Estimating genetic relatedness from ancient DNA

Divyaratan Popli, Stéphane Peyrégne, and Benjamin M. Peter

Zusammenfassung

Die Schätzung genetischer Verwandtschaft anhand von alter DNA

Die Analyse von genetischer Verwandtschaft prähistorischer Individuen kann einen Einblick in ihre soziale Hierarchie und Kultur ermöglichen. Dies ist auch wichtig für die nachgelagerten Genetikanalysen. Jedoch ist es schwierig, Verwandtschaft anhand alter DNA mit unterschiedlicher Herkunft wegen des niedrigen Sequenzumfangs und der Kontamination der DNA herzuleiten. Wir stellen hier KIN (Kinship INference) vor, eine Methode, die auf die Verwandtschaft eines Individuenpaares schlussfolgert, indem sie die Segmente, die identisch nach Abstammung sind und durch beide aufgewiesen werden, identifiziert. KIN kann unter der Verwendung von ≥ 0.05x Sequenzumfang akkurat Verwandte bis zum 3. Grad bestimmen und Eltern-Kind-Paare von Geschwistern differenzieren. Wir liefern zusätzliche Modelle, um DNA-Kontaminationen zu korrigieren sowie um Inzucht aufzudecken und so die Klassifikationsgenauigkeit zu verbessern.

Summary

Analyzing genetic relatedness of ancient individuals may provide a glimpse into their social hierarchy and culture, and is important for downstream genetic analyses. However, inferring relatedness from ancient DNA is a difficult problem due to low sequence coverage and DNA contamination from various sources. We present KIN (Kinship INference), a method that infers the relatedness of a pair of individuals by identifying the identical-by-descent segments that they share. KIN can accurately classify up to 3rd-degree relatives using ≥ 0.05x sequence coverage and can differentiate parent-child pairs from siblings. We provide additional models to correct for DNA contamination and detect inbreeding, improving classification accuracy.

READ (Monroy Kuhn et al. 2018) and lcMLkin (Lipatov et al. 2015) are the most commonly used software algorithms that infer relatedness with low-coverage ancient DNA data. READ is able to identify up to 2nd-degree relatives with a sequence coverage as low as 0.05x, but does not distinguish between siblings and parent-child pairs. In comparison, lcMLkin can also classify higher degrees of relatedness, while distinguishing between parent-child and siblings, but requires at least 2x sequence coverage to work efficiently (Tab. 1). There are additional challenges that these methods do not address, which affect relatedness estimates, such as DNA contamination from modern, living people and the presence of long runs of homozygosity (ROH) due to inbreeding.

Here, we present a method, KIN, which uses Hidden Markov Models (HMMs) to infer genetic relatedness between a pair of individuals. The method identifies the parts of the genomes that they share by descent and takes contamination and ROH in each individual into account (Popli et al. 2023). These shared parts of the genome are referred to as ›Identical By Descent‹ (IBD) segments. KIN consists of ten different HMMs, each modeling a different genomic distribution of IBD segments given a particular case of relatedness. We considered the following ten cases: Unrelated, 5th degree, 4th degree, 3rd degree, grandparent-grandchild, avuncular, half-siblings, parent-child, siblings, and identical. We then identify the best fitting case of relatedness for a pair of indi-

Feature	READ	lcMLkin	KIN
Coverage	≥ 0.05x	≥ 2x	≥ 0.05x
Degrees of relatedness	Up to 2nd degree	Up to 5th degree	Up to 3rd degree
Differentiation between siblings and parent-child	No	Yes	Yes
Accounts for ROH (runs of homozygosity)	No	No	Yes
Contamination Correction	No	No	Yes

Tab. 1 Comparison of READ, lcMLkin and KIN.

Tab. 1 *Vergleich von READ, lcMLkin und KIN.*

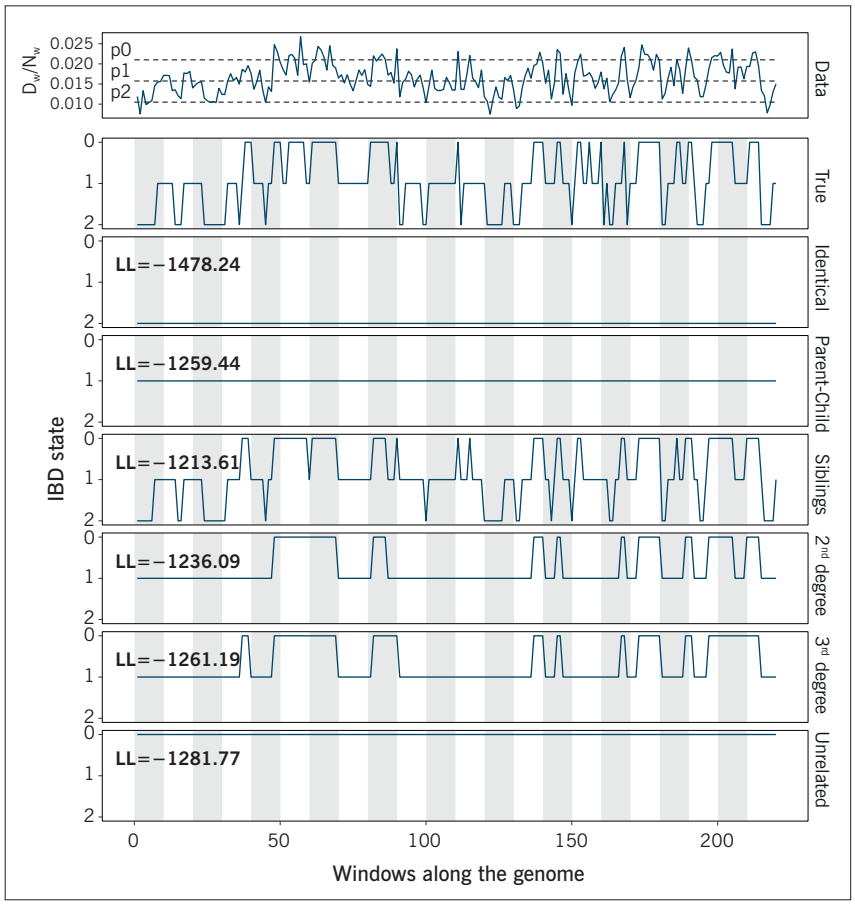

Fig. 1 Comparison of pairwise difference data and inferred IBD fragments (Identical By Descent). The top panel shows the proportion of differences in each window along the genome for a pair of simulated siblings. Dashed lines represent p_0, p_1 and p_2 estimates. The second panel shows the true IBD state for each window. The remaining panels show the IBD states predicted by particular relatedness models. The log-likelihood value for each model is shown in the upper left corner of the panel. Light and shaded backgrounds represent distinct chromosomes.

Abb. 1 *Vergleich der paarweisen unterschiedlichen Daten und daraus geschlussfolgerter IBD-Fragmente (identisch nach Abstammung). Das obere Feld zeigt die Proportion von Unterschieden für ein Paar simulierter Geschwister in jedem Fenster entlang der Genome. Gestrichelte Linien stellen p_0-, p_1- und p_2-Schätzungen dar. Das zweite Feld gibt den wahren IBD-Status für jedes Fenster wieder. Die verbleibenden Felder zeigen den IBD-Status, der durch spezifische Verwandtschaftsmodelle vorhergesagt wird. Der logarithmische Wahrscheinlichkeitswert für jedes Model wird in der oberen linken Ecke des Feldes gezeigt. Helle und schattierte Hintergründe geben eindeutige Chromosome an.*

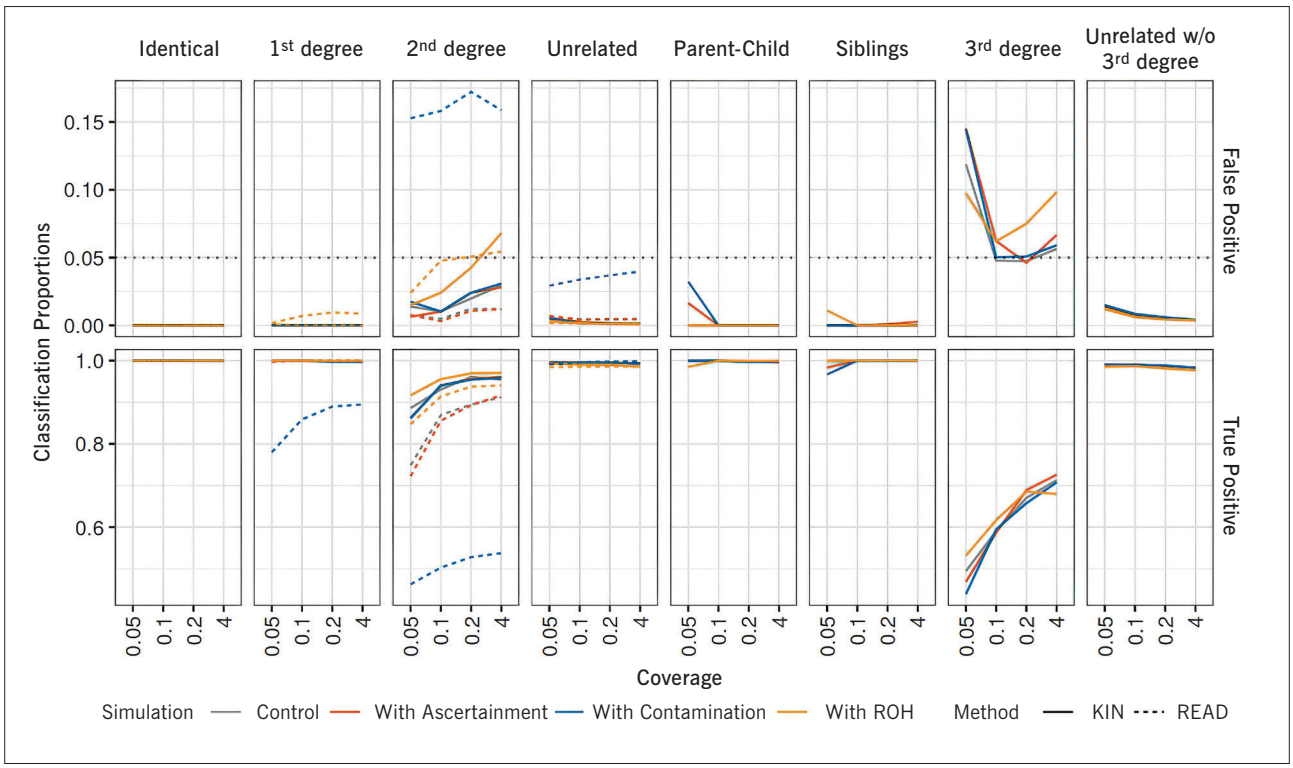

Fig. 2 Comparison of KIN with READ using simulations with different coverages and different cases of ascertainment, contamination, and ROH (runs of homozygosity). ›Unrelated‹ label here refers to KIN performance results when all Unrelated, 5th degree, 4th degree, 3rd degree pairs are labelled as Unrelated (for fair comparison with READ). ›Unrelated w/o 3rd degree‹ refers to the performance results when 3rd degree is classified separately from the unrelated individuals.

Abb. 2 *Vergleich von KIN und READ unter der Verwendung von Simulationen mit verschiedenen Erfassungsbereichen und verschiedenen Fällen von Erhebung, Kontamination und ROH (Homozygotiestrukturen). Die Bezeichnung ›Unrelated‹ beschreibt die KIN-Funktionsergebnisse, wenn alle nicht verwandt sind. Zudem werden im 5., 4. und 3. Grad Verwandte und Paare, die nicht verwandt sind (zum fairen Vergleich mit READ), aufgeführt. ›Unrelated w/o 3rd degree‹ beschreibt die Funktionsergebnisse, wenn sie separat von den nicht verwandten Individuen klassifiziert werden.*

viduals by comparing the likelihood of all the models. We noticed that KIN does not have the power to differentiate relatedness above the 3rd degree, particularly for low coverage data, so we classify these as unrelated. Similarly, we merge grandparent-grandchild, avuncular, half-siblings as 2nddegree relatives.

We evaluate the performance of KIN using simulations, and compare it to that of READ (Fig. 2). We show that both the methods perform well in control simulations (with no ROH and contamination), except in the case of 2nd-degree relatives, for which KIN has more power than READ. In simulations with contamination, we see that READ's performance is greatly affected, while KIN's performance is similar to the control. In simulations with ROH, we find that the true- and false-positive rates increase for both KIN and READ, although READ shows slightly more false positives, particularly for 2nd degree relatives with low coverage data. This effect is perhaps because ROH makes closely related individuals seem closer than they actually are. We find that both methods are not sensitive to ascertainment bias in simulations.

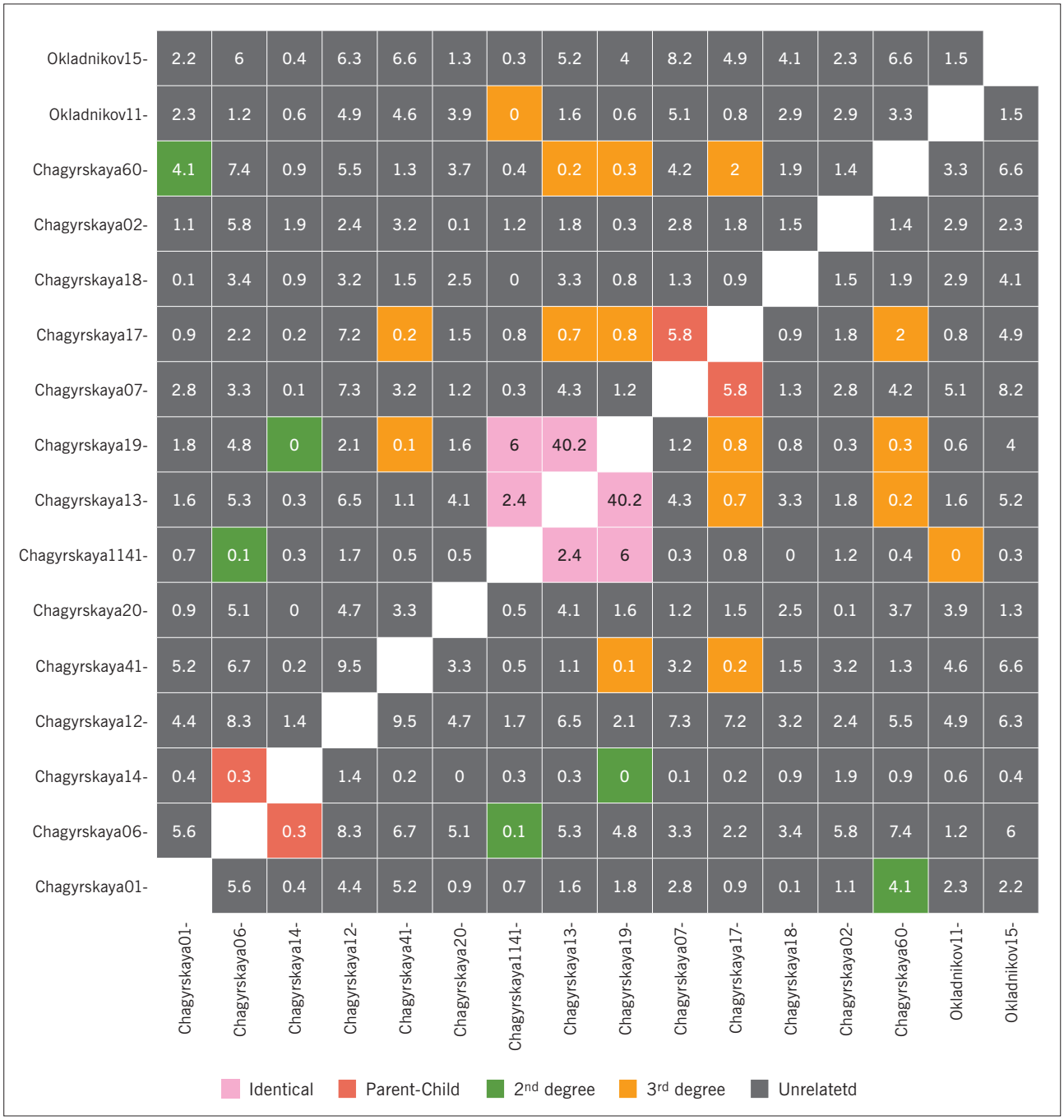

Fig. 3 Application of KIN to Neandertal remains from Chagyrskaya and Okladnikov caves, both Altai Krai (Russia). The color of a square represents the relatedness, while the number within denotes log likelihood ratio (ΔLL) between the two maximum likelihood models.

Abb. 3 *Die Anwendung von KIN an Neandertalerüberresten aus den Chagyrskaya- und Okladnikov-Höhlen, beide Region Altai (Russland). Die Farbe eines Quadrats repräsentiert die Verwandtschaft, während die Ziffer darin das logarithmische Wahrscheinlichkeitsverhältnis (ΔLL) zwischen den zwei größten Wahrscheinlichkeitsmodellen bezeichnet.*

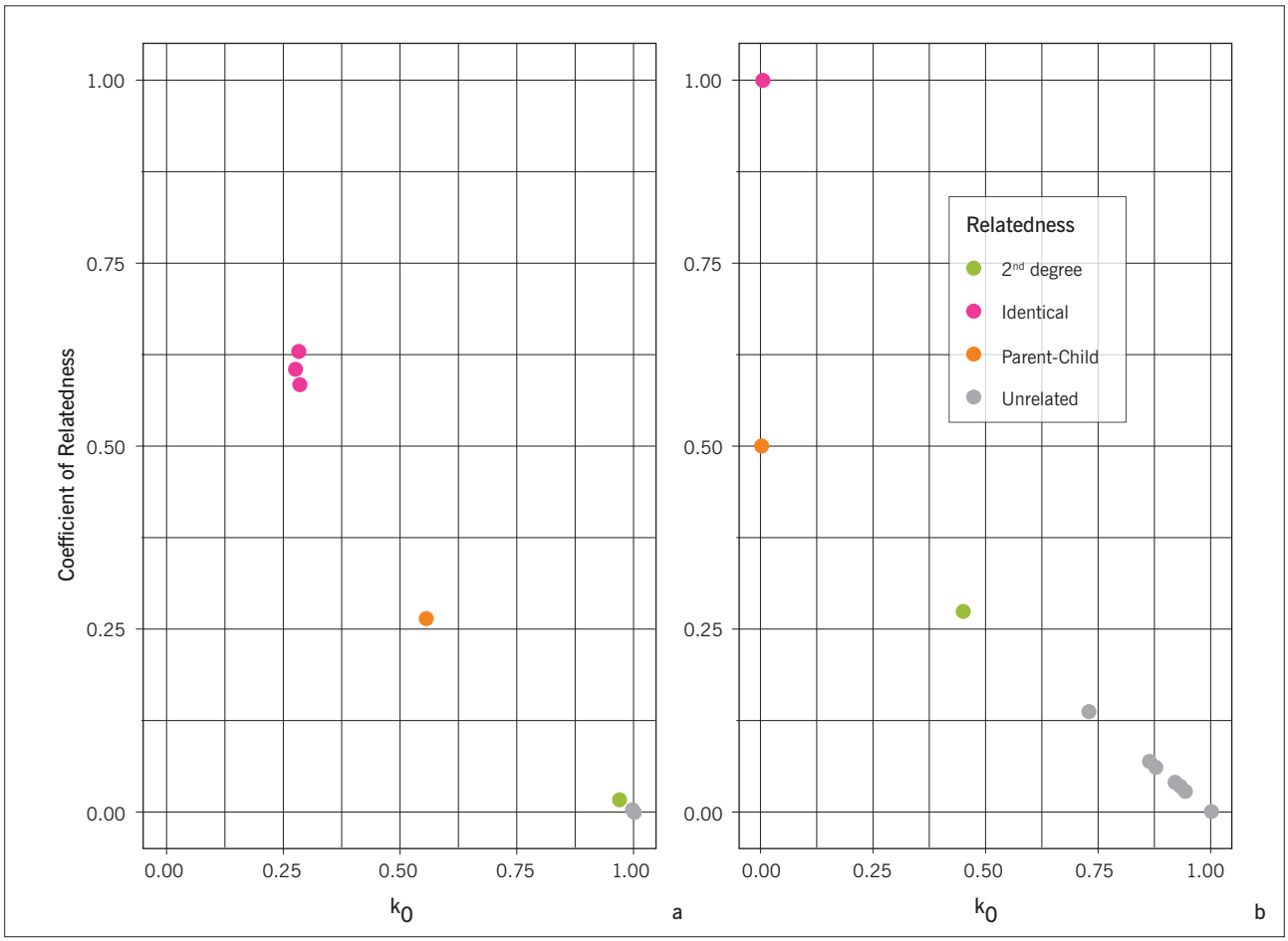

Fig. 4a–b Comparison of IBD states estimated for Chagyrskaya specimens using (a) lcMLkin and (b) KIN. Each plot shows the relatedness coefficient (calculated as k_2+k_1/2) on the y-axis plotted against k_0 (proportion of the genome with no chromosomes shared). The relatedness shown with different colors is estimated with both READ and KIN.

Abb. 4a–b *Vergleich von IBD-Zuständen, geschätzt für die Chagyrskaya-Proben unter der Verwendung von (a) lcMLkin und (b) KIN. Jeder Punkt zeigt den Verwandtschaftskoeffizienten (berechnet als k_2+k_1/2), auf der Y-Achse eingetragen gegen k_0 (Proportion der Genome ohne geteilte Chromosome). Die Verwandtschaft, die durch verschiedene Farben ausgedrückt ist, ist sowohl mit READ als auch mit KIN geschätzt.*

We apply KIN to ancient DNA data from 16 Neandertal specimens from Chagyrskaya and Okladnikov caves, Russia. This data was generated by targeted enrichment for 713 000 positions across the genome (Kolobova et al. 2020; Skov et al. 2022). The coverage ranged from 0.01x to 12.34x, and several libraries showed evidence of contamination. We find long ROH in most individuals, and are able to detect specimens that come from the same individual, as well as from three pairs of related individuals, a parent-child pair, and 2nd- and 3rd-degree relatives (Fig. 3). Our results match that of READ, except that READ does not identify the 3rd-degree relatives and does not distinguish the parent-child pair from siblings.

We next compare the relatedness coefficient (r) and the proportion of the genome in the unrelated state (k_0) obtained from KIN to those from lcMLkin for all pairs for whom the results of READ and KIN matches. We find that lcMLkin identifies four different clusters corresponding to different relatedness cases (identical, parent–child, 2nd degree and unrelated). However, the observed values for these clusters strongly deviates from expected, suggesting that the results of lcMLkin are perhaps affected by the low coverage and the presence of ROH. For example, specimens from the same individual are expected to be at $r = 1$ and $k_0 = 0$, but are observed at $r \approx 0.6$ and $k_0 \approx 0.3$ for lcMLkin (Fig. 4).

Relatedness results from KIN are in the form of a table that shows, for each pair, the most likely and second best models, along with the confidence level represented by the log likelihood ratio. This tabular format makes it easy to automatize the processing of KIN results for large datasets. We also provide a python package (KINgaroo) to make KIN input files from bam files, while optionally estimating ROH and adjusting for the contamination estimates. KINgaroo and KIN are available along with their documentation on github: <https://github.com/DivyaratanPopli/Kinship_Inference> (21.03.2023).

Bibliography

Monroy Kuhn et al. 2018
J. M. Monroy Kuhn/M. Jakobsson/T. Günther, Estimating genetic kin relationships in prehistoric populations. PLOS ONE 13,4, 2018, e0195491, <https://doi.org/10.1371/journal.pone.0195491> (21.03.2023).

Lipatov et al. 2015
M. Lipatov/K. Sanjeev/R. Patro/K. R. Veeramah, Maximum Likelihood Estimation of Biological Relatedness from Low Coverage Sequencing Data. bioRxiv 2015, <https://doi.org/10.1101/023374> (21.03.2023).

Popli et al. 2023
D. Popli/S. Peyrégne/B. M. Peter, KIN: a method to infer relatedness from low-coverage ancient DNA. Genome Biol. 24,10, 2023, <https://doi.org/10.1186/s13059-023-02847-7> (21.03.2023).

Skov et al. 2022
L. Skov/S. Peyrégne/D. Popli/L. N. M. Iasi/T. Devièse et al., Genetic insights into the social organization of Neanderthals. Nature 610, 2022, 519–525, <https://doi.org/10.1038/s41586-022-05283-y> (21.03.2023).

Kolobova et al. 2020
K. A. Kolobova/R. G. Roberts/V. P. Chabai/Z. Jacobs/M. T. Krajcarz et al., Archaeological evidence for two separate dispersals of Neanderthals into southern Siberia. Proc. Nat. Acad. Scien. USA 117, 2020, 2879–2885, <https://doi.org/10.1073/pnas.1918047117> (21.03.2023).

Source of Figures

1–4 D. Popli

Tab. 1 D. Popli

Addresses

Divyaratan Popli, M. Sc.
Max Planck Institute for Evolutionary Anthropology
Department of Evolutionary Genetics
Deutscher Platz 6
04103 Leipzig
Germany
divyaratan_popli@eva.mpg.de

Dr. Stéphane Peyrégne
Max Planck Institute for Evolutionary Anthropology
Department of Evolutionary Genetics
Deutscher Platz 6
04103 Leipzig
Germany
stephane_peyregne@eva.mpg.de

Dr. Benjamin M. Peter
Max Planck Institute for Evolutionary Anthropology
Department of Evolutionary Genetics
Deutscher Platz 6
04103 Leipzig
Germany
benjamin_peter@eva.mpg.de