

The Manuscript as Dataset: Biological Evidence and Computational Analysis in Medieval Palimpsests

James B. Harr III¹ 

¹ Department of Data Science, The College of William and Mary

Abstract

Palimpsests offer rich textual layers for scholars of medieval manuscripts, but the physical signals of reuse often escape traditional visual or spectral detection. In this talk, I present a novel interdisciplinary approach that bridges molecular biology, computational genomics, and digital humanities to investigate whether mitochondrial DNA (mtDNA) preserved in parchment can serve as a material trace of palimpsesting processes and how machine learning might leverage such biological data to assist in computational classification. I begin by situating this work within broader digital humanities debates on material digital scholarship, where physical artifacts are reimagined as data sources through non-invasive sampling and computational analysis. Using non-destructive surface brushing, I collected DNA from parchment leaves and applied high-throughput sequencing and a robust bioinformatics pipeline to generate high-coverage mitochondrial genomes. Preliminary results show that palimpsested and single-use parchment yield nearly indistinguishable mitochondrial profiles, indicating that historical chemical and physical erasure techniques do not systematically degrade mtDNA. I then explore how supervised learning models, including neural networks, PCA-based analysis, and class imbalance strategies like SMOTE and sequence augmentation, perform on this biological dataset. While classifiers struggled to distinguish palimpsests from single-use samples based solely on mtDNA, the computational effort highlights the limits and potentials of biological data as a digital resource in manuscript studies. This project contributes to digital humanities by demonstrating how biological provenance data can enrich the interdisciplinary study of textual artifacts and by critically reflecting on the role of machine learning in interpreting complex cultural heritage data.

Keywords: palimpsests, biocodicology, mitochondrial DNA, machine learning, medieval manuscripts, ancient DNA

1 Introduction and Motivation

Palimpsests occupy a singular place in the medieval archive. Produced when scribes scraped or chemically washed parchment to remove existing text before writing anew, these layered manuscripts preserve multiple textual lives within a single material object. They have attracted sustained scholarly attention precisely because what was erased often proves more historically significant than what replaced it, from Cicero's *De re publica* recovered beneath a patristic commentary to the mathematical treatises of Archimedes hidden beneath a Byzantine prayer book.

Yet for all their cultural significance, palimpsests remain difficult to identify with confidence. Multispectral imaging and ultraviolet illumination have transformed detection and recovery of erased text, but they are not infallible, and many palimpsests likely remain unidentified in collections worldwide [7]. The challenge is both visual and material: once thoroughly processed, reused parchment may betray no obvious signs of its prior life to the naked eye, and even advanced imaging techniques depend on the survival of some chemical trace of the original ink.

James B. Harr III. "The Manuscript as Dataset: Biological Evidence and Computational Analysis in Medieval Palimpsests." In: *From Manuscript to Megabyte: Building Sustainable and Open Futures in the Digital Humanities*, ed. by James B. Harr III, Emma Stanley, Jenna Rinalducci, and Donna Kain. Vol. 5. Anthology of Computers and the Humanities. 2025, 14–21. <https://doi.org/10.63744/CtkiLEdTpnTq>.

© 2025 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

This project pursued a different line of inquiry. Parchment is animal skin, and animal skin carries DNA. If the mechanical or chemical processes of erasure altered the mitochondrial genome of the parchment’s source animal in consistent, measurable ways, a biological signal might exist where a visual one does not. The question, then, was whether palimpsesting leaves a biological trace, one that might be measured, modeled, and eventually used as a detection tool [10].

The manuscript chosen for study was UPenn Ms. Codex 1629, *Commentaria ad Rhetoricam Ciceronis* [4], a fourteenth-century codex held at the Kislak Center for Special Collections at the University of Pennsylvania. It contains both single-use and palimpsested parchments with generous margins enabling non-invasive sampling. The study was designed in conversation with the emerging field of biocodicology, which integrates molecular biology, computational analysis, and traditional codicology to reconstruct manuscript production, provenance, and reuse [8; 19].

2 Biological Foundations and Prior Work

Parchment’s durability as a writing surface is also what makes it valuable for biological analysis. Composed primarily of collagen, a dense protein matrix, parchment binds and protects mitochondrial DNA even across centuries of storage and handling. When a scribe scraped or washed a surface for reuse, the collagen at deeper layers of the skin remained largely intact, meaning that mtDNA is not only present in medieval parchment but often remarkably well-preserved [16].

The field of biocodicology has advanced considerably over the past decade. Early work by Fiddyment et al. [8] demonstrated that species identification was possible through peptide mass fingerprinting without destructive sampling. Teasdale et al. [19] extended this approach to the York Gospels, revealing biological diversity and evidence of reuse across the codex’s constituent parchments and establishing each manuscript as a kind of layered biological record. Scheible et al. [16] refined minimally invasive surface-brushing techniques to extract mtDNA from fifteenth-century manuscripts, including chemically treated palimpsests, achieving over 90% mtDNA coverage and implementing rigorous authentication pipelines.

Phylogenetic reference datasets underpin species and lineage identification in parchment analysis. Naderi et al. [14] identified six major goat mitochondrial haplogroups across Eurasia, while Daly et al. [6] traced ancient goat genomes to mosaic domestication in the Fertile Crescent. Complementary ovine mtDNA work by Meadows et al. [13] and comparative sequencing studies [2] provide baselines for assigning parchment samples to regional populations. Bioinformatics tools such as MapDamage, PMDtools, and schmutzi address the post-mortem deamination and contamination issues characteristic of ancient DNA, enabling authentication, modeling, and visualization with high precision [15].

The intersection of machine learning and manuscript studies has further expanded the analytical toolkit available to biocodicologists. Easton et al. [7] applied multispectral imaging and pattern recognition to reveal erased texts in the Archimedes Palimpsest. Fiorucci et al. [9] surveyed artificial intelligence applications across cultural heritage contexts more broadly, and digital codicology initiatives [17; 18] have illuminated the potential for automated classification of reused parchment and biological pattern detection.

3 Methodology

Prior to sampling, a comprehensive condition assessment of Codex 1629 was conducted to inform sample site selection and ensure long-term preservation. The manuscript features a limp vellum binding with the textblock sewn on three supports; the textblock consists entirely of parchment, written in dark brown-black and green-black inks with substantial margins, decorative red ink initials, and numerous marginal annotations. The parchment leaves display minor planar distortions,

embedded grime, localized staining, and minimal insect damage, but no visible evidence of prior conservation treatment that might have introduced protein-based contamination.

Based on this assessment, sampling locations were selected from the bottom margins of leaves, avoiding the gutter, which accumulates dirt and debris, and the corners and edges most frequently handled by patrons. A total of 99 samples were collected from 52 folios, with priority given to folios identified as potential palimpsests following raked light and multispectral light inspections.

All collection followed a standardized contamination-prevention protocol. Sterile nitrile gloves were replaced frequently, and surgical-grade face masks were worn throughout. Genetic material was collected using EndoCervix brushes applied to the substrate surface for a minimum of 60 seconds across an area of $2.5\text{ cm} \times 10\text{ cm}$. The brush heads were aseptically detached and transferred to DNA LoBind tubes for transport to the North Carolina State University College of Veterinary Medicine. Post-sampling inspection confirmed that the binding, sewing, and parchment remained intact and consistent with the manuscript's pre-sampling condition.

DNA isolation was performed in batches of 16 or fewer samples, each containing a reagent blank to monitor contamination. For each sample, isolation buffer was added to the brush head, followed by incubation at 50°C with agitation at 800 rpm for approximately 24 hours. DNA concentration was quantified using the Qubit dsDNA High Sensitivity Assay Kit. Sequencing libraries were prepared using the KAPA Hyper Prep Kit, with KAPA unique dual-indexed adapters used to individually barcode each sample for multiplexed sequencing. A single pool of 67 libraries was prepared and enriched using hybridization capture with commercially available cow, goat, and sheep mitochondrial genome baits, following established protocols [16]. Two sequential rounds of enrichment at 60°C for approximately 48 hours each were performed before sequencing on an Illumina MiniSeq System. Raw sequence reads were processed through the bioinformatics workflow established by Scheible et al. [16], using CLC Genomics Workbench. This pipeline filtered and aligned reads, reconstructed mitochondrial genomes, and applied authentication tools, including schmutzi [15] and ANGSD [12], designed to distinguish genuine ancient DNA from modern contamination.

Following the generation of mtGenome consensus sequences, a Python-based neural network was developed to test whether an algorithm could distinguish between single-use and palimpsested parchment samples. A reference dataset was compiled by retrieving approximately 500 ancient mtGenome sequences of *Capra hircus* (goat) and *Ovis aries* (sheep) from GenBank [2; 13; 14]. All sequences were trimmed to the 16042–16600 bp region, isolating the displacement loop (d-loop) to maximize comparison consistency. The final training and testing set comprised 124 samples: 5 palimpsests and 119 single-use or untested specimens. To address class imbalance, two interventions were applied: Synthetic Minority Over-sampling Technique (SMOTE) [3], which generated synthetic palimpsest samples by interpolating between existing data points; and sequence augmentation, which subdivided each mtGenome into shorter segments to expand the effective dataset while maintaining biological plausibility. Principal component analysis [11] was used to visualize feature separability. Two supervised classifiers (logistic regression and random forest) were evaluated using ROC-AUC scores, confusion matrices, and feature importance. All work was implemented in Python via Google Colab, prioritizing methodological transparency and reproducibility [1; 17; 18].

4 Results

The primary sequencing results demonstrated remarkably similar performance across both parchment types. Palimpsested samples ($n = 22$) yielded an average consensus length of 16,504 base pairs with 86,304 mapped reads, achieving 99% mtGenome coverage and a mean read depth of 574 (± 235). Single-use samples ($n = 26$) produced an average consensus length of 15,925 base pairs with 75,141 mapped reads, achieving 96% mtGenome coverage and a mean read depth of 521

(± 172). Statistical analysis revealed no significant differences between parchment types in either mtGenome coverage ($p = 0.0799$) or mean read depth ($p = 0.8397$). All 22 palimpsested samples exceeded the 90% genome coverage quality threshold established by Scheible et al. [16], and all 22 surpassed the $>10\times$ mean read depth threshold. Among single-use samples, 23 of 26 met both thresholds; the same three specimens failed both metrics, suggesting localized issues with those particular samples rather than any systematic problem with single-use parchment sequencing. The consistently high performance of palimpsested samples is particularly significant given that these parchments underwent additional physical and chemical processing, including the historical application of chemical washes intended to remove the original text, yet retained genetic material at levels statistically indistinguishable from single-use specimens. This finding is itself a meaningful result for biocodicology. It confirms that parchment’s biological record survives the palimpsesting process, opening new possibilities for genetic investigation of manuscript production, circulation, and reuse patterns across medieval Europe.

To verify species attribution, each consensus mtGenome was queried against GenBank using a default blastn search. All samples except one single-use specimen returned a best match to *Ovis aries* (sheep) with 100% query coverage, an E-value of 0, and $>99\%$ sequence identity. The single exception mapped to *Capra hircus* (goat), indicating that Codex 1629 was produced primarily from sheep parchment with at least one goat skin incorporated into its construction. This mixed-species composition is consistent with medieval parchment production practices, where scribes utilized available animal resources without strict species uniformity, and it confirms that the sequencing was successful and the data reliable.

The baseline neural network identified three of five palimpsests correctly but misclassified two as single-use. The model achieved perfect precision for palimpsests (1.00), meaning that when it labeled a sample as a palimpsest, it was always correct, but recall remained limited at 0.60, indicating that 40% of actual palimpsests were missed. Overall accuracy was 98% across the full test set ($n = 124$), but this figure is misleading given the extreme class imbalance: a model that predicted “single-use” for every sample would achieve nearly identical overall accuracy while detecting no palimpsests at all.

To understand this conservative behavior, principal component analysis was conducted on the balanced training set. The visualization revealed the fundamental challenge underlying palimpsest detection: substantial overlap between classes in feature space. While PC1 explained 99.1% of the variance, palimpsest and single-use samples were intermingled throughout the distribution with no meaningful cluster separation. One notable exception was a broader spread among palimpsest samples, with some appearing as outliers at the margins of the distribution, suggesting that there may be genuine biological variability associated with reuse, but it is subtle and not captured in the dominant axes of variation. The PCA confirmed that the classifiers were not failing due to poor model design. They were accurately reflecting the structure of the underlying data: palimpsesting does not alter k-mer features of the mitochondrial genome in ways that produce distinguishable clusters.

To address class imbalance, SMOTE was applied, generating over 3 million synthetic palimpsest samples. PCA of the resulting balanced dataset showed synthetic palimpsest points spread widely along PC1 (which captured 99.4% of variance), while single-use samples remained tightly grouped near the origin. While SMOTE reduced the statistical class imbalance and enhanced the model’s ability to learn palimpsest-like patterns, the extensive overlap between classes persisted. Classification recall for palimpsests improved only modestly despite massive oversampling. As an alternative strategy, a sequence augmentation approach was implemented, extracting random sub-sequences from each of the 24 palimpsest samples. Sequence-level cross-validation prevented data leakage. With 538 single-use and 538 palimpsest sample segments, retraining the model achieved a balanced accuracy of 0.989. However, a critical caveat applies:

without additional independent palimpsest samples for external validation, these results remain preliminary rather than conclusive. The segments derive from the same small set of original palimpsest specimens, and true confirmation of generalizability would require testing on entirely new, previously unseen palimpsest parchments.

Two supervised classifiers were trained on a reduced feature set consisting of GC content, sequence length, and CpG count. ROC curve analysis revealed that both models achieved only marginally better-than-random performance: logistic regression yielded an AUC of 0.55 and random forest an AUC of 0.57. An AUC of 0.5 is equivalent to a coin flip; a score above 0.7 is generally considered meaningful. The two classifiers exhibited opposite failure modes. Logistic regression showed strong bias toward predicting the majority class, resulting in numerous false negatives. Random forest predicted nearly all samples as positive, producing a large number of false positives while offering no improvement in palimpsest identification. These divergent behaviors illustrate that both classifiers were responding primarily to class priors rather than meaningful biological features. Feature importance values showed that GC content and sequence length contributed most to predictions, with CpG count carrying comparatively less weight, yet even the variation in these features was insufficient to support reliable classification.

5 Interpretation and Implications

It would be easy to characterize these machine learning results as failures. The more accurate framing is that they constitute evidence. The models are not failing because they are poorly designed; they are accurately reflecting the biology. Palimpsesting does not alter the mitochondrial genome in ways that machine learning can detect. That is a finding, and an important one. It establishes a meaningful boundary condition for biocodicology: mtDNA integrity is preserved through the palimpsesting process, but that preservation is precisely what makes classification difficult. The biological record survives erasure; it simply does not bear a legible mark of it. Taken together, the sequencing results and the machine learning results reinforce the same conclusion from complementary directions. The sequencing showed no significant difference in genome coverage or read depth between parchment types. Machine learning confirmed that the genetic features available (k-mer frequencies, GC content, sequence length, CpG count) do not differ in ways that correspond to parchment history. Both lines of evidence converge on the same biological reality: medieval erasure practices transformed the textual surface without transforming the genetic substrate. This finding has significant practical implications for the field. It means that the biological record in even the most heavily processed manuscripts may be recoverable and reliable for purposes such as species identification, provenance tracing, and reconstruction of animal husbandry and trade networks. It also means, however, that mtDNA alone cannot serve as a palimpsest detection tool, at least at the level of the whole mitochondrial genome.

The most significant constraint on this research is structural rather than methodological: confirmed palimpsests are rare, and the field cannot easily manufacture more of them. The training set contained only five confirmed palimpsest specimens, creating an extreme class imbalance that no oversampling technique can fully overcome. SMOTE generates synthetic data by interpolation; it cannot introduce biological signal that is absent from the original samples. The choice to focus on the d-loop region, while motivated by its relatively high genetic variability, introduced an additional constraint. The d-loop is non-coding and does not differ in ways that correspond to the historical processing of the parchment; its variability reflects population-level diversity rather than individual specimen history. The genetic similarity across all samples also narrowed the signal space available for classification: both parchment types derive from closely related domesticated species whose mitochondrial genomes are highly conserved across individuals and breeds. These are not failures of method but structural features of working at the intersection of archive and biology. They are worth naming clearly so that future researchers can design studies that address

them, for instance, by sampling across multiple manuscripts, seeking access to larger collections of confirmed palimpsests, or expanding the feature set beyond mtDNA to include nuclear DNA markers, proteomic data, or material analysis.

6 Future Directions

The path forward for biocodicology is integrative. No single method (imaging, codicology, or genomics) will detect all palimpsests independently, and the results of this study clarify the specific contributions and limitations of genetic analysis. What the project demonstrates is that biological data can be collected, authenticated, and analyzed at scale without harming manuscripts, and that the resulting data is reliable even in heavily processed substrates. The next step is to combine that data with multispectral imaging, material analysis, and codicological description in multimodal models.

Several specific directions merit further investigation. Deep learning architectures capable of detecting subtle patterns in raw sequence data may outperform the k-mer and summary-statistic approaches used here. Alternative feature engineering strategies, including whole-genome alignments, epigenetic markers, or nuclear DNA features, might capture dimensions of parchment biology that mtDNA cannot. Integration with existing digital manuscript platforms could enable scalable screening protocols in which computational predictions guide targeted multispectral imaging, combining the low cost of genetic screening with the proven recovery capabilities of spectral analysis.

The interdisciplinary collaboration required to pursue these directions (bringing together medievalists, conservation specialists, molecular biologists, and data scientists) models the kind of team-based digital humanities research increasingly necessary for complex cultural heritage questions. By conducting all neural network applications in Python via Google Colab and documenting methods transparently, this project aims to lower technical barriers for manuscript scholars considering computational approaches [1; 17; 18].

7 Conclusion

Two findings anchor this project. First, mitochondrial DNA survives the palimpsesting process. Chemical washing and mechanical scraping, the tools of medieval text removal, did not eliminate or significantly degrade the recoverable genetic material in Codex 1629’s parchment leaves. This confirms that even heavily processed manuscripts may retain analyzable genetic signatures, and it establishes a reliable foundation for future biocodicological research on manuscript production, circulation, and reuse across medieval Europe. Second, the biological record preserved in parchment cannot, on its own and at the level of the mitochondrial genome, reliably distinguish reused parchment from parchment written on once. Machine learning models applied to mtDNA sequence features produced AUC scores barely above chance, and PCA confirmed that the two classes of parchment are not meaningfully separated in feature space. These results are not a failure of the computational method; they accurately reflect the biology. Palimpsesting does not leave a legible mark at the genetic level.

Together, these findings define realistic expectations for biocodicology as a detection science and open a productive conversation about what integrative, multimodal approaches to manuscript study might accomplish. The manuscript is a dataset. What it encodes, and what computational methods can extract from it, remains an open and generative question.

References

- [1] Camps, Jean-Baptiste. *La “chanson d’Otinel”: Edition complète du corpus manuscrit et prolégomènes à l’édition critique*. HAL archives-ouvertes. PhD thesis. Université Paris-Sorbonne, 2016. URL: <https://hal.science/tel-01664932>.
- [2] Cassidy, Lara M., Teasdale, Matthew D., Carolan, Seán, Enright, Rita, Werner, Ruth, Bradley, Daniel G., Finlay, Emma K., and Mattiangeli, Valeria. “Capturing goats: Documenting two hundred years of mitochondrial DNA diversity among goat populations from Britain and Ireland”. In: *Biology Letters* 13, no. 3 (2017), p. 20160876. DOI: 10.1098/rsbl.2016.0876.
- [3] Chawla, Nitesh V., Bowyer, Kevin W., Hall, Lawrence O., and Kegelmeyer, W. Philip. “SMOTE: Synthetic minority over-sampling technique”. In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357. DOI: 10.1613/jair.953.
- [4] Cicero, Marcus Tullius. “Commentaria ad Rhetoricam Ciceronis (Ms. Codex 1629)”. Manuscript, Kislak Center for Special Collections, University of Pennsylvania. 1342. URL: <https://find.library.upenn.edu/catalog/9958752123503681>.
- [5] Clemens, Raymond and Graham, Timothy. *Introduction to Manuscript Studies*. Cornell University Press, 2007.
- [6] Daly, Kevin G. et al. “Ancient goat genomes reveal mosaic domestication in the Fertile Crescent”. In: *Science* 361, no. 6397 (2018), pp. 85–88. DOI: 10.1126/science.aas9411.
- [7] Easton, Roger L., Knox, Keith T., and Christens-Barry, William A. “Multispectral imaging of the Archimedes Palimpsest”. In: *32nd Applied Imagery Pattern Recognition Workshop, 2003. Proceedings.* 2003, pp. 111–116. DOI: 10.1109/AIPR.2003.1284258.
- [8] Fiddymment, Sarah et al. “Animal origin of 13th-century uterine vellum revealed using non-invasive peptide fingerprinting”. In: *Proceedings of the National Academy of Sciences* 112, no. 49 (2015), pp. 15066–15071. DOI: 10.1073/pnas.1512264112.
- [9] Fiorucci, Marco, Khoroshiltseva, Marina, Pontil, Massimiliano, Traviglia, Arianna, Bue, Alessio Del, and James, Stuart. “Machine learning for cultural heritage: A survey”. In: *Pattern Recognition Letters* 133 (2020), pp. 102–108. DOI: 10.1016/j.patrec.2020.02.017.
- [10] III, James B. Harr, Blong, Meredith E., Gadomski, Timothy, Scheible, Melissa, Meiklejohn, Kelly A., and Jr., William E. Gundling. “Manuscripts, machine learning, and mitochondrial DNA: A computational approach to palimpsest detection in medieval manuscripts”. Manuscript submitted for publication (under review). 2025.
- [11] Jolliffe, Ian T. and Cadima, Jorge. “Principal component analysis: A review and recent developments”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, no. 2065 (2016), p. 20150202. DOI: 10.1098/rsta.2015.0202.
- [12] Korneliussen, Thorfinn Sand, Albrechtsen, Anders, and Nielsen, Rasmus. “ANGSD: Analysis of next generation sequencing data”. In: *BMC Bioinformatics* 15, no. 1 (2014), p. 356. DOI: 10.1186/s12859-014-0356-4.
- [13] Meadows, Jennifer R. S., Cemal, Ibrahim, Karaca, Orhan, Gootwine, Elisha, and Kijas, James W. “Five ovine mitochondrial lineages identified from sheep breeds of the Near East”. In: *Genetics* 175, no. 3 (2007), pp. 1371–1379. DOI: 10.1534/genetics.106.068353.

- [14] Naderi, Saeid, Rezaei, Hamid-Reza, Taberlet, Pierre, Zundel, Stefan, Rafat, Seyed-Abbas, Naghash, Hamid-Reza, el-Barody, Mohamed A., Ertugrul, Onur, Pompanon, François, and Econogene Consortium. “Large-scale mitochondrial DNA analysis of the domestic goat reveals six haplogroups with high diversity”. In: *PLoS ONE* 2, no. 10 (2007), e1012. DOI: 10.1371/journal.pone.0001012.
- [15] Renaud, Gabriel, Slon, Viviane, Duggan, Ana T., and Kelso, Janet. “Schmutzi: Estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA”. In: *Genome Biology* 16 (2015), p. 224. DOI: 10.1186/s13059-015-0776-0.
- [16] Scheible, Melissa, Stinson, Timothy L., Breen, Matthew, Callahan, Benjamin J., Thomas, Rachael, and Meiklejohn, Kelly A. “The development of non-destructive sampling methods of parchment skins for genetic species identification”. In: *PLoS ONE* 19, no. 3 (2024), e0299524. DOI: 10.1371/journal.pone.0299524.
- [17] Schöch, Christof. “Big? Smart? Clean? Messy? Data in the humanities”. In: *Journal of Digital Humanities* 2, no. 3 (2013), pp. 2–13. URL: <https://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities>.
- [18] Stokes, Peter A. “Computer-aided palaeography, present and future”. In: *Kodikologie und Paläographie im digitalen Zeitalter – Codicology and Palaeography in the Digital Age*, ed. by Malte Rehbein, Patrick Sahle, Torsten Schaßan, Philipp Schopper, and Andrea Vogeler. Books on Demand, 2009, pp. 309–338.
- [19] Teasdale, Matthew D., Fiddymment, Sarah, Vnouček, Jiří, et al. “The York Gospels: A 1000-year biological palimpsest”. In: *Royal Society Open Science* 4 (2017), p. 170988. DOI: 10.1098/rsos.170988.
- [20] Yang, Dongya Y., Eng, Barry, Wayne, John S., Dудар, J. Christopher, and Saunders, Shelley R. “Improved DNA extraction from ancient bones using silica-based spin columns”. In: *American Journal of Physical Anthropology* 105 (1998), pp. 539–543.