MASTER´S THESIS

Course code: BIO5011 Name: Sebastian L. Berg

GOT STRUCTURE?

Population Genetic Structure and Demographic History of *Calanus hyperboreus*

Date: July 25, 2024 Total number of pages: 42

NORD University

www.nord.no

I. Foreword

It’s funny how quickly time passes once you get absorbed into everyday occurrences. Looking back now, I realize how many people I owe my thanks to for helping me reach this point. My dear Mom and Dad, thank you for nurturing my curiosity from a young age. I still remember the amazement I felt when we lived in New Zealand, seeing the kiwis, blue penguins, and all manners of other beautiful creatures. It was those experiences that ignited my passion for biology. You gifted me not only a telescope to gaze at the night sky and books about dinosaurs to learn about the past, but also the invaluable tools of an open and inquisitive mind.

Trond and Thi, thank you for always being my cornerstones. I know I’ll always have your care and support throughout this journey.

I am deeply grateful to my supervisor, Galice, whose expertise in evolutionary biology shaped my own interests, skills, and research. Irina’s help was invaluable whenever I got stuck. Her extensive and detailed feedback significantly refined my thinking and arguments. Apollo’s guidance on all things bioinformatics. All of it proved indispensable during my analysis. I’ve learned so much under your mentorship these past few years. You have inspired my passion for bioinformatics and genomics; I am truly grateful for that.

To my wonderful friends, thank you for making this journey a joy. I cherish our late-night study sessions in the master’s room and library, the overly intense table tennis breaks, and, most importantly, your continued love and kindness. I couldn’t have done this without your camaraderie.

And to my dear partner, thank you for your love and support. You were my constant through the late nights and the moments of self-doubt, always there, steadfast and unwavering, even as our lives took us to different places.

Thank you all; you have enriched my life and made this achievement possible. I am excited to carry your influence with me as I embark on the next chapter, ever mindful that learning is a lifelong journey. As Isaac Asimov wisely said, *“People think of education as something they can finish.”*

II. Table of Contents

1.0 Abstract

Calanus hyperboreus is a keystone zooplankton species in Arctic food webs. Despite this, the investigation of its population genetic structure and demography has been hindered due to the challenges of sequencing its large and repetitive genome combined with the difficulty of obtaining samples from central Arctic regions. This study used target capture sequencing (TCS) developed for related *C. finmarchicus* to overcome these challenges. Initially, 4,901,092 SNPs were identified. However, after stringent filtering to account for batch effects and to ensure sufficient coverage depth across all individuals, the final dataset consisted of 389 SNPs from 54 individuals across 11 locations in the Atlantic and Arctic oceans. Population genetic structure analyses revealed high levels of gene flow across the species' range, suggesting that *C. hyperboreus* is a single panmictic population. Pairwise Wier-Cockerham F_{ST} comparisons detected subtle but statistically significant differentiation between certain locations. Tajima's *D* and Site Frequency Spectrum (SFS) analysis revealed a skew toward low-frequency variants, suggesting a recent population expansion. These findings have important implications for understanding this keystone zooplankton species' evolutionary history and future.

2.0 Introduction

2.1 Zooplankton and Expected/Observed Genetic Connectivity

Zooplankton are critical to marine ecosystems, providing essential services such as nutrient cycling, carbon flux, and trophic transfer of energy from primary producers to higher-level consumers. Understanding the factors that shape their population genetic structure is essential for predicting their responses to environmental change and for effective population management (Carstensen *et al.*, 2012). Zooplankton species have life stages that drift in ocean currents, leading to expectations of high gene flow and low genetic differentiation. Furthermore, their broad geographic distributions, often spanning ocean basins, suggest significant genetic mixing. However, the high evolutionary potential of zooplankton, resulting from a rapid response to the selection of beneficial mutations given large population sizes (Peijnenburg and Goetze, 2013), challenges such a view. Studies using molecular markers such as mitochondrial DNA and microsatellites have revealed that many zooplankton species exhibit significant genetic structure

(Gómez *et al.*, 2002; Blanco-Bercial and Bucklin, 2016). Isolation by distance or the presence of oceanographic barriers such as major currents, gyres, and fronts can impede dispersal and isolate populations. Life history traits, such as differences in reproductive strategies, can further affect dispersal and gene flow. Additionally, local adaptation to varying environmental conditions, such as temperature, salinity, light penetration, and nutrient availability, can lead to genetic divergence between populations. This interplay of factors highlights the need for further investigation into the population genetics of key zooplankton species (Peijnenburg and Goetze, 2013).

2.2 Ecological Importance of *Calanus*

Copepods of the genus *Calanus*, such as *C. finmarchicus* (Gunnerus, 1770), *C. glacialis* (Jaschnov, 1995), *C. helgolandicus* (Claus, 1863), and *C. hyperboreus* (Krøyer, 1838), dominate the Arctic and Atlantic oceans, contributing up to 80% of zooplanktonic biomass in the Barents Sea (Skjoldal and Aarflot, 2023). These copepods are integral to the trophic networks due to their ability to accumulate and store lipids, primarily wax esters, in internal lipid sacs (Falk-Petersen *et al.*, 2009; Aarflot *et al.*, 2018; Skjoldal and Aarflot, 2023). The lipids fuel essential processes like growth and reproduction (Irigoien, 2004; Falk-Petersen *et al.*, 2009). *Calanus* hatch as nauplii from eggs and pass through five copepodite stages (CI-CV) before adulthood. Nutrient availability, temperature, and photoperiod influence their life cycles (Hansen *et al.*, 2003; Falk-Petersen *et al.*, 2009).

2.3 Population Structure of *Calanus*

Population genetic structure within *Calanus* varies by species. Some *Calanus*, such as the Pacific *Calanus sinicus* (Brodsky, 1962), feature little genetic differentiation and no strong population genetic structure, likely through more substantial gene flow or periodic replacement (Bucklin, 2000; Huang, Liu and Chen, 2014). Previous research into the population genetic structure of *Calanus finmarchicus* and *Calanus glacialis* has yielded conflicting results. Allelic variation at 24 SNPs indicates basin-scale population genetic differentiation of *C. finmarchicus* (Unal and Bucklin, 2010), while high levels of gene flow indicate the western North Atlantic Ocean constitutes a single population (Bucklin and Kocher, 1996), differentiated from samples collected in the Norwegian Sea (Bucklin, Sundt and Dahle, 1996). Microsatellite genotyping revealed no population genetic structure for C. *glacialis* in the fjords of Svalbard, White Sea, and Amundsen

Gulf (Coelho *et al.*, 2016). An investigation using SNPs (*C. finmarchicus*: 34,449 SNPs, *C. glacialis*: 17,035 SNPs) from Target Capture Sequencing (TCS) found no noticeable differentiation in *C. finmarchicus*, while *C. glacialis* exhibited differentiation corresponding to two locations (Isfjord and Skjerstadfjord) in Norwegian fjords (Choquet *et al.*, 2019).

Despite extensive population genetic studies on other *Calanus* species, *C. hyperboreus* remains poorly understood due to challenges in sequencing its large, repetitive genome, coupled with sampling difficulties and historically limited research interest. Developing and applying genetic markers for *C. hyperboreus* is, therefore, crucial to advancing future research.

2.4 Calanus hyperboreus

Calanus hyperboreus (Krøyer, 1838) is one of several *Calanus* species that dominate the Arctic and Northern Atlantic oceans (Carstensen *et al.*, 2012). Recent transcriptomics-based and microsatellite-based phylogeny places *C. hyperboreus* in a monophyletic group (Lizano *et al.*, 2022). As with its relatives such as *C. finmarchicus* and *C. glacialis*, *C. hyperboreus* has adapted to the predictable changes in food availability caused by seasonal algal blooms (Hobbs *et al.*, 2020; Skottene *et al.*, 2020; Kvile, Prokopchuk and Stige, 2022). It is a keystone zooplankton species in Arctic food webs, acting as a vital link between primary producers (phytoplankton) and higher trophic levels, including fish, seabirds, and marine mammals (Falk-Petersen *et al.*, 2009; Wold *et al.*, 2011)

C. hyperboreus is uniquely adapted to thrive in the harsh and highly seasonal Arctic environment. It has a prolonged and variable life cycle compared to its relatives, ranging from two to six years (Scott *et al.*, 2000; Falk-Petersen *et al.*, 2009). This extended life cycle incorporates three overwintering diapause phases during the CIII, CIV, and CV copepodite stages (Falk-Petersen *et al.*, 2009). Diapause length varies depending on environmental conditions such as temperature (Falk-Petersen *et al.*, 2009; Maps, Record and Pershing, 2014). During these diapause periods, *C. hyperboreus* migrates to deeper waters and reduces its metabolic rate, relying on stored lipid reserves for survival (Hirche, 1996; Maps, Record and Pershing, 2014).

The energy reserves also fuel reproduction (Falk-Petersen *et al.*, 2009). *C. hyperboreus* employs a capital breeding strategy at depth (Falk-Petersen *et al.*, 2009), ensuring young nauplii spawn in time to capitalize on the spring phytoplankton bloom. Bioenergetic modeling suggests *C. hyperboreus'* lipid reserves can sustain diapause for over a year, highlighting the significant investment in this breeding approach (Schmid, Maps and Fortier, 2018). Additionally, *C. hyperboreus* females are believed to be iteroparous (Hirche *et al.*, 2024), which is unusual among copepods. The success of *C. hyperboreus* in the Arctic Ocean is linked to the connectivity between the outer continental shelf region and the deep basin, as different life stages of *C. hyperboreus* utilize different habitats (Hirche *et al.*, 2024). The outer shelf region is more productive and favors development, while the deep basin has reduced predation pressure, which benefits adults. The deep basin serves as a potential site for *C. hyperboreus* reproduction. These adaptations, along with larger body size, allow *C. hyperboreus* to cope with the harsh and seasonal Arctic (Broms, Melle and Kaartvedt, 2009; Falk-Petersen *et al.*, 2009).

The ecological significance of *C. hyperboreus* extends beyond its role as a food source. It contributes to nutrient cycling by producing detritus (marine snow) and vertical migration in the water column. Furthermore, *C. hyperboreus* participates in the biological carbon pump by packaging carbon into fecal pellets that sink into the deep ocean, removing carbon from surface waters and sequestering it (Visser, Grønning and Jónasdóttir, 2017). Sloppy feeding can further increase carbon cycling by releasing dissolved organic carbon (DOC) into the water column (Møller, Thor and Nielsen, 2003).

2.5 Molecular Tools in *Calanus* Research

2.5.1 Molecular markers

Population genetic studies of zooplankton rely heavily on molecular markers, polymorphic DNA sequences present in both nuclear and mitochondrial DNA (Anne, 2006; Frías-López *et al.*, 2016).

Mitochondrial DNA (mtDNA) markers are maternally inherited and exhibit a higher mutation rate than nuclear DNA. However, it is important to note that the ratio between nuclear and mitochondrial DNA mutation rates can vary across taxa (Allio *et al.*, 2017). Furthermore, mtDNA can be prone to homoplasy, where identical variants arise from independent evolutionary origins, potentially leading to misleading phylogenetic interpretations if used without complementary nuclear markers (Anne, 2006). In comparison, nuclear DNA (nDNA) markers are biparentally inherited and feature a lower mutation rate than mtDNA markers (Anne, 2006).

Both mtDNA and nDNA can harbor various categories of molecular markers, including microsatellites, single nucleotide polymorphisms (SNPs), and InDels. Microsatellites are useful as molecular markers because of their high variability across individuals. Extensive polymorphism and co-dominant inheritance patterns make them valuable for population genetic analyses. However, it's crucial to consider the potential complications when using them (Choquet *et al.*, 2023). Homoplasy can lead to misidentification. Additionally, null alleles, which do not amplify during PCR, can result in misinterpretation of genotypes. Issues with paralogous loci, which are duplicated in the genome, can lead to amplification of the wrong locus. Locus duplication, where multiple primer pairs target the same locus, can also cause problems. Finally, multiple reverse primer sites within the same locus can amplify multiple PCR products. These complications need to be carefully considered when applying microsatellites in population genetic studies (Selkoe and Toonen, 2006; Choquet *et al.*, 2023). SNPs are single-base-pair variations in DNA sequences that are ubiquitous throughout the genome. High-throughput sequencing technologies facilitate their efficient identification through genotyping-by-sequencing (GBS; Bucklin *et al.*, 2020). SNPs are predominantly biallelic and well-suited for discerning regional to large-scale population structure (Bucklin *et al.*, 2020).

The selection of appropriate molecular markers, whether mitochondrial or nuclear, is guided by the specific research objectives, the availability of genomic resources for the species in question, and considerations of cost-effectiveness associated with genotyping methodologies.

2.5.2 Reduced Representation Sequencing and Target Capture Sequencing

The known genome sizes of *Calanus spp.* are huge and highly repetitive (e.g., *C. hyperboreus,* 12.2 Gb, *C. glacialis,* 11.83 Gb*,* haploid; McLaren et al., 1988). The genome size and complexity introduce significant challenges to sequencing efforts and bioinformatics in genomic studies (Weydmann *et al.*, 2017; Lizano *et al.*, 2022). Next-generation sequencing (NGS) has dramatically improved the field of genomics, population structure, molecular marker development, and whole genome sequencing. However, generating NGS data for large repetitive genomes still often leads to high error rates (Frías-López *et al.*, 2016; Morton *et al.*, 2020). Whole-genome sequencing can also be prohibitively expensive and computationally intensive for species such as *C. hyperboreus*. Due to the small body size of copepods, these methods are often hindered by the limited amount of DNA that can be extracted. Therefore, alternative approaches are necessary to overcome these challenges.

One set of approaches is reduced representation sequencing (RRS). RRS are genomic subsampling procedures such as for example, Target-capture sequencing (TCS), which offers distinct advantages over WGS, especially for phylogenetic studies (Davey *et al.*, 2011). Target-capture sequencing selectively enriches and sequences specific genomic regions of interest in all samples, reducing costs and data complexity (Jones and Good, 2016). Enriching targeted regions is done by designing biotinylated capture baits that hybridize (bind) with the targeted region. The biotin causes the bound sequences to stick to beads while non-hybridized fragments are washed away. This is particularly beneficial for large and repetitive genomes.

The design of TCS probes necessitates prior knowledge of the genomic regions of interest, which may be lacking when sequencing non-model species like those in the *Calanus* genus. In such cases, reference genomes from closely related species, or de novo transcriptome assembly or draft genome assembly, can be used to guide the design of the capture bait (Choquet *et al.*, 2019; Andermann *et al.*, 2020). This thesis uses SNPs called from capture-sequenced nDNA sequences to investigate the population genetic structure of *C. hyperboreus*. Studies using the same capture probes have been performed for *C. finmarchicus* and *C. glacialis* (Choquet *et al.*, 2019).

2.6 Research Question

- What was the efficacy of TCS probes designed for other *Calanus* on *C. hyperboreus?*
- Does *Calanus hyperboreus* exhibit population structure across the sampling range?
- What are the potential drivers of population differentiation?

- What is the demographic history of *C. hyperboreus?*

3.0 Materials and Methods

Target Capture Sequencing data of *Calanus hyperboreus* specimens was pooled from prior work on species boundaries in four *Calanus* species (Choquet *et al.*, 2023). The work was conducted in two batches (a pilot and a full-scale batch) performed by different people. The pair-end sequencing data generated from these batches were then analyzed for this master project. The pilot batch covered four sampling sites where *C. hyperboreus* was present, while the full-scale batch collected *C. hyperboreus* from eleven sites. The sampling area of both included locations across the Atlantic and Arctic oceans, with some overlap (Figure 1). The number of samples per site and coordinates are detailed in Table 1.

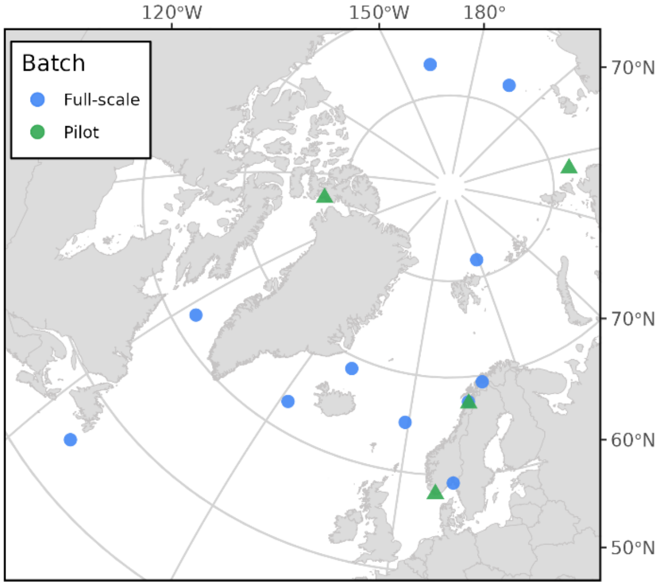


Figure 1. Map of sampling locations across the Northern Hemisphere. Locations from the full-scale and pilot study are colored blue and green respectively.

Table 1. Sampling Locations, abbreviation, sampling location coordinates, batch membership, and number of samples per location.

Location	**Abbreviated**	**Latitude**	**Longitude**	**Study**	***n***
Norwegian Sea	Norwe.	65°03 N	00°51 W	Full-scale	4
East Greenland / North Iceland	E.Gre.	68°48 N	18°23 W	Full-scale	4
Greenland Sea	Green.	62°50 N	28°17 W	Full-scale	4
Labrador Sea	Labra.	62°13 N	57°21 W	Full-scale	5
Balsfjord	Balsf.	69°21 N	19°13 E	Full-scale	4
Chukchi Sea	Chuckc.	76°24 N	162°14 W	Full-scale	5
Mistfjord	Mistf.	67°27 N	14°50 E	Full-scale	4
Oslofjord	Oslof.	59°12 N	10°38 E	Full-scale	4
Off-Quebec	O.Que.	45°05 N	53°44 W	Full-scale	4
North of East-Siberian Sea	E.Sib.	77°27 N	163°58 E	Full-scale	5
Northern Barents Sea	Baren.	81°50'46.32 N	28°47'8.159 E	Full-scale	5
Laptev Sea	Lapte.	78°17'60 N	113°0'36 E	Pilot	5
Lenefjord	Lenef.	58°4'48 N	7°9'42.119 E	Pilot	5
Skjerstadfjord	Skjer.	67°16'45 N	14°53'18.999 E	Pilot	4
West Greenland Sea	W.Gre.	77°51'2 N	77°49'3 W	Pilot	5

3.1 Sampling and DNA Extraction

Library preparation and target capture sequencing followed the methodology described in Choquet *et al.* (2023). Both batches used the same procedure but were performed by two different people. *C. hyperboreus* samples were collected using 150-200 µm vertical tow nets, preserved in 80-90% ethanol, stored at -20°C after 24 hours, and identified using nuclear InDel markers (Smolina *et al.*, 2014). Genomic DNA was extracted using the E.Z.N.A. Insect DNA kit (Omega Bio-Tek).

3.2 Target Capture Probe Design

A set of probes was designed based on a preliminary draft genome assembly of *C. finmarchicus*. This genome-based design used 80-mer probes synthesized by MYcroarray MYbaits, targeting 2,656 unique contigs (2,106,591 bp). The capture efficiency of the probes was previously evaluated for *C. finmarchicus* and *C. glacialis* (38% and 23%, respectively) when mapped directly to the draft assembly (Choquet *et al.*, 2019).

3.3 Target Capture Sequencing

The pilot study prepared DNA libraries using the NEXTflex Rapid Pre-Capture Combo Kit (Bioo Scientific). Individually indexed libraries were then pooled, and two sequence capture reactions were performed following the MYcroarray MYbaits protocol with the modifications described in Choquet *et al*. (2019). Paired-end sequencing of the captured *C. hyperboreus* library pool was done on a NextSeq 500 (Illumina) platform, using NextSeq 500/550 2x150 bp mid-output kits v2.5 (Choquet *et al.*, 2023). Bcl2fastq v1.8.4 (Illumina) was used to demultiplex the sequences.

Libraries for the full-scale study were also prepared using the NEXTflex Rapid Pre-Capture Combo Kit (Bioo Scientific). Four sequence capture reactions were done following the protocol described in Choquet *et al.* (2019). The pooled capture was sent to the Oslo Sequencing Center for sequencing.

3.4 Bioinformatic Pipeline

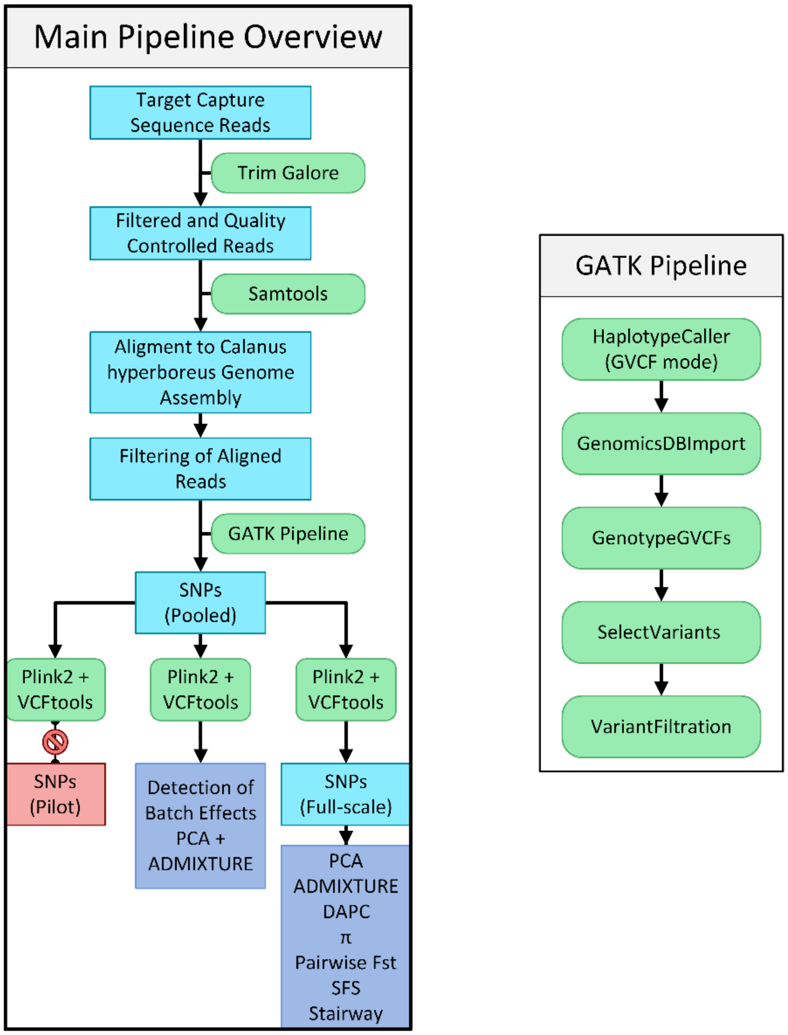


Figure 2. Conceptual framework of the bioinformatic pipeline. Programs are displayed in rounded green boxes, workflow steps as blue rectangles, and analyses as purple rectangles. Population genetic structure analyses and population history analyses were only performed for the full-scale samples. The GATK workflow is illustrated on the right. Please refer to the methodology section below for filter variables and steps.

3.4.1 Read Processing

Adapter removal and quality filtering were performed using Trim Galore v0.6.10 (Krueger, 2024), a wrapper for Cutadapt v1.18 (Martin, 2011) and FastQC v0.11.5 (Andrews, 2010), with default parameters for Illumina sequencing data. This included trimming adapter sequences, removing low-quality bases, discarding short reads (<20 bp), and filtering out reads with low average quality (Phred score < 20). FastQC was used to assess sequencing quality before and after trimming.

3.4.2 Alignment

Trimmed reads were aligned to an unpublished and unannotated *Calanus hyperboreus* genome assembly (7.49 Gb; Choquet *et al*., In prep.), consisting of 130,619 contigs and 1,066 scaffolds using BWA-MEM v1.7 (Li, 2013) with default parameters. The resulting SAM files were converted to BAM format and sorted using Samtools v1.7 (Danecek *et al.*, 2021). Alignments with mapping quality scores below 20 (Q < 20) were filtered out using Samtools view. Additionally, mate-pair information was corrected using Samtools fixmate, and a BWA index was generated for the reference genome assembly to ensure compatibility with Picard tools v3.1.1 ('Picard toolkit', 2019). Duplicate reads were marked using Picard's MarkDuplicates and removed using the AddOrReplaceReadGroups.

3.4.3 Variant Calling and Filtering

Variant calling and filtering were performed on the pooled data. The GATK v4.5.0.0 Best Practices workflow for germline short variant discovery was implemented with modifications based on the pipeline described in Choquet *et al.* (2023).

First, the *C. hyperboreus* reference genome assembly was indexed using Samtools faidx. Then, alignment files (BAM) were validated for integrity using Picard's ValidateSamFile tool. To correct for potential alignment artifacts, reads were split and soft-clipped at InDel positions using the SplitNCigarReads tool from GATK v4.5.0.0 (Auwera and O'Connor, 2020).

Next, raw variant calls were generated for each sample using GATK HaplotypeCaller in GVCF mode. The resulting genomic VCF files (gVCFs) were combined using GenomicsDBImport, and joint genotyping was performed with GenotypeGVCFs to create a single multi-sample VCF file.

Finally, stringent filtering criteria were applied to ensure high-quality variant calls. Variants were filtered using GATK VariantFiltration with hard filter thresholds based on quality metrics (Table 2). These thresholds were derived from an empirical assessment of the data and recommendations from previous studies (Choquet et al., 2023).

Table 2. GATK VariantFiltration values.

Filter	Value
QD	< 4.0
FS	> 60.0
MQ	< 45.0
MQRankSum	< -5.0
ReadPosRankSum	< -5.0

Additional filtering steps were performed to refine the variant dataset further. Variants with genotype missingness rates exceeding 80% and variants with mean read depths below five were filtered out using VCFtools v0.1.15 (Danecek *et al.*, 2011). Furthermore, insertion-deletions (InDels) and non-biallelic variants were removed. Variants with deviations from Hardy-Weinberg equilibrium (HWE; p-value < 0.0001) and variants with strong linkage disequilibrium were removed using PLINK2 v2.0 (Chang *et al.*, 2015; Purcell and Chang, 2017). The differences in sequencing metrics (properly paired reads before and after filtering) between the pilot study and the full-scale study were investigated in R v4.4.1 (R Core Team, 2024) using Mann-Whitney U tests, given the lack of homogeneity of variance and normal distributions (Levene's test; Shapiro-Wilk's test; Fox and Weisberg, 2019).

3.5 Detection of Batch Effects

Principal Component Analysis (PCA) was performed on the pooled samples from the pilot and full-scale batches. Principal components were generated using PLINK2. The first two principal components, explaining the highest proportion of genetic variance, were visualized using the Tidyverse v2.0 (Wickham *et al.*, 2019) and ggplot2 v3.5.1 (Wickham, 2016) R packages.

Admixture analyses were conducted on the pooled dataset using ADMIXTURE v1.3.0 (Alexander, Novembre and Lange, 2009) to infer population structure and potential admixture. A range of ancestral populations (*K*) from 1 to 5 was evaluated, and cross-validation error was used to select the *K* value that best fitted the data. Individual ancestry proportions were plotted using R with the ggplot2 package.

3.6 Full-scale Study

3.6.1 Population Structure

Complementary analyses, including PCA, Admixture analysis, and Discriminant Analysis of Principal Components (DAPC), were used to investigate the population structure and genetic differentiation among *C. hyperboreus* individuals from the full-scale batch.

Principal components were generated from the full-scale batch using PLINK2. Again, the two principal components explaining the highest proportion of variance were visualized using the Tidyverse and ggplot2 R packages. Sampling locations were indicated by color to assess clustering patterns.

Admixture analyses were performed on the full-scale dataset. Ancestral populations (*K*) of 1 to 5 were evaluated, and cross-validation error was calculated for *K*. Individual ancestry proportions were plotted using R with the ggplot2 package.

Discriminant Analysis of Principal Components (DAPC; Jombart, Devillard and Balloux, 2010) was performed using the Adegenet R package v2.1.10 (Jombart, 2008). DAPC combines PCA with Discriminant analysis, maximizing genetic differentiation among groups while minimizing variation within groups. The sampling location was used as the prior group assignment for individuals.

The Mean nucleotide diversity (π) was calculated per sampling location with VCFtools, using a 10kb sliding window to investigate differentiation across the sampled area.

Pairwise Weir and Cockerham's (1984) F_{ST} values between sampling locations were calculated from the full-scale dataset using the Hierfstat v0.5.11 (Goudet *et al.*, 2022) package for R. The pairwise comparisons' p-values were estimated through permutation testing (2,000 permutations) and adjusted using the Bonferroni correction.

3.6.2 Demographic History

To investigate the demographic history of *C. hyperboreus*, the Tajima's *D* statistic was calculated using VCFtools with a 10 kb sliding window. In addition, a folded site frequency spectrum (SFS) was generated from the full-scale dataset SNPs using the vcf2sfs.py python script (Van Rossum and Drake, 2009; Liu *et al.*, 2018) and visualized using R with the ggplot2 package. Site frequency spectrums can provide insights into past demographic events, such as population expansion or bottlenecks, by revealing deviations from the expected allele frequency distribution under neutrality. In the absence of a known recombination rate for *C. hyperboreus,* several values (1.0 × 10^{-7}, 1.0 × 10^{-9}, 1.0 × 10^{-10}, 1.0 × 10^{-11}) were tested and compared, indicating SFS's sensitivity to changes. Using fastsimcoal2 (Excoffier *et al.*, 2021), three population history models (constant population size, instantaneous growth occurring 4,000 generations in the past, and instantaneous decline occurring 4,000 generations in the past) were simulated for comparison with the observed SFS, assuming a 3-year generation time.

Stairway plots were generated for the SFSs with Stairway Plot v2 (Liu and Fu, 2015, 2020), a coalescent-based method that estimates changes in effective population size over time. Due to the lack of a reliable nuclear mutation rate for *C. hyperboreus*, multiple stairway plots were generated assuming different mutation rates based on those observed in *Daphnia* (3.8 × 10^{-9}; Keith *et al.*, 2016) and *Alpheus* (2.64 × 10^{-9}; Silliman *et al.*, 2021). This approach allowed for the sensitivity of the demographic models to variations in mutation rate to be assessed.

4.0 Results

4.1 Descriptive Statistics

An average (mean) of 11,872,927 reads across seventy-three pair-end libraries was retained after adapter trimming and quality control. After alignment, the variance of the number of reads did not differ between the batches (Mann-Whitney U test, W = 589, p-value = 0.3425; Table 3) when excluding the outlier "ind.03". Alignment rates were high for most samples (mean = 70.62%), with a mean of 7,590,615 reads. However, six samples sequenced in the full-scale study had low alignment rates, of which three samples (Norwegian Sea, ind.05; East-Greenland, ind.10; Greenland Sea, ind.14) were excluded from variant downstream analysis. The remaining samples were retained to meet the required number for the Linkage-disequilibrium calculations, as downstream PINK2 eigenvector functions would not allow for less than 50 samples. Removal of duplicated reads removed a substantial portion of the reads (the mean percentage of reads removed by deduplication was 71.143%), specifically in the samples from the full-scale study (89.621% removed). In contrast, a higher proportion (18.62% removed) of reads from the pilot study were retained after deduplication, highlighting the difference between the batches (Mann-Whitney U test, W = 6, p-value = 1.17 × 10^{-10}). The mean percentage of reads retained after applying all filters was 16.35% (per batch; full-scale 5.01%, and pilot 48.60%).

Variant calling of the aligned reads resulted in 4,938,241 variants. After removing InDels and non-biallelic variants and excluding the mentioned three low-alignment samples, 4,901,092 variants were retained. The variants with a mean minimum depth of less than five were excluded, after which 2,382 remained (Supplementary Figure 2). After HWE and Linkage-disequilibrium pruning, a total of 389 SNPs remained.

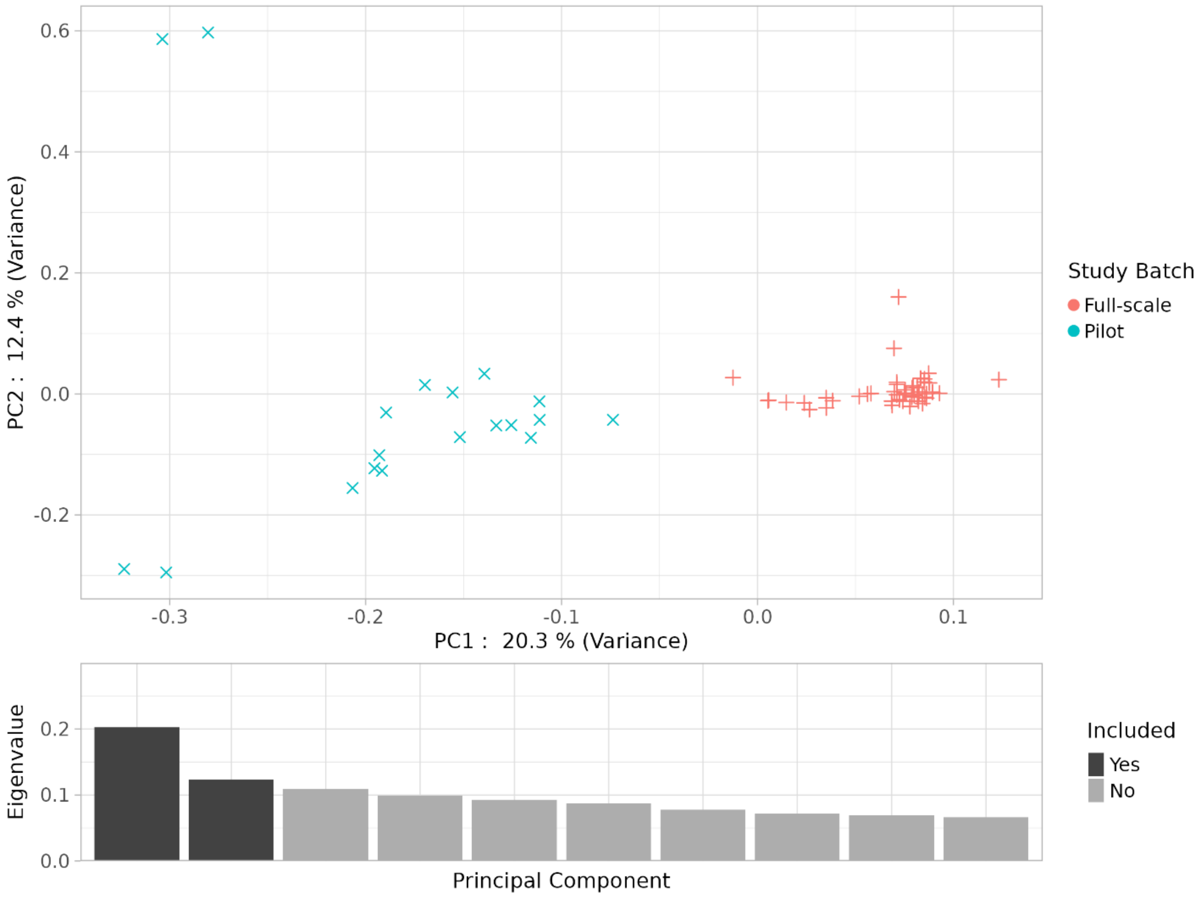


Figure 3. Principal component analysis of the pooled samples. Study batch membership is indicated by color. The eigenvalue per principal component is shown in the bar plot, with included PCs colored.

4.2 Detection of Batch Effects

PCA of the pooled datasets detected two distinct clusters along the first principal component, coinciding with the sample's study batch membership (Figure 3). This grouping was supported by the admixture analysis, where two populations (*K*=2) resulted in the lowest cross-validation error value (Figure 4). The downstream analyses were applied to individuals from the full-scale

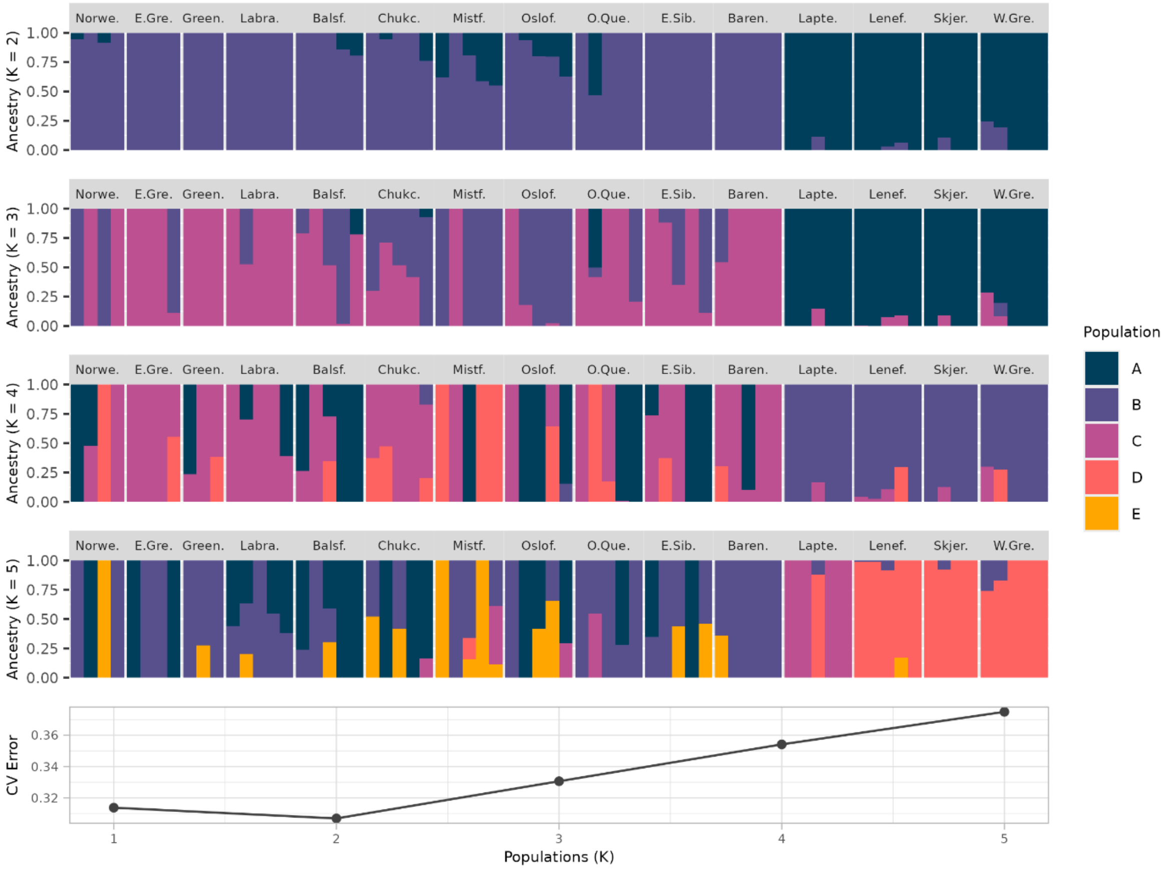


Figure 4. Admixture analysis of the pooled samples with ancestries of 2 to 5 populations (K). The proportions of ancestry memberships per individual are indicated by color. The cross-validation error is indicated for each K-value.

study to mitigate potential biases. The full-scale dataset was chosen for its broader geographic coverage.

4.3 Population Structure

Principal component analysis (PCA) performed on individuals from the full-scale study using SNPs from the combined dataset did not reveal distinct genetic clusters corresponding to sampling locations (Figure 5a). The first two principal components showed substantial overlap among individuals from separate locations.

While DAPC revealed subtle clustering of *C. hyperboreus* individuals, with Chukchi samples clustering by themselves, most of the sampling locations' clusters overlapped (Figure 5b).

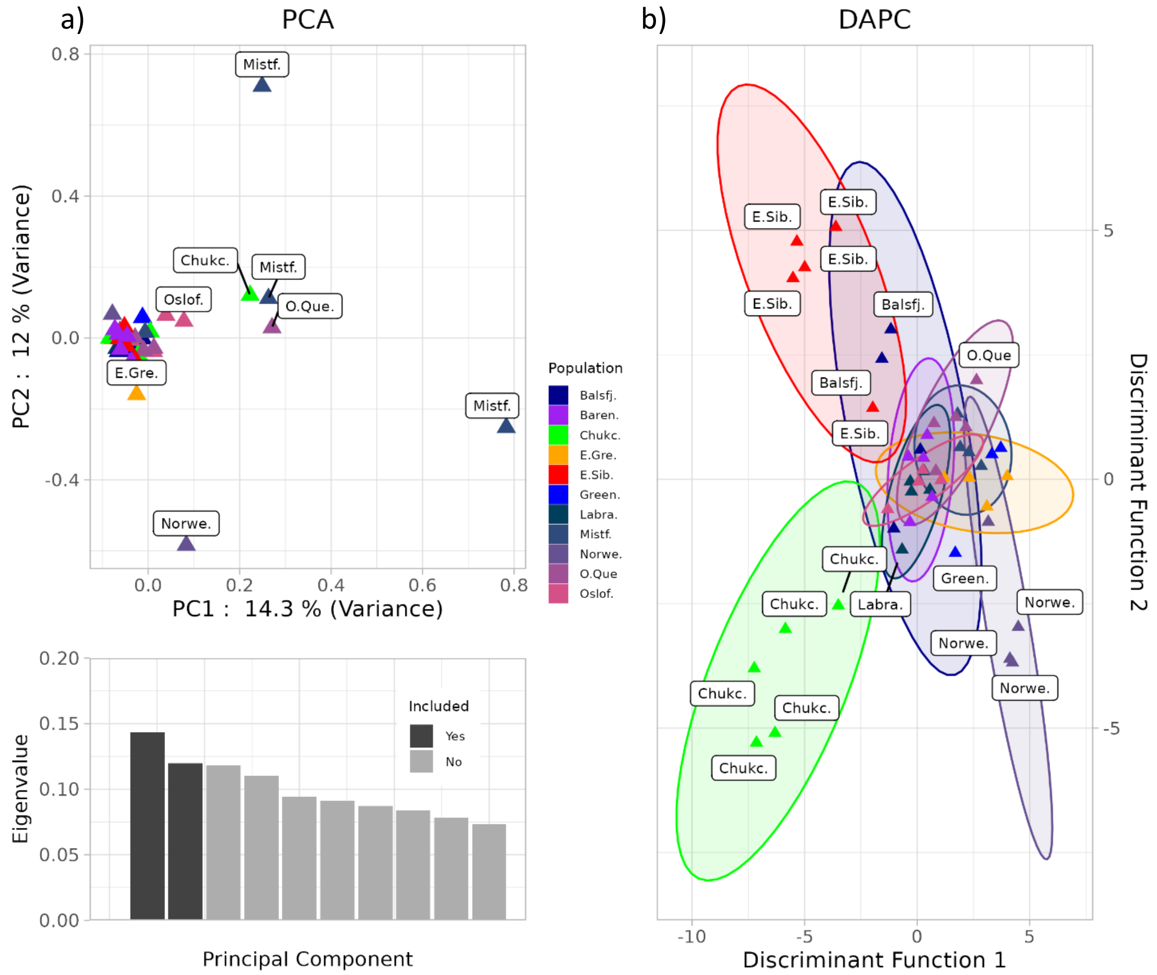


*Figure 5. Principal component analysis of the samples from the full-scale study (**a**) Sampling locations are indicated for points outside the cluster. Sample location membership is indicated by color. The bar plot indicates the eigenvalues of PCs, with included PCs shaded darker. Discriminant analysis of principal components of the full-scale study (**b**). Sampling location membership is indicated by color. Inertial ellipses representing 95% confidence intervals are shown in the respective colors.*

Using cross-validation to assess the optimal number of genetic clusters (*K*), the ADMIXTURE analysis supported a model with *K* = 1 (Figure 6).

Nucleotide diversity (π) was consistent across sampling locations (mean $\pi = 4.68 \times 10^{-5}$; Figure 7b). In contrast, the pairwise F_{ST} (Weir-Cockerham) of the sampling locations revealed weak genetic differentiation between the sites (Figure 8; Table 4).

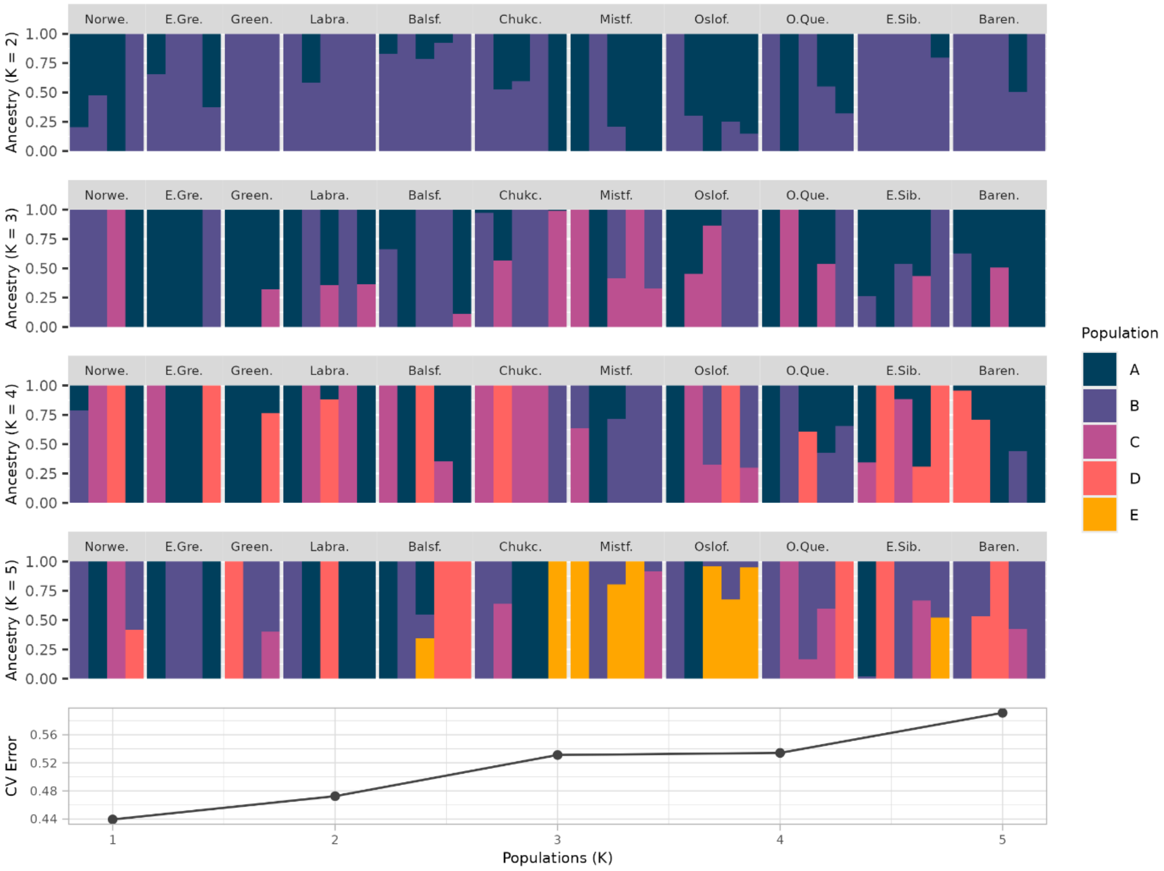


Figure 6. Admixture analysis of full-scale samples for ancestries of 2 to 5 (k). The ancestry proportion per individual is indicated by color. Cross-validation error values for each k-value are indicated.

4.4 Demographic History

The distribution of Tajima's *D* values across 10kb windows for the study was skewed towards negative values (mean Tajima's *D* = -0.9; Figure 7a).

The site frequency spectrum generated from the SNPs featured a similar distribution of derived allele frequencies as in the model simulating instantaneous population growth when using a recombination rate of 1×10^{-9} (Figure 9a).

The stairway plot generated from the observed SNPs followed similar trends to those using the population growth model but differed in the timing of the population increase (Figures 10a and 10d). Models with higher mutation rates (3.8×10^{-9}) shifted the timescale towards the present (Figure 10d, 10e, and 10f).

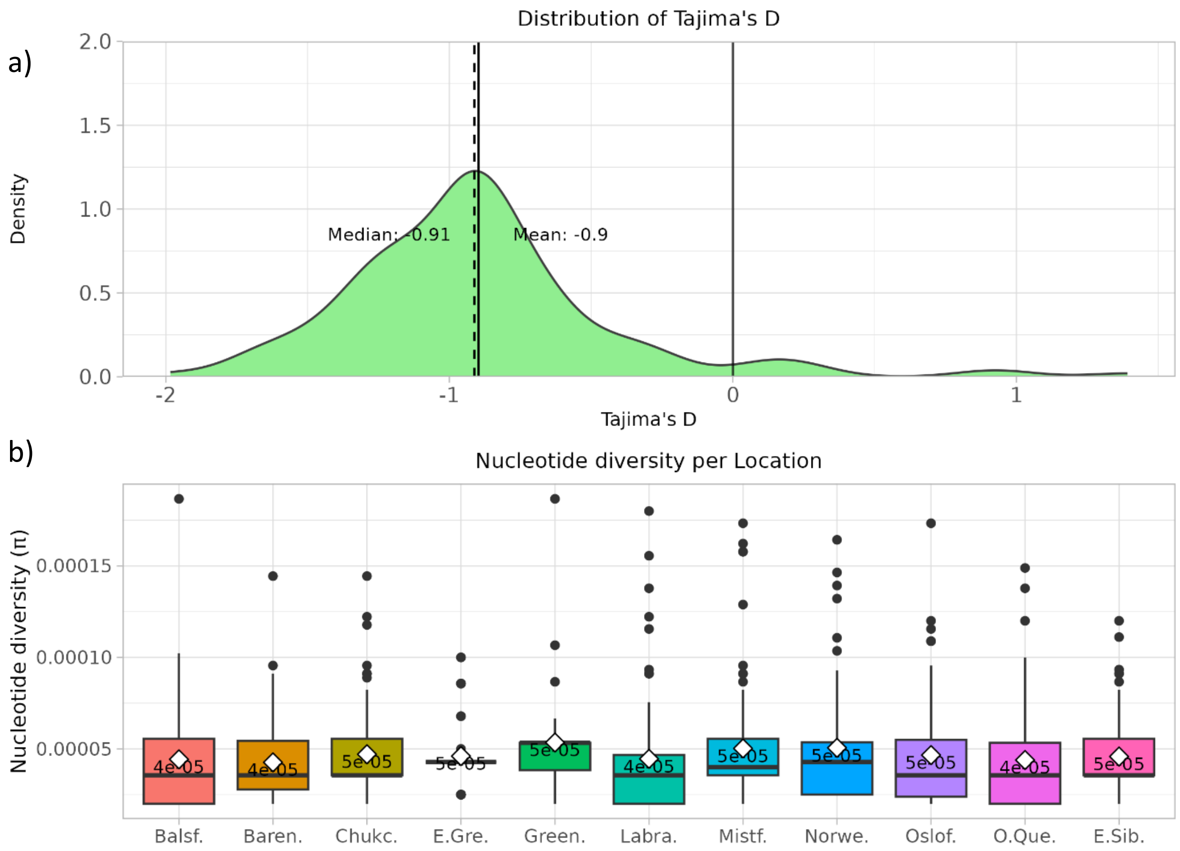


*Figure 7. The distribution of Tajima's D values was calculated using a 10kb sliding window from the full-scale dataset (**a**). Mean and Median values are given and indicated with a solid and dashed line, respectively. Tajima's D = 0 is marked with a black line. Comparison of Nucleotide Diversity (π) per Samling Site calculated using a 10kb sliding window (**b**). Mean values per sampling location are indicated with a white diamond, and mean values are given below.*

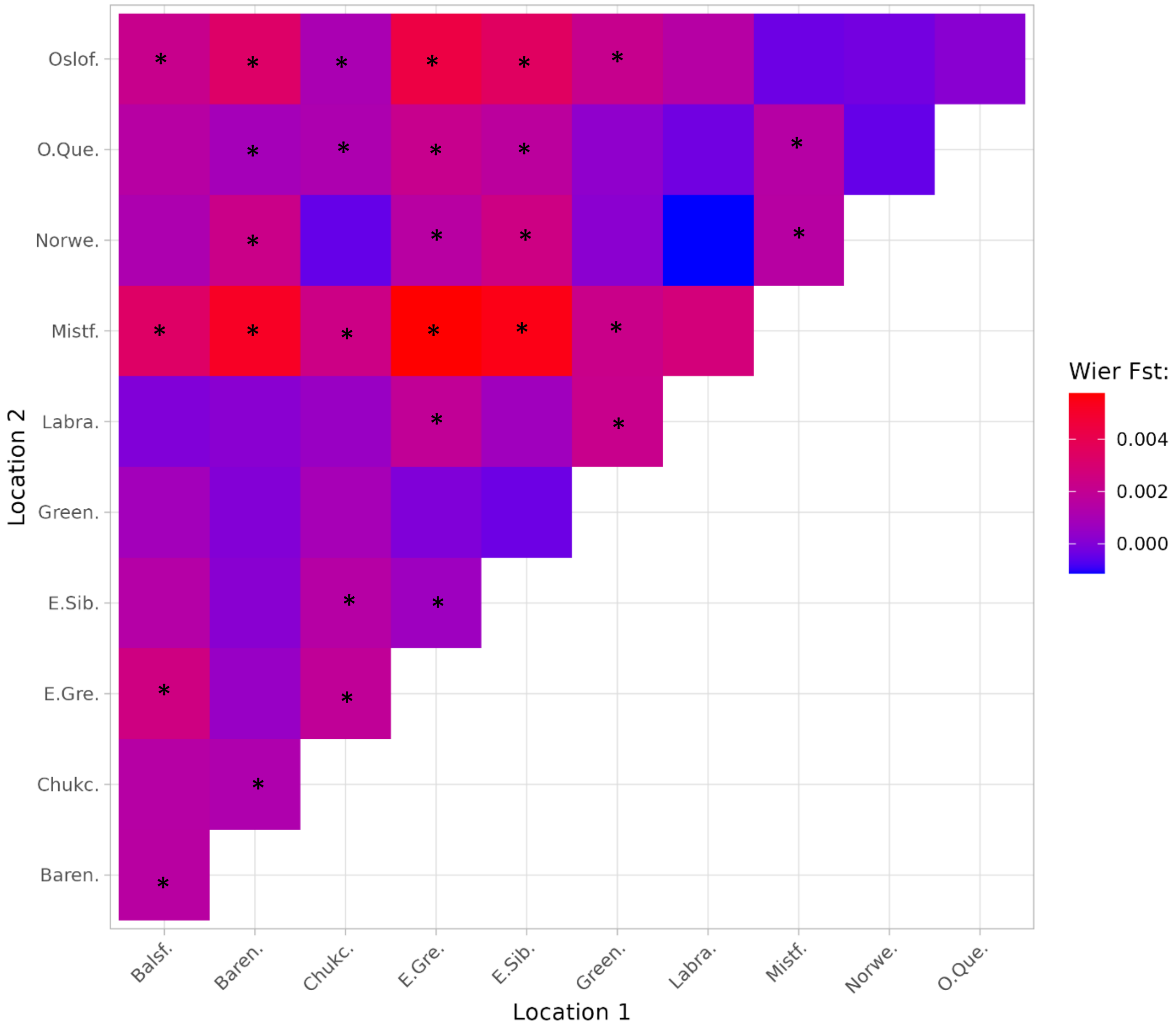


Figure 8. Pairwise Wier-Cockerham F_{ST} Comparisons between sampling locations. Significance ($p < 0.05$) is shown with a star. See Table 4 for F_{ST} values and Bonferroni adjusted p-values estimated through permutation testing ($n = 2,000$).

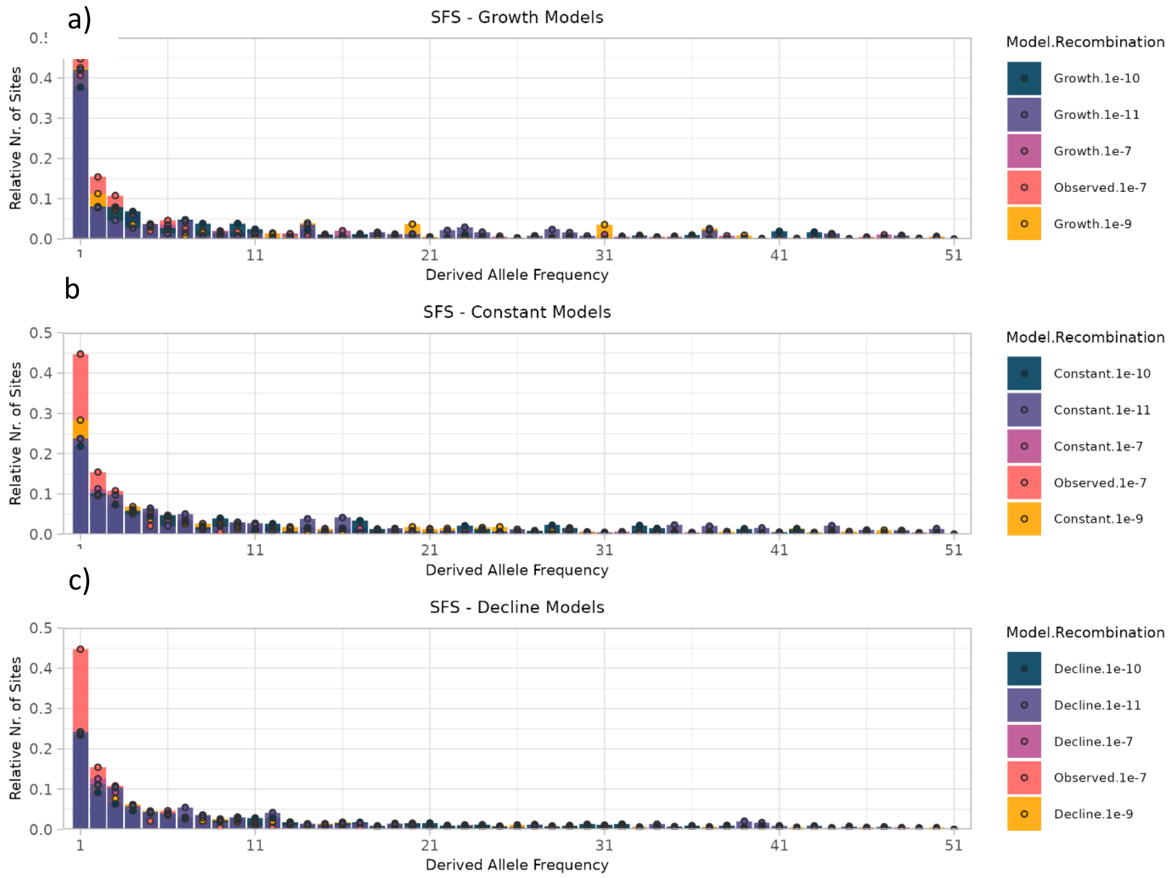


*Figure 9. Site frequency spectrum showing derived allele frequencies calculated from the full-scale dataset, compared SFS generated assuming a growth, constant, and decline models, using several recombination rates (1.0 × 10^{-7}, 1.0 × 10^{-9}, 1.0 × 10^{-10}, and 1.0 × 10^{-11}). The relative number of sites is marked with a black circle. The growth model simulates an instantaneous population growth 4,000 generations ago (**a**). The constant model simulates a stable population (**b**). The decline model simulates an instantaneous population reduction 4,000 generations ago (**c**). Model and recombination rates are indicated by color.*

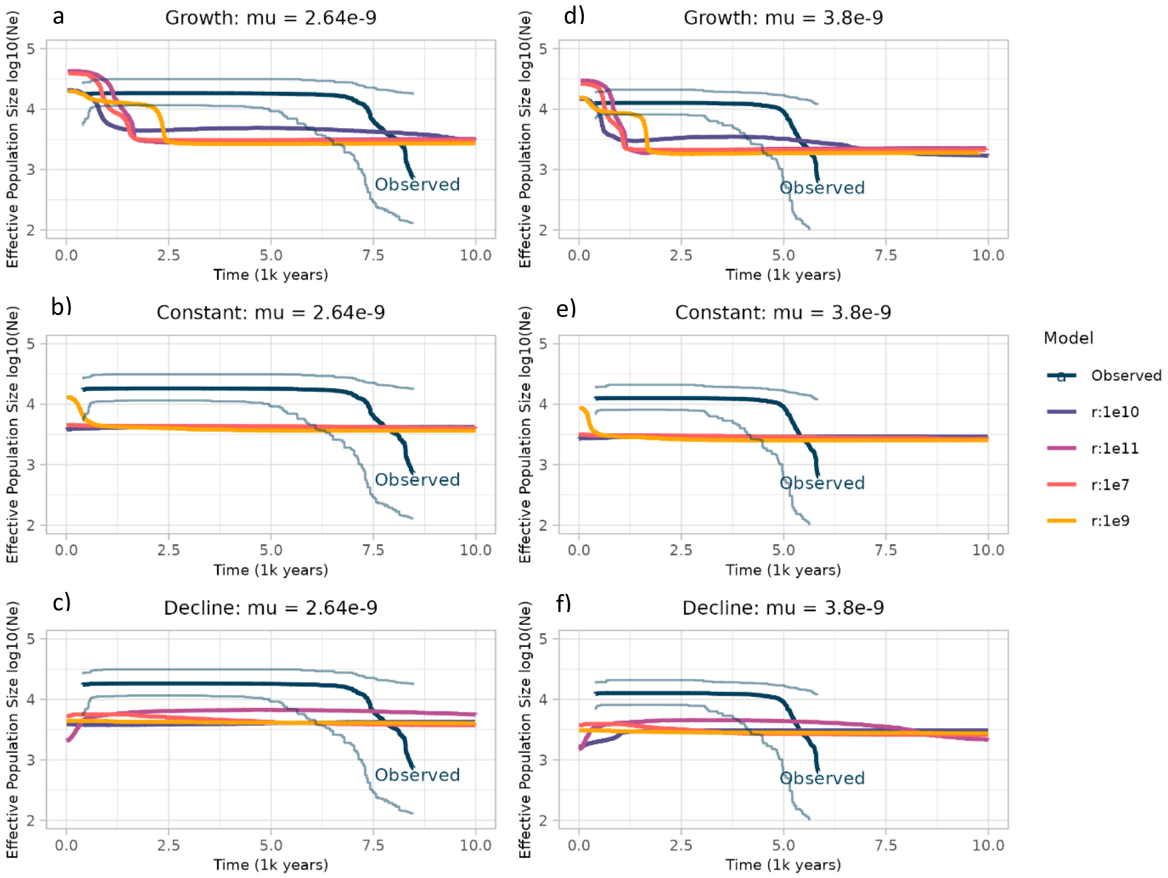


*Figure 10. Stairway plots based on the SFS calculated from the full-scale dataset and simulated population history events using a mutation rate of 2.64 × 10^{-9}, modeling growth (**a**), constant population size (**b**), and population decline (**c**). Stairway plots with mutation rates of 3.8 × 10^{-9}, based on the simulated instant growth 4,000 generations ago (**d**), constant population size(**e**), and instant population decline 4,000 generations ago (**f**). The thin, light blue lines show the 95% confidence interval of the observed trendline.*

5.0 Discussion

5.1 Batch Effects

Batch effects in population genomics arise from technical variations in data generation, potentially confounding analyses, and leading to erroneous conclusions about population structure, genetic diversity, and evolutionary history (Maceda and Lao, 2021). Several factors contribute to these variations, including different sequencing chemistries, read types and lengths, DNA quality, and sequencing depth (Tom *et al.*, 2017; Lou and Therkildsen, 2022). These can lead to systematic differences in base quality scores, alignment accuracy, and representation of genomic regions. Addressing batch effects is crucial and can be achieved through bioinformatic approaches such as read trimming, SNP filtering, PCA visualization, and specialized software like Combat (Tom *et al.*, 2017; Zhang *et al.*, 2018; Lou *et al.*, 2021; Lou and Therkildsen, 2022). Removing samples or batches may be necessary, but this reduces sample size and potentially valuable information.

The sequencing data from the pilot and full-scale datasets were pooled together to improve SNP mining and increase the data's geographical coverage. Population structure analyses of the pooled dataset revealed two distinct clusters. Initial batch-effect-naive PCA revealed strong clustering on the first principal component, coinciding with study batch membership (Figure 3). This was supported by cross-validation error testing, where two clusters (*K* = 2; Figure 4) best represented the data. Admixture analysis (*K* = 2; Figure 4) confirmed this pattern, assigning individuals to their respective study batches.

To investigate the origin of these observed clusters, statistical comparisons of several sequencing metrics were performed: the number of raw reads, the number of proper pair reads, the number of retained reads after deduplication, and the proportion of retained reads after filtering (Table 5). Statistical testing of the sequencing metrics confirmed the presence of batch effects for reads post-deduplication and post-filtering. Namely, a Mann-Whitney U test demonstrated complete separation between batches when assessing the number of reads remaining after deduplication (Table 5). The deduplication step of the bioinformatic pipeline is crucial for removing PCR duplicates, identical copies of DNA fragments that can arise during library preparation or target capture sequencing (Marx, 2017). A higher proportion of duplicates can artificially inflate sequencing depth and potentially skew population genetic parameter estimates. The observed difference in retained reads between the pilot and full-scale studies may stem from several biases. The full-scale study might have had lower amounts of DNA or library input during preparation or capture, respectively, leading to a higher proportion of duplicates. The higher number of individuals pooled in the full-scale study could have influenced duplicate rates due to uneven representation. To avoid such problems in future studies, ensuring sufficient starting DNA (Rochette *et al.*, 2023), optimizing PCR cycles, carefully controlling library input, and strategically planning pooling strategies are recommended. The proportion of reads retained after all filters had been applied also showed complete separation between the two batches (Table 5). These findings align with box-plot visualization of the respective metrics (supplementary Figure 1c and 1d). Similarly, the mean SNP coverage per sample varied by study batch (Supplementary Figure 2b).

The full-scale samples were selected for analysis given their broader geographical representation (11 sampling sites compared to 4; Figure 1), with sites spanning the North Atlantic and the Arctic. Individuals from the pilot study were filtered out of the dataset to mitigate the identified batch effects on downstream analyses. Splitting the dataset reduced the overall sample size and limited the statistical power of the analyses. This trade-off was deemed necessary to minimize the impact of batch effects and ensure their reliability.

5.2 Population Structure

Population genetic structure in copepods can vary dramatically depending on the species and its life history traits. For example, the copepod *Pseudocalanus minutus* in the Okhotsk Sea shows no significant population genetic structure, indicating high gene flow across its range (Hirai, Katakura and Nagai, 2023). This contrasts with its coastal congeners, *P. acuspes* and *P. newmani*, which exhibit strong population structuring (Hirai, Katakura and Nagai, 2023). Similar contrasting patterns have been observed within the *Calanus* genus. *C. finmarchicus*, which inhabits the North Atlantic Ocean, exhibits no significant population genetic structure (Choquet *et al.*, 2019). However, *C. glacialis*, which shares a similar geographic distribution with *C. hyperboreus* in the Arctic, displays distinct population structuring (Choquet *et al.*, 2019; Lizano, 2022). These contrasting patterns highlight the potential for genetic differentiation within Arctic *Calanus* species.

The present study did not reveal any strong population structure in *C. hyperboreus*. The PCA results showed substantial overlap between individuals from different sampling locations, forming a single, undifferentiated cluster (Figure 5a). This suggests high gene flow across the sampled range of *C. hyperboreus*, as supported by the lowest cross-validation error when assuming a single genetic population (*K* = 1; Figure 6). Similarly, nucleotide diversity (π) was consistent across the sampled locations (Figure 7b).

The contrasting findings between *C. hyperboreus* and *C. glacialis* could be attributed to differences in their life histories or the geographic distribution of the sampled populations. *C. glacialis* may experience limited gene flow between populations due to geographic barriers or differences in environmental selection pressures, forming distinct genetic groups. In contrast, *C. hyperboreus* may have a more continuous distribution, along with ontogenetic migrations (Hirche *et al.*, 2024), providing sufficient connectivity to limit differentiation.

While no distinct populations were identified, subtle genetic differentiation within the sampled locations was found. Discriminant Analysis of Principal Components (DAPC) revealed a tendency for individuals to cluster more closely with others from the same sampling location, particularly evident for the Chukchi Sea samples (Figure 5b). This suggests that subtle genetic differences may exist between locations, likely due to factors like geographic distance or localized adaptation. The pairwise F_{ST} comparison supports this notion by detecting subtle but statistically significant differentiation between certain locations (Figure 8 and Table 4).

This pattern of weak differentiation with subtle localized differences suggests that while dispersal and gene flow are dominant forces shaping the genetic structure of *C. hyperboreus*, factors such as geographic distance or localized adaptation may be contributing to divergence

5.3 Demographic History

Analysis of Tajima's *D* statistic and the Site Frequency Spectrum (SFS) revealed intriguing patterns in the demographic history of *C. hyperboreus*. Tajima's *D* exhibited a predominately negative distribution across the contigs, with a -0.91 median value (Figure 7a). This indicates an excess of rare alleles compared to what would be expected under neutral evolution, suggesting possible influences of recent population expansion or selective pressures. While Tajima's *D* can be affected by factors like recombination, the SFS provides further support for this interpretation. The SFS derived from the SNP data exhibited a skew toward low-frequency variants (e.g., singletons and doubletons; Figure 9), a pattern often associated with demographic events like expansions or bottlenecks. Notably, the observed SFS displayed a more pronounced skew than the SFS generated under the instantaneous population growth model (Figure 9a), suggesting a more complex demographic history.

This pattern could be explained by several scenarios, including a recent population expansion of *C. hyperboreus*, a population bottleneck followed by recovery, or positive selection acting on rare variants. However, positive selection is considered less likely given that the SNPs were filtered to remove those showing significant deviations from Hardy-Weinberg equilibrium and linkage disequilibrium, as these deviations can be indicative of selective processes. A recent population expansion of *C. hyperboreus* is a compelling explanation for the observed genetic patterns, especially considering that the closely related *C. glacialis* may have undergone a post-glacial expansion approximately 10,000 years ago (Weydmann *et al.*, 2018). Given their shared distribution across the Arctic and North Atlantic Oceans, it is plausible that *C. hyperboreus* experienced a parallel demographic expansion, driven by increased habitat availability following the retreat of glacial ice. This expansion could have resulted in an increase in genetic diversity and a skew towards rare alleles, as observed in the data.

In contrast, *C. finmarchicus* appears to have maintained stable effective population sizes through various climate events, likely by shifting its distribution to more habitable areas (Provan *et al.*, 2008). This difference in demographic responses highlights the varying strategies that copepods have employed to cope with environmental change.

Future studies could explore alternative demographic models, such as exponential growth or models incorporating migration, to further investigate the demographic history of *C. hyperboreus*.

5.4 Limitations

This study has several limitations: First, the TCS capture baits were developed for *C. finmarchicus* and are less effective for *C. hyperboreus* due to its phylogenetic distance and repetitive genome. This results in the low capture rates observed in *C. hyperboreus*, leading to low sequencing depth and fewer genomic locations represented. Indeed, target capture performance is lower for *C. hyperboreus* than other *Calanus* (see supporting information in Choquet *et al.*, 2023). Consequently, this likely led to the underrepresentation of unique genomic regions or the overrepresentation of conserved ones, potentially explaining the low number of SNPs retained after filtering. Furthermore, the limited population sample sizes, coupled with the low number of SNPs, hindered the detection of subtle population structure differences. While SNPs generally require fewer samples than other markers for population structure analysis, achieving accurate estimations of parameters like F_{st}, especially with smaller sample sizes, necessitates a sufficiently high number of SNPs (Willing, Dreyer and Oosterhout, 2012; Nazareno *et al.*, 2017). A higher SNP density provides a more comprehensive representation of the genome and increases the power to detect subtle genetic variations between populations. However, the combined challenges of Arctic sampling and low capture rates in this study resulted in a low number of SNPs. This highlights the importance of balancing sample size and marker density for robust population genetic inferences, particularly in challenging study systems. Finally, inferences about the population history of *C. hyperboreus* are limited by the lack of a reliable mutation rate and recombination rate. The rate of recombination is highly variable between genomic regions (Wilfert, Gadau and Schmid-Hempel, 2007; Peñalba and Wolf, 2020). The TCS probed targeted 5' UTRs, which tend to have lower recombination rates (Hasan and Ness, 2020). While the data suggest a recent population expansion, further research with empirically determined values for these rates is needed to pinpoint the precise timing and magnitude of this event. Future studies should prioritize increased sampling per site and broader geographic coverage. Additionally, higher capture efficiency resulting in more SNPs will improve resolution and reveal finer-scale patterns.

5.5 Implications for Conservation and Management

This study provides insights into the population structure and dynamics of *C. hyperboreus*, offering valuable implications for conservation and management strategies. The lack of strong population structure suggests high genetic connectivity across the species' range, which could enhance the resilience of the entire Arctic food web to environmental change. However, the subtle genetic differentiation between locations, particularly evident in the Chukchi Sea samples, highlights the potential for local adaptation and the need for region-specific conservation efforts.

The observed high gene flow indicates that environmental pressure in one area could affect populations across a broader region. Therefore, continued monitoring of *C. hyperboreus* populations is crucial, especially given the uncertainties surrounding its demographic history. This monitoring can provide early warning signs of potential population declines and help assess the impact of environmental change. Conservation efforts should consider the broader ecosystem, recognizing the vital role of *C. hyperboreus* in the Arctic food web. Protecting its habitat and ensuring the health of phytoplankton populations are essential. The findings of this study contribute to our understanding of *C. hyperboreus* and can inform effective conservation strategies to ensure its long-term health and genetic diversity.

6.0 Tables

Table 3. Alignment and filtering statistics

Sample	Study	Reads Pairs	Trimmed Read Pairs	Proper Pairs Aligned	Proper Pairs Aligned (%)	Pre-deduplication	Post-deduplication	Post-deduplication (%)	Post-filter (%)
01-155	Full-scale	4.99E+06	4.37E+06	3.21E+06	73.31	2.26E+06	2.62E+05	11.59	6.00

02-155	Full-scale	1.10E+07	9.65E+06	6.72E+06	69.66	4.77E+06	6.33E+05	13.27	6.56
03-155	Full-scale	1.25E+08	1.08E+08	5.63E+07	52.28	3.10E+07	2.74E+06	8.84	2.55
04-155	Full-scale	3.80E+06	3.79E+06	9.28E+05	24.50	4.72E+05	2.37E+04	5.02	0.62
05-155	Full-scale	3.69E+06	3.68E+06	4.41E+05	11.99	2.39E+05	1.37E+04	5.73	0.37
06-165	Full-scale	5.45E+06	4.88E+06	3.01E+06	61.63	2.08E+06	2.56E+05	12.33	5.25
07-165	Full-scale	6.05E+06	5.13E+06	3.68E+06	71.84	2.53E+06	3.58E+05	14.16	6.97
08-165	Full-scale	5.35E+05	5.34E+05	1.39E+05	25.99	7.27E+04	3.62E+03	4.98	0.68
09-165	Full-scale	9.49E+06	8.63E+06	4.57E+06	52.96	3.10E+06	3.39E+05	10.91	3.92
10-165	Full-scale	2.83E+06	2.82E+06	6.26E+05	22.16	3.24E+05	1.92E+04	5.93	0.68
11-168	Full-scale	7.65E+06	6.70E+06	4.63E+06	69.08	3.28E+06	4.06E+05	12.38	6.05
12-168	Full-scale	5.89E+06	5.15E+06	3.60E+06	69.85	2.56E+06	3.13E+05	12.25	6.08
13-168	Full-scale	7.22E+06	6.36E+06	4.44E+06	69.88	3.18E+06	4.00E+05	12.57	6.28
14-168	Full-scale	3.23E+06	3.22E+06	8.21E+05	25.53	4.46E+05	1.88E+04	4.23	0.59
15-176	Full-scale	5.64E+06	5.03E+06	3.68E+06	73.28	2.63E+06	2.37E+05	9.02	4.71
16-176	Full-scale	1.64E+07	1.47E+07	1.07E+07	72.83	7.60E+06	6.28E+05	8.27	4.28
17-176	Full-scale	1.10E+07	9.71E+06	7.11E+06	73.29	5.09E+06	4.25E+05	8.36	4.38
18-176	Full-scale	1.34E+07	1.20E+07	8.82E+06	73.20	6.32E+06	5.25E+05	8.29	4.36
19-176	Full-scale	1.69E+07	1.51E+07	1.08E+07	71.43	7.64E+06	6.09E+05	7.96	4.04
20-Bal	Full-scale	9.40E+06	8.42E+06	6.24E+06	74.15	4.55E+06	3.55E+05	7.80	4.21
21-Bal	Full-scale	1.13E+04	1.04E+04	6.76E+03	65.04	5.43E+03	1.01E+03	18.59	9.72
22-Bal	Full-scale	1.51E+07	1.35E+07	9.88E+06	73.20	7.07E+06	6.37E+05	9.02	4.72
23-Bal	Full-scale	2.06E+07	1.83E+07	1.33E+07	72.52	9.53E+06	7.94E+05	8.33	4.33
24-Bal	Full-scale	1.72E+07	1.54E+07	1.11E+07	71.73	7.99E+06	6.86E+05	8.59	4.45
25-Chk	Full-scale	8.79E+06	7.88E+06	5.69E+06	72.22	3.97E+06	2.92E+05	7.35	3.70
26-Chk	Full-scale	1.38E+07	1.22E+07	8.82E+06	72.03	6.29E+06	5.29E+05	8.40	4.32
27-Chk	Full-scale	7.86E+06	6.96E+06	5.15E+06	73.94	3.63E+06	3.12E+05	8.59	4.48
28-Chk	Full-scale	1.64E+07	1.47E+07	1.03E+07	70.25	7.39E+06	6.92E+05	9.36	4.71
29-Chk	Full-scale	2.04E+07	1.87E+07	1.43E+07	76.39	1.05E+07	1.22E+06	11.67	6.56
30-Mis	Full-scale	2.00E+07	1.83E+07	1.46E+07	79.59	1.07E+07	1.34E+06	12.52	7.30
31-Mis	Full-scale	5.12E+03	4.58E+03	3.70E+03	80.70	2.62E+03	1.26E+02	4.81	2.75
32-Mis	Full-scale	1.76E+07	1.61E+07	1.27E+07	78.86	9.36E+06	1.08E+06	11.57	6.72
33-Mis	Full-scale	2.27E+07	2.07E+07	1.62E+07	78.25	1.19E+07	1.40E+06	11.79	6.78
34-Mis	Full-scale	2.27E+07	2.08E+07	1.59E+07	76.64	1.17E+07	1.41E+06	12.02	6.80
35-Osl	Full-scale	4.72E+03	4.33E+03	1.95E+03	44.96	9.52E+02	3.14E+02	32.98	7.26
36-Osl	Full-scale	1.17E+07	1.08E+07	8.42E+06	78.27	6.21E+06	7.49E+05	12.06	6.97
Sample	**Study**	**Reads Pairs**	**Trimmed Read Pairs**	**Proper Pairs Aligned**	**Proper Pairs Aligned (%)**	**Pre-deduplication**	**Post-deduplication**	**Post-deduplication (%)**	**Post-filter (%)**
37-Osl	Full-scale	1.80E+07	1.65E+07	1.30E+07	78.82	9.51E+06	1.14E+06	11.99	6.92
38-Osl	Full-scale	2.08E+07	1.90E+07	1.46E+07	76.49	1.07E+07	1.24E+06	11.62	6.53
39-Osl	Full-scale	1.64E+07	1.50E+07	1.16E+07	77.37	8.60E+06	1.02E+06	11.81	6.79
40-Que	Full-scale	7.43E+04	6.85E+04	5.52E+04	80.63	3.00E+04	4.57E+03	15.20	6.67
41-Que	Full-scale	2.31E+07	2.12E+07	1.61E+07	76.17	1.18E+07	1.41E+06	12.00	6.67

42-Que	Full-scale	1.21E+07	1.09E+07	7.98E+06	73.55	5.74E+06	5.70E+05	9.93	5.25
43-Que	Full-scale	1.27E+07	1.13E+07	8.25E+06	72.95	5.93E+06	5.35E+05	9.01	4.73
44-Que	Full-scale	1.81E+07	1.60E+07	1.17E+07	73.29	8.51E+06	7.87E+05	9.25	4.91
45-Ess	Full-scale	6.10E+06	5.53E+06	4.26E+06	77.02	3.12E+06	3.27E+05	10.46	5.90
46-Ess	Full-scale	6.24E+06	5.60E+06	4.23E+06	75.48	3.03E+06	2.74E+05	9.03	4.89
47-Ess	Full-scale	1.17E+07	1.05E+07	7.93E+06	75.50	5.69E+06	5.33E+05	9.36	5.07
48-Ess	Full-scale	9.71E+06	8.76E+06	6.66E+06	76.05	4.62E+06	4.86E+05	10.52	5.55
49-Ess	Full-scale	1.25E+07	1.12E+07	8.13E+06	72.71	5.85E+06	6.05E+05	10.35	5.41
50-Bar	Full-scale	9.24E+06	8.33E+06	6.05E+06	72.59	4.12E+06	3.72E+05	9.02	4.46
51-Bar	Full-scale	7.18E+06	6.41E+06	4.69E+06	73.15	3.36E+06	3.37E+05	10.01	5.25
52-Bar	Full-scale	7.32E+06	6.52E+06	4.85E+06	74.34	3.48E+06	3.16E+05	9.08	4.85
53-Bar	Full-scale	6.76E+06	6.02E+06	4.32E+06	71.74	3.10E+06	3.22E+05	10.37	5.34
54-Bar	Full-scale	8.94E+06	7.99E+06	5.82E+06	72.81	4.11E+06	4.06E+05	9.87	5.08
Lapt17	Pilot	4.99E+06	4.59E+06	3.73E+06	81.19	2.81E+06	2.26E+06	80.28	49.14
Lapt1	Pilot	8.24E+06	7.56E+06	6.18E+06	81.78	4.65E+06	3.72E+06	79.97	49.21
Lapt2	Pilot	4.57E+06	4.18E+06	3.24E+06	77.68	2.44E+06	1.97E+06	80.52	47.12
Lapt31	Pilot	1.11E+07	1.01E+07	7.69E+06	75.97	5.87E+06	4.75E+06	80.81	46.85
Lapt3	Pilot	1.81E+07	1.66E+07	1.31E+07	78.85	9.90E+06	7.98E+06	80.53	48.15
Lene1	Pilot	7.58E+06	6.95E+06	5.56E+06	80.05	4.22E+06	3.37E+06	79.98	48.54
Lene22	Pilot	6.28E+06	5.71E+06	4.46E+06	78.13	3.38E+06	2.70E+06	80.08	47.38
Lene23	Pilot	6.43E+06	5.85E+06	4.49E+06	76.76	3.37E+06	2.66E+06	79.18	45.56
Lene45	Pilot	7.50E+06	6.84E+06	5.49E+06	80.19	4.13E+06	3.29E+06	79.63	48.04
Lene7	Pilot	4.79E+06	4.38E+06	3.38E+06	77.13	2.55E+06	2.04E+06	80.09	46.58
Skj10	Pilot	1.31E+07	1.20E+07	9.08E+06	75.54	6.96E+06	5.71E+06	81.96	47.46
Skj12	Pilot	1.32E+07	1.21E+07	8.96E+06	74.13	6.88E+06	5.96E+06	86.67	49.35
Skj14	Pilot	1.61E+07	1.49E+07	1.11E+07	74.53	8.36E+06	6.86E+06	81.98	46.02
Skj24	Pilot	6.06E+06	5.55E+06	4.26E+06	76.74	3.25E+06	3.03E+06	93.25	54.53
Wgr15	Pilot	4.86E+06	4.46E+06	3.43E+06	77.07	2.61E+06	2.13E+06	81.53	47.76
Wgr18	Pilot	6.29E+06	5.80E+06	4.83E+06	83.28	3.64E+06	2.91E+06	79.80	50.14
Wgr19	Pilot	8.09E+06	7.43E+06	5.87E+06	79.02	4.43E+06	3.55E+06	80.15	47.85
Wgr21	Pilot	1.20E+07	1.11E+07	9.66E+06	87.38	7.24E+06	5.74E+06	79.34	51.98
Wgr2	Pilot	1.61E+07	1.48E+07	1.27E+07	85.60	9.55E+06	7.68E+06	80.39	51.72

Table 4. Pairwise Weir-Cockerham F_{ST}-values (below diagonal) and Bonferroni adjusted P-values (above diagonal). P-values were estimated by permutation (n = 2,000), shuffling sampling location memberships between the individuals.

	Balsf.	**Baren.**	**Chukc.**	**E.Gre.**	**E.Sib.**	**Green.**	**Labra.**	**Mistf.**	**Norwe.**	**O.Que.**	**Oslof.**
Balsf.		0.0700	0.0850	*0.0100	0.0850	0.0750	0.1649	*0.0250	0.0800	0.0800	*0.0450
Baren.	0.0016		*0.0250	0.0550	0.1149	0.1149	0.1549	*0.0050	*0.0300	*0.0150	*0.0500
Chukc.	0.0015	*0.0012		*0.0100	*0.0150	0.0600	0.1399	*0.0050	0.1799	*0.0300	*0.0300
E.Gre.	*0.0025	0.0005	*0.0019		*0.0450	0.0950	*0.0100	*0.0050	*0.0450	*0.0100	*0.0050
E.Sib.	0.0014	0.0002	*0.0015	*0.0007		0.1249	0.1349	*0.0050	*0.0250	*0.0050	*0.0400
Green.	0.0008	0.0000	0.0010	0.0000	-0.0004		*0.0150	*0.0100	0.1299	0.1099	*0.0150
Labra.	0.0000	0.0002	0.0006	*0.0019	0.0007	*0.0022		0.0550	0.2249	0.1899	0.0950
Mistf.	*0.0034	*0.0052	*0.0024	*0.0058	*0.0055	*0.0023	0.0028		*0.0450	*0.0200	0.1299
Norwe.	0.0012	*0.0023	-0.0005	*0.0016	*0.0025	0.0002	-0.0011	*0.0015		0.1799	0.1649
O.Que.	0.0015	*0.0009	*0.0012	*0.0022	*0.0017	0.0003	-0.0003	*0.0015	-0.0005		0.1749
Oslof.	*0.0022	*0.0033	*0.0011	*0.0044	*0.0035	*0.0022	0.0015	-0.0004	-0.0002	0.0002	

Table 5. Statistical tests comparing the pilot and full-scale batches.

	Levene		Shapiro-Wilks		Man-Whitney U	
Data \ Statistic	*F-statistic*	Pr(>F)	W	p-value	W	p-value
Raw reads	4.3862	0.03985*	0.96148	0.02677*	589	0.3425
Total pairs	4.0644	0.04764*	0.9593	0.02017*	578	0.4174
Post-deduplication	36.788	6.01E-08*	0.72055	2.10E-01*	6	1.92E-10*
Retained pairs (%)	0.3629	0.5488	0.6427	5.94E-12*	0	1.17E-10*

7.0 References

Aarflot, J.M. *et al.* (2018) 'Contribution of Calanus species to the mesozooplankton biomass in the Barents Sea', *ICES Journal of Marine Science*. Edited by D. Fields, 75(7), pp. 2342–2354. Available at: https://doi.org/10.1093/icesjms/fsx221.

Alexander, D.H., Novembre, J. and Lange, K. (2009) 'Fast model-based estimation of ancestry in unrelated individuals', *Genome Research*, 19, pp. 1655–1664. Available at: https://doi.org/10.1101/gr.094052.109.

Allio, R. *et al.* (2017) 'Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker', *Molecular Biology and Evolution*, 34(11), pp. 2762–2772. Available at: https://doi.org/10.1093/molbev/msx197.

Andermann, T. *et al.* (2020) 'A Guide to Carrying Out a Phylogenomic Target Sequence Capture Project', *Frontiers in Genetics*, 10. Available at: https://doi.org/10.3389/fgene.2019.01407.

Andrews, S. (2010) 'FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]'. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Anne, C. (2006) 'Choosing the right molecular genetic markers for studying biodiversity: from molecular evolution to practical aspects', *Genetica*, 127(1), pp. 101–120. Available at: https://doi.org/10.1007/s10709-005-2485-1.

Auwera, G.V. der and O'Connor, B.D. (2020) *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. First edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly.

Blanco-Bercial, L. and Bucklin, A. (2016) 'New view of population genetics of zooplankton: RAD-seq analysis reveals population structure of the North Atlantic planktonic copepod Centropages typicus', *Molecular Ecology*, 25(7), pp. 1566–1580. Available at: https://doi.org/10.1111/mec.13581.

Broms, C., Melle, W. and Kaartvedt, S. (2009) 'Oceanic distribution and life cycle of Calanus species in the Norwegian Sea and adjacent waters', *Deep Sea Research Part II: Topical Studies in Oceanography*, 56(21), pp. 1910–1921. Available at: https://doi.org/10.1016/j.dsr2.2008.11.005.

Bucklin, A. (2000) 'Population genetics of drifting (Calanus spp.) and resident (Acartia clausi) plankton in Norwegian fjords', *Journal of Plankton Research*, 22(7), pp. 1237–1251. Available at: https://doi.org/10.1093/plankt/22.7.1237.

Bucklin, A. *et al.* (2020) 'Population Genomics of Marine Zooplankton', in M.F. Oleksiak and O.P. Rajora (eds) *Population Genomics: Marine Organisms*. Cham: Springer International Publishing, pp. 61–102. Available at: https://doi.org/10.1007/13836_2017_9.

Bucklin, A. and Kocher, T.D. (1996) 'Source regions for recruitment of *Calanus finmarchicus* to Georges Bank: evidence from molecular population genetic analysis of mtDNA', *Deep Sea*

Research Part II: Topical Studies in Oceanography, 43(7), pp. 1665–1681. Available at: https://doi.org/10.1016/S0967-0645(96)00059-8.

Bucklin, A., Sundt, R.C. and Dahle, G. (1996) ‘The population genetics of *Calanus finmarchicus* in the North Atlantic’, *Ophelia*, 44(1–3), pp. 29–45. Available at: https://doi.org/10.1080/00785326.1995.10429837.

Carstensen, J. *et al.* (2012) ‘Effects of environmental conditions on the biomass of Calanus spp. in the Nordic Seas’, *Journal of Plankton Research*, 34(11), pp. 951–966. Available at: https://doi.org/10.1093/plankt/fbs059.

Chang, C.C. *et al.* (2015) ‘Second-generation PLINK: rising to the challenge of larger and richer datasets’, *GigaScience*, 4(1), pp. s13742-015-0047–8. Available at: https://doi.org/10.1186/s13742-015-0047-8.

Choquet, M. *et al.* (2019) ‘Towards population genomics in non-model species with large genomes: a case study of the marine zooplankton Calanus finmarchicus’, *Royal Society open science*, 6(2), p. 180608. Available at: https://doi.org/10.1098/rsos.180608.

Choquet, M. *et al.* (2023) ‘Unmasking microsatellite deceptiveness and debunking hybridization with SNPs in four marine copepod species of Calanus’, *Molecular Ecology*, 32(24), pp. 6854–6873. Available at: https://doi.org/10.1111/mec.17183.

Coelho, N.C. *et al.* (2016) ‘Pan-Arctic population of the keystone copepod Calanus glacialis’, *Polar Biology*, 39(12), pp. 2311–2318. Available at: https://doi.org/10.1007/s00300-016-1898-x.

Danecek, P. *et al.* (2011) ‘The variant call format and VCFtools’, *Bioinformatics*, 27(15), pp. 2156–2158. Available at: https://doi.org/10.1093/bioinformatics/btr330.

Danecek, P. *et al.* (2021) ‘Twelve years of SAMtools and BCFtools’, *GigaScience*, 10(2), p. giab008. Available at: https://doi.org/10.1093/gigascience/giab008.

Davey, J.W. *et al.* (2011) ‘Genome-wide genetic marker discovery and genotyping using next-generation sequencing’, *Nature Reviews Genetics*, 12(7), pp. 499–510. Available at: https://doi.org/10.1038/nrg3012.

Excoffier, L. *et al.* (2021) ‘fastsimcoal2: demographic inference under complex evolutionary scenarios’, *Bioinformatics*, 37(24), pp. 4882–4885. Available at: https://doi.org/10.1093/bioinformatics/btab468.

Falk-Petersen, S. *et al.* (2009) ‘Lipids and life strategy of Arctic *Calanus*’, *Marine Biology Research*, 5(1), pp. 18–39. Available at: https://doi.org/10.1080/17451000802512267.

Fox, J. and Weisberg, S. (2019) *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. Available at: https://www.john-fox.ca/Companion/.

Frías-López, C. *et al.* (2016) ‘DOMINO: development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms’, *Bioinformatics*, 32(24), pp. 3753–3759. Available at: https://doi.org/10.1093/bioinformatics/btw534.

Gómez, A. *et al.* (2002) 'The interplay between colonization history and gene flow in passively dispersing zooplankton: microsatellite analysis of rotifer resting egg banks', *Journal of Evolutionary Biology*, 15(1), pp. 158–171. Available at: https://doi.org/10.1046/j.1420-9101.2002.00368.x.

Goudet, J. *et al.* (2022) 'hierfstat: Estimation and Tests of Hierarchical F-Statistics'. Available at: https://cran.r-project.org/web/packages/hierfstat/index.html (Accessed: 30 October 2024).

Hansen, B.W. *et al.* (2003) 'Differences in life-cycle traits of Calanus finmarchicus originating from 60°N and 69°N, when reared in mesocosms at 69°N', *Marine Biology*, 142(5), pp. 877–893. Available at: https://doi.org/10.1007/s00227-002-1004-5.

Hasan, A.R. and Ness, R.W. (2020) 'Recombination Rate Variation and Infrequent Sex Influence Genetic Diversity in Chlamydomonas reinhardtii', *Genome Biology and Evolution*, 12(4), p. 370. Available at: https://doi.org/10.1093/gbe/evaa057.

Hirai, J., Katakura, S. and Nagai, S. (2023) 'Comparisons of genetic population structures of copepods Pseudocalanus spp. in the Okhotsk Sea: the first record of P. acuspes in coastal waters off Japan', *Marine Biodiversity*, 53(1), p. 12. Available at: https://doi.org/10.1007/s12526-022-01323-y.

Hirche, H.-J. (1996) 'Diapause in the marine copepod, Calanus finmarchicus — A review', *Ophelia*, 44(1–3), pp. 129–143. Available at: https://doi.org/10.1080/00785326.1995.10429843.

Hirche, H.J. *et al.* (2024) 'From fringe to basin: unravelling the survival strategies of Calanus hyperboreus and C. glacialis in the Arctic Ocean', *Marine Ecology Progress Series*, 745, pp. 41–57. Available at: https://doi.org/10.3354/meps14665.

Hobbs, L. *et al.* (2020) 'Eat or Sleep: Availability of Winter Prey Explains Mid-Winter and Spring Activity in an Arctic Calanus Population', *Frontiers in Marine Science*, 7, p. 541564. Available at: https://doi.org/10.3389/fmars.2020.541564.

Huang, Y., Liu, G. and Chen, X. (2014) 'Molecular phylogeography and population genetic structure of the planktonic copepod Calanus sinicus Brodsky in the coastal waters of China', *Acta Oceanologica Sinica*, 33(10), pp. 74–84. Available at: https://doi.org/10.1007/s13131-014-0542-2.

Irigoien, X. (2004) 'Some ideas about the role of lipids in the life cycle of Calanus finmarchicus', *Journal of Plankton Research*, 26(3), pp. 259–263. Available at: https://doi.org/10.1093/plankt/fbh030.

Jombart, T. (2008) 'adegenet: a R package for the multivariate analysis of genetic markers', *Bioinformatics*, 24(11), pp. 1403–1405. Available at: https://doi.org/10.1093/bioinformatics/btn129.

Jombart, T., Devillard, S. and Balloux, F. (2010) ‘Discriminant analysis of principal components: a new method for the analysis of genetically structured populations’, *BMC Genetics*, 11(1), p. 94. Available at: https://doi.org/10.1186/1471-2156-11-94.

Jones, M.R. and Good, J.M. (2016) ‘Targeted capture in evolutionary and ecological genomics’, *Molecular Ecology*, 25(1), pp. 185–202. Available at: https://doi.org/10.1111/mec.13304.

Keith, N. *et al.* (2016) ‘High mutational rates of large-scale duplication and deletion in Daphnia pulex’, *Genome Research*, 26(1), pp. 60–69. Available at: https://doi.org/10.1101/gr.191338.115.

Krueger, F. (2024) ‘FelixKrueger/TrimGalore’. Available at: https://github.com/FelixKrueger/TrimGalore (Accessed: 21 April 2024).

Kvile, K.Ø., Prokopchuk, I.P. and Stige, L.C. (2022) ‘Environmental effects on *Calanus finmarchicus* abundance and depth distribution in the Barents Sea’, *ICES Journal of Marine Science*. Edited by D. Fields, 79(3), pp. 815–828. Available at: https://doi.org/10.1093/icesjms/fsab133.

Li, H. (2013) ‘Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM’. arXiv. Available at: https://doi.org/10.48550/arXiv.1303.3997.

Liu, S. *et al.* (2018) ‘Genomic parallelism and lack thereof in contrasting systems of three-spined sticklebacks’, *Molecular Ecology*, 27(23), pp. 4725–4743. Available at: https://doi.org/10.1111/mec.14782.

Liu, X. and Fu, Y.-X. (2015) ‘Exploring population size changes using SNP frequency spectra’, *Nature Genetics*, 47(5), pp. 555–559. Available at: https://doi.org/10.1038/ng.3254.

Liu, X. and Fu, Y.-X. (2020) ‘Stairway Plot 2: demographic history inference with folded SNP frequency spectra’, *Genome Biology*, 21(1), p. 280. Available at: https://doi.org/10.1186/s13059-020-02196-9.

Lizano, A.M. *et al.* (2022) ‘Insights into the species evolution of *Calanus* copepods in the northern seas revealed by *de novo* transcriptome sequencing’, *Ecology and evolution*, 12(2), p. e8606. Available at: https://doi.org/10.1002/ece3.8606.

Lizano, A.M.D. (2022) ‘Examining challenges in species-level taxonomy among Calanus copepods in the Northern seas using genome and transcriptome data (PhD thesis)’, *Nord University* [Preprint], (45). Available at: https://nordopen.nord.no/nord-xmlui/handle/11250/3030081.

Lou, R.N. *et al.* (2021) ‘A beginner’s guide to low-coverage whole genome sequencing for population genomics’, *Molecular Ecology*, 30(23), pp. 5966–5993. Available at: https://doi.org/10.1111/mec.16077.

Lou, R.N. and Therkildsen, N.O. (2022) ‘Batch effects in population genomic studies with low-coverage whole genome sequencing data: Causes, detection and mitigation’, *Molecular*

Ecology Resources, 22(5), pp. 1678–1692. Available at: https://doi.org/10.1111/1755-0998.13559.

Maceda, I. and Lao, O. (2021) ‘Analysis of the Batch Effect Due to Sequencing Center in Population Statistics Quantifying Rare Events in the 1000 Genomes Project’, *Genes*, 13(1), p. 44. Available at: https://doi.org/10.3390/genes13010044.

Maps, F., Record, N.R. and Pershing, A.J. (2014) ‘A metabolic approach to dormancy in pelagic copepods helps explaining inter- and intra-specific variability in life-history strategies’, *Journal of Plankton Research*, 36(1), pp. 18–30. Available at: https://doi.org/10.1093/plankt/fbt100.

Martin, M. (2011) ‘Cutadapt removes adapter sequences from high-throughput sequencing reads’, *EMBnet.journal*, 17(1), pp. 10–12. Available at: https://doi.org/10.14806/ej.17.1.200.

Marx, V. (2017) ‘How to deduplicate PCR’, *Nature Methods*, 14(5), pp. 473–476. Available at: https://doi.org/10.1038/nmeth.4268.

McLaren, I.A., Sevigny, J.M. and Corkett, C.J. (1988) ‘Body sizes, development rates, and genome sizes among Calanus species’, *Hydrobiologia*, 167(1), pp. 275–284. Available at: https://doi.org/10.1007/BF00026315.

Møller, E.F., Thor, P. and Nielsen, T.G. (2003) ‘Production of DOC by Calanus finmarchicus, C. glacialis and C. hyperboreus through sloppy feeding and leakage from fecal pellets’, *Marine Ecology Progress Series*, 262, pp. 185–191. Available at: https://doi.org/10.3354/meps262185.

Morton, E.A. *et al.* (2020) ‘Challenges and Approaches to Genotyping Repetitive DNA’, *G3 Genes|Genomes|Genetics*, 10(1), pp. 417–430. Available at: https://doi.org/10.1534/g3.119.400771.

Nazareno, A.G. *et al.* (2017) ‘Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species’, *Molecular Ecology Resources*, 17(6), pp. 1136–1147. Available at: https://doi.org/10.1111/1755-0998.12654.

Peijnenburg, K.T.C.A. and Goetze, E. (2013) ‘High evolutionary potential of marine zooplankton’, *Ecology and Evolution*, 3(8), pp. 2765–2781. Available at: https://doi.org/10.1002/ece3.644.

Peñalba, J.V. and Wolf, J.B.W. (2020) ‘From molecules to populations: appreciating and estimating recombination rate variation’, *Nature Reviews Genetics*, 21(8), pp. 476–492. Available at: https://doi.org/10.1038/s41576-020-0240-1.

‘Picard toolkit’ (2019) *Broad Institute, GitHub repository*. Broad Institute. Available at: https://broadinstitute.github.io/picard/.

Provan, J. *et al.* (2008) ‘High dispersal potential has maintained long-term population stability in the North Atlantic copepod Calanus finmarchicus’, *Proceedings of the Royal Society B: Biological Sciences*, 276(1655), pp. 301–307. Available at: https://doi.org/10.1098/rspb.2008.1062.

Purcell, S.M. and Chang, C.C. (2017) ‘PLINK 2.0’. Available at: www.cog-genomics.org/plink/2.0/.

R Core Team (2024) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Rochette, N.C. *et al.* (2023) ‘On the causes, consequences, and avoidance of PCR duplicates: Towards a theory of library complexity’, *Molecular Ecology Resources*, 23(6), pp. 1299–1318. Available at: https://doi.org/10.1111/1755-0998.13800.

Schmid, M.S., Maps, F. and Fortier, L. (2018) ‘Lipid load triggers migration to diapause in Arctic Calanus copepods—insights from underwater imaging’, *Journal of Plankton Research*, 40(3), pp. 311–325. Available at: https://doi.org/10.1093/plankt/fby012.

Scott, C.L. *et al.* (2000) ‘Lipids and life strategies of Calanus finmarchicus , Calanus glacialis and Calanus hyperboreus in late autumn, Kongsfjorden, Svalbard’, *Polar Biology*, 23(7), pp. 510–516. Available at: https://doi.org/10.1007/s003000000114.

Selkoe, K.A. and Toonen, R.J. (2006) ‘Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers’, *Ecology Letters*, 9(5), pp. 615–629. Available at: https://doi.org/10.1111/j.1461-0248.2006.00889.x.

Silliman, K. *et al.* (2021) ‘Base-substitution mutation rate across the nuclear genome of Alpheus snapping shrimp and the timing of isolation by the Isthmus of Panama’, *BMC Ecology and Evolution*, 21(1), p. 104. Available at: https://doi.org/10.1186/s12862-021-01836-3.

Skjoldal, H.R. and Aarflot, J.M. (2023) ‘Abundance and biomass of copepods and cladocerans in Atlantic and Arctic domains of the Barents Sea ecosystem’, *Journal of Plankton Research*, 45(6), pp. 870–884. Available at: https://doi.org/10.1093/plankt/fbad043.

Skottene, E. *et al.* (2020) ‘Lipid metabolism in Calanus finmarchicus is sensitive to variations in predation risk and food availability’, *Scientific Reports*, 10(1), p. 22322. Available at: https://doi.org/10.1038/s41598-020-79165-6.

Smolina, I. *et al.* (2014) ‘Genome- and transcriptome-assisted development of nuclear insertion/deletion markers for Calanus species (Copepoda: Calanoida) identification’, *Molecular Ecology Resources*, 14(5), pp. 1072–1079. Available at: https://doi.org/10.1111/1755-0998.12241.

Tom, J.A. *et al.* (2017) ‘Identifying and mitigating batch effects in whole genome sequencing data’, *BMC Bioinformatics*, 18(1), p. 351. Available at: https://doi.org/10.1186/s12859-017-1756-z.

Unal, E. and Bucklin, A. (2010) ‘Basin-scale population genetic structure of the planktonic copepod *Calanus finmarchicus* in the North Atlantic Ocean’, *Progress in Oceanography*, 87(1), pp. 175–185. Available at: https://doi.org/10.1016/j.pocean.2010.09.017.

Van Rossum, G. and Drake, F.L. (2009) *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Visser, A.W., Grønning, J. and Jónasdóttir, S.H. (2017) ‘Calanus hyperboreus and the lipid pump’, *Limnology and Oceanography*, 62(3), pp. 1155–1165. Available at: https://doi.org/10.1002/lno.10492.

Weir, B.S. and Cockerham, C.C. (1984) ‘Estimating F-Statistics for the Analysis of Population Structure’, *Evolution*, 38(6), pp. 1358–1370. Available at: https://doi.org/10.2307/2408641.

Weydmann, A. *et al.* (2017) ‘Mitochondrial genomes of the key zooplankton copepods Arctic Calanus glacialis and North Atlantic Calanus finmarchicus with the longest crustacean non-coding regions | Scientific Reports’, *Scientific Reports*, 7(1), p. 13702. Available at: https://doi.org/10.1038/s41598-017-13807-0.

Weydmann, A. *et al.* (2018) ‘Postglacial expansion of the Arctic keystone copepod Calanus glacialis’, *Marine Biodiversity*, 48(2), pp. 1027–1035. Available at: https://doi.org/10.1007/s12526-017-0774-4.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org.

Wickham, H. *et al.* (2019) ‘Welcome to the Tidyverse’, *Journal of Open Source Software*, 4(43), p. 1686. Available at: https://doi.org/10.21105/joss.01686.

Wilfert, L., Gadau, J. and Schmid-Hempel, P. (2007) ‘Variation in genomic recombination rates among animal taxa and the case of social insects’, *Heredity*, 98(4), pp. 189–197. Available at: https://doi.org/10.1038/sj.hdy.6800950.

Willing, E.-M., Dreyer, C. and Oosterhout, C. van (2012) ‘Estimates of Genetic Differentiation Measured by FST Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers’, *PLOS ONE*, 7(8), p. e42649. Available at: https://doi.org/10.1371/journal.pone.0042649.

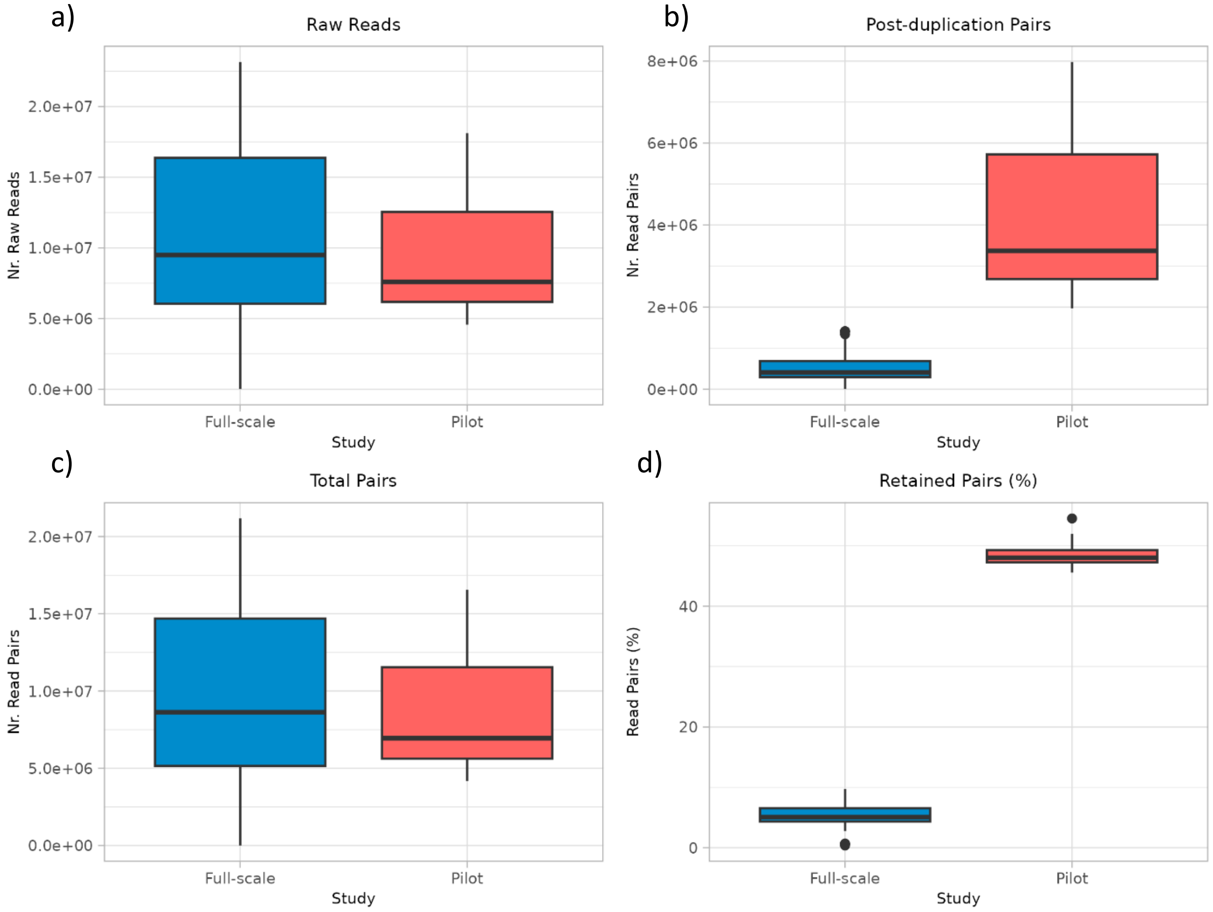
Wold, A. *et al.* (2011) ‘Arctic seabird food chains explored by fatty acid composition and stable isotopes in Kongsfjorden, Svalbard’, *Polar Biology*, 34(8), pp. 1147–1155. Available at: https://doi.org/10.1007/s00300-011-0975-4.

Zhang, Y. *et al.* (2018) ‘Alternative empirical Bayes models for adjusting for batch effects in genomic studies’, *BMC Bioinformatics*, 19(1), p. 262. Available at: https://doi.org/10.1186/s12859-018-2263-6.

Appendix

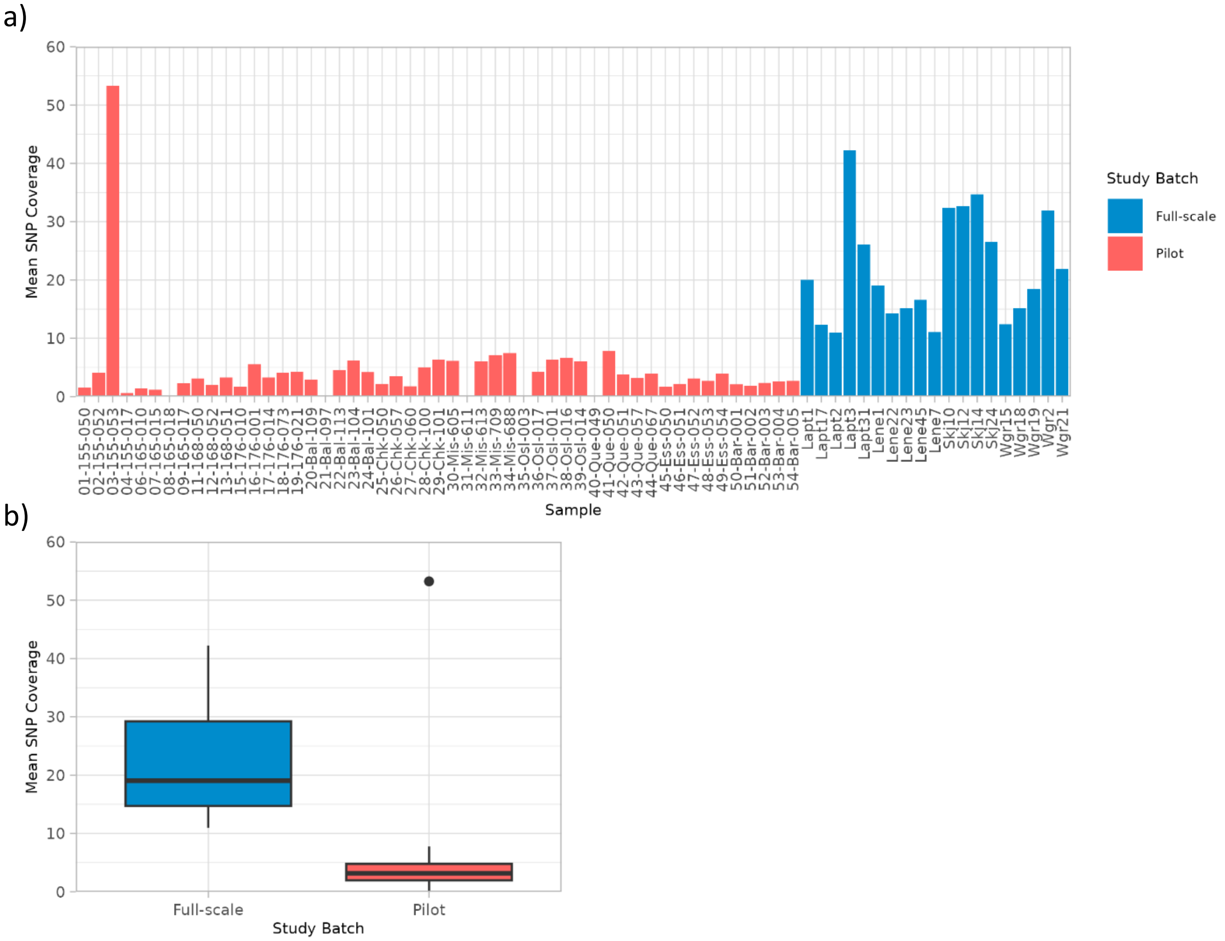
Supplementary Information

A. Figures



Supplemental Figure 1. Boxplots comparing sequencing metrics of the pilot and full-scale studies. The number of raw reads (a). The number of total pairs (b). The number of read pairs retained after deduplication (c). The percentage of reads retained after all filters were applied (d).

Appendix



Supplemental Figure 2. Bar-plot of the mean SNP coverage (before HWE and LD filtering) per individual (a). Box-plot comparing the mean SNP coverage (before HWE and LD filtering) of the pilot and full-scale batches (b).