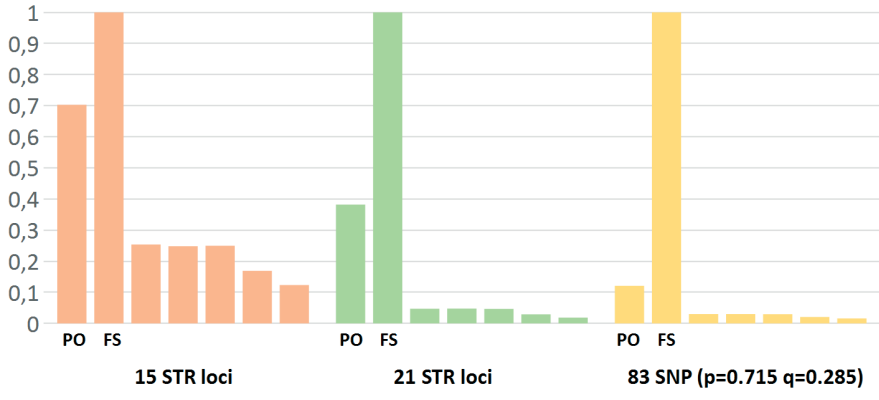

Figure 2. Leralaakh compared probabilities

(3) "Ieralaakh", towards a statistically decisive result and finer levels of kinship

The case of the "Ieralaakh" pair shows how Relatedness can in some cases eliminate possible relationship reliably (with a 5% risk of error). While LR varies according to allelic frequencies, which themselves are subject to the influence of numerous phenomena, the discriminative power of Relatedness steadily increases (given markers with comparable variability) with the augmentation of the number of markers. Moreover, LR results must be interpreted within the context of a larger framework, constituted by the large meta-population in which the subjects can be included. Objective metrics such as Relatedness, Exclusions or IBD can be used in isolation, relying on statistical models that permit the quantification of the probability that a result is accurate or otherwise.

CONCLUSION

The study of kinship using ancient DNA from a remote population meets with specific issues that arise from its isolation. Ultimately, it is not the degraded nature of genetic material that constitutes an obstacle (especially for Yakut burials, that benefit from exceptional conditions of conservation) but the unavailability of a large meta-population in which test values can be understood, and specifically, reliable allelic frequencies computed.

When interested solely in Parent-Offspring or Full-Sibling relationships, Likelihood Ratios based on STR profiles and frequencies are usually efficient, even when based on tentative reference data. However, a direct count of the number of excluding loci is sometimes as precise as those computations in calling kinship. The analysis remains a qualitative one, with similar profiles classified as "Parent-Offspring" when no exclusions can be observed, "Full-Siblings" when only a few exclusions are observed, and "other" or "Unrelated" when too many exclusions are observed (or when other data excludes PO or FS). The subtle distinction between second-degree relationships is made difficult when the approximation intrinsic to the allelic frequencies (computed on small or ill-defined populations) throws doubt on all test values that are somewhat indistinct.

The future of precise kinship studies in ancient populations resides in the multiplication of markers and the creation of models that account for cultural and social tendencies, that directly influence the genetics of the human group. It is also now possible to envisage the study of very precise levels of kinship, where

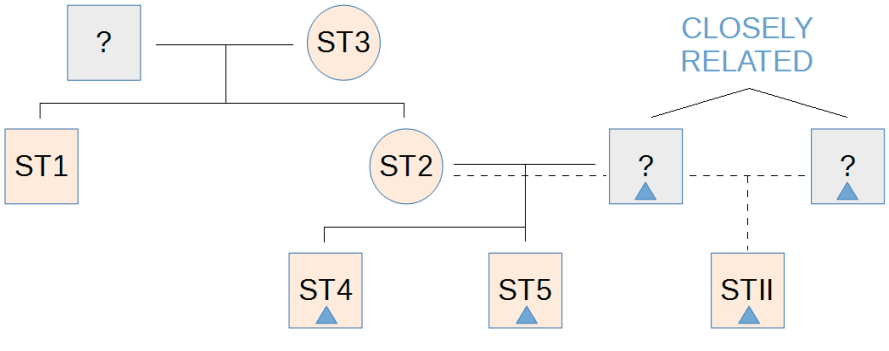

Figure 3. Shamanic Tree genealogy Alternative Hypothesis

for example the difference between siblings whose parents are unrelated and siblings whose parents are first cousins can be accurately described theoretically and observed practically. The progress to sequencing chips that gather data at millions of SNP loci, combined with the use of kinship tests based on objective similarity values, carefully calibrated for a specific social structure, will allow for finer and more accurate assessments of ancient kinship.

ACKNOWLEDGEMENTS

This work was supported by the French Archaeological Missions in Oriental Siberia (Ministry of Foreign and European Affairs, France), Project №6.1766.2017/4.6 of the Ministry of Education and Science of Russia, the Human Adaptation programme of the French Polar Institute Paul Emile Victor and the French National Agency of Research (ANR jcjc-0115 «Sibérie»).

REFERENCES

1. I. Clisson et al., « Genetic analysis of human remains from a double inhumation in a frozen kurgan in Kazakhstan (Berel site, Early 3rd Century BC) », *Int J Legal Med*, 2002.

2. Fanny Mendisco et al., « Genetic Diversity of a Late Prehispanic Group of the Quebrada de Humahuaca, Northwestern Argentina: DNA from Ancient Northwestern Argentineans », *Annals of Human Genetics* 78, nº 5 (September 2014): 367-80, doi:10.1111/ahg.12075.

3. Christine Keyser-Tracqui, Eric Crubézy, et Bertrand Ludes, « Nuclear and Mitochondrial DNA Analysis of a 2,000-Year-Old Necropolis in the Egyin Gol Valley of Mongolia », *The American Journal of Human Genetics* 73, nº 2 (August 2003): 247-60, doi:10.1086/377005.

4. M. Lacan et al., « Ancient DNA Reveals Male Diffusion through the Neolithic Mediterranean Route », *Proceedings of the National Academy of Sciences* 108, nº 24 (June 2011): 9788-91, doi:10.1073/pnas.1100723108.

5. Christine Keyser et al., « The ancient Yakuts: a population genetic enigma », *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 370, nº 1660 (December 2014), doi:10.1098/rstb.2013.0385.

6. James T Robinson et al., « Integrative genomics viewer », Nature Biotechnology 29, no 1 (January 2011): 24 26, doi:10.1038/nbt.1754.

7. K. Belkhir et al., « GENETIX 4.0. 5.2., Software under Windows™ for the genetics of the populations », 2004.

8. Steven T. Kalinowski, Aaron P. Wagner, et Mark L. Taper, « Ml-Relate: A Computer Program for Maximum Likelihood Estimation of Relatedness and Relationship », *Molecular Ecology Notes* 6, nº 2 (June 2006): 576-79, doi:10.1111/j.1471-8286.2006.01256.x.

9. Laurent Excoffier et Heidi E. L. Lischer, « Arlequin Suite Ver 3.5: A New Series of Programs to Perform Population Genetics Analyses under Linux and Windows », *Molecular Ecology Resources* 10, nº 3 (May 2010): 564-67, doi:10.1111/j.1755-0998.2010.02847.x.

10. Daniel Kling, Andreas O. Tillmar, et Thore Egeland, « Familias 3 – Extensions and New Functionality », Forensic Science International: Genetics 13 (November 2014): 121-27, doi:10.1016/j.fsigen.2014.07.004.

11. R core Team, « R: A language and environment for statistical computing », 2014, http://www.R-project.org/.

12. Albert Jacquard et André Chaventré, *Génétique des populations humaines* (Presses universitaires de France, 1974).

13. Michael S. Blouin, « DNA-Based Methods for Pedigree Reconstruction and Kinship Analysis in Natural Populations », *Trends in Ecology & Evolution* 18, nº 10 (October 2003): 503-11, doi:10.1016/S0169-5347(03)00225-8.

14. C. Hollard et al., « First Application of the Investigator DIPplex Indels Typing Kit for the Analysis of Ancient DNA Samples », *Forensic Science International: Genetics Supplement Series* 3, nº 1 (December 2011): e393-94, doi:10.1016/j.fsigss.2011.09.058.

Corresponding author:

Vincent Zvénigorosky, Institut de Médecine Légale, 11 rue Humann, 67000 Strasbourg, France

z.vincent@live.fr / zvenigorosky@unistra.fr

Tel.: +33(0)3.68.85.33.48

Author coordinates:

Sardana A. Fedorova – sardaanafedorova@mail.ru

Laboratory of Molecular Genetics, Yakut Research Center of Complex Medical Problems, Yakutsk, Sakha Republic, Russia , Laboratory of Molecular Biology, North-Eastern Federal University, Sakha Republic, Russia

Clémence Hollard – hollard@unistra.fr

Institut de Médecine Légale, Université de Strasbourg, Strasbourg, France

Tel.: +33(0)3.68.85.33.48

Angéla Gonzalez – agonzalezmartin@unistra.fr

Institut de Médecine Légale, Université de Strasbourg, Strasbourg, France

Tel.: +33(0)3.68.85.33.48

Anatoly N. Alexeev – secretar@igi.ysn.ru

Institut des sciences humaines et des problèmes des peuples minoritaires du Nord, Branche Sibérienne de l'Académie des sciences de Russie (Yakoutsk)

Rosa I. Bravina – bravinari@bk.ru

Institut des sciences humaines et des problèmes des peuples minoritaires du Nord, Branche Sibérienne de l'Académie des sciences de Russie (Yakoutsk)

Eric Crubézy – Crubezy.eric@free.fr

Laboratoire AMIS, CNRS UMR 5288, Université de Toulouse, Toulouse, France

Christine Keyser – ckeyser@unistra.fr

Institut de Médecine Légale, Université de Strasbourg, Strasbourg, France

Laboratoire AMIS, CNRS UMR 5288, Université de Toulouse, Toulouse, France

Tel.: +33(0)3.68.85.33.48