

Exome Sequencing Provides Evidence of Polygenic Adaptation to a Fat-Rich Animal Diet in Indigenous Siberian Populations

PingHsun Hsieh,¹ Brian Hallmark,² Joseph Watkins,³ Tatiana M. Karafet,⁴ Ludmila P. Osipova,^{5,6} Ryan N. Gutenkunst,^{*,7} and Michael F. Hammer^{*,1,4}

¹Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ

²Interdisciplinary Program in Statistics, University of Arizona, Tucson, AZ

³Department of Mathematics, University of Arizona, Tucson, AZ

⁴ARL Division of Biotechnology, University of Arizona, Tucson, AZ

⁵Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

⁶Novosibirsk State University, Novosibirsk, Russia

⁷Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ

*Corresponding authors: E-mails: rgutenk@email.arizona.edu; mfh@email.arizona.edu.

Associate editor: Brenna Henn

Abstract

Siberia is one of the coldest environments on Earth and has great seasonal temperature variation. Long-term settlement in northern Siberia undoubtedly required biological adaptation to severe cold stress, dramatic variation in photoperiod, and limited food resources. In addition, recent archeological studies show that humans first occupied Siberia at least 45,000 years ago; yet our understanding of the demographic history of modern indigenous Siberians remains incomplete. In this study, we use whole-exome sequencing data from the Nganasans and Yakuts to infer the evolutionary history of these two indigenous Siberian populations. Recognizing the complexity of the adaptive process, we designed a model-based test to systematically search for signatures of polygenic selection. Our approach accounts for stochasticity in the demographic process and the hitchhiking effect of classic selective sweeps, as well as potential biases resulting from recombination rate and mutation rate heterogeneity. Our demographic inference shows that the Nganasans and Yakuts diverged ~12,000–13,000 years ago from East-Asian ancestors in a process involving continuous gene flow. Our polygenic selection scan identifies seven candidate gene sets with Siberian-specific signals. Three of these gene sets are related to diet, especially to fat metabolism, consistent with the hypothesis of adaptation to a fat-rich animal diet. Additional testing rejects the effect of hitchhiking and favors a model in which selection yields small allele frequency changes at multiple unlinked genes.

Key words: adaptation, diet, Siberia, extreme climates, polygenic selection.

Introduction

Recent fossil evidence uncovered from Upper Paleolithic sites within the Siberian Arctic (71°N, dated ~28,000 years ago; 72°N, dated ~45,000 years ago) indicates that Siberia was one of the earliest places in Eurasia that anatomically modern humans (AMH) successfully colonized (Pitulko et al. 2004; Gillingham et al. 2007; Pitulko et al. 2016). The AMH fossils discovered at Mal'ta (dated ~24,000 years ago) and Ust'-Ishim (dated ~45,000 years ago) in South/Western Siberia (Fu et al. 2014; Raghavan et al. 2014) suggest that humans successfully adapted to a cold environment shortly after their expansion out of Africa. While their relationships to the Upper-Paleolithic fossils are debatable (Raghavan et al. 2014; Wong et al. 2016), today, the 31 indigenous populations living across Siberia are distinguished by language, culture, lifestyle, and means of subsistence (Karafet et al. 2002; Uinuk-Ool et al. 2003) with many still practicing a foraging subsistence strategy. The extremely cold climate (mean

January temperature from −32 to −41°C) (Kathiresan et al. 2008) and dramatic seasonal variation in photoperiod cause unpredictable and limited food resources (Uinuk-Ool et al. 2003). These environmental (polar eco-region) and ecological (subsistence/diet) constraints have likely imposed strong selective pressures (Hancock et al. 2010).

A recent study of Northeastern Siberians found evidence for a classic selective sweep associated with fat metabolism, suggesting an adaptation to diet (Clemente et al. 2014). In addition, genome-scan studies in Siberians have reported candidate loci for classic selective sweeps that are associated with genes involved in energy metabolism, vascular smooth muscle contraction, and DNA damage response against UV radiation (Cardona et al. 2014). However, physiological traits such as metabolism are likely regulated by complex biological pathways involving tens or hundreds of genes (Pritchard et al. 2010; Devlin 2015). Thus, it is important not only to assess the effects of strong selection acting on single genes, but also

to investigate models that involve simultaneous selection promote allele frequency changes at multiple loci (Pritchard and Di Rienzo 2010; Pritchard et al. 2010). In addition, in order to identify true signals of selection, one must control for possible biases resulting from non-selective forces, including stochasticity of the demographic process as well as recombination and mutation rate heterogeneity (Hsieh et al. 2016).

Genetic studies of Siberian populations have focused on data from uniparentally-inherited mtDNA (Fedorova et al. 2013; Derenko et al. 2014), the Y chromosome (Karafet et al. 2002; Fedorova et al. 2013), as well as on genome-wide single nucleotide polymorphisms (Fedorova et al. 2013; Cardona et al. 2014; Pugach et al. 2016; Wong et al. 2016). These studies have supported the hypothesis of a Central Asian/South Siberian origin of modern Siberians (Karafet et al. 2002; Fedorova et al. 2013; Derenko et al. 2014), and have revealed geographic clines of genetic admixture across indigenous Siberians with respect to East-Asian and European ancestries (Fedorova et al. 2013; Cardona et al. 2014; Pugach et al. 2016). However, despite these studies describing relationships among Siberian populations and their neighbors, a model-based framework that allows for hypothesis testing and that jointly incorporates effective population size, population divergence/isolation, and gene flow is lacking.

In this study, we developed a hypothesis-testing approach for inferring both the demographic and adaptive histories of indigenous Siberians. We generated whole-exome data from two indigenous Central Siberian populations, the Nganasans (nomadic hunters) from the Taymyr Peninsula in the Arctic Ocean, and the Yakuts (herders) of North-Central Siberia. We chose to study these two populations as representatives for modern-day Siberians because (1) the Nganasans, the northernmost Eurasian population, lived in isolation until the mid-1970s (Goltsova et al. 2005), and are considered to be the direct descendants of the Paleolithic hunters who inhabited Taymyr Peninsula (Simchenko 1976); and (2) the Yakuts, one of the largest populations (>450,000 individuals) in North-Central Siberia who live in extreme subarctic climate. To construct demographic models for these two Siberian populations that incorporate divergence, isolation, and gene flow, we also included exome sequencing data from East-Asian and European populations in the 1000 Genome Project (1000 Genomes Project Consortium 2012). We then searched for signals of polygenic selection using a Mann–Whitney U (MWU)-based test to identify gene sets especially enriched with highly differentiated variants compared with the rest of data. We further validated each identified gene set using a parametric test based on coalescent simulations to provide additional controls for demographic history, gene clustering, and mutation/recombination rate heterogeneity in the genome. Our study design used biologically meaningful gene sets to detect polygenic selection signals arising from not only small allele frequency shifts at many loci, but also combinations of hard/soft sweeps at multiple loci (Chevin and Hospital 2008; Pritchard and Di Rienzo 2010). This study design requires no a priori information about selective traits and is fundamentally different from other studies (e.g., Berg and

Coop 2014) that utilizes phenotypic information presumably relevant to adaptation for the population of interest.

Results

Demographic Relationships Among Central Siberians, East Asians, and Europeans

To infer demographic histories based on exome sequence data derived from two indigenous Siberian populations, the Nganasans (NGA, $N = 21$) and Yakuts (YAK, $N = 21$) (see Materials and Methods), as well as exome data from Europeans (CEU, $N = 97$) and East Asians (CHB, $N = 95$) (1000 Genomes Project Consortium 2012), we used the diffusion-based demographic inference tool *∂a∂i* (Gutenkunst et al. 2009). Population joint site frequency spectra (SFSs) were inferred directly from exome BAM files using ANGSD (Korneliusson et al. 2014), which equips a probabilistic model to accurately infer allele frequencies of variants for low-coverage sequencing data without explicitly calling genotypes. We began our modeling strategy with pairwise two-population (2D) models and then used these findings to choose a formal hypothesis-testing framework for three-population (3D) models. Supplementary tables S1–S3, Supplementary Material online, list the 2D models that we considered for each population pair and their maximum likelihood estimates for model parameters. The best-fit model for the Siberian population pair NGA-YAK (supplementary table S1, Model S1_1, Supplementary Material online) has these two populations isolated from each other since their divergence. For both the pairs NGA-CHB and YAK-CHB, the best-fit models (supplementary table S2, Model S2_4 and table S3, Model S3_4, Supplementary Material online) have the ancestors of the pairs experiencing population size bottlenecks, followed by exponential recoveries. There is also evidence of symmetric gene flow between the two populations in each of the two pairs after their divergence (supplementary table S2, Model S2_4 and table S3, Model S3_4, Supplementary Material online). We found that the split time of the NGA-YAK pair is more recent than between either of these populations and CHB. We then attempted to identify the most likely outgroup population by inferring the best-fit model topologies for each of the three population trios: CHB-YAK-NGA, CEU-CHB-NGA, and CEU-CHB-YAK (Supplementary Materials online). We found that CHB is an outgroup to the two Siberian populations, and CEU is an outgroup to the other three populations. (supplementary table S4, Model S4_1–3, table S5, Model S5_1–3, and table S6, Model S6_1–3, supplementary Material online). We did not test CEU-YAK-NGA because the inference results from the aforementioned three population trios are sufficient to determine the tree topology of the four populations.

Complex Three-Population Demographic Models

It was infeasible to estimate a joint four-population SFS using ANGSD, due to the large dimensionality of the resulting matrix ($195 \times 191 \times 43 \times 43 = 68,866,005$ entries). In order to gain additional insight into the demographic prehistory of the four populations, we instead opted to build complex

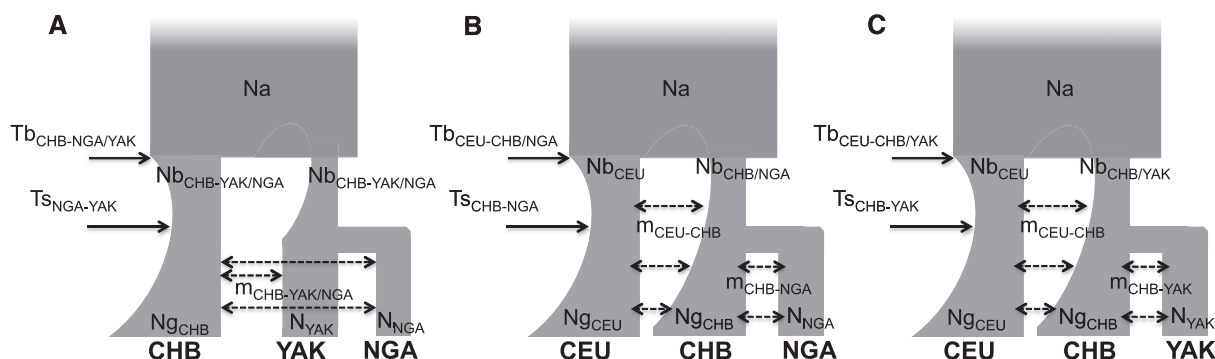


Fig. 1. Best-fit three-population demographic models for Nganasans (NGA), Yakuts (YAK), East-Asians (CHB), and Europeans (CEU). (A) CHB–YAK–NGA, (B) CEU–CHB–NGA, (C) CEU–CHB–YAK. Maximum likelihood estimates for model parameters and corresponding site-frequency spectra for each model can be found in [table 1](#) and supplementary figs. S1–S3, Supplementary Material online, respectively.

three-population models using the SFSs for the three population trios: CHB–YAK–NGA ([fig. 1A](#), [table 1](#), and supplementary table S4, Supplementary Material online), CEU–CHB–NGA ([fig. 1B](#), [table 1](#), and supplementary table S5, Supplementary Material online), and CEU–CHB–YAK ([fig. 1C](#), [table 1](#), and supplementary table S6, Supplementary Material online). Because the likelihood calculated by $\partial a \partial i$ is composite, we determined the best-fit models using a likelihood ratio test (LRT) with statistical correction, which asymptotically adjusts the LRT statistic in order to conduct the conventional chi-squared test (see Materials and Methods; also see [Coffman et al. 2016](#)). [Figure 1](#) and [table 1](#) show the best-fit models among those we tested and their maximum likelihood parameter estimates, respectively, for the three population trios (supplementary tables S4–S6, Supplementary Material online). Parameter estimates are calculated using the mutation rate of 2.35×10^{-8} ([Gutenkunst et al. 2009](#), compatible with [Nachman and Crowell 2000](#)) and a generation time of 25 years (see supplementary table S7, Supplementary Material online, for values calculated using the mutation rate of 1.5×10^{-8} per base per generation and a generation time of 29 years). In the case of CHB–YAK–NGA, the best-fit model (nine parameters, log-likelihood = $-221,308$, Model-A in [fig. 1](#) and [table 1](#)) indicates that the ancestors of the CHB and the two Siberian populations diverged $\sim 14,000$ years ago (kya) (95% CI 11–15 kya), and that both experienced the same $\sim 87\%$ population bottleneck in size (95% CI 86–88%) followed by an exponential recovery in size with a rate of 0.8% per generation (95% CI 0.74–0.94%). Our inference also shows evidence of continuous symmetric gene flow after the divergence of the YAK and the NGA at ~ 9.6 kya (95% CI 7.7–11 kya) with the same magnitude between the CHB and each of the two Siberian groups (adjusted Chi-square statistic = 3.784; adjusted LRT P -value = 0.026 when compared with a model with no gene flow: Supplementary table S5, Model S5_4; Methods, Supplementary Material online).

As for the cases of CEU–CHB–NGA (Model-B in [fig. 1](#)) and CEU–CHB–YAK (Model-C in [fig. 1](#)), the best-fit models have the same tree topology and statistically compatible parameter estimates ([table 1](#)). The divergence of the CEU and the

ancestors of the CHB and Siberians occurred at ~ 14 – 15 kya (95% CI in Model-B: 14–17 kya; 95% CI in Model-C: 12–15 kya), followed by population bottlenecks with size reductions of ~ 82 – 84% in the CEU (95% CI in Model-B: 79.8–83.9%; 95% CI in Model-C: 81.6–86.3%) and of ~ 92.6 – 93.2% in the ancestors of the CHB and Siberians (95% CI in Model-B: 91.8–93.4%; 95% CI in Model-C: 92.7–93.7%). Following divergence and bottleneck events, the CEU effective population size increases exponentially with per-generation rates of 0.38% (Model-B, 95% CI 0.35–0.43%) and 0.46% (Model C, 95% CI 0.41–0.55%). Similarly, for the CHB and the Siberian populations, effective population size increases exponentially with per-generation rates of 0.68% (Model-B, 95% CI 0.63–0.75%) and 0.82% (Model-C, 95% CI 0.74–0.91%). In addition, consistent with our results for CHB–YAK–NGA, we also found that the CHB diverged from the ancestors of the two Siberian groups ~ 12 kya (Model-B, 95% CI 11–14 kya) and 13 kya (Model-C, 95% CI 11–12 kya). Similar to previous studies ([Gutenkunst et al. 2009](#); [Gravel et al. 2011](#)), we found evidence of continuous symmetric gene flow with similar magnitude between the CEU and the CHB since their divergence ([fig. 1](#) and [table 1](#)). Furthermore, we also found that the best-fit models, which incorporate continuous symmetric gene flow between the CHB and either of the two Siberian groups, are statistically better than models without gene flow (adjusted Chi-square statistics: 3.993 and 18.558 and adjusted LRT P -values 0.023 and $< 2.2 \times 10^{-16}$ for models with the NGA and YAK, respectively; supplementary tables S5 and S6, Supplementary Material online).

Our estimated divergence times between Europeans and East Asians are more recent than previously reported times based on Out-of-Africa (OOA) models ([Gutenkunst et al. 2009](#); [Gravel et al. 2011](#)). To test whether this difference arose from our use of low-coverage Siberian data, we used only the high-coverage CEU and CHB exome data from the 1000 Genomes Project to infer the portion of the OOA model that includes Europeans and East Asians. The resulting divergence time estimates was compatible with our best-fit models (supplementary table S7, Supplementary Material online), suggesting that our low-coverage Siberian exome data did not strongly bias our inferences.

Table 1. Maximum Likelihood Parameter Estimates and Confidence Intervals for the Best-Fit Demographic Models for Each of the Population Trios: CHB–YAK–NGA, CEU–CHB–NGA, and CEU–CHB–YAK (fig. 1).

Parameters	CHB–YAK–NGA		CEU–CHB–NGA		CEU–CHB–YAK	
	Model-A		Model-B		Model-C	
	Estimates	95% CI	Estimates	95% CI	Estimates	95% CI
N_a	7,396	6,543–8,360	8,562	7,613–9,510	8,615	7,769–9,461
$Nb_{CHB-YAK/NGA}$	939	894–983	–	–	–	–
Nb_{CEU}	–	–	1,558	1,378–1,729	1,387	1,180–1,585
$Nb_{CHB/NGA}$	–	–	633	565–702	–	–
$Nb_{CHB/YAK}$	–	–	–	–	585	542–628
Ng_{CEU}	–	–	17,004	15,505–18,493	19,323	18,177–20,469
Ng_{CHB}	82,546	64,485–105,674	45,507	37,724–53,298	55,161	50,664–59,658
N_{YAK}	16,707	2,285–121,967	–	–	3,316	1,679–4,953
N_{NGA}	2,840	2,499–3,232	1,926	1,840–2,003	–	–
$T_{CHB-YAK/NGA}$	14,052	11,463–15,901	–	–	–	–
$T_{CEU-CHB/NGA}$	–	–	15,838	14,127–17,124	–	–
$T_{CEU-CHB/YAK}$	–	–	–	–	14,214	12,491–15,507
$T_{YAK-NGA}$	9,614	7,765–11,094	–	–	–	–
$T_{CHB-NGA}$	–	–	13,271	11,986–14,555	–	–
$T_{CHB-YAK}$	–	–	–	–	12,061	11,199–12,922
$m_{CHB-YAK/NGA}$	3.02×10^{-5}	2.99×10^{-5} – 3.05×10^{-5}	–	–	–	–
$m_{CEU-CHB}$	–	–	7.6×10^{-5}	6.94×10^{-5} – 8.35×10^{-5}	6.5×10^{-5}	5.89×10^{-5} – 7.12×10^{-5}
$m_{CHB-NGA}$	–	–	2.2×10^{-5}	1.87×10^{-5} – 2.59×10^{-5}	–	–
$m_{CHB-YAK}$	–	–	–	–	2.5×10^{-5}	2.13×10^{-5} – 2.85×10^{-5}
P_{flip}	0.025	0.024–0.027	0.019	0.018–0.020	0.019	0.017–0.020

NOTE.—Parameter estimates are calculated using the mutation rate of 2.35×10^{-8} per base per generation (Gutenkunst et al. 2009) and a generation time of 25 years. N is effective population size, where N_a is the ancestral effective population size; T is the time of a demographic event (years), m is migration rate (per chromosome per generation). The parameter P_{flip} models the proportion of variants with ancestral state misidentification. 95% confidence intervals were estimated using Godambe Information matrix (see Materials and Methods).

Candidate Gene Sets for Polygenic Selection in Indigenous Siberian Populations

To search for Siberian-specific signals of polygenic selection, we applied an MWU test using the population branch statistic (PBS; Yi et al. 2010) to each of the two Siberian population samples: NGA (focal population) versus CHB/CEU and YAK (focal population) versus CHB/CEU (supplementary fig. S5, Supplementary Material online). In short, our approach for detecting polygenic selection aims to determine whether the PBS distribution of single nucleotide variants (SNVs) in a set of genes is significantly shifted toward larger values than that of the remaining data (see Materials and Methods). Among the 1,206 curated human-specific gene sets from the NCBI BioSystems Database (<http://www.ncbi.nlm.nih.gov/biosystems/>), a total of 9 and 26 gene sets for the NGA and YAK tests, respectively, have Bonferroni-corrected P -values < 0.05 . Because the Bonferroni-corrected P -values between the tests of these two populations are highly correlated (Pearson correlation = 0.64, $P < 2.2 \times 10^{-16}$; fig. 2A), unless stated otherwise, for simplicity we choose to report P -values and false discovery rates from polygenic analyses based on NGA. Overall, we found little to no correlation between these P -values and either the number of SNVs (Pearson correlation = -0.04 , $P = 0.13$, supplementary fig. S7A, Supplementary Material online) or the number of genes in the gene sets (Pearson correlation = -0.025 , $P = 0.37$, supplementary fig. S7B, Supplementary Material online). Thus, the observed significant signals are not biased by the size of the gene sets. To make our inference more

conservative, we focused on the eight gene sets that are found to be significant in both the NGA and YAK samples (fig. 2A, table 2). We also excluded Glycerophospholipid Biosynthesis from further analysis, because genes in that set are a subset of those in Phospholipid Metabolism (fig. 2B), leaving seven candidate gene sets (table 2). To validate the significance of these seven candidate gene sets, we performed additional tests based on non-parametric permutations of genes and parametric genome simulations (see Materials and Methods). In short, the non-parametric permutation test controls for composition of gene sets, and the parametric simulation-based test controls for demographic history, as well as variable rates of mutation and recombination across the genome. For all seven candidate gene sets, the P -values of both the non-parametric permutation and parametric simulation tests are < 0.05 . This suggests that our candidates are unlikely to be false positives due to gene set composition, demographic history, or heterogeneity of mutation and recombination rates in the genome (table 2).

Among the candidate gene sets, we found that many are diet-related, including Fatty Acid Metabolism (BSID: 868084; Bonferroni's $P = 1.7 \times 10^{-2}$, FDR = 2.4×10^{-3} , fig. 3), which plays significant roles in energy production and the structure of cellular membranes (KEGG: hsa01212; Turner et al. 2014); Pancreatic Secretion (BSID: 169306; Bonferroni's $P = 1.2 \times 10^{-2}$, FDR = 2.1×10^{-3} , fig. 3), which includes genes encoding digestive enzymes (KEGG: hsa04972); and Protein Digestion and Absorption (BSID: 172847; Bonferroni's $P = 1.9 \times 10^{-2}$, FDR = 2.4×10^{-3} , fig. 3), which

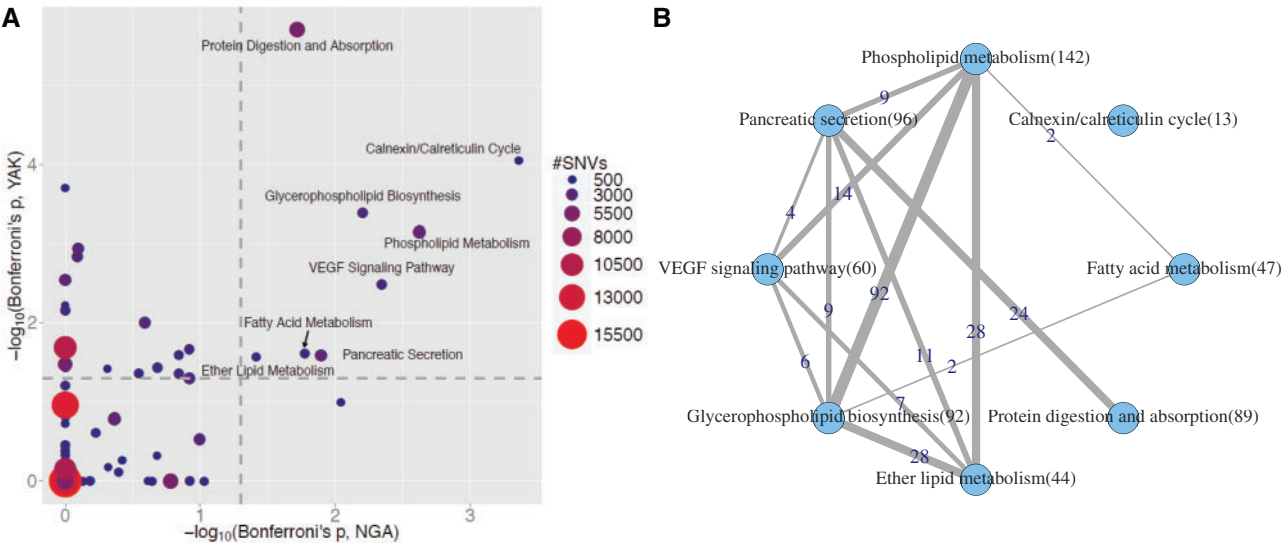


Fig. 2. Candidate gene sets of Siberian-specific polygenic selection. (A) Joint distribution of Bonferroni's *P*-values (in $-\log_{10}$ scale) from the polygenic tests using the NGA- and YAK-specific PBS scores (see Results). Each gene set is plotted as a circle, whose size and color indicate the number of SNVs in that gene set. The dash gray lines are the significance cutoffs, $-\log_{10}(\text{Bonferroni's } P = 0.05)$, for both the polygenic tests. Note that the Bonferroni-corrected *P*-values for the two Siberian populations are highly correlated (Pearson correlation = 0.64, $P < 2.2 \times 10^{-16}$). (B) Genes shared between candidate gene sets. The numbers in the parentheses and on the edges are the numbers of genes within a gene set and shared between gene sets, respectively. The width of an edge reflects the number of genes shared between the two vertices (gene sets).

Table 2. Statistical Significance of the Siberian-Specific Candidate Gene Sets for Polygenic Adaptation.

	# Genes	Bonferroni's <i>P</i> -value	Permutation <i>P</i> -value	Parametric simulation <i>P</i> -value	Major contributing genes (MWU <i>P</i> -value)
Calnexin/calreticulin cycle	13	4.4×10^{-4}	2.57×10^{-2}	1.15×10^{-3}	MAN1B1, UGGT2 (1.30×10^{-20})
Phospholipid metabolism	142	2.3×10^{-3}	1.47×10^{-2}	4.50×10^{-4}	n.a.
VEGF signaling pathway	60	4.5×10^{-3}	1.22×10^{-2}	5.00×10^{-4}	n.a.
Pancreatic secretion	96	1.2×10^{-2}	2.08×10^{-2}	5.50×10^{-4}	PRSS1, PRSS2, PRSS3, CELA3B (4.45×10^{-19})
Fatty acid metabolism	47	1.7×10^{-2}	1.05×10^{-2}	4.00×10^{-4}	ACACA, ECHS1, HADHA, HADHB, FADS1, FADS2, ACAA2 (1.27×10^{-15})
Protein digestion and absorption	89	1.9×10^{-2}	1.94×10^{-2}	1.20×10^{-3}	COL4A3, COL5A1, COL6A1, COL15A1, CPB2, MEP1A, PRSS1, PRSS2, PRSS3, CELA3B, COL18A1 (3.28×10^{-31})
Ether lipid metabolism	44	3.8×10^{-2}	1.68×10^{-2}	7.00×10^{-4}	PAFAH1B2, PAFAH2, PLA2G5, PLA2G7, PLA2G6, AGPS, PLA2G4C, PLD3, PLA2G2E, PLA2G3, LPCAT1, EPT1, PLD4, PLA2G4E, PLA2G4F (2.12×10^{-12})

NOTE.—The (Bonferroni's) *P*-values were calculated based on MWU test, which compares the distributions of PBS between an NCBI BioSystems gene set and the remaining data. Permutation and parametric simulation *P*-values were obtained based on 10,000 random permutations and 10,000 parametric simulations for each of the CEU–CHB–NGA and CEU–CHB–YAK analyses as described in Materials and Methods. Major contributing genes were identified using the leave-one-out procedure described in the main text, and MWU *P*-values were calculated by comparing the PBS distributions of major contributing genes to the remaining data.

includes genes that carry out a series of degradative processes on ingested proteins and affect the absorption of amino acids (KEGG: hsa04974).

The other candidate gene sets are Phospholipid Metabolism (BSID: 1270053, Bonferroni's $P = 2.3 \times 10^{-3}$, FDR = 1.1×10^{-3} , supplementary fig. S9, Supplementary Material online), which includes members that are major constituents of cellular membranes and also involved in important cellular signaling cascades and membrane trafficking activities (Demirkan et al. 2012); Ether Lipid Metabolism (BSID: 82990; Bonferroni's $P = 3.8 \times 10^{-2}$, FDR = 4.3×10^{-3} , fig. 3), which regulates the lifecycle of ether lipids (KEGG: hsa00565), a special class of phospholipids (Brites et al. 2004); VEGF (vascular endothelial growth factor) Signaling Pathway (BSID: 83066; Bonferroni's $P = 4.5 \times 10^{-3}$,

FDR = 1.5×10^{-3} , supplementary fig. S8A, Supplementary Material online), which primarily mediates angiogenesis and has been associated with diseases, including coronary heart diseases and diabetes (Sasso et al. 2005); and Calnexin/Calreticulin Cycle (BSID: 1268722; Bonferroni's $P = 4.4 \times 10^{-4}$, FDR = 4.4×10^{-4} , supplementary fig. S8B, Supplementary Material online), which ensures the proper folding of proteins/glycoproteins in the endoplasmic reticulum and serves as a quality control system by degrading misfolded proteins (REACTOME: R-HSA-532668).

Identification of Major Contributing Genes in Candidate Gene Sets

To investigate whether the observed significant signals of polygenic selection are driven by a subset of important genes

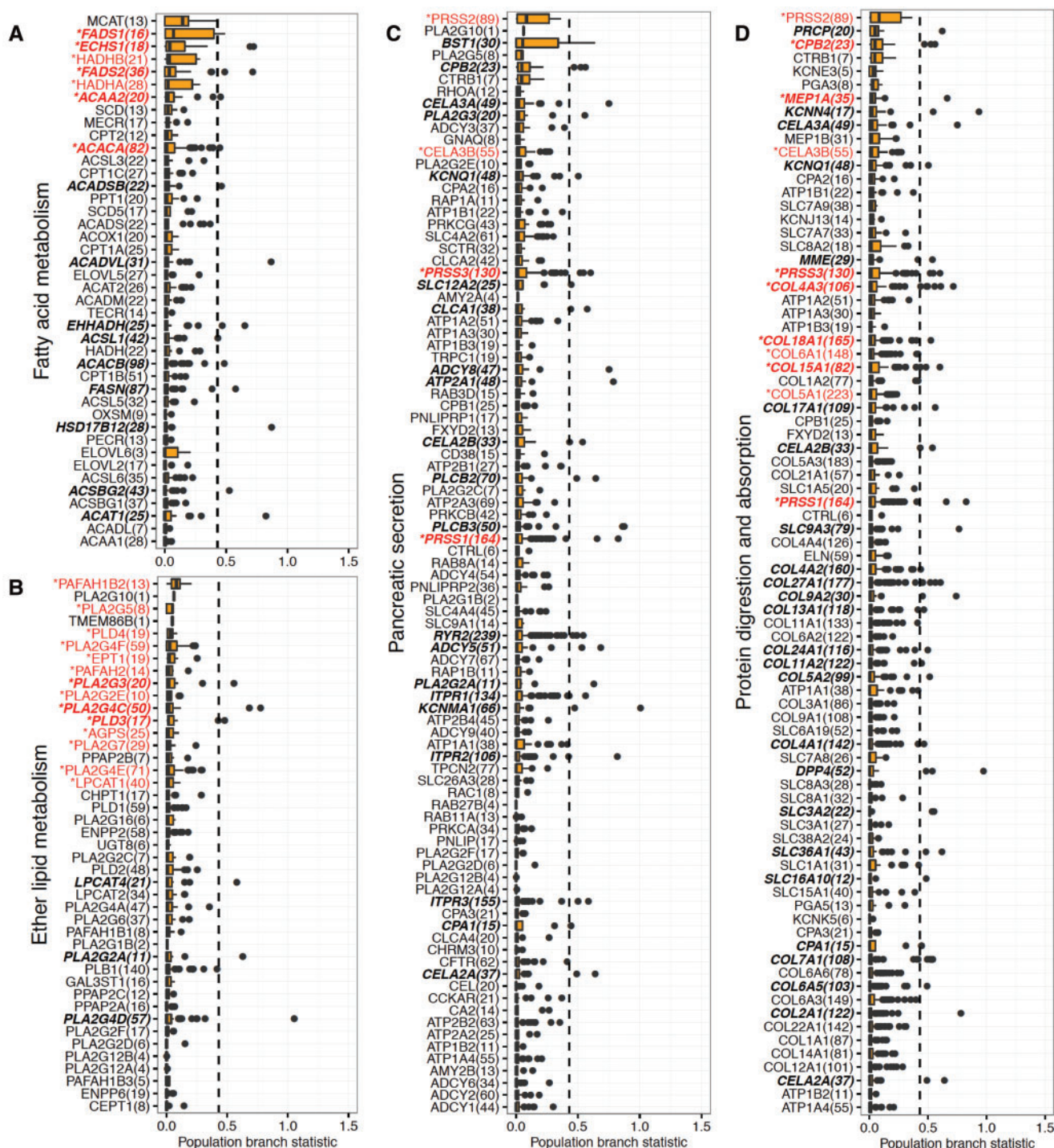


FIG. 3. PBS distributions of genes in candidate gene sets. Each boxplot represents the PBS distribution for a given gene. The number inside the parentheses is the number of SNPs in a gene. The black dashed line indicates the exome-wide top 1% PBS outlier cutoff (PBS > 0.43). Gene names marked with * are the major contributing genes identified by the leave-out-one procedure described in main text (table 2). In addition, genes that have PBS outliers are highlighted in bold-italics. (A) Fatty acid metabolism, (B) Ether lipid metabolism, (C) Pancreatic secretion, and (D) Protein digestion and absorption. See supplementary fig. S8, Supplementary Material online for the plots of the other three candidate gene sets.

within a candidate gene set, we used a leave-one-out analysis to assess the contribution of each individual gene to the polygenic signal (see Materials and Methods). We defined a gene to be a major contributor to a candidate gene set if that gene set is no longer statistically significant when that gene is excluded from the analysis. Interestingly, neither the VEGF

Signaling Pathway nor Phospholipid Metabolism gene sets have any major contributing genes (supplementary figs. S8A and S9, Supplementary Material online, table 2). This suggests that genes in each of these two sets contribute to the observed polygenic selection signal as a whole. On the other hand, we found multiple major contributing genes in the

remaining five candidate gene sets (table 2). Indeed, in each case a much smaller *P*-value was obtained when only these major contributing genes were included in the gene set in the polygenic test (table 2). While we cannot distinguish from which form of polygenic selection (many small allele frequency shifts vs. multiple hard/soft sweeps at many loci) the observed signal arose, our results imply that within each candidate gene set, the major contributing genes might be the primary targets of selection in these Siberians.

Testing for the Hitchhiking Effect of a Selective Sweep on a Single Gene

The signal of polygenic selection may be confounded by the hitchhiking effect of a selective sweep on a single gene. The confounding effect of hitchhiking is even more profound in our MWU test when the favorable gene is in close physical proximity to other genes in the same gene set. To investigate this possibility, we used PBS outliers to detect classical selective sweeps. We defined an SNV as a PBS outlier if its value was in the top 1% of the PBS distribution for our data set ($PBS > 0.43$) and examined the hitchhiking effect around the 10 kb radius of each PBS outlier.

As our first validation, we hypothesized that the observed signal of each candidate gene set was generated by hitchhiking on strongly selected variants in a single gene. If this were true, the observed polygenic signal should be significantly diminished when we excluded PBS outliers and variants surrounding them (10 kb radius of the outliers) in a single gene from our MWU polygenic test. As a second validation, we hypothesized that the observed signal of each of the seven candidate gene sets was generated by hitchhiking on strongly selected variants in nearby genes in the same gene set. To test this scenario, we considered each gene and all genes within 1 Mb of that gene in each gene set and excluded PBS outliers and surrounding variants (in 10 kb radius of the outliers) from these genes. We applied both these filtering procedures to every gene in each candidate gene set and found that, in almost all cases, the candidate gene sets remained significant (Bonferroni's $P < 0.05$; supplementary tables S9–S15, Supplementary Material online). Tests only became non-significant when SNVs were heavily removed from a major contributing gene (supplementary tables S9–S15, Supplementary Material online). As a third validation, we pruned our data by dividing the genome into windows, selecting the largest PBS value variant from each window, and rerunning the MWU test for each candidate gene set using the pruned data. Using window sizes of 100 kb and 10 kb left ~4 and ~15% of the original 508,160 SNVs in the data, respectively. When using these pruned data, three of the seven candidate gene sets (Calnexin/calreticulin cycle, Fatty acid metabolism, as well as Protein digestion and absorption) remained significant (supplementary table S16, Supplementary Material online).

Finally, we performed a gene-based PBS selection scan, using the largest PBS value in each gene as the test statistic, to determine candidate genes under the classic selective sweep model. To account for possible biases due to differences in gene size, the *P*-value of each candidate gene was

defined as the proportion of simulations that had their largest PBS values for the same genomic region equal to or more extreme than that observed from the data. There were 20 genes with *P*-values $< 1 \times 10^{-4}$ in one or both of the analyses for the two population trios CEU–CHB–NGA and CEU–CHB–YAK (supplementary table S17, Supplementary Material online). We chose these 20 genes as our selective sweep candidates because their largest PBS values are highly unlikely to result from the neutral processes accounted for in our simulations. We found no overlap between these genes and those within any of our polygenic candidate gene sets, suggesting the observed polygenic signals are not confounded by classic selective sweeps. While the *P*-values of the *FADS* genes for this test are not significant (supplementary table S18, Supplementary Material online), it has been reported that these genes underwent a strong selective sweep in arctic Greenlandic Inuits (Fumagalli et al. 2015). To test whether this sweep also affected the two Siberian populations, we genotyped the highest PBS variant reported in Fumagalli et al. (2015), rs74771917, in a subset of our Nganasan ($n = 19$) and Yakut ($n = 15$) samples. The selected T allele was at 33 and 40% frequency in the Nganasan and Yakut samples, respectively. This is much lower than the 98% reported in the Inuits, suggesting that different types of selection have been acting in the *FADS* region in different arctic populations. Together, the results of these validations support our hypothesis that selection acting on multiple biologically related genes is likely the driving force for the observed significant signals, rather than the hitchhiking effects of a selective sweep on a single gene.

Discussion

Diet-Related Polygenic Adaptation in Indigenous Central Siberians

In this study, we aimed to search for the signatures of Siberian-specific adaptation by detecting significant allele frequency shifts of variants at multiple genes within a biologically meaningful gene set. Such signatures can arise from small allele frequency changes at many loci or a combination of hard/soft selective sweeps at a moderate number of loci (Chevin and Hospital 2008; Pritchard and Di Rienzo 2010). The rationale for this study design is that many observed physiological adaptations in indigenous Siberians are metabolism-related (Leonard et al. 2005), and thus possibly polygenic (Voruganti et al. 2012). We found seven candidate gene sets for Siberian-specific polygenic selection after Bonferroni correction (fig. 2 and table 2). In order to validate these candidate gene sets, we used a permutation test and found that these signals are unlikely in random gene sets (table 2). In addition, we also performed whole-exome coalescent simulations and showed that our polygenic selection inferences are unlikely to result from non-selective forces, such as demographic history and variation in mutation/recombination rates across the genome (Tajima 1989; Andolfatto and Przeworski 2000; Wall et al. 2002; Schaffner et al. 2005; Teshima et al. 2006; Hsieh et al. 2016). This in turn supports the possible action of selection in our samples.

In addition, using a series of validation tests, we showed that the observed polygenic signal in each candidate gene set is not likely due to the hitchhiking effect of a single selective sweep. While most of the results from these validations suggest the act of selection at many loci on these candidate gene set, our approach cannot determine if the signals are due to purely small allele frequency shifts or sweeps acting in concert with small allele frequency changes at multiple loci (Pritchard and Di Rienzo 2010). Interestingly, when only the top PBS values from the individual 100 and 10 kb windows were used in the MWU tests, only three gene sets (Calnexin/calreticulin cycle, Fatty acid metabolism, as well as Protein digestion and absorption) remained significant, while the selection signals were mostly diminished for the other four (supplementary table S16, Supplementary Material online). Given that only 4 and 15% of the original data remain in the analyses and the MWU *P*-values become evidently larger using the pruned data, it is likely that the tests are underpowered. Nevertheless, we consider these seven candidate gene sets as true positives of polygenic adaptation. Interestingly, many of our candidates are dietary related, especially with respect to the metabolism of fat. Some (major contributing) genes are in more than one candidate gene set. It is possible that the pleiotropic effects of changes in these genes enabled more efficient adaptations. Determining the biological importance of each gene among these candidate gene sets is, however, beyond the scope of the current study.

The major contributing genes *FADS1* and *FADS2* in the candidate set Fatty Acid Metabolism are involved in the synthesis of DHA and AA long chain polyunsaturated fatty acids (LC-PUFAs). Long-chain fatty acids are commonly found in foods such as milk and fish oils and are one of the major energy sources for many tissues. LC-PUFAs are essential for normal development in humans, especially in the brain, and their precursors for synthesis, linoleic acid and alpha-linolenic acid, must be supplied through dietary intake (Darios and Davletov 2006). *FADS1* and *FADS2* are also associated with the levels of low-density lipoprotein (LDL) and high-density lipoprotein (HDL), total cholesterol, type 2 diabetes, and cardiovascular heart diseases (Illig et al. 2010; Demirkan et al. 2012). Different variants in the two genes were reported recently as classic selective sweep candidates for cold adaptation or local dietary requirements in Alaskan/Greenland Inuit (Voruganti et al. 2012; Fumagalli et al. 2015), African (Mathias et al. 2012), European (Ameur et al. 2012; Buckley et al. 2017), Native American (Amorim et al. 2017), and ancient Eurasian populations (Mathieson et al. 2015; Ye et al. 2017). Interestingly, we genotyped the reported selected variant, rs74771917, of the *FADS* region in the arctic Inuits (Fumagalli et al. 2015) for our NGA and YAK samples, but observed much lower frequencies for the favorable allele compared to the Inuits (Results). Thus, it is possible that different types of selection (e.g., a single selective sweep vs. polygenic selection) act on the *FADS* cluster in different arctic populations. Moreover, two other major contributing genes *HADHA* and *HADHB* encode subunits of the enzyme complex mitochondrial trifunctional protein (MTP), which is required to break down long-chain fatty acids, such as LC-PUFAs.

Mutations in both genes are known to lead to MTP deficiency, a metabolic disorder, due to the inability to metabolize long-chain fatty acids for energy production in tissues (Choi et al. 2007). Individuals with MTP deficiency are known to have abnormal body composition (e.g., higher fat mass) and elevated levels of leptin, a hormone secreted by adipose tissue and involved in energy balance through signaling energy sufficiency to the brain and promoting energy expenditure in the periphery (Gillingham et al. 2007). In a recent *in vitro/in vivo* rabbit physiological study, a significant increase in the enzyme level of *HADHA* was observed in rabbits with a diet rich in unsaturated fatty acids (e.g., DHA/AA). A diet rich in unsaturated fatty acids is also associated with low incidences of cardiovascular death in both rabbits and humans (Den Ruijter et al. 2012). *ACAA2*, on the other hand, is associated with both HDL and LDL levels in humans (Choi et al. 2007). A recent study showed that there is a marked reduction in the expression level of *ACAA2* in mice with a high-fat diet, which is known to be associated with cardiovascular diseases, compared to those with normal diet (Soh et al. 2011).

The other two lipid/fat-metabolism-related candidate gene sets found in our analysis are Phospholipid Metabolism and Ether Lipid Metabolism. The major contributing genes, *PLD2* and *AGPAT1*, in the candidate set Phospholipid Metabolism are known to be associated with the levels of phospholipids and sphingolipids in blood plasma, respectively. Altered levels of circulating phospholipids have been implicated in the pathology of type 2 diabetes and cardiovascular diseases (Demirkan et al. 2012). Of the enzymes encoded by the members of Ether Lipid Metabolism, plasmalogens are known to serve as storage for PUFAs. The degradation of plasmalogens, which in turn releases free DHA and AA, is mediated by phospholipidase A2 encoded by the gene family *PLA2*, which includes 8 out of 15 major polygenic signal contributing genes found in our analysis (table 2). In addition, mutations in another major contributing gene, *APGS*, can cause severe growth and neurological abnormalities, and carriers do not survive past childhood (Shi et al. 2015). Interestingly, although not in humans, a recent mutagenesis study in *C. elegans* suggests the importance of ether lipids for the survival of *C. elegans* at cold temperatures (Shi et al. 2015).

The strongest polygenic selection signal in our analysis is from the candidate set Calnexin/Calreticulin Cycle (BSID: 1268722). In the endoplasmic reticulum (ER), genes (e.g., the major contributing gene *MAN1B1*) in this set ensure proper folding of proteins/glycoproteins and serve as a quality control system by degrading misfolded proteins. Interestingly, the membrane of the ER plays an essential role in lipid and protein biosynthesis. It produces all the transmembrane proteins and the lipids for most of the cell's organelles and almost all lipids for mitochondrial and peroxisomal membranes (Alberts et al. 2000). This highlights the important biochemical role of this gene set in both fat and protein metabolisms.

As early modern humans dispersed from their ancestral African homeland into much colder environments, biological adaptations must have occurred to deal with severe cold stress, dramatic variation in photoperiod, and new food resources. Because a hunting and foraging lifestyle provides

the most reliable subsistence in the challenging conditions in Siberia, today many indigenous Siberians, including both the Nganasans and Yakuts, still practice reindeer herding, hunting, trapping, and fishing (Karafet et al. 2002; Weber et al. 2002; Rolland 2014). Although we were not able to directly identify the selected phenotypes using our data, the fact that our candidate gene sets are unequivocally related to diet suggests that these indigenous Siberian people have adapted to their high-fat and protein-rich animal food diets. In particular, physiological studies have shown that indigenous Siberians have markedly higher basal metabolic rates (BMRs) and relatively lower blood lipid levels (e.g., HDL/LDL and triglycerides) (Leonard et al. 2005). Both elevation in BMRs and low blood lipid levels have been attributed to an adaptation to maintain stable lipid levels in the blood plasma and cope with severe cold stress by increasing heat production (Leonard et al. 2005; Cardona et al. 2014). In addition, it is hypothesized that the elevated BMRs among indigenous Siberians may be protective against more dramatic increases in blood lipid levels due to urbanization and increasing consumption of processed market food (Leonard et al. 2005; Snodgrass et al. 2010).

In addition to *FADS1* and *FADS2*, which are associated with long-chain fatty-acid metabolism and have been reported as selective sweep candidates in different human populations (Ameur et al. 2012; Mathias et al. 2012; Voruganti et al. 2012; Clemente et al. 2014; Fumagalli et al. 2015; Mathieson et al. 2015), our analysis revealed several other genes (e.g., *ACAA2*, *HADHA*, *HADHB*, and genes in the *PLA2* family) that may have played roles in the adaptation of energy homeostasis in our Central Siberian samples (Soh et al. 2011; Den Ruijter et al. 2012). Another member in the set of Fatty Acid Metabolism, *CPT1A*, was also recently reported as a candidate classic selective sweep for cold adaptation in Northeastern Siberians (Clemente et al. 2014). The enzyme that *CPT1A* encodes is responsible for transporting long-chain fatty acids into mitochondria for energy production (Lemas et al. 2012). Interestingly, however, this gene was not one of the major contributing genes identified in our polygenic analysis, and in our sample the PBS scores of SNVs in this gene are all below 0.43 (the exome-wise significance cutoff; fig. 3A, table 2). The fact that multiple genes identified in several worldwide populations using different methods and models (polygenic vs. classic selective sweep models) are involved in the same biological function (long-chain fatty acid metabolism) underscores the complexity of adaptation to diet in people at high latitude, perhaps through convergent evolution, similar to high-altitude adaptation in humans (Yi et al. 2010; Huerta-Sanchez et al. 2013; Udpa et al. 2014).

Complex Demographic History of Indigenous Central Siberians

Our exome sequence-based inferences provide the most comprehensive demographic models to date for the Nganasans and Yakuts, as well as new insight into the dynamics of AMH settlement in North Asia. Our best-fit models consistently show that both the Nganasans and the Yakuts are more distantly related to Europeans than to East Asians,

suggesting a common origin of Central Siberians (fig. 1; supplementary tables S4–S6, Supplementary Material online). This finding is consistent with the hypothesis that Central Siberians originated in South Siberia/Central Asia, the area adjacent to Mongolia/North China (Karafet et al. 2002; Derenko et al. 2014; Pugach et al. 2016). This is also consistent with observations of substantial East-Asian ancestry in most Siberians from several recent studies (Fedorova et al. 2013; Cardona et al. 2014; Pugach et al. 2016; Wong et al. 2016). Our best-fit models also show continuous gene flow between the two Siberian groups and East Asians (fig. 1 and table 1). While Yakuts are known to have substantial East-Asian ancestry, relatively low East-Asian ancestry has been observed in the Nganasans (Fedorova et al. 2013; Cardona et al. 2014; Pugach et al. 2016). The inferred gene flow in our models from East-Asians to both the Nganasans and Yakuts implies a stronger genetic influence from East-Asians on Central Siberians than previously thought, but further research is required to test this hypothesis.

Consistent with a recent study (Wong et al. 2016), we found that the ancestors of East Asians and of the ancestors of the two Central Siberian groups diverged ~12–14 kya, after the end of Last Glacial Maximum (LGM, ~26–19 kya) and at the beginning of Holocene (~12 kya). The much milder climate in the Holocene might have permitted humans to expand northward into North Central Siberia (Mayewski et al. 2004). Together with previous evidence for the Southern origin of modern Central Siberians, our estimated divergence time of the ancestors of East Asians and the two Central North Siberian groups supports the hypothesis of post-LGM population expansion from South toward North Central Siberia (Karafet et al. 2002; Derenko et al. 2014; Pugach et al. 2016). In addition, because the 24,000-year-old ancient genome from the Mal'ta site in South Siberia shows no genetic affinity to modern South/Central Siberians (Raghavan et al. 2014), it has been suggested that the population to which this Upper-Paleolithic individual belonged was replaced by the ancestors of present-day Siberians. Our recent divergence estimate between East-Asians and the NGA/YAK supports this replacement model.

Both the contemporary effective population sizes of the Nganasans and the Yakuts estimated in our best-fit models are relatively small. While the Nganasans live in relatively small, isolated groups on the Taymyr Peninsula, the Yakuts are a much larger ethnic group of >450,000 people (Uinuk-Ool et al. 2003; Pugach et al. 2016). Effective population size estimates of the Yakuts vary widely between our best-fit models, but their 95% CIs overlap, so they are statistically compatible (table 1). Two recent studies suggested a historical population bottleneck in the Yakuts, as recent as 1000 years ago (Zlojutro et al. 2009; Pugach et al. 2016). While recent events are indeed important to understand the demographic history of Siberians, we did not attempt to model events less than 5000 years before present, because this would require a larger sample size and higher coverage data to confidently capture rare variants (Robinson et al. 2014). The population sizes inferred in our models are thus best interpreted as long-term effective population sizes.

There are several points that are worth discussing regarding our demographic inferences. First, both our models and most of the parameter estimates that relate to the East Asians and Europeans are highly consistent with those from previous studies (Gutenkunst et al. 2009; Gravel et al. 2011), except for the relatively more recent divergence time between Europeans and East Asians in our models. We refit the Europeans and East Asians portion of the OOA model using only the high-coverage CEU and CHB exome data from 1000 Genomes Project. The resulting divergence time estimate was still younger than previously reported (supplementary table S7, Supplementary Material online), suggesting that the low-coverage Siberian data do not strongly bias our inferences. Gravel et al. (2011) combined both CHB and JPT samples to represent East Asians, which could lead to an increase in low-frequency variants relative to that in CHB alone. Moreover, the previous OOA models incorporate migration from Africans into both Europeans and East Asians, which would also increase the amount of low-frequency variants. We did not include this demographic feature in our models, because we focused on demographic processes taking place among Siberians and the two major Eurasian populations. The discrepancy in the estimated divergence times between our models and previous studies is thus likely due to differences in the population samples analyzed and the underlying models employed, rather than due to our low-coverage Siberian data. In addition, our inferred dates are based on a generation time of 25 years and a phylogeny-based mutation rate of 2.35×10^{-8} per-site per-generation (Gutenkunst et al. 2009; compatible with Nachman and Crowell 2000). We used the phylogenetic estimate because of its long history in population genetics. Our date estimates would be about two times older if we used the mean rate of $\sim 1.5 \times 10^{-8}$ per-site per-generation estimated by recent pedigree-based exome sequencing studies (Sally 2016) as well as a longer generation time of 29 years (Langergraber et al. 2012) (supplementary table S8, Supplementary Material online). For example, the divergence time between the ancestors of East-Asians and the two Siberian groups would be pushed back further to ~ 24 – 28 kya, a time around the LGM. This new date would not affect the aforementioned implications for the hypotheses of post-LGM population expansion and replacement. Nevertheless, caution is still advised when comparing population genetic date estimates with the fossil record.

Materials and Methods

Human Samples and Exome Sequencing

DNA samples of both Nganasans (NGA, 21 samples) and Yakuts (YAK, 21 samples) were collected during field expeditions in Siberia by two of us (L.P.O. and T.M.K.) and our collaborators from the Institute of Cytology and Genetics, Novosibirsk, Russia. Informed consent was obtained from all collected human subjects, and this study was approved by the Human Subjects Committee at the University of Arizona (#HSC A93.05) and the Institute of Cytology and Genetics. Exomes of these samples were captured using the Agilent SureSelect All Exon v5 no UTRs capture kit and

pooled sequenced using Illumina HiSeq 2000. After applying the standard quality-control procedures of the University of Arizona Genetics Core, read mate pairs were aligned to the human reference genome (hg19), resulting in $\sim 6\times$ and $\sim 4\times$ mean coverage on target for the NGA and YAK samples, respectively. We also downloaded BAM files for 97 European (CEU) and 95 Han Chinese (CHB) exomes from the 1000 Genomes Project (1000 Genomes Project Consortium 2012). As several different exome capture kits were used, we created a BED file containing the union of all possible exome targets padded by 100 bp on either side, resulting in $\sim 81,441,327$ bp of potential target.

Because of the low average coverage of our Siberian exomes, we developed an inference pipeline based on the program ANGSD, which was designed to compute common statistics for low coverage data (Korneliussen et al. 2014). For all computations, we only used sites from autosomal regions where we had coverage for all populations. We iteratively computed the 1-D, 2-D, and 3-D unfolded joint SFSs for the populations, pairs, and trios, using ANGSD. Ancestral states were determined using the Enredo-Pecan-Ortheus (EPO) 6-way primate alignment (1000 Genomes Project Consortium 2012). F_{ST} between pairs of populations was estimated at each site using the program ngsTools (Fumagalli et al. 2014). Sites with <0.95 posterior probability of being variable among all populations were removed, resulting in a total of 508,160 SNVs in both exonic and intronic sequences. Computation of the PBS and other analyses and processing were performed with custom scripts in R and Python. This inference pipeline explicitly accounts for genotyping uncertainty and infers statistics of interest without explicitly calling genotypes. To ensure the quality of statistical inference, in both population trios, CEU–CHB–NGA and CEU–CHB–YAK, we further excluded sites with pairwise $F_{ST} > 0$, but with $<90\%$ pairwise posterior probability of being variable, as estimated by ngsTools. This resulted in a total of 426,195 variants in the remaining data for polygenic selection tests.

Inference of Demographic History Using $\partial a \partial i$

We used the SFS based demographic inference tool $\partial a \partial i$ (Gutenkunst et al. 2009) to build and fit our demographic models. To avoid possible biases in our demographic inferences due to natural selection, we annotated our data using ANNOVAR (Wang et al. 2010) and built three unfolded joint SFSs for the population trios CHB–YAK–NGA, CEU–CHB–NGA, and CEU–CHB–YAK using intronic sequences from our data. This resulted in $\sim 227,321$, $\sim 270,604$, and $\sim 272,990$ intronic SNVs in the SFSs of CHB–YAK–NGA, CEU–CHB–NGA, and CEU–CHB–YAK, respectively, from a total of 42,325,419 intronic base pairs. The per-sample mean coverage is lower at these intronic sites, with an average of $2.5\times$ for NGA and $2.4\times$ for YAK. However, ANGSD estimates SFS at each site using pooled reads from all samples, and the per-population mean coverages of our intronic data are $74\times$ and $50.3\times$ for our NGA and YAK population samples, respectively. To account for possible ancestral state misidentification, in our $\partial a \partial i$ models, we included an additional

parameter P_{flip} for the fraction of ancestral states that were misidentified (Ragsdale et al. 2016).

To estimate demographic parameters, the derivative-based BFGS algorithm was used to optimize the composite log-likelihood. We used the Godambe Information Matrix to estimate confidence intervals for model parameters and adjusted the test statistics of LRT for linkage, comparing with a null distribution that was an equal mixture of Chi-squared distributions with degrees of freedom 0 and 1 (Coffman et al. 2016), as implemented in *∂a∂i*. Because this statistical approach requires bootstraps of SFS that preserve linkage, for each of the population trios we generated 100 Bootstraps of the SFS by summing over sampled spectra (with replacement) generated from 124 25-Mb non-overlapping regions across the entire sequence data. Each of the 25-Mb SFS was estimated using the same inference pipeline described above.

We used the phylogenetic-based mutation rate of 2.35×10^{-8} per-site per-generation (Gutenkunst et al. 2009; compatible with Nachman and Crowell 2000) and a generation time of 25 years to calculate parameter estimates in physical units.

Coalescent Whole-Exome Simulations

We used MaCS (Chen et al. 2009) for our coalescent simulations, because of its ability to efficiently perform large genomic scale simulations. We used the HapMap genetic map (The International HapMap Consortium 2007) to model recombination rate variation across the entire exome. To model mutational heterogeneity, we carried out a simulation procedure similar to that described in Hsieh et al. (2016). In brief, for each locus j , we performed MaCS simulations using a mutation parameter $\hat{\theta}_j$, defined as $\left(\frac{S_j}{\hat{S}}\right)$, where S_j is the number of variants in the locus and $\hat{S}$ is the expected number of variants given the best-fit demographic model with mutation parameter $\theta = 1$. All simulations presented here model the effects of demography, recombination rate heterogeneity, and mutation rate heterogeneity. Only regions with sequence coverage in our exome data were simulated. Supplementary fig. S6, Supplementary Material online, shows that this simulation approach indeed reproduces the pattern of local mutation density in the real data well (Pearson correlation = 0.88).

Test for Classic and Polygenic Adaptation

To search for signals of polygenic selection, we assessed if any gene set was enriched with highly differentiated SNVs in the population of interest using the PBS. PBS is defined as in Yi et al. (2010), which estimates the branch length of the focal population A under neutrality by transforming the pairwise F_{ST} of three populations A, B, and C (supplementary fig. S5, Supplementary Material online).

$$\text{PBS}_A = \frac{T(F_{\text{ST-AB}}) + T(F_{\text{ST-AC}}) - T(F_{\text{ST-BC}})}{2}, \quad (1)$$

where $F_{\text{ST-AB}}$, $F_{\text{ST-AC}}$, and $F_{\text{ST-BC}}$ are the F_{ST} for population pairs A–B, A–C, and B–C, respectively, and $T = -\log(1 - F_{\text{ST}})$. In short, PBS measures the branch length that leads to the focal population of interest for each SNV; the longer the branch length leading to the focal population, the larger

the allele frequency change is for this population compared to others, which can be a result of natural selection. For both the data and simulations, we calculated PBS values for all SNVs within the same genomic regions that are defined as genes in the real data (RefSeq, downloaded from USCS Genome Browser in May 2015). We downloaded the curated human-related (organism: *Homo sapiens*) gene sets from the NCBI BioSystems database (<http://www.ncbi.nlm.nih.gov/bio> systems/, as of May 2015). We excluded gene sets with <10 genes, and this resulted in 1,206 gene sets. In addition, in each gene set, we also excluded sex-linked genes, because our demographic inference and simulations were specifically designed for autosomal data and do not control for possible biases due to differences in evolutionary history between autosomes and sex chromosomes. One-sided (alternative distribution is greater than the null) MWU tests were performed in R (R Core Team 2015) to quantify the difference in PBS distribution between a gene set (the alternative) and the rest of data (the null). This test was applied to the PBS estimates inferred from the two population trios: CEU–CHB–NGA and CEU–CHB–YAK. Therefore, we called a gene set a Siberian-specific candidate gene set if it was found to be significant (Bonferroni's $P < 0.05$) in both analyses.

Validation of Candidate Gene Sets

To validate each of the candidate gene sets, we calculated two types of P -value: (1) non-parametric permutation P -value, and (2) parametric simulation P -value. For the permutation P -values, in order to generate pseudo-gene sets that had similar numbers of SNVs as the data, we generated 10,000 non-parametric permutations by shuffling and re-assigning gene labels only within genes with similar amount of SNVs: <11, 11–20, 21–50, 51–100, and >101 SNVs. For the parametric P -values, we ran 10,000 whole-exome simulations under each of the best-fit demographic models of the two population trios of CEU–CHB–NGA and CEU–CHB–YAK (fig. 1 and table 1) using the method described in the previous section. To account for the uncertainty of parameter estimates, the demographic parameters for each simulation were drawn from their confidence intervals, assuming a multivariate normal distribution. Thus, the parametric P -value of a gene set was calculated as:

$$\frac{\sum_{s \in \text{Simulations}} I(\text{being significant under } s)}{|\text{Simulations}|}, \quad (2)$$

where $|\text{Simulations}|$ is the total number of simulations, s is a given simulation, and I is an indicator function of being significant under s using the same MWU based polygenic test (Bonferroni's $P < 0.05$).

We performed a leave-one-out procedure, in which we defined that a gene is a major contributing gene for a candidate gene set if the non-parametric (MWU) polygenic test is not statistically significant when the SNVs of that gene were excluded. In addition, it is possible that PBS outliers (defined as if $\text{PBS} > 0.43$, which is the top 1% cutoff for the PBS distribution of the entire data) at a locus might result from a classic selective sweep and/or its hitchhiking effect and in turn drive the observed signal of polygenic selection. To

test this hypothesis, we re-examined statistical significance by removing PBS outliers and all SNVs within 10 kb radius of each PBS outlier for each single gene in the candidate gene set. We also repeated the procedure, but excluded PBS outliers in genes that are within 1 Mb radius of the gene set. In addition, we assessed if a gene is a candidate for a classic selective sweep using the largest PBS value of the SNVs in the gene as the test statistic. In this case, for each gene the statistical significance of this test is defined as the fraction of the simulations that have their largest PBS values for the same region greater than the observed largest PBS value.

Data Accessibility and Databases

The full exome data for the two Siberian population samples are available from the NCBI Sequence Read Archive with accession BioProjectID: PRJNA389435. A variety of databases that were used in this study, including the NCBI BioSystems (<http://www.ncbi.nlm.nih.gov/biosystems/>, as of May 2015), the Kyoto Encyclopedia of Genes and Genomes (KEGG: <http://www.genome.jp/kegg/>), and REACTOME (<http://www.reactome.org/>).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This study was supported by National Science Foundation grants to T.M.K. and M.F.H. (PLR-1203874), to R.N.G. (DEB-1146074), and by the State Research Project (No. 0324-2016-0002) to L.P.O.

References

- 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.
- Alberts B, Lewis J, Bray D. 2000. *Molecular biology of the cell*. New York: Garland Science.
- Ameur A, Enroth S, Johansson A, Zaboli G, Igl W, Johansson AC, Rivas MA, Daly MJ, Schmitz G, Hicks AA. 2012. Genetic adaptation of fatty-acid metabolism: a human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *Am J Hum Genet*. 90(5):809–820.
- Amorim CEG, Nunes K, Meyer D, Comas D, Bortolini MC, Salzano FM, Hunemeier T. 2017. Genetic signature of natural selection in first Americans. *Proc Natl Acad Sci U S A*. 114:2195–2199.
- Andolfatto P, Przeworski M. 2000. A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics* 156(1):257–268.
- Berg JJ, Coop G. 2014. A population genetic signal of polygenic adaptation. *PLoS Genet*. 10(8):e1004412.
- Brites P, Waterham HR, Wanders RJ. 2004. Functions and biosynthesis of plasmalogens in health and disease. *Biochimica Et Biophysica Acta* 1636(2–3):219–231.
- Buckley MT, Racimo F, Allentoft ME, Jensen MK, Jonsson A, Huang H, Hormozdiari F, Sikora M, Marnetto D, Eskin E, et al. 2017. Selection in Europeans on Fatty Acid Desaturases Associated with Dietary Changes. *Mol Biol Evol*. 34(6):1307–1318.
- Cardona A, Pagani L, Antao T, Lawson DJ, Eichstaedt CA, Yngvadottir B, Shwe MTT, Wee J, Romero IG, Raj S, et al. 2014. Genome-wide analysis of cold adaptation in indigenous Siberian populations. *PLoS One* 9(5):e98076.
- Chen GK, Marjoram P, Wall JD. 2009. Fast and flexible simulation of DNA sequence data. *Genome Res*. 19(1):136–142.
- Chevin LM, Hospital F. 2008. Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics* 180(3):1645–1660.
- Choi JH, Yoon HR, Kim GH, Park SJ, Shin YL, Yoo HW. 2007. Identification of novel mutations of the HADHA and HADHB genes in patients with mitochondrial trifunctional protein deficiency. *Int J Mol Med*. 19:81–87.
- Clemente FJ, Cardona A, Inchley CE, Peter BM, Jacobs G, Pagani L, Lawson DJ, Antao T, Vicente M, Mitt M, et al. 2014. A selective sweep on a deleterious mutation in CPT1A in Arctic populations. *Am J Hum Genet*. 95(5):584–589.
- Coffman AJ, Hsieh PH, Gravel S, Gutenkunst RN. 2016. Computationally efficient composite likelihood statistics for demographic inference. *Mol Biol Evol*. 33(2):591–593.
- Darios F, Davletov B. 2006. Omega-3 and omega-6 fatty acids stimulate cell membrane expansion by acting on syntaxin 3. *Nature* 440(7085):813–817.
- Demirkan A, van Duijn CM, Ugocsai P, Isaacs A, Pramstaller PP, Liebisch G, Wilson JF, Johansson A, Rudan I, Aulchenko YS, et al. 2012. Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. *PLoS Genet*. 8(2):e1002490.
- Den Ruijter HM, Verkerk AO, Schumacher CA, Houten SM, Belterman CNW, Baartscheer A, Brouwer IA, van Bilsen M, de Roos B, Coronel R. 2012. A diet rich in unsaturated fatty acids prevents progression toward heart failure in a rabbit model of pressure and volume overload. *Circ-Heart Fail*. 5(3):376–384.
- Derenko M, Malyarchuk B, Denisova G, Perkova M, Litvinov A, Grzybowski T, Dambueva I, Skonieczna K, Rogalla U, Tsybovsky I, et al. 2014. Western Eurasian ancestry in modern Siberians based on mitogenomic data. *BMC Evol Biol*. 14(1):217.
- Devlin MJ. 2015. The “Skinny” on brown fat, obesity, and bone. *Am J Phys Anthropol*. 156:98–115.
- Fedorova SA, Reidla M, Metspalu E, Metspalu M, Rootsi S, Tambets K, Trofimova N, Zhadanov SI, Kashani BH, Olivieri A, et al. 2013. Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *BMC Evol Biol*. 13(1):127.
- Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PL, Aximu-Petri A, Prufer K, de Filippo C, et al. 2014. Genome sequence of a 45, 000-year-old modern human from western Siberia. *Nature* 514(7523):445–449.
- Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jorgensen ME, Korneliusen TS, Gerbault P, Skotte L, Linneberg A, et al. 2015. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* 349(6254):1343–1347.
- Fumagalli M, Vieira FG, Linderoth T, Nielsen R. 2014. ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics* 30(10):1486–1487.
- Gillingham MB, Purnell JQ, Jordan J, Stadler D, Haqq AM, Harding CO. 2007. Effects of higher dietary protein intake on energy balance and metabolic control in children with long-chain 3-hydroxy acyl-CoA dehydrogenase (LCHAD) or trifunctional protein (TFP) deficiency. *Mol Genet Metab*. 90(1):64–69.
- Goltsova TV, Osipova LP, Zhadanov SI, Vilems R. 2005. The effect of marriage migration on the genetic structure of the Taimyr Nganasan population: genealogical analysis inferred from MtDNA markers. *Genetika* 41(7):954–965.
- Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Genomes P, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A*. 108:11983–11988.
- Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations

- from multidimensional SNP frequency data. *PLoS Genet.* 5(10):e1000695.
- Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, Sukernik R, Utermann G, Pritchard J, Coop G, et al. 2010. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci U S A.* 107(Suppl 2):8924–8930.
- Hsieh P, Veeramah KR, Lachance J, Tishkoff SA, Wall JD, Hammer MF, Gutenkunst RN. 2016. Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. *Genome Res.* 26(3):279–290.
- Huerta-Sanchez E, Degiorgio M, Pagani L, Tarekgn A, Ekong R, Antao T, Cardona A, Montgomery HE, Cavalleri GL, Robbins PA, et al. 2013. Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations. *Mol Biol Evol.* 30:1877–1888.
- Illig T, Gieger C, Zhai GJ, Romisch-Margl W, Wang-Sattler R, Prehn C, Altmaier E, Kastenmuller G, Kato BS, Mewes HW, et al. 2010. A genome-wide perspective of genetic variation in human metabolism. *Nat Genet.* 42(2):137–U166.
- Karafet TM, Osipova LP, Gubina MA, Posukh OL, Zegura SL, Hammer MF. 2002. High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Hum Biol.* 74(6):761–789.
- Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, Cooper GM, Roos C, Voight BF, Havulinna AS, et al. 2008. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet.* 40(2):189–197.
- Korneliusson TS, Albrechtsen A, Nielsen R. 2014. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 15:356.
- Langergraber KE, Prufer K, Rowney C, Boesch C, Crockford C, Fawcett K, Inoue E, Inoue-Muruyama M, Mitani JC, Muller MN, et al. 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci U S A.* 109(39):15716–15721.
- Lemas DJ, Wiener HW, O'Brien DM, Hopkins S, Stanhope KL, Havel PJ, Allison DB, Fernandez JR, Tiwari HK, Boyer BB. 2012. Genetic polymorphisms in carnitine palmitoyltransferase 1A gene are associated with variation in body composition and fasting lipid traits in Yup'ik Eskimos. *J Lipid Res.* 53(1):175–184.
- Leonard WR, Snodgrass JJ, Sorensen MV. 2005. Metabolic adaptation in indigenous Siberian populations. *Annu Rev Anthropol.* 34(1):451–471.
- Mathias RA, Fu W, Akey JM, Ainsworth HC, Torgerson DG, Ruczinski I, Sergeant S, Barnes KC, Chilton FH. 2012. Adaptive evolution of the FADS gene cluster within Africa. *PLoS One* 7(9):e44926.
- Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528(7583):499.
- Mayewski PA, Rohling EE, Stager JC, Karlen W, Maasch KA, Meeker LD, Meyerson EA, Gasse F, van Kreveld S, Holmgren K, et al. 2004. Holocene climate variability. *Quaternary Res.* 62(03):243–255.
- Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156(1):297–304.
- Pitulko VV, Nikolsky PA, Giryay EY, Basilyan AE, Tumskoy VE, Koulakov SA, Astakhov SN, Pavlova EY, Anisimov MA. 2004. The Yana RHS site: humans in the Arctic before the last glacial maximum. *Science* 303(5654):52–56.
- Pitulko VV, Tikhonov AN, Pavlova EY, Nikolskiy PA, Kuper KE, Polozov RN. 2016. Paleoanthropology. Early human presence in the Arctic: evidence from 45,000-year-old mammoth remains. *Science* 351(6270):260–263.
- Pritchard JK, Di Rienzo A. 2010. Adaptation—not by sweeps alone. *Nat Rev Genet.* 11(10):665–667.
- Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol.* 20(4):R208–R215.
- Pugach I, Matveev R, Spitsyn V, Makarov S, Novgorodov I, Osakovsky V, Stoneking M, Pakendorf B. 2016. The complex admixture history and recent southern origins of Siberian populations. *Mol Biol Evol.* 33(7):1777–1795.
- R Core Team. 2015. R: A language and environment for statistical computing.
- Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Jr., Orlando L, Metspalu E, et al. 2014. Upper palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* 505:87–91.
- Ragsdale AP, Coffman AJ, Hsieh P, Struck TJ, Gutenkunst RN. 2016. Triallelic population genomics for inferring correlated fitness effects of same site nonsynonymous mutations. *Genetics* 203(1):513–523.
- Robinson JD, Coffman AJ, Hickerson MJ, Gutenkunst RN. 2014. Sampling strategies for frequency spectrum-based population genomic inference. *BMC Evol Biol.* 14:254.
- Rolland N. 2014. The pleistocene peopling of the north: paleolithic milestones and thresholds horizons in Northern Eurasia: Part I: lower paleolithic antecedents. *Archaeol Ethnol Anthropol Eurasia* 42(1):2–17.
- Sasso FC, Torella D, Carbonara O, Ellison GM, Torella M, Scardone M, Marra C, Nasti R, Marfella R, Cozzolino D, et al. 2005. Increased vascular endothelial growth factor expression but impaired vascular endothelial growth factor receptor signaling in the myocardium of type 2 diabetic patients with chronic coronary heart disease. *J Am Coll Cardiol.* 46(5):827–834.
- Scally A. 2016. The mutation rate in human evolution and demographic inference. *Curr Opin Genet Dev.* 41:36–43.
- Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15(11):1576–1583.
- Shi X, Tarazona P, Brock TJ, Browse J, Feussner I, Watts JL. 2015. A C. elegans model for ether lipid biosynthesis and function. *J Lipid Res.* 57:265–275.
- Simchenko YB. 1976. Cultura okhotnikov na oleney severnoy Yevrazy (Culture of the hunters of cervines in northern Eurasia). Moscow: Science.
- Snodgrass JJ, Leonard WR, Tarskala LA, Egorova AG, Maharova NV, Pinigina IA, Halyev SD, Matveeva NP, Romanova AN. 2010. Impaired fasting glucose and metabolic syndrome in an indigenous Siberian population. *Int J Circumpol Heal.* 69(1):87–98.
- Soh J, Kwon DY, Cha YS. 2011. Hepatic gene expression profiles are altered by dietary unsalted Korean fermented soybean (chongkuk-jang) consumption in mice with diet-induced obesity. *J Nutr Metab.* 2011:260214.
- Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.
- Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* 16(6):702–712.
- The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.
- Turner N, Cooney GJ, Kraegen EW, Bruce CR. 2014. Fatty acid metabolism, energy expenditure and insulin resistance in muscle. *J Endocrinol.* 220(2):T61–T79.
- Udpa N, Ronen R, Zhou D, Liang J, Stobdan T, Appenzeller O, Yin Y, Du Y, Guo L, Cao R, et al. 2014. Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. *Genome Biol.* 15(2):R36.
- Uinuk-Ool TS, Takezaki N, Klein J. 2003. Ancestry and kinships of native Siberian populations: the HLA evidence. *Evol Anthropol.* 12(5):231–245.
- Voruganti VS, Higgins PB, Ebbesson SO, Kennish J, Goring HH, Haack K, Laston S, Drigalenko E, Wenger CR, Harris WS, et al. 2012. Variants in CPT1A, FADS1, and FADS2 are associated with higher levels of estimated plasma and erythrocyte Delta-5 desaturases in Alaskan Eskimos. *Front Genet.* 3:86.

- Wall JD, Andolfatto P, Przeworski M. 2002. Testing models of selection and demography in *Drosophila simulans*. *Genetics* 162(1): 203–216.
- Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38(16):e164.
- Weber AW, Link DW, Katzenberg MA. 2002. Hunter-gatherer culture change and continuity in the Middle Holocene of the Cis-Baikal, Siberia. *J Anthropol Archaeol.* 21(2):230–299.
- Wong EH, Khrunin A, Nichols L, Pushkarev D, Khokhrin D, Verbenko D, Evgrafov O, Knowles J, Novembre J, Limborska S, et al. 2016. Reconstructing genetic history of Siberian and Northeastern European populations. *Genome Res.* 27(1):1.
- Ye K, Gao F, Wang D, Bar-Yosef O, Keinan A. 2017. Dietary adaptation of FADS genes in Europe varied across time and geography. *Nat Ecol Evol.* 1:0167.
- Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329:75–78.
- Zlotutro M, Tarskaia LA, Sorensen M, Snodgrass JJ, Leonard WR, Crawford MH. 2009. Coalescent simulations of Yakut mtDNA variation suggest small founding population. *Am J Phys Anthropol.* 139(4):474–482.