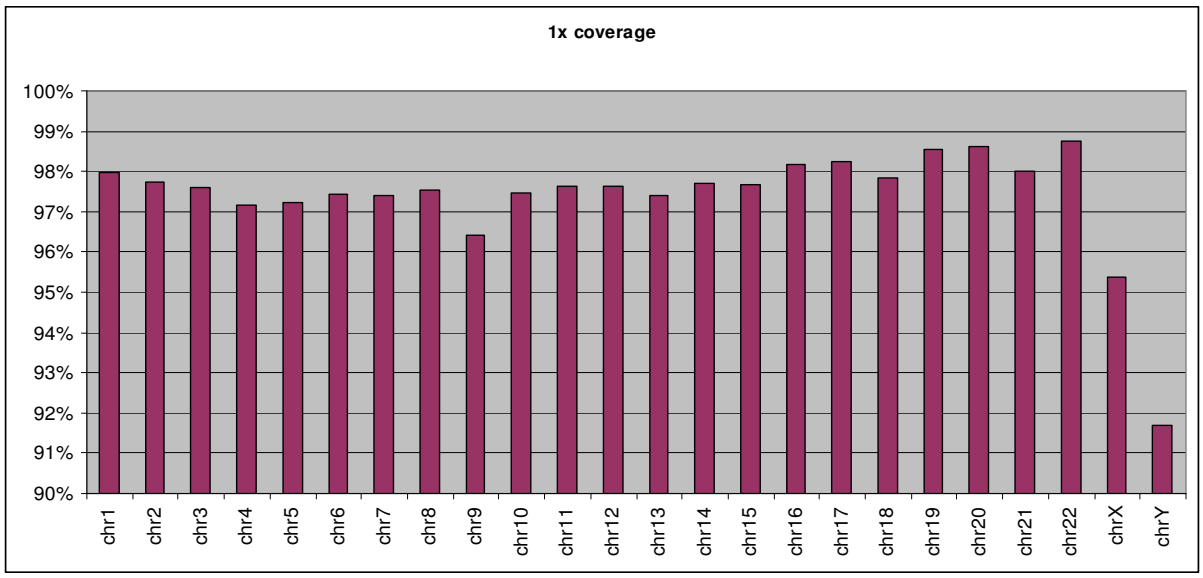
Supplementary Information

New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing

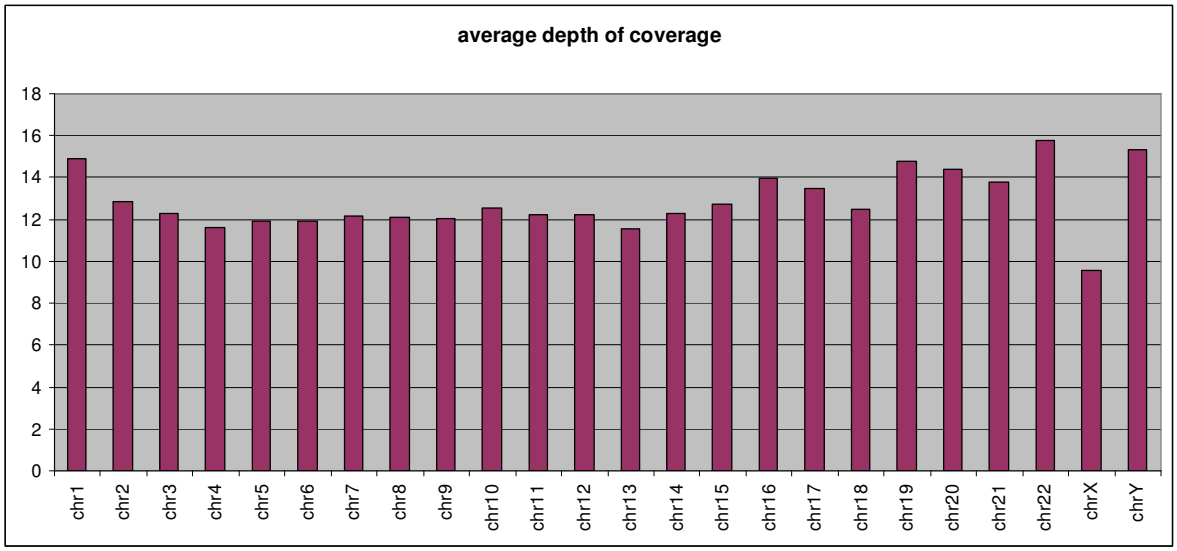
Andreas Keller, Angela Graefen, Markus Ball, Mark Matzas, Valesca Boisguerin, Frank Maixner, Petra Leidinger, Christina Backes, Rabab Khairat, Michael Forster, Björn Stade, Andre Franke, Jens Mayer, Jessica Spangler, Stephen McLaughlin, Minita Shah, Clarence Lee, Timothy T. Harkins, Alexander Sartori, Andres Moreno-Estrada, Brenna Henn, Martin Sikora, Ornella Semino, Jacques Chiaroni, Siiri Rootsi, Natalie M. Myres, Vicente M. Cabrera, Peter A.Underhill, Carlos D. Bustamante, Eduard Egarter Vigl, Marco Samadelli, Giovanna Cipollini, Jan Haas, Hugo Katus, Brian D. O'Connor, Marc R.J. Carlson, Benjamin Meder, Nikolaus Blin, Eckart Meese, Carsten M. Pusch, Albert Zink

Supplementary Figures

a

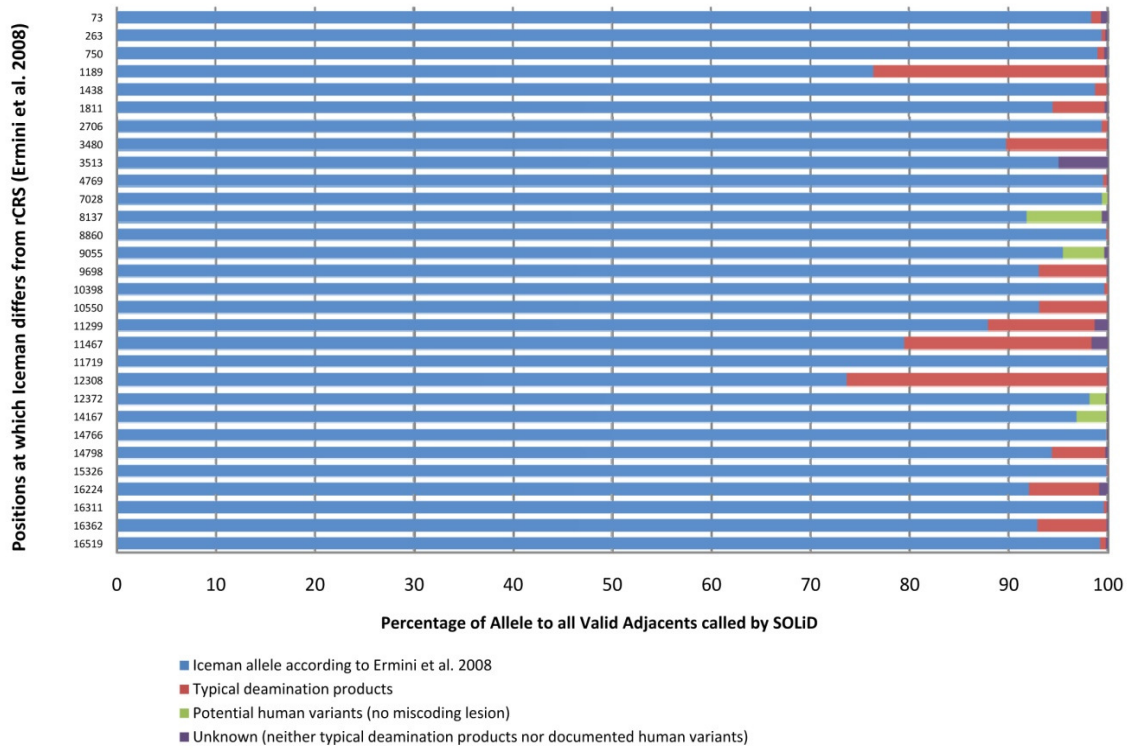


b

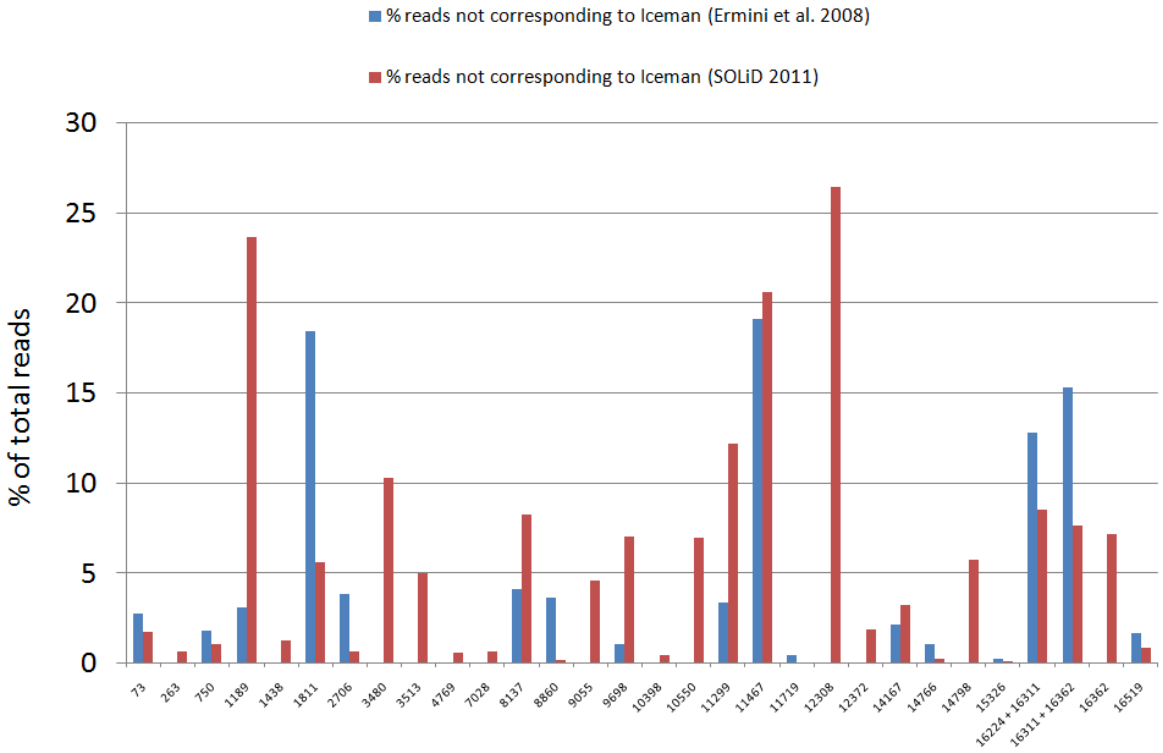


Supplementary Figure S1: Percentage of covered chromosomes and average coverage of each chromosome. a. For each chromosome, the fraction of the chromosome covered by at least a single read is shown. **b.** For each chromosome, the averaged depth of coverage, i.e., the number of reads mapping on average to that chromosome, is shown**.**

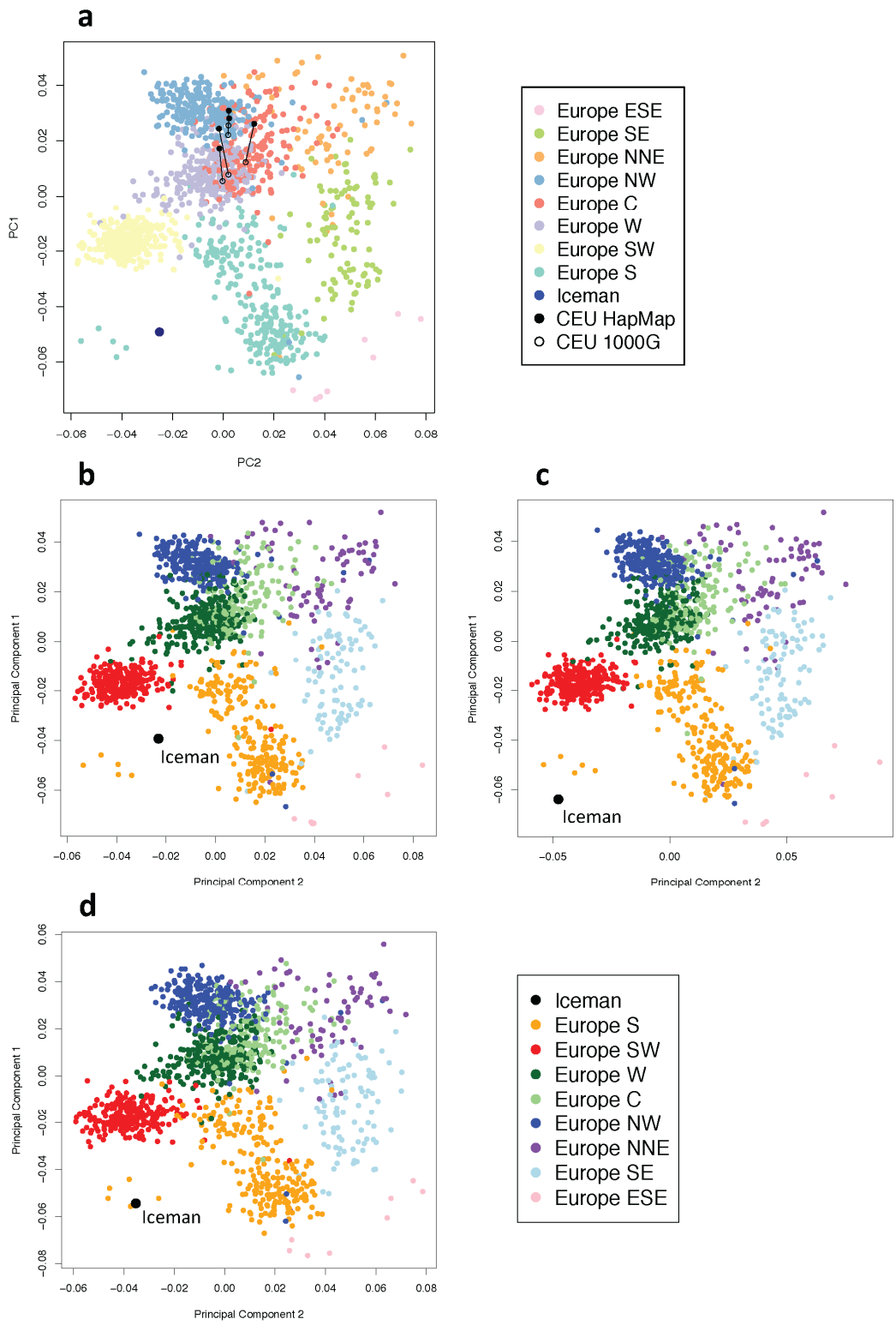
a



b



Supplementary Figure S2: **Comparison of the data generated in our study with previously published Iceman data. a.** Comparison of SOLiD4 Iceman mitochondrial allelic ratios with reviously published Iceman genomic data (Ermini et al. 2008)[3]. **b.** Comparison of non-Iceman allele reads in SOLiD4 Iceman mitochondrial data and previously published Iceman data (Ermini et al. 2008)[3].



Supplementary Figure S3: Principal Component Analysis of the Iceman genome combined with 1,387 Europeans from PopRes using different filter settings. a. The Iceman and five CEU individuals from HapMap and the low-coverage 1000 Genomes pilot 1 data projected onto the principal components inferred from the full dataset. Black dots indicate the position of the CEU individuals using the HapMap SNP array data. Black circles connected with a line indicate the position of the same individuals using the unfiltered genotypes called from the sequencing data. **b.** Unfiltered PCA using 193,723 SNPs after merging with PopRes data, including 31.6% of missing genotypes in the Iceman. **c.** PCA filtered for missing genotypes at 6x in the Iceman (132,981 SNPs used). **d.** PCA filtered for missing genotypes at 12x in the Iceman (49,597 SNPs used).

Supplementary Tables

Supplementary Table S1: Summary of sequencing results. Calculations are based on the number of unambiguous reads in the genome assembly hg18.

	uncovered bases	%covered	%uncovered	N-adjusted avg coverage (paired coverage)	N-adjusted avg coverage (singleton)	N-adjusted avg coverage (paired+singleton)
chr1	29535229	96.76%	3.24%	4.84	3.22	8.06
chr2	13679277	96.45%	3.55%	4.64	3.07	7.72
chr3	12072420	96.26%	3.74%	4.41	3	7.41
chr4	12208405	95.60%	4.40%	4.36	2.89	7.24
chr5	10543891	95.84%	4.16%	4.42	2.91	7.33
chr6	10314957	96.00%	4.00%	4.48	2.91	7.39
chr7	10104940	95.98%	4.02%	4.63	2.9	7.53
chr8	9099189	96.19%	3.81%	4.7	2.91	7.6
chr9	26088777	95.04%	4.96%	4.67	2.83	7.5
chr10	8806644	96.16%	3.84%	5.25	2.92	8.17
chr11	8123968	96.34%	3.66%	4.74	2.9	7.64
chr12	6980634	96.21%	3.79%	4.63	2.89	7.52
chr13	22595293	95.80%	4.20%	4.25	2.86	7.11
chr14	21371189	96.27%	3.73%	4.62	2.92	7.54
chr15	21932506	96.39%	3.61%	4.95	2.93	7.88
chr16	12302608	97.01%	2.99%	6.14	3.02	9.16
chr17	3395086	96.89%	3.11%	5.28	2.95	8.23
chr18	4125016	96.43%	3.57%	4.73	2.93	7.66
chr19	9610374	97.16%	2.84%	5.68	2.94	8.63
chr20	4391975	97.54%	2.46%	5.52	3.08	8.59
chr21	13922038	96.64%	3.36%	5.28	2.95	8.23
chr22	15628184	97.74%	2.26%	6.11	3	9.1
chrX	14848723	92.72%	7.28%	2.73	2.03	4.76
chrY	34784695	89.61%	10.39%	9.05	1.74	10.8
mitochondrial	0	100.00%	0.00%	1177.09	0.92	1178.01

Supplementary Table S2: Results of the SNP analysis for the Iceman NGS data

iceman	A/C	A/G	A/T	C/G	C/T	G/T	all
heterozygous	72946	321889	57564	62892	320556	72720	908567
homozygous	65539	269863	49796	72034	269430	65655	792317
nonsense	36	111	54	9	115	32	357
readthrough	2	6	6	5	2	4	25
startcancelled	1	7	1	1	4	0	14
missense	793	3844	491	688	3853	841	10510
synonymous	345	3836	177	339	3747	376	8820
Splice Site	8	124	2	3	136	6	279
UTR	8818	38924	6453	8958	39221	8844	111218
intronic	41337	187664	30829	41327	187107	41093	529357

Supplementary Table S3: Summary of the SNP analysis

Total number of SNPs	1700884
Total number of NOVEL SNPs	528975
all synonymous-coding SNPs	8820
all missense SNPs	10510
all canceledstart SNPs	14
all readthrough SNPs	25
all nonsense SNPs	357
NOVEL missense SNPs	6155
NOVEL canceled start SNPs	7
NOVEL readthrough SNPs	10
NOVEL nonsense SNPs	330
SNPs in Acceptor	67
SNPs in Donor	212
SNPs in 5'UTR	3625
SNPS in 3'UTR	12699
SNPs in UTR-Splicesites	94894
SNPS in Introns	529357
unknown/intergenic SNPs	1040305
Polyphen Predictions	3953
SNAP Predictions	1644
SIFT Predictions	7310
SNPs3d Predictions	3835
BAD Polyphen Predictions	407
possibly BAD Polyphen Predictions	628
BAD Snap Predictions	704
BAD SIFT Predictions	604
BAD SNPs3d Predictions	383
SNPs associated with an disease (HGMD)	290
SNPs in CpG-Islands	13735
SNPs in promotors	198
heteroygosity	0.534173
Ti/Tv ratio	1.67544

Supplementary Table S4: Comparison between Iceman mitochondrial genomic data generated in this study and variants identified in Iceman mitochondrial genome by Ermini et al. 2008[3]

rCRS Position (NC_012920)	rCRS (and other known human variants according to	Iceman allele as published by Ermini	Allele count SOLiD Iceman mitochondrial genomic data generated in this study						% of alleles corresponding to published Iceman genome (Ermini et al. 2008)[3]	% of alleles constituting typical deamination products or human variants	% of alleles representing known human mtDNA sequences (no deamination products)	Other (e.g. bacterial origin, sequencing errors)
			A	C	G	T	N	Total				
73	A	G	20	0	1954	0	14	1988	98,29	1,01	0,00	0,70
263	A	G	41	14	11669	18	1	11743	99,37	0,35	0,00	0,28
750	A	G	56	0	8061	7	25	8149	98,92	0,69	0,00	0,39
1189	T	C	1	1178	4	360	0	1543	76,34	23,33	0,00	0,32
1438	A	G	31	0	2590	3	0	2624	98,70	1,18	0,00	0,11
1811	A	G	149	3	2692	7	0	2851	94,42	5,23	0,00	0,35
2706	A	G	26	1	4538	2	1	4568	99,34	0,57	0,00	0,09
3480	A	G	222	1	1959	1	1	2184	89,70	10,16	0,00	0,14
3513	C	T	0	66	6	1361	0	1433	94,98	0,00	0,00	5,02
4769	A	G	12	1	2904	1	2	2920	99,45	0,41	0,00	0,14
7028	C	T	1	27	2	4694	0	4724	99,36	0,00	0,57	0,06
8137	C	T	3	244	2	2943	16	3208	91,74	0,00	7,61	0,65
8860	A	G	7	0	3609	0	0	3616	99,81	0,19	0,00	0,00
9055	G	A	1995	2	87	0	7	2091	95,41	0,00	4,16	0,43
9698	T	C	0	2036	2	151	1	2190	92,97	6,89	0,00	0,14
10398	A	G	25	0	6006	1	0	6032	99,57	0,41	0,00	0,02
10550	A	G	74	0	1000	1	0	1075	93,02	6,88	0,00	0,09
11299	T	C	0	1650	0	202	26	1878	87,86	10,76	0,00	1,38
11467	A	G	154	3	648	8	3	816	79,41	18,87	0,00	1,72
11719	G	A	6896	0	0	1	2	6899	99,96	0,00	0,00	0,04
12308	A	G	667	0	1866	2	1	2536	73,58	26,30	0,00	0,12
12372	G	A	727	1	12	0	1	741	98,11	0,00	1,62	0,27
14167	C	T	1	87	0	2773	4	2865	96,79	0,00	3,04	0,17
14766	C(G)	T	4	1	0	2247	0	2252	99,78	0,00	0,04	0,18
14798	T	C	0	859	1	49	2	911	94,29	5,38	0,00	0,33
15326	A	G	4	0	3018	0	0	3022	99,87	0,13	0,00	0,00
16224	T	C	0	2035	2	156	19	2212	92,00	7,05	0,00	0,95
16311	T	C	0	1836	0	6	3	1845	99,51	0,33	0,00	0,16
16362	T(G)	C	1	1026	0	77	1	1105	92,85	6,97	0,00	0,18
16519	T	C	5	5464	8	33	1	5511	99,15	0,60	0,00	0,25
Avg									**94,49**	**4,46**	**2,13***	**0,49**

* In the case of C>T and G>A variation, it is not possible to determine whether this is a deamination of the Iceman sequence or a potential human contaminant (as the allele produced by deaminated sites is also a known human variant in each of the listed positions). Therefore, to determine the average ratio of potential human contamination, only those positions less prone to miscoding lesions via deamination (namely those where the Iceman has an A or T allele) were regarded. Specific changes to these sites may be caused by human contamination, but also by sequencing errors or non-human contamination.

Supplementary Table S5: Calling of heterozygous and homozygous SNPs

	Number of SNPs	SNPs in dbSNP	% in dbSNP	Hetero	Hetero in dbSNP	% Hetero in dbSNP	Homo	Homo in dbSNP	% Homo in dbSNP
F3 singletons	812,780	683,577	84.10%	145,800	88,564	60.74%	666,980	595,013	89.21%
F5 singletons	5,146	3,054	59.35%	353	137	38.81%	4,793	2,917	60.86%
F3 singletons + paired	2,209,840	1,602,354	72.51%	1,112,383	601,677	54.09%	1,097,457	1,000,677	91.18%
F3 singletons + F5 singletons + paired	2,218,163	1,610,200	72.59%	1,147,380	630,400	54.94%	1,070,783	979,800	91.50%

Supplementary Table S6: Identified SNPs of clinical/functional relevance

Chr.	dbSNP#	1000 Genomes minor allele frequency	HG18 position	Gene	Coverage Iceman					SNP Association	Reference	Type
					A	C	G	T	N			
chr1	rs228648	A=0.476	7836017	UTS2	23	0	1	0	2	Diabetes, type 2, association with	HGMD[41] #CM078165	mis
chr1	rs1801133	A=0.325	11778965	MTHFR	10	0	1	0	0	Cardiovascular disease, association with	HGMD #CM950819	mis
chr1	rs1764391	T=0.354	35033356	GJA4	0	4	0	9	0	Atherosclerosis, association with	HGMD #CM994122	mis
chr1	rs945508	T=0.277	155173705	ARHGEF11	0	2	0	3	0	Diabetes, type 2, association with ?	HGMD #CM074015	mis
chr2	rs4988235	A=0.219	136325116	MCM6	0	0	14	0	0	Lactase nonpersistence	[20]	int
chr2	rs6544718	T=0.098	43958429	ABCG8	0	16	0	4	0	Incr. serum cholesterol, in low-chol. consumers, assoc. with	HGMD #CM067333	mis
chr2	rs1143634	A=0.146	113306861	IL1B	5	0	8	0	0	Malaria, severity, association with	HGMD #CM040228	syn
chr2	rs1990760	T=0.353	162832297	IFIH1	0	3	0	18	0	Diabetes, type 1, assoc. with	HGMD #CM066881	mis
chr2	rs16858808	A=0.029	218737177	CXCR1	5	0	12	0	0	AIDS progression, protection, association with	HGMD #CM066576	mis
chr3	rs3732379	T=0.163	39282260	CX3CR1	0	14	0	5	0	HIV infection, susceptibility to, association	HGMD #CM000504	mis
chr3	rs1126478	T=0.408	46476217	LTF	0	5	0	12	1	Periodontitis, aggressive, association with	HGMD #CM096382	mis
chr3	rs4917	T=0.281	187820407	AHSG	0	11	0	4	1	Leanness, association with	HGMD #CM052829	mis
chr4	rs3755863	T=0.388	23424620	PPARGC1A	0	2	0	11	0	Diabetes, type 2, association with	HGMD #CM035706	syn
chr4	rs8192678	T=0.298	23424760	PPARGC1A	0	2	0	10	0	Diabetes, type 2, association with	HGMD #CM013959	mis
chr4	rs1870377	A=0.241	55667731	KDR	8	0	0	9	0	Coronary heart disease, association with	HGMD #CM074306	mis
chr4	rs35771241	T=0.065	56197061	NMU	0	0	4	3	0	Obesity, association with	HGMD #CM066152	mis
chr4	rs1054627	A=0.168	88951716	IBSP	5	0	15	0	0	Hip bone mineral density, assoc. with	HGMD #CM093421	mis
chr4	rs1047214	C=0.275	156355126	NPY2R	0	20	0	8	0	Severe obesity, in men, association with	HGMD #CM057405	syn
chr4	rs2880415	C=0.309	156355477	NPY2R	0	8	0	1	0	Severe obesity, in men, association with	HGMD #CM057404	syn
chr5	rs16891982	G=0.454	33987450	SLC45A2	0	0	17	0	0	Pigmentation variation, assoc. with (G allele light skin colour)	HGMD #CM051556	mis
chr6	rs1799945	G=0.077	26199158	HFE	0	8	8	0	0	Haemochromatosis, association with	HGMD #CM960827	mis

chr6	rs1046089	A=0.395	31710946	BAT2	11	0	2	0	0	Malaria, severe, association with	HGMD #CM090830	mis
chr6	rs1049353	T=0.146	88910354	CNR1	0	13	0	11	0	Abdominal adiposity in obese men, association with ?	HGMD #CM074755	syn
chr6	rs6929137	A=0.347	151978370	C6orf97	16	0	11	0	0	Bone mineral density, association with ?	HGMD #CM093417	mis
chr7	rs1726866	A=0.428	141319174	TAS2R38	7	0	2	0	0	Phenylthiocarbamide taste sensitivity, association	HGMD #CM031369	mis
chr7	rs713598	C=0.499	141319814	TAS2R38	0	11	6	0	0	Phenylthiocarbamide taste sensitivity, association	HGMD #CM031368	mis
chr8	rs11549147	G=0.029	11703746	FDFT1	27	0	8	1	0	Increased total cholesterol, association with	HGMD #CM081612	mis
chr9	rs10757274	G=0.396	22086055	CDKNBAS	1	0	8	0	0	Ischaemic stroke, sudden cardiac death, association with	[23-25]	NA
chr9	rs2383206	G=0.459	22105026	CDKNBAS	0	0	8	0	0	Coronary heart disease, association with	[26]	NA
chr9	rs8176719	no data	135122729	ABO	(Coverage 7 reads)					Blood group O, association with	[32,51]	del
chr9	rs505922	C=0.364	135139050	ABO	0	0	0	5	0	Blood group O, association with	[32]	mis
chr9	rs8176746	T=0.123	135121143	ABO	0	0	9	0	0	Blood group A, association with	[51]	mis
chr10	rs4536103	G=0.442	71002210	NEUROG3	3	0	8	0	0	Diabetes, type 2, association with ?	HGMD #CM068025	mis
chr11	rs3829241	A=0.230	68611939	TPCN2	6	0	42	0	0	Hair colour, association with (brown rather than blond)	HGMD #CM083200	mis
chr13	rs5351	T=0.436	77373314	EDNRB	0	14	0	20	1	Atherosclerosis, association with	[29]	syn
chr15	rs4778138	G=0.470	26009415	OCA2	6	0	1	0	0	Non-blue eye colour, association with	[29,30,52]	int
chr15	rs4778241	A=0.489	26012308	OCA2	9	2	0	0	0	Non-blue eye colour, association with	[29,30,52]	int
chr15	rs7495174	G=0.266	26017833	OCA2	10	0	3	0	0	Non-blue eye colour, association with (A/G and G/G)	[29,30,52]	int
chr15	rs1129038	T=0.297	26030454	OCA2	0	6	0	0	0	Non-blue eye colour, association with	[29,53]	int
chr15	rs12913832	G=0.297	26039213	HERC2	31	0	0	0	0	Non-blue eye colour, association with	[29,53]	int
chr15	rs916977	C=0.423	26186959	HERC2	0	10	0	12	2	Non-blue eye colour, association with (C/T and T/T)	[29,30]	-
chr15	rs1801449	A=0.241	40468491	CAPN3	5	0	3	0	0	Muscular dystrophy	HGMD #CM099258	mis
chr16	rs5716	C=0.083	20703888	ACSM3	0	6	8	0	0	Hypertension, association with	HGMD #CM070002	mis
chr17	rs16940674	T=0.114	41266288	CRHR1	0	10	0	6	0	Increased body mass index, association with	HGMD #CM044865	syn

chr17	rs197922	A=0.333	42363569	GOSR2	9	0	1	0	0	Hypertension, association with	HGMD #CM091892	mis
chr18	rs1058396	A=0.437	41573517	SLC14A1	7	0	1	0	0	Kidd blood group variant	HGMD #CM973381	mis
chr19	rs1047286	A=0.083	6664262	C3	4	0	1	0	0	Age-related macular degeneration, association with	HGMD #CM099911	mis
chr19	rs3745297	C=0.318	54350021	HRC	9	6	0	0	0	Ventricular arrhythmias, assoc. with ?	HGMD #CM087519	mis
chr19	rs2070745	G=0.347	56941759	FPR1	0	8	11	0	0	Periodontitis, aggressive, association with	HGMD #CM074876	mis
chr19	rs1613662	G=0.141	60228407	GP6	4	0	3	0	0	Myocardial infarction, age related, association	HGMD #CM013737	mis
chr20	rs7121	C=0.363	56912202	GNAS	0	4	0	19	0	Essential hypertension, association with	HGMD #CM023931	syn

Supplementary Table S7: Y-chromosomal Haplogroup G marker status in the Iceman

Haplogroup	Marker	dbSNP#	Pos. HG18	Ancestral	Derived	Iceman	Iceman derived allele?	Reference
G	M201	rs2032636	13536923	G	T	7T	yes	Underhill 2001[48]
G	P257/U6	rs2740980	12942936	G	A	5A	yes	Karafet 2008[42]
G	U33	rs1125978	rs1125978	C	G	16G	yes	Sims 2009[15]
G	U17	rs34742138	2906401	C	T	2T	yes	Sims 2009[15]
G1	M342	n.n.	21653330	C	T	7C	no	Cinnioglu 2004[49]
G2	P287	rs4116820	20531485	G	T	6T,1G	yes	Karafet 2008[42]
G2a	P15	n.n.	21653414	C	T	2T, 1A	yes	Hammer 2000[50]
G2a	U5	rs2178500	rs2178500	T	G	4G	yes	Sims 2009[15]
G2a1	P16	n.n.	19434578	A	T	10A	no	Sims 2009[15]
G2a1a	P18	n.n.	25751219	C	T	4C	no	Sims 2009[15]
G2a2	M286	rs13447379	21151187	G	A	20G	no	Cinnioglu 2004[49]
G2a3	U8	rs7892988	13202247	T	C	7T	no	Sims 2009[15]
G2a4	L91	n.n.	20104943	G	C	5C	yes	ISOGG 2011[14]

Supplementary Table S8: Genes with affected SNPs

A: Start-/Stop-Codons

Chr.	Position	dbSNP#	Gene	Function	Effect
chr1	1876879	rs28548017	KIAA1751	StopCodon	rdt
chr1	12698931	rs3000859	AADACL3	StartCodon	scl
chr1	114925733		BCAS2	StartCodon	scl
chr1	148799295		ADAMTSL4	StopCodon	syn
chr1	220788009		HHIPL2	StartCodon	scl
chr1	224478006		MIXL1	StartCodon	scl
chr1	228895760		COG2	StopCodon	rdt
chr1	245486132	rs1778540	VN1R5	StopCodon	rdt
chr2	33641307		RASGRP3	StopCodon	rdt
chr2	171889017	rs10205459	METTL8	StopCodon	rdt
chr2	179153014	rs4145333	TTN	StopCodon	rdt
chr2	220170884	rs681747	STK11IP	StartCodon* (R1M)	mis
chr2	231918049		ARMC9	StopCodon	syn
chr3	48700711	rs1048940	IP6K2	StopCodon	syn
chr3	50130891		RBM5	StopCodon	rdt
chr3	51397806	rs13091931	ARMET	StartCodon* (R1M)	mis
chr3	51993123		ACY1	StartCodon	scl
chr3	99209437	rs832032	GABRR3	StopCodon	rdt
chr3	197023084		MUC4	StartCodon	scl
chr4	100487213	rs283413	ADH1C	StopCodon	rdt
chr4	152432053	rs2407221	ESSPL	StopCodon	rdt
chr5	37100751		NIPBL	StopCodon	rdt
chr5	180210999	rs705441	ZFP62	StartCodon	scl
chr6	57575059	rs4535533	PRIM2	StopCodon	rdt
chr6	132901302	rs2842899	TAAR9	StopCodon	rdt
chr7	75459763	rs4732519	TMEM120A	StopCodon	rdt
chr8	57391458		SDR16C5	StartCodon	scl
chr11	55870092	rs1905055	OR8K1	StartCodon	scl
chr11	57739770	rs7103033	OR1S1	StopCodon	rdt
chr11	58141755		ZFP91	StopCodon	syn
chr12	46559162	rs2228570	VDR	StartCodon	scl
chr12	111932671	rs15895	OAS2	StopCodon	rdt
chr12	113277720		TBX5	StopCodon	syn
chr14	95800832		BDKRB1	StopCodon	rdt
chr15	29156415	rs4779816	TRPM1	StartCodon	scl
chr15	39459676	rs7168431	NUSAP1	StopCodon	syn
chr15	97471323	rs2305445	SYNM	StopCodon	rdt
chr16	53921718		IRX6	StopCodon	syn
chr16	79703308	rs8054182	PKD1L2	StopCodon	rdt
chr17	1371645		PITPNA	StopCodon	rdt
chr17	38956740		DHX8	StopCodon	rdt
chr18	12244816	rs7505615	CIDEA	StartCodon	scl
chr19	10259860		ICAM4	StopCodon	rdt
chr19	50014827		BCAM	StopCodon	rdt
chr19	61191091	rs306457	NLRP8	StopCodon	rdt
chr22	36417288		NOL12	StopCodon	rdt
chrX	12834747	rs3764880	TLR8	StartCodon	scl
chrX	54852492		MAGED2	StartCodon	scl

* wildcard Start-Aminoacid is an Arginin, substituted AA is a Methionin

syn synonymous
mis missense
rdt readthrough
scl start canceled (start lost)

B: readthrough

chr.	Position	dbSNP#	Gene
chr1	1876879	rs28548017	KIAA1751
chr1	228895760		COG2
chr1	245486132	rs1778540	VN1R5
chr2	33641307		RASGRP3
chr2	171889017	rs10205459	METTL8
chr2	179153014	rs4145333	TTN
chr3	50130891		RBM5
chr3	99209437	rs832032	GABRR3
chr4	100487213	rs283413	ADH1C
chr4	152432053	rs2407221	ESSPL
chr5	37100751		NIPBL
chr6	57575059	rs4535533	PRIM2
chr6	132901302	rs2842899	TAAR9
chr7	75459763	rs4732519	TMEM120A
chr11	57739770	rs7103033	OR1S1
chr12	111932671	rs15895	OAS2
chr14	95800832		BDKRB1
chr15	97471323	rs2305445	SYNM
chr16	79703308	rs8054182	PKD1L2
chr17	1371645		PITPNA
chr17	38956740		DHX8
chr19	10259860		ICAM4
chr19	50014827		BCAM
chr19	61191091	rs306457	NLRP8
chr22	36417288		NOL12

Supplementary Table S9: Number of hits and coverage of the genomic reads to selected *Borrelia* genome and plasmid sequences

Reference name	Reference acc #	reads	basecalls	ref length	reads / bp ref	average depth cov	1x coverage	5x coverage	10x coverage	20x coverage
Borrelia burgdorferi ZS7, complete genome	NC_011728.1	905460	30407293	906707	1	33,54	57,50%	9,90%	3,90%	1,90%
Borrelia burgdorferi B31 plasmid lp17	NC_001849.1	1675	43142	16823	0,1	2,56	67,10%	13,00%	4,80%	1,60%
Borrelia burgdorferi B31 plasmid lp25	NC_001850.1	2432	62609	24177	0,1	2,59	63,80%	12,30%	4,10%	0,90%
Borrelia burgdorferi B31 plasmid cp9	NC_001904.1	869	22313	9386	0,09	2,38	62,50%	11,50%	4,70%	1,40%
Borrelia burgdorferi plasmid cp26	NC_001903.1	2230	57314	26498	0,08	2,16	57,10%	11,20%	3,80%	1,30%
Borrelia burgdorferi B31 plasmid lp38	NC_001856.1	4129	107207	38829	0,11	2,76	56,50%	9,90%	3,60%	1,60%
Borrelia burgdorferi B31 plasmid lp28-4	NC_001854.1	2515	64478	27323	0,09	2,36	58,10%	9,70%	3,70%	0,80%
Borrelia burgdorferi B31 plasmid lp28-3	NC_001853.1	2080	53665	28601	0,07	1,88	55,50%	9,30%	2,70%	1,20%
Borrelia burgdorferi B31 plasmid lp54	NC_001857.1	4164	109170	53561	0,08	2,04	57,30%	8,70%	2,70%	1,10%
Borrelia burgdorferi B31 plasmid lp28-1	NC_001851.1	1888	48829	26921	0,07	1,81	46,20%	7,80%	2,70%	0,90%
Borrelia burgdorferi B31 plasmid lp28-2	NC_001852.1	1661	42708	29766	0,06	1,43	50,20%	7,20%	2,30%	0,60%
Borrelia burgdorferi B31 plasmid cp32-1	NC_000948.1	1789	45844	30750	0,06	1,49	50,20%	6,30%	2,20%	0,90%
Borrelia burgdorferi B31 plasmid lp36	NC_001855.1	2097	53816	36849	0,06	1,46	47,60%	6,00%	2,30%	0,80%
Borrelia burgdorferi B31 plasmid lp56	NC_000956.1	2630	67425	52971	0,05	1,27	30,30%	5,10%	2,00%	0,80%
Borrelia burgdorferi B31 plasmid lp21	NC_000955.1	614	15830	18753	0,03	0,84	24,90%	4,60%	1,10%	0,30%
Borrelia burgdorferi ZS7 plasmid ZS7_cp32-3+10	NC_011720.1	1792	46714	48168	0,04	0,97	30,90%	3,70%	1,40%	0,30%
Borrelia burgdorferi B31 plasmid cp32-7	NC_000952.1	1174	31630	30800	0,04	1,03	23,50%	3,60%	1,50%	0,50%
Borrelia burgdorferi ZS7 plasmid ZS7_lp28-1	NC_011780.1	672	17266	23422	0,03	0,74	26,00%	3,60%	1,10%	0,10%
Borrelia burgdorferi B31 plasmid cp32-3	NC_000949.1	929	24949	30223	0,03	0,83	23,70%	3,10%	0,90%	0,40%
Borrelia burgdorferi B31 plasmid cp32-4	NC_000950.1	980	25770	30299	0,03	0,85	30,50%	2,90%	0,80%	0,50%
Borrelia burgdorferi ZS7 plasmid ZS7_lp28-4	NC_011785.1	561	14426	28885	0,02	0,5	16,90%	2,80%	0,80%	0,20%
Borrelia burgdorferi B31 plasmid cp32-9	NC_000954.1	742	19181	30651	0,02	0,63	19,90%	2,60%	0,90%	0,40%
Borrelia burgdorferi B31 plasmid cp32-6	NC_000951.1	668	17103	29838	0,02	0,57	20,20%	2,50%	0,80%	0,40%
Borrelia burgdorferi B31 plasmid lp5	NC_000957.1	114	2929	5228	0,02	0,56	26,70%	2,10%	0,10%	0,00%
Borrelia burgdorferi ZS7 plasmid ZS7_cp32-4	NC_011736.1	691	18006	30964	0,02	0,58	12,00%	1,70%	1,10%	0,40%
Borrelia burgdorferi ZS7 plasmid ZS7_cp26	NC_011724.1	365	9405	26514	0,01	0,35	12,60%	1,60%	0,60%	0,20%
Borrelia burgdorferi ZS7 plasmid ZS7_cp32-12	NC_011735.1	844	21786	29806	0,03	0,73	13,80%	1,50%	0,70%	0,40%
Borrelia burgdorferi ZS7 plasmid ZS7_lp17	NC_011782.1	253	6536	17266	0,01	0,38	14,30%	1,50%	0,70%	0,10%
Borrelia burgdorferi B31 plasmid cp32-8	NC_000953.1	331	8522	30885	0,01	0,28	8,60%	1,30%	0,30%	0,20%
Borrelia burgdorferi ZS7 plasmid ZS7_cp32-1	NC_011731.1	378	9656	30330	0,01	0,32	12,90%	1,20%	0,20%	0,20%
Borrelia burgdorferi ZS7 plasmid ZS7_lp25	NC_011783.1	242	6163	24326	0,01	0,25	10,60%	1,10%	0,30%	0,10%
Borrelia burgdorferi ZS7 plasmid ZS7_lp54	NC_011784.1	457	11686	53615	0,01	0,22	10,70%	0,60%	0,30%	0,10%
Borrelia burgdorferi ZS7 plasmid ZS7_lp28-3	NC_011781.1	97	2477	28414	0	0,09	4,10%	0,40%	0,10%	0,10%
Borrelia burgdorferi ZS7 plasmid ZS7_lp28-2	NC_011779.1	183	4686	29758	0,01	0,16	5,20%	0,30%	0,10%	0,10%
Borrelia burgdorferi ZS7 plasmid ZS7_cp32-9	NC_011722.1	93	2375	30467	0	0,08	4,50%	0,30%	0,10%	0,00%
Borrelia burgdorferi ZS7 plasmid ZS7_lp36	NC_011778.1	145	3677	36852	0	0,1	6,50%	0,20%	0,10%	0,00%

Supplementary Methods

Sampling

The bone sample used in DNA extraction for whole genome sequencing was taken from the interior of the Iceman's left ilium with a sterilised biopsy needle (HS Trapsystem) and immediately frozen at -20°C. Bone sampling was carried out under sterile conditions at a temperature of 4°C in an antechamber of the Iceman's conservation cell. All persons present wore protective overalls, sterile gloves and facemasks to prevent exogenous contamination. To identify potential contamination, DNA samples were taken from all persons handling the mummy or the sample.

PCR analysis

Gel electrophoresis was performed using native extract to estimate DNA quantity and fragment length. Isolated DNA was in a size range between 100 bp and 600 bp. The Iceman's hypervariable region 1 of the human mitochondrial genome was screened for the presence of haplogroup K-specific SNPs. PCR was subsequently carried out as an extraction quality control, using a 1:10 dilution of the native extract. Some of the characteristic SNPs of the Iceman's mitochondrial genome (rCRS positions 16224, 16311, 16362) were screened in the Tübingen laboratories using two pairs of mitochondrial primers (Table 2). The PCR protocol was as follows for a total reaction volume of 25 µl: 1x FastStart PCR Buffer (Roche), 0.04mM dNTP mix (Roche), 0.04mM forward primer (Invitrogen), 0.04mM reverse primer (Sigma Aldrich), 1.5u FastStart polymerase (Roche), 1µl diluted Iceman DNA extract, and UV-irradiated PCR-grade HPLC water (Roth) in a total volume of 25 µl. Cycling conditions in a GeneAmp PCR System 2700 (Applied Biosystems) were as follows: 5 min. 94°C; 45 cycles: 30s 94°C, 30s 55°C, 30s 72°C; 5 min. 72°C final elongation. After visualization on

agarose gels the obtained PCR products were purified with the ExoSAP-IT® PCR clean-up kit (Affymetrix) according to the manufacturer's protocol. Cycle-sequencing reaction setup was done with 2 µl BigDye® Terminator v1.1 buffer, 1x BigDye® Terminator v1.1 dye mix (Applied Biosystems), 1.6pmol sequencing primer, 1-3 µl of the PCR product, PCR-grade HPLC water (Roth) in a total volume of 10 µl reaction mixture. Cycling conditions consisted of 25 cycles of 30s 94°C, 30s 55°C, 4 min. 60°C. Cycle-sequencing products were purified by means of Centri-Sep® columns (Princeton Separations) according to the manufacturer's protocol. Capillary electrophoresis was performed on an ABI PRISM®310 genetic analyzer (Applied Biosystems). DNA sequences were subsequently compared to the Iceman's genomic consensus sequence using the Lasergene SeqMan Pro software version 8.0.2 (DNASTAR®). The Iceman-derived sequences were also compared to the mitochondrial genome sequences of the laboratory staff for authentication purposes. None of the lab members were tested positive for mitochondrial haplotype K. The remaining DNA extract was re-frozen until NGS library preparation.

Sequence Alignment

We used a custom workflow in the SeqWare Pipeline tool to align the Iceman genome SOLiD reads using BFAST version 0.6.4e with the indexing strategy and parametrization recommended by the tool's manual.

The resulting BAM alignment files were further processed using the SRMA tool version 0.1.15, which performed realignments around indels to ensure the highest accuracy in calling this variant type. The resulting re-aligned BAM files were further processed by PicardTools version 1.48 to annotate read group and other anotation information as well as fixing mate pair information. These refined BAM files were then used by variant calling with SamTools version 0.1.17 using the mpileup program and a -C parameter of 50 to scale mapping quality scores. Variants were filtered for downstream analysis by requiring a minimum coverage of 3

and a maximum coverage of 30. The parametrizations used by this pipeline have been shown to have a very high variant true positive call rate for SNVs (95.5-98.6%) and low false positive rate (<0.1-3.1%) based on data simulations. Once called, variants were further annotated using the Annovar program downloaded on 20110506 and annotation tracks provided by that project. These tracks included dbSNP 132, the variants called using Life Technologies LifeScope v1.0, OMIM genes, variants from the 1000 genomes project, and SIFT scores. The resulting variants and annotations where then loaded into a database for realtime analysis using the SeqWare Query Engine project. This tool allowed for variants to be filtered by annotations and directly visualized or further annotated by other downstream tools.

The SeqWare project (http://seqware.sf.net), specifically the SeqWare Pipeline sub-project, was used to execute the complex analytical pipeline entirely on the Amazon EC2 cloud environment using software contributions by Nimbus Informatics. Both the Iceman genome website (http://icemangenome.net) and the Query Engine for searching variants are fully hosted on the Amazon cloud.

Subsequent Data Verification

For data verification purposes, DNA was extracted from new sample material (tissue from the femoral muscle) at the ancient DNA facilities at the EURAC Institute for Mummies and the Iceman, Bolzano, Italy. The samples were frozen at -20°C until extraction. Pre- and post-PCR areas are spatially separate, and strict anti-contaminant measures were adhered to. All surfaces and appliances were treated with 3% sodium hypochlorite bleach or DNA awaysolution (Roth) prior to and after use. Tubes were irradiated under UV light. As the sample was taken from the Iceman under sterile conditions, no further surface irradiation was carried out so as to avoid possible aerial contamination. The samples were manually reduced

with a sterile scalpel and an irradiated and bleach-treated mortar and pestle to avoid heat-induced damage[54]. Sample digestion and subsequent DNA extraction was carried out using the phenol-chloroform method described by Burger *et al*[55], using Amicon Ultra-15 Centrifugal Filter Units (Millipore) for concentration. For washing steps, PCR grade, UV-irradiated HPLC water (tested for contamination prior to use) was used. Blank controls were set up during extraction to monitor potential contamination.

Mitochondrial sequences were amplified by PCR to verify the authenticity of the extracted DNA, using both standard HVS1 primers and primers specifically developed to target specific mutations in the Iceman's haplotype, as reported in previous studies[3,4] (Table 2). PCR primers were designed for most of the SNPs using the DNAStar LaserGene® software, taking into account the short fragment size of the residual nuclear DNA (Table 1). Primer specificity was examined using NCBI PrimerSelect. Laboratory staff, handling the sample and the water used to rehydrate the mummy at specific intervals, were genotyped for the Iceman's specific mtDNA SNPs (rCRS positions 8137, 16224, 16311, 16362) and the non-clinical, Y-chromosomal SNP rs2032636.

PCR setup protocol was as follows for a total reaction volume of 50 µl: 5µl $MgCl_2$ (Applied Biosystems), 5µl GeneAmp 10x PCR Gold Buffer (Applied Biosystems), 1µl 10mM dNTP mix (Roche or Applied Biosystems), 1µl purified BSA 100x (New England BioLabs), 1µl forward primer (20mM, Invitrogen), 1µl reverse primer (20mM, Invitrogen), 0.3µl AmpliTaq Gold® DNA polymerase (Applied Biosystems), 4µl template DNA extract solution, rest UV-irradiated, PCR grade HPLC water to a total volume of 50 µl. Cycling conditions were as follows: initial denaturation 6 minutes at 94°C, 50 cycle steps (40 for mitochondrial loci): denaturation phase 40 seconds/94°C, annealing phase 40 seconds/given primer temperature (55°C unless otherwise stated, see Table 1), elongation phase 40 seconds/72°C. After visualisation on agarose gel, PCR products were purified with the Promega Wizard® PCR

clean-up kit, according to manufacturer's protocol. Cycle sequencing reaction setup was as follows: 1.5 µl BigDye® Terminator v1.1 buffer, 1 µl BigDye® Terminator v1.1 dye mix, 1 µl primer, 1-3 µl PCR product (depending on band strength), rest PCR grade HPLC water to a total volume of 10 µl. Cycling conditions were denaturation phase of 30 seconds at 92°C, annealing phase at 15 seconds at the given primer optimum temperature (see Table 1), elongation phase at 2.5 minutes for a total of 25 cycles. Cycle sequencing products were purified by means of Centri-Sep® columns (Princeton Separations) according to manufacturer's protocol. Capillary electrophoresis was performed on the ABI PRISM® 310 genetic analyzer (Applied Biosystems), with varied instrument settings according to fragment size (data upon request). Primary sequence visualisation (inspection of correctly called bases) was carried out with the ABI PRISM sequence analysis 3.4.1 software, sequences were subsequently aligned with NCBI dbSNP reference sequences, the Iceman's genomic consensus sequence and (in the case of non-clinical SNPs) genotyped sequences from laboratory staff to handle the samples, using DNASTAR® Lasergene SeqMan Pro software version 8.0.2. Sequencing results are listed in Table 1.

Supplementary References

48 Underhill, P. A. *et al.* The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* **65**, 43-62, (2001).

49 Cinnioglu, C. *et al.* Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet* **114**, 127-148, (2004).

50 Hammer, M. F. *et al.* Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc Natl Acad Sci U S A* **97**, 6769-6774, (2000).

51 Melzer, D. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet* **4**, e1000072, (2008).

52 Kayser, M. *et al.* Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. *Am J Hum Genet* **82**, 411-423, (2008).

53 Duffy, D. L. *et al.* A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am J Hum Genet* **80**, 241-252, (2007).

54 Adler, C. J., Haak, W., Donlon, D. & Cooper, A. The Genographic Consortium, Survival and recovery of DNA from ancient teeth and bones. *Journal of Archaeological Science* **38**, 956-964, (2011).

55 Burger, J. *et al.* Molecular phylogeny of the extinct cave lion Panthera leo spelaea. *Mol Phylogenet Evol* **30**, 841-849, (2004).