

## **Millennium-old Pathogenic Mendelian Mutation Discovery for Multiple Osteochondromas from a Gaelic Medieval Graveyard.**

Iseult Jackson<sup>1,2</sup>, Valeria Mattiangeli<sup>1</sup>, Lara M Cassidy<sup>1</sup>, Eileen Murphy<sup>3</sup>, Daniel G Bradley<sup>1</sup>

1. Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland

2. The SFI Centre for Research Training in Genomics Data Science, NUI Galway, Ireland

3. Archaeology and Palaeoecology, School of Natural and Built Environment, Queen's University Belfast, Belfast, Northern Ireland

\*Corresponding authors: Daniel G Bradley: [dbradley@tcd.ie](mailto:dbradley@tcd.ie); Eileen Murphy: [eileen.murphy@qub.ac.uk](mailto:eileen.murphy@qub.ac.uk);

## Summary

Only a limited number of genetic diseases are diagnosable in archaeological individuals and none have had causal mutations identified in genome-wide screens. Two individuals from the Gaelic Irish Medieval burial ground of Ballyhanna, Co. Donegal, showed evidence of bone tumors consistent with the autosomal dominant condition multiple osteochondromas. Genome sequencing of the earlier individual uncovered a missense mutation in the second exon of *EXT1*, a specific lesion that has been identified in several modern patients. The later individual lacked this but displayed a novel frameshift mutation leading to a premature stop codon and loss of function in the same gene. These molecular confirmations of a paleopathological diagnosis within a single rural ancient context are surprisingly disjunct, given the observation of clusters of this disease in modern isolated populations and a *de novo* mutation rate of only 10%.

## Main Text

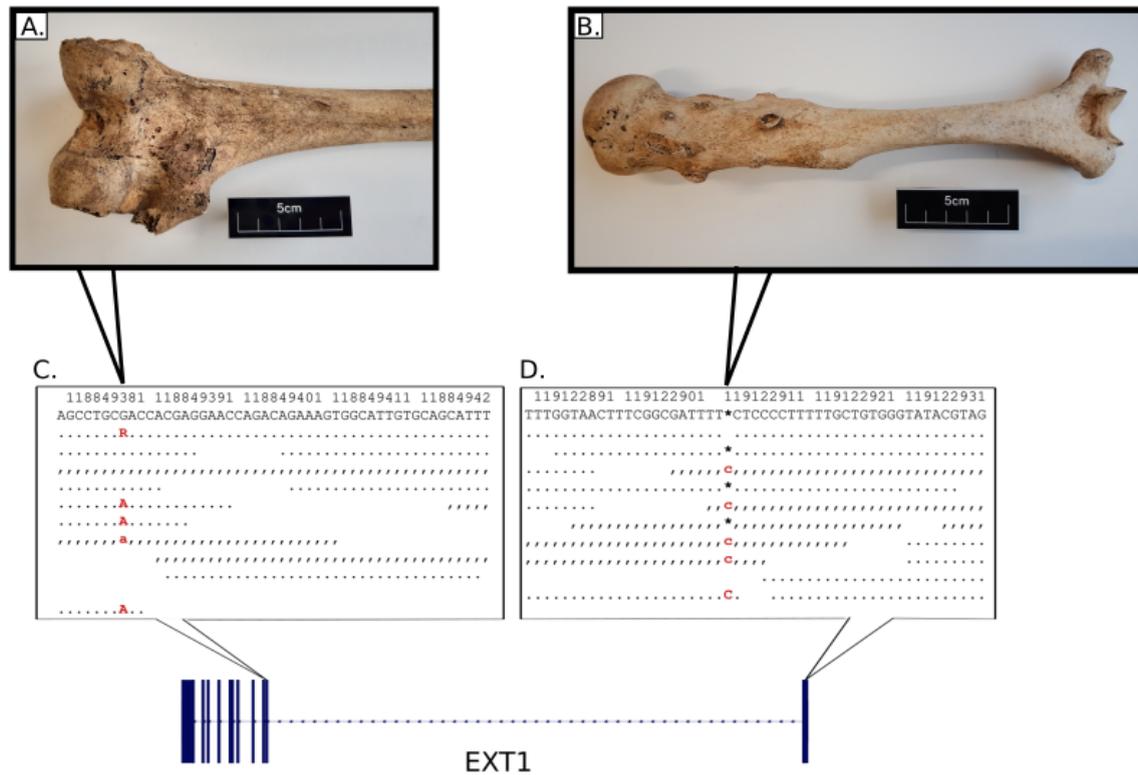
Cases of Mendelian genetic disease have been uncovered in the archaeological record, but diagnosis is restricted to osteologically visible conditions, with more commonly encountered conditions including achondroplasia, multiple epiphyseal dysplasia and Léri-Weill dyschondrosteosis<sup>1</sup>.

Our study focuses on two adult male individuals (Sk197 and Sk331) buried in a Medieval graveyard at Ballyhanna, near the town of Ballyshannon in Co. Donegal, Ireland. Each displayed multiple bony tumors suggestive of multiple osteochondromas (MO), a rare, autosomal dominant bone condition<sup>2</sup>. Although these tumors are normally benign, the condition can result in limb deformity, reduced stature, compression of nerves, and, more rarely, malignancy<sup>3</sup>.

The rural burial ground at Ballyhanna was associated with a small Medieval church, constructed after the middle of the 13th century AD, although it is possible that an earlier wooden church was present on the site. The land at that time would have been owned by a bishop and the estate lands would have been managed by an *erenagh* (estate manager). Radiocarbon dating showed that the earliest burials dated to the late 7th to early 8th century, but the vast majority of individuals were interred between AD 1200 and 1650, when the area around Ballyshannon was under the autonomous control of the Ó Domnaill Gaelic lords. As such, the graveyard at Ballyhanna can be considered to have essentially contained the remains of a Gaelic Medieval population. Those buried at Ballyhanna would have comprised the lower classes and included tenant farmers, labourers, merchants, artisans, clergy and the very poor<sup>4</sup>.

Radiocarbon dating revealed that the two individuals were definitely not contemporaneous and were potentially separated by several hundred years. Of the two, Sk331 lived more recently (dated AD 1031-1260; UBA-11442) (**Figure S1; Table S1**) and was the more severe case. He displayed extensive bilateral osteochondromas, both sessile and pedunculated in form, on most bones throughout his skeleton (**Figure 1 b.**). He also had a short stature compared to other adult males at Ballyhanna (158.3 cm), displayed a major deformity of his left forearm due to shortening of the ulna (Type 1<sup>5</sup>), had unequal bone lengths due to the lesions, as well as a range of orthopaedic deformities that affected his hips, knees and left ankle; all of which are consistent with this condition<sup>6</sup>. He died as a young adult (18-25 years). Sk197 was dated to AD 689-975 (UBA-11443) and was slightly older (30-40 years) when he died. While multiple osteochondromas were evident throughout his skeleton, they were generally less pronounced than those evident in Sk331. Limb length discrepancy was present in his forearm bones, his sacro-iliac joints displayed ankylosis and he would have had genu valgum during life. Unlike Sk331, he was estimated to have been of roughly average height for the population (166.8 cm). Neither individual appears to have suffered from any tumors that progressed to malignancy<sup>4</sup>.

In order to identify likely causative mutations, we used an unbiased genome wide approach involving shotgun sequencing and ancient DNA protocols (see **Methods**). Petrous bone samples from each individual yielded intermediate levels of endogenous DNA preservation (Sk331: 12.2%; Sk197: 13.9%) and were sequenced to a mean depth of coverage of 4.2× and 5.1× respectively (see **Table 1** and **Table S2** for summary and sequencing statistics). Likely pathogenic mutations in *EXT1* were identified from an exome-wide scan in both individuals. All possible variants were initially filtered for quality (minimum allelic depth 3; maximum read depth twice the mean genomic coverage; minimum genotype quality 50) and predicted molecular impact using SnpEff and SnpSift (high or moderate impact)<sup>7,8</sup>. All qualifying variants were heterozygous calls. Therefore, only variants in genes with a high probability of loss of function intolerance (pLI > 90%) according to gnomADv2.1.1 were retained, in order to filter out variants that were unlikely to cause a pathological phenotype in a heterozygous state<sup>9</sup>. These variants were then assessed based on allele frequencies (<1% in gnomAD), *in silico* predicted pathogenicity (according to SIFT and Polyphen2) and publicly available data from ClinVar<sup>9-11</sup>. Only one mutation in each individual had the level of support required by the American College of Medical Genetics guidelines to classify them as pathogenic (**Figure 1; Table S2**)<sup>12,13</sup>.



**Figure 1. Examples of lesions and alignment of NGS data at putative disease variants.** EXT1 gene structure with disease variants highlighted in red and bold typeface. First line: hs37d5 reference sequence. Second line: consensus for each individual. Dots: bases match reference on forward strand. Commas: bases match reference on reverse strand. R: consensus base is an A or G. A) Example of large sessile osteochondroma on the posterior surface of the distal end of Sk197's left femur. B) Example of sessile and pedunculated osteochondromas on the posterior surface of Sk331's left humerus. C) G>A transition in exon 2 of EXT1 in Sk197, resulting in Arginine to Cysteine mutation at the corresponding amino acid residue. D) C insertion in exon 1 of EXT1 in Sk331.

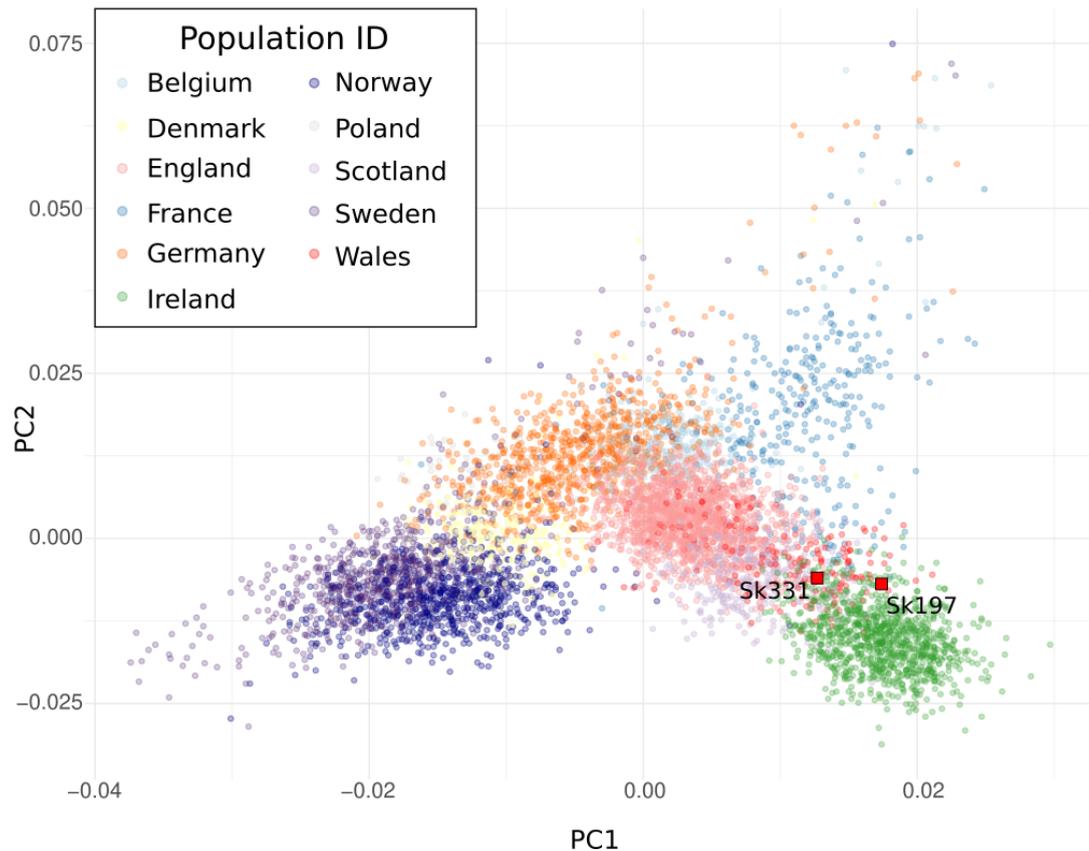
The predicted pathogenic mutation identified in Sk197 is a missense mutation in the second exon of *EXT1* (NC\_000008.10:g.118849385G>A; NP000118:p.Arg340Cys) (**Figure 1c**). This specific mutation has been identified in at least three patients<sup>14-16</sup>, and different missense mutations at the same amino acid residue have been identified as pathogenic (R340G,L,H; accessions VCV000988576, VCV000002495, VCV000265129), as have substitutions at nearby residues. Arginine to Cysteine is a non-conservative amino acid substitution, and at this position has been demonstrated to disrupt EXT1/EXT2 complex activity, consistent with what is known about disease mechanism<sup>17,18</sup>. Therefore, there are multiple lines of evidence supporting a pathogenic role for this mutation in Sk197: the same amino acid change as an established disease variant is observed; functional studies have shown a deleterious effect; this mutation is absent in population databases; computational evidence supports a damaging effect on the gene product; finally, the phenotype (MO) is highly specific for this gene.

Although the G>A mutation is not found in Sk331, a novel predicted pathogenic mutation is observed. This mutation is a C insertion within the first exon of *EXT1* (NC\_000008.10:g.119122909\_119122910insC; NP\_000118:p.(Lys126Argfs\*63)), resulting in a frameshift mutation and premature stop codon (**Figure 1d**). This is a very severe mutation, resulting in a complete loss of the protein product of one copy of *EXT1* and is consistent with what is known about pathogenic mutations associated with this disease<sup>19</sup>. Although this mutation has not been observed in modern patients (as reported in ClinVar; date accessed 14-05-2021), there are 6 frameshift variants predicted to be pathogenic within 50bp of this site, and 3 nonsense SNVs predicted to be pathogenic in this region (**Table S4**). Therefore, this is a null mutation where loss of function is a known mechanism of disease, in a known disease gene, at a locus where there is modern clinical data supporting pathogenicity. Computational tools predict this variant to be deleterious. In sum, there is a high level of evidence to support the pathogenicity of this variant.

90% of cases of MO are caused by mutations in the exostosin genes (*EXT1* and *EXT2*), which are involved in heparan sulfate chain synthesis and assembly<sup>20</sup>. These chains interact with a wide range of signalling molecules, and deficiencies in these interactions lead to altered signalling pathways<sup>17</sup>. Most of these mutations are classified as inactivating mutations; they result in a complete loss of the protein product from the copy of the gene carrying the mutation<sup>19</sup>. As *EXT1* and *EXT2* form a complex to carry out their molecular function, such mutations in either gene result in a reduction in the quantity of functional complex, and therefore a reduction in heparan sulfate chain production. Although this does have a physiological effect (for example, defects in lipid metabolism and clearance), it is insufficient for tumor formation<sup>20</sup>. MO is a dominant condition and it is thought that a complete somatic loss of *EXT1/2* function is necessary for disease, which has been reported as being due to a loss of heterozygosity or aneuploidy<sup>20,21</sup>, although studies in mice have demonstrated that compound *EXT1/2* heterozygosity can mimic the human MO phenotype<sup>22</sup>.

To date only stature phenotypes in recent historical samples have led to successful identification of causal Mendelian lesions; none with a genome-wide approach. Two 18th and 19th century skeletons of extremely tall individuals had sufficiently intact DNA for targeted PCR of gigantism-associated loci<sup>23,24,25</sup>. A pathological achondroplasia mutation in the *FGFR3* gene has also been identified in 180-year old remains<sup>26</sup>.

We also used projection principal components analysis with modern northwest European populations to test affinities of the two Ballyhanna genomes<sup>27-29</sup>. Both individuals fall at the overlap between Scottish and Irish samples, consistent with what we might expect for modern individuals from the northern part of Ireland<sup>27</sup> (**Figure 2**).



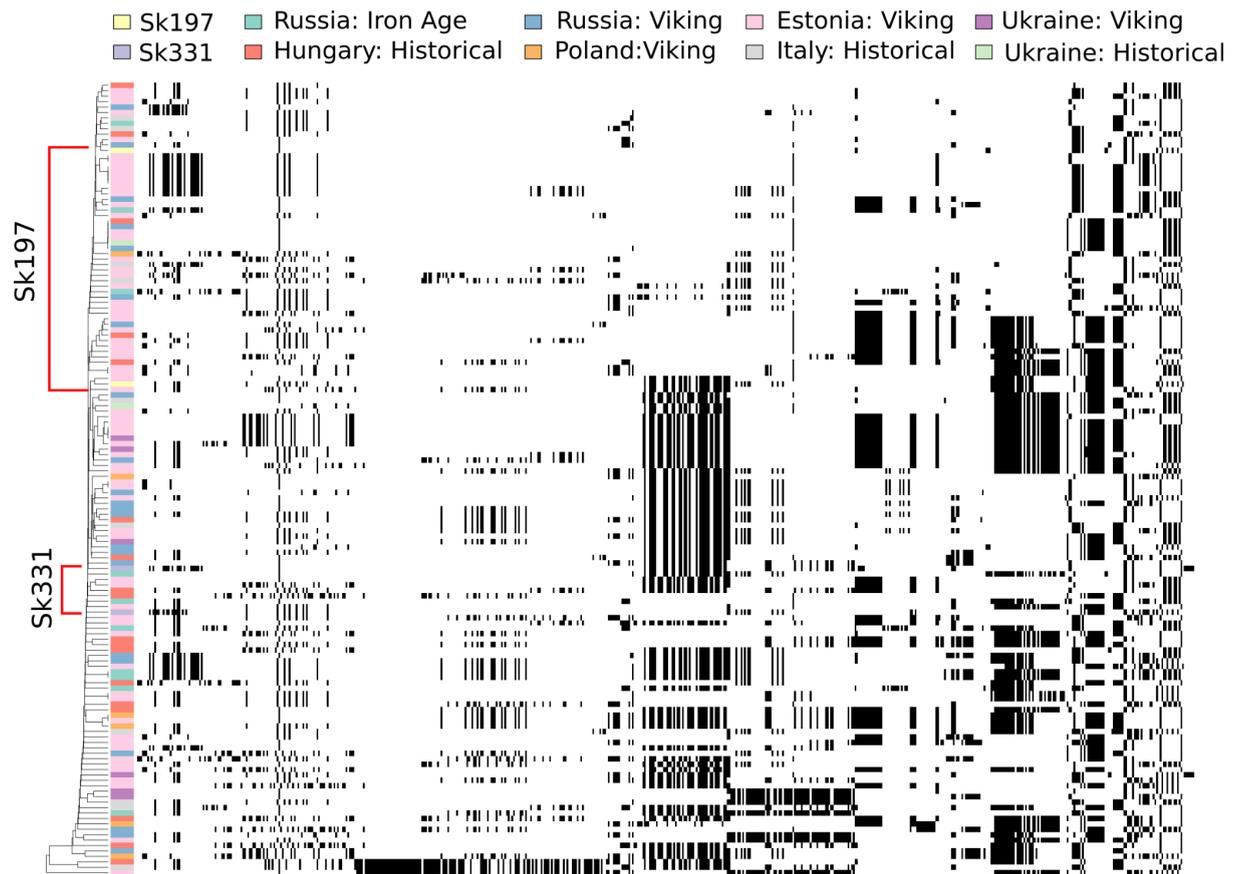
**Figure 2. Principal Component Analysis of North-west European populations.** Sk331 and Sk197 projected onto a plot of North-west European populations. Modern populations: Ireland (green); Scotland (lilac); Wales (red); England (pink); France (blue); Germany (orange); Belgium (pale blue); Norway (dark purple); Sweden (purple); Denmark (yellow); Poland (grey).

These men had different mitochondrial haplogroups but fell into the same clade of Y chromosome haplotypes, although Sk331 had a slightly more derived Y haplogroup (**Table 1**). Both grouped in the cluster R1b-M222, which is known to have its highest frequency in the same northwestern region in modern Ireland<sup>30,31</sup>. Interestingly, from modern examination of surnames, the Ó Domnaill Gaelic lords whose clan held power in this region would have been expected to display this haplotype<sup>31</sup>.

ID	Genomic Coverage	Sex	mtDNA Haplogroup	Y Haplogroup
Sk331	4.2×	XY	H13a1a	R1b1a1b1a1a2c1a1a1a1a3a (SNPs: M207; M343; M269; P312; S245; L21; M222; DF106; DF104; DF105; A259; A260)
Sk197	5.1×	XY	H4a1a2a	R1b1a1b1a1a2c1a1a1a1a1a (SNPs: M207; M343; M269; P312; S245; L21; M222; DF106; DF104; DF105)

**Table 1. Sequencing Summary.** Summary of coverage, sex and uniparental haplogroups assigned using MitoMaster and Phylotree 17 (mitochondrial) and ISOGG v.15.58 (Y).

The haplotype background of each Ballyhanna pathogenic mutation was assessed using diploid data imputed using GLIMPSE and phased with a dataset of 72 imputed, published Iron Age and Medieval genomes using SHAPEITv2 (see **Table S5** for details)<sup>32–36</sup>. The region around *EXT1*, including 50kB up- and downstream were visualised using haplostrips<sup>37</sup> (**Figure 3**). The Ballyhanna haplotypes do not cluster together, further supporting two independent MO-causing mutations in these individuals.



**Figure 3. *EXT1* haplotype clustering in imputed ancient data.** *EXT1* haplotypes were clustered using the tool haplostrips. Colours at the side of the plot indicate approximate geographic location and period of the individuals; Sk197 (pale yellow) and Sk331 (pale purple) are highlighted.

MO cases have been osteologically identified in the palaeopathological record from the Middle Bronze Age to the post-medieval period<sup>6,38</sup>. Ten of these cases are isolated examples, but four countries have multiple individuals with MO: Gotland (Sweden) (n=2: mother and unborn infant); Jordan (n=3, of which two are broadly contemporaneous), England (n=3) and Ireland (n=4, including the 2 probands here)<sup>6</sup>. The estimated incidence of this condition in modern individuals is approximately 1/50000, although higher incidences have been reported in the isolated populations of the Chamorros of Guam (1/1000) and the northern Ojibway community in Manitoba (13/1000)<sup>39,40</sup>. The *de novo* rate of mutation in this condition is low at 10%<sup>41</sup>, and this clustering in modern, restricted populations strongly suggests founder effects typical of severe dominant genetic diseases. Such would have been expected within the same graveyard in the north-western corner of Gaelic Medieval Ireland, and therefore it is striking that our analyses demonstrate these two rare cases arise from separate mutational events.

## Acknowledgements

We wish to thank Transport Infrastructure Ireland and Donegal County Council who funded the Ballyhanna Research Project and the National Museum of Ireland for granting the necessary licences for the study. This work was funded by the Science Foundation Ireland/Health Research Board/Wellcome Trust Biomedical Research Partnership Investigator Award No. 205072 to DGB, “Ancient Genomics and the Atlantic Burden” and IJ was supported by The Science Foundation Ireland Centre for Research Training in Genomics Data Science (18/CRT/6214). We acknowledge Trinseq for sequencing support and the DJEI/DES/SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities.

## Declaration of interests

The authors declare no competing interests

## Data and Code availability

Raw FASTQ and aligned BAM files are available from the European Nucleotide Archive (ENA) under accession number XXX.

## Supplemental information

Document S1

Tables S1-S5

## References

1. Lewis, M. (2019). Skeletal Dysplasias and Related Conditions. Ortner’s Identification of Pathological Conditions in Human Skeletal Remains 615–637.
2. McKenzie, C., and Murphy, E. (2011). Health in medieval Ireland : the evidence from Ballyhanna, Co. Donegal. In Past Times, Changing Fortunes: Proceedings of a Public Seminar on Archaeological Discoveries on National Road Schemes, S. Conran, E. Danaher, and M. Stanley, eds. (National Roads Authority), pp. 131–144.
3. Pedrini, E., Jennes, I., Tremosini, M., Milanese, A., Mordenti, M., Parra, A., Sgariglia, F., Zuntini, M., Campanacci, L., Fabbri, N., et al. (2011). Genotype-phenotype correlation study in 529 patients with multiple hereditary exostoses: Identification of “protective” and “risk” factors. *Journal of Bone and Joint Surgery - Series A* 93, 2294–2302.
4. McKenzie, C.J., and Murphy, E.M. (2018). Life and Death in Medieval Gaelic Ireland: The Skeletons from Ballyhanna, Co. Donegal (Four Courts Press).

5. Masada, K., Tsuyuguchi, Y., Kawai, H., Kawabata, H., Noguchi, K., and Ono, K. (1989). Operations for forearm deformity caused by multiple osteochondromas. *J. Bone Joint Surg. Br.*
6. Murphy, E.M., and McKenzie, C.J. (2010). Multiple osteochondromas in the archaeological record: A global review. *J. Archaeol. Sci.* *37*, 2255–2264.
7. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* *6*, 80–92.
8. Cingolani, P., Patel, V.M., Coon, M., Nguyen, T., Land, S.J., Ruden, D.M., and Lu, X. (2012). Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Genet.* *3*, 35.
9. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* *581*, 434–443.
10. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* *4*, 1073–1082.
11. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* *Chapter 7*, Unit7.20.
12. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* *17*, 405–424.
13. Ellard, S., Baple, E.L., Owens, M., Eccles, D.M., Abbs, S., and Zandra, C. (2020). ACGS Best Practice Guidelines for Variant Classification 2020 v4.01.
14. Zuntini, M., Pedrini, E., Parra, A., Sgariglia, F., Gentile, F.V., Pandolfi, M., Alberghini, M., and Sangiorgi, L. (2010). Genetic models of osteochondroma onset and neoplastic progression: Evidence for mechanisms alternative to EXT genes inactivation. *Oncogene* *29*, 3827–3834.
15. Delgado, M.A., Martinez-Domenech, G., Sarrión, P., Urreiziti, R., Zecchini, L., Robledo, H.H., Segura, F., De Kremer, R.D., Balcels, S., Grinberg, D., et al. (2014). A broad spectrum of genomic changes in latinamerican patients with EXT1/EXT2-CDG. *Sci. Rep.* *4*, 1–7.
16. Philippe, C., Porter, D.E., Emerton, M.E., Wells, D.E., Simpson, A.H., and Monaco, A.P. (1997). Mutation screening of the EXT1 and EXT2 genes in patients with hereditary multiple exostoses. *Am. J. Hum. Genet.* *61*, 520–528.
17. McCormick, C., Leduc, Y., Martindale, D., Mattison, K., Esford, L.E., Dyer, A.P., and Tufaro, F. (1998). The putative tumour suppressor EXT1 alters the expression of cell- surface heparan sulfate. *Nat. Genet.* *19*, 158–161.
18. McCormick, C., Duncan, G., Goutsos, K.T., and Tufaro, F. (2000). The putative tumor suppressors EXT1 and EXT2 form a stable complex that accumulates in the Golgi apparatus and catalyzes the synthesis of heparan sulfate. *Proc. Natl. Acad. Sci. U. S. A.* *97*, 668–673.
19. Jennes, I., Pedrini, E., Zuntini, M., Mordenti, M., Balkassmi, S., Asteggiano, C.G., Casey, B., Bakker, B., Sangiorgi, L., and Wuyts, W. (2009). Multiple osteochondromas: Mutation update and description of the Multiple Osteochondromas Mutation Database (MOdb). *Human Mutation* *30*, 1620–1627.
20. Pacifici, M. (2017). Hereditary Multiple Exostoses: New Insights into Pathogenesis, Clinical Complications and Potential Treatments. *Curr. Osteoporos. Rep.* *15*, 142–152.

21. Reijnders, C.M.A., Waaijer, C.J.F., Hamilton, A., Buddingh, E.P., Dijkstra, S.P.D., Ham, J., Bakker, E., Szuhai, K., Karperien, M., Hogendoorn, P.C.W., et al. (2010). No haploinsufficiency but loss of heterozygosity for EXT in multiple osteochondromas. *Am. J. Pathol.* *177*, 1946–1957.
22. Zak, B.M., Schuksz, M., Koyama, E., Mundy, C., Wells, D.E., Yamaguchi, Y., Pacifici, M., and Esko, J.D. (2011). Compound heterozygous loss of Ext1 and Ext2 is sufficient for formation of multiple exostoses in mouse ribs and long bones. *Bone* *48*, 979–987.
23. Tatsi, C., and Stratakis, C.A. (2020). The Genetics of Pituitary Adenomas. *Journal of Clinical Medicine* *9*, 30.
24. Chahal, H.S., Stals, K., Unterländer, M., Balding, D.J., Thomas, M.G., Kumar, A.V., Besser, G.M., Atkinson, A.B., Morrison, P.J., Howlett, T.A., et al. (2011). AIP Mutation in Pituitary Adenomas in the 18th Century and Today. *N. Engl. J. Med.* *364*, 43–50.
25. Beckers, A., Fernandes, D., Fina, F., Novak, M., Abati, A., Rostomyan, L., Thiry, A., Ouafik, L., 'housine, Pasture, B., Pinhasi, R., et al. (2017). Paleogenetic study of ancient DNA suggestive of X-linked acroigantism. *Endocr. Relat. Cancer* *24*, L17–L20.
26. Boer, L.L., Naue, J., de Rooy, L., and Oostra, R.-J. (2017). Detection of G1138A Mutation of the FGFR3 Gene in Tooth Material from a 180-Year-Old Museological Achondroplastic Skeleton. *Genes* *8*,
27. Byrne, R.P., Martiniano, R., Cassidy, L.M., Carrigan, M., Hellenthal, G., Hardiman, O., Bradley, D.G., and McLaughlin, R.L. (2018). Insular Celtic population structure and genomic footprints of migration. *PLoS Genet.* *14*, e1007152.
28. Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., Lawson, D.J., et al. (2015). The fine-scale genetic structure of the British population. *Nature* *519*, 309–314.
29. International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C.A., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., Su, Z., et al. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* *476*, 214–219.
30. Batini, C., Hallast, P., Zadik, D., Delsler, P.M., Benazzo, A., Ghirotto, S., Arroyo-Pardo, E., Cavalleri, G.L., de Knijff, P., Dupuy, B.M., et al. (2015). Large-scale recent expansion of European patrilineages shown by population resequencing. *Nat. Commun.* *6*, 7152.
31. Moore, L.T., McEvoy, B., Cape, E., Simms, K., and Bradley, D.G. (2006). A Y-Chromosome Signature of Hegemony in Gaelic Ireland. *Am. J. Hum. Genet.* *78*, 334–338.
32. Rubinacci, S., Ribeiro, D.M., Hofmeister, R.J., and Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* *53*, 120–126.
33. O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genet.* *10*, e1004234.
34. Margaryan, A., Lawson, D.J., Sikora, M., Racimo, F., Rasmussen, S., Moltke, I., Cassidy, L.M., Jørsboe, E., Ingason, A., Pedersen, M.W., et al. (2020). Population genomics of the Viking world. *Nature* *585*, 390–396.
35. Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. *Nature* *522*, 167–172.
36. Krzewińska, M., Kılınc, G.M., Juras, A., Koptekin, D., Chyleński, M., Nikitin, A.G., Shcherbakov, N.,

- Shuteleva, I., Leonova, T., Kraeva, L., et al. (2018). Ancient genomes suggest the eastern Pontic-Caspian steppe as the source of western Iron Age nomads. *Science Advances* *4*, eaat4457.
37. Marnetto, D., and Huerta-Sánchez, E. (2017). Haplostrips: revealing population structure through haplotype visualization. *Methods Ecol. Evol.* *8*, 1389–1392.
38. Antunes-Ferreira, N., Cunha, E., and Marques, C. (2014). Multiple osteochondromas in a 16th–19th century individual from Setúbal (Portugal). *Anthropol. Sci.* *122*, 157–163.
39. Krooth, R.S., Macklin, M.T., and Hilbish, T.F. (1961). Diaphysial aclasis (multiple exostoses) on Guam. *Am. J. Hum. Genet.* *13*, 340–347.
40. Black, B., Dooley, J., Pyper, A., and Reed, M. (1993). Multiple Hereditary Exostoses An Epidemiologic Study of an Isolated Community in Manitoba. *Clin. Orthop. Relat. Res.* *287*, 212–217.
41. Wuyts, W., Schmale, G.A., Chansky, H.A., and Raskind, W.H. (2020). Hereditary Multiple Osteochondromas (Seattle (WA): University of Washington, Seattle).

## Methods

### Sampling and Sequencing Preparation

The petrous bones of two individuals with Multiple Osteochondromas from Ballyhanna, Co. Donegal (Excavation number 03E1384) were sampled for DNA extraction and shotgun sequencing (Licence to Alter Number 6861, National Museum of Ireland).

Bones were photographed before modification, and exposed to UV light for 30 minutes on either side to remove surface contaminants. The surface of each bone was cleaned using a drill bit, and a triangular section of the otic capsule was cut. Half was stored as a bone fragment, and the other half was pulverised in a Mixer Mill 400 (Retsch). Extracts were prepared from 0.1g of this powder. One extract of Sk331 was prepared as in Boessenkool et al (2017)<sup>1</sup>: bone powder was initially incubated in 0.5% bleach for 15 minutes to remove contaminants. After centrifuging for a few seconds at 13000 rpm, the bleach supernatant was discarded, and the remaining pellet was washed three times with UV water. The pellet was then incubated in UV-ed EDTA at 37°C for 30 minutes at 900rpm. This was then centrifuged at 13000 rpm for 10 minutes and the supernatant was removed and stored. The remaining pellet was incubated in UV-ed extraction buffer with proteinase K for 48 hours at 37°C and 900rpm. This was centrifuged at 13000 rpm for 10 minutes, and the supernatant was transferred to an Amicon Ultra-4 Centrifugal Filter unit 30kDa, diluted in 3mL 10 mM Tris-EDTA buffer and centrifuged at 5000 rpm until 100uL of solution remained. This volume was then added to a silica column (MinElute PCR purification kit, Qiagen, Hilden, Germany) and purified according to the manufacturer's instructions. Sk197 was extracted following a similar protocol but with some modifications as in Dabney et al. (namely, shorter incubation with proteinase K (24hr), lower centrifugation speeds and the use of Zymo-Spin V columns (Zymo Research) rather than the Amicon columns)<sup>2</sup>. An additional extract from the initial EDTA wash was prepared for Sk197.

3 sequencing libraries were prepared from 16.25uL of each extract (total: 9 libraries). Each of these were USER-treated to mitigate the effects of post-mortem damage. Libraries were prepared as in Meyer and Kircher (2010), with modifications as in Gamba et al (2014)<sup>3,4</sup>. Libraries and controls were amplified using Accuprime Pfx (Life technology); indexing primers p5 and p7; 10x reaction buffer and 3uL of library/water with a total reaction volume of 25uL. The first libraries for each extract were amplified for 14 cycles for an initial screen. The remaining PCRs were amplified for 10-13 cycles for high coverage sequencing. PCRs were purified with the MinElute PCR purification kit, and the concentrations of these products were quantified using the Agilent TapeStation system with a D1000 screentape (Sk331) or a QuBit with the High Sensitivity dsDNA assay (Invitrogen) (Sk197). Samples were screened and sequenced to a high read depth on a Novaseq 6000 platform using paired-end 50bp reads.

### Alignment and Data Processing

The quality of demultiplexed fastq files was assessed using the FASTQC suite<sup>5</sup>. Adapter sequences were trimmed using AdapterRemoval v2.2.2 and paired end reads were collapsed<sup>6</sup>. Trimmed reads were aligned to the hs37d5 reference genome using bwa aln with relaxed parameters and the seed disabled (-o 2 -n 0.01 -l 16500)<sup>7</sup>. PCR and optical duplicates were removed using PicardTools v2.22.1 MarkDuplicates (<http://broadinstitute.github.io/picard>), and reads were filtered for mapping quality > 20 using samtools<sup>8</sup>. Bam files were merged to sample

level and indel realignment was performed using GATK<sup>9</sup>. Reads were “softclipped” by reducing the quality score of the first and last two bases of each read to a PHRED score of two. Bam files were then fished for exact p7 index matches, as one base-pair mismatches had been allowed during demultiplexing.

Mitochondrial and X chromosome contamination estimates were used to assess authenticity of this data. Mitochondrial contamination was estimated using the frequency of heteroplasmic sites when aligned to the revised Cambridge Reference sequence, both including and excluding possible postmortem damage. X chromosome contamination levels were estimated using ANGSD as in Rasmussen et al. (2011)<sup>10</sup>.

### Uniparental Haplogroups

Mitochondrial consensus sequences were estimated using samtools mpileup and vcfutils<sup>11</sup>. Each site had a minimum read depth of 5 and base quality 30. Consensus fasta files were used as input to Mitomaster, which was used to estimate mitochondrial haplogroups<sup>12</sup>.

Y haplogroups were investigated by calling representative SNPs of likely haplogroups from the International Society of Genetic Genealogy 2020 Y-DNA tree (v. 15.58)<sup>13</sup>. Base calls across these loci were generated for both samples using GATK’s pileup tool, and loci with derived alleles were flagged. These were then manually assessed to call Y haplogroups.

### Pseudohaploid SNP calling

GATK was used to pileup SNPs of interest for pseudohaploid analyses from the final BAM files (see **Table S6** for details). These sites were filtered for minimum base quality 30, and one read was sampled at random to generate pseudohaploid calls.

### Modern Population Affinities

#### *Outgroup $f_3$ statistics*

Shared drift with modern human populations was initially assessed using outgroup  $f_3$  statistics using Admixtools<sup>14</sup>. Pseudohaploid SNPs were called as above, samples were merged with the Human Origins dataset, and statistics of the form  $f_3(\text{Mbuti}; \text{Ballyhanna}, \text{Modern\_Population})$  were calculated.

#### *Principal Components Analysis*

Principal components were calculated on a dataset of modern Northwestern Europeans (including Irish individuals)<sup>15–17</sup>, and Sk331 and 197 were projected onto this using smartpca v8000 with shrinkmode:YES<sup>18,19</sup>. No outlier iterations were performed, and SNPs in LD were eliminated with an  $r^2$  threshold  $> 0.2$ .

### Exome Scan

A scan to search for possible dominant acting variants was developed. Bcftools v1.10.2 was used to call possible variant sites in autosomal exons from bam files aligned to hs37d5<sup>11</sup>, and calls were filtered for quality using the command

```
“bcftools mpileup -R  $\{\text{bed}\}$  --count-orphans --annotate FORMAT/AD,INFO/AD --ignore-RG --fasta-ref  $\{\text{ref}\}$   $\{\text{bamfile}\}$  -Ou | bcftools call -mv --format-fields GQ,GP -Ou | bcftools filter -s LowQual -Oz -o  $\{\text{id}\}$ .exome.unsorted.vcf.gz”.
```

Co-ordinates of autosomal exons were downloaded from the UCSC table browser tool on 15-09-2020 in BED format using the NCBI RefSeq database, including 2bp up- and downstream of the coding sequences<sup>20</sup>.

SnEff and SnpSift were used to annotate these possible variants in terms of molecular impact on canonical transcripts<sup>21,22</sup>. Possible variants were filtered for minimum allelic depth 3 and maximum read depth less than twice the mean genomic coverage. A minimum genotyping quality of 50 was used. Sites were then filtered for high or moderate molecular impact, and variants in genes with a high probability of loss of function intolerance ( $pLI > 90\%$ ) according to gnomADv2.1.1 were retained<sup>23</sup>. These variants were filtered for low ( $<1\%$ ) allele frequency in gnomad and predicted pathogenic effect by both SIFT and Polyphen2 (SNPs) or just SIFT (indels)<sup>24,25</sup>. The ACMG classification guidelines were used to assess support for the candidate pathogenic variants<sup>26</sup>.

## Imputation

Autosomal imputation was performed on both individuals using GLIMPSE<sup>27</sup>. Genotype likelihoods for autosomal biallelic SNPs in the 1000 genomes project phase 3 panel were called for each individual using bcftools v1.10.2 as in the GLIMPSE documentation<sup>11,27</sup>. Individuals were merged on a per-chromosome basis and imputed in chunks of 2Mb with a 200kB buffer region, using the Phase 3 panel of the 1000 genomes project as a reference dataset<sup>28</sup>. A genotype probability filter of 99% was used for all downstream analyses of this data.

## EXT1 Haplotypes

Relationships between the haplotypes at the disease gene (*EXT1*) in affected individuals was investigated in the context of a larger imputed dataset of published Iron Age and Historical samples<sup>29-32</sup>. This data was phased using SHAPEITv2 with the 1000 Genomes Project Phase 3 reference panel<sup>28,33</sup>. The region spanned by *EXT1* exons and 50kB up- and down-stream were visualised using the tool haplostrips<sup>34</sup>.

## Radiocarbon Dating

Radiocarbon dating was undertaken on two rib samples. The dates were repeated to ensure accuracy.

Lab Code	Sample/Context	Years BP	Calibrated Date Ranges
UBA-11442	Rib from Sk197	1223 +/- 27	AD 708-877 (1 $\sigma$ ); AD 689-885 (2 $\sigma$ )
UBA-11442 Repeat		1162 +/- 24	AD 775-950 (1 $\sigma$ ); AD 774-975 (2 $\sigma$ )

UBA-11443	Rib from Sk331	856 +/- 24	AD 1175-1221 (1 $\sigma$ ); AD 1158-1260 (2 $\sigma$ )
UBA-11443 Repeat		928 +/- 40	AD 1042-1162 (1 $\sigma$ ); AD 1031-1211 (2 $\sigma$ )

Calibration: Reimer et al, 2020 <sup>35</sup>

## References

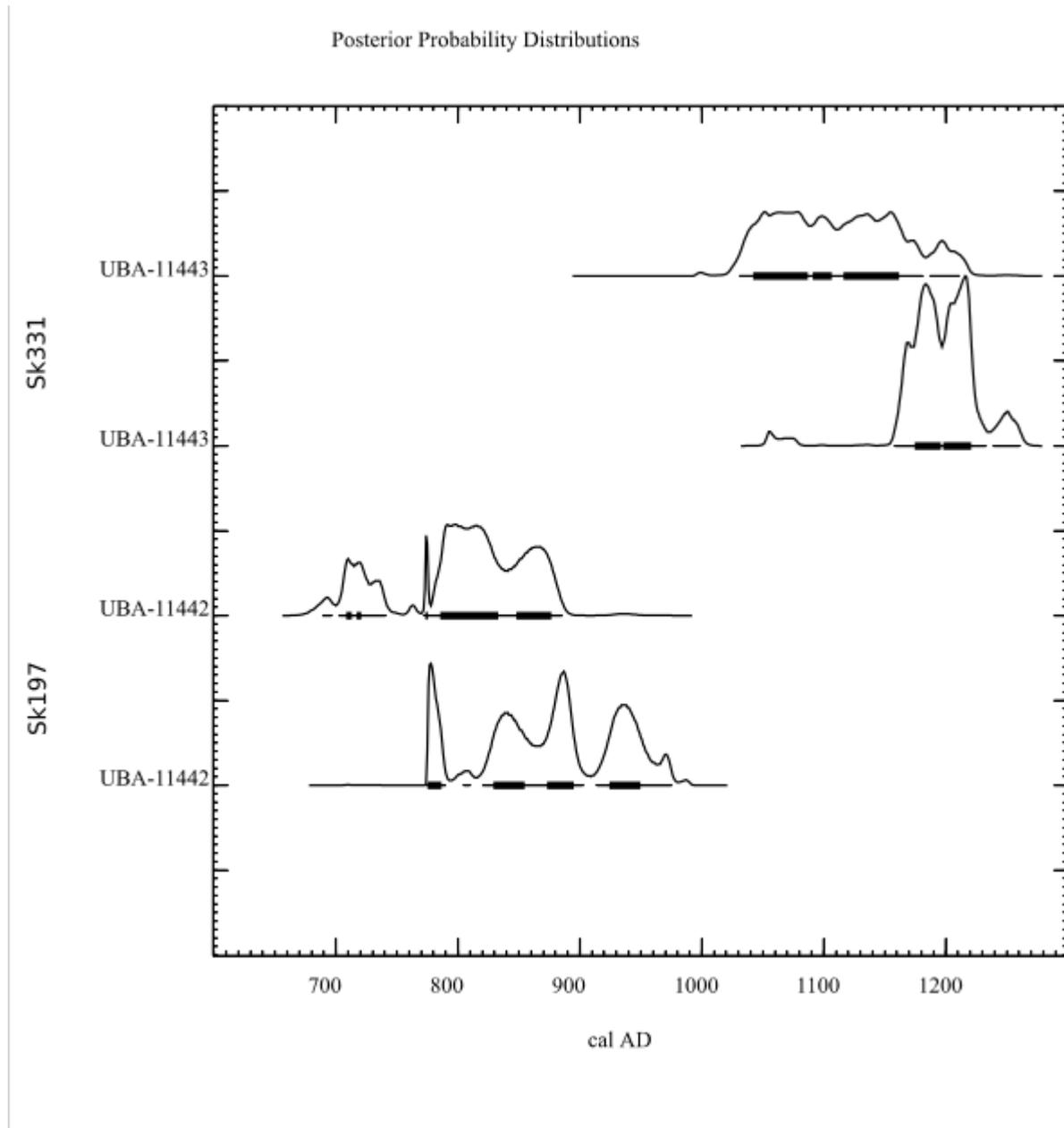
1. Boessenkool, S., Hanghøj, K., Nistelberger, H.M., Der Sarkissian, C., Gondek, A.T., Orlando, L., Barrett, J.H., and Star, B. (2017). Combining bleach and mild predigestion improves ancient DNA recovery from bones. *Mol. Ecol. Resour.* 17, 742–751.
2. Dabney, J., Knapp, M., Glocke, I., Gansauge, M.T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.L., et al. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* 110, 15758–15763.
3. Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 5,.
4. Gamba, C., Jones, E.R., Teasdale, M.D., McLaughlin, R.L., Gonzalez-Fortes, G., Mattiangeli, V., Domboróczki, L., Kővári, I., Pap, I., Anders, A., et al. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* 5, 5257.
5. Andrews, S. (2010). FastQC.
6. Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9, 88.
7. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.
8. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
9. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303.
10. Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., Jombart, T., et al. (2011). An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* 334, 94–98.
11. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993.
12. Lott, M.T., Leipzig, J.N., Derbeneva, O., Xie, H.M., Chalkia, D., Sarmady, M., Procaccio, V., and Wallace, D.C. (2013). mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr. Protoc. Bioinformatics* 44, 1.23.1–1.23.26.
13. International Society of Genetic Genealogy (2020). Y-DNA Haplogroup Tree 2020, Version: 15.58,

Date: 21 May 2020.

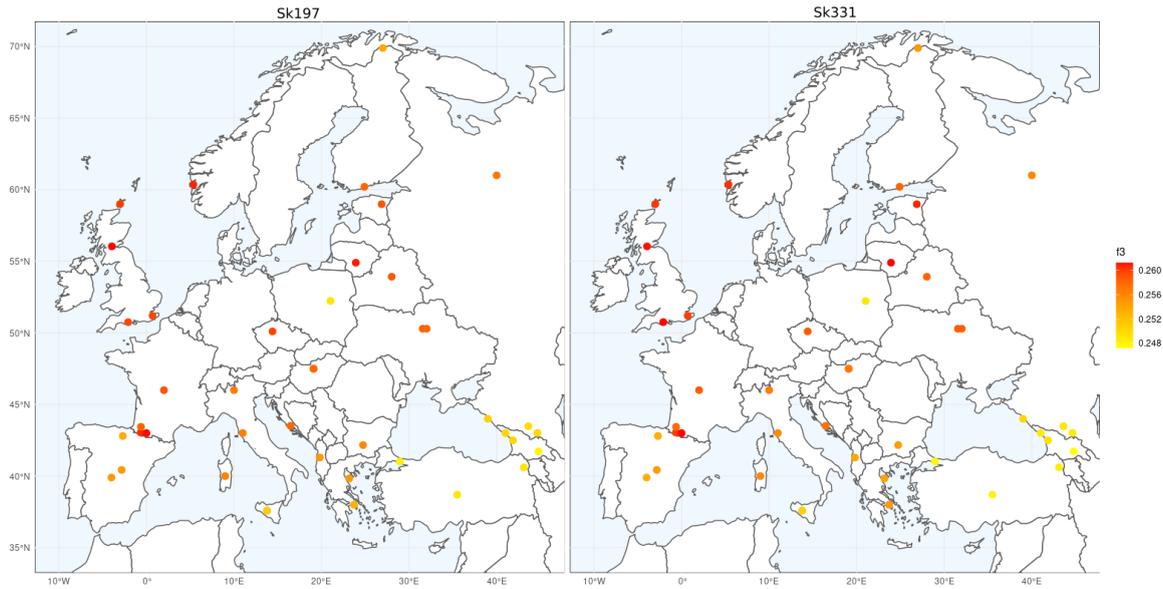
14. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. *Genetics* *192*, 1065–1093.
15. Byrne, R.P., Martiniano, R., Cassidy, L.M., Carrigan, M., Hellenthal, G., Hardiman, O., Bradley, D.G., and McLaughlin, R.L. (2018). Insular Celtic population structure and genomic footprints of migration. *PLoS Genet.* *14*, e1007152.
16. Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., Lawson, D.J., et al. (2015). The fine-scale genetic structure of the British population. *Nature* *519*, 309–314.
17. International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C.A., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., Su, Z., et al. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* *476*, 214–219.
18. Patterson, N., Price, A.L., and Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genet.* *2*, e190.
19. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* *38*, 904–909.
20. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* *32*, D493–D496.
21. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* *6*, 80–92.
22. Cingolani, P., Patel, V.M., Coon, M., Nguyen, T., Land, S.J., Ruden, D.M., and Lu, X. (2012). Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Genet.* *3*, 35.
23. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* *581*, 434–443.
24. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* *4*, 1073–1082.
25. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* *7*, 248–249.
26. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* *17*, 405–424.
27. Rubinacci, S., Ribeiro, D.M., Hofmeister, R.J., and Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* *53*, 120–126.
28. The 1000 Genomes Project Consortium, Gibbs, R.A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., et al. (2015). A global reference for human genetic variation. *Nature* *526*, 68–74.

29. Amorim, C.E.G., Vai, S., Posth, C., Modi, A., Koncz, I., Hakenbeck, S., La Rocca, M.C., Mende, B., Bobo, D., Pohl, W., et al. (2018). Understanding 6th-century barbarian social organization and migration through paleogenomics. *Nat. Commun.* *9*, 3547.
30. Margaryan, A., Lawson, D.J., Sikora, M., Racimo, F., Rasmussen, S., Moltke, I., Cassidy, L.M., Jørsboe, E., Ingason, A., Pedersen, M.W., et al. (2020). Population genomics of the Viking world. *Nature* *585*, 390–396.
31. Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. *Nature* *522*, 167–172.
32. Krzewińska, M., Kılınc, G.M., Juras, A., Koptekin, D., Chyleński, M., Nikitin, A.G., Shcherbakov, N., Shuteleva, I., Leonova, T., Kraeva, L., et al. (2018). Ancient genomes suggest the eastern Pontic-Caspian steppe as the source of western Iron Age nomads. *Science Advances* *4*, eaat4457.
33. O’Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genet.* *10*, e1004234.
34. Marnetto, D., and Huerta-Sánchez, E. (2017). Haplostrips: revealing population structure through haplotype visualization. *Methods Ecol. Evol.* *8*, 1389–1392.
35. Reimer, P.J., Austin, W.E.N., Bard, E., Bayliss, A., Blackwell, P.G., Ramsey, C.B., Butzin, M., Cheng, H., Lawrence Edwards, R., Friedrich, M., et al. (2020). The IntCal20 Northern Hemisphere Radiocarbon Age Calibration Curve (0–55 cal kBP). *Radiocarbon* *62*, 725–757.

## Supplementary figures



**Figure S1. Radiocarbon dates for Sk197 and Sk331.** Posterior distribution of estimated age is displayed, where the x-axis is in calendar years AD. UBA-11442 is Sk197 and UBA-11443 is Sk331.



**Figure S2. Outgroup  $f_3$  statistics with European Populations.** Shared drift with modern European populations assessed using the statistic  $f_3$ (Mbuti; Ballyhanna, Modern\_Population). Higher values are closer to red and indicate more shared drift.