Open access • Posted Content • DOI:10.1101/2020.02.18.933622

A whole-genome sequenced control population in northern Sweden reveals subregional genetic differences — Source link

Daniel Svensson, Matilda Rentoft, Anna M. Dahlin, Emma Lundholm ...+7 more authors

Institutions: Umeå University, Science for Life Laboratory, Swedish Defence Research Agency

Published on: 19 Feb 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Population and Genetic variation

RESEARCH ARTICLE

A whole-genome sequenced control population in northern Sweden reveals subregional genetic differences

Daniel Svensson[1☯], **Matilda Rentoft**[2☯], **Anna M. Dahlin**[3], **Emma Lundholm**[4], **Pall I. Olason**[5], **Andreas Sjödin**[1,6], **Carin Nylander**[3], **Beatrice S. Melin**[3], **Johan Trygg**[1], **Erik Johansson**[2]*

1 Department of Chemistry, Computational Life Science Cluster, Umeå University, Umeå, Sweden, **2** Department of Medical Biochemistry and Biophysics, Umeå University, Umeå, Sweden, **3** Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden, **4** Centre for Demography and Ageing, Umeå University, Umeå, Sweden, **5** Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden, **6** Division of CBRN Security and Defence, FOI–Swedish Defence Research Agency, Umeå, Sweden

☯ These authors contributed equally to this work.
* erik.tm.johansson@umu.se

OPEN ACCESS

Citation: Svensson D, Rentoft M, Dahlin AM, Lundholm E, Olason PI, Sjödin A, et al. (2020) A whole-genome sequenced control population in northern Sweden reveals subregional genetic differences. PLoS ONE 15(9): e0237721. https://doi.org/10.1371/journal.pone.0237721

Editor: Ludmila Prokunina-Olsson, National Cancer Institute, UNITED STATES

Received: March 2, 2020

Accepted: July 31, 2020

Published: September 11, 2020

Data Availability Statement: Due to legal restriction imposed by the data protection law that is reviewed by the Swedish Data Protection Authority (https://www.datainspektionen.se/other-lang/in-english/) we are not allowed to share individual-level data from human subject whole genome-sequencing by full open access, as this can be equated with personal information. Instead, any researcher interested in the individual-level data may apply for access through a specific application board at the Department of Biobank Research at Umeå University (ebf@umu.se). The

Abstract

The number of national reference populations that are whole-genome sequenced are rapidly increasing. Partly driving this development is the fact that genetic disease studies benefit from knowing the genetic variation typical for the geographical area of interest. A whole-genome sequenced Swedish national reference population (n = 1000) has been recently published but with few samples from northern Sweden. In the present study we have whole-genome sequenced a control population (n = 300) (ACpop) from Västerbotten County, a sparsely populated region in northern Sweden previously shown to be genetically different from southern Sweden. The aggregated variant frequencies within ACpop are publicly available (DOI 10.17044/NBIS/G000005) to function as a basic resource in clinical genetics and for genetic studies. Our analysis of ACpop, representing approximately 0.11% of the population in Västerbotten, indicates the presence of a genetic substructure within the county. Furthermore, a demographic analysis showed that the population from which samples were drawn was to a large extent geographically stationary, a finding that was corroborated in the genetic analysis down to the level of municipalities. Including ACpop in the reference population when imputing unknown variants in a Västerbotten cohort resulted in a strong increase in the number of high-confidence imputed variants (up to 81% for variants with minor allele frequency < 5%). ACpop was initially designed for cancer disease studies, but the genetic structure within the cohort will be of general interest for all genetic disease studies in northern Sweden.

Introduction

The challenge for all studies on genetic diseases is to disentangle disease-causing genetic variation. This task is inherently dependent on knowledge of the genetic variation in the

application is reviewed by the Biobank expert committee, and through amendment of ethical permission and relevant data sharing agreements, the data can be shared. Aggregated data (variant frequencies) can be freely accessed through the National Bioinformatics Infrastructure Sweden site (https://swefreq.nbis.se/dataset/ACpop), DOI: 10.17044/NBIS/G000005.

Funding: This work was funded by a project grant (2011.0042) from the Knut and Alice Wallenberg Foundation (E.J., B.S.M., J.T.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

population. Even though a number of large-scale projects [1–3] have mapped much of the common global human genetic variation, the map of human genetic diversity is far from complete, in particular with respect to rare variants. The recent explosive population growth has resulted in an abundance of rare variants [4, 5] which, because of their relatively recent origin [6, 7] tend to be geographically clustered to a higher degree than variants that are common in different parts of the world [8]. Therefore, in order to obtain a better overview of all genetic variation, it is important to sequence all regions and populations. In recent years, several such projects have mapped the genetic variation in national populations, providing important information regarding locally occurring rare variation [9–11].

The recent release of a whole-genome sequenced Swedish national reference population, SweGen, represents an additional such national population [12]. The genetic structure of the Swedish population is characterized by a pronounced genetic difference between the northernmost counties and the rest of Sweden [13, 14]. For example, the genetic distance, as measured by F_{ST}, between the northern and southern parts of Sweden is larger than that between southern Sweden and other populations of Northern European descent [13]. Furthermore, the northernmost counties exhibit an elevated level of homozygosity [13, 14] and an increase in the genetic distance as a function of the geographical distance [14], compared to southern Sweden. In agreement with these observations, Y chromosomal haplotype frequencies also showed that the population in Västerbotten County in northern Sweden is differentiated from that in the southern Swedish counties [15].

Our studies focus on this genetically pronounced northern Swedish region, specifically Västerbotten County. The county is characterized by vast distances between population centers; the geographical footprint is large (54,665 km^2) [16], but the population size is small (268,278) [17]. The demographic history of the county highlights a sparsely populated region with an estimated population of as few as 7,304 individuals in 1571 who mainly inhabited the coastal region [18]. Historically, the inland region has predominantly been inhabited by the indigenous Sami population, but settlement programs were initiated by the state during the mid- to late-17th century with the main part of the in-migration taking place from mid-18th to mid-19th century [19]. It was also in this later period that settlers formed the majority in relation to the Sami population [20]. Previous work based on marriage patterns and distribution of allele frequencies of eight protein markers indicated the presence of a genetic substructure within Västerbotten County [21]. The hypothesized substructure was associated with the major river valleys that stretch from the mountains in the west to the coast in the east in northern Sweden.

In summary, the genetics of the Västerbotten population have been shown to be distinct compared to the rest of Sweden. To improve the likelihood of conducting successful studies on genetic disease in this region, there is a need to acquire a better understanding of the genetic variation in the population. While focusing on establishing a control population to be used in cancer disease studies, we have whole-genome sequenced a population sample named ACpop from the county of Västerbotten. The whole-genome sequenced population, ACpop, provides a high-resolution map of the genetic variation which agrees with earlier studies of protein markers and marriage patterns. Imputation of rare variants in genotyping data from the region is also evaluated.

Materials and methods

Sample selection

The study was approved by the ethical board in Umeå, Sweden (dnr 2014-290-31 and dnr 2017-370-32). The ethical board is a Swedish counterpart to an Institutional Review Board and is governed by the Swedish Ethical Review Authority [22, 23]. Written informed consent was obtained from participants. The experiments were performed in accordance with relevant guidelines and regulations. The 300 samples were selected from the Västerbotten Intervention Programme [24] (VIP) cohort of the Northern Sweden Health and Disease Study (NSHDS). Blood from residents were collected in health care units in each municipality and a donated sample was considered to represent a resident of that municipality. Half of the samples to be sequenced were spread evenly over the 15 municipalities of the county. The remaining half were dedicated to increase the frequency of sequenced individuals in the three most highly populated municipalities (Lycksele, Skellefteå, and Umeå), as well as two municipalities that is located on the border (S1 Fig). An equal number of men and women were selected in each municipality. To maximize the diversity among selected individuals and to minimize selection bias, 27 phenotypic, health, and lifestyle-related variables were extracted from the VIP [25], and a principal component (PC) model was used to select the individuals to be sequenced from each municipality. Principal component (PC) models were calculated separately for each gender and municipality. Samples to be sequenced were selected from the two first PCs according to a full factorial design in two levels with one center point. For municipalities where sample size was substantially increased (Lycksele, Umeå, and Skellefteå), we extended the experimental design with another full factorial design around each of the corner points, as well as with another four center points [25]. For the municipalities where sample size was moderately increased (Storuman and Malå), we extended the experimental design simply by adding another full factorial design in two levels and one center point but angled 45 degrees to the original one.

Whole genome sequencing, raw data processing and calling of variants

DNA samples were obtained from the NSHDS and paired-end sequenced (2x150 bp) to a depth of at least 30x using the Illumina HiSeqX system at NGI-U (Uppsala, Sweden). Samples underwent a quality check by the sequencing facility. PCR-free library preparation kits were used for all samples.

Raw data processing was performed in accordance with the GATK best practices [26, 27]. Briefly, reads were aligned to the 1000g fasta reference (b37) using BWA (v0.7.10-r789) [28] Sorting, indexing, and marking of duplicates was done using Picard (v.1.118) [29], and realignment around indels was done using GATK (v3.3.0) [30]. Qualimap (v2.0.2) [31] was used to assure sample quality and to identify any deviating samples. SNPs and small indels were called for each sample separately using HaplotypeCaller (GATK v.3.3.0). The resulting gVCFs were then jointly genotyped into a single VCF file using GenotypeGVCFs (GATK v. 3.3.0). The called variants were quality filtered in accordance with GATK recommendations using VQSR. Guided by the Ti/Tv ratio, the truth sensitivity cutoff was set to 99.7 and 99.0 for SNPs and small indels, respectively. Additionally, variants were annotated with the result from a Hardy-Weinberg equilibrium (HWE) test. Variants with a Phred-scaled p-value above 60 ($p \leq 10^{-6}$) were removed in all analyses.

Annotation of variants

Functional annotation of novel variants, as defined by not being present in dbSNP v.150, was performed with the Variant Effect Predictor (VEP, v. 82) [32]. Combined Annotation Dependent Depletion (CADD) scores for the novel variants were calculated using the CADD software (v. 1.6) [33, 34].

Cryptic relations analysis

The software KING [35] was used to screen for unknown pairwise relationships between all samples in ACpop. KING was run with the *kinship* flag and using all biallelic SNPs with a successful genotyping rate of at least 99.9%. Pairwise relationships were inferred using the kinship coefficient as suggested by [35].

PCA

Both the joint (ACpop + 1000g populations) and the ACpop PCA models were produced using smartpca [36] (v. 13050) of the EIGENSOFT (v. 6.0.1) software package. The allowed maximum missing genotype rate was 0.1% and 1% for variants and individuals, respectively. Two full sibling relations (4 individuals) were identified by the relationship analysis, and one individual from each of these relations was excluded prior to PCA. For the joint PCA of ACpop and the 1000 genomes (1000g) project European super population, autosomal biallelic SNPs with a frequency of at least 5% in ACpop and each European subpopulation were used. For the PCA of only ACpop, autosomal biallelic SNPs with a frequency of at least 5% were used. Variants overlapping long-range linkage disequilibrium (LD) regions were excluded as suggested by EIGENSOFT. An LD pruning step was performed before the PCAs using PLINK [37] (v1.90b3) with a window size of 20,000, step size of 2,000 and an r^2 threshold of 0.2. In total, 118,544 variants were used for the joint PCA and 156,680 variants for the ACpop PCA. The 1000g genotypes (release v5 20130502) were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

Imputation

Imputation was performed in two different cohorts. One cohort of individuals from Västerbotten county (n = 500) and one cohort of Swedish individuals (n = 214). Both cohorts have previously been used as control populations in a GWAS of glioma [38], and were genotyped with the Illumina 660W array. All individuals had a call rate >94% and were unrelated (PI-HAT< 0.2). SNPs were filtered based on call rate (>90%), minor allele frequency (>0.01), and HWE ($p>1x10^{-6}$). A/T and G/C SNPs were removed. All subjects were phased together using SHAPEIT [39] and untyped genotypes on chromosome 20 were imputed with IMPUTE2 [40, 41] (v2.3.2) using three different reference populations– 1000g, ACpop, and combination of the two (1000g + ACpop). Imputation was performed on genomic regions less than 5 Mb in size with 1 Mb buffer regions. To create the ACpop reference panel we first created a VCF by jointly calling variants in the ACpop samples and in 64 whole genome sequenced individuals from two families. The families originate from the region. Raw data processing and calling of variants was performed in the same way as was described for ACpop. Phasing was performed with SHAPEIT2 [42], using all biallelic SNPs with a genotyping rate above 1%. SHAPEIT2 was run with the duohmm option active. The number of conditioning states was set to 1000, the window size was set to 300, and the MCMC iterations were set to have 15 burn-in and pruning iterations and 50 main iterations. After phasing, we extracted the haplotypes from the 300 ACpop samples and used them for subsequent imputation analysis. The 1000g reference panel was downloaded from https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html and is based on 2,504 samples from phase 3 (v5 20130502).

FST

Pairwise F_{ST} was calculated using smartpca [36] (v. 13050) from the EIGENSOFT (v. 6.0.1) software package. The same two sibling samples that were excluded in the PCA were also excluded in this analysis. Variants with a missing genotype rate of 0.1% or more were excluded from the analysis. For each pairwise calculation, only autosomal biallelic SNPs that were polymorphic (alternative allele frequency > 0) in both investigated populations were used. Pairwise F_{ST} was calculated between ACpop and the European subpopulations of the 1000g project (release v5 20130502) as well as between regionally defined populations within ACpop.

Inbreeding analysis

Calculation of the inbreeding coefficient for ACpop was performed using FSuite [43] (v. 1.0.4). We followed the procedure outlined in Gazal et al [44] and included the 26 populations of the 1000g project in the analysis. We included autosomal biallelic SNPs with a frequency of at least 5% in each of the 27 populations and which did not deviate from HWE ($p < 10^{-5}$) in any of the 27 populations. The HWE test was performed using PLINK [37, 45] (v1.90b3). Additionally, we removed variants with a missing genotype rate of 5% or more, as suggested by FSuite manual. Prior to calculating the population specific allele frequencies, 14 individuals from 15 reported [44] first-cousin or closer relationships (including one trio) were excluded from the 1000g data (release v5 20130502), as well as the same two ACpop individuals that were excluded in the PCA and F_{ST} analyses due to close relationships. FSuite relies on the creation of several random sparse genome maps (submaps) in order to avoid markers in linkage disequilibrium [46]. We selected markers between recombination hotspots for a total of 100 submaps. The inbreeding coefficient reported by FSuite is the median inbreeding coefficient from the 100 submaps. Pairwise Mann-Whitney U tests [47] were performed using the mannwhitneyu function of the stats module in SciPy [48] (v. 1.4.1).

Runs of Homozygosity (ROH)

PLINK [37] (v1.90b3) was used to screen for runs of homozygous genotypes in all ACpop and 1000g project samples. For each population, autosomal biallelic variants with an allele frequency of at least 5% in the studied population were included. Population-specific allele frequencies were calculated as described in the methods section for the inbreeding analysis. The minimum length of a ROH segment (parameter:—homozyg-kb) was set to 500 kb. All other parameter settings were kept at their default values. Pairwise Mann-Whitney U tests [47] were performed using the mannwhitneyu function of the stats module in SciPy [48] (v. 1.4.1).

Results

Selection of the cancer control samples results in a population that reflects diverse Västerbotten origins

We have whole genome sequenced 300 individuals intended to be used as a control population in genetic disease studies in the Swedish county of Västerbotten and northern Sweden. Based on a previous study [21], it was clear that the control population would benefit from a selection of individuals from all parts of Västerbotten County. The longitudinal and large-scale health study cohort, the Västerbotten Intervention Programme [24] (VIP) cohort of the Northern Sweden Health and Disease Study (NSHDS), provided approximately 95,000 samples from which a selection could be made. Through the VIP, individuals across Västerbotten were invited to undergo a health examination, complete a questionnaire on health and lifestyle, and donate a blood sample for research. Furthermore, our primary focus was to use the control population for cancer studies, thus only individuals who had reached at least 80 years of age in 2014 without being diagnosed with cancer were included from the VIP cohort. This resulted in 2079 women and 1423 men from which the final selection of 150 women and 150 men was made (Fig 1A). The included individuals had donated their blood samples at the latest in 1995 and were at least 60 years of age when making the donation.

The municipality of birth records for all residents in Västerbotten County in 1995 who were born 1934 and earlier were utilized to investigate the origins of the population (available in the Linnaeus database, Centre for Demography and Ageing, Umeå University). An analysis of the population demography revealed that 82% of the Västerbotten population aged 60 years

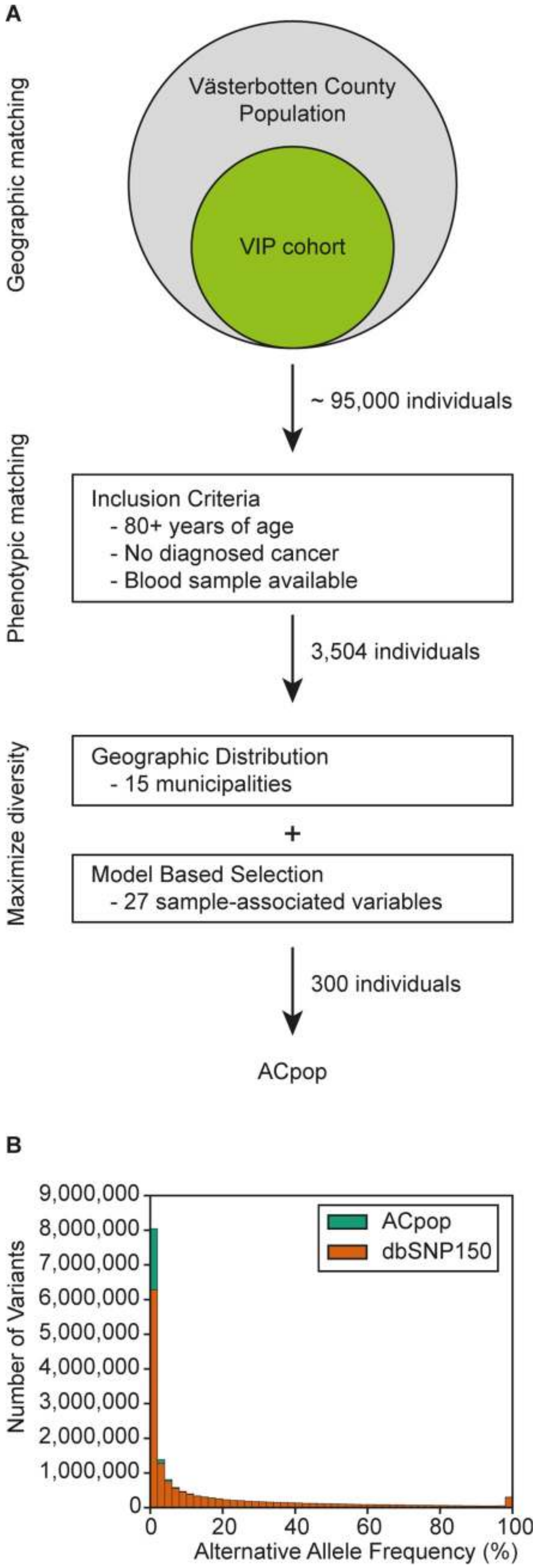
A
Geographic matching
Västerbotten County Population
VIP cohort
~ 95,000 individuals
Phenotypic matching
Inclusion Criteria
- 80+ years of age
- No diagnosed cancer
- Blood sample available
3,504 individuals
Maximize diversity
Geographic Distribution
- 15 municipalities
+
Model Based Selection
- 27 sample-associated variables
300 individuals
ACpop
B
Number of Variants
9,000,000
8,000,000
7,000,000
6,000,000
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
ACpop
dbSNP150
0
20
40
60
80
100
Alternative Allele Frequency (%)

Fig 1. Selection process and allele frequency distribution. The initial aim of ACpop was to sequence a geographically and phenotypically matched cancer control population for Västerbotten County. The selection was made using three steps; selection of cohort from which individuals were identified, inclusion criteria, and maximization of diversity. (**A**). The alternative allele frequency distribution of the variants contained in ACpop, stratified over 50 bins of 2% each. The share of the variants in each bin that does not overlap with dbSNP v. 150 is indicated by green, while the opposite is indicated by orange (**B**).

https://doi.org/10.1371/journal.pone.0237721.g001

or older in 1995 was born in Västerbotten, and 59.4% of the population still resided in the same municipality where they were born (Table 1). Only 10.3% of the population had migrated from a neighboring county, and 7.7% of the population came from other parts of Sweden (further south) or were immigrants. The inflow is expected to increase with successive cohorts. When comparing to a more recent cohort of 60 years and older in 2013 it is found that the share of the population originating from other parts of Sweden and immigrants have risen to 13% and 11% respectively. Mobility into Västerbotten has been low in the 1995 cohort but mobility out from Västerbotten was higher as 35% of those born in the county lived outside Västerbotten in 1995. Based on the low mobility, the 300 individuals were selected from all 15 municipalities in Västerbotten (S1 Fig) to obtain a sample of the genetic variation from all parts of the county. A selection strategy was employed for each municipality to maximize diversity and avoid random selection bias. In short, the strategy consisted of using principal component analysis (PCA) to summarize 27 sample-associated features in conjunction with design of experiments.

Overview of the ACpop variant dataset

The final call set comprised 17,344,482 variants, distributed over 16,522,463 variant sites, after quality control. Of these variants, 14,513,111 and 2,831,371 were SNPs and indels, respectively. A large number of variants (6,176,006) were uncommon, having an allele frequency (AF) of below 1%, and roughly half of these (2,941,762) were seen only once (singletons, allele count of 1) (Fig 1B). As sequencing efforts continue around the world, the total number of discovered genetic variants increases. The most comprehensive collection of genetic variants is dbSNP, which provides updated releases on a regular basis. The latest release of dbSNP (v. 150) contains a total of over 325.7 M variants, more than double the size of the previous release (v. 149). We compared the ACpop variant set against the latest release of dbSNP (v. 150) and found that a total of 2,022,713 variants were not represented in the collection (Table 2). Even though many (938,209) of these unique variants were singletons, as many as 471,744 and 115,682 variants had an AF $\geq$ 1% and AF $\geq$ 5%, respectively. A summary of functional effect annotations and deleterious scores for the novel variants are available in S1 Table, while a

Table 1. Origins of the Västerbotten population aged 60 or older in 1995.

Geographical region of birth	Men (%)	Women (%)	Total (%)
Västerbotten County, same municipality	15,281 (63.8)	16,556 (55.9)	31,837 (59.4)
Västerbotten County, other municipality	4,861 (20.3)	7,221 (24.4)	12,082 (22.6)
Neighboring counties	2,236 (9.3)	3,287 (11.1)	5,523 (10.3)
Non-neighboring counties in Sweden	1,117 (4.7)	1,311 (4.4)	2,428 (4.5)
Other country	455 (1.9)	1,234 (4.2)	1,689 (3.2)
Total (%)	**23,950 (100)**	**29,609 (100)**	**53,559 (100)**

The origin (geographical region of birth) of individuals that were living in Västerbotten County in the year 1995 and that were aged 60 or more at the time.

https://doi.org/10.1371/journal.pone.0237721.t001

Table 2. Summary table of the ACpop dataset.

		ACpop
Number of samples		300
Minimum sequencing depth		30X
Number of variants		17,344,482
Average no. of variants per genome		4,202,229
Novel variants (%)		2,022,713 (11.66%)
Average Het/Hom ratio (SNPs/indels)		1.68 (1.54/2.53)
Ts/Tv ratio		2.08
SNPs		
	Total number	14,513,111
	Average per genome	3,435,205
	Novel variants (%)	1,177,731 (8.11%)
	Singletons (allele count = 1)	2,606,782
Indels		
	Total number	2,831,371
	Average per genome	767,024
	Novel variants (%)	844,982 (29.84%)
	Singletons (allele count = 1)	334,980

Novel variants is the number of variants in ACpop that are not present in the latest version of dbSNP (v. 150). Ts/Tv ratio is the ratio of SNP transitions and transversions, and Het/Hom ratio is the ratio between the number of heterozygous and homozygous occurrences of the variants.

https://doi.org/10.1371/journal.pone.0237721.t002

complete record of the novel variant annotations are available at https://github.com/clicumu/ACpop. Compared to the recently released whole genome sequenced Swedish national reference (SweGen), ACpop contains 1,813,725 unique variants, and a substantial number (794,800) of variants that are common in ACpop (AF > 1%) but are not represented or rare (AF ≤ 1%) in SweGen. The average genome in ACpop consists of 4.20 M variants (3.98 M sites), where 3.44 M are SNPs and 0.77 M are indels, and carries 6,742 variants not found in dbSNP (v. 150).

Subregional genetic structure of ACpop

In a global setting, the genetic variation of ACpop tends to show similarity with other European populations from the 1000g project, as evident by their co-localization within the first and second components of a joint PCA (Fig 2A). To investigate the relationship between ACpop and other European populations in greater detail, a second PCA was performed with only the European subpopulations of the 1000g project. The ACpop samples displayed a distinct separation from members of the other populations in the first two principal components, indicating that there is genetic variation in ACpop that is not captured by sequencing efforts on other European populations (Fig 2B). The internal spread of ACpop in the first and second principal components is in part due to the fact that ACpop is a comparatively large sample in the context of 1000g project European populations. In addition, the pairwise F_{ST} between ACpop and the European populations was calculated. Of the five investigated European populations, ACpop appeared to be genetically closest to the British (GBR) population rather than the geographically neighboring Finnish (FIN) population (Table 3).

It was earlier demonstrated that there is a north-south gradient, where genetic subpopulations are located along the large river valleys, which extends from the mountains in the west to

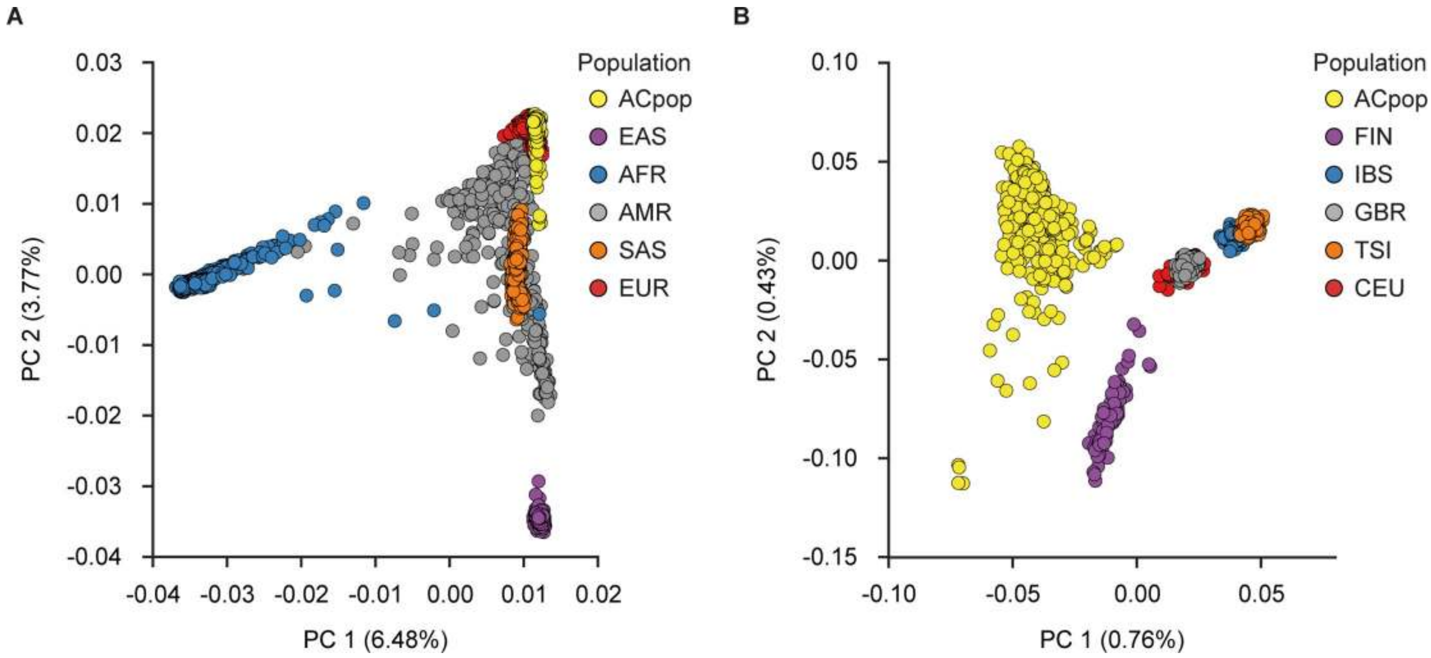

Fig 2. PCA of ACpop and the 1000g. Joint PCA of ACpop and the super populations (EUR: European, EAS: East Asian, AFR: African, AMR: Ad Mixed American, SAS: South Asian) of the 1000g project (**A**). Joint PCA of ACpop and the European populations (FIN: Finnish in Finland, CEU: Utah Residents (CEPH) with Northern and Western European Ancestry, GBR: British in England and Scotland, IBS: Iberian Population in Spain, TSI: Tuscany in Italy) of the 1000g project (**B**). The proportion of the variance explained by each principal component is indicated in the figure axis titles.

https://doi.org/10.1371/journal.pone.0237721.g002

the coast in the east. The colonization of the inland and mountain regions with a relatively small founder population together with our contemporary analysis showing low mobility among individuals born in 1934 (Table 1), led us to ask if there is a genetic difference between the western and eastern populations. Västerbotten County can be divided into three geographical regions, a mountainous region, an inland region, and a coastal region (Fig 3). To further investigate any genetic structure within the county, a PCA was performed with the ACpop samples. By dividing the samples into groups corresponding to the three geographical regions, a subtle pattern emerged in which samples are placed along a gradient going from the coast, through the inland area and to the mountains (Fig 3). The same geographical division was used for calculating pairwise F_{ST}. The intra-county genetic distance between the group of individuals associated with the mountainous region on the one hand, and the group of individuals associated with the coastal region on the other hand, is comparable to that of the genetic

Table 3. Pairwise F_{ST} of European populations.

	ACpop	GBR	CEU	IBS	TSI	FIN
ACpop	--	0.004 (0.00006)	0.004 (0.000055)	0.008 (0.000097)	0.009 (0.000176)	0.006 (0.000076)
GBR		--	0.0 (0.000041)	0.002 (0.000061)	0.004 (0.000109)	0.007 (0.000108)
CEU			--	0.002 (0.000073)	0.004 (0.000134)	0.006 (0.000106)
IBS				--	0.002 (0.000051)	0.01 (0.000139)
TSI					--	0.012 (0.000192)
FIN						--

The European populations are the subpopulations of the European super population of the 1000g project. The standard error of the corresponding F_{ST} value is given in parentheses.

https://doi.org/10.1371/journal.pone.0237721.t003

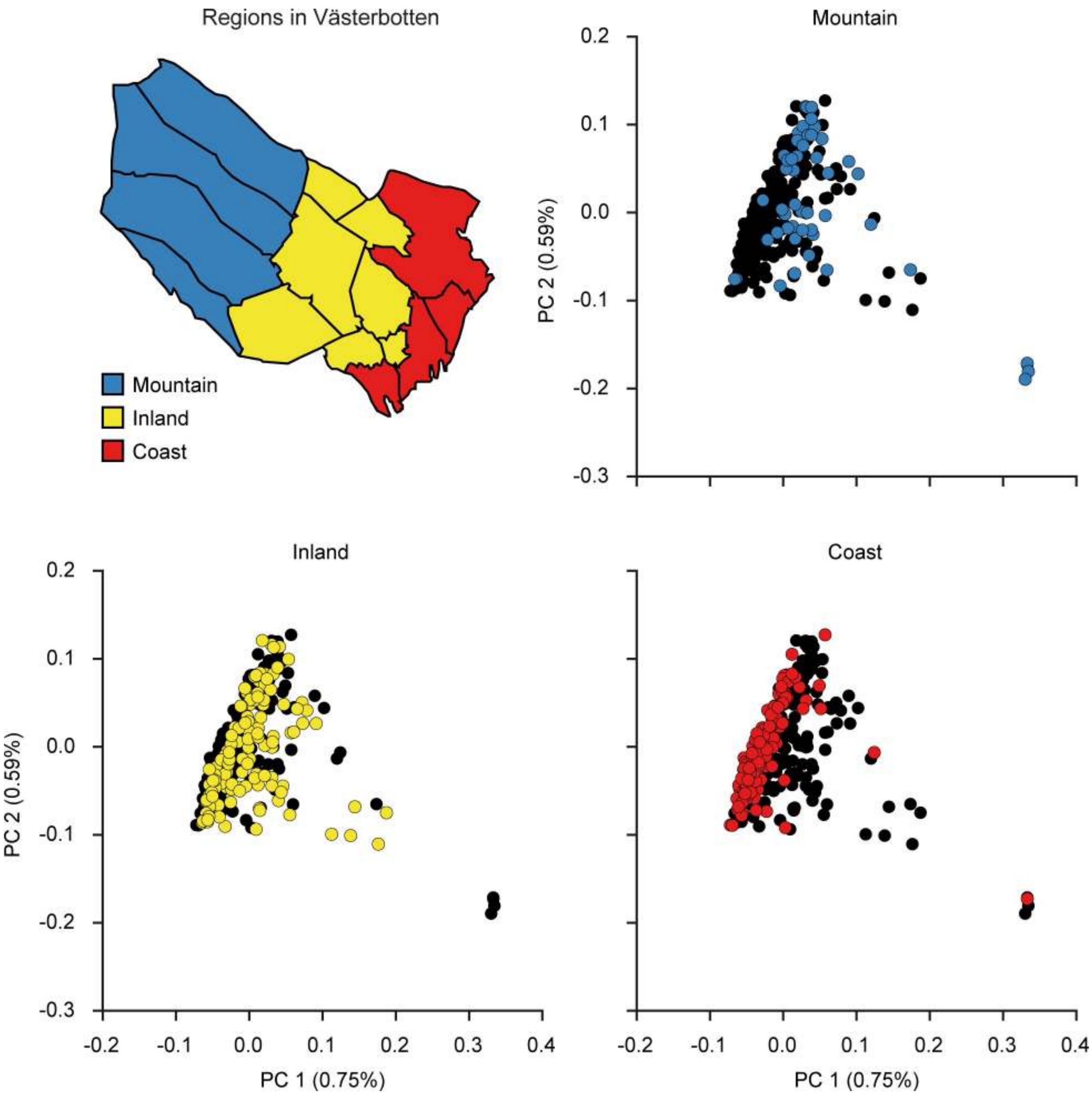

Fig 3. PCA of ACpop samples. The samples were plotted along principal components 1 and 2 and are colored according to three geographical regions in Västerbotten County. Top left panel: a map of Västerbotten County showing the 15 municipalities and the regional subdivision. The proportion of the variance explained by each principal component is indicated in the figure axis titles.

https://doi.org/10.1371/journal.pone.0237721.g003

distance between the British (GBR) and Spanish (IBS) populations in the 1000g project (Table 3 and Table 4). The emerging structure within the county could be further stratified by the municipal association of the samples. Samples tend to cluster according to municipality in the first and second principal components in a way that mimics the geographical associations of the municipalities. That is to say, neighboring municipalities tend to be represented in the PCA by neighboring samples (Fig 4) which correlates with our demographic analysis (Table 1).

Earlier array-based studies of the northern Swedish population have suggested an elevated level of autozygosity [14] and an increased number of homozygous segments [13]. With access

Table 4. Pairwise F_{ST} of Västerbotten subpopulations.

	Coast	Inland	Mountain
Coast	--	0.001 (0.000035)	0.002 (0.000083)
Inland		--	0.001 (0.000067)
Mountain			--

The FST values were calculated pairwise between three geographically defined regions in Västerbotten County, the mountains, inland and coast. Parentheses contain the standard error of the corresponding FST value.

https://doi.org/10.1371/journal.pone.0237721.t004

to a whole-genome sequenced, carefully selected and comparatively large population sample from a focused northern Swedish geographical area, we wanted to provide a better understanding of the existing homozygosity pattern. The degree of autozygosity can be estimated using the inbreeding coefficient *f* which represents the proportion of the genome that is homozygous by descent. The results showed that the average *f* value for ACpop was 0.015, and that the proportion of individuals classified as showing signs of autozygosity ($f > 0.001$) was 67% (S2 Table). The result of the inbreeding analysis stratified over the three regions indicates a gradual increase (non-significant, p>0.05) of *f* from coast to mountain (S3 Table). Furthermore, a ROH analysis showed that ACpop has a higher median total sum of homozygous segments compared to other European populations (panel B in S2 Fig). Both the number of ROH segments and the total length (in base pairs) of ROH segments are higher in ACpop compared to the European populations in the 1000g project (p-values < 0.05, S4 Table, S5 Table).

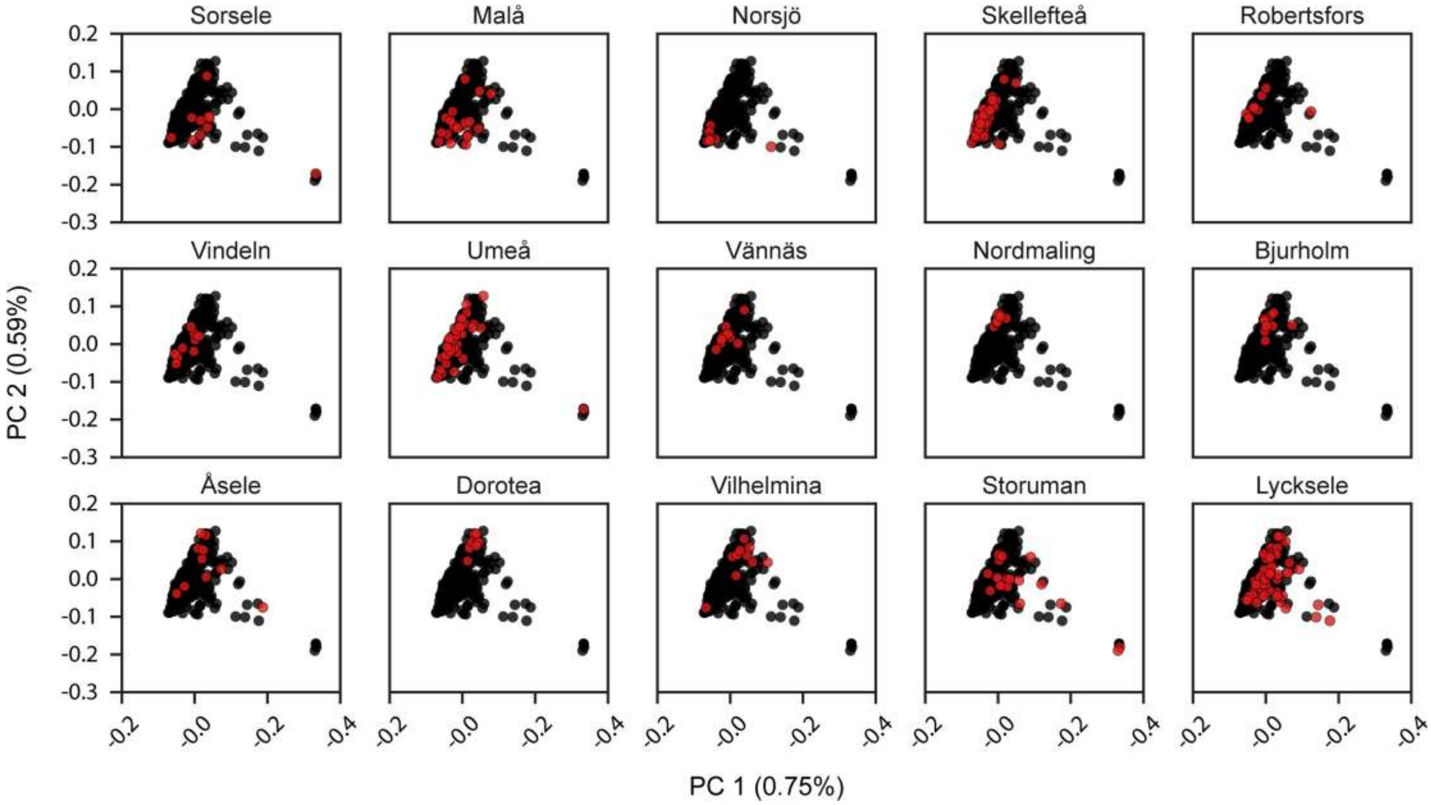

Fig 4. PCA of ACpop samples. The samples were plotted along the first and second components in 15 subplots, each corresponding to one of the municipalities in Västerbotten County. Samples associated with the municipality are colored red, and the rest are colored black. The proportion of the variance explained by each principal component is indicated in the figure axis titles.

https://doi.org/10.1371/journal.pone.0237721.g004

Effect of ACpop in imputation of national and regional cohorts

A whole genome sequenced reference population can be used as a reference panel for imputation in genome-wide association studies (GWAS). Previous studies have shown that population-specific reference panels can improve imputation accuracy in matching cohorts [49, 50]. To investigate the imputation performance of ACpop, we compared the number of variants on chromosome 20 that were imputed with a high *info* score, which is given by the imputation software as a measure of confidence of imputation of individual variants, when using (i) the 1000g reference panel, (ii) the ACpop reference panel, and (iii) a combination of the two reference panels. The number of variants imputed with high confidence (info $\geq$ 0.8) increased by 37% in a cohort from Västerbotten County (Fig 5A) and by 11% in a Swedish national cohort (Fig 5B) when using the combination of the reference panels, compared to using only the 1000g reference panel. For variants with minor allele frequency $\leq$ 5% the corresponding increase was 81% and 23% for the Västerbotten and Swedish cohort, respectively (Fig 5). For variants that were directly genotyped by array technology in the Västerbotten County and Swedish national cohorts, accuracy of imputation was assessed by means of r^2 values (i.e. the correlation between the genotyped and imputed values for a particular variant). Accuracy of imputation of genotyped variants increased with 3.4% in the Västerbotten county cohort (mean r^2 = 0.98) and 1.1% in the Swedish cohort (mean r^2 = 0.95) when using the combination of 1000g and ACpop, compared to using the 1000g panel alone (S3 Fig).

Discussion

In this study, we have presented and characterized ACpop, a whole genome sequenced Västerbotten County population sample of 300 individuals which corresponds to 0.11% of the population. The selection of ACpop aimed in part to include as much as possible of the genetic diversity in Västerbotten County. Because geographic origin is a determining factor of genetic

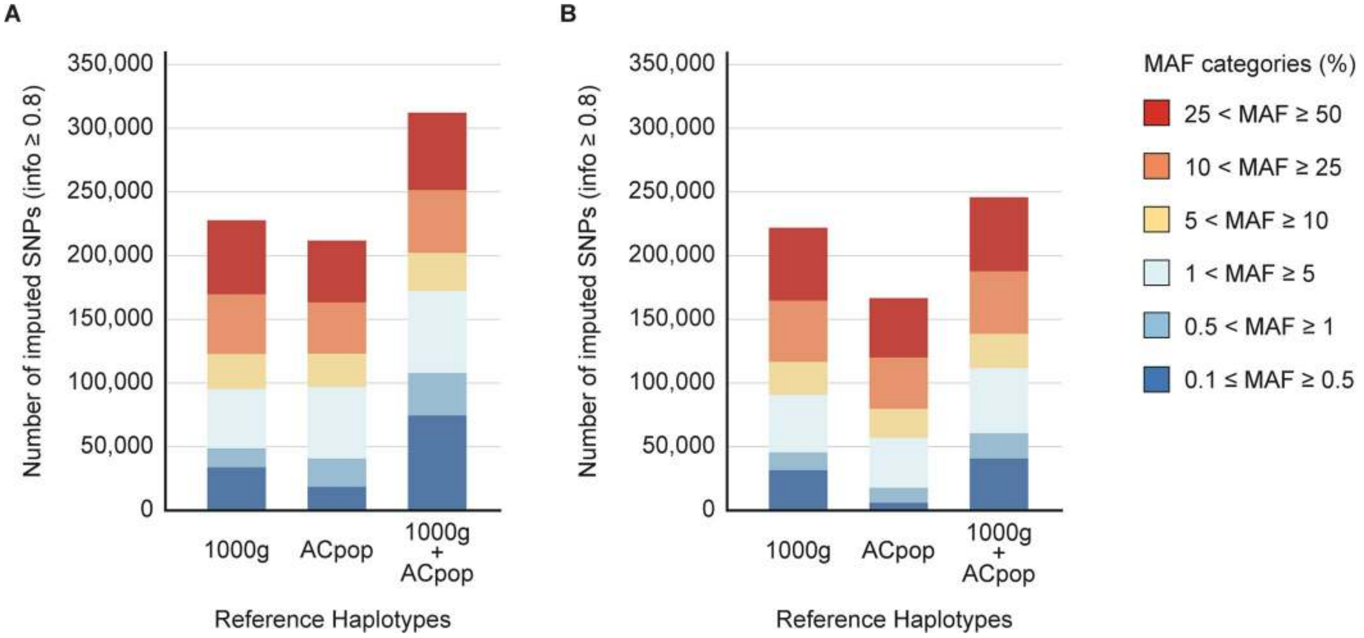

Fig 5. Imputation result. Unknown genotypes on chromosome 20 were imputed using either only the 1000g reference panel, only the ACpop reference panel, or using a combined reference panel from both panels. The number of high quality SNPs (info $\geq$ 0.8), aggregated by minor allele frequency (MAF), that were imputed in a Västerbotten cohort (A) and a Swedish national cohort (B) is indicated.

https://doi.org/10.1371/journal.pone.0237721.g005 variation, a main component in the selection was to ensure a geographical spread across the county. In addition, using statistical design we selected diverse individuals with respect to health and life style associated variables in the biobank. Demographic investigations indicate that the population we selected from was, to a large extent, living in the same municipality in which they were born. Therefore, by basing the selection on the municipal association of the samples, we are confident that ACpop is representative of the genetic variation in Västerbotten. Furthermore, ACpop includes a large number of rare local variants that were not represented in dbSNP or SweGen. In all, ACpop is a resource that provides an unprecedented level of detail of the genetic landscape of Västerbotten County.

Previous studies describing the Swedish genetic landscape have highlighted genetic differences between northern Sweden and both southern Sweden [15] and other European populations [13, 14]. In the PCA that included all of the 1000g project global populations, ACpop showed a clear clustering with the European samples (Fig 2A), as expected. A few samples appear to be drifting toward the South Asia (SAS) population, which could indicate individual-level ancestry from that population. The F_{ST} analysis indicated that ACpop was genetically closer to the CEU and GBR populations than to the FIN population. Although the FIN population is geographically closer to a northern Swedish population such as ACpop, this observation is in line with previous findings using marker data [13, 14]. ACpop demonstrates in this study that not only is there a genetic substructure within Sweden but also within Västerbotten county. This is in line with indications of an elevated positive correlation of genetic and geographic distances demonstrated in a previous array-based study [14]. The intra-county genetic distances between the mountain and coast regions as measured by F_{ST} are comparable to differences between major European populations in the 1000g project (Table 4). This substructure is also supported in the principal component analyses where the two largest components reflect the municipal and regional associations of the samples, respectively (Fig 3, Fig 4). In fact, our results suggest that association studies in Västerbotten might show an improvement if matched controls are taken not only from Västerbotten, but from the same municipality as each affected individual. Similar conclusions were drawn from studies on the Icelandic population with a genealogy that show some similarities to Västerbotten [51]. Iceland was colonized by 8,000–16,000 settlers 1,100 years ago, maintained a population size of about 50,000 around 1850, followed by an expansion to its current size of about 300,000. This resulted in a population with genetic substructures even if the population was assumed to be homogeneous. The comparatively high coefficient of inbreeding for ACpop as a whole might be the result of a small gene pool associated with the small number of founders, in combination with population isolation. In addition, one study [52] that was limited to one coastal municipality suggested an increased occurrence of consanguineous marriages during the 19th century which, if this is representative for the entire county, will influence the inbreeding coefficient of our samples.

Imputation is a method that is often used to increase the number of genetic variants that can be investigated in genetic association studies, and the 1000g is often the reference panel of choice for these analyses. But although the 1000g includes several genetically diverse populations, we found that including ACpop ($n = 300$) together with the 1000g reference panel ($n = 2{,}504$), had a positive impact on imputation performance in Västerbotten samples. The increase of confidently imputed variants was highest for variants with minor allele frequency $\leq 5\%$. Rare variants are underrepresented on the array and in our analyses of imputation accuracy. However, previous studies have shown that the addition of a genetically well-match reference panel increase imputation accuracy also for rare variants [50]. The increase in imputation performance was not as striking in a Swedish national cohort, which again is likely to be the result of the genetic differences between the north and south of Sweden. Imputation performance is expected to increase as haplotypes are added to the imputation reference panel

[53]. However, the modest increase in imputation performance in the Swedish national cohort suggests that the distinct increase in imputation performance in the Västerbotten cohort is, to a large extent, due to the addition of reference samples that are genetically well-matched. Our results suggest that association studies in populations that are not well represented by panels such as 1000g would greatly benefit by the addition of reference subjects for imputation who are drawn from the same geographical region as the studied individuals.

Supporting information

S1 Fig. Geographical overview of Västerbotten County. The location of Västerbotten County in Sweden (grey) (left panel). A map of Västerbotten County with the location of all 15 municipalities and their names (right panel). The number of samples drawn from each municipality is indicated within parentheses below the name of the corresponding municipality.
(JPG)

S2 Fig. Runs of homozygosity (ROH) analysis. The ROH analysis is performed on a per sample-basis. We plot the total length of ROH segments (in kb) against the number of ROH segments for the ACpop samples (A). In (B), the median total ROH lengths for ACpop and the subpopulations of the 1000g project are visualized. The 1000g project subpopulations in (B) are colored according to superpopulation.
(JPG)

S3 Fig. Imputation accuracy. The mean r^2 value for imputed and genotyped variants when using 1) only ACpop, 2) only 1000g, and 3) both ACpop and 1000g as the reference panel. The results are binned according to the expected minor allele frequency (MAF).
(JPG)

S1 Table. Novel variant annotations. Summary of novel variant functional effect annotations from the Variant Effect Predictor (VEP). The consequence types listed are as defined by the Sequence Ontology. The most severe consequences (i.e. at most one consequence per variant) are included. In addition, annotations generated with Combined Annotation Dependent Depletion (CADD) are presented. Specifically, the number of variants with a scaled CADD score above 20 and 30 are presented. A variant with a scaled CADD-score of 20 is predicted to be among the top 1% of most damaging mutations that can be done to the human genome. Conversely, a scaled score of 30 corresponds to the top 0.1% of the most damaging mutations.
(XLSX)

S2 Table. Summary of the inbreeding analysis (ACpop + 1000g). The mean *f* value (column: f_mean) and the proportion of samples that were classified as inbred by FSuite (column: fsuite_inbred) are listed for ACpop and the subpopulations of the 1000g project. Standard deviations are in parentheses. Values in the superpop and pop columns corresponds to the 1000g project superopulation and population codes, respectively.
(XLSX)

S3 Table. Summary of the inbreeding analysis (ACpop regions). The results are presented in total and divided into three geographic regions. The inbreeding coefficient *f* can be considered to be the ratio of the genome of an individual that is homozygous by descent. The average *f* value is the average taken across the individuals in each category. Standard deviations are in parentheses. The inbred status of a sample is the result of testing if *f* is significantly different from 0.001 using a likelihood test.
(XLSX)

S4 Table. Summary of the runs of homozygosity (ROH) analysis. The means and medians of the total number (columns: nseg_median, nseg_mean) and total length of ROH segments (columns: kb_median, kb_mean) are listed for ACpop and the subpopulations of the 1000g project. Standard deviations are in parentheses. Values in the superpop and pop columns corresponds to the 1000g project superopulation and population codes, respectively.
(XLSX)

S5 Table. Summary of p-values. Pairwise Mann-Whitney U tests were performed using three different variables (inbreeding coefficient *f*, total ROH length, and number of ROH segments) between ACpop and the European subpopulations of the 1000g project. The p-values of all pairwise tests are listed in this file. The p-values for variable tested is presented in its own separate sheet.
(XLSX)

Acknowledgments

We are grateful to all participants of the VIP cohort of the Northern Sweden Health and Disease Study (NSHDS), the staff at the Department of Biobank Research, Umeå University. Illumina sequencing was performed by the National Genomics Infrastructure (NGI), which is hosted by SciLifeLab in Stockholm and Uppsala. The computations were performed using resources provided by the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). P.I.O. was supported by a grant from the Knut and Alice Wallenberg Foundation to the Wallenberg Advanced Bioinformatics Infrastructure.

Author Contributions

Conceptualization: Daniel Svensson, Matilda Rentoft, Anna M. Dahlin, Beatrice S. Melin, Johan Trygg, Erik Johansson.

Data curation: Daniel Svensson, Matilda Rentoft, Anna M. Dahlin, Pall I. Olason, Carin Nylander.

Formal analysis: Daniel Svensson, Matilda Rentoft, Anna M. Dahlin, Emma Lundholm, Pall I. Olason.

Funding acquisition: Beatrice S. Melin, Johan Trygg, Erik Johansson.

Investigation: Daniel Svensson, Matilda Rentoft, Anna M. Dahlin, Emma Lundholm.

Methodology: Daniel Svensson, Matilda Rentoft.

Project administration: Carin Nylander, Johan Trygg, Erik Johansson.

Resources: Emma Lundholm, Carin Nylander, Beatrice S. Melin, Johan Trygg.

Supervision: Andreas Sjödin, Johan Trygg, Erik Johansson.

Validation: Daniel Svensson, Matilda Rentoft, Anna M. Dahlin, Emma Lundholm.

Visualization: Matilda Rentoft, Anna M. Dahlin, Emma Lundholm.

Writing – original draft: Daniel Svensson, Matilda Rentoft, Anna M. Dahlin, Emma Lundholm, Beatrice S. Melin, Johan Trygg, Erik Johansson.

Writing – review & editing: Daniel Svensson, Matilda Rentoft, Anna M. Dahlin, Emma Lundholm, Andreas Sjödin, Beatrice S. Melin, Johan Trygg, Erik Johansson.

References

1. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Sep 30; 526(7571):68–74.
2. Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 17; 536(7616):285–91.
3. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 Sep 2; 467(7311):52–8.
4. Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun. 2010 Nov 30; 1 (8):131–6.
5. Keinan A, Clark AG. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. Science. 2012; 336:740–3.
6. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2012 Nov 28; 493(7431):216–20.
7. Nelson MR, Wegmann D, Ehm MG, Kessner D, St. Jean P, Verzilli C, et al. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science. 2012 Jul 6; 337 (6090):100–4.
8. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci. 2011; 108(29):11983–8.
9. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 2014; 46(8):818–25.
10. Gudbjartsson DF, Helgason H, Gudjonsson S a, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015 Mar 25; 47(5):435–44.
11. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun. 2015; 6:1–13.
12. Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, et al. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. Eur J Hum Genet. 2017 Nov 23; 25(11):1253–60.
13. Humphreys K, Grankvist A, Leu M, Hall P, Liu J, Ripatti S, et al. The genetic structure of the Swedish population. PLoS One. 2011; 6(8):e22547.
14. Salmela E, Lappalainen T, Liu J, Sistonen P, Andersen PM, Schreiber S, et al. Swedish population substructure revealed by genome-wide single nucleotide polymorphism data. PLoS One. 2011 Jan 9; 6(2): e16747.
15. Karlsson AO, Wallerström T, Götherström A, Holmlund G. Y-chromosome diversity in Sweden–A long-time perspective. Eur J Hum Genet. 2006 Aug 24; 14(8):963–70.
16. Sweden Statistics. Land and water area 1 January by region and type of area. Year 2012–2020 [Internet]. 2020 [cited 2020 Feb 22]. Available from: http://www.statistikdatabasen.scb.se/pxweb/en/ssd/START__MI__MI0802/Areal2012N/
17. Statistics Sweden. Population 1 November by region, age and sex. Year 2002–2017 [Internet]. 2017 [cited 2018 Jan 19]. Available from: http://www.statistikdatabasen.scb.se/pxweb/en/ssd/START__BE__BE0101__BE0101A/FolkmangdNov/?rxid=4d6d2f0f-c0fc-4ab6-8cd9-77032b5e2ee2
18. Palm LA. Folkmängden i Sveriges socknar och kommuner 1571–1997: med särskild hänsyn till perioden 1571–1751. Göteborg: L. A. Palm; 2000. 199 p.
19. Bylund E. Koloniseringen av Botniaregionen. In: Edlund L-E, Beckman L, editors. Botnia: en nordsvensk region. Höganäs: Bra böcker; 1994. p. 86–98.
20. Sköld P, Axelsson P. The northern population development; colonization and mortality in Swedish Sápmi, 1776–1895. Int J Circumpolar Health. 2008; 67(1):27–42.
21. Einarsdottir E, Egerbladh I, Beckman L, Holmberg D, Escher SA. The genetic population structure of northern Sweden and its implications for mapping genetic diseases. Hereditas. 2007; 144(5):171–80.
22. The Swedish Ethical Review Authority [Internet]. [cited 2020 Jul 13]. Available from: https://etikprovningsmyndigheten.se/
23. The Swedish Ethical Review Authority—Application for ethical approval [Internet]. [cited 2020 Jul 13]. Available from: https://etikprovningsmyndigheten.se/wp-content/uploads/2020/06/00-Ansokan-om-etikprovning-_-Allma_-n-information-och-underskrifter_eng.pdf
24. Norberg M, Blomstedt Y, Lönnberg G, Nyström L, Stenlund H, Wall S, et al. Community participation and sustainability–evidence over 25 years in the Västerbotten Intervention Programme. Glob Health Action. 2012 Dec 17; 5(1):19166.

25. Rentoft M, Svensson D, Sjödin A, Olason PI, Sjöström O, Nylander C, et al. A geographically matched control population efficiently limits the number of candidate disease-causing variants in an unbiased whole-genome analysis. Toft M, editor. PLoS One. 2019 Mar 27; 14(3):e0213350.

26. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May; 43 (5):491–8.

27. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma. 2013; 43(1110):11.10.1–33.

28. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15; 25(14):1754–60.

29. Picard [Internet]. Available from: http://broadinstitute.github.io/picard

30. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep; 20(9):1297–303.

31. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2015 Oct 1; 32(2):btv566.

32. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016 Jun 6; 17(1):122.

33. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar; 46(3):310–5.

34. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019 Jan 8; 47(D1):D886–94.

35. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W. Robust relationship inference in genome-wide association studies. 2010; 26(22):2867–73.

36. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2 (12):2074–93.

37. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4:7.

38. Rajaraman P, Melin BS, Wang Z, McKean-Cowdin R, Michaud DS, Wang SS, et al. Genome-wide association study of glioma and meta-analysis. Hum Genet. 2012 Dec 11; 131(12):1877–88.

39. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011 Dec 4; 9(2):179–81.

40. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009 Jun; 5(6):e1000529.

41. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012 Jul 22; 44(8):955–9.

42. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013; 10(1):5–6.

43. Gazal S, Sahbatou M, Babron MC, Génin E, Leutenegger AL. FSuite: Exploiting inbreeding in dense SNP chip and exome data. Bioinformatics. 2014; 30(13):1940–1.

44. Gazal S, Sahbatou M, Babron M-C, Génin E, Leutenegger A-L. High level of inbreeding in final phase of 1000 Genomes Project. Sci Rep. 2015; 5(1):17453.

45. Graffelman J, Moreno V. The mid p-value in exact tests for Hardy-Weinberg equilibrium. Stat Appl Genet Mol Biol. 2013 Jan 1; 12(4):433–48.

46. Leutenegger AL, Sahbatou M, Gazal S, Cann H, Génin E. Consanguinity around the world: What do the genomic data of the HGDP-CEPH diversity panel tell us? Eur J Hum Genet. 2011; 19(5):583–7.

47. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann Math Stat. 1947 Jul 10; 18(1):50–60.

48. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020 Mar 1; 17(3):261–72.

49. Surakka I, Kristiansson K, Anttila V, Inouye M, Barnes C, Moutsianas L, et al. Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. Genome Res. 2010 Oct; 20(10):1344–51.

50. Mitt M, Kals M, Pärn K, Gabriel SB, Lander ES, Palotie A, et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet. 2017; 2551(10):869–76.

51. Helgason A, Yngvadóttir B, Hrafnkelsson B, Gulcher J, Stefánsson K. An Icelandic example of the impact of population structure on association studies. Nat Genet. 2005 Jan 19; 37(1):90–5.

52. Bittles AH, Egerbladh I. The influence of past endogamy and consanguinity on genetic disorders in northern Sweden. Ann Hum Genet. 2005; 69(5):549–58.

53. Howie B, Marchini J, Stephens M. Genotype Imputation with Thousands of Genomes. G3; Genes| Genomes|Genetics. 2011 Nov 4; 1(6):457–70.