

## Article

# Genetic architecture in Greenland is shaped by demography, structure and selection

<https://doi.org/10.1038/s41586-024-08516-4>

Received: 6 July 2024

Accepted: 11 December 2024

Published online: 12 February 2025

Open access

 Check for updates

Frederik Filip Stæger<sup>1</sup>, Mette K. Andersen<sup>2</sup>, Zilong Li<sup>1</sup>, Jasmin Pernille Hjerresen<sup>2</sup>, Shixu He<sup>1</sup>, Cindy G. Santander<sup>1</sup>, Rasmus Tanderup Jensen<sup>2</sup>, Karsten Fleischer Rex<sup>3,4</sup>, Anne Cathrine Baun Thuesen<sup>2</sup>, Kristian Hanghøj<sup>1</sup>, Inge Høst Seiding<sup>5</sup>, Emil Jørsboe<sup>2,6,7</sup>, Sara Elizabeth Stinson<sup>2</sup>, Malthe Sebro Rasmussen<sup>1</sup>, Renzo F. Balboa<sup>1</sup>, Christina Viskum Lytken Larsen<sup>8,9</sup>, Peter Bjerregaard<sup>8</sup>, Mikkel Schubert<sup>2</sup>, Jonas Meisner<sup>2</sup>, Allan Linneberg<sup>10,11</sup>, Niels Grarup<sup>2</sup>, Eleftheria Zeggini<sup>12,13</sup>, Rasmus Nielsen<sup>14,15</sup>, Marit E. Jørgensen<sup>8,9,16</sup>, Torben Hansen<sup>2</sup>✉, Ida Moltke<sup>1</sup>✉ & Anders Albrechtsen<sup>1</sup>✉

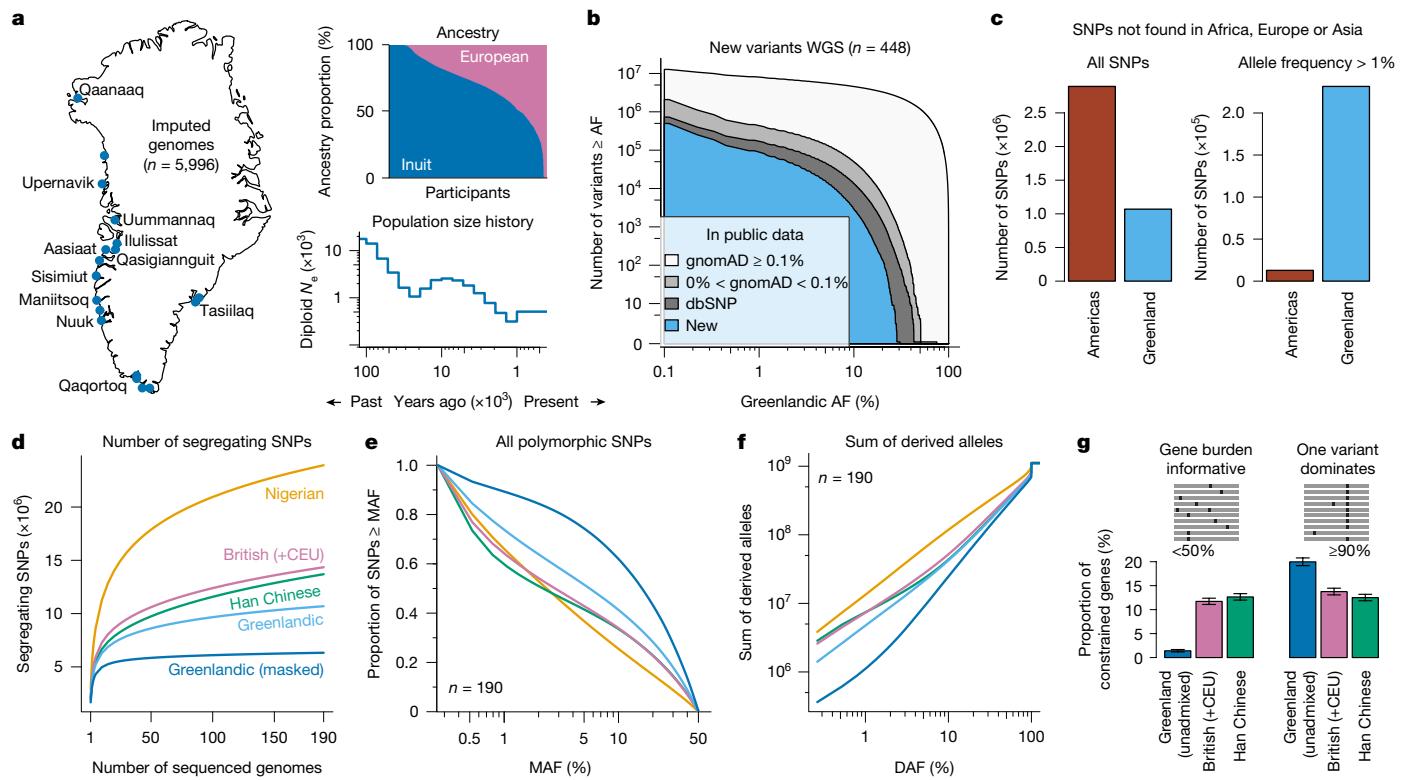
Greenlandic Inuit and other indigenous populations are underrepresented in genetic research<sup>1,2</sup>, leading to inequity in healthcare opportunities. To address this, we performed analyses of sequenced or imputed genomes of 5,996 Greenlanders with extensive phenotypes. We quantified their historical population bottleneck and how it has shaped their genetic architecture to have fewer, but more common, variable sites. Consequently, we find twice as many high-impact genome-wide associations to metabolic traits in Greenland compared with Europe. We infer that the high-impact variants arose after the population split from Native Americans and thus are Arctic-specific, and show that some of them are common due to not only genetic drift but also selection. We also find that European-derived polygenic scores for metabolic traits are only half as accurate in Greenlanders as in Europeans, and that adding Arctic-specific variants improves the overall accuracy to the same level as in Europeans. Similarly, lack of representation in public genetic databases makes genetic clinical screening harder in Greenlandic Inuit, but inclusion of Greenlandic data remedies this by reducing the number of non-causal candidate variants by sixfold. Finally, we identify pronounced genetic fine structure that explains differences in prevalence of monogenic diseases in Greenland and, together with recent changes in mobility, leads to a predicted future reduction in risk for certain recessive diseases. These results illustrate how including data from Greenlanders can greatly reduce inequity in genomic-based healthcare.

Genetics is increasingly important for diagnosis, risk assessment and treatment of disease. However, genetics research has been carried out predominantly in people of European genetic ancestry. This introduces inequity in healthcare opportunities, as population-specific disease variants remain undiscovered, clinical genetic screens for rare diseases lack reference panels with population-specific allele frequencies and the accuracy of polygenic risk assessment is reduced significantly<sup>3–5</sup>. Fortunately, genetic research in currently underrepresented populations, such as the Indigenous African, Mexican and Australian populations, is now being prioritized<sup>6–11</sup>. In the Greenlandic population, single nucleotide polymorphism (SNP)-array studies have identified several Arctic-specific high-impact variants, including variants associated

with diabetes in *TBC1D4* (refs. 12,13) and *HNF1A*<sup>14</sup> and obesity in *ADCY3* (ref. 15). Linkage studies have identified recessive lethal variants in *ATP8B1* and *PCCB*<sup>16,17</sup>. The *TBC1D4* variant, and also an *S*l variant causing congenital sucrase-isomaltase deficiency (CSID)<sup>12,13,18</sup>, are already being used to diagnose and treat disease in Greenland and other Arctic populations. However, an in-depth characterization of whole genomes from Inuit, including Greenlanders, is lacking. For this reason, the genetic architecture of disease in Greenland, and related Arctic populations, has not been described and these populations are not represented in genetic reference databases.

The Greenlandic population has been historically small and isolated. The Inuit ancestors entered northwestern Greenland from Arctic

<sup>1</sup>Department of Biology, Section for Computational and RNA Biology, University of Copenhagen, Copenhagen, Denmark. <sup>2</sup>Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. <sup>3</sup>Department of Internal Medicine, Queen Ingrid's Hospital, Nuuk, Greenland. <sup>4</sup>Department of Clinical Medicine, Arctic Health Research Centre, Aalborg University Hospital, Aalborg, Denmark. <sup>5</sup>Ilisimatusarfik - University of Greenland, Nuuk, Greenland. <sup>6</sup>Li Ka Shing Centre for Health Information and Discovery, Big Data Institute, University of Oxford, Oxford, UK. <sup>7</sup>Nuffield Department of Population Health, University of Oxford, Oxford, UK. <sup>8</sup>Centre for Public Health in Greenland, National Institute of Public Health, University of Southern Denmark, Copenhagen, Denmark. <sup>9</sup>Greenland Center for Health Research, Institute for Health and Nature, University of Greenland, Nuuk, Greenland. <sup>10</sup>Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospital, The Capital Region of Denmark, Copenhagen, Denmark. <sup>11</sup>Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. <sup>12</sup>Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany. <sup>13</sup>TUM School of Medicine and Health, Technical University of Munich (TUM) and Klinikum Rechts der Isar, Munich, Germany. <sup>14</sup>Departments of Integrative Biology and Statistics, University of California at Berkeley, California, CA, USA. <sup>15</sup>GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark. <sup>16</sup>Steno Diabetes Center Greenland, Nuuk, Greenland. ✉e-mail: torben.hansen@sund.ku.dk; ida@bio.ku.dk; aalborgtsen@bio.ku.dk



**Fig. 1 | Genetic architecture comparisons.** **a**, Sample locations in Greenland, inferred admixture proportions and historic effective population size estimated from an ARG of the Greenlandic (masked,  $n = 150$ ) Inuit using Relate. **b**, Number of variants (both SNPs and insertion/deletions) as a function of allele frequency (AF), for example, at AF = 10%, we show the number of variants with AF  $\geq 10\%$ . The variant counts are grouped by whether the variant is new, found only in dbSNP, in gnomAD with AF  $< 0.1\%$  or in gnomAD (v.3.0) with AF  $\geq 0.1\%$ . Note the logarithmic axis. **c**, Number of SNPs not found in any African, European or Asian populations in gnomAD for both 448 1KG people from the Americas and 448 Greenlanders. Note the different y axis scale on the two barplots. **d**, Number of segregating SNPs as a function of how many genomes from a given population were sequenced. **e**, Proportion of polymorphic SNPs as a function of minimum

MAF, for example, at MAF = 5%, we show the proportion of polymorphic SNPs with MAF  $\geq 5\%$ . **f**, Average participants cumulative sum of derived alleles including fixed derived. All lines are extended slightly beyond 100% DAF to show that they all end up at a similar level. **g**, Proportion (95% CI) of constrained genes where the most common predicted deleterious SNP contributes less than 50% of the gene burden (gene burden informative, Greenland (unadmixed) = 1.4% (1.2–1.7%); British (+CEU) = 11.7% (11.1–12.4%) and Han Chinese = 12.7% (12.0–13.3%)) and where the gene burden is dominated by a single common variant (one variant dominates, Greenland (unadmixed) = 20.0% (19.2–20.8%), British (+CEU) = 13.8% (13.1–14.5%) and Han Chinese = 12.5% (11.9–13.2%)) as illustrated by schematics above the corresponding barplots.

Canada less than 1,000 years ago, and migrated south following the west coast, and then north following the east coast<sup>19</sup>. In the past few generations, a large amount of admixture with Europeans has occurred, so the present-day Greenlandic population has on average 75% Inuit genetic ancestry and 25% European<sup>20</sup>, but with large regional differences<sup>19</sup>. The ancestral Inuit population went through a more severe population bottleneck<sup>21</sup> than any European population, including Icelanders and Finns<sup>22,23</sup>, and are genetically distant from other populations<sup>24</sup>. Additionally, the traditional Greenlandic diet, rich in fat and protein, has led to natural selection in a genomic region on chromosome 11 encompassing *CPT1A* and *FADS2* (refs. 25,26). Therefore, our understanding of the genetic architecture of disease, which is based mainly on Europeans, might not apply to Greenland. Thus, characterization of the Greenlandic genome can expand our understanding of genetic diseases and has a large potential for improving genetically driven healthcare in Greenland and in other Arctic populations.

Based on deep whole-genome sequencing (WGS) of 448 Greenlanders and whole-genome imputation of an additional 5,558 Greenlanders, both with extensive phenotypes, we quantify the historical population bottleneck and how it has shaped the genetic architecture of the Greenlandic population. We show the consequences of the genetic architecture, and highlight how much Greenlandic WGS data improve disease mapping, diagnosis and prediction, and show that the most impactful variants are Arctic-specific. These results can be leveraged to improve healthcare both across different regions in Greenland and

in Arctic populations in general. For a non-peer reviewed Greenlandic summary of the article, see Supplementary Text.

## Genetic architecture in Greenland

To characterize the genetic architecture of the Greenlandic Inuit, we generated WGS data for 448 participants (average sequencing depth of 35 $\times$ ). Using these data, we imputed the genomes of an additional 5,548 participants who were genotyped using genome-wide SNP arrays (MEGA-chip, Illumina), in total comprising 14% of the adult population in Greenland in 2023<sup>27</sup>. The participants were sampled from representative towns and settlements across the entire country (Fig. 1a; Methods). Consistent with previous studies, most participants were admixed with both Inuit and European genetic ancestry (Fig. 1a). For this reason, all analyses presented below are performed on (1) ‘Greenlandic’, which is all participants; (2) ‘Greenlandic (unadmixed)’, which is the subset of participants with only Inuit genetic ancestry; or (3) ‘Greenlandic (masked)’, which is all participants excluding any genotype with European ancestry.

We first quantified the population bottleneck using an estimated ancestral recombination graph (ARG) from masked Greenlandic Inuit WGS data. This indicated an effective population size ( $N_e$ ) of less than 300 at the time of migration from Siberia to Greenland approximately 1,000 years ago (Fig. 1a). In the WGS data, we identified 496,963 new variants and an additional 232,525 variants present in dbSNP, but not in gnomAD (Fig. 1b). Importantly, 15,766 of these new variants were

very common ( $AF > 5\%$ ) in Greenland. A further 31,046 common variants were present in dbSNP, but not in gnomAD, and 96,032 common variants were found to be rare ( $AF < 0.1\%$ ) in gnomAD. This is both an unusually high number of new common variants and an unusually low number of new rare variants. For comparison, in the same number of people from the 1000 Genomes Project (1KG) with Native American genetic ancestry, we found almost three times more SNPs that were not found in any African, European or Asian gnomAD populations than in the Greenlandic whole genomes (Americas, 2,893,077; Greenland, 1,069,501; Fig. 1c). However, only 0.4% of those SNPs were common in the Americas, whereas 21.7% were common in Greenland ( $AF > 1\%$ ; Americas, 12,814; Greenland, 231,594; Fig. 1c). The lower number of variants in Greenland is also seen in Fig. 1d, where each additional sequenced Greenlandic Inuit (Greenlandic masked) participant adds markedly fewer new SNPs than an added participant in any of the other populations (Extended Data Fig. 1).

The allele frequency distribution, visualized by the proportion of SNPs in 190 people above a given minor allele frequency (MAF), showed a very distinct distribution in the Greenlandic Inuit population compared with the Nigerian (YRI + ESN), British (GBR + CEU) and Han Chinese (CHB + CHS) populations from 1KG (Fig. 1e). Approximately 74% of all SNPs in the Greenlandic (masked) Inuit were common, with a  $MAF \geq 5\%$ , compared with 37–43% in the other populations. A similar difference is seen compared with all other 1KG populations, including the Finnish (Extended Data Fig. 2). The pattern was the same across functional categories (Extended Data Fig. 3), but with a reduced proportion of common alleles for missense and predicted high-confidence loss-of-function (LoF) (pLoF (HC)) SNPs in all populations. This indicates that, while genetic drift has been the dominating factor in shaping the genetic architecture in Greenland, negative selection has also had an impact on deleterious alleles. Notably, despite different distribution of allele frequencies, the total sum of derived alleles is similar across all populations (Fig. 1f), indicating that it is only the distribution of alleles, including deleterious ones, that is different and not the amount.

In line with the relatively high proportion of common SNPs in Greenlandic (masked) Inuit, we found that, for unadmixed Greenlanders, a large proportion of constrained genes with a putative high gene burden is dominated by a single common SNP. For example, if we estimate gene burden as the number of people that carry at least one predicted deleterious allele (predicted missense or pLoF SNPs) then, in 20.0% of the constrained genes, the most common predicted deleterious SNP contributed more than 90% to the gene burden (Fig. 1g and Extended Data Fig. 4). For comparison, this was the case for only 13.8% and 12.5% of the constrained genes in British (+CEU) and Han Chinese 1KG samples, respectively. Hence, for most genes, rare variants contribute less to diseases in Greenland compared with in Europe and East Asia and, as a consequence, gene burden tests will have limited power in Greenland.

## Disease mapping and prediction

### Clinical screening of monogenic diseases

A potential large source of inequity in genetic healthcare is the ability to diagnose monogenic diseases based on genetic screening. To quantify this inequity, we analysed WGS data from Greenland and the Nigerian, British and Han Chinese samples from 1KG and compared the number of non-causal pLoF variants each person carries after filtering away variants with  $MAF > 0.1\%$  in any gnomAD population. In a clinical setting, all of the non-filtered variants are potentially causal, and the more a person has, the harder it is to diagnose the disease and the higher the risk of misdiagnosis<sup>4</sup>. We find, on average, 12 ( $12.3 \pm 0.16$  s.e.m.;  $n = 448$ ) presumably non-causal pLoF variants in the Greenlandic population compared with 4 ( $4.36 \pm 0.1$  s.e.m.;  $n = 207$ ), 3 ( $2.94 \pm 0.15$  s.e.m.;  $n = 190$ ) and 3 ( $3.37 \pm 0.15$  s.e.m.,  $n = 208$ ) in the African, European and East

Asian populations, respectively (Fig. 2a). However, when adding WGS data from the 448 Greenlandic samples, as a Greenlandic reference panel for filtering variants common in Greenland, but rare in gnomAD, the number of remaining variants is reduced to 2 ( $1.52 \pm 0.09$  s.e.m.;  $n = 448$ ). When restricting the analysis to unadmixed Greenlandic participants, the number of non-causal variants is reduced from 13 ( $13.3 \pm 0.60$  s.e.m.;  $n = 31$ ) to 1 ( $1.26 \pm 0.21$  s.e.m.;  $n = 31$ ), when adding the Greenlandic reference panel. The pattern is independent of the maximum MAF threshold used for filtering (Fig. 2b and Extended Data Fig. 5a). Note that these analyses were carried out with the same MAF threshold across all populations but, for a specific disease, the thresholds could be adjusted according to the different disease prevalence of each population (Extended Data Fig. 5b).

### Linkage disequilibrium and imputation

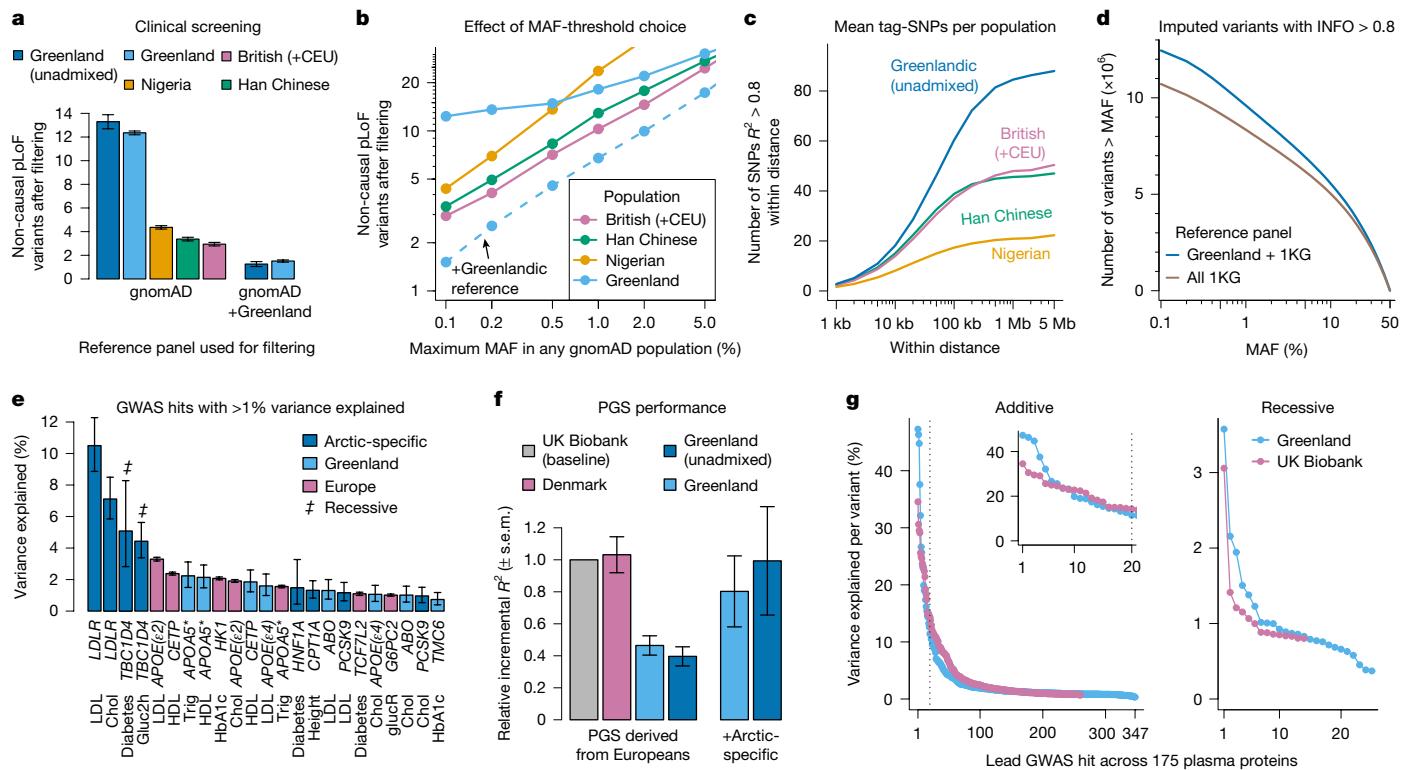
Population bottlenecks affect linkage disequilibrium (LD) and, consistent with previous findings<sup>19</sup>, we observed higher and more extended LD in Greenland compared with all 1KG populations (Supplementary Fig. 1). In the unadmixed Greenlanders, we found, on average, 88 tag-SNPs ( $R^2 > 0.8$ ) within 5 Mb of the focal SNP compared with 50, 47 and 22 for British, Han Chinese and Nigerian populations, respectively (Fig. 2c). This indicates that, although far fewer variants are needed to capture association signals across the genome, identifying the causal variant can be more difficult in the unadmixed Greenlanders. The overall pattern is consistent on most chromosomes across all 1KG populations and the unadmixed Greenlandic population, but we saw the largest number of tag-SNPs on chromosome 11, which contains the *CPT1A* and *FADS* genes (Supplementary Fig. 2).

Imputation from SNP-chip data is used widely to infer whole-genome variation before running a genome-wide association study (GWAS). This relies on matching haplotypes, which are affected by LD, and we therefore quantified the benefit of including Greenlandic data in the reference panel. We imputed the genomes of 5,548 SNP-chipped Greenlandic participants with the Greenland WGS participants added to the reference panel. This resulted in an additional 1.7 million accurately called (INFO score  $> 0.8$ ) variants, of which 0.7 million were common ( $MAF > 5\%$ ). This indicates that high quality imputation is possible only with representation of Greenlanders in the reference panel (Fig. 2d and Supplementary Fig. 3).

### GWAS and polygenic score performance

To further explore the consequences for disease mapping, we performed GWAS on 13 metabolic traits in 5,996 Greenlanders based on their imputed whole genomes (QQ plots in Supplementary Fig. 4), and compared the results with GWAS of the same 13 metabolic traits from large European cohorts. Among the 23 associations that explained more than 1% of the phenotypic variance, 7 were found in Europe and 16 were found in Greenland, of which 8 were Arctic-specific (Fig. 2e and Supplementary Table 1). For Greenlanders, the highest proportion of variance was explained by the previously reported Arctic-specific *LDLR* variant associated with low-density lipoprotein (LDL)-cholesterol, explaining 10.5% (95% confidence interval (CI), 8.9–12.3%)<sup>28,29</sup>. For Europeans, the highest amount of variance was explained by the *APOE*(ε2) variant also associated with LDL-cholesterol, explaining 3.3% (95% CI, 3.2–3.4%). The non Arctic-specific variants, and thus the variants present in both Greenland and Europe, contributed a similar amount in both populations, with the differences being driven by moderate differences in allele frequencies for *HK1*, *APOA5*, *APOE* and *ABO* (Supplementary Table 2).

We also assessed how well polygenic scores (PGS) constructed on European data performed in Greenlanders. We compared incremental  $R^2$  (incr.  $R^2$ ) coefficients on the same 13 metabolic phenotypes between non-British Europeans from the UK Biobank (Europe), admixed Greenlanders (Greenland), unadmixed Greenlanders (Inuit) and the Danish Inter99 cohort (Denmark). We included the latter because it is



**Fig. 2 | Consequences of genetic architecture for disease mapping.** **a**, Mean number of non-causal pLoF variants ( $\pm$  s.e.m.) remaining after removing variants present at MAF  $> 0.1\%$  in any population in the reference panel (gnomAD v.3.0.0). **b**, Same as **a**, but using different MAF thresholds in the reference panel. **c**, Mean number of tag-SNPs ( $R^2 > 0.8$ ) as a function of distance from focal SNP. **d**, Number of imputed variants with an INFO score greater than 0.8 and MAF above threshold given on the x axis. Imputation was performed with either the merged reference panel of Greenlandic WGS plus 1KG ( $n = 448 + 3,202$ ) or only the 1KG reference panel ( $n = 3,202$ ). **e,f**, Comparison of largest GWAS in Greenland and Europe on 13 metabolic traits. **e**, Genome-wide associations explaining more than 1% variance in the largest GWAS across 13 metabolic traits in both Europe and Greenland (95% CI). Gene names are given below bars, with phenotype associated with the

variant listed below gene name. For diabetes, we used liability-scale variance explained. Asterisk, the causal gene in this region is uncertain. Chol, total cholesterol; Gluc2h, glucose (2 h); GlucR, glucose (random); HbA1c, haemoglobin 1Ac; HDL, high-density lipoprotein; Trig, triglycerides. **f**, Mean incr.  $R^2$  ( $\pm$  s.e.m.) of European-derived PGS predicting the corresponding 13 metabolic traits normalized to UK Biobank for all Greenland participants, only unadmixed Greenlandic participants or Danish participants. The two bars to the right are the mean incr.  $R^2$  after adding the Arctic-specific variants. **g**, Variance explained for lead SNP in genome-wide significant associations on 175 plasma proteins (Olink) in Greenland and UK Biobank ordered by variance explained and grouped by the model yielding the lowest  $P$  value;  $n = 3,707$ . The inset shows a zoom-in of the first 20 GWAS hits.

comparable with the Greenlandic cohort with a similar protocol for phenotypes and similar mean participant age, which is younger than the UK Biobank (mean age Greenland, 45 years; Inter99, 46 years; UK Biobank, 64 years). Incr.  $R^2$  is the increase in  $R^2$  when including PGS in the model (Extended Data Fig. 6). To compare across traits, we normalized incr.  $R^2$  to the incr.  $R^2$  of the UK Biobank (see Extended Data Fig. 6 for unnormalized values). On average, we saw a comparable performance between Europe and Denmark (mean relative incr.  $R^2 = 1.03 \pm 0.11$ ), but 54% less variance explained in admixed Greenlanders (mean relative incr.  $R^2 = 0.46 \pm 0.06$ ), and further reduced performance in the unadmixed Greenlanders (mean relative incr.  $R^2 = 0.40 \pm 0.06$ ; Fig. 2f). Simply adding the Arctic-specific variants explaining more than 1% variance from the GWAS, along with the previously identified Arctic variant in *S100* (Supplementary Table 3) with the PGS, markedly improved the overall performance of the PGS in both the admixed (mean relative incr.  $R^2 = 0.80 \pm 0.22$ ) and unadmixed Greenlanders (mean relative incr.  $R^2 = 0.99 \pm 0.34$ ). For the traits affected most by the few added Arctic-specific variants, for example, LDL-cholesterol, total cholesterol, 2-hour glucose and type 2 diabetes, the PGS in unadmixed Greenlanders outperforms that of the UK Biobank baseline after adding these variants (Extended Data Fig. 6 and Supplementary Table 4). In the most extreme case, it improves the prediction more than fourfold compared with UK Biobank (type 2 diabetes, UK Biobank incr.  $R^2 = 0.014$  (95% CI, 0.013–0.015) and unadmixed Greenlanders incr.  $R^2 = 0.066$  (95% CI, 0.055–0.078)).

## Matched GWAS on 175 plasma proteins

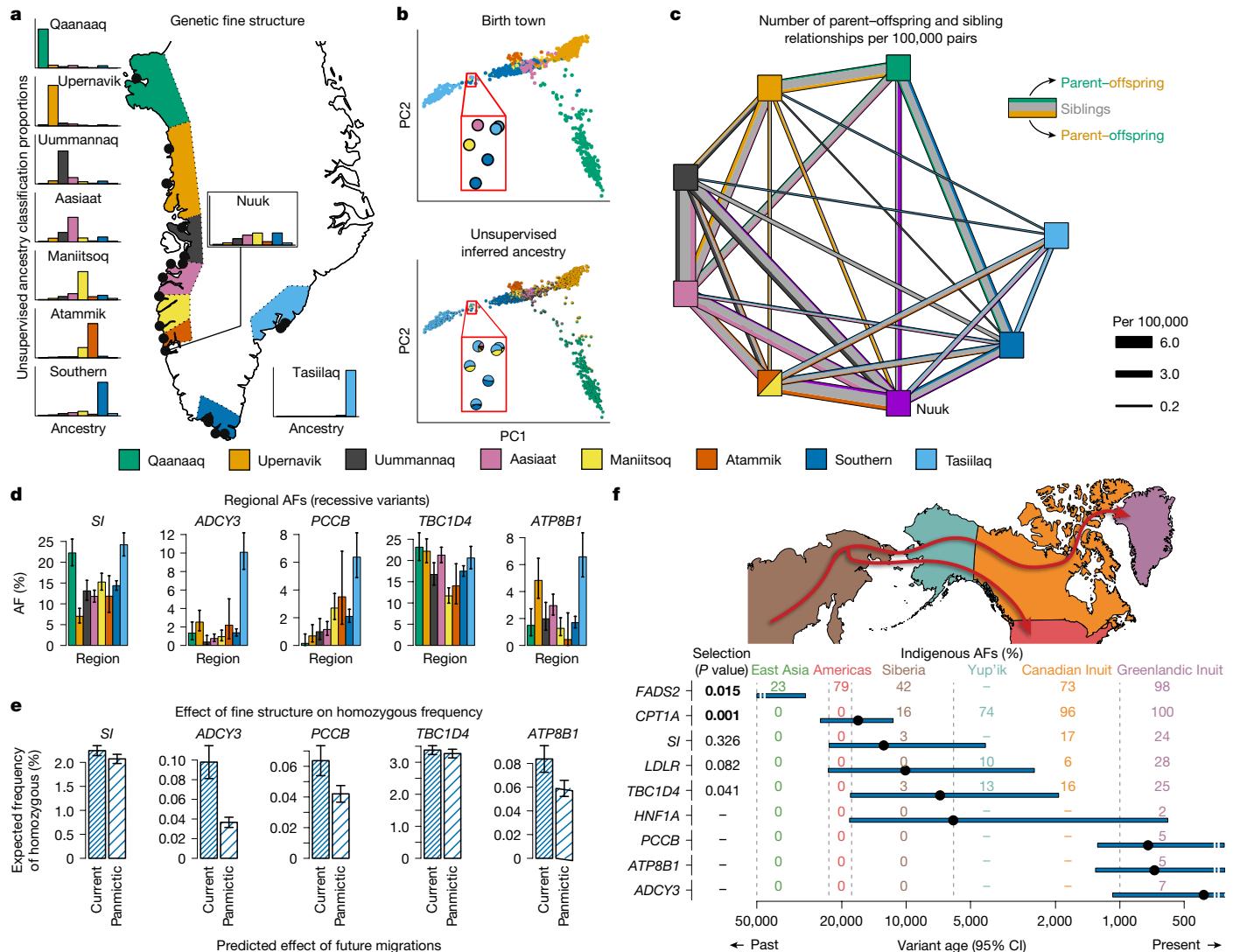
To replicate the association patterns observed in the GWAS of the 13 metabolic traits, we performed a sample-size-matched GWAS in UK Biobank and compared it with GWAS results from Greenland ( $n = 3,707$  in both populations) on amounts of 175 plasma proteins. There were 368 independent genome-wide significant ( $P < 5 \times 10^{-8}$ ) signals in the Greenlandic cohort, whereas we identified 271 signals in the UK Biobank (Supplementary Tables 5 and 6). Signals were classified as mainly either additive or recessive, depending on which model showed the lowest  $P$  value. We observed 343 mainly additive and 25 mainly recessive signals in Greenland, compared with 257 mainly additive and 14 mainly recessive signals in the UK Biobank. As expected, we observed the highest impact variants in Greenland. For the additive model, the three variants explaining most variance in Greenlanders explained 47.3%, 46.3% and 44.8% compared with UK Biobank, where the top three variants explained 34.5%, 30.6% and 29.5% (Fig. 2g). It should be noted that if the sample sizes were larger, we would expect to find more low impact variants to affect disease burden in Europe than in Greenland.

## Fine structure and disease prevalence

### Limited historical migration

Even though the Inuit have been living in Greenland for less than 1,000 years, there are large regional differences in disease prevalence<sup>32</sup>.

# Article



**Fig. 3 | Genetic fine structure.** **a**, Unsupervised genetic clustering grouped by region. Mean estimated ancestry ( $K=8$ ) proportions for all samples in each region are shown in barplots. Map of Greenland with regions coloured (exaggerated inland for visibility) and sample locations indicated with black dots. **b**, People with birth and sample location in the same region ( $n=1,921$ ) visualized on the first two principal components coloured either by birth town region or inferred ancestry proportions. The enlarged areas highlight the ancestry assignment for people between clusters. **c**, Number of parent-offspring and full sibling relations inferred from genetic data per 100,000 possible relationships between each pair of regions. Grey lines represent sibling relationships, with line width indicating the inferred number of sibling relationships per 100,000 possible relationship pairs. Coloured lines represent parent-offspring relationships, where the colour indicates from which region the parent

was sampled and the line width indicates the inferred number of sibling relationships per 100,000 possible relationship pairs. **d**, Regional differences in AFs for five highly penetrant Arctic-specific recessive variants (coefficient of variation ( $AF^2$ ), *SI* = 74%, *ADCY3* = 251%, *PCCB* = 164%, *TBC1D4* = 42% and *ATP8B1* = 149%). **e**, Expected frequency of homozygous participants for each variant with and without the current fine structure. **f**, Estimated variant age along with 95% CI of the eight Arctic-specific variants and the variant in *FADS2*. Coloured percentages are Indigenous allele frequencies of the different regions. Vertical dashed lines are split times between the populations<sup>39</sup>. Map shows a schematic illustration of the migration routes for the ancestral population that gave rise to the Greenlandic Inuit. *P* values for directional selection are unadjusted and bold *P* values indicate significance after FDR(BH), see also Supplementary Table 9.

To explore whether this could be explained by genetic fine structure within Greenland, we used the unsupervised neural network framework, HaploNet<sup>33</sup>, to first infer and mask the European genetic ancestry component on a local haplotype level, and then estimate the fine structure in the ancestral Inuit haplotypes. We saw a pronounced genetic fine structure corresponding well with geographical location (Fig. 3a and Extended Data Fig. 7). Assuming eight ancestral components, we observed almost exclusively one ancestry in each of Qaanaaq, Upernivik, Southern and Tasiilaq. By contrast, the capital, Nuuk, has a diverse mix of genetic ancestries. To evaluate how well the inferred ancestry components correspond to geographical location, we visualized participants with more than 50% Inuit genetic ancestry who had matching birth town and sample location (Fig. 3b and Extended Data Fig. 8). On a

principal component analysis plot, the unsupervised inferred ancestry matched almost perfectly with the reported birth town for people with one predominant ancestry. By contrast, people with ancestry from several regions were positioned in-between the clusters, and the reported birth town was, as expected, from one of the regions of their inferred ancestry (Fig. 3b). The fact that the inferred ancestry clusters fits so well with birth locations indicates that migration between the inferred regions in Greenland has historically been very limited before the last two to three generations. This is consistent with historical records<sup>34,35</sup> showing that the population, up until the 1960s, stayed in small permanent settlement groups, where generations of deep local knowledge allowed for efficient hunting and secured their livelihood.

## Recent changes in mobility

In the first quarter of the twentieth century, an urbanization process, including population concentration in towns, started. After Greenland was formally integrated into the Kingdom of Denmark in 1953, the process was boosted, resulting in closure of small settlements, further population concentration in towns and increased mobility between regions<sup>35,36</sup>. Consistently, we find a large amount of mobility in the last few generations; 26% of participants reported that they were born in a different region than they were living in while taking part in the population health surveys (on average 43 years old). We further used the fact that many participants have close relatives in the cohort with 7.2 parent–offspring and 13.4 full sibling relationships per 100,000 pairs of participants, which are high rates compared to UK Biobank (0.0035 parent–offspring and 0.0178 full sibling per 100,000 pairs of participants<sup>37</sup>). The connectivity between regions, measured as the number of close relative pairs per 100,000 pairs of participants, differed quite a bit, and geographically adjacent regions were better connected than regions further apart (Fig. 3c), with the exception of Nuuk, which was well connected across longer distances. Taken together, the strong genetic clustering indicates that the regions have been very isolated historically but, within the last few generations, we observe a large amount of mobility and the regions have become very connected, especially Nuuk, indicating that the population is becoming more panmictic.

## Predicted consequences for disease

The fine structure affects disease prevalence in Greenland but the recent increase in mobility will affect prevalence in the future. We illustrate this on the highly penetrant (odds ratio > 8) Arctic-specific recessive variants in the genes *ADCY3* (obesity), *TBC1D4* (diabetes), *SI* (CSID), *ATP8B1* (Cholestasis Familiaris Groenlandica (CFG)) and *PCCB*<sup>16</sup> (propionic aciduria). Some of the variants showed much bigger regional differences than others (Fig. 3d), with similar results when masking European ancestry (Supplementary Fig. 5). The fine structure explained previous observations, where CFG prevalence is higher in East Greenland<sup>32</sup>. The variants in *ATP8B1* and *PCCB* are often lethal in homozygous form and, currently, all pregnant women in Greenland are offered genetic screening for these variants<sup>38</sup>. We estimated the expected frequency of homozygous carriers both with the current observed fine structure and assuming panmixia (Fig. 3e). The panmixia scenario represents a predicted near-future scenario without fine structure. As expected, the variants in *ADCY3*, *ATP8B1* and *PCCB*, with pronounced differences in regional allele frequencies indicated by coefficient of variation (Fig. 3d), were predicted to undergo a large reduction in homozygous carriers, and thus disease prevalence, even in the absence of negative selection and external gene flow. By contrast, we do not expect a similar future reduction in CSID or diabetes caused by the *TBC1D4* variant.

## Arctic-specific variants and selection

The regional allele frequency differences indicate that some of the variants have arisen recently. To show this, and to test whether the high-impact variants denoted to be Arctic-specific in Fig. 2e are truly Arctic-specific, we estimated whether the ages of the variants are younger than the split from Native Americans (approximately 18,000 years ago<sup>39</sup>) using ARGs estimated from the Greenlandic WGS data. We did not include the *PCSK9* variant because previous studies indicate that we do not know the causal variant<sup>40</sup>. However, we did include the variant in *FADS2* as it has been indicated previously to be under selection in the Arctic and because it serves as a control as it is a variant that we know is not Arctic-specific despite being almost fixed in Greenland.

The estimated variant ages indicate that all variants are Arctic-specific, except for *FADS2* (Fig. 3f and Supplementary Table 7). The *ADCY3*, *ATP8B1* and *PCCB* variants were the youngest and were all

estimated to be less than 1,000 years old, which indicates that they arose during migrations within Greenland. This also explains the large regional allele frequency differences. The other variants were older (older than 1,000 years) and should thus be shared with Inuit populations in North America and for the oldest (older than 6,000 years) also with Siberians<sup>39</sup>. This is corroborated by the presence or absence of the variants in a range of contemporary populations (Fig. 3f) with available allele frequency information. We can use the estimated variant ages to identify geographical regions in which it would be relevant to look at specific variants to inform healthcare decisions. For example, the results indicate the presence of the *TBC1D4* variant in Siberia and Alaskan Yup'ik, which means that there is opportunity for improvement of diagnosis and treatment of diabetes<sup>12,41</sup> in these regions. Also, the *SI* variant, which is suggested to be present in Alaskan Yup'ik, could be important for treatment of severe gastrointestinal problems in infants<sup>30,42</sup> in this population. That the variants are truly Arctic-specific is corroborated by their absence in the 1KG worldwide populations, with the exception of one Japanese person carrying one copy of the *TBC1D4* variant (Supplementary Table 8). However, a comparison of the haplotypes of the Greenlandic carriers and the Japanese carrier showed completely different haplotypes, indicating that it is a recurrent mutation (Extended Data Fig. 9). Some of the suggested Arctic-specific variants were found in the newest unpublished version of gnomAD study (v.4.1.0) at ultra-low frequencies (less than 0.06%), mainly in the unassigned ancestry category 'Remaining' (Supplementary Table 8). Unlike 1KG, the gnomAD ancestry assignments are based on PC cutoffs and are not well studied. Thus, we interpret the presence of these variants as unaccounted Arctic ancestry among the 800,000 participants.

Given that almost all the variants are Arctic-specific, and thus quite young, it is surprising that they are found at such high frequencies. Strong evidence for positive selection for *CPT1A*<sup>43</sup> and *FADS2* (ref. 25) has been shown previously and suggestive evidence has been presented for *TBC1D4* (ref. 13). The previous studies in Inuit relied mainly on allele frequency differences from other populations<sup>25,26</sup>. However, with the WGS data, we could test for selection by estimating the frequency trajectory using CLUES<sup>44</sup>. We excluded variants with a derived allele frequency (DAF) < 10%, resulting in five remaining variants to test. For these 5 variants, we matched with 999 other variants with DAF within ±10% of the variant to obtain empirical *P* values (Supplementary Table 9). The results confirmed strong evidence for additive selection ( $s_{add}$ ) on the *CPT1A* variant ( $P_{empirical} = 0.001$ , false discovery rate (Benjamini-Hochberg) (FDR(BH)) = 0.005,  $s_{add} = 0.016$ ). Moreover, we found moderate evidence for selection on the *FADS2* variant ( $P_{empirical} = 0.015$ , FDR(BH) = 0.038,  $s_{add} = 0.006$ ) and weak evidence for recessive selection ( $s_{rec}$ ) on the *TBC1D4* variant ( $P_{empirical} = 0.041$ , FDR(BH) = 0.068,  $s_{rec} = 0.077$ ). We found no evidence for selection on the *LDLR* variant ( $P_{empirical} = 0.082$ , FDR(BH) = 0.103), or for recessive selection on the *SI* variant ( $P_{empirical} = 0.326$ , FDR(BH) = 0.326). Posterior mean of allele frequencies for the 5 variants and 20 of the random DAF-matched variants can be found in Extended Data Fig. 10. These results independently confirm that selection has had an impact on the genetic architecture in Greenland and in the Arctic.

## Discussion

This work demonstrates that including indigenous populations in genetic research can help alleviate the inequalities in genomic healthcare. We showed that, despite having a comparable sum of derived alleles, the frequency distribution of variants in Greenland is different to other populations as a consequence of having gone through an extensive population bottleneck. This means that, although we do not expect the overall genetic disease burden to be different compared with other populations<sup>21</sup>, there are differences for specific diseases, and for complex diseases we expect a larger contribution from high-impact variants. We show that many common variants in

Greenland are Arctic-specific, and thus can be studied only with inclusion of Arctic populations in research. Moreover, we show that owing to lack of representation in reference panels, clinical screenings and risk prediction for Inuit are currently difficult, but can be improved easily by including WGS data from a moderate number of Greenlanders. For disease mapping, we show a limited potential for gene-based association tests in Greenland, but an increased potential for moderately sized GWAS owing to the increased number of proxy-SNPs and the presence of higher impact variants for both metabolic traits and protein abundance. These variants, like the previously reported *TBC1D4*, *LDLR* and *SI* variants, not only have large effects, but are also common. This means that these variants are easier to study further than rare or low effect variants and provide great potential for genetic-based early prevention and treatment at population level. Our analyses of genetic architecture were limited to SNPs and small insertion/deletions identified through short-read sequencing, but we expect the genetic architecture to affect other inherited variants similarly.

We detected pronounced genetic fine structure in Greenland and were able to perform unsupervised clustering that matched almost perfectly with the region of birth, and thus indicate historically very limited migration between the regions. This structure explains some of the differences in disease prevalence in Greenland. The population becoming more panmictic will lead to a decrease in prevalence for certain recessive diseases in the next couple of generations. By estimating the variant ages we found that many are Arctic-specific, including all of the high-impact variants, which was consistent with the observed presence and absence in contemporary populations. Finally, consistent with previous studies, we found strong evidence for positive selection on *CPT1A* and *FADS2*, and weak evidence for *TBC1D4*.

All these results indicate that the genetic architecture of diseases in the Greenlandic population is shaped by its demographic history, fine structure and natural selection. The results highlight the potential benefits for the Greenlandic and other Arctic populations of being represented in genetic research and we hope that these results can contribute to a more globally inclusive and equitable research in genomic medicine.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-024-08516-4>.

- Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
- Fatumo, S. et al. A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
- Hindorf, L. A. et al. Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2017).
- Manrai, A. K. et al. Genetic misdiagnoses and the potential for health disparities. *N. Engl. J. Med.* **375**, 655–665 (2016).
- Patel, A. P. et al. A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat. Med.* **29**, 1793–1803 (2023).
- Choudhury, A. et al. High-depth African genomes inform human migration and health. *Nature* **586**, 741–748 (2020).
- Wonkam, A. Sequence three million genomes across Africa. *Nature* **590**, 209–211 (2021).
- Sohail, M. et al. Mexican Biobank advances population and medical genomics of diverse ancestries. *Nature* **622**, 775–783 (2023).
- Ziyatdinov, A. et al. Genotyping, sequencing and analysis of 140,000 adults from Mexico City. *Nature* **622**, 784–793 (2023).
- Silcock, M. et al. Indigenous Australian genomes show deep structure and rich novel variation. *Nature* **624**, 593–601 (2023).
- Reis, A. L. M. et al. The landscape of genomic structural variation in Indigenous Australians. *Nature* **624**, 602–610 (2023).
- Manousaki, D. et al. Toward precision medicine: *TBC1D4* disruption is common among the Inuit and leads to underdiagnosis of type 2 diabetes. *Diabetes Care* **39**, 1889–1895 (2016).
- Moltke, I. et al. A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193 (2014).
- Thuesen, A. C. B. et al. A novel splice-affecting *HNF1A* variant with large population impact on diabetes in Greenland. *Lancet Reg. Health Eur.* **24**, 100529 (2023).

- Gruarup, N. et al. Loss-of-function variants in *ADCY3* increase risk of obesity and type 2 diabetes. *Nat. Genet.* **50**, 172–174 (2018).
- Ravn, K. et al. High incidence of propionic acidemia in Greenland is due to a prevalent mutation, 1540insCCC, in the gene for the beta-subunit of propionyl CoA carboxylase. *Am. J. Hum. Genet.* **67**, 203–206 (2000).
- Klomp, L. W. et al. A missense mutation in *FIC1* is associated with Greenland familial cholestasis. *Hepatology* **32**, 1337–1341 (2000).
- Andersen, M. K. et al. Loss of sucrase-isomaltase function increases acetate levels and improves metabolic health in Greenlandic cohorts. *Gastroenterology* **162**, 1171–1182.e3 (2021).
- Moltke, I. et al. Uncovering the genetic history of the present-day Greenlandic population. *Am. J. Hum. Genet.* **96**, 54–69 (2015).
- Waples, R. K. et al. The genetic history of Greenlandic–European contact. *Curr. Biol.* **31**, 2214–2219.e4 (2021).
- Pedersen, C.-E. T. et al. The effect of an extreme and prolonged population bottleneck on patterns of deleterious variation: insights from the Greenlandic Inuit. *Genetics* **205**, 787–801 (2017).
- Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
- Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
- Rasmussen, M. et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010).
- Fumagalli, M. et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343–1347 (2015).
- Skotte, L. et al. Missense mutation associated with fatty acid metabolism and reduced height in Greenlanders. *Circ. Cardiovasc. Genet.* **10**, e001618 (2017).
- Statbank Greenland. Population January 1st by place of birth, gender, age and time. <https://bank.stat.gl/443/sq/46b0d929-17fa-4929-861f-fe6cbdb46727> (2024).
- Jørsbø, E. et al. An *LDLR* missense variant poses high risk of familial hypercholesterolemia in 30% of Greenlanders and offers potential for early cardiovascular disease intervention. *HGG Adv.* **3**, 100118 (2022).
- Dubé, J. B. et al. Common low-density lipoprotein receptor p.G116S variant has a large effect on plasma low-density lipoprotein cholesterol in circumpolar Inuit populations. *Circ. Cardiovasc. Genet.* **8**, 100–105 (2015).
- Marcadier, J. L. et al. Congenital sucrase-isomaltase deficiency: identification of a common Inuit founder mutation. *CMAJ* **187**, 102–107 (2015).
- Stinson, S. E. et al. Genetic regulation of the plasma proteome and its link to cardiometabolic disease in Greenlandic Inuit. Preprint at medRxiv <https://doi.org/10.1101/2024.07.03.24309577> (2024).
- Eiberg, H., Nørregaard-Pedersen, B. & Nielsen, I. M. Cholestasis Familiaris Groenlandica/Björn-like disease in Greenland—a population study. *Int. J. Circumpolar Health* **63**, 189–191 (2004).
- Meisner, J. & Albrechtsen, A. Haplotype and population structure inference using neural networks in whole-genome sequencing data. *Genome Res.* **32**, 1542–1552 (2022).
- Frandsen, N. H. et al. *Danmark og kolonierne: Grønland* (Gads Forlag, 2017).
- Kjær Sørensen, A. *Danmark-Greenland in the Twentieth Century*, Vol. 341 (Museum Tusculanum Press, 2007).
- Bjerregaard, P. & Viskum Lytken Larsen, C. Health aspects of colonization and the post-colonial period in Greenland 1721 to 2014. *J. North. Stud.* **10**, 85–106 (2017).
- Hofmeister, R. J. et al. Parent-of-origin inference for biobanks. *Nat. Commun.* **13**, 6668 (2022).
- Nielsen, I. M., Kern, P. & Eiberg, H. From research to prevention in Greenland. *Greenland Medical Society. Ugeskr. Laeger* **169**, 1105 (2007).
- Flegontov, P. et al. Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America. *Nature* **570**, 236–240 (2019).
- Senftleber, N. K. et al. GWAS of lipids in Greenlanders finds association signals shared with Europeans and reveals an independent PCSK9 association signal. *Eur. J. Hum. Genet.* **32**, 215–223 (2024).
- Schnurr, T. M. et al. Physical activity attenuates postprandial hyperglycaemia in homozygous *TBC1D4* loss-of-function mutation carriers. *Diabetologia* **64**, 1795–1804 (2021).
- Treem, W. R. Congenital sucrase-isomaltase deficiency. *J. Pediatr. Gastroenterol. Nutr.* **21**, 1–14 (1995).
- Clemente, F. J. et al. A selective sweep on a deleterious mutation in *CPT1A* in Arctic populations. *Am. J. Hum. Genet.* **95**, 584–589 (2014).
- Stern, A. J., Wilton, P. R. & Nielsen, R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genet.* **15**, e1008384 (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

## Methods

### Inclusion and ethics statement

The study was approved by the Scientific Ethics Committee in Greenland (project 505-42, 505-95, project 2011–13 (ref. no. 2011–056978), project 2014–08 (ref. no. 2014–098017), project 2017–5582, project 2015–22 (ref. no. 2015–16426) and project 2021–09) and was conducted in accordance with the Declaration of Helsinki, second revision. All participants gave written informed consent.

User studies were conducted as public meetings, and in shared circles in relation to the population studies in Greenland and in relation to genotype-based intervention studies in CSID and among carriers of the *TBC1D4* variant<sup>45,46</sup>. In brief, the conclusions from these user studies indicate that research on chronic diseases is considered important and necessary. This includes understanding how heredity and lifestyle interact. There is a demand for research on the significance of traditional lifestyle for health, including the importance of traditional practices in treatment. Additionally, it is crucial to explore areas where no treatment currently exists. Research serves as a prerequisite for discovering new methods of prevention and treatment, and explanations can help reduce anxiety associated with diseases and symptoms. Finally, there is broad support for investigating hereditary causes of health and disease. Users express that identifying known genetic causes of illness can enable early prevention and treatment, especially for family members. Lack of representation is considered a significant issue.

### Participants and representative sampling

We included Greenlanders living in Greenland from the population health surveys B99 (ref. 47) (recruited 1999–2001;  $n = 1,317$ ), Inuit Health in Transition<sup>48</sup> (IHIT; recruited 2005–2010;  $n = 3,108$ ), and B2018 (refs. 49,50) (recruited 2017–2019;  $n = 2,539$ ). We also included a cohort of Greenlanders living in Denmark, collected as part of the B99 survey (BBH;  $n = 739$ ). Some participants were part of several surveys, resulting in a total of 5,996 unique participants across the different cohorts. As previously described<sup>47–49</sup>, participants were recruited in representative towns and settlements of each region in Greenland where random samples were drawn from the central population registers.

### Genetic data

The WGS data were handled as described in ref. 14, where it was used for screening of new variants in 14 genes related to maturity-onset diabetes in the young. In brief, the 448 WGS samples were sequenced using Illumina 150 paired-end sequencing and had an average sequencing depth of 35×. After adaptor trimming, reads were mapped with BWA-MEM to GRCh38, and genotype calling was carried out with GATK haplotype caller. Only variants in the T98 tranche from variant quality score recalibration were used. Moreover, we used genotype data on an additional 5,548 people generated from the Multi-Ethnic Global Array (MEGA-chip, Illumina); of these, 4,182 were described previously in ref. 20 (deposited with European Genome-phenome Archive, accession no. EGAD00010002057) and the remaining 1,366 were part of the newer B2018 study. The two datasets were merged and, after variant quality control, the data consisted of 1.6 million autosomal variants.

### Phasing and imputation

To prepare a good reference panel for the downstream imputation analyses, we phased the WGS data using Shapeit2 (ref. 51) with trio and duo information. We call this the Greenlandic WGS panel and it includes all sites found to be polymorphic in the Greenlandic WGS data. Moreover, we prepared another reference panel consisting of all participants in the 1KG Project, which we called the 1KG panel. For this panel we included all overlapping sites with the Greenlandic panel and all sites with a MAF > 1% in the populations with East Asian and European ancestry (CDX, CEU, CHB, CHS, GBR, TSI and IBS<sup>52</sup>). We then imputed the MEGA-chip data with IMPUTE2 (refs. 51,53) with option

‘merge\_ref\_panels’, which can leverage both prepared reference panels to improve imputation performance. Finally, we created two datasets with and without phasing information by merging the imputed (and phased) MEGA data with the phased WGS data, which resulted in 5,996 participants. For downstream analyses, we did quality control on variants with respect to genotype missingness and allele frequency, as noted in each analysis. Additionally, to investigate the improvement of imputation accuracy using the Greenlandic WGS panel, we also ran IMPUTE v.2 with only the 1KG reference panel for comparison and showed only imputation performance on the overlapping sites.

### Variant annotation

All variants were annotated using VEP<sup>54</sup> with the following additional custom annotations: dbSNP (build 155), gnomAD v.3.0.0 allele frequencies and genome coverage (gnomADOver15: proportion of gnomAD participants with depth > 15, gnomADOver50: proportion of gnomAD participants with depth > 50, gnomAD filter: the gnomAD filter column being either PASS, AC0, AS\_VQSR, InbreedingCoeff or NA if not polymorphic in gnomAD), 1KG frequencies (v.20201028\_3202\_phased)<sup>52</sup>, ancestral allele from the Ensembl *Homo sapiens* ancestral sequence, and Loss-Of-Function Transcript Effect Estimator<sup>55</sup>. We did not use GnomAD v.4 because it did not have information about depth and filters for non-called sites.

To find new variants, we kept only variants with good coverage in gnomAD (gnomADOver15 > 80%, gnomADOver50 < 10% and gnomAD filter = PASS or NA). Moreover, we did not allow for spanning deletions (variant call format (VCF) asterisk allele), variants with missingness greater than 20% or several variants 1 bp apart in the Greenlandic WGS data. All these filters were used to minimize the number of false-positive variants. For plotting, variants were polarized according to the minor allele in gnomAD African populations.

For comparing the number of SNPs not found in other populations we used only SNPs with good coverage in gnomAD (as defined above) and found the maximum allele frequency in any of the following gnomAD populations: AFR, NFE, FIN, SAS, EAS and ASJ, covering African, European and Asian populations. SNPs with a maximum gnomAD frequency more than 0.01%, were excluded and the remaining SNPs were counted as SNPs not in Africa, Europe or Asia for both the 448 admixed Greenlandic WGS samples and a random sample of 448 samples from the Americas in the 1KG PEL, CLM, MXL and PUR populations.

### Measurements and assays

Height, weight, systolic blood pressure, diastolic blood pressure, and hip and waist circumference were measured, and body mass index and waist-hip ratio were calculated. All IHIT participants above 18 years, B99 participants above 35 years and a subset of B2018 participants underwent an oral glucose tolerance test, where blood samples were drawn after an overnight fast of at least 8 h, and at 30 min (only for B2018) and 2 h after receiving 75 g glucose. Plasma glucose was measured at fasting, 30 min and 2 h, and haemoglobin 1Ac at fasting, as previously described<sup>56</sup>. Concentrations of serum cholesterol, high-density lipoprotein cholesterol and triglycerides were measured, and LDL-cholesterol calculated. Type 2 diabetes was defined based on the World Health Organization 1999 criteria<sup>57</sup> and controls were defined as normal glucose tolerant based on the oral glucose tolerance test data.

The Olink protein data for the Greenlandic participants used for protein quantitative trait loci analysis are from ref. 31. Using the Olink Target 96 Inflammation and Cardiovascular II panels, relative plasma levels of 184 proteins were measured in 3,732 participants across the population surveys. The 2 batches were bridged and normalized based on 16 control samples using the OlinkAnalyze R package (<https://cran.r-project.org/web/packages/OlinkAnalyze/index.html>). Normalized protein expression values on a log<sub>2</sub> scale were inverse-rank normalized, including normalized protein expression data below the limit of detection. Samples with a quality control warning were excluded.

# Article

The Danish data used in assessment of polygenic scores consist of 6,182 people from the Danish population-based Inter99 cohort, where metabolic phenotypes were measured as described previously<sup>58</sup>, and genotyping was done using the Infinium OmniExpress-24 v.1.3 Chip (Illumina Inc.). Genotypes were called using GenCall in GenomeStudio (v.2011.1; Illumina) and quality control was performed according to a standard procedure as previously described<sup>59</sup>. Data were imputed using the Haplotype Reference Consortium reference panel on the Michigan imputation server.

## Admixture, haplotype masking and fine structure

Inuit and European admixture proportions were calculated using the software ADMIXTURE<sup>60</sup> on a subset of variants with MAF > 5%, missingness less than 1% and LD-pruned within 1 Mb removing variants with  $R^2 > 0.8$  using Plink v.1.9.0 (ref. 61).

For fine structure analysis of the Inuit ancestry, we used the neural network framework, HaploNet<sup>33</sup>, on the phased and imputed data of all 5,996 Greenlandic participants on SNPs with MAF > 5%. First, we performed window-based haplotype clustering using a Gaussian mixture variational autoencoder. A window size of 1,024 SNPs was used to generate haplotype cluster likelihoods for all samples, which we leveraged to infer fine-scale population structure through both ancestry estimation and principal component analysis. We performed unsupervised ancestry estimation allowing for two ancestral sources ( $K = 2$ ) and ran it with several seeds to ensure that the expectation maximization algorithm of HaploNet had converged. The convergence criterion was defined as having two runs within five log-likelihood units of the best seed. The two ancestry sources were assumed to reflect Inuit and European ancestry.

Next, we used HaploNet Fatash to infer the local ancestry of the haplotypes. Fatash obtains posterior probabilities of the local ancestry per genomic window per haplotype based on haplotype likelihoods and genome-wide admixture proportion estimates obtained from HaploNet train and HaploNet admix. The model is based on a hidden Markov model with instantaneous rate change. Fatash is implemented in Python/Cython as a submodule in the HaploNet software suite. Local ancestry tracts were inferred using three different window sizes (1,024; 512 and 256 variants) to increase accuracy proximate to recombination events. We used only the smaller windows (512 or 256 variants) if the fit was more than 50 log-likelihood units better to balance overfitting and detection of true recombination events.

Finally, we masked the haplotype clusters in genomic windows for each haplotype with posterior probability less than 0.95 for Inuit ancestry based on the local ancestry inferred by Fatash. After genomic masking, we excluded people with less than 0.90 missingness and genomic windows with haplotype frequencies less than 0.01 after masking from downstream analyses. The masked dataset was used for fine structure analyses of the Inuit ancestry using HaploNet PCA and HaploNet Admix. For HaploNet Admix, we used the same convergence criterion as described above and modelled from two to eight ancestry sources. The admixture model with nine ancestry sources did not fulfil our convergence criteria.

## Local ancestry and homozygous ancestry masking of WGS data

From the imputed and phased data, we first excluded variants with missingness less than 1% and MAF < 5%. Then, we inferred local ancestry for the admixed participants using RFmix v.2 with unadmixed Greenlandic Inuit ( $N = 99$ ) and participants with European genetic ancestry ( $N = 313$ , CEU, IBS and TSI) from the 1KG project as reference populations. This gives us inference of local ancestry tracts for each person. To analyse the genetic architecture of the Greenlandic Inuit, we constructed a masked WGS dataset, using vcffppR<sup>62</sup>, where regions with any European local ancestry were excluded resulting in an average of 240 (minimum, 201; maximum, 285) Greenlandic Inuit participants at each site. To create the masked dataset, for each person we only keep sites in a local ancestry region inferred to be of Inuit ancestry on both alleles. In this way we

can also assess rare variants without relying on correct phasing. We manually excluded the HLA region (Chr. 9:28510120–33480577), local ancestry tracts with fewer than 200 variants per megabase and local ancestry tracts with extreme inferred mean-Inuit ancestry (less than 63% or more than 71%) as they are all potentially problematic regions in terms of inferring local ancestry accurately.

## Site frequency spectra

Reference and alternative allele counts were counted using Plink v.1.9.0 (ref. 61) keeping allele order and projected to the wanted number of participants using the formula  $\text{binom}(m,j) \times \text{binom}(n-m, k-j)/\text{binom}(n,k)$ , where  $k$  is the observed number of alternative alleles,  $n$  is the number of total alleles,  $m$  is the number of alleles to project to (that is, two times the number of participants) and  $j$  is the site frequency spectra (SFS)-bin. For each site we get the probability that we would observe  $j$  alternative alleles in a subsample of  $m$  alleles. The probabilities were summed across variants and folded to get the folded SFS.

The derived sum was calculated based on a SFS polarized by ancestral or derived allele only using SNPs with a high-confidence ancestral allele match. Derived alleles were counted and projected to the needed number of participants as above SFS, but not folded.

To measure the number of segregating SNPs as a function of the number of participants sequenced, we projected the SFS to the wanted number of participants, folded the SFS and summed across all the non-zero SFS-bins. In this way, we get the number of segregating SNPs for all possible subsamples of participants from our data.

## Clinical screening

Only variants with good coverage in gnomAD v.3.0.0 as described above and predicted to be LoF on the canonical transcript were used. For each participant, variants with a MAF in any gnomAD population above 0.1% were excluded. For Greenlandic people, an additional count was made where the variants could be excluded based on either MAF from any gnomAD population or MAF from the Greenlanders excluding itself. Figure 2b shows the results for different MAF thresholds. Numbers of people in genetic ancestry groups were as follows: Greenlandic, 448; Greenlandic (unadmixed), 31; Nigerian (YRI + ESN), 207; Han Chinese (CHB + CHS), 208 and British (+CEU) (GBR + CEU), 190.

## Gene burden

A total of 190 participants with the least European ancestry were sampled from the Greenlandic WGS data (on average 12% and at most 20% European ancestry) and 190 participants were sampled randomly from the East Asian populations, CHB and CHS. The European GBR + CEU group had a total of 190 unrelated samples. All variants were annotated with an allele frequency based on the African populations from 1KG without European admixture (LWK, ESN, YRI, MSL and GWD)<sup>52</sup>. Only SNPs with good coverage in gnomAD as defined above, predicted to be missense or LoF variants in the canonical transcript of a constrained gene, and with an African MAF < 0.01% were kept. Constrained genes were defined to be genes with expected number of pLoF and LOEUF score estimated by Karczewski et al.<sup>55</sup> being less than 10 and 1, respectively. This resulted in 9,533 constrained genes from canonical transcripts. In each gene, two burdens were calculated per person: (1) the gene burden, which is 0 if no SNPs were present and 1 if at least one SNP was present and (2) the most common burden, which is 1 if the person carries the most common variant in the gene and 0 otherwise. If the most common burden was less than 50% of the gene burden, we categorize the gene burden as informative and if the most common burden was at least 90% of the gene burden, we categorize the gene as one variant dominates. All calculations were done independently in each of the three populations.

## LD decay and tag-SNPs

We randomly sampled 190 unadmixed (inferred European ancestry less than 1%) and unrelated Greenlanders. From these, we calculated

pairwise LD using Plink v.1.9.0 in a 10-Mb region. The average number of tag-SNPs was calculated by counting the number variants that were in high LD ( $R^2 > 0.8$ ) in regions of varying sizes.

### Genome-wide association studies

Association tests were run using a linear mixed model, with the estimated genetic relationship matrix (GRM) as a random effect, taking population structure and relatedness into account, using GEMMA v.0.98.5 (ref. 63). The GRM was estimated from a set of ‘good’ variants with MAF > 5%, missing less than 1%, INFO > 0.95, and in Hardy–Weinberg equilibrium taking population structure into account (PCAngsd, excluding sites with |SiteF| less than 0.05 and  $P < 1 \times 10^{-6}$ ). We used a leave one chromosome out scheme where the GRM used for testing associations on a given chromosome was estimated using all the other chromosomes. All phenotypes were tested using a score-test after sex-stratified rank-based inverse normal transformation and including age, sex and cohort as covariates. As in previous studies<sup>14</sup>, the odds ratio for diabetes was calculated using the logistic mixed model implemented in GMMAT<sup>64</sup>. For each phenotype, the GRMs were estimated only for people with no missing phenotype or covariate information. To identify independent association signals in the Greenlandic cohorts, we first calculated the LD-adjust<sup>65</sup> correlation,  $R^2$  between pairwise SNPs accounting for the population structure. Then, we performed LD clumping on the above association signals using PCAone v.0.4.4 (ref. 66) adjusted for five principal components with the options ‘--clump-p1 1e-6 --clump-p2 1e-6 --clump-r2 0.001 --clump-bp 10000000 --pc 5’.

European independent genome-wide significant ( $P < 5 \times 10^{-8}$ ) signals within 10 Mb were extracted from summary statistics for the 13 metabolic traits using the extract\_instruments function with default parameters from the R package TwoSampleMR v.0.5.10 (ref. 67) (Supplementary Table 10). The independent signals of type 2 diabetes were extracted from the curated list of Mahajan et al.<sup>68</sup>.

For the sample-size-matched GWAS with UK Biobank, we randomly sampled 5,996 people with no missingness on the phenotypes. For associations on protein abundances, we randomly sampled 3,707 of the 5,996, matching the number of people in the Greenlandic cohorts with protein abundances. The GRMs on the UK Biobank were estimated on variants with MAF > 5% and missing less than 1%, resulting in a total of 4.5 million variants. For all protein abundance associations, we extracted the variant with the lowest  $P$  value, removed all variants within  $\pm 2.5$  Mb of that variant and repeated until no variants were left with a  $P$  value  $< 5 \times 10^{-8}$ . This was done independently for the additive and recessive model, but assigned to the model with the lowest  $P$  value. To avoid duplicated signals of strong associations that were found in both models, we excluded associations within  $\pm 1$  Mb of an association with a lower  $P$  value.

For continuous traits under the additive model, the variance explained was calculated as  $PVE = \beta_{add}^2 2AF(1 - AF)$ , where  $\beta_{add}$  is the additive effect size and AF is the allele frequency of the effect allele<sup>69</sup>. For the recessive model variance explained was calculated as  $PVE = \beta_{rec}^2 F_{hom}(1 - F_{hom})$ , where  $\beta_{rec}$  is the recessive effect size and  $F_{hom}$  is the expected homozygous frequency  $F_{hom} = AF^2$ . This method was used for plotting as we could calculate variance explained CIs using the same formulas but with the lower and upper CI for the effect size. For choosing the variants with more than 1% variance explained we used the formula  $PVE = \beta^2 / (\beta^2 + SE(\beta)^2 \times N)$ , where  $SE(\beta)$  is the standard error of  $\beta$  and  $N$  is the number of participants<sup>70</sup>. For binary traits we calculated the liability-scale variance explained using the R package Mangrove (v.1.21) as previously described<sup>14</sup>.

### Polygenic scores

PGS files from Weissbrod et al.<sup>71</sup> were downloaded from the PGS catalogue (Supplementary Table 11). All variants were lifted over from Hg19 to Hg38 using the UCSC LiftOver command line interface while keeping track of strand-flipping. For both the Greenlandic, Danish and UK Biobank (non-British Europeans) datasets, overlapping variants were

identified by matching on chromosome, position and effect, and other allele on both alternative and reference allele. The PGS was calculated on the genotype dosages using the score function in Plink2 v.2.0.0 (ref. 61). Each phenotype was rank-based inverse normal transformed separately for males and females. Next, two linear models were fitted: (1) Null model, phenotype is approximately age × sex + PC1–10 and (2) PGS model, phenotype is approximately age × sex + PC1–10 + PGS. Then, we calculated the incr.  $R^2 = \text{adjusted } R^2(\text{PGS model}) - \text{Adjusted } R^2(\text{Null model})$ . For the Greenlandic cohorts, an additional covariate for cohort was included and for the UK Biobank an additional covariate for assessment centre was included. Body mass index was included as a covariate for waist:hip ratio. The CIs for the incr.  $R^2$  were calculated using the Olkin and Finn’s approximation for s.e. To summarize across all phenotypes, we calculated the relative incr.  $R^2$  using the UK Biobank as baseline. The increased performance on incr.  $R^2$  on LDL-cholesterol for the Danish cohort was probably due to differences in age distributions (age range, UK Biobank, 46–80 years; Danish, 30–60 years). The population-based Greenlandic cohort design followed the population-based design of the Danish Inter99. The UK Biobank had a completely different cohort design and is not completely compatible; for example, mean age in UK Biobank is 64 years, whereas mean age in the Danish and Greenlandic cohort is 46 and 45 years, respectively.

### Relatedness analysis and relation graph

As described previously<sup>72</sup>, we estimated relatedness using a filtered set of genetic variants with MAF > 5%, missingness < 5% and LD-pruned (Plink v.1.9.0 indep-pairwise 1,000 kb 1 0.8) along with the inferred admixture proportions as input to NGSremix<sup>73</sup>. For each pair, NGSremix calculates pairwise relatedness as the fraction of loci sharing zero, one or two alleles identical by descent (represented by k0, k1 and k2, respectively). Parent–offspring pairs were defined as relationships with  $k_1 + k_2 > 0.95$  and  $k_1 > 0.75$  and the parent was inferred using the age of the participants. Full sibling pairs were defined as relationships with  $0.3 < k_1 < 0.7$  and  $k_2 > 0.1$ . Out of 5,828 people with age and location, we identified 1,727 parent–offspring and 1,841 full sibling relationships. Relationships were normalized to the number of possible pairs. For relationships between regions, for example, region 1 and 2, the number of possible pairs was calculated as  $n_{\text{possible}(1,2)} = n_{\text{region1}} \times n_{\text{region2}}$ , where  $n_{\text{region1}}$  and  $n_{\text{region2}}$  is the number of participants in region 1 and 2, respectively. Within region, the number of possible pairs of parent–offspring relationships was calculated as  $n_{\text{possible}(1,1)} = n_{\text{region1}} \times (n_{\text{region1}} - 1)$ . Within region, the number of possible pairs of full sibling relationships was calculated as  $n_{\text{possible}(1,1)} = (n_{\text{region1}} \times (n_{\text{region1}} - 1)) / 2$ .

### Expected homozygous frequency

To calculate the expected frequency of homozygous carriers with the current fine structure,  $f_{hom(\text{structure})}$ , we estimated the regional allele frequencies,  $AF_{\text{region}}$ , based on sample location, calculated the expected number of homozygous carriers in each region as  $n_{hom(\text{region})} = n_{\text{region}} \times AF_{\text{region}}^2$ , calculated the total sum of homozygous carriers,  $n_{hom} = n_{hom(\text{region1})} + n_{hom(\text{region2})} + \dots + n_{hom(\text{region8})}$  and divided with the total number of participants  $f_{hom(\text{structure})} = n_{hom} / n_{\text{total}}$ . The expected frequency of homozygous carriers as a panmictic population,  $f_{hom(\text{panmictic})}$ , was estimated as  $f_{hom(\text{panmictic})} = AF^2$ . CIs of the homozygous frequencies were estimated as the s.d. of the frequency estimates from 10,000 bootstrap samples.

### Estimation of allele frequencies in Greenland and in other datasets

We estimated frequencies for different variants in the Greenlandic Inuit population using the masked participants from this study. Moreover, we surveyed the allele frequencies for those variants in a range of available datasets<sup>12,17,26,29,30,74–81</sup> from across the world (Supplementary Tables 7 and 8). We used SAMtools<sup>82</sup> and BGTE<sup>83</sup> to extract the allele counts of the relevant variants from those datasets.

# Article

## Ancestral recombination graph

The ARG was estimated in two versions: (1) using all 448 WGS Greenlanders (used for analysis of *FADS2* and *CPT1A*) and (2) dividing the genome into 6-Mb chunks, finding people with only Inuit ancestry in that chunk and randomly subsampling 150 of those people, resulting in a Greenlandic Inuit masked ARG from 150 mosaic participants (used for all other analyses). For the masked version, only chunks with more than one variant per 5,000 bp were used. Both ARGs were estimated using the default Relate v.1.2.1 pipeline; for each chromosome (or chunk, in the masked version), convert phased VCF to haps/sample file format, prepare input using the provided PrepareInputFiles script, with the high-confidence Ensembl *H. sapiens* ancestral sequence as ancestral reference and estimate the ARG for the chromosome/chunk using the provided RelateParallel.sh script with Effective population size of haplotypes set to 2,000 and mutation rate per generation to  $2.5 \times 10^{-8}$ . The overall population size was then estimated using the provided EstimatePopulationSize script on all chromosomes/chunks, resulting in the estimated population size history.

To estimate variant age, we made 10,000 branch length samples of the local tree from the estimated ARG using the provided SampleBranchLengths script from Relate. Note that Relate does not allow for changes in the tree topology when resampling. From each sample, we extracted the time to the most recent common ancestor for the variant as well as the preceding ancestor. This yields a minimum and maximum age of the variant measured in generations for each branch length sample. From the intervals, we estimated the age of the variants and the credibility interval. We calculated a probability density as the weighted average of the intervals where the weight is the reciprocal of each interval length. By doing so we assumed that the age is equally likely to lie anywhere within each interval and we gave equal weight to each of the 10,000 sampled branch lengths. The variant estimate was estimated to be the median of the probability density and the 95% credible interval was calculated as 2.5% and 97.5% quantiles. The age in generations was converted to years by multiplying with the assumed generation time of 28 years per generation.

## Selection

As with variant age estimates above, we tested for selection using branch length samples of the local tree from the estimated ARG. These branch length samples were used as input to CLUES to infer allele frequency trajectories and test for selection<sup>44</sup>. To obtain empirical *P* values, we tested for selection for 999 additional variants matched to have within  $\pm 10\%$  derived allele frequency of each variant. The empirical *P* value was then calculated as log-likelihood rank divided by the total number of variants tested. Random variants were not sampled within  $\pm 5$  Mb of any of the tested variants.

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Data are archived securely at the European Genome-phenome Archive; MEGA-chip data in IHIT/B99 (EGAD00010002057), MEGA-chip in B2018 (EGAD50000000934) and WGS data (EGAD50000000933). Data access is restricted and any access is contingent on approval by the Research Ethics Committee of Greenland (nun@nanoq.gl) and subsequent acceptance by the dataowner (Department of Health, Greenland Government; pn@nanoq.gl). We used the publicly available final phase 31000 Genomes data found here: [https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\\_collections/1000G\\_2504\\_high\\_coverage/working/20201028\\_3202\\_phased/](https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/). Maps in Fig. 1a and Fig. 3a were drawn using the R-library maps v.3.4.1.1 (ref. 84), which relies on data from the Natural Earth data (<https://www.naturalearthdata.com/>).

## Code availability

Code used for analysis is available at <https://github.com/popgenDK/greenlandWGS2024>. Most analyses were run in parallel using GNU parallel 20190922 (ref. 85).

45. Olesen, I., Hansen, N. L., Jørgensen, M. E. & Larsen, C. V. L. *Perspektiver på genetiske undersøgelser af diabetes, belyst gennem Sharing Circles: En kvalitativ undersøgelse i Qasigiannguit* (Syddansk Universitet, Statens Institut for Folkesundhed, 2020).
46. Sønftleber, N. K. et al. The effect of sucrase-isomaltase deficiency on metabolism, food intake and preferences: protocol for a dietary intervention study. *Int. J. Circumpolar Health* **82**, 2178067 (2023).
47. Bjerregaard, P. et al. Inuit health in Greenland: a population survey of life style and disease in Greenland and among Inuit living in Denmark. *Int. J. Circumpolar Health* **62**, 3–79 (2003).
48. Bjerregaard, P. *Inuit Health in Transition Greenland Survey 2005–2009* (Syddansk Universitet, Statens Institut for Folkesundhed, 2010).
49. Bjerregaard, P. et al. The Greenland population health survey 2018—methods of a prospective study of risk factors for lifestyle related diseases and social determinants of health amongst Inuit. *Int. J. Circumpolar Health* **81**, 2090067 (2022).
50. Larsen, C. V. L. et al. *Befolkningsundersøgelsen i Grønland 2018. Levevilkår, livsstil og helbred: Oversigt over indikatorer for folkesundhed* (Syddansk Universitet, Statens Institut for Folkesundhed, 2019).
51. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
52. The 1000 Genomes Project Consortium A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
53. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
54. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
55. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
56. Andersen, M. K. The derived allele of a novel intergenic variant at chromosome 11 associates with lower body mass index and a favorable metabolic phenotype in Greenlanders. *PLOS Genet.* **16**, e1008544 (2020).
57. Alberti, K. G. & Zimmet, P. Z. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: definition and classification of diabetes mellitus provisional report of a WHO consultation. *Diabet. Med.* **15**, 539–553 (1998).
58. Jørgensen, T. et al. A randomized non-pharmacological intervention study for prevention of ischaemic heart disease: baseline results Inter99. *Eur. J. Cardiovasc. Prev. Rehabil.* **10**, 377–386 (2003).
59. Grarup, N. et al. Identification of novel high-impact recessively inherited type 2 diabetes risk variants in the Greenlandic population. *Diabetologia* **61**, 2005–2015 (2018).
60. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
61. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
62. Li, Z. vcffp: a C++ API for rapid processing of the variant call format. *Bioinformatics* **40**, btae049 (2024).
63. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
64. Chen, H. et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.* **98**, 653–666 (2016).
65. Bercovich, U., Rasmussen, M. S., Li, Z., Wiuf, C. & Albrechtsen, A. Measuring linkage disequilibrium and improvement of pruning and clumping in structured populations. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.05.02.592187> (2024).
66. Li, Z., Meisner, J. & Albrechtsen, A. Fast and accurate out-of-core PCA framework for large scale biobank data. *Genome Res.* **33**, 1599–1608 (2023).
67. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human genome. *eLife* **7**, e34408 (2018).
68. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
69. Park, J.-H. et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).
70. Shim, H. et al. A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. *PLoS ONE* **10**, e0120758 (2015).
71. Weissbrod, O. et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).
72. Lin, L. et al. Analysis of admixed Greenlandic siblings shows that the mean genotypic values for metabolic phenotypes differ between Inuit and Europeans. *Genome Med.* **16**, 71 (2024).
73. Nøhr, A. K. et al. NGremix: a software tool for estimating pairwise relatedness between admixed individuals from next-generation sequencing data. *G3 (Bethesda)* **11**, jkab174. (2021).
74. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
75. Cann, H. M. et al. A human genome diversity cell line panel. *Science* **296**, 261–262 (2002).
76. Malyarchuk, B. A., Derenko, M. V., Denisova, G. A. & Litvinov, A. N. Distribution of the arctic variant of the *CPT1A* gene in indigenous populations of Siberia. *Vavilov J. Genet. Breed.* **20**, 571–575 (2016).
77. Malyarchuk, B. A., Derenko, M. V. & Denisova, G. A. The frequency of inactive sucrase-isomaltase variant in indigenous populations of Northeast Asia. *Russ. J. Genet.* **53**, 1052–1054 (2017).

78. Wong, E. H. M. et al. Reconstructing genetic history of Siberian and Northeastern European populations. *Genome Res.* **27**, 1–14 (2017).
79. Lemas, D. J. et al. 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**, 175–184 (2012).
80. Bell, R. R., Draper, H. H. & Bergan, J. G. Sucrose, lactose, and glucose tolerance in northern Alaskan Eskimos. *Am. J. Clin. Nutr.* **26**, 1185–1190 (1973).
81. Parajuli, R. P. et al. Variation in biomarker levels of metals, persistent organic pollutants, and omega-3 fatty acids in association with genetic polymorphisms among Inuit in Nunavik, Canada. *Environ. Res.* **200**, 111393 (2021).
82. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
83. Li, H. BGT: efficient and flexible genotype query across many samples. *Bioinformatics* **32**, 590–592 (2016).
84. Becker, R. A. et al. maps: draw geographical maps V.3.4.1.1 <https://CRAN.R-project.org/package=maps> (2023).
85. Tange, O. Gnu parallel 2018. Zenodo <https://doi.org/10.5281/zenodo.1146013> (2018).

**Acknowledgements** We sincerely thank all the study participants. We thank E. Hjerresen for help with the classification of birth towns into regions. The Novo Nordisk Foundation Centre for Basic Metabolic Research is an independent research centre at the University of Copenhagen partially supported by an unrestricted donation from the Novo Nordisk Foundation (NNF23SA0084103). Steno Diabetes Center Greenland is partly supported by the Novo Nordisk Foundation (NNF20SA0064190) and the population surveys in Greenland are partly supported by the Novo Nordisk Foundation (NNF17OC0028136). Novo Nordisk Foundation supported F.F.S., R.F.B., S.H., Z.L. and A.A. (NNF20OC0061343), A.A. (NNF23OC0084422), S.E.S. (NNF18CC0033668) and A.C.B.T. (NNF20SA0064340). F.F.S., R.F.B. and I.M. were supported by a Villum Young Investigator

grant (VIL19114). A.A. and M.S.R. were supported by the Independent Research Fund Denmark (8021-00360B). C.G.S. and I.M. were supported by the European Research Council (ERC-2018-STG-804679).

**Author contributions** Conceptualization: T.H., I.M. and A.A. Methodology: F.F.S., K.H., M.S., J.M., R.N., I.M. and A.A. Software: F.F.S., Z.L., S.H., K.H., M.S.R. and J.M. Validation: K.F.R. and I.H.S. Formal analysis: F.F.S., Z.L., S.H., C.G.S., R.T.J., K.H., E.J., M.S.R., R.F.B., M.S. and J.M. Investigation: F.F.S., A.L., N.G., E.Z., M.E.J. and T.H. Resources: C.V.L.L., P.B., M.E.J., T.H., I.M. and A.A. Data curation: F.F.S., C.V.L.L., P.B., M.K.A., Z.L., J.P.H., K.F.R., A.C.B.T., E.J., S.E.S., M.S., N.G., I.M. and A.A. Writing—original draft: F.F.S., M.K.A., M.E.J., T.H., I.M. and A.A. Writing—review and editing: F.F.S., M.K.A., Z.L., J.P.H., S.H., C.G.S., R.T.J., K.F.R., A.C.B.T., K.H., I.H.S., E.J., S.E.S., M.S.R., R.F.B., C.V.L.L., P.B., M.S., J.M., A.L., N.G., E.Z., R.N., M.E.J., T.H., I.M. and A.A. Visualization: F.F.S., S.H. and A.A. Supervision: R.N., T.H., I.M. and A.A. Project administration: M.K.A., A.C.B.T., N.G., M.E.J., C.V.L.L., P.B., T.H., I.M. and A.A. Funding acquisition: E.Z., M.E.J., T.H., I.M. and A.A.

**Competing interests** M.E.J., I.M. and T.H. hold shares in Novo Nordisk. N.G., K.H. and M.S.R. are now employed at Novo Nordisk. The remaining authors declare no competing interests.

#### Additional information

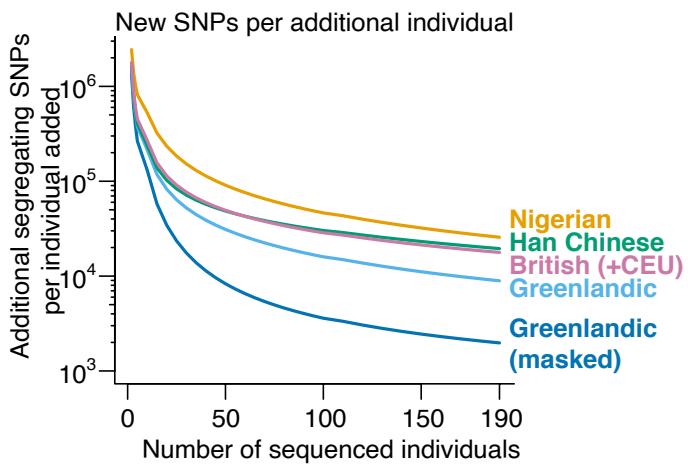
**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41586-024-08516-4>.

**Correspondence and requests for materials** should be addressed to Torben Hansen, Ida Moltke or Anders Albrechtsen.

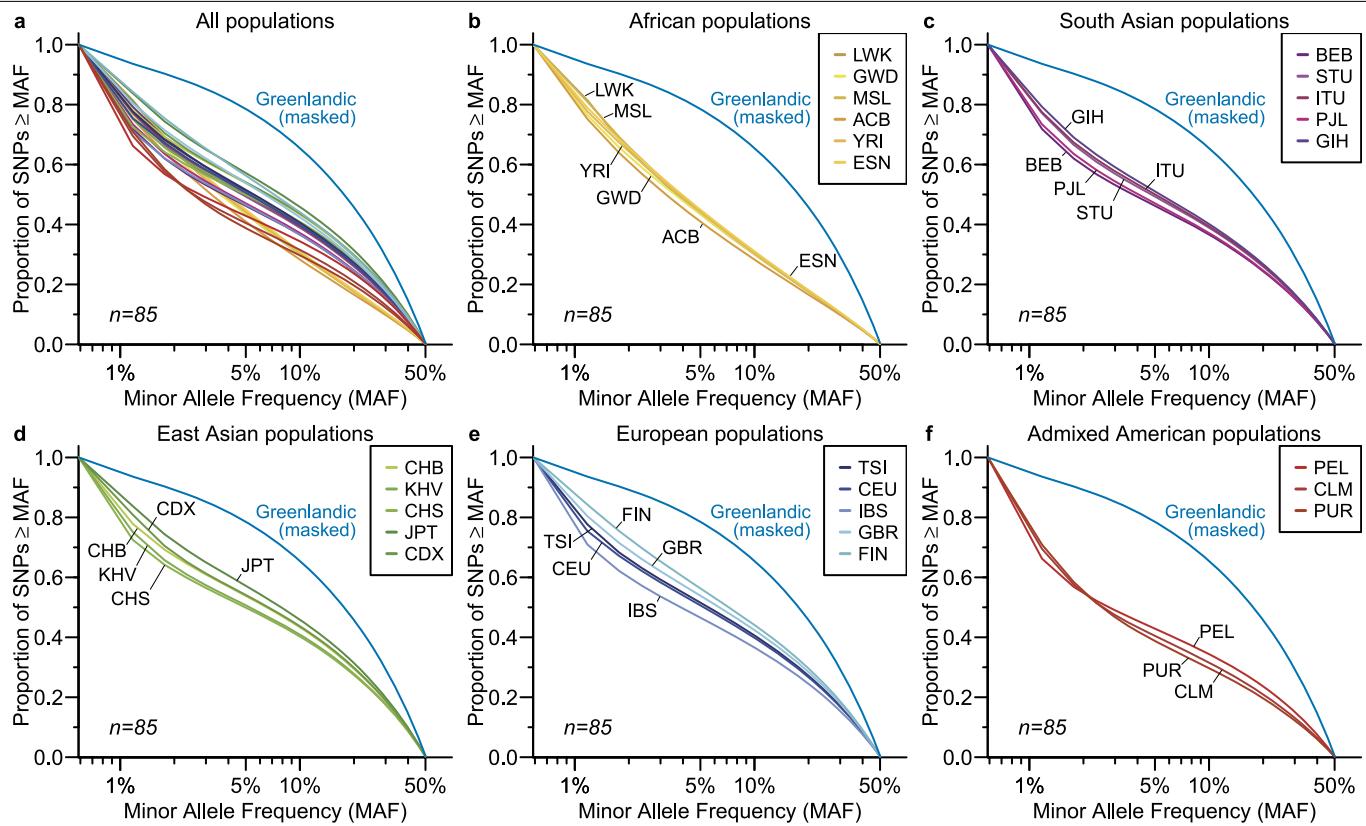
**Peer review information** *Nature* thanks Arun Durvasula and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

## Article



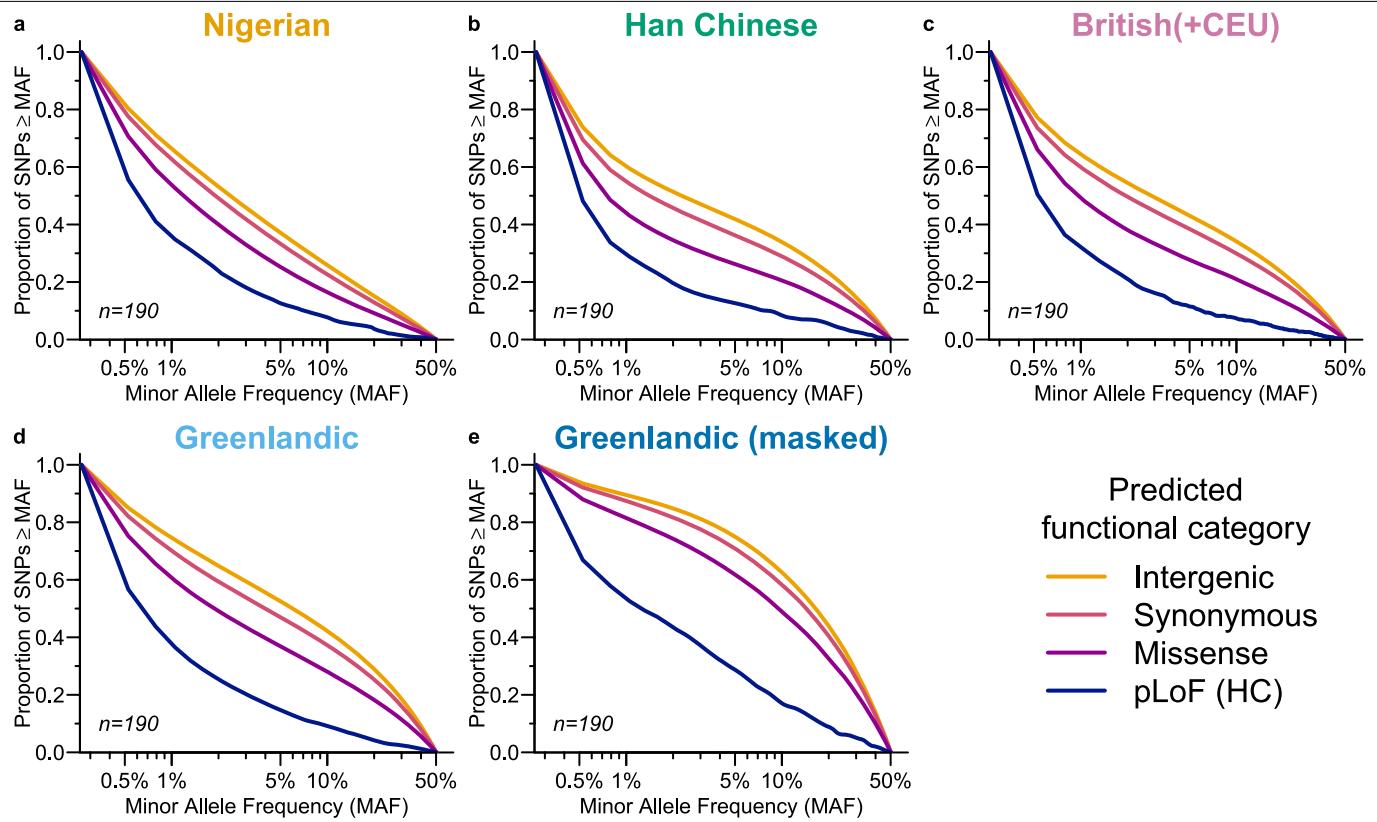
**Extended Data Fig. 1 | Additional SNPs per sequenced individual.** Number of new SNPs added per additionally sequenced individual for Greenlanders compared to Nigerian, British(+CEU) and Han Chinese samples from IKG.



**Extended Data Fig. 2 | Allele frequency distributions.** **a**, Proportion of polymorphic SNPs for all 1KG populations with at least 85 individuals projected to 85 individuals. The proportion of SNPs is shown as a function of the allele frequency, i.e. at  $\text{MAF} = 5\%$ , the proportion of polymorphic SNPs with  $\text{MAF} \geq 5\%$

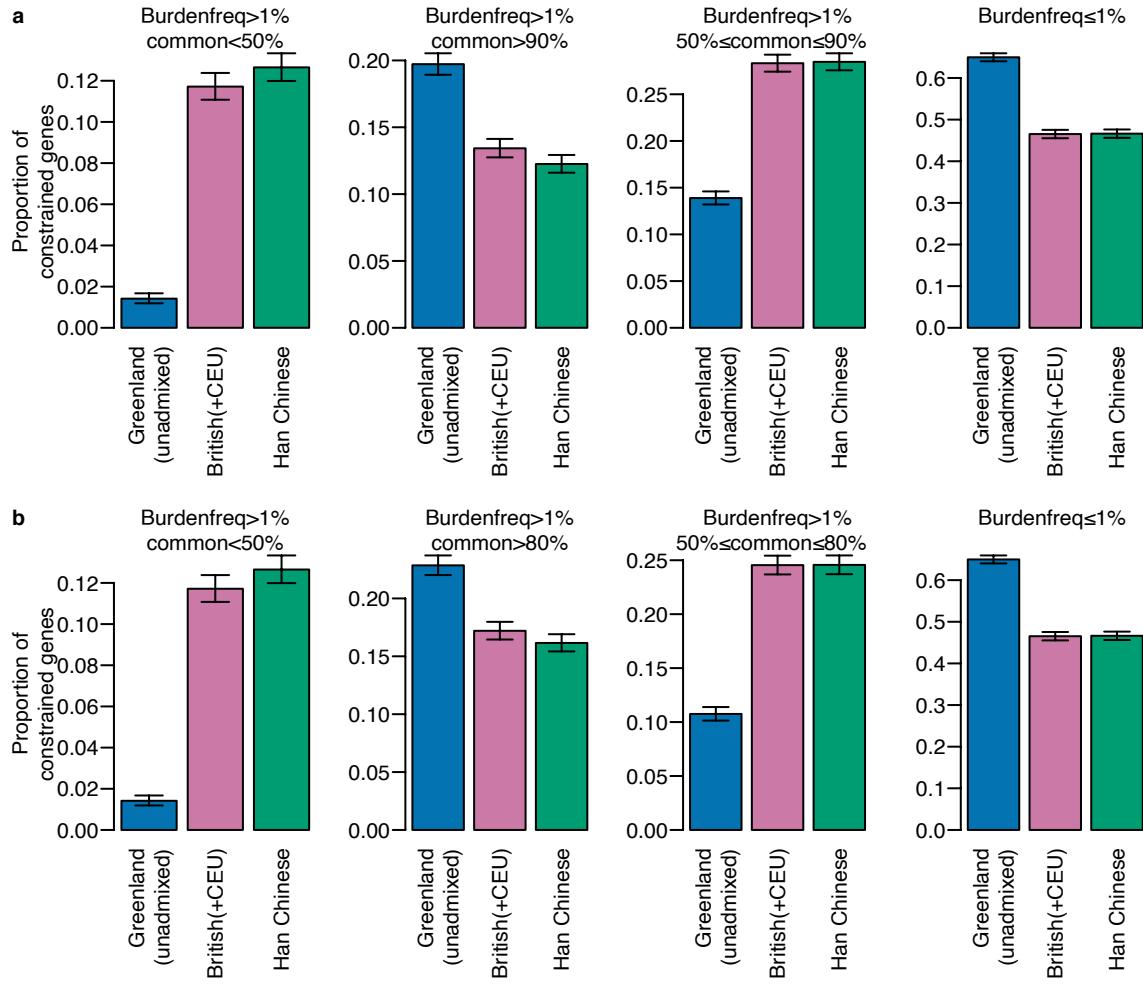
is shown. **a**, All 1KG populations, **b-f**, Populations in the African, South Asian, East Asian, European, or Admixed American 1KG superpopulation, respectively.

# Article



**Extended Data Fig. 3 | Allele frequency distributions in predicted functional groups.** Proportion of polymorphic SNPs grouped by predicted functional categories for the a-e, Nigerian, Han Chinese, British(+CEU),

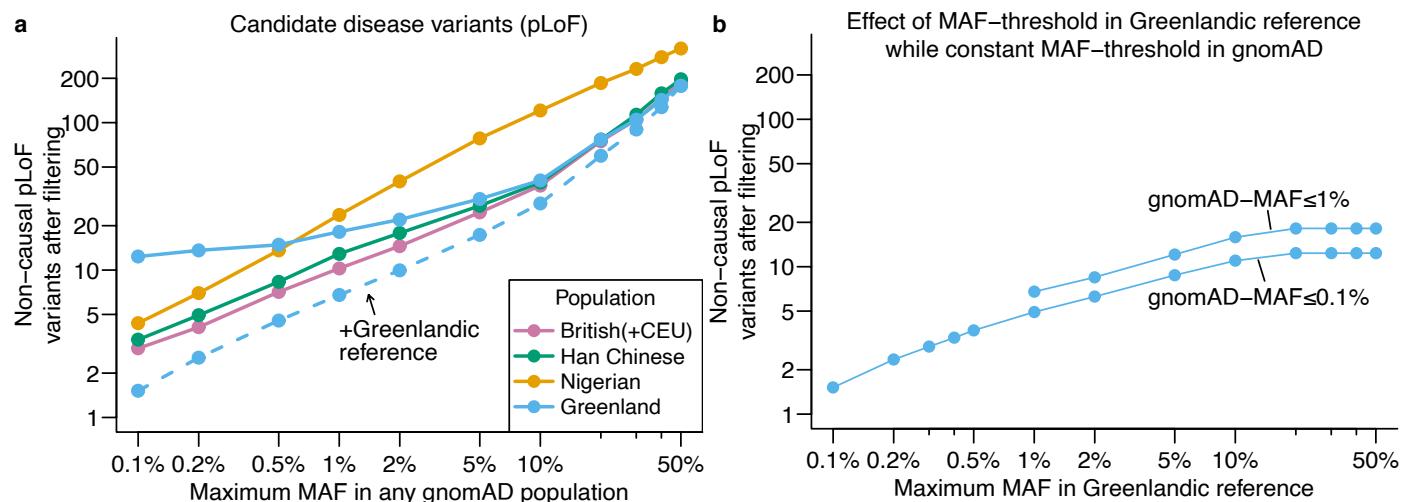
Greenlandic, and masked Greenlandic, respectively. pLoF (HC) is the predicted loss of function SNPs LOFTEE high confidence.



**Extended Data Fig. 4 | Predicted Gene burden distribution in constrained genes.** The gene burden frequency, Burdenfreq, is the proportion of individuals carrying one or more predicted deleterious variants in a gene and ‘common’ is the proportion of the gene burden which is attributed to the most common predicted deleterious variant in the gene. The predicted deleterious variants were here defined as being a missense or LoF variant with an allele frequency lower than 0.01% in African individuals (1KG populations: LWK, ESN, YRI, MSL,

and GWD). Proportion (95%CI) of constrained genes. **a**, The two first panels are the same as Fig. 1g, the third panel shows the proportion of constrained genes where the most common predicted deleterious variant contributes between 50% to 90% of the gene burden, and the fourth panel shows the proportion of genes with a gene burden frequency  $\leq 1\%$ . **b**, similar to a, but changing the common-threshold to 80% instead of 90%.

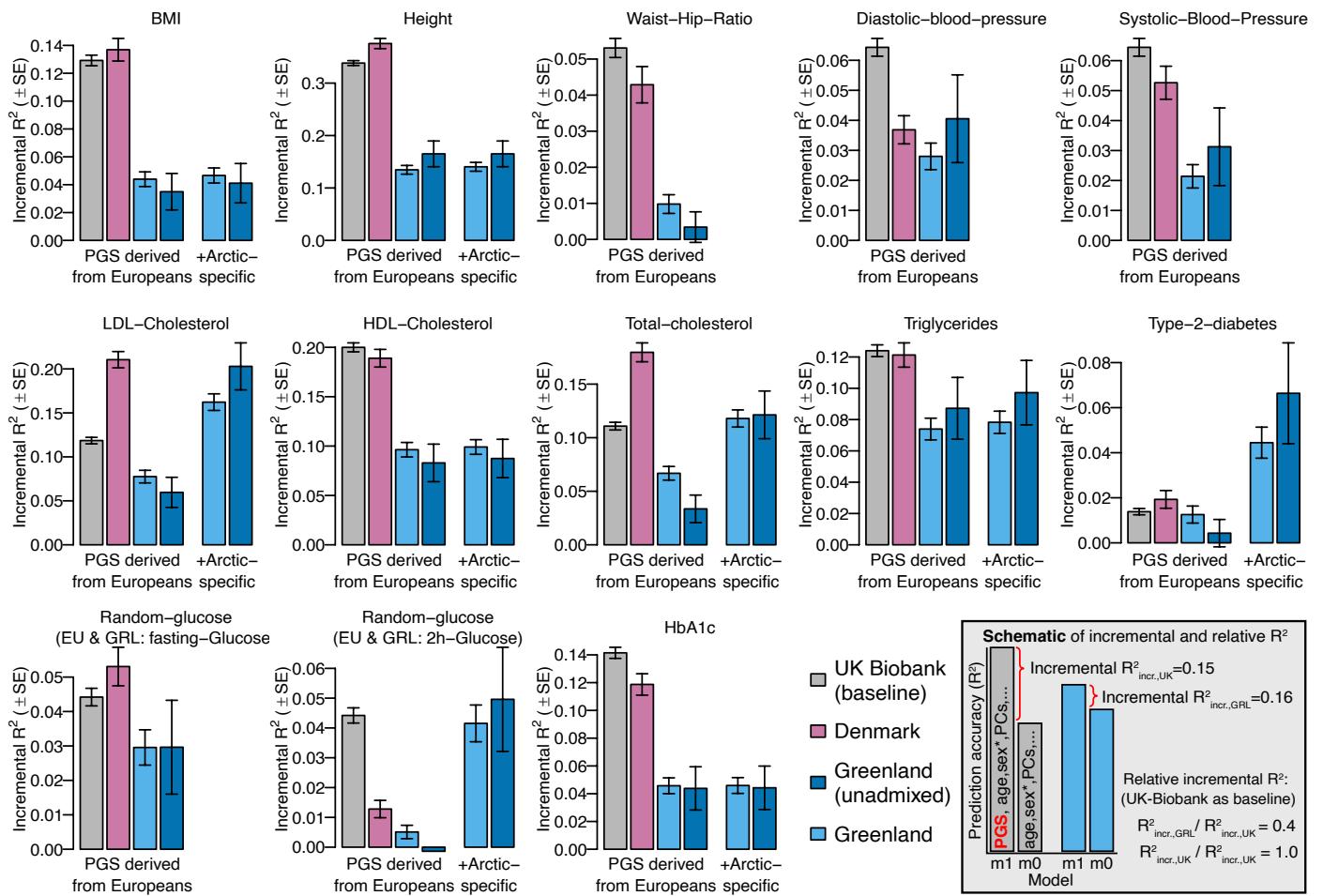
# Article



**Extended Data Fig. 5 | Number of non-causal variants in clinical screening setting.**

**a**, Mean number of non-causal pLoF variants remaining after removing variants with given frequency in gnomAD. The dashed blue line is the Greenlandic population after removing variants with above given frequency in either

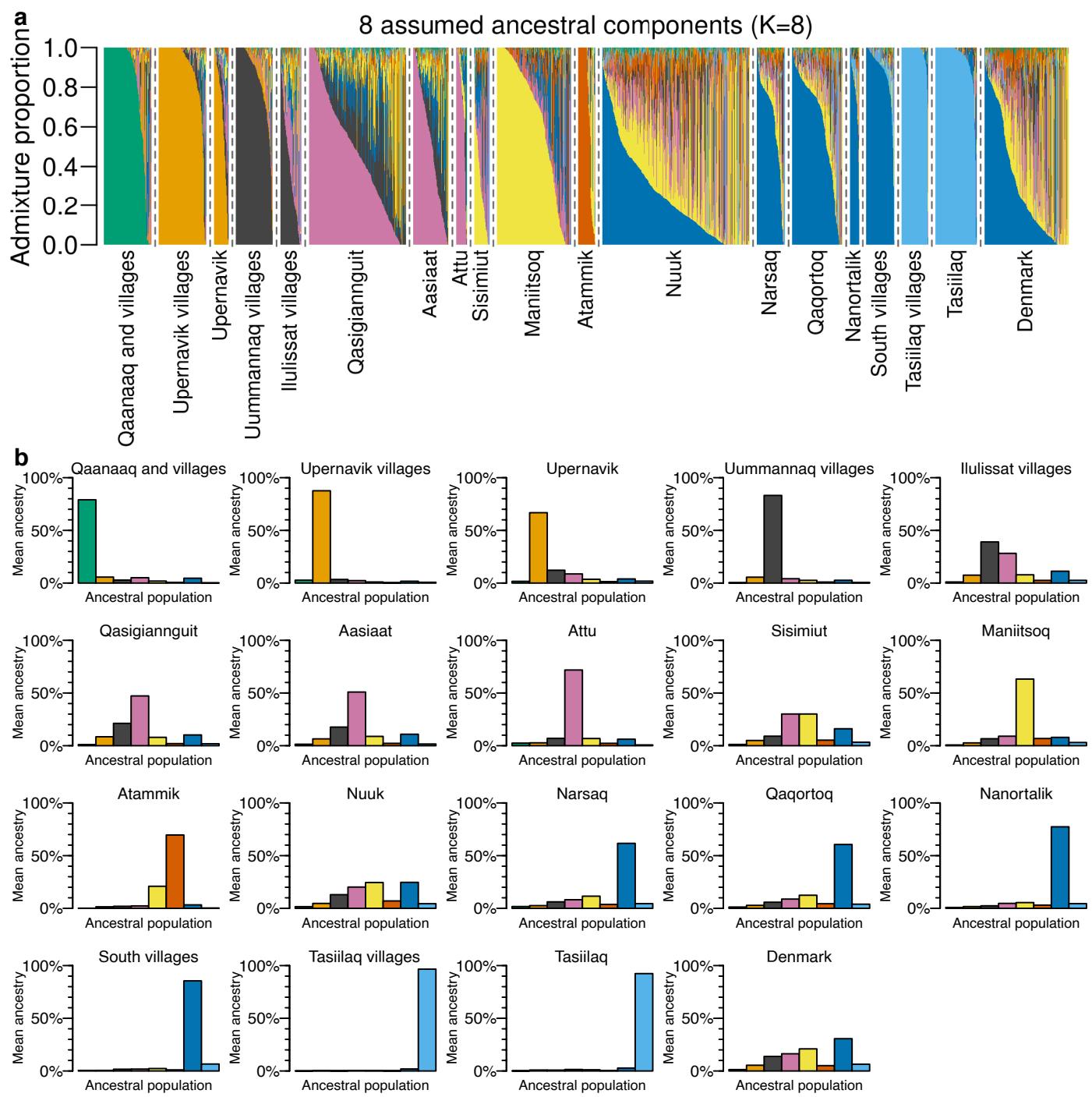
gnomAD or in the Greenlethic reference. **b**, same as panel a, but only for the Greenlandic population at two different fixed gnomAD-MAF thresholds and varying MAF threshold (x-axis) in the Greenlethic reference.



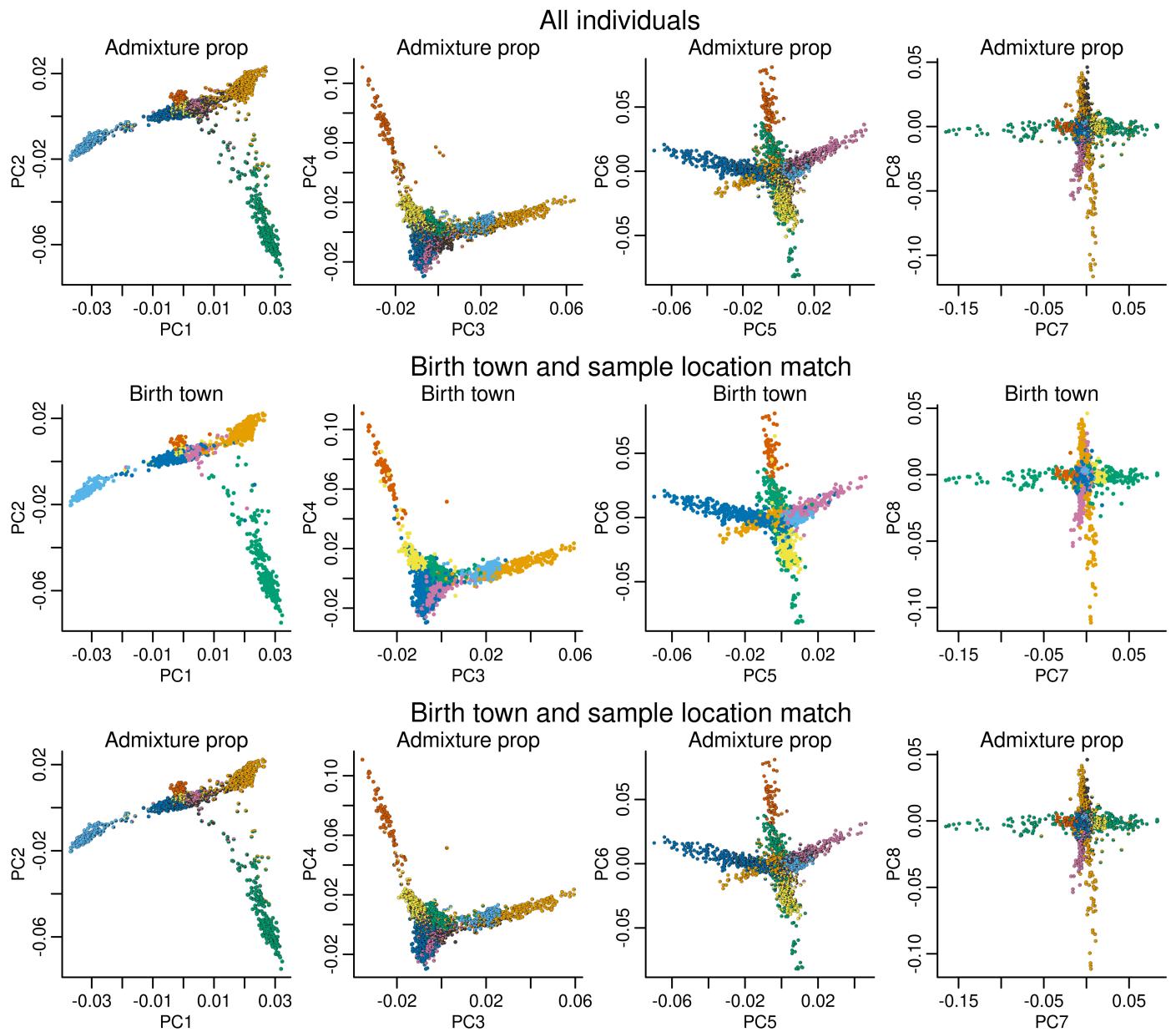
**Extended Data Fig. 6 | Polygenic score predictive performance.** Incremental  $R^2$  of polygenic score (PGS) prediction of phenotypes for UK biobank (non-British Europeans), Denmark (Inter99), Greenland and Greenland unadmixed. For most phenotypes, one or more Arctic-specific variants could be added to the PGS and improved the prediction. List of Arctic-specific variants and their effect in Supplementary table 4. To test whether the incremental  $R^2$  is improved with the

added Arctic-specific variants, we performed a two-sided paired t-test on the incremental  $R^2$  value in Greenland with and without the Arctic-specific variants in the PGS. The improvement was significant both when the test was both done on all traits ( $p$ -value = 0.04067) and only on the traits with Arctic-specific variants ( $p$ -value = 0.03647). \*Traits were rank-based inverse normal transformed separately for each sex and sex was also included as a covariate in the model.

# Article



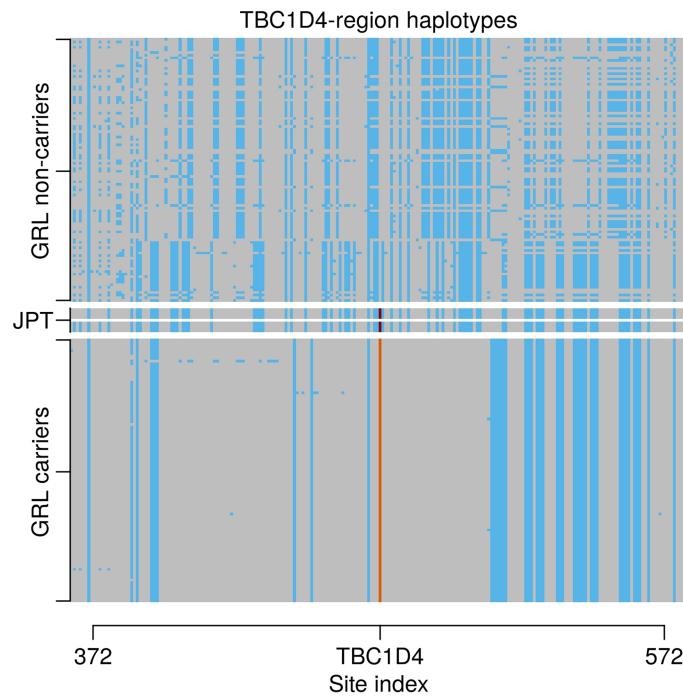
**Extended Data Fig. 7 | Inferred admixture proportion of non-European ancestry.** **a**, Estimated individual admixture proportions from *Haplotype Admix* on masked haplotypes excluding any European ancestry. **b**, Mean admixture proportions of all samples in sample location.



**Extended Data Fig. 8 | HaploNet principal components of non-European ancestry.** Principal component analysis from HaploNet PCA. **a**, All samples visualised with individual ancestry proportions as pie charts. **b**, samples with

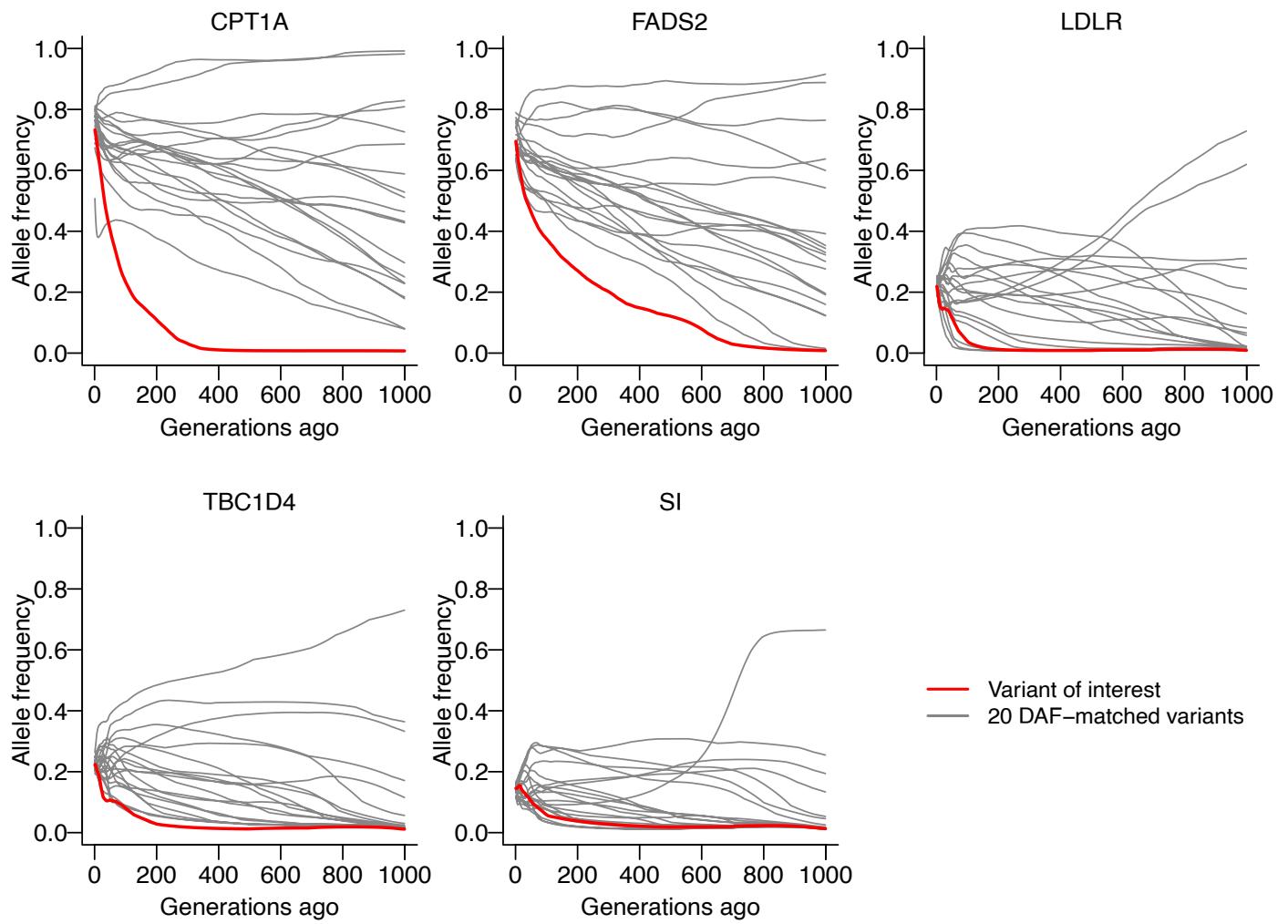
<50% European and matching birth town and sample location, coloured by birth town region. **c**, same samples as in a but only for individuals with birth town information and where location and birth town was in the same region.

# Article



## Extended Data Fig. 9 | TBC1D4 haplotypes in Greenland and 1KG-JPT.

Haplotype plot of the genomic region around the variant in *TBC1D4* for Greenlandic heterozygous carriers ( $n = 116$ ) and the Japanese individual (JPT) from 1KG carrying the variant. For the Greenlandic individuals, their haplotypes without the *TBC1D4* variant are shown on top and their haplotypes with the variant are shown on the bottom. Grey means reference allele, blue means alternative allele, and orange/red highlight the *TBC1D4* variant. Since the variant is tri-allelic, the Japanese individual did not have phasing information for that variant. Notice that the Japanese individual matches haplotypes without the *TBC1D4* variant and not the haplotype with the variant which suggests that it is a recurrent mutation.



**Extended Data Fig. 10 | Clues-inferred allele frequency trajectories.** Posterior mean of allele frequency trajectories inferred by Clues on the five variants tested (in red) and 20 random DAF-matched variants (in grey).

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection     The R-package TwoSampleMR v.0.5.10 was used to extract independent association signals for European GWAS.

Data analysis     All code for data analysis is provided here: <https://github.com/popgenDK/greenlandWGS2024>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Genetic data is archived at the European Genome-phenome Archive; MEGA-chip data in IHIT/B99 (EGAD00010002057), MEGA-chip in B2018 (EGAD50000000934), and whole genome sequencing data (EGAD50000000933). Use of the data is contingent on approval by the Research Ethics Committee of Greenland (nun@nanoq.gl) and subsequent acceptance by the dataowner (Department of Health, Greenland Government; pn@nanoq.gl).

We used the publicly available final phase 3 1000 genomes data found here: [https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\\_collections/1000G\\_2504\\_high\\_coverage/working/20201028\\_3202\\_phased/](https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/)

## Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

### Reporting on sex and gender

Sex was used as a covariate in association studies and in assessment of polygenic score performance  
No information on gender was collected.

### Reporting on race, ethnicity, or other socially relevant groupings

We only use genetically inferred ancestry categories and we estimate genetic admixture proportions. This is used throughout the manuscript to understand the historical and contemporary effect of demography on genetic architecture of disease. We do not use the terms race or ethnicity nor do we use self identified labels.

### Population characteristics

The individuals in this study were included in one or more of three different population health surveys from Greenland. The age distribution is from 18 to 95 and mean age of 45.

### Recruitment

Quote from manuscript: "individuals were recruited in representative towns and settlements of each region in Greenland where random samples were drawn from the central population registers"

### Ethics oversight

Science Ethics Committee in Greenland ( <https://nun.gl> )

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences

Behavioural & social sciences

Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

### Sample size

Sample sizes were based on the practical limit of the teams conducting the population health surveys (see Bjerregaard, P. "Inuit Health in Transition Greenland survey 2005-2009").

### Data exclusions

Duplicated samples (inferred from genetics) were excluded. Samples with more than 2% missing data on the SNPchip were excluded. Genotyped variants with more than 5% missing data were excluded.

### Replication

No replication experiments were carried out in this study.

### Randomization

Individuals were randomly samples from the population registers. This study does not include any experimental groups.

### Blinding

No blinding. No treatment or experimental groups.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Involved in the study                                  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies                    |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data                 |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Plants                        |

### Methods

- |                                     |   |
|-------------------------------------|---|
| n/a                                 | Involved in the study                           |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq               |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

## Plants

### Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

### Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

### Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.