Characterising the microbiome from host shotgun sequencing data: bacterial and diatom community dynamics derived from killer whale skin

Rebecca Hooper[1]*, Jaelle Brealey[1]*, Tom van der Valk[1], Antton Alberdi[2], John W. Durban[3], Holly Fearnbach[4], Kelly M. Robertson[3], Robin W. Baird[5], M. Bradley Hanson[6], Paul Wade[7], M. Thomas P. Gilbert[2,8], Phillip A. Morin[3], Jochen B.W. Wolf[9,10], Andrew D. Foote[11]**, Katerina Guschanski[1]**

*These authors contributed equally to this work.

**These authors contributed equally to this work and are co-senior authors.

[1]Animal Ecology, Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36, Uppsala, Sweden
[2]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Volgade 5-7, Copenhagen K 1350, Denmark
[3]Marine Mammal and Turtle Division, Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 8901 La Jolla Shores Dr., La Jolla, CA 92037, U.S.A.
[4]SR3, SeaLife Response, Rehabilitation, and Research P.O. Box 1404 Mukilteo, WA 98275, USA
[5]Cascadia Research, 4th Avenue, Olympia, Washington 98501, USA
[6]Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Boulevard East, Seattle, Washington 98112, USA
[7]National Marine Mammal Laboratory, Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 7600 Sand Point Way NE, Seattle, Washington 98115, USA
[8]NTNU University Museum, N-7491 Trondheim, Norway
[9]Science of Life Laboratories and Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, Uppsala SE-752 36, Sweden
[10]Section of Evolutionary Biology, Faculty of Biology, LMU Munich, Großhaderner Strasse 2, Planegg-Martinsried 82152, Germany
[11]Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Bangor University, Bangor, Gwynedd, LL57 2UW, UK

Correspondence
Rebecca Hooper, Tremough House, University of Exeter Cornwall Campus, Penryn, Cornwall, TR10 9FE, UK. Email: rebecca.hooper@evobio.eu

Jaelle Brealey, Animal Ecology, Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36, Uppsala, Sweden. Email: jaelle.brealey@ebc.uu.se
Andrew Foote, Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Bangor University, Bangor, Gwynedd, LL57 2UW, UK. Email: FooteAD@gmail.com
Katerina Guschanski, Animal Ecology, Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36, Uppsala, Sweden. Email: katerina.guschanski@ebc.uu.se

Abstract

Recent exploration into the interactions and relationship between hosts and their microbiota has revealed a connection between many aspects of the host's biology and associated microorganisms. Whereas amplicon sequencing has traditionally been used to characterise the microbiome, the increasing number of published population genomics datasets can potentially be exploited to opportunistically study microbial profiles from the host shotgun sequencing data. Here, we use sequence data originally generated from killer whale *Orcinus orca* skin biopsies for population genomics, to characterise the skin microbiome and investigate how host social and geographic factors influence the microbial community composition. We identified 845 microbial species from 2.4 million reads that did not map to the killer whale reference genome. After accounting for limitations and biases of such a dataset (e.g. low and uneven sequencing depth, and contamination), we detected ecologically relevant signals. We found that both ecotypic and geographic factors influence community composition of killer whale skin microbiomes. Furthermore, we identified key species that drive community composition of the microbiome and showed that they are embedded in unique species networks, one of which is tentatively linked to diatom presence and poor skin condition. These results further add to the hypothesis that the episodic migrations of Antarctic killer whales to warmer waters are associated with skin turnover and may control the effects of potentially pathogenic bacteria such as *Tenacibaculum dicentrarchi*. Our work

demonstrates the feasibility of microbiome studies from host shotgun sequencing data and highlights the importance of metagenomics in understanding the relationship between host and microbial ecology.

KEYWORDS

Cetacea, Contamination, Metagenomics, Microbiota, *Orcinus orca*

1 | INTRODUCTION

The skin microbiome is an ecosystem comprised of trillions of microbes sculpted by ecological and evolutionary forces acting on both the microbes and their host (McFall-Ngai, & Eliot, 2005; Byrd et al., 2018). Recent explorations into how these factors shape the microbial community have revealed a tight connection between almost every aspect of the host's biology and the associated microbial community (Bordenstein, & Theis, 2015). Although numerous intrinsic and extrinsic factors that influence the skin microbiome composition have been identified, the relative importance of these factors often appear to differ even between closely related host taxa (Kueneman et al., 2014; McKenzie, Bowers, Fierer, Knight, & Lauber, 2012; Wolz et al., 2017). Intrinsically, the host's evolutionary history, age, sex, and health appear significant (Leyden, McGiley, Mills, & Kligman, 1975; Cho & Blaser, 2012; McKenzie et al., 2012; Phillips et al., 2012; Apprill et al., 2014; Ying et al., 2015; Chng et al., 2016). Extrinsically, both environmental factors, where a sub-selection of environmental microbes colonise host skin (Apprill et al., 2014; Wolz et al., 2017; Walke et al., 2014; Ying et al., 2015), and socioecological factors, such as a host's social group and the level of interaction with conspecifics (Jain, 2011; Lax et al., 2014; Song et al., 2013; Kolodny et al. 2017), can play important roles.

Most microbiome studies to date are based on 16S ribosomal RNA gene sequences, a highly conserved region of the bacterial genome (Hamady & Knight, 2009). However, in addition to potential biases in PCR amplification, in which low reliability of quantitative estimations arise due to PCR stochasticity and different numbers of 16S gene copies in each bacterial species (Alberdi et al., 2018), metagenomic analysis of the 16S region can limit functional and taxonomic classification (Quince, Walker, Simpson, Loman, & Segata, 2017). In contrast,

shotgun metagenomics can facilitate both high-resolution taxonomic and functional analyses (Ranjan, Rani, Metwally, McGee, & Perkins, 2016; Quince, Walker, Simpson, Loman, & Segata, 2017; Koskella et al., 2017). The advent of affordable high-throughput sequencing has seen an ever-increasing number of population genomics studies in a wide range of study systems (e.g. Jones et al., 2012; Poelstra et al., 2014; Der Sarkissian et al., 2015; Nater et al., 2017). Yet, to the best of our knowledge, only a few human studies have exploited their sequencing data to secondarily investigate the microbial communities associated with the sampled tissue of their host (Ames et al., 2015; Zhang et al., 2015; Mangul et al., 2016).

Here, we explore the relative importance of extrinsic factors on the skin microbiome of killer whales (*Orcinus orca*) using shotgun sequencing data derived from skin biopsy samples of five ecologically specialised populations, or ecotypes (Foote et al., 2016). Given the widespread geographic range (Forney & Wade, 2007) and variation in ecological specialisation of killer whales, even in sympatry (Durban et al., 2017; Ford et al., 1998), this species provides a good study system for exploring the effects of both geographic location and ecotype (a proxy for both sociality and phylogenetic history) on the skin microbiome. However, the opportunistic use of such data is also fraught with potential pitfalls. We therefore describe in detail measures taken to disentangle potential sources of contamination from the true skin microbiome, thus providing a useful roadmap for future host-microbiome studies that exploit host-derived shotgun sequencing data.

2 | MATERIALS AND METHODS

2.1 | Study system

Throughout most of the coastal waters of the North Pacific two ecotypes of killer whales are found in sympatry: the mammal-eating '*transient*' and fish-eating '*resident*' ecotype (Ford et al., 1998; Saulitis et al., 2000; Matkin et al., 2007; Filatova et al., 2015). Four decades of field studies have found that they are socially and genetically isolated (Hoelzel & Dover, 1991; Barrett-Lennard, 2000; Hoelzel et al., 2007; Ford, 2009; Parsons et al., 2013; Filatova et al., 2015; Morin et al., 2015; Foote & Morin, 2016). Killer whales have also diversified into several ecotypes in the waters around the Antarctic continent, including a form commonly observed hunting seals in the pack-ice of the Antarctic peninsula (*type B1*), a form that feeds on penguins in the coastal waters of the Antarctic peninsula (*type B2*), and a dwarf form thought to primarily feed on fish in the dense pack-ice of the Ross Sea (*type C*) (Pitman & Ensor, 2003; Pitman & Durban, 2010, 2012; Durban et al., 2017; Pitman et al. 2018).

2.2 | Sample collection and data generation

Here, we used the unmapped reads from a published population genomics study of killer whale ecotypes (European Nucleotide Archive, www.ebi.ac.uk/ena, accession numbers: ERS554424-ERS554471; Foote et al. 2016). In that study, low coverage genomes were sequenced from a total of 49 wild killer whales, corresponding to five ecotypes; 10 samples each of the North Pacific fish-eating *resident* and sympatric mammal-eating *transient* ecotypes, and 8, 11 and 10 samples respectively from Antarctic types *B1*, *B2* and *C* (see Figure 1a for the sampling locations). DNA was extracted from epidermal biopsies collected by firing a lightweight dart with a sterilised stainless-steel cutting tip from a sterilised projector (e.g. Palsbøll et al., 1991; Barrett-Lennard et al., 1996) at the flank of the killer whale. Biopsies

were stored in sterile tubes at -20°C. At no point were biopsy samples in direct contact with human skin.

DNA extraction, library building, and sequencing have been previously described (Foote et al., 2016). All lab work was conducted in a sterile flow hood to prevent contamination, and the negative controls were not sequenced. Sequencing was done at the Danish National High-Throughput DNA Sequencing Centre within the University of Copenhagen. The facility is specifically geared for low DNA quantity library sequencing from ancient and environmental DNA. Samples of the same ecotype were pooled and sequenced across multiple sequencing lanes. Samples of different ecotypes were always run on different sequencing lanes, with the exception of several *type B1* and *B2* samples, which were sequenced on shared lanes.

2.3 | Sequencing read pre-processing

As a means to enrich the dataset for bacterial sequences, we first used SAMtools v1.5 (Li et al., 2009) to remove all sequencing reads that mapped to the killer whale reference nuclear genome (Oorca1.1, GenBank: ANOL00000000.2; Foote et al., 2015) and mitochondrial genome (GU187176.1) with BWA-mem (Li & Durbin, 2009). The remaining reads were adapter-trimmed using AdapterRemoval V2.1.7 (Schubert, Lindgreen, & Orlando, 2016). We then removed duplicates generated during library indexing PCR by merging reads with identical sequences using in-house python scripts (available upon request). All reads with an average quality score <30 were filtered out using prinseq V0.20.4 (Schmieder & Edwards, 2011), and all reads of <35bp were removed using AdapterRemoval V2.1.7.

2.4 | Investigating contamination

Despite the precautions outlined above, contamination can be introduced at several stages of the sequence data generation, and subsequently mistaken for the genuine host-associated microbiome signal. Contaminating DNA can be present in laboratory reagents and extraction kits (Lusk et al., 2014; Salter et al., 2014). For example, silica in some commercial DNA spin columns is derived from diatom cells and therefore can be a potential source of contamination with diatom DNA (Naccache et al., 2013). However, the Qiagen QIAquick spin columns used in this study do not contain silica from biological material, according to the manufacturer. Cross contamination can also occur between samples processed in the same sequencing centre (Ballenghien et al., 2017). The impact of contamination increases in samples with small amounts of true exogenous DNA and can swamp the signal from the host's actual microbiome (Lusk et al., 2014; Salter et al., 2014). Here, we implement a workflow that attempts to differentiate between contaminant and real exogenous DNA from host species shotgun sequencing data.

The contamination of microbial reference genomes by PhiX, which is used as a control in Illumina sequencing, is a known potential source of error in metagenomics studies using shotgun sequencing data (Mukherjee et al., 2015). Therefore, to avoid erroneous mapping of PhiX-derived reads to contaminated genomes, we removed all reads mapping to the PhiX genome used by Illumina (RefSeq ID: NC_001422) with BWA-mem 0.7.15 (Li, 2013) on standard parameters.

If the amount of contamination (derived from laboratory reagents or environment) is relatively equal among samples, we expect the relative proportion of contaminant sequencing reads to be inversely correlated to the quantity of sample-derived DNA: low DNA

quantity samples will be disproportionately affected by contaminant DNA sequences compared with high-quantity samples (Lusk et al., 2014; Salter et al., 2014). We therefore estimated the correlation between the proportion of the total sequencing reads assigned to each microbial taxon and total DNA read count (both prior to PCR duplicate removal). Microbial taxa that consistently increased in abundance in low-quantity samples were flagged as potential contaminants.

To account for the possibility of contamination with human-associated microorganisms, we next quantified the amounts of human DNA in our samples and used this as a proxy of human-derived microbial contamination. To this end, reads that did not map to the killer whale reference genome were mapped to the human genome (GRCh38.p12) using BWA-mem with default parameters. Only reads uniquely mapping to a single region of the genome with high quality (samtools -q 30 -F 4 -F 256) were retained, and we removed all duplicates using samtools rmdup in single-end mode. Human contamination levels were estimated by calculating the percentage of reads mapped to the human genome out of the total reads that did not map to the killer whale genome. We included these values as a covariate in statistical models as a way to, at least partially, control for contamination with human-associated microorganisms.

Finally, we investigated whether specific bacterial taxa have previously been reported to be likely contaminants. Following read-based analyses, we found that our samples were dominated by *Cutibacterium* (*Propionibacterium*) *acnes*, which is abundant on human skin (Byrd et al., 2018) and a known contaminant of high-throughput sequencing data (Lusk et al., 2014; Mollerup et al., 2016). We therefore investigated the distribution of sequence identity

between our *C. acnes* reads and the *C. acnes* reference genomes, with the expectation that human or laboratory contaminants would show high (close to 100) percent identity, whereas killer whale-derived *C. acnes* would be more divergent. Additionally, we analysed data from a North Pacific killer whale sequenced at ~20x coverage in a published study, in which sample collection, DNA extraction, and sequencing was entirely independent of our data production (accession number: SRP035610; Moura et al., 2014). If *C. acnes* was present in these data, it would suggest that either it was a real component of the killer whale skin microbiome, or it was independently introduced as contamination in both studies.

2.5 | Taxonomic assignment

We used MALT (MEGAN Alignment Tool) version 0.3.8 (Herbig et al., 2016) to create a reference database of bacterial genomes downloaded from the NCBI FTP server (ftp://ftp.ncbi.nlm.nih.gov/genbank, accessed 26th January 2017). We performed a semi-global nucleotide-nucleotide alignment against the reference database. Sequence identity threshold was set to 95% and only taxa with five or more aligned reads were included in subsequent analysis.

The nucleotide alignments produced in MALT were further analysed in MEGAN version 6.7.6 (Huson, Auch, Qi, & Schuster, 2007). We used the default parameters to assign reads to the lowest common ancestor (LCA) of the set of taxa that these reads were assigned to during the MALT alignment (Huson et al., 2007). Alignments with uneven coverage across bacterial genomes were flagged using a custom python script, and manually assessed in MEGAN. This step was necessary to identify spurious and incorrectly supported bacterial taxa, which were removed from further analysis if they represented highly abundant species (Warinner et al.,

2017). Taxonomic composition of the samples was also interactively explored in MEGAN, and the raw count data (number of reads assigned to each taxon) was exported for subsequent analysis.

Taxonomic assignment was also carried out using an assembly-based approach. Filtered metagenomic sequences of all samples were merged to perform a co-assembly using Megahit 1.1.1 (Li et al., 2015) with default settings and the following k-list: 21,29,39,59,79. Assembly quality was assessed using Quast 4.5 (Gurevich et al., 2013). Contigs were subsequently mapped to reference bacterial genomes with MGMapper (Petersen et al., 2017) using best mode to assign taxonomy. The assembly file was indexed using BWA-mem and samtools was used to map the reads of each sample back to the assembly contigs. Individual coverage values were calculated with Bedtools 2.26.0 (Quinlan et al., 2010) and contig coverage table normalised using Cumulative Sum Scale (CSS) as implemented in MetagenomeSeq (Paulson et al., 2013).

In addition, we explored how low sequencing depth may affect the inferred bacterial profiles: The sequencing data used is this study is rather shallow in terms of coverage of microbial taxa, corresponding to low coverage killer whale genomes (mean of 2x). To this end, we used an independently sequenced 20x coverage *resident* killer whale genome (Moura et al., 2014). By drawing a random subset of reads from this genome using samtools, we compared the taxonomic composition of the microbiome of the same individual at 20x, 10x, 5x and 2x mean sequence coverage depth.

2.6 | Diversity analyses

We calculated all diversity measures in Vegan 2.4-3 (Oksanen et al., 2017), using reads that were assigned to the species level in MEGAN. By focussing on taxa at the species-level, we were able to explore the skin microbiome at a high resolution, an advantage of shotgun over amplicon-based analyses. However, results of this analysis should be interpreted in light of a species-level focus, where we are exploring a small, potentially enriched in pathogens and common environmental bacteria, yet well-resolved representation of the microbiome, rather than a holistic representation of the entire ecosystem.

To control for bias introduced by varying genome size (species with larger genomes show higher read counts, which are translated into higher abundance scores; Warinner et al., 2017), we divided all read counts by the size of the respective full bacterial genome. If the taxon was mapped to the level of the strain, we divided the read count by the published genome size of that strain; if identified to the species level, we divided the read count by the average genome size across all published strains of that species.

Beta diversity was explored using two dissimilarity matrices in the R-package Vegan: abundance-based Bray-Curtis and presence/absence-based binary Jaccard distances. To assess the strength and significance of ecotype and geographic location (longitude and latitude) in describing variation in community composition, we conducted permutational multivariate analysis using the function *Adonis* in Vegan. We controlled for differing depths of coverage between samples using two techniques. First, we used genome size-controlled data (see above) and included the number of reads mapping to the species level as a

covariate. Second, Total Sum Scaling (TSS) normalisation of the genome-size controlled data was conducted, followed by conversion to the Bray-Curtis distance matrix. TSS normalisation is irrelevant for presence absence data, as only species presence, rather than species abundance, is retained in the binary presence absence matrix. As a result, three models were explored: two Bray-Curtis models with differing depth-control techniques, and one Jaccard model using the first described depth-control technique. Each model consisted of the following covariates: latitude (numeric), longitude (numeric), ecotype (categorical), and percentage human contamination (numeric), with library size included only when TSS normalisation was not used. For each model, residuals were permuted 9999 times. We used the function *Betadisper* (Vegan 2.4-3) followed by ANalysis Of VAriance (ANOVA) to test for homogeneity of group dispersions. *Betadisper* can be used to ensure that i) the *Adonis* model results are not confounded by heterogeneous variances (Anderson, 2001) and ii) to make biological inferences about between-group variance in community composition.

We used the function *capscale* from the Vegan package to perform Principal Co-ordinate Analysis (PCoA). The six bacterial species that described the most variation on PCoA1, and the six that described the most variation on PCoA2, were designated as 'driving species'. Due to overlap of three driving species between the two axes, we classified a total of nine unique driving species that define individual differences in microbiome composition (Table S1).

2.7 | Network analysis

To venture beyond single microbial species and explore microbial interactions that include interspecific dynamics, we expanded our analyses to networks of bacterial communities associated with the driving species identified through the PCoA. Using netassoc 0.6.3

(Morueta-Holme et al., 2016), we compared the observed partial correlation coefficients between species with a null distribution estimated from identical species richness and abundances as the observed data. Species co-occurrence scores were calculated between all species pairs in the raw dataset. We set the number of null replicates to 999, and corrected p-values for multiple comparisons using the FDR method. From the resulting matrix of significant co-occurrence scores, we selected the 10 species with the highest absolute co-occurrence score for each of the 9 unique driving species. We created a new matrix including only these species and visualised co-occurrence networks with the package igraph 1.0.1 (Csárdi & Nepusz, 2006).

2.8 | Functional profiling

Community composition can be a poor predictor of the functional traits of the microbiome, due to processes such as horizontal gene transfer (HGT) between bacterial taxa, which can decouple species composition and function (Koskella et al., 2017). Shifting focus from the taxonomic composition to the genic composition of the microbiome reduces the impact of HGT on functional characterisation (Koskella et al., 2017).

To explore functional profiles of the samples, we used DIAMOND v0.9.10 with default parameters (Buchfink, Xie & Huson, 2015) to create a reference database of non-redundant protein sequences from fully sequenced bacterial genomes downloaded from the NBCI FTP server (ftp://ftp.ncbi.nlm.nih.gov/genbank, accessed 9th March 2017). Nucleotide-to-amino-acid alignments of the sample reads to the reference database was performed in DIAMOND and the top 10% of alignments per query reported. The MEGAN tool daa-meganizer was then used to assign reads to proteins based on the DIAMOND alignments and to assign functional

roles to these proteins using the SEED (Overbeek et al., 2005) and eggNOG (Huerta-Cepas et al., 2017) databases. Since one protein can have more than one function, it is possible for one read to be assigned to multiple functional subsystems. The raw count data (number of reads assigned to functional subsystem) were exported from MEGAN and further processed in R. To control for differences in library depth, read counts per functional group were normalised by total read numbers mapping to SEED or eggNOG terms. We used principal components analysis (PCA) performed in the R package *prcomp* to visualise differences in functional groups between individuals.

We additionally performed an assembly-based functional profiling to overcome the individual weaknesses of both assembly and read-based methodologies (Quince et al., 2017). *Ab initio* gene prediction was performed over the metagenomic assembly using Prodigal 2.6.3 (Hyatt et al., 2010). The list of predicted gene sequences was indexed using BWA and samtools was used to map the reads of each sample back to the gene sequences. We used Bedtools 2.26.0 (Quinlan et al., 2010) to calculate individual coverages. Gene coverage table was subsequently Cumulative Sum Scale (CSS) normalised using MetagenomeSeq (Paulson et al., 2013).

2.9 | Diatom-association analyses

Antarctic killer whales are often observed to have a yellow hue, which has been attributed to diatom coverage (Berzin & Vladimirov, 1983; Pitman & Ensor, 2003) and identifiable individuals have been observed to transit from this yellow skin colouration to a 'clean' skin condition (Durban & Pitman, 2012). This change is hypothesised to occur during brief migrations to subtropical latitudes, where turnover of the outer skin layer takes place with a

reduced thermal cost (Durban & Pitman, 2012). If this hypothesis is correct, diatom abundance should be correlated to skin age and colouration (Hart, 1935; Konishi et al. 2008; Durban & Pitman, 2012). Inter-individual variation in microbiome profiles within the Antarctic ecotypes could therefore reflect variation in the age of the outer skin layer. During network analysis, we identified a possible association between a key bacterial species driving between-sample differences in community composition (*Tenacibaculum dicentrarchi*) and bacterial species associated with diatoms. Following from our observations that three samples from Antarctic ecotypes had high abundances of *T. dicentrarchi* and that in the PCoA these samples were differentiated from most other samples, we investigated the link between observed diatom coverage, abundance of *T. dicentrarchi* and abundance of other algae-associated bacterial species. We conducted qualitative colour grading of *type B1* and *type B2* individuals using photographs taken at the time of biopsy collection, ranging from 'clean' through to 'prominent' yellow colouration.

Only a limited number of diatom reference genomes are available and therefore we used two methodologies to quantify the level of diatom DNA in our samples. First, we used MALT and MEGAN in the same taxonomic pipeline as previously described, but with a reference database comprised of NCBI RefSeq nucleotide sequences from the diatom phylum Bacillariophyta (downloaded 30th October 2017). To date, only seven diatom reference genomes are available, thus identification at the species level was not attempted. Instead, numbers of reads mapping to Bacillariophyta were exported and further processed in R. Raw diatom read counts were converted to proportion of total number of sequencing reads per sample. Second, we aligned all reads against the Silva release 128 database using BWA and used uclust (Edgar, 2010) in QIIME (Caporaso et al., 2010) to assign taxonomy based on the

Silva 18S database. From the resulting OTU table we retained reads that matched to known diatom taxonomy.

We explored the correlation between latitude (by grouping North Pacific and Antarctic ecotypes) and proportion of reads per sample mapping to diatoms using a generalised linear model with a quasipoisson error structure and log link. As covariates, we included longitude, number of reads mapping to the bacterial species level to control for library size, and number of human reads to control for human-associated microbial contamination. Using the same model structure, we then tested the correlation between proportion of reads per sample mapping to diatoms and the presence/abundance of *T. dicentrarchi* reads, as well as the correlation with presence of known algae-associated bacterial species (including *T. dicentrarchi, Cellulophaga baltica*, *Formosa* sp. *Hel1_33_131*, *Winogradskyella* sp., *Marinovum algicola*, *Agarivorans gilvus*, *Pseudoalteromonas atlantica* and *Shewanella baltica*: Amin et al., 2012; Goecke et al., 2013a,b, incorporated as a binary variable).

3 | RESULTS

Metagenomic profiles from the skin microbiome of 49 killer whales from five ecotypes (Figure 1) were successfully reconstructed using shotgun sequencing data from DNA extracted from skin biopsies. Of the reads retained following our stringent filtering procedure, but before our investigations into *C. acnes* as a possible contaminant, 8.20% (n = 7,984,195) were assigned to microbial taxa using the read-based approach, with 2.45% (n = 2,384,587) assigned at the species level (see Dryad repository doi:10.5061/dryad.c8v3rv6). Overall, 845 species of microbes were identified. The co-assembly yielded a 33.01 Mbp-long metagenome comprised of 45,934 contigs (N50: 970 bp, Average: 730 bp, Max: 48,182 bp). Taxonomy was assigned

to 41.73% of the contigs. Results from the assembly-based approach were concordant with the read-based results, we therefore report only the latter.

3.1 | Investigating contamination

On average 0.16% of reads (range 0.008-5.43%) reads mapped to the human genome (Table S2), suggesting presence of human contamination and making it possible that human-derived bacteria were present in our dataset. After correcting for multiple testing, we found no significant negative correlation between the proportion of reads assigned to each bacterial taxon and the total number of sequenced reads (Supporting Information, Fig. S1). Negative trends (although not significant) between some bacterial taxa and the total number of sequenced reads were largely driven by one outlier sample with the lowest coverage (B1_124047). Following the deduplication step of our processing pipeline, these taxa were no longer present in the dataset, as they fell below our defined threshold of five aligned reads in MALT (Fig. S2).

C. acnes was identified as the most abundant bacterial species, with an average abundance of 39.57% (standard deviation = 24.65; Fig. S3a), but it may have been introduced via human or laboratory contamination (Lusk et al., 2014). Percent identity to the human-derived *C. acnes* genome was 100% for 245 and over 97% for 505 of the 527 contigs identified as *C. acnes* by MGMapper (Fig. S4), supporting the idea of a likely exogenous source of *C. acnes*. Killer whale samples pooled by ecotype were sequenced across multiple sequencing lanes, allowing us to investigate if contamination with *C. acnes* was introduced at the sequencing step. Relative *C. acnes* abundance per sample was highly similar between sequencing lanes (Coefficient of Variation = 0.076; Fig. S5), suggesting that the contamination occurred prior to

sequencing. However, this species was also present to a high abundance (18.06% of reads aligning at species level) in the independently sequenced *resident* killer whale (Moura et al., 2014), suggesting that contamination with *C. acnes* was not specific to our workflow. We concluded that there was a high probability that *C. acnes* was a lab contaminant and therefore removed all *C. acnes* reads/contigs from our dataset before continuing with analysis.

3.2 | Taxonomic exploration

Read-based and assembly-based approaches produced concordant taxonomic profiles. The most abundant constituents of the killer whale skin microbiome at the phylum level were Proteobacteria, Actinobacteria, Bacteroidetes and Firmicutes (Fig. S3a), which have been identified in previous studies of baleen whale skin microbiota (Appril et al., 2014; Shotts et al., 1990). At the species level, we found a high level of inter-individual variation (Figure 1b, S3b): three samples, all from Antarctic ecotypes *B2* and *C*, were dominated by *Tenacibaculum dicentrarchi,* with abundances ranging from 13.63 – 43.34%. Another three Antarctic ecotype samples (*B1* and *B2*) were dominated by *Pseudoalteromonas haloplanktis* (from 44.52 – 86.03%), which is common in seawater. Finally, one Antarctic sample was dominated by *Staphylococcus saprophyticus* (81.3%).

Subsetting an independently sequenced *resident* killer whale genome (Moura et al., 2014) to lower sequencing depth, we inferred that while five most common taxa were found in similar proportions in high and low coverage data, the identification of rarer taxa became more stochastic at lower sequencing depths (Table S3). Our results may therefore suffer from this bias associated with low coverage data, which would be most prominent in

presence/absence-based analyses. As a means to control for this bias, we include library size as a covariate in models investigating beta diversity.

3.3 | Diversity analyses

Human contamination was significant in one of three models exploring beta diversity (Table 1), explaining less than 4% of the variation in species composition in all models. Latitude and ecotype were significant in all models, with latitude explaining less than 5% of variation in the data in all models, and ecotype explaining 10-12% of variation (Table 1). Longitude was not significant in any of the models. *Betadisper* analysis revealed no significant heterogeneity in the variation of community composition between ecotypes (non-TSS normalised Bray-Curtis: df = 4, F = 0.62, p = 0.64; TSS normalised Bray-Curtis: df = 4, F = 1.05, p = 0.39; binary Jaccard: df = 4, F = 0.96, p = 0.44). This suggests that between-individual variation in microbial composition is similar between ecotypes.

The Bray-Curtis PCoA explained more variation than Jaccard (26.12% versus 17.57% on the first two axes) and therefore we focused on Bray-Curtis results. A network based on significant co-occurrences between nine bacterial species driving variation at the individual level (Table S1) and the top 10 co-occurring species for each of the driving taxa, showed clearly differentiated and distinct community groups (Figure 2). Further investigation into the top 10 species that significantly co-occurred with the driving taxa found that two of the species showing the highest co-occurrence scores with the driving taxon *T. dicentrarchi* (*Formosa sp. Hel1_33_131* and *Algibacter alginolytica*) are associated with algae (Becker et al., 2017; Sun et al., 2016).

3.4 | *T. dicentrarchi* and diatoms

Both approaches to diatom identification produced concordant results (Fig. S6, Table S4). Antarctic ecotypes had a significantly higher abundance of diatom DNA than North Pacific ecotypes ($\beta = 0.65$, SE = 0.29, $p = 0.03$; Figure 3a). Individuals with "prominent" yellow colouration showed higher diatom abundance (Figure 3b), supporting the link between skin colour and diatom presence in Antarctic killer whales. Furthermore, the abundance of diatom DNA per sample was significantly positively correlated with the abundance and presence of *T. dicentrarchi* reads (number of *T. dicentrarchi* reads: $\beta = 0.014$, SE = 0.003, $p = <0.001$; presence of *T. dicentrarchi*: $\beta = 0.915$, SE = 0.207, $p <0.001$; Figure 3c) and the presence of at least one algae-associated bacterial species ($\beta = 0.98$, SE = 0.17, $p = <0.001$; Figure 3d).

3.5 | Functional analysis

In the read-based functional analysis a total of 3,611,441 reads mapped to eggNOG functions and 1,440,371 reads mapped to SEED functions. In the contig-based functional analysis we identified 56,042 potential genes in our metagenome, out of which EggNog function was assigned to 35,182. Both approaches identified energy production and conversion (class C) and amino acid metabolism and transport (class E) as the most abundant eggNOG functions in our dataset. The eggNOG PCA revealed high variability between individuals (Figure 4), however, a cluster of Antarctic whales was observed in principal component 2. These samples had high abundances of *T. dicentrarchi* (Figure 4) and were associated with functions corresponding to the COG functional categories J (translation, ribosomal structure and biogenesis, $\beta = 0.007$, SE = 0.002, $p = <0.001$), F (nucleotide transport and metabolism, $\beta =$

0.005, SE = 0.002, p = 0.004) and I (lipid transport and metabolism, β = 0.004, SE = 0.002, p = 0.03). The same cluster of high abundance *T. dicentrarchi* Antarctic samples was also identified in the SEED PCA (Fig. S7). These samples had increased numbers of reads mapping to DNA metabolism, amino acids and derivatives and cofactors/vitamins, although none of these functions were significantly correlated with *T. dicentrarchi* abundance.

4 | DISCUSSION

Our study highlights that communities of exogenous or host-associated microbiota can be genetically characterised from shotgun sequencing of DNA extracted from the host tissue. We were able to identify key species interactions, ecological networks and community dynamics of the microbes and diatoms that colonise killer whale skin by utilising unmapped reads from shotgun sequencing data generated from skin biopsies. A key advantage of this approach over amplicon-based sequencing is the ability to assess functional variation based on gene content, and to identify taxa to species level (Koskella et al., 2017; Quince, Walker, Simpson, Loman, & Segata, 2017). However, despite ongoing efforts to describe bacterial species diversity, the breadth of the reference database is a limiting factor in the unbiased characterisation of bacterial composition. Thus, taxa identified in our analyses are necessarily limited to species with available genomic information, and in some cases are likely to represent their close phylogenetic relatives (Tessler et al., 2017). We also demonstrate the impact of contamination on the low numbers of reads from true host-associated microbes, which can dilute the signal of biologically meaningful variation among samples.

Social and geographic factors have been found to influence microbiota diversity in terrestrial and semi-terrestrial animals (Koskella et al., 2017). However, there is less understanding of

how these factors interplay in a wide-ranging social marine mammalian system (Nelson et al., 2015). We found that beta diversity of the killer whale skin microbiome was significantly influenced by ecotype and latitude. The effect of ecotype as the most important tested variable highlights the significance of social factors in shaping microbiome richness and composition. In addition, it underscores that although killer whale skin is influenced by the local environment (Romano-Bertrand et al., 2015), it represents a unique ecosystem that is separate from that of the surrounding habitat. Killer whales are highly social mammals, and thus are likely to have a high potential for horizontal transfer of microbes between individuals during contact (Nelson et al., 2015). Ecotype-specific social behaviour, organisation and population structure (Lax et al., 2014; Tung et al., 2015; Li et al. ,2016), as well as other variables related to ecotype ecology, such as range size, and diet (due to transmission of bacteria from different prey species (Menke et al., 2017)), are all likely to affect the diversity of microbial species that individuals are exposed to, and also influence the level of horizontal transfer of microbes between whales. The phylogenetic and phylogeographic history of ecotypes is also likely to play a role, where the phylogeny of vertically transmitted bacterial species is likely to reflect that of the host (Ley et al., 2008, but see Rothschild et al., 2018). It is also likely to be influenced by the host's evolutionary history, including secondary contact between ecotypes (Foote & Morin, 2016), where both vertical and horizontal transmission of microbes between ecotypes is possible.

Despite the significance of ecotypes as a variable explaining skin microbiome diversity in killer whales, there is a strong overlap among ecotypes in the PCoA (Figure 1c), suggesting a shared core microbiome. However, the PCoA also reveals substantial inter-individual variation within

ecotypes. Among Antarctic ecotypes, individual variation was associated with diatom presence and a discrete subnetwork of microbial taxa. The occurrence of a 'yellow slime' attributed to diatoms on the skin of whales, including killer whales, was recorded as early as a century ago (Bennett 1920; Pitman et al., 2018). The extent of diatom adhesion on Antarctic whales is thought to correlate with latitude and the time the whale has spent in cold waters (Hart, 1935; Konishi et al. 2008). Satellite tracking of *type B1* and *B2* killer whale movements, Durban & Pitman (2012) documented rapid return migrations to subtropical latitudes, in which individuals travelled up to 9,400 km in 42 days. Based on the strong directionality and velocity of travel during these migrations, Durban & Pitman (2012) hypothesised that they were not associated with breeding or feeding behaviour. Instead, they argued that these migrations could be driven by the need to leave the frigid Antarctic waters and temporarily move to warmer waters, to allow for physiological maintenance including the regeneration of the outer skin layer (Durban & Pitman, 2012). The identification of the same individuals in Antarctic waters, sometimes with a thick accumulation of diatoms, and other times appearing 'clean', supports the hypothesis that skin regeneration is an intermittent rather than continuous process (Durban & Pitman, 2012). Furthermore, bite marks from cookiecutter sharks *Isistius spp*., which are typically found in warm and temperate waters, are regularly observed on *type C* killer whales (Pitman et al., 2018), which have also been directly tracked to warm sub-tropical waters (Durban et al., 2013).

Our findings lend credence to the hypothesis of Durban & Pitman (2012), that 'clean' and yellow-tinted *type B1* and *B2* killer whales represent differences in diatom load. In addition, we provide the first evidence that the extent of diatom coverage is also associated with a significant variation in the skin microbiome community, which includes an increase in

bacterial species that are known to form yellowish coloured colonies, e.g. *Stenotrophomonas maltophilia* and *T. dicentrarchi* (Buller et al., 2014; Avendaño-Herrera et al., 2016). Perhaps most significantly, diatom abundance was positively correlated with the abundance of *T. dicentrarchi*, a known pathogen in several fish species, which is associated with skin lesions and severe tail and fin rot (Piñeiro-Vidal et al., 2012; Habib et al., 2014; Avendaño-Herrera et al., 2016). Our analyses revealed that samples with high abundances of *T. dicentrarchi* show distinct functional profiles. Whether *T. dicentrarchi* represents a pathogen to killer whale hosts is unknown. However, Type B1 killer whales in apparently poor health and with heavy diatom loads have been observed with severe skin conditions (skin peeling and lesions; Figure 3e). Skin maintenance may thus represent a balancing act for Antarctic killer whales of managing the costs of pathogen load, thermal regulation, reduced foraging time and long-range movement. Research into the skin microbiome should therefore continue to form a component of the ongoing holistic and multidisciplinary research programme to investigate the health of Antarctic killer whale populations.

Ongoing field efforts provide the opportunity to further explore the relationships and interactions between killer whale hosts, their skin microbiome, other exogenous symbionts such as diatoms, and the environment. Our community-based analyses suggest the presence of a distinct environmental species network centred on *P. haloplanktis* as a driving species (Fig. 2). Collection and metagenomic characterization of environmental samples, such as sea water, alongside host biological samples would allow further explorations into the contribution of local ecological factors to the host microbiome. As a means of reducing the impact of contamination with DNA from laboratory environment, microbiome characterisation can be conducted by means of RNA sequencing. This has an additional

advantage of generating metatranscriptomic data, which, in combination with the metagenomic data, can facilitating the comparison/contrast between community function (using RNA transcript) and community taxonomic composition (using DNA sequence; Koskella et al., 2017). This may further reduce the potential impact of common lab contaminants, allowing the exploration of the bacterial functional repertoire that is in use in a given ecological context, including reconstruction of metabolic pathways (Bashiardes, Zilberman-Schapira, Elinav 2016). Contamination in the laboratory could be further controlled for and characterised through inclusion of extraction, library preparation and PCR blanks as negative controls (Lusk et al., 2014; Salter et al., 2014) and measures such as double-indexing (Rohland & Reich, 2012; van der Valk et al., 2018), which can then inform the emerging downstream filtering methods for separating true microbiomes from contamination (Delmont & Eren, 2016; Davis et al., 2018). Lastly, the advances in long-read sequencing using portable nanopore-based platforms (Garalde et al., 2018) make it possible to generate data suitable for reconstructing complete bacterial genomes while in the field, including in the Antarctic (Johnson et al., 2017). This is a promising development with respect to improving the breadth of host taxa from which bacterial taxa are derived and should improve future mapping of metagenomics data and taxonomic assignment.

Acknowledgements
We would like to thank Bob Pitman who was involved in the collection of many of the samples used in this study and who has contributed much to our understanding of the variation among killer whale types. We'd further like to thank David Studholme for pointing out the rich literature surrounding diatom microbiomes, as well as potential diatom-related contamination. Linda Rhodes, Morten Limborg and the microbiome journal club of the Evogenomics section at the Centre for GeoGenetics provided valuable feedback on this work and an earlier draft of this manuscript. We acknowledge The Danish National High-Throughput DNA Sequencing Centre for sequencing the samples and, particularly, Andaine Seguin-Orlando, Lillian Petersen, Cecilie Demring Mortensen, Kim Magnussen and Ian

Lissimore for technical support. Sample collection from killer whales in Antarctica was supported by the Lindblad Expeditions-National Geographic Conservation Fund. This work was funded by European Research Council grant ERCStG-336536 to J.B.W.W.; a Danish National Research Foundation grant DNRF94 to M.T.P.G, the Welsh Government and Higher Education Funding Council for Wales through the Sêr Cymru National Research Network for Low Carbon, Energy and Environment, and from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 663830 to AF, the FORMAS grant (project 2016-00835) to KG. A.A. was supported by the Danish Council for Independent Research - DFF (grant 5051-00033) and Lundbeckfonden (grant R250-2017-1351). Finally, we would like to thank the Linnaeus Scholarship Foundation at the University of Uppsala for providing a scholarship to RH, the Erasmus Mundus Master Programme in Evolutionary Biology (MEME) for facilitating the collaborative supervision of RH, and to Vanessa and Peter Hooper for providing financial support to RH throughout this project.

References

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2017). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, 9, 134–147.

Ames S. K., Gardner S. N., Marti J. M., Slezak T. R., Gokhale M. B., Allen J. E. (2015) Using populations of human and microbial genomes for organism detection in metagenomes. *Genome Research*, 25, 1056–1067.

Amin, S. A., Parker, M. S., & Armbrust, E. V. (2012). Interactions between diatoms and bacteria. Microbiology and Molecular Biology Reviews, 76(3), 667-684.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. Austral ecology, 26(1), 32-46.

Apprill, A., Robbins, J., Eren, A. M., Pack, A. A., Reveillaud, J., Mattila, D., Moore, M., Niemeyer, M., Moore, K.M. & Mincer, T. J. (2014). Humpback whale populations share a core skin bacterial community: towards a health index for marine mammals?. *PLoS One*, *9*(3), e90785.

Avendaño-Herrera, R., Irgang, R., Sandoval, C., Moreno-Lira, P., Houel, A., Duchaud, E., ... & Ilardi, P. (2016). Isolation, characterization and virulence potential of Tenacibaculum dicentrarchi in salmonid cultures in Chile. Transboundary and emerging diseases, 63(2), 121-126.

Ballenghien, M., Faivre, N., & Galtier, N. (2017). Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC biology*, *15*(1), 25.

Barrett-Lennard, L. G. (2000). *Population structure and mating patterns of killer whales (Orcinus orca) as revealed by DNA analysis* (Doctoral dissertation, University of British Columbia).

Barrett-Lennard, L., Smith, T. G., & Ellis, G. M. (1996). A cetacean biopsy system using lightweight pneumatic darts, and its effect on the behavior of killer whales. *Marine Mammal Science*, *12*(1), 14-27.

Bashiardes S, Zilberman-Schapira G, Elinav E. (2016). Use of Metatranscriptomics in Microbiome Research. *Bioinformatics and Biology Insights, 10*, 19–25.

Becker, S., Scheffel, A., Polz, M. F., & Hehemann, J. H. (2017). Accurate quantification of laminarin in marine organic matter with enzymes from marine microbes. Applied and environmental microbiology, 83(9), e03389-16.

Belden, L. K., Hughey, M. C., Rebollar, E. A., Umile, T. P., Loftus, S. C., Burzynski, E. A., Minbiole, K. P., House, L. L., Jensen, R. V., Becker, M. H. & Walke, J. B. (2015). Panamanian frog species host unique skin bacterial communities. *Frontiers in microbiology*, *6*, 1171.

Bennett, A. G. (1920). On the occurrence of diatoms on the skin of whales. In Proc. R. Soc. Lond. B (Vol. 91, No. 641, pp. 352-357).

Berzin, A. A., & Vladimirov, V. L. (1983). A new species of killer whale (Cetacea, Delphinidae) from the Antarctic waters. Zoologichesky Zhurnal, 62(2), 287-295.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–20.

Bordenstein, S. R., & Theis, K. R. (2015). Host biology in light of the microbiome: ten principles of holobionts and hologenomes. *PLoS biology*, *13*(8), e1002226.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature methods, 12(1), 59.

Buller, N. B. (2014). Bacteria and fungi from fish and other aquatic animals: a practical identification manual. Cabi.

Byrd, A. L., Belkaid, Y., & Segre, J. A. (2018). The human skin microbiome. *Nature Reviews Microbiology*.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... & Huttley, G. A. (2010). QIIME allows analysis of high-throughput community sequencing data. Nature methods, 7(5), 335.

Chng, K. R., Tay, A. S. L., Li, C., Ng, A. H. Q., Wang, J., Suri, B. K., Matta, S. A., McGovern, N., Janela, B., Wong, X. F. C. C. & Sio, Y. Y. (2016). Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. *Nature microbiology*, *1*(9), 16106.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, Complex Systems, 1695(5), 1-9.

Davis, N. M., Proctor, D., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2017). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. bioRxiv, 221499.

Delmont, T. O., & Eren, A. M. (2016). Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. PeerJ, 4, e1839.

Durban, J. W., Fearnbach, H., Burrows, D. G., Ylitalo, G. M., & Pitman, R. L. (2017). Morphological and ecological evidence for two sympatric forms of Type B killer whale around the Antarctic Peninsula. *Polar Biology*, *40*(1), 231-236.

Durban, J. W., & Pitman, R. L. (2012). Antarctic killer whales make rapid, round-trip movements to subtropical waters: evidence for physiological maintenance migrations?. *Biology Letters*, *8*(2), 274-277.

Durban JW, Pitman RL (2013) Out of Antarctica: dive data support 'physiological maintenance migration' in Antarctic killer whales. 20th Biennial Conference on the Biology of Marine Mammals. Society for Marine Mammalogy (9–13 December 2013, Dunedin, New Zealand).

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), 2460-2461.

Filatova, O. A., Borisova, E. A., Shpak, O. V., Meshchersky, I. G., Tiunov, A. V., Goncharov, A. A., Fedutin, I. D. & Burdin, A. M. (2015). Reproductively isolated ecotypes of killer whales Orcinus orca in seas of the Russian far east. *Biology Bulletin of the Russian Academy of Sciences*, *42*, 674-681.

Foote, A. D., & Morin, P. A. (2016). Genome-wide SNP data suggest complex ancestry of sympatric North Pacific killer whale ecotypes. Heredity, 117(5), 316.

Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., ... & Robertson, K. M. (2016). Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nature communications*, *7*, 11693.

Ford, J. K., Ellis, G. M., Barrett-Lennard, L. G., Morton, A. B., Palm, R. S., & Balcomb III, K. C. (1998). Dietary specialization in two sympatric populations of killer whales (Orcinus orca) in coastal British Columbia and adjacent waters. *Canadian Journal of Zoology*, *76*(8), 1456-1471.

Ford, J. K. (2009). Killer whale: Orcinus orca. In *Encyclopedia of Marine Mammals (Second Edition)* (pp. 650-657).

Forney, K. A., & Wade, P. R. (2006). Worldwide distribution and abundance of killer whales. *Whales, whaling and ocean ecosystems*, 145-162.

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., ... & Jordan, M. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nature methods*.

Goecke, F., Labes, A., Wiese, J., & Imhoff, J. F. (2013a). Phylogenetic analysis and antibiotic activity of bacteria isolated from the surface of two co-occurring macroalgae from the Baltic Sea. European journal of phycology, 48(1), 47-60.

Goecke, F., Thiel, V., Wiese, J., Labes, A., & Imhoff, J. F. (2013b). Algae as an important environment for bacteria–phylogenetic relationships among new bacterial species isolated from algae. Phycologia, 52(1), 14-24.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072-1075.

Habib, C., Houel, A., Lunazzi, A., Bernardet, J. F., Olsen, A. B., Nilsen, H., ... & Duchaud, E. (2014). Multilocus sequence analysis of the marine bacterial genus Tenacibaculum suggests parallel evolution of fish pathogenicity and endemic colonization of aquaculture systems. Applied and environmental microbiology, 80(17), 5503-5514.

Hamady, M., & Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research*, *19*(7), 1141-1152.

Hart, T. J. (1935). On the diatoms of the skin film of whales, and their possible bearing on problems of whale movements. University Press.

Herbig, A., Maixner, F., Bos, K. I., Zink, A., Krause, J., & Huson, D. H. (2016). MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. bioRxiv, 050559.

Hoelzel, A. R., & Dover, G. A. (1991). Genetic differentiation between sympatric killer whale populations. *Heredity*, *66*(2), 191.

Hoelzel, A. R., Hey, J., Dahlheim, M. E., Nicholson, C., Burkanov, V., & Black, N. (2007). Evolution of population structure in a highly social top predator, the killer whale. *Molecular Biology and Evolution*, *24*(6), 1407-1415.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Molecular biology and evolution, 34(8), 2115-2122.

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome research*, *17*(3), 377-386.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics, 11(1), 119.

Johnson, S. S., Zaikova, E., Goerlitz, D. S., Bai, Y., & Tighe, S. W. (2017). Real-time DNA sequencing in the Antarctic dry valleys using the Oxford Nanopore sequencer. Journal of Biomolecular Techniques: JBT, 28(1), 2.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S. et al. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55.

Kircher, M., Sawyer, S., & Meyer, M. (2011). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic acids research, 40(1), e3-e3.

Kolodny, O., Weinberg, M., Reshef, L., Harten, L., Hefetz, A., Gophna, U., Feldman, M. W., & Yovel, Y. (2017). Who is the host of the host-associated microbiome? Colony-level dynamics overshadow individual-level characteristics in the fur microbiome of a social mammal, the Egyptian fruit-bat. *bioRxiv*, 232934.

Kong, H. H., Oh, J., Deming, C., Conlan, S., Grice, E. A., Beatson, M. A., Nomicos, E., Polley, E.C., Komarow, H.D., Murray, P.R. & Turner, M. L. (2012). Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome research*, *22*(5), 850-859.

Kong, H. H., & Segre, J. A. (2012). Skin microbiome: looking back to move forward. *Journal of Investigative Dermatology*, *132*(3), 933-939.

Konishi, K., Tamura, T., Zenitani, R., Bando, T., Kato, H., & Walløe, L. (2008). Decline in energy storage in the Antarctic minke whale (Balaenoptera bonaerensis) in the Southern Ocean. Polar Biology, 31(12), 1509-1520.

Koskella, B., Hall, L. J., & Metcalf, C. J. E. (2017). The microbiome beyond the horizon of ecological and evolutionary theory. *Nature ecology & evolution*, *1*(11), 1606.

Kueneman, J. G., Parfrey, L. W., Woodhams, D. C., Archer, H. M., Knight, R., & McKenzie, V. J. (2014). The amphibian skin-associated microbiome across species, space and life history stages. *Molecular Ecology*, *23*(6), 1238-1250.

Lax, S., Smith, D. P., Hampton-Marcell, J., Owens, S. M., Handley, K. M., Scott, N. M., Gibbons, S.M., Larsen, P., Shogan, B.D., Weiss, S., & Metcalf, J. L. (2014). Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science*, *345*(6200), 1048-1052.

Lemieux-Labonté, V., Tromas, N., Shapiro, B. J., & Lapointe, F. J. (2016). Environment and host species shape the skin microbiome of captive neotropical bats. *PeerJ*, *4*, e2430.

Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R., & Gordon, J. I. (2008). Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology*, *6*(10), 776.

Leyden, J. J., McGiley, K. J., Mills, O. H., & Kligman, A. M. (1975). Age-related changes in the resident bacterial flora of the human face. *Journal of Investigative Dermatology*, *65*(4), 379-381.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Li, H., Qu, J., Li, T., Li, J., Lin, Q., & Li, X. (2016). Pika population density is associated with the composition and diversity of gut microbiota. *Frontiers in microbiology*, *7*, 758.

Lusk, R. W. (2014). Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PloS one*, *9*(10), e110808.

Mangul S. , M Olde Loohuis L., Ori A., Jospin G., Koslicki D., Yang H. T., Wu T., Boks M. P., Lomen-Hoerth C., Wiedau-Pazos M, … & Ophoff, R. A. (2016) Total RNA Sequencing reveals microbial communities in human blood and disease specific effects. *bioRxiv*. 057570

Matkin, C. O., Barrett-Lennard, L. G., Yurk, H., Ellifrit, D., & Trites, A. W. (2007). Ecotypic variation and predatory behavior among killer whales (*Orcinus orca*) off the eastern Aleutian Islands, Alaska. *Fishery Bulletin*, *105*(1), 74-88.

McFall-Ngai, M. J., Henderson, B., & Ruby, E. G. (Eds.). (2005). *The influence of cooperative bacteria on animal host biology* (Vol. 9). Cambridge University Press.

McKenzie, V. J., Bowers, R. M., Fierer, N., Knight, R., & Lauber, C. L. (2012). Co-habiting amphibian species harbor unique skin bacterial communities in wild populations. *The ISME journal*, *6*(3), 588.

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, 31(10), 1674-1676.

Martinez Arbizu, P. (2017). pairwiseAdonis: Pairwise multilevel comparison using adonis. R package version 0.0.1.

Menke, S., Melzheimer, J., Thalwitzer, S., Heinrich, S., Wachter, B., & Sommer, S. (2017). Gut microbiomes of free-ranging and captive Namibian cheetahs: diversity, putative functions, and occurrence of potential pathogens. *Molecular ecology*.

Mollerup, S., Friis-Nielsen, J., Vinner, L., Hansen, T. A., Richter, S. R., Fridholm, H., ... & Mourier, T. (2016). Propionibacterium acnes: disease-causing agent or common contaminant? Detection in diverse patient samples by next-generation sequencing. *Journal of clinical microbiology*, *54*(4), 980-987.

Morin, P. A., Archer, F. I., Foote, A. D., Vilstrup, J., Allen, E. E., Wade, P., ... & Bouffard, P. (2010). Complete mitochondrial genome phylogeographic analysis of killer whales (Orcinus orca) indicates multiple species. *Genome research*, *20*(7), 908-916.

Morueta-Holme, N., Blonder, B., Sandel, B., McGill, B. J., Peet, R. K., Ott, J. E., ... & Svenning, J. C. (2016). A network approach for inferring species associations from co-occurrence data. Ecography, 39(12), 1139-1150.

Moura, A. E., Janse van Rensburg, C., Pilot, M., Tehrani, A., Best, P. B., Thornton, M., ... & Dahlheim, M. E. (2014). Killer whale nuclear genome and mtDNA reveal widespread population bottleneck during the last glacial maximum. *Molecular biology and evolution*, *31*(5), 1121-1131.

Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C., & Pati, A. (2015). Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Standards in genomic sciences*, *10*(1), 18.

Naccache, S. N., Greninger, A. L., Lee, D., Coffey, L. L., Phan, T., Rein-Weston, A., ... & Chiu, C. Y. (2013). The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *Journal of virology*, *87*(22), 11966-11977.

Nater, A., Mattle-Greminger, M. P., Nurcahyo, A., Nowak, M. G., de Manuel, M., Desai, T., ... & Lameira, A. R. (2017). Morphometric, behavioral, and genomic evidence for a new Orangutan species. *Current Biology*, *27*(22), 3487-3498.

Nelson, T. M., Apprill, A., Mann, J., Rogers, T. L., & Brown, M. V. (2015). The marine mammal microbiome: current knowledge and future directions. *Microbiology Australia*, *36*(1), 8-13.

Oksanen, J., Guillaume Blanchet, F., Kindt, R., & Legendre, P. (2017). others. 2016. vegan: Community ecology package. R package version 2.3–5.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., ... & Fonstein, M. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research*, *33*(17), 5691-5702.

Palsbøll, P. J., F. Larsen, And E. S. Hansen. 1991. Sampling of skin biopsies from free-ranging large cetaceans in West Greenland; Development of new biopsy tips and bolt designs. *Rep. Int. Whaling Comm.*, Special Issue, **13**: 71–79.

Parsons, K. M., Durban, J. W., Burdin, A. M., Burkanov, V. N., Pitman, R. L., Barlow, J., ... & Wade, P. R. (2013). Geographic patterns of genetic differentiation among killer whales in the northern North Pacific. *Journal of Heredity*, *104*(6), 737-754.

Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. Nature methods, 10(12), 1200.

Petersen, T. N., Lukjancenko, O., Thomsen, M. C. F., Sperotto, M. M., Lund, O., Aarestrup, F. M., & Sicheritz-Pontén, T. (2017). MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads. *PloS one*, *12*(5), e0176469.

Phillips, C. D., Phelan, G., Dowd, S. E., McDonough, M. M., Ferguson, A. W., Delton Hanson, J., Siles, L., Ordóñez-Garza, N.I.C.T.É., San Francisco, M. & Baker, R. J. (2012). Microbiome analysis among bats describes influences of host phylogeny, life history, physiology and geography. *Molecular ecology*, *21*(11), 2617-2627.

Piñeiro-Vidal, M., Gijón, D., Zarza, C., & Santos, Y. (2012). Tenacibaculum dicentrarchi sp. nov., a marine bacterium of the family Flavobacteriaceae isolated from European sea bass. International journal of systematic and evolutionary microbiology, 62(2), 425-429.

Pitman, R. L., & Durban, J. W. (2010). Killer whale predation on penguins in Antarctica. *Polar Biology*, *33*(11), 1589-1594.

Pitman, R. L., & Durban, J. W. (2012). Cooperative hunting behavior, prey selectivity and prey handling by pack ice killer whales (Orcinus orca), type B, in Antarctic Peninsula waters. *Marine Mammal Science*, *28*(1), 16-36.

Pitman, R. L., & Ensor, P. (2003). Three forms of killer whales (Orcinus orca) in Antarctic waters. *Journal of Cetacean Research and Management*, *5*(2), 131-140.

Pitman, R. L., Fearnbach, H. & Durban, J. W. (2018). Abundance and population status of Ross Sea killer whales (*Orcinus orca*, type C) in McMurdo Sound, Antarctica: Evidence for impact by commercial fishing in the Ross Sea? Polar Biology DOI:/10.1007/s00300-017-2239-4

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Müller, I., ... & Wolf, J. B. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, *344*(6190), 1410-1414.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature biotechnology*, *35*(9), 833.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841-842.

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and biophysical research communications*, *469*(4), 967-977.

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome research, 22(5), 939-946.

Romano-Bertrand, S., Licznar-Fajardo, P., Parer, S., & Jumas-Bilak, E. (2015). Impact de l'environnement sur les microbiotes: focus sur l'hospitalisation et les microbiotes cutanés et chirurgicaux. *Revue Francophone des Laboratoires*, *2015*(469), 75-82.

Rothschild, D., Weissbrod, O., Barkan, E., Korem, T., Zeevi, D., Costea, P. I., ... & Pevsner-Fischer, M. (2018). Environmental factors dominate over host genetics in shaping human gut microbiota composition. *Nature* doi:10.1038/nature25973

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., ... & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology*, *12*(1), 87.

Der Sarkissian, C., Ermini, L., Schubert, M., Yang, M. A., Librado, P., Fumagalli, M., ... & Petersen, B. (2015). Evolutionary genomics and conservation of the endangered Przewalski's horse. *Current Biology*, *25*(19), 2577-2583.

Saulitis, E., Matkin, C., Barrett-Lennard, L., Heise, K., & Ellis, G. (2000). Foraging strategies of sympatric killer whale (Orcinus orca) populations in Prince William Sound, Alaska. *Marine mammal science*, *16*(1), 94-109.

Scharschmidt, T. C., & Fischbach, M. A. (2013). What lives on our skin: ecology, genomics and therapeutic opportunities of the skin microbiome. *Drug Discovery Today: Disease Mechanisms*, *10*(3-4), e83-e89.

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863-864.

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC research notes*, *9*(1), 88.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. Genome biology, 12(6), R60.

Shotts Jr, E.B., Albert, T. F., Wooley, R. E., & Brown, J. (1990). Microflora associated with the skin of the bowhead whale *(Balaena mysticetus*). *Journal of Wildlife Disease,* 26, 351-359.

Song, S. J., Lauber, C., Costello, E. K., Lozupone, C. A., Humphrey, G., Berg-Lyons, D., Caporaso, J. G., Knights, D., Clemente, J. C., Nakielny, S. & Gordon, J. I. (2013). Cohabiting family members share microbiota with one another and with their dogs. *elife*, *2*.

Sun, C., Fu, G. Y., Zhang, C. Y., Hu, J., Xu, L., Wang, R. J., ... & Zhang, X. Q. (2016). Isolation and complete genome sequence of Algibacter alginolytica sp. nov., a novel seaweed-degrading Bacteroidetes bacterium with diverse putative polysaccharide utilization loci. Applied and environmental microbiology, 82(10), 2975-2987.

Tessler, M., Neumann, J. S., Afshinnekoo, E., Pineda, M., Hersch, R., Velho, L. F. M., ... & Mason, C. E. (2017). Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Scientific reports, 7(1), 6589.

Tung, J., Barreiro, L. B., Burns, M. B., Grenier, J. C., Lynch, J., Grieneisen, L. E., ... & Archie, E. A. (2015). Social networks predict gut microbiome composition in wild baboons. *Elife*, *4*.

van der Valk T, Vezzi F, Ormestad M, Dalén L, Guschanski K. Low rate of index hopping on the Illumina HiSeq X platform. Biotechniques. BioRxiv, doi: 10.1101/179028).

Walke, J. B., Becker, M. H., Loftus, S. C., House, L. L., Cormier, G., Jensen, R. V., & Belden, L. K. (2014). Amphibian skin may select for rare environmental microbes. *The ISME journal*, *8*(11), 2207.

Warinner, C., Herbig, A., Mann, A., Fellows Yates, J. A., Weiß, C. L., Burbano, H. A., ... & Krause, J. (2017). A robust framework for microbial archaeology. *Annual review of genomics and human genetics*, *18*, 321-356.

Wright, E. S., & Vetsigian, K. H. (2016). Inhibitory interactions promote frequent bistability among competing bacteria. Nature communications, 7, 11274.

Wright, E. S., & Vetsigian, K. H. (2016). Quality filtering of Illumina index reads mitigates sample cross-talk. BMC genomics, 17(1), 876.

Wolz, M., Carly, R., Yarwood, S. A., Grant, E. H. C., Fleischer, R. C., & Lips, K. R. (2017). Effects of host species and environment on the skin microbiome of Plethodontid salamanders. *Journal of Animal Ecology*.

Ying, S., Zeng, D. N., Chi, L., Tan, Y., Galzote, C., Cardona, C., Lax, S., Gilbert, J. & Quan, Z. X. (2015). The influence of age and gender on skin-associated microbial communities in urban and rural human populations. *PloS one*, *10*(10), e0141842.

Zhang C., Cleveland K., Schnoll-Sussman F., McClure B., Bigg M., Thakkar P., Schultz N., Shah M. A., Betel D (2015) Identification of low abundance microbiome in clinical samples using whole genome sequencing. *Genome Biology*, 16(1).

FIGURES

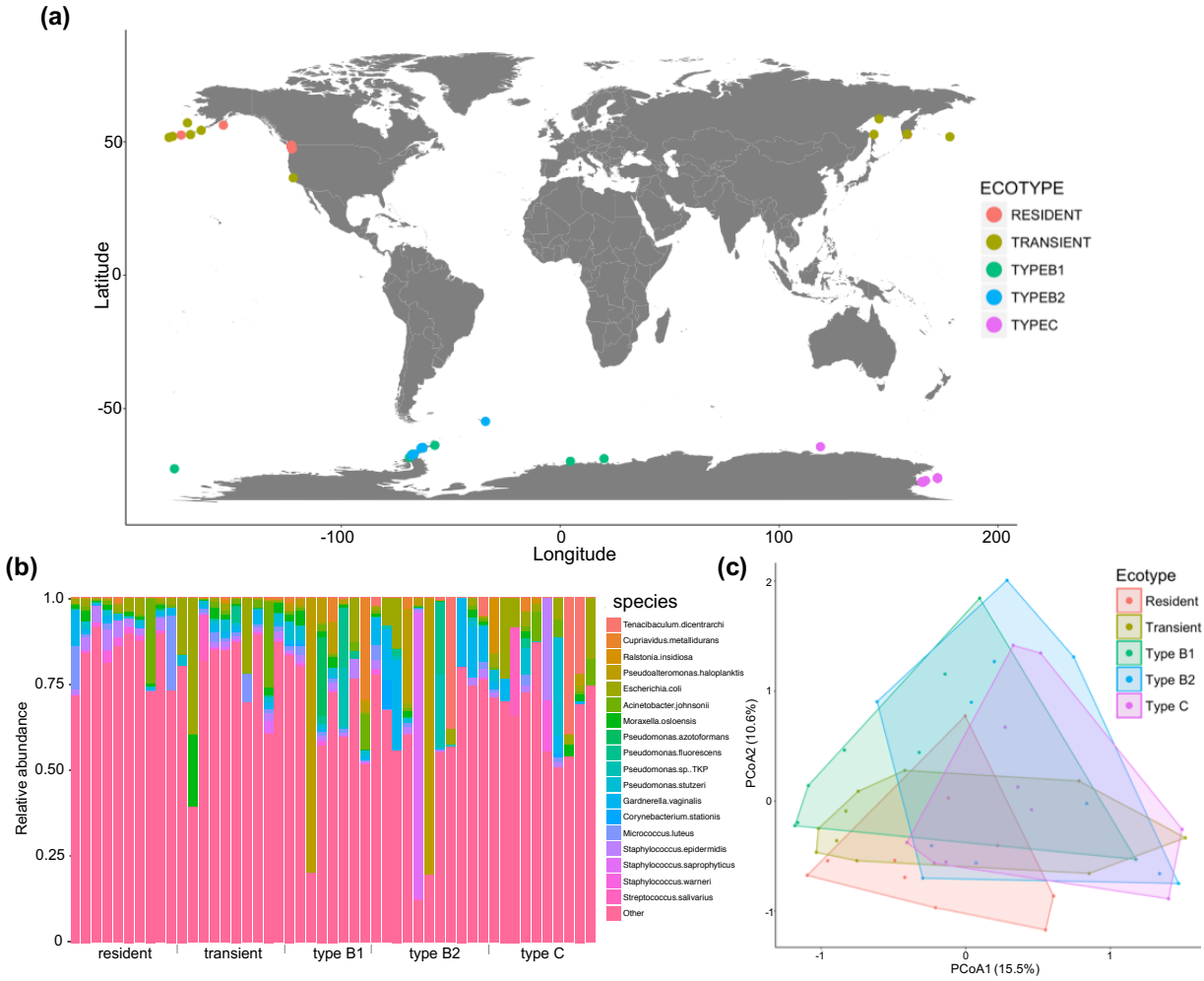

FIGURE 1 Summary of geographic sample origins and microbial diversity measures. (a) Map of sampling locations of the 49 samples of five killer whale ecotypes, from which skin microbiomes were included in this study. The Antarctic ecotypes primarily inhabit waters 8–16°C colder than the North Pacific ecotypes. (b) Alpha diversity after the removal of *C. acnes*. Individuals, represented by columns, are grouped by ecotype, and the relative proportions of bacterial taxa are indicated by column shading. (c) Beta diversity between ecotypes illustrated as a Bray-Curtis PCoA estimated from read counts.

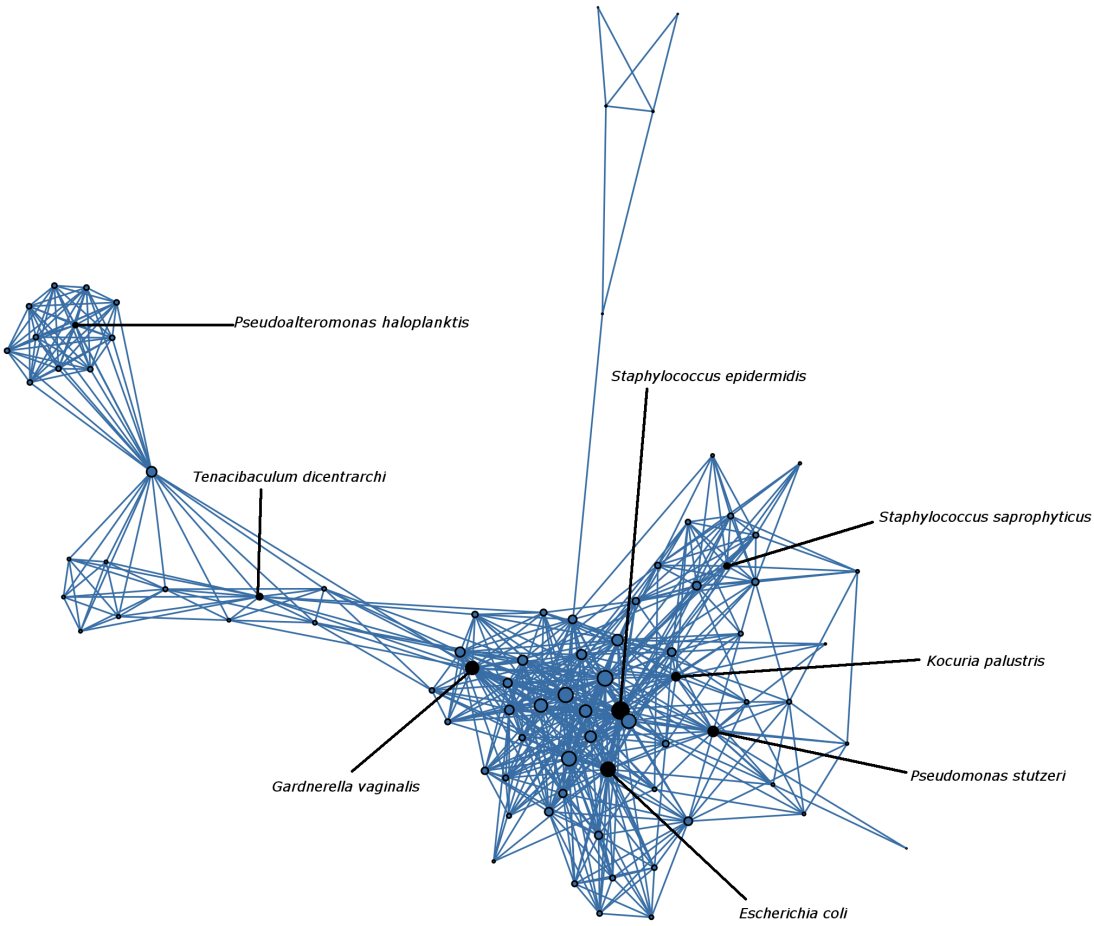

FIGURE 2 **Positive co-occurrence network** built from a co-occurrence matrix of all species, subsetted to the 9 driving taxa (black nodes) and their top 10 positive and significant co-occurring species. Only species with a co-occurrence score of >500 are shown. One driving species, *Salinibacter ruber,* is not included because it did not have any co-occurrences exceeding the score 500.

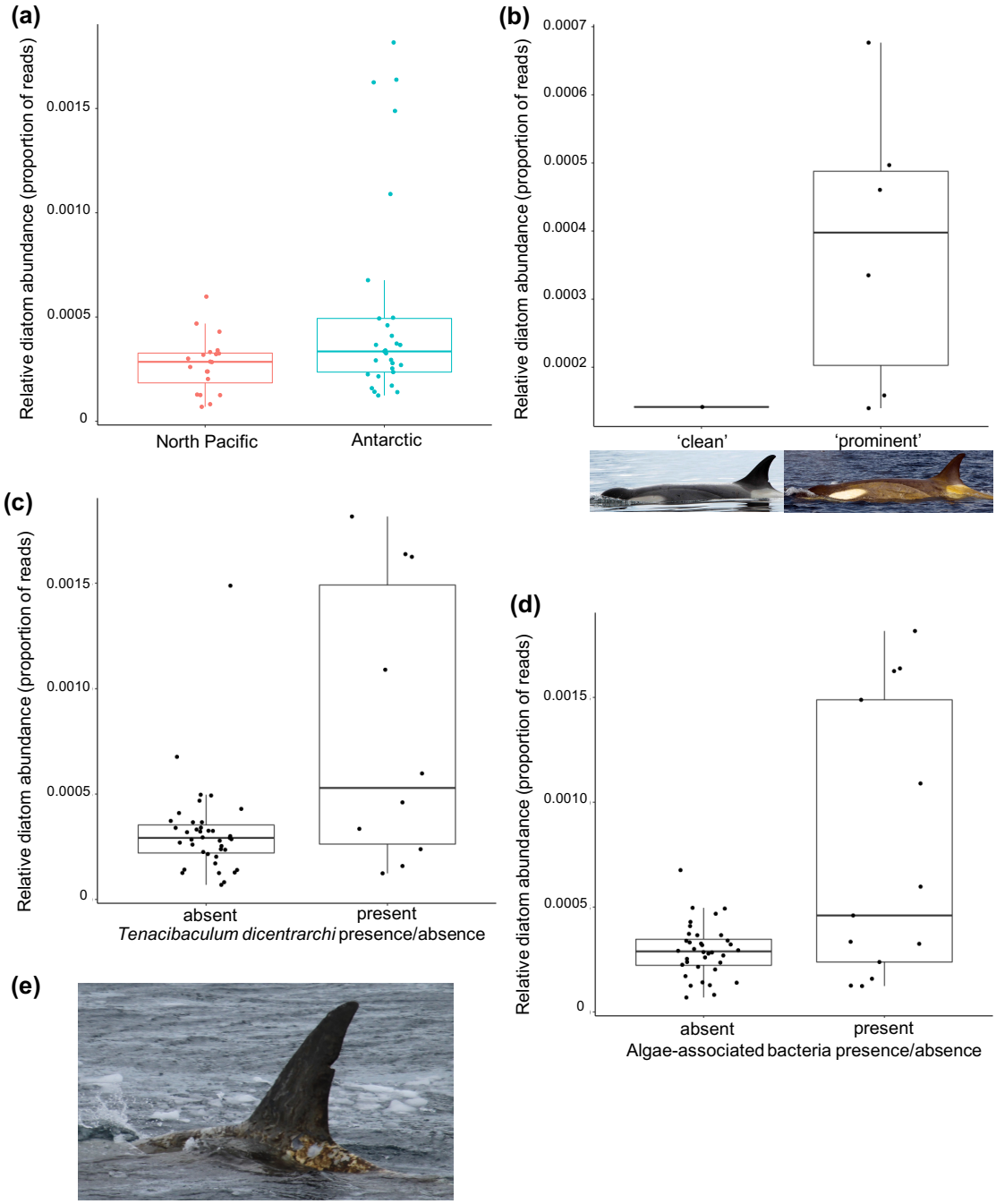

FIGURE 3 The influence of diatom abundance on skin microbiome community composition. (a) Relative diatom abundance is significantly higher in Antarctic killer whales than North Pacific, but this is largely driven by a subset of outlier Antarctic individuals. (b) Within Antarctic *type B1* and *type B2* specimens, the relative diatom abundance is significantly associated with skin colouration of the host killer whale, with the yellowish hue being a reliable indicator of diatom load. Inset images are of the same *type B2* killer whale displaying extreme variability in diatom coverage, both photographs by John Durban. Relative diatom abundance is significantly associated with (c) the presence of *Tenacibaculum dicentrarchi* and (d) several algae-associated bacteria, including *T. dicentrarchi*. (e) Photograph of a *type B1* killer whale in the Gerlache Strait of the Antarctic Peninsula on the 4th December 2015 with high diatom coverage and poor skin condition. Photograph by Conor Ryan.

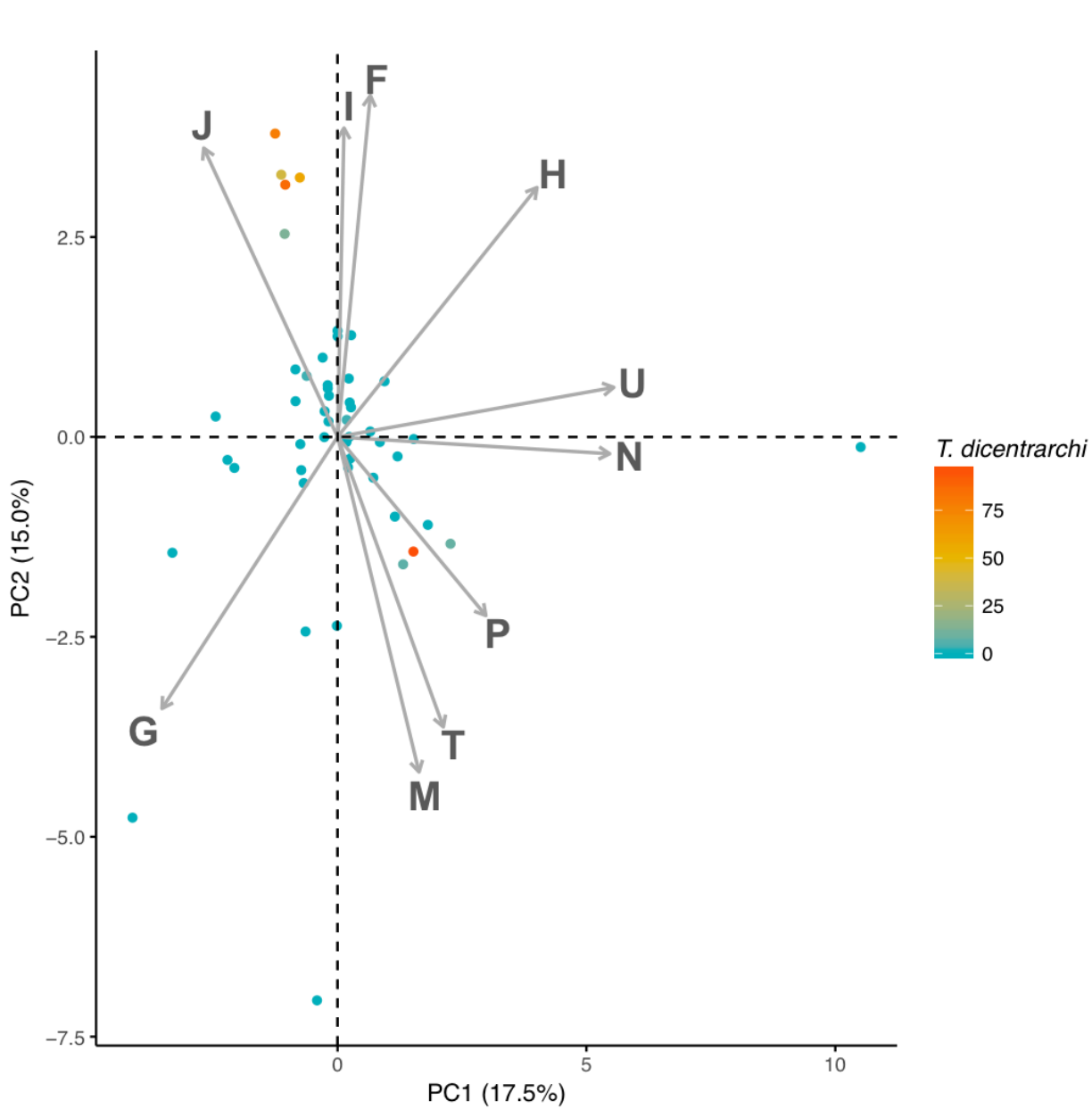

FIGURE 4 PCA of variation in functional COGs between individuals, coloured by *T. dicentrarchi* abundance. Individuals with high relative abundances of *T. dicentrarchi* generally cluster with high values in principal component 2. The top 10 COGs contributing to PCA variation are shown in grey arrows (J: translation, ribosomal structure and biogenesis, I: lipid transport and metabolism, F: nucleotide transport and metabolism, H: coenzyme transport and metabolism, U: intracellular trafficking, secretion and vesicular transport, N: cell motility, P: inorganic ion transport and metabolism, T: signal transduction mechanisms, M: cell wall membrane envelope biogenesis, G: carbohydrate transport and metabolism).

TABLES

Table 1. Factors influencing the killer whale skin microbiome. Results of Adonis models using genome-size controlled species data. (a) Bray Curtis model with library size included as a covariate; (b) TSS normalised Bray Curtis model; (c) Binary Jaccard model with library size included as a covariate. Significant factors are highlighted in bold.

	A) Bray Curtis			**B) Bray Curtis (TSS normalised)**			**C) Binary Jaccard**		
	F.Model	R^2	P	F.Model	R^2	P	F.Model	R^2	P
Latitude	1.9	0.04	**0.01**	2.19	0.04	**<0.01**	1.6	0.03	**0.03**
Longitude	1.08	0.02	0.32	1.09	0.02	0.32	1.09	0.02	0.27
Ecotype	1.53	0.12	**<0.01**	1.57	0.12	**<0.01**	1.35	0.11	**0.01**
Library size	1.56	0.03	**0.01**	-	-	-	1.33	0.03	0.11
Human contamination	1.72	0.03	**0.01**	1.17	0.02	0.26	1.33	0.03	0.1
Residuals	0.76			0.79			0.79		
Total		1			1			1	