scientific reports

OPEN

A collaborative and near-comprehensive North Pacific humpback whale photo-ID dataset

Ted Cheeseman[1,2]✉, Ken Southerland[1], Jo Marie Acebes[3], Katherina Audley[4], Jay Barlow[5], Lars Bejder[6], Caitlin Birdsall[7,8], Amanda L. Bradford[9], Josie K. Byington[10], John Calambokidis[11], Rachel Cartwright[12], Jen Cedarleaf[13], Andrea Jacqueline García Chavez[4], Jens J. Currie[14], Joëlle De Weerdt[15], Nicole Doe[7], Thomas Doniol-Valcroze[16], Karina Dracott[8,17], Olga Filatova[18], Rachel Finn[19], Kiirsten Flynn[11], John K. B. Ford[16], Astrid Frisch-Jordán[20], Christine M. Gabriele[21,22], Beth Goodwin[23], Craig Hayslip[24], Jackie Hildering[7], Marie C. Hill[9,25], Jeff K. Jacobsen[26], M. Esther Jiménez-López[27], Meagan Jones[28], Nozomi Kobayashi[29], Edward Lyman[19], Mark Malleson[30], Evgeny Mamaev[31], Pamela Martínez Loustalot[32], Annie Masterman[33], Craig Matkin[34], Christie J. McMillan[7,16], Jeff E. Moore[5], John R. Moran[33], Janet L. Neilson[21], Hayley Newell[1], Haruna Okabe[29], Marilia Olio[1], Adam A. Pack[35,36], Daniel M. Palacios[24,37], Heidi C. Pearson[13], Ester Quintana-Rizzo[38], Raul Fernando Ramírez Barragán[4], Nicola Ransome[39], Hiram Rosales-Nanduca[27], Fred Sharpe[40], Tasli Shaw[30], Stephanie H. Stack[14], Iain Staniland[41], Jan Straley[13], Andrew Szabo[40], Suzie Teerlink[42], Olga Titova[43], Jorge Urban R.[32], Martin van Aswegen[18], Marcel Vinicius de Morais[1], Olga von Ziegesar[44], Briana Witteveen[45], Janie Wray[17], Kymberly M. Yano[9,25], Denny Zwiefelhofer[1] & Phil Clapham[46]

We present an ocean-basin-scale dataset that includes tail fluke photographic identification (photo-ID) and encounter data for most living individual humpback whales (*Megaptera novaeangliae*) in the North Pacific Ocean. The dataset was built through a broad collaboration combining 39 separate curated photo-ID catalogs, supplemented with community science data. Data from throughout the North Pacific were aggregated into 13 regions, including six breeding regions, six feeding regions, and one migratory corridor. All images were compared with minimal pre-processing using a recently developed image recognition algorithm based on machine learning through artificial intelligence; this system is capable of rapidly detecting matches between individuals with an estimated 97–99% accuracy. For the 2001–2021 study period, a total of 27,956 unique individuals were documented in 157,350 encounters. Each individual was encountered, on average, in 5.6 sampling periods (i.e., breeding and feeding seasons), with an annual average of 87% of whales encountered in more than one season. The combined dataset and image recognition tool represents a living and accessible resource for collaborative, basin-wide studies of a keystone marine mammal in a time of rapid ecological change.

[1]Happywhale, Santa Cruz, California, USA. [2]Southern Cross University, Lismore, NSW, Australia. [3]BALYENA.ORG, Jagna, Philippines. [4]Whales of Guerrero, Barra de Potosí, Mexico. [5]NOAA Southwest Fisheries Science Center, San Diego, California, USA. [6]Hawai‘i Institute of Marine Biology, University of Hawai‘i at Manoa, Kaneohe, Hawai‘i, USA. [7]Marine Education and Research Society, Port McNeill, British Columbia, Canada. [8]Ocean Wise, Vancouver, British Columbia, Canada. [9]NOAA Fisheries Pacific Islands Fisheries Science Center, Honolulu, Hawai‘i, USA. [10]Pacific Wildlife Foundation, Port Moody, British Columbia, Canada. [11]Cascadia Research Collective, Olympia, Washington, USA. [12]Keiki Kohola Project, Delray Beach, Hawai‘i, USA. [13]University of Alaska Southeast, Juneau, Alaska, USA. [14]Pacific Whale Foundation, Maui, Hawai‘i, USA. [15]Association ELI-S, Gujan-Mestras, France. [16]Fisheries and Oceans Canada, Nanaimo, British Columbia, Canada. [17]North Coast Cetacean Society, Hartley Bay, British Columbia, Canada. [18]University of Southern Denmark, Odense, Denmark. [19]NOAA Hawaiian Islands Humpback Whale National Marine Sanctuary, Kihei, Maui, Hawaii, USA. [20]Ecología y Conservación de Ballenas, Puerto Vallarta, Mexico. [21]Glacier Bay National Park and Preserve, Gustavus, Alaska, USA. [22]Hawai‘i

Marine Mammal Consortium, Kamuela, Hawaiʻi, USA. [23]Eye of the Whale Marine Mammal Research, Kamuela, Hawaiʻi, USA. [24]Marine Mammal Institute, Oregon State University, Newport, Oregon, USA. [25]Cooperative Institution of Marine and Atmospheric Research, Research Corporation of the University of Hawaiʻi, Honolulu, Hawaiʻi, USA. [26]VE Enterprises, Arcata, California, USA. [27]Departamento Académico de Ingeniería en Pesquerías, Universidad Autónoma de Baja California Sur, La Paz, Baja California Sur, México. [28]Whale Trust, Puunene, Hawaiʻi, USA. [29]Okinawa Churashima Foundation, Kunigami-gun, Japan. [30]Humpback Whales of the Salish Sea, Duncan, British Columbia, Canada. [31]Commander Islands National Park, Kamchatka Krai, Russian Federation. [32]Universidad Autónoma de Baja California Sur, La Paz, Mexico. [33]NOAA Alaska Fisheries Science Center, Juneau, Alaska, USA. [34]North Gulf Oceanic Society, Homer, Alaska, USA. [35]University of Hawaiʻi at Hilo, Hilo, Hawaiʻi, USA. [36]The Dolphin Institute, Hilo, Hawaiʻi, USA. [37]Department of Fisheries, Wildlife, and Conservation Sciences, Oregon State University, Newport, Oregon, USA. [38]Emmanuel College, Boston, Massachusetts, USA. [39]Murdoch University, Perth, WA, Australia. [40]Alaska Whale Foundation, Petersburg, Alaska, USA. [41]International Whaling Commission, Impington, UK. [42]NOAA Fisheries Alaska Regional Office, Juneau, Alaska, USA. [43]Severtsov Institute of Ecology and Evolution, Moscow, Russian Federation. [44]Winged Whale Research, Homer, Alaska, USA. [45]University of Alaska Fairbanks, Fairbanks, Alaska, USA. [46]Seastar Scientific, Vashon Island, Washington, USA. ✉email: ted@happywhale.com

Understanding the population ecology of a species is crucial for conservation management, but studies of most migratory marine species are compromised by data deficiency. Individual identification through techniques such as photographic identification (photo-ID), radio telemetry, and genetic sequencing allow researchers to track individual animals over time. This enables population modeling, revealing movement patterns, social interactions, and reproductive success rates. Photo-ID, in which a photograph of persistently identifiable features of an individual is recorded together with its date and location, offers an efficient and non-invasive data collection method[1]. For long-lived migratory species, effective population studies require extensive data collection, including the additional challenges of collaboration across regional and international boundaries.

The humpback whale, *Megaptera novaeangliae*, is a globally distributed baleen whale species with a complex population structure and major ecosystem impacts[2–4]. Individuals engage in extensive seasonal migrations between high-latitude feeding areas during the spring, summer, and fall, and low-latitude tropical waters to mate and calve in winter and spring[2,4–8]. The long-distance migrations undertaken by humpback whales expose populations to diverse management regimes, anthropogenic risks, and ecological conditions[9]. For example, a very large marine heatwave in the North Pacific from late 2013–2016[10–12] caused major negative impacts on humpback whale food resource availability. This resulted in sharp declines in abundance, survival, and reproductive success of humpback whales in Hawaiʻi and Southeast Alaska[13–17]. In a changing oceanic ecosystem, a cost-effective and non-invasive technique that repeatedly samples most living individuals can offer valuable insights into the status of the species and its ecosystem.

Humpback whale populations worldwide were severely depleted by extensive commercial whaling until late in the twentieth century. This species was listed under the U.S. Endangered Species Act (ESA) in 1970 due to an estimated 31,785 killed in the North Pacific from 1900 to 1979[18–20]. Following a global ban on humpback whale catches by the International Whaling Commission in 1966, and the cessation of Soviet illegal whaling in the following decade[19], the humpback whale population has grown. Two studies have evaluated the abundance of humpback whales in the full North Pacific: first in the 1990s[4], then a study entitled Structure of Populations, Levels of Abundance and Status of Humpback Whales (SPLASH) conducted from 2004 to 2006[8]. These studies estimated total North Pacific humpback whale abundance at 21,063 individuals in 2006, with an annual growth rate of 8.1% between the two study periods[21]. A major portion of SPLASH relied on the identification and resighting of individual humpback whales through photo-ID. This method involved trained observers visually matching photographs of the ventral side of each whale's tail (flukes) based on unique white and black pigmentation patterns, together with unique fluke trailing edge contours[22,23]. SPLASH documented 7,640 individual humpback whales in 18,469 unique encounters (defined as a single sighting of a unique individual supported by a referenced photo-ID image, recorded on a specific day at a specific location); these encounters occurred across all known breeding and feeding areas. SPLASH reinforced the value of broad-scale data sharing and collaboration, and exposed gaps in knowledge of humpback whale status in the North Pacific.

In 2016, NOAA Fisheries, pursuant to the ESA, defined 14 humpback Distinct Population Segments (DPSs) globally using photo-ID data and other lines of evidence[24]. DPS designations are based on theoretically discrete breeding areas where many whales show long-term site fidelity[25]. In feeding areas, whales also show high site fidelity and arguably face greater biological and anthropogenic stressors[26]. Four DPSs occur in the North Pacific, with breeding occurring in waters off Central America, Mexico, Hawaiʻi, and the Western North Pacific (Mariana Islands, the Philippines, and Japan). Based on varying rates of recovery, the Central America and Western North Pacific DPS remain listed as Endangered (s), the Mexico DPS is considered Threatened, and the Hawaiʻi DPS has been deemed to not warrant listing[27]. Ironically, removal of the Hawaiʻi DPS's endangered status by the US coincided with the 2013–2016 marine heatwave that negatively affected population health[13–16,28].

Individual photo-ID data have advanced the understanding of humpback whale behavior, ecology and conservation issues based on many regional study efforts[13,14,25,29–41]. However, after SPLASH ended in 2006, local and regional photo-ID datasets were seldom integrated with one another. This was in part due to prohibitively time-intensive visual matching of individual ID fluke photos in ever-growing catalogs. The current study established the North Pacific Humpback Whale Photo-ID (NPPID) collaboration. The goal of this collaboration was to integrate and advance knowledge of humpback whale population structure and migratory movement in the North Pacific through creation of a shared repository of resighting data for individual whales across the full study region. A central objective of the effort was to implement a collaborative framework to facilitate data availability, access, and readiness. Given the large amount of data involved and the difficulty of obtaining long-term funding, to be successful the system needed to drive the incremental cost of acquisition of each successive datapoint to near zero. Such a system required effective technology and web-based data management to submit, quality-control, identify, and curate encounter data for a growing set of known individual whales. The NPPID was built on newly established automated fluke photo-ID matching technology. This technology achieves a measured 97–99% accuracy with good- to high-quality images and is orders of magnitude faster than manual visual matching[42]. However, a system is not technology alone; the system needed to sustainably nurture positive collaboration practices to bring together the many contributors responsible for tens of thousands of whale encounters per year. Therefore, the NPPID was developed as a shared effort utilizing the user-friendly and interactive web-based platform, www.Happywhale.com (Happywhale). Here we describe the process of building this ocean-basin-wide ongoing photo-ID collaboration involving 43 research groups and thousands of public contributors (also known as "community scientists" or "citizen scientists"). This approach has enabled rapid feedback for population and longitudinal studies of humpback whales across the North Pacific. The process and framework described here have broader practical relevance for navigating the use of complex multi-contributor datasets.

Materials and methods

The North Pacific humpback whale Photo-ID (NPPID) collaboration.

This effort began in 2018 as a data-sharing initiative to revive the collaboration established with the 2004–2006 SPLASH study[8], supplemented by photo-ID images from community scientists. We built upon the SPLASH dataset, study methodology, and collaboration, but did not have a budget for data acquisition or fieldwork. All SPLASH collaborators known to be active in North Pacific humpback whale studies were invited to join, along with all known newer regional researchers and organizations. Data collection relied on existing archives and ongoing field efforts by the individual collaborators. All dedicated data collection by study collaborators was carried out in accordance with permitting requirements of respective authorities (permit details are listed in acknowledgements). Data collection from community scientists was sourced primarily from whale watch companies operating under regulations and guidelines of respective national, regional, and local authorities. A primary incentive for participation in the NPPID collaboration was the promise of novel and fully automated image-recognition technology[42] that effectively eliminated the cumbersome, time-intensive visual matching process from photo-ID data management.

Through a memorandum of agreement (MOA, Supplementary Material I), all research organizations in the NPPID committed to sharing photo-ID images and associated supporting data for every available encounter, with a focus on a 2001–2021 study period. The specific research aim was to further knowledge of population structure and migratory movement via photographic mark-recapture population model development [e.g.21,43]. The broader aim was to create an ongoing, living dataset for continued population monitoring. Under the MOA terms, each data contributor chose whether their data were publicly visible via Happywhale or visible only to collaborators who had signed the MOA. The MOA limited data use to a defined set of publications about population status and migratory patterns; any additional use required agreement from all collaborators. The infrastructure, compiled data, and collaborator connections will remain after the period of the current MOA. Therefore, its use needs to be addressed with further agreement among collaborators if the dataset is going to be an ongoing, living entity.

Data integration and quality control.

Humpback whale encounter data were delivered to Happywhale data managers from collaborators in a wide range of states of reconciliation, from unmatched original scans and digital photos to fully edited sets of images (i.e., exposure adjusted as needed and cropped tightly around flukes), with IDs assigned to each individual whale. The minimum data required for each encounter were: date, location, and photo-ID image or confirmed individual ID. All encounters of each whale were preserved, and all available supporting attribute data were maintained with the encounter; this could include filename, date, time, location, individual ID from the collaborator's naming/numbering system, observer names, vessel name, observed whale sex, age class, health, behavior, group composition and any further observations. Because the state of every dataset varied at the time of delivery, all data were managed through the following standard series of steps:

1. **Image management and matching**: Images were quality-controlled through cropping tightly around the flukes and, if necessary, exposure adjustment to facilitate algorithmic ID followed by visual ID confirmation. All images were quality-scored on a 0–5 scale as described in a previous study[42], where 0 represented photos in which no photo-ID features were visible, and 1–5 represented very poor to excellent quality photos, respectively. All photo-ID images were matched to a progressively growing set of known whales via an automated image recognition system[42]. Every match proposed by the system was manually confirmed by a trained observer. All matches that could be visually confirmed by a trained observer were maintained regardless of image quality. A previous study established that 97–99% of potential matches are found by this method for good- to high-quality images[42].
2. **Supporting attribute data curation**: Given the diversity of supporting data formats received, standardization was necessary for dataset management. Locations were categorized as general (confident of location within 200 km [km]), approximate (confident of location within 20 km), or precise (confident of location within 2 km). Within the precise location category, location data source was categorized as: (a) camera GPS embedded into the image, (b) synchronous GPS track, (c) pinpoint recorded from a GPS unit, (d) pinpoint recorded via a mobile app, or (e) manually transcribed record. For encounters without a known date, an approximate date to month, season or year was assigned as information allowed, with date precision noted in encounter attributes. Encounters without a date known at least to year or location known confidently within 200 km were excluded. Descriptive observational data and contextual information such as whale sex, age class, behavior, mother/calf relationships or group composition, and scarring (e.g., from entanglement, ship

strike, killer whales) were recorded with each encounter when available and without standardization. Data quality was reviewed on import, with an opportunity for review by both data managers and data contributors before entry into a relational database.

3. **Efficiency with large datasets**: To increase efficiency for collaborators with large, well-curated datasets, some encounters were accepted with an individual ID name/number and supporting date, location, and attribute data, without a photo-ID image linked to every encounter. These encounters were linked to known individuals represented in one or more catalog photos.

Many-to-one reference catalogs.

All images were automatically matched against all individual humpback whales known at the time of each respective dataset integration. Across the NPPID study area, 39 separate catalog systems were received that had collaborator-specific individual IDs (Table 1). These ID naming systems were accommodated into a many-to-one ID structure so that any individual could be tracked via any of the multiple catalog IDs assigned to them.

Reconciliation of duplicate IDs.

Every image was matched within and among all collaborator catalogs. One individual ID per catalog was allowed. Thus, if individuals were found with duplicate IDs due to false negatives (where a previously undetected match of one whale with two or more separate IDs within a collaborator catalog was found), the contributor chose the persisting ID (typically the lowest of a sequential ID series). Each duplicate ID was noted in the attributes for the individual whale. Newly detected (i.e., unmatched) individuals were added to the continually growing reference set, with the collaborator ID, if available, or with a newly assigned Happywhale catalog ID. False positives (where two different whales were combined into one individual record) were minimized through trained observer review of every match.

Community science data contributions.

Opportunistic images submitted through Happywhale were also matched against all known whales, supplementing the research collaboration with community science-sourced encounter data. The same image and data quality control standards were applied as described above. All community science data contributors implicitly acknowledged their choice of data usage rights during the submission process and had the option of changing usage rights settings among established levels of Creative Commons usage rights (https://en.wikipedia.org/wiki/Creative_Commons_license). Unlike research collaborators participating under the terms of the MOA, public contributors did not have the option of restricting public visibility. Public contributors had access to an encounter comment system whereby suspected data errors and outliers could be brought to the attention of data managers, creating a feedback loop for review and error detection.

Information system structure and development.

The NPPID data management system integrated a workflow of image processing, individual identification, and recording and curating encounter and individual attribute information. Data were structured through units of contributors (i.e., "users"), images, encounters, individual humpback whales, and surveys (i.e., "voyages"), linked by a series of workflow processes (Fig. 1). The cloud-based information architecture was composed of a dedicated server for the Java Spring application using a PostgreSQL database populated with Darwin Core compatible fields[44]. Submitted binary media were stored in a Simple Storage Service (S3) system for global retrieval. The ID system used a combination of a Node server and a Python Flask app to run the PyTorch-based ID algorithm.

During the collaboration, ongoing system development brought enhanced functionality and sophistication to data management processes within the Happywhale.com web platform. In 2021, the automated image recognition system was rebuilt to deliver results in under 0.1 s per image. This efficiency reduced server load, which has accommodated direct access by collaborators to batch process photo-ID images directly via web and mobile app interfaces in the lab or field. Near-instantaneous access to image processing was adopted by many collaborators to facilitate more efficient and effective internal data management.

NPPID collaborators were invited to directly manage their data import process and ongoing curation, with training, feedback, and quality control oversight by system managers. Some collaborators used the system as a principal repository of their data while others maintained their own separate data management systems during the study. As import and management tools developed in a constantly evolving system, collaborators were increasingly enabled and encouraged to manage their own data.

System use, public outreach, and data accessibility.

The FAIR Principles (Findable, Accessible, Interoperable, and Reusable) for scientific data[45] guided system design. Public awareness of the opportunity to contribute to whale conservation science was spread through word of mouth, social media, and documentary films. The primary focus of outreach was to seek and reach naturalists, whale watch guides and enthusiasts already familiar with the concept of marine mammal photo-ID, and equipped with camera gear sufficient to create quality images. Community scientists and NPPID collaborators were promised they would be rewarded with knowledge. This was accomplished through a notification system with alerts to novel developments regarding individuals they had encountered (e.g., initial identification typically within a few days of submission, discovery of duplicate IDs, and ongoing resightings). Would-be contributors were directed to Happywhale with little guidance beyond a request for humpback whale photo-ID photos from any date and location, as long as the contributor could confirm the date and location. The data upload process sought to balance ease of access with rigor for data quality, with data validation dependent upon the image management process.

Data are searchable and accessible in 'map view' (Fig. 2) and 'list view' formats via Happywhale. Users can expand a search from a set of encounters (for example, all encounters contributed by one user or all encounters in a geographic area in a defined time period) to all sightings globally of individuals within the found set. This

Reference catalog	Code	Individuals	Number of 2001–2021 encounters	Regions
Alaska Whale Foundation	AWF	780	10,508	Southeast Alaska
BALYENA.ORG[75]	BALYENA	228	1193	Philippines
Bree Witteveen Alaska Catalog	BREE	1906	7959	Alaska Peninsula and Gulf of Alaska
MML Cetacean Assessment and Ecology Program	CAEP	412	3889	Pelagic North Pacific
Association ELI-S	CCN	124	1713	Nicaragua
Canadian Pacific Humpback Collaboration	CPHC	1197	14,569	British Columbia
Cascadia Research Collective[4,76–79]	CRC	7017	66,967	British Columbia to Central America, mostly US West Coast
Clayoquot and Barkley Sound	CS	530	6094	Central British Columbia
DFO Canada	DFO	1982	26,494	British Columbia
Eye of the Whale	EOTW	172	1589	Gulf of Alaska and Hawai‘i
ECOBAC FIBB Catalog	FIBB	2818	27,130	Mexico
Glacier Bay National Park & Preserve	GBNP	72	732	Glacier Bay and Icy Strait, Southeast Alaska
Gulf Watch Alaska	GWAK	361	2216	Kenai Fjords and Prince William Sound, Gulf of Alaska
Hawaiian Islands Humpback Whale National Marine Sanctuary	HIHWNMS	1051	7368	Hawai’i
Happywhale[42]	HW	11,851	78,070	Pan-Pacific
Juneau Flukes	JUNEAU	19	1180	Juneau, Southeast Alaska
Kachemak Bay Whales	KBAY	391	2188	Kachemak Bay, Gulf of Alaska
Keiki Kohola Project	KKP	33	198	Hawai’i
Marine Education and Research Society	MERS	308	7288	Central and southern British Columbia
Humpbacks of the Salish Sea (HWSS)	HWSS	483	7853	Salish Sea
Marine Mammals of Oaxaca	MMO	65	781	Oaxaca, Mexico
North CoastCetacean Society	NCCS	329	5169	Northern British Columbia
North Coast Cetacean Research Initiative	NCCRI	188	1860	Northern British Columbia
North Gulf Oceanic Society	NGOS	39	396	Gulf of Alaska
Okinawa Churashima Foundation[75]	OCF	1732	6322	Okinawa, Japan
Oregon State University Marine Mammal Institute Whale Habitat, Ecology, and Telemetry Laboratory[80]	OSUWTG	1242	14,765	Central and eastern Pacific
Pacific Islands Fisheries Science Center, NOAA Fisheries[35]	PIFSC	92	245	Mariana Islands and Northwestern Hawaiian
Pacific Whale Foundation	PWF	5022	33,346	Hawai’i-focused, pan-Pacific
Prince William Sound Catalog	PWS	184	1570	Prince William Sound, Alaska
Russian Cetacean Habitat Project[75,81]	RCHP	2028	6545	Russia
Sayulita Humpback Whale	SAYU	145	1376	Sayulita, Mexico
Southeast Alaska Humpback Whale Catalog	SEAK	2571	34,284	Southeast Alaska
SPLASH Project[8,21]	SPLASH	7838	64,144	Pan-Pacific
The Dolphin Institute[25]	TDI	1055	5970	Hawai’i and Southeast Alaska
The Marine Mammal Center	TMMC	67	1827	San Francisco Bay, California
Programa de Investigación de Mamiferos Marinos, Universidad Autónoma de Baja California Sur	UABCS	601	3950	Baja California Sur and Revillagigedo Islands, Mexico
Universidad Nacional Autónoma de Mexico	UNAM	1053	9045	Mexico
Whales of Guerrero	WGRP	329	5242	Guerrero, Mexico
Whale Trust	WTM	1994	15,708	Hawai’i

Table 1. Humpback whale photo-ID reference catalog naming systems integrated in this study. A single unified dataset allows cross-referencing of all known IDs for each individual. Some collaborating research groups share naming systems; all IDs are accounted for only once in the table below.

allows quick visual exploration of migratory connections for any set of whales. For collaborators, data are available for export into a standard comma-separated value (CSV) format, translatable to downstream analytical and research processes in GIS or statistical software.

Analytics—Documenting detection probability. The 2004–2006 SPLASH project actively developed collaborations and supported field efforts in all known (at the time) North Pacific humpback whale breeding and feeding areas, in pursuit of comprehensively representative sample sizes[8]. In contrast, the NPPID project relied on contributions from existing datasets, ongoing field efforts, and community science image contributions. With successive integration of datasets, detection probabilities progressively increased, leading to a predominance of resightings (documenting an individual multiple times) and fewer new whales added to the comprehensive catalog. This caused a shift in methodology from predominantly cataloging new whales to confirming matches

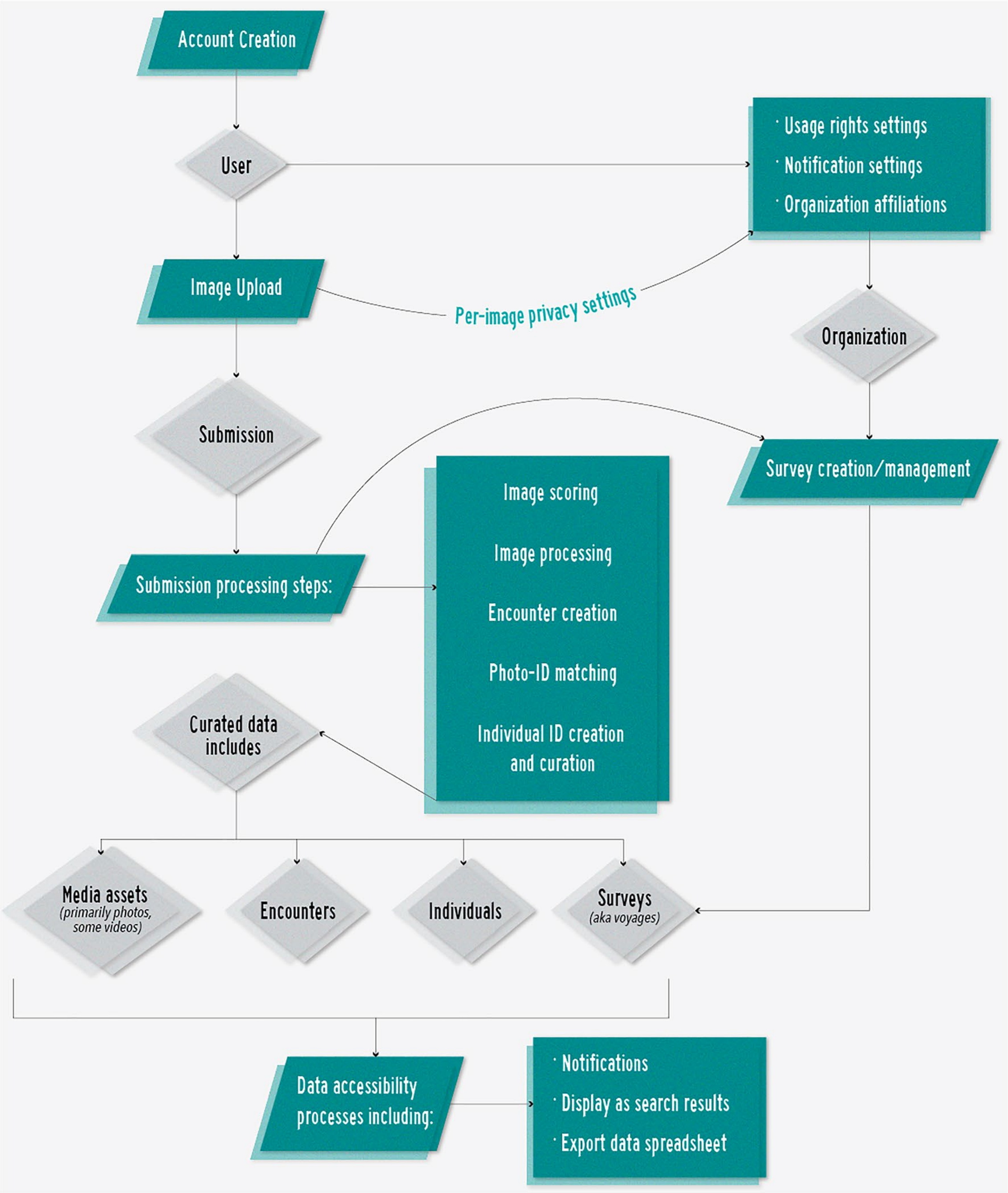
Account Creation
User
· Usage rights settings
· Notification settings
· Organization affiliations
Image Upload
Per-image privacy settings
Organization
Submission
Survey creation/management
Image scoring
Image processing
Encounter creation
Photo-ID matching
Individual ID creation and curation
Submission processing steps:
Curated data includes
Media assets (primarily photos, some videos)
Encounters
Individuals
Surveys (aka voyages)
Data accessibility processes including:
· Notifications
· Display as search results
· Export data spreadsheet

Figure 1. Happywhale simplified data process and elements. Processes (in teal) of user creation, media upload, submission processing and accessibility lead to creation of elements (in gray) of users, organizations, submissions, and curated data.

of known whales. To understand the proportion of the populations sampled in our growing known dataset, we plotted a discovery curve of new versus total identified individuals (Fig. 3), and a modified discovery curve of individuals identified over time (Fig. 4), in order to describe effort over the course of the history of the dataset.

Results

The NPPID collaboration involved 43 research organizations and included data from all nations around the North Pacific rim where humpback whales are known to regularly occur (Tables 1 and 2, Fig. 5). The complete NPPID collaboration ocean-basin dataset totaled 30,100 individual whales (February 1977 through August 2022 encompassing all available data). A total of 27,956 unique individuals were documented in 157,350 encounters during the 2001–2021 study period (Table 2, Fig. 2). Effort was variable over time: it was much higher in some areas relative to others, and skewed to the central and eastern North Pacific. However, data collection occurred

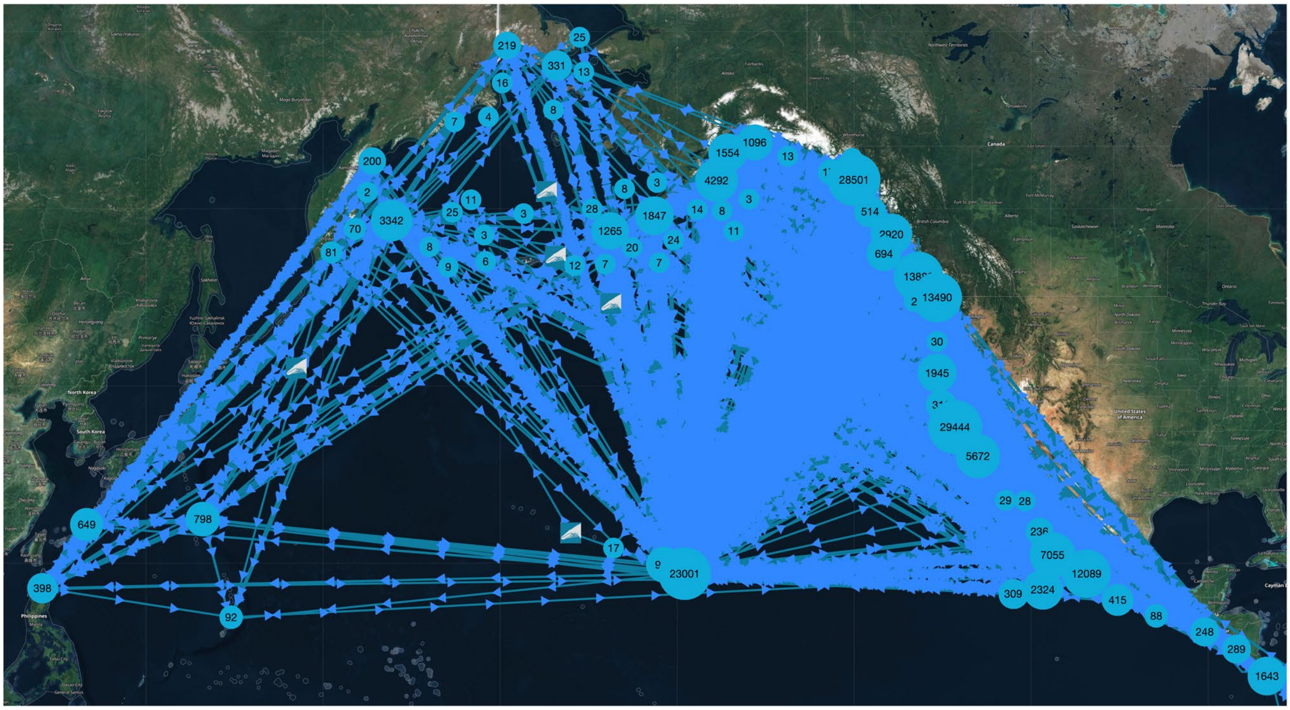

Figure 2. All North Pacific humpback whale encounters and migratory connections as viewable in Happywhale map view for all data collected through August 2022. Numbers in blue circles are counts of individual encounters aggregated by area, while the humpback whale icon represents a single encounter. Blue lines and arrows represent migratory connections of whales sighted in more than one location, not actual travel paths. Map created using Happywhale, built on a basemap reproduced with permission from Maptiler (www.maptiler.com) and OpenStreetMap (www.openstreetmap.org).

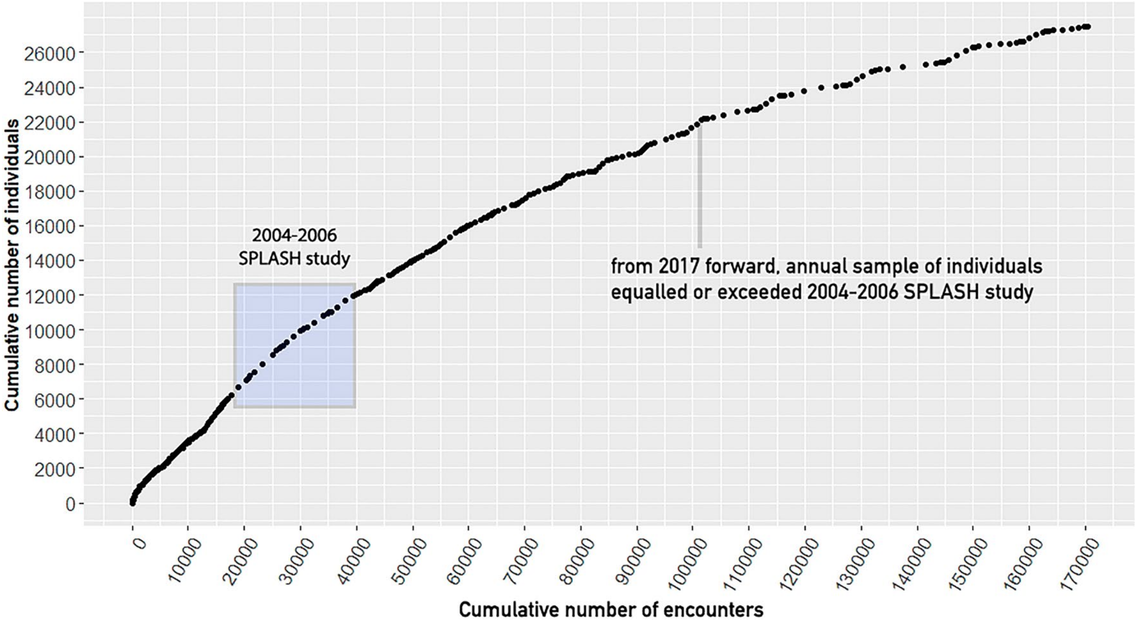

Figure 3. Discovery curve of cumulative number of North Pacific individual humpback whales versus cumulative number of encounters for all data collected through August 2022. Each dot represents one month of effort. The 2004 through 2006 SPLASH study resulted in a large increase in known whales during the study's three years. From 2017 forward, at 101,000 cumulative encounters, the annual number of individuals identified matched or exceeded SPLASH annual sample sizes, yet the cumulative number of individuals increased by an average of only 5% annually, compared to 21% during SPLASH.

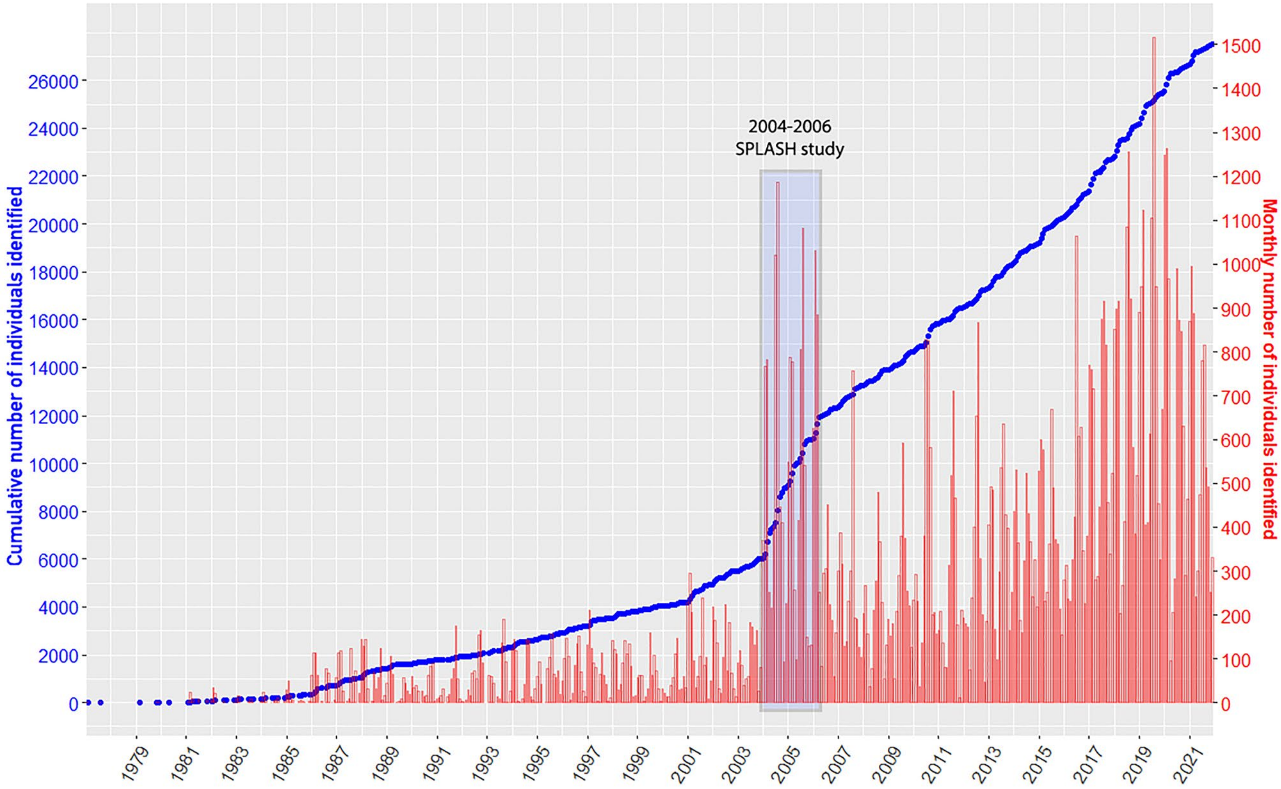

Figure 4. Cumulative individual identifications over time for the number of uniquely identified individual humpback whales documented in the North Pacific for all available photo-ID records collected through August 2022. Dates refer to the time when whales were photographed. Field effort during the 2004–2006 SPLASH study, highlighted in light blue, resulted in a steep increase in the total number of individuals identified.

in all known humpback whale breeding and feeding areas, with high rates of individual resighting throughout (Table 3, Fig. 5). Approximately two-thirds of encounters were represented by a single photo-ID image, while the remaining third contained additional supporting images (e.g., multiple views of the flukes, dorsal fin to fluke series and/or behavioral and anatomical images of the same individual). Naming/numbering protocols for 39 reference catalogs were combined into one unified set, with an average of 1.96 IDs per individual (range: 1–10). Most encounters (66%, documenting 24,049 individuals) were sourced from NPPID collaborators, with the remaining 34% submitted by community scientists (documenting 15,298 individuals); these are shown by region in Fig. 5. The community science-sourced component of the dataset was contributed by 3413 Happywhale users (Supplementary Material II). By volume, most community science-sourced images were contributed by whale watch tour naturalists, who consistently photographed and uploaded photo-ID images of every whale they were able to photograph. Most encounters (66%) were made publicly visible, with the remainder visible only to NPPID collaboration members (Tables 1 and 2 by region and research group). An additional 6318 humpback whale encounters (4% of total North Pacific encounters, primarily from public contributors), remained unidentified to individual due to poor image quality.

An annual average of 87% of individuals (84–92%) were documented in more than one season (Table 3, by region Fig. 5), averaging 5.6 seasons of encounters per individual. During the three-year SPLASH study, the cumulative number of individuals documented increased annually by an average of 21%. By contrast, from 2017 forward, with a comparable or greater number of individuals identified per year, cumulative individuals increased by an average of 5%, due to the documentation of a higher proportion of living individuals (Fig. 3). Data collection temporarily surged during the 2004–2006 SPLASH study, then increased gradually from 2007 and 2014 and more strongly from 2015 (Fig. 4).

Automated image recognition with manual review of each proposed match detected approximately 2,300 duplicate IDs (false negatives) within the 39 collaborator catalogs: these represent cases where the same whale was given multiple IDs within one catalog due to an undetected match (8% of total individuals). The range of false negatives across collaborator catalogs of greater than 100 individuals was 0.1–11%. In the SPLASH dataset of 7971 total individuals, 331 (4%) previously undetected false negatives were found. False positive errors, where two or more whales were confused as one individual, were far less likely than false negatives, prevented by manual review of each proposed match. False positives error rates were estimated to be below 0.1%. Over 5700 encounter comments were received through Happywhale's online comment fields from researchers and community scientists, in many cases alerting data managers to potential errors in date, location and/or whale identities.

NPPID collaborating organization	Identified encounters	Individuals	Resighting ratio	% Publicly visible
Alaska Whale Foundation	3887	1387	2.8	26
Association ELI-S	133	129	1.0	98
BALYENA.ORG	479	231	2.1	57
Cascadia Research Collective	44,310	7117	6.2	99
Commander Islands National Park	579	564	1.0	100
Ecologia y Conservación de Ballenas, A.C. ECOBAC	6892	2998	2.3	48
Eye of the Whale Marine Mammal Research	247	229	1.1	100
Department of Fisheries and Oceans Canada	4680	1669	2.8	37
Glacier Bay National Park & Preserve	10,753	649	16.6	6
Happywhale	5085	1762	2.9	100
Hawai‘i Marine Mammal Consortium	2702	1894	1.4	18
Humpback Whales of the Salish Sea	1990	517	3.8	1
International Whaling Commission	168	156	1.1	100
Juneau Flukes	377	73	5.2	89
Marine Education and Research Society	17,400	541	32.2	2
Oregon State University Marine Mammal Institute Whale Telemetry Group	1789	1281	1.4	100
Murdoch University	3985	2198	1.8	96
Pacific Islands Fisheries Science Center, NOAA Fisheries	142	99	1.4	65
NOAA Fisheries Southwest Science Center	1265	971	1.3	100
NOAA Fisheries Science Center, Alaska	2052	1373	1.5	91
NOAA Hawai‘ian Islands Humpback Whale National Marine Sanctuary	2309	1940	1.2	39
North Coast Cetacean Research Initiative, Ocean Wise	1080	280	3.9	0
North Coast Cetacean Society	5562	453	12.3	9
North Gulf Oceanic Society	1315	639	2.1	100
Okinawa Churashima Foundation	5989	1735	3.5	10
Pacific Whale Foundation	9730	5065	1.9	100
Pacific Wildlife Foundation, Canada	755	494	1.5	100
Russian Cetacean Habitat Project	3692	2057	1.8	61
Simmons University/ Emmanuel College	108	99	1.1	0
The Dolphin Institute	3706	2407	1.5	27
The Keiki Kohola Project	350	311	1.1	88
Universidad Autónoma de Baja California Sur (PRIMMA-UABCS)	2518	2100	1.2	100
University of Alaska Fairbanks	4934	2186	2.3	98
University of Alaska Southeast	5200	1945	2.7	34
University of Hawai‘i at Mānoa, Hawai‘i Institute of Marine Biology	341	305	1.1	100
VE Enterprises	1257	699	1.8	98
Whale Trust	2984	2183	1.4	93
Whales of Guerrero	698	571	1.2	100
Winged Whale Research	477	218	2.2	100

Table 2. Dataset detail by NPPID collaborating research organization for all data contributed. The resighting ratio statistic reports the average number of encounters of each identified individual.

Discussion

The NPPID collaboration established a comprehensive, broad-scale, and rich dataset made possible by a rapid and rewarding feedback process connecting collaborator and community science data around the North Pacific Ocean basin. The NPPID collaboration is the first of its kind to develop a long-term individual ID database on this scale. This effort established a unique dataset foundation well-suited for humpback whale population modeling, as well as for any research benefitting from individual identification, such as longitudinal studies of individual health.

This study began during the development of fast and accurate automated image recognition for humpback whale flukes and demonstrated the scalability for the algorithm used. We could not initially predict how comprehensively we might document the populations of humpback whales across the NPPID study area. However, in a relatively short period the results exceeded expectations. As of August 2022, 56 months after the creation of this study, 30,100 individual North Pacific humpback whales had been documented. Some regions are now extremely well sampled. For example, in Southeast Alaska and northern British Columbia for 2011–2019, fewer than 6% of individuals encountered each year were unique (encountered in only one season) (Table 3, Fig. 5).

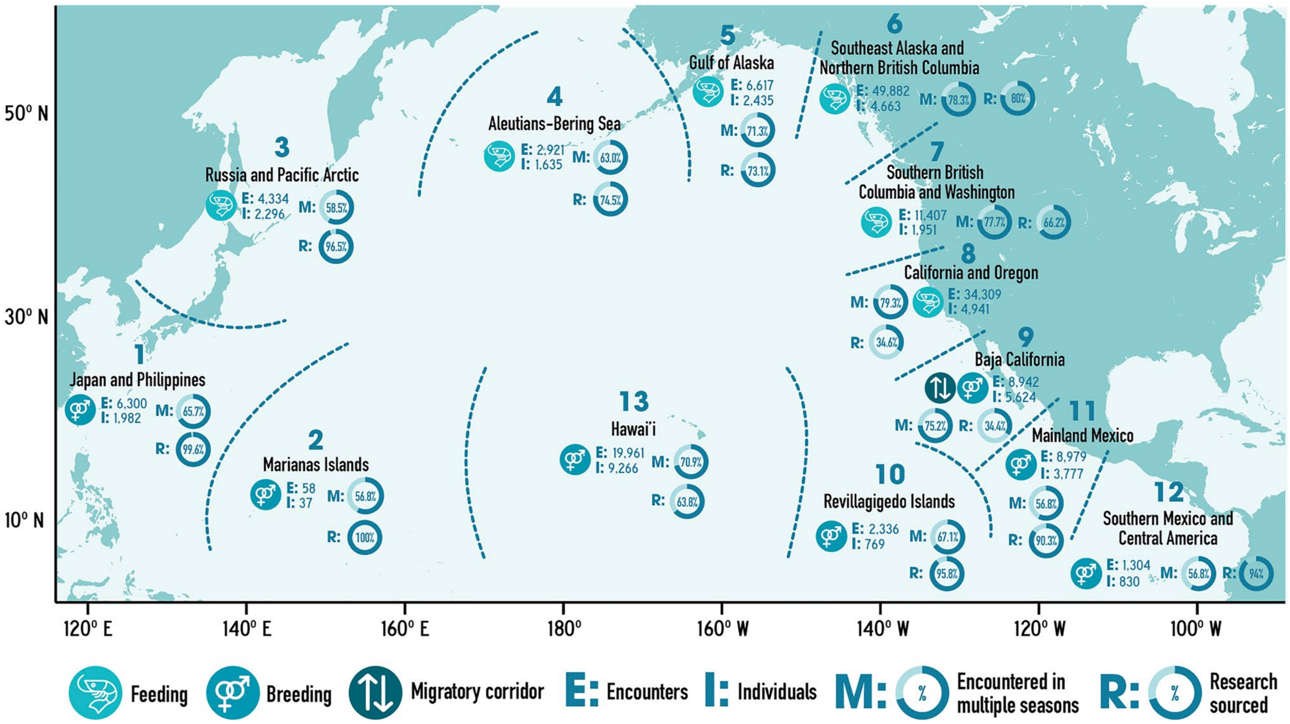

Figure 5. Humpback whale photo-ID data collections by region across the North Pacific Ocean. Region boundaries are indicated by dashed lines, with numbering that corresponds with Table 3. Data for each region includes: a symbol indicating feeding, breeding, or migratory corridor, E: a count of all encounters (trimmed to one encounter per individual per season) documented in each region, I: a total count of individuals documented in each region, M: the percentage of individuals encountered in more than one sampling season, and R: the percentage of data sourced from research collaborators versus community science. Map created with Adobe Illustrator 27.5 on an open source basemap from Freepik.com.

The annual set of newly documented individuals includes recruitment of calves and juveniles, and a progressively smaller proportion of previously undocumented adults.

Data gaps exist, particularly in the western North Pacific, in remote feeding areas such as the Aleutian Islands, and in the Mexican offshore breeding area of the Revillagigedo Islands, where effort was far less than in most breeding, feeding and migratory corridor areas of the central and eastern Pacific. In the Northwestern Hawaiian Islands archipelago, recent acoustic-based surveys including those using wave-glider technology have revealed substantial singing and thus humpback whale abundance with relatively little fluke ID effort[46–48]. It remains to be determined if the majority of these whales use this region as a terminal breeding ground, or whether they mix during a breeding season with those whales in the main Hawaiian Islands. However, even in these least-sampled regions, over 50% of individuals were encountered in more than one season, in the same or in different regions. Thus, we believe that the great majority of individuals in all the North Pacific, including the less sampled regions, are documented in the NPPID dataset. By extensively resampling populations in breeding grounds, migratory corridors, and feeding areas, the impact of effort bias on population models can be reduced[21]. We believe applies to the NPPID dataset.

Accessibility and user agreements.

Data collection should not be an end unto itself, and sharing is a core tenet of good data management[49]. The Happywhale web platform was developed to make data accessible by design, aiming for a user experience that is both easy and rewarding. Users were motivated to contribute more and higher-quality data by a simple user interface to upload images, which then rewarded them with rapid results of information about "their" individual whales. Accessibility creates a public good as a resource for research, education, resource management, and science communication. In the existing NPPID dataset, 66% of all North Pacific humpback whale encounter data are publicly visible. Researchers and community scientists can explore migratory connections across the North Pacific via the web platform (Fig. 2). For research collaborators, this has inspired studies that would not have been possible without the large collective investment in building a platform and populating it with a comprehensive and contemporary dataset[50,51]. As of December 2022, the NPPID had contributed data to seven other collaborative peer-reviewed publications[13,37,38,52–55]. Accessible information about North Pacific humpback whale individuals has also proven very useful for resource managers, for example in tracking fishing gear entanglement cases, and individual identification and past sighting histories of dead or stranded whales[56].

We recognize that including many actors and an open-science stance can add complexity to a collaboration[57] with concerns for misuse of shared or public data[58]. Successful aspects of this collaboration bring opportunities

Map region#: region		2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
1: Japan and Philippines* (Excluding Hokkaido migratory corridor)	Encounters	108	130	113	361	482	665	300	254	359	314	448	387	345	403	454	326	290	293	258	2	8	6300
	Individuals	108	130	113	277	310	397	300	252	359	309	443	383	340	395	445	322	288	290	252	2	7	1982
	Research sourced (%)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	98.6	95.3	50.0	0	99.6
	Publicly visible (%)	3.7	10.8	22.1	77.0	82.0	88.3	3.3	5.1	1.1	4.1	1.6	2.6	3.5	3.5	6.8	3.7	1.0	1.4	14.0	50.0	100	23.5
	Encountered multiple seasons (%)	98.1	93.8	91.2	81.6	81.0	82.6	88.7	90.9	92.8	93.9	88.3	92.7	94.4	90.1	89.9	89.9	86.1	79.3	75.8	100	71.4	65.7
2: Marianas Islands	Encounters	0	0	0	0	0	0	4	0	0	0	0	0	0	0	5	2	22	11	2	12	0	58
	Individuals	0	0	0	0	0	0	4	0	0	0	0	0	0	0	3	2	17	3	2	8	0	37
	Research Sourced (%)	–	–	–	–	–	–	100	–	–	–	–	–	–	–	100	100	100	100	100	100	–	100
	Publicly Visible (%)	–	–	–	–	–	–	100	–	–	–	–	–	–	–	100	100	100	100	100	100	–	100
	Encountered multiple seasons (%)	–	–	–	–	–	–	75.0	–	–	–	–	–	–	–	33.3	100	70.6	33.3	50	37.5	–	56.8
3: Russia and Pacific Arctic (Includes Commander Islands)	Encounters	0	2	1	57	94	24	9	68	261	934	412	433	513	202	337	260	287	244	72	18**	106**	4334
	Individuals	0	2	1	55	81	9	9	54	240	683	336	346	403	165	313	248	216	175	70	16**	106**	2296
	Research sourced (%)	–	100	100	96.5	100	16.7	55.6	63.2	83.9	100	100	100	100	100	100	100	98.3	83.2	86.1	100**	98.1**	96.5
	Publicly visible (%)	–	100	100	98.2	100	95.8	100	91.2	96.2	74.8	63.1	47.1	47.4	45.5	51.9	79.2	74.6	51.6	65.3	100**	100**	66.6
	Encountered multiple seasons (%)	–	100	100	85.5	80.2	88.9	77.8	77.8	72.1	68.5	77.4	80.1	76.9	83.0	76.4	70.2	66.7	71.4	65.7	31.2**	42.5**	58.5
4: Aleutians – Bering Sea (All Aleutians to Alaska Peninsula, boundary at 157°W)	Encounters	117	43	10	654	681	0	192	253	84	343	87	279	49	1	7	7	35	10	64	0	5	2921
	Individuals	85	41	10	524	447	0	153	194	71	277	82	182	47	1	7	7	33	10	64	0	5	1635
	Research sourced (%)	100	100	0.0	43.3	57.7	–	100	100	100	99.4	100	100	100	100	85.7	0.0	94.3	70.0	12.5	–	0.0	74.5
	Publicly visible (%)	100	100	100	100	100	–	100	100	100	100	100.	100	100	100	100	100	100	100	100	–	100	100
	Encountered multiple seasons (%)	80.0	87.8	90.0	63.4	71.4	–	84.3	80.4	87.3	73.3	68.3	83.0	55.3	100	100	85.7	51.5	40.0	75.0	–	100	63.0
5: Gulf of Alaska (Between 157° and 141°W)	Encounters	174	237	156	1130	805	179	644	243	125	247	446	499	254	316	278	131	238	84	214	90	127	6617
	Individuals	155	203	146	770	497	146	413	191	97	215	299	342	195	207	222	111	144	52	104	60	83	2435
	Research sourced (%)	53.4	67.9	60.9	67.3	61.6	81.6	90.2	72.8	89.6	89.9	97.5	92.8	96.5	94.6	64.4	53.4	59.2	27.4	45.3	21.1	15.7	73.1
	Publicly visible (%)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	85.6	99.2	99.8
	Encountered multiple seasons (%)	85.8	82.3	84.2	79.9	84.1	84.9	86.7	84.3	89.7	85.6	89.6	84.5	75.9	82.6	85.6	92.8	93.7	92.3	94.2	93.3	92.8	71.3
6: Southeast Alaska and Northern British Columbia	Encounters	79	54	246	2478	1808	1670	1623	1363	2031	1761	1938	2147	2495	3087	3159	3325	3129	3456	6226	3855	3952	49882
	Individuals	75	51	218	1355	879	741	723	543	799	690	729	536	581	777	902	1013	786	838	1595	1219	1222	4663
	Research Sourced (%)	87.3	85.2	89.8	74.1	82.6	95.6	96.9	97.4	89.0	97.8	88.4	92.4	96.3	86.9	93.1	82.9	89.2	79.4	61.5	69.2	43.6	80.0
	Publicly visible (%)	41.8	40.7	32.5	79.5	84.5	6.9	5.4	6.4	13.3	10.9	19.8	8.8	8.7	29.5	18.6	18.4	16.6	21.8	42.6	31.9	62.6	29.9
(Alaska E of 141°W and British Columbia N of Vancouver Is.)	Encountered multiple seasons (%)	89.3	90.2	92.7	87.9	90.6	92.8	91.4	93.9	92.2	94.5	96.6	97.0	95.5	94.6	95.7	96.5	96.3	95.5	95.5	94.2	92.1	78.3
7: Southern British Columbia and Washington (North tip of Vancouver Island to Washington/ Oregon Border)	Encounters	50	54	22	137	343	150	385	195	328	144	139	275	204	118	380	849	1077	2002	2265	1542	748	11407
	Individuals	42	44	21	116	191	101	230	157	195	117	113	223	155	66	125	270	369	565	633	427	388	1951
	Research sourced (%)	86.0	85.2	86.4	54.7	64.1	97.3	94.8	90.3	75.9	85.4	68.3	74.9	77.9	66.9	75.0	76.2	68.8	70.8	56.8	65.9	21.1	66.2
	Publicly visible (%)	100	100	100	99.3	98.0	42.0	37.1	73.8	47.3	44.4	89.2	87.6	82.4	64.4	74.2	67.7	55.6	55.1	65.9	50.6	79.8	63.2
	Encountered multiple seasons (%)	97.6	100	90.5	94.0	92.7	91.1	93.0	93.6	90.3	91.5	92.9	90.1	89.0	92.4	93.6	93.7	93.8	89.4	89.6	90.2	79.6	77.7

(continued)

Map region#: region		2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
8: California and Oregon (Oregon Border to Mexican Border)	Encounters	301	372	421	407	606	302	197	333	679	818	634	873	1175	1730	2397	2469	4548	5838	3836	2954	3419	34309
	Individuals	240	299	355	302	369	208	177	296	462	507	381	509	565	584	654	1266	1304	1575	1244	1117	1139	4941
	Research sourced (%)	91.7	89.5	84.8	76.2	67.5	87.7	83.2	82.0	36.1	36.7	26.5	25.9	75.6	73.%	43.8	35.3	31.2	38.7	9.9	11.4	2.3	34.6
	Publicly visible (%)	100	100	100	100	100	100	100	100	100	100	99.7	99.8	100	100	99.3	100	99.9	100	99.0	99.4	100	99.8
	Encountered multiple seasons (%)	97.1	97.0	97.7	96.7	91.6	96.6	93.2	95.6	97.8	97.0	95.3	95.3	97.9	96.4	88.2	83.7	94.1	91.9	93.6	91.0	86.2	79.3
9: Baja California	Encounters	0	4	22	257	304	220	150	45	102	64	95	63	418	569	598	350	1100	1242	1109	1433	797	8942
	Individuals	0	4	21	248	264	188	142	44	96	63	93	63	390	538	563	321	946	1135	1022	1287	697	5624
	Research sourced (%)	–	100	100	75.1	65.1	71.4	58.7	93.3	83.3	89.1	93.7	60.3	63.2	63.8	63.0	36.3	23.7	24.6	9.3	16.2	9.2	34.4
	Publicly visible (%)	–	0.0	95.5	96.5	97.7	94.1	85.3	40.0	41.2	57.8	44.2	68.3	93.1	97.5	97.2	99.4	100	99.3	99.8	100	100	96.5
	Encountered multiple seasons (%)	–	100	95.2	73.0	86.7	88.3	84.5	97.7	96.9	88.9	92.5	90.5	82.6	77.9	76.9	77.6	87.6	88.6	81.7	78.3	76.6	75.2
10: Revillagigedo Islands	Encounters	0	0	0	1060	549	586	0	0	0	1	1	0	2	1	1	0	15	0	20	48	52	2336
	Individuals	0	0	0	346	246	280	0	0	0	1	1	0	2	1	1	0	15	0	20	46	45	769
	Research sourced (%)	–	–	–	96.8	95.1	95.1	–	–	–	0.0	0.0	–	0.0	0.0	0.0	–	100	–	100	97.9	100	95.8
	Publicly visible (%)	–	–	–	100	100	100	–	–	–	100	100	–	100	100	100	–	100	–	100	100	100	100
	Encountered multiple seasons (%)	–	–	–	71.7	77.6	73.6	–	–	–	100	0.0	–	100	100	100	–	80.0	–	90.0	80.4	75.6	67.1
11: Mainland Mexico (All mainland Mexico to south of Colima)	Encounters	243	195	162	433	484	528	251	337	249	333	260	448	790	666	353	269	609	669	562	710	428	8979
	Individuals	223	182	139	289	317	412	230	308	236	306	244	373	599	508	333	217	493	537	470	549	378	3777
	Research sourced (%)	100	99.5	98.1	94.7	94.4	99.2	100	100	98.8	100	99.6	98.7	99.2	100	96.9	76.6	65.8	70.3	81.7	82.4	79.7	90.3
	Publicly visible (%)	43.6	46.2	46.3	74.6	78.5	68.9	38.2	44.2	41.8	33.9	43.5	61.2	77.7	77.8	62.9	89.2	93.8	82.1	100	100	100	73.5
	Encountered multiple seasons (%)	89.2	86.8	94.2	94.1	92.1	93.2	95.2	92.2	94.1	97.1	96.3	92.2	90.7	93.7	87.7	91.2	90.7	94.2	94.0	93.6	93.4	85.8
12: Southern Mexico and Central America (All mainland Mexico from Guerrero through Panama)***	Encounters	21	10	8	23	79	76	9	60	30	55	27	13	8	30	74	65	225	204	36	70	181	1304
	Individuals	16	7	7	20	66	61	7	56	29	39	20	11	8	29	68	63	204	183	36	65	167	830
	Research sourced (%)	100	100	100	95.7	96.2	98.7	100	100	100	100	100	100	100	100	98.6	98.5	100	96.6	66.7	80.0	79.0	94.0
	Publicly visible (%)	100	100	100	100	93.7	94.7	77.8	41.7	23.3	40.0	51.9	84.6	12.5	90.0	97.3	95.4	91.6	88.2	38.9	78.6	91.2	82.5
	Encountered multiple seasons (%)	93.7	100	100	100	92.4	93.4	100	94.6	89.7	100	100	100	87.5	93.1	89.7	93.7	88.2	86.9	94.4	93.8	86.2	87.2
13: Hawai'i	Encounters	642	375	192	1149	1310	1688	1369	838	201	467	227	239	653	535	844	427	877	1195	2056	2225	2452	19961
	Individuals	597	350	190	975	1076	1287	1168	751	180	435	215	219	598	494	765	394	776	981	1597	1638	1802	9266
	Research sourced (%)	100	98.9	27.6	70.2	73.3	72.4	61.9	84.7	75.6	66.6	84.6	52.7	75.5	72.9	72.5	68.4	51.9	53.6	62.1	48.0	45.3	63.8
	Publicly visible (%)	0.6	1.3	75.5	92.6	89.2	88.7	38.2	15.6	27.4	43.7	39.6	54.0	64.3	69.0	81.2	63.2	72.9	87.8	93.3	81.7	82.2	71.1
	Encountered multiple seasons (%)	82.7	81.1	83.2	84.2	85.9	86.2	85.2	81.8	81.7	81.8	80.0	85.4	80.6	82.0	77.3	83.0	83.9	85.7	83.2	84.9	83.0	70.9
Total	Encounters	1735	1476	1353	8146	7545	6088	5133	3989	4449	5481	4714	5656	6906	7658	8887	8480	12452	15248	16720	12959	12275	157350
	Individuals	1471	1293	1136	4824	4397	3612	3390	2715	2648	3460	2911	3035	3622	3546	4226	3912	4936	5564	6384	5726	5542	27956
	Research Sourced (%)	92.9	90.8	77.5	75.4	76.9	88.0	85.3	91.6	80.6	86.0	83.3	81.3	89.1	83.5	74.9	66.3	56.8	56.2	46.6	46.7	31.0	66.3
	Publicly visible (%)	46.7	57.5	71.2	90.2	91.6	65.7	39.7	36.6	39.9	50.2	46.6	48.6	51.3	60.2	59.7	58.2	69.9	71.7	71.4	70.4	83.0	65.5
	Encountered multiple seasons (%)	87.4	88.2	91.1	80.9	84.5	87.1	87.8	88.1	90.3	85.6	89.8	89.8	86.9	88.4	85.1	86.3	88.1	87.9	87.2	85.9	83.2	62.6

Table 3. 2001–2021 humpback whale dataset with sample size and characteristics presented by region andyear. Encounters: a count of the total number of photo-ID documented encounters of individuals, defined as one encounter per individual whale per day. Individuals: count of unique identified individuals. Research sourced: percent subset of encounters sourced from NPPID collaborators, while all remaining encounters were sourced from community science efforts. Public: percent subset of encounters visible to all users of Happywhale, while all remaining encounters are visible only to NPPID collaborators. Multiple seasons: a count of identified individuals documented in more than one season during the study, overall total for all years in final column. Seasons: north of 32.55°N = feeding regions (blue) data by calendar year; south of 32.55°N = migratory corridor (tan) and breeding regions (orange) data by breeding season, defined as 1 August through 31 July (i.e., a December 2015 encounter in Hawai'i is considered to be in the 2016 breeding season). *Japan and Philippines data limited to one encounter per individual per year. **Russia 2020 and 2021 data limited to new individuals. ***Central America data from Nicaragua to Panama excludes Southern Hemisphere migrants encountered November through April of each year.

but also pose two challenges that the collaboration must address: (1) How do we encourage contributing researchers to allow public visibility of data to allow the widest possible benefit, while ensuring data are used correctly in context, with proper credit preserved? (2) How do we simplify and clarify co-authorship policies to be effective, meaningful, and not so complex as to hinder publication?

An ideal collaboration builds datasets that directly answer present biological and management questions, and simultaneously creates data-sharing readiness. Data readiness for study of ecological change depends on both standardized repositories and aligned research interests[13,59,60]. The NPPID dataset has been successfully applied in this context, contributing to challenging management issues such as the US West Coast Dungeness crab fishery. Here, researchers can readily determine the proportion of whales in the Endangered Central American DPS[51,61–63]. The NPPID collaboration began with a MOA, offering co-authorship to contributors in a series of publications investigating humpback whale migratory patterns and population status in the North Pacific. Collaborators wishing to address additional research questions must seek permission from all relevant data contributors. While the communication required is a cost imposed on prospective studies, community is built around mutually beneficial collaboration. The MOA created an effective working group and context for this study through the completion of the specified series of publications. Future success will require clear use, sharing, and management policy, with oversight and funding maintained into the future.

Data quality improved by accessibility. Accessibility adds value as part of the FAIR Principles for scientific data[45] that guided this study design. Accessibility also serves the immediate practical value of improving data quality, consistency, and repeatability. Active collaboration and public access to data make knowledge gaps more visible and encourages effort to fill them[64]. With many eyes reviewing the dataset, from curious public enthusiasts exploring encounters of "their" whales or an area of their personal interest to research collaborators pursuing diverse lines of inquiry, an ongoing collaborative quality control process frequently detects data discrepancies. Happywhale user comments—over 5,700 as of August 2022—alerted NPPID data managers to enough errors that public accessibility to review might be considered as a systematic method of quality control, worthy of attention for its own value and efficiency.

All datasets will contain errors; more accurate image recognition, repeatedly applied, and review of data by diverse users will continually detect some, but not all errors. The SPLASH study estimated a 9–10% rate of missed matches using trained human matchers, the largest model error correction factor in the associated mark-recapture population estimate[21]. This kind of accuracy assessment rarely appears in photo-ID based mark-recapture studies, yet missed matches were detected in every dataset larger than 100 individuals involved in this study. Our finding of 331 false negatives in 7971 (4%) total individuals in the SPLASH study, when added to algorithm error rates for good-to-high quality images of 1–3%[42], suggests the 9–10% error estimation was high by 3–4%. In our most accurately matched large dataset, the 2004–2020 whales of Glacier Bay National Park and Preserve, Alaska, missed matches accounted for only 0.15% (1 of 633 individuals, a first-summer calf to adult match with substantial fluke pigment change). All other datasets of more than 100 individuals showed from 2 to 11% detectable false negative missed match rates. Considering this range and other sources of error and bias, it is important to understand and account for limitations in any dataset, including ours.

Effort bias and appropriate use. Ideally, a dataset should be created with its specific use in mind a priori, following a good data management plan[49] with an optimized data workflow[65]. However, because we built a dataset gathered from post-SPLASH photo-ID archives and opportunistic efforts, standardization had to stand in for a priori data management plans. The effort was geographically and temporally heterogeneous, and any study design or interpretation of data must account for this to ensure appropriate use. It would be easy, for example, to falsely interpret the lower effort in the western North Pacific as evidence of smaller whale populations. Datasets cannot be assumed to provide an error-free documentation of humpback whale presence in the study area (i.e., devoid of effort bias); no clear rule can be set a priori to identify the appropriate application of an evolving dataset of this nature. It is therefore imperative that any potential data user actively engage directly with collaborating researchers to understand data limitations and potential. Data contributors can also be the primary data users, a group that will benefit from increased knowledge of and aptitude with the data management system built through Happywhale.

Because there could not be a comprehensive data collection plan across this large scale of a study area and time period, the full dataset might be considered opportunistic, a sum of collected efforts of dedicated research, research from platforms of opportunity, and community science contributions. Figure 4 demonstrates a large increase in data collection over time, elevated during the 2004–2006 SPLASH study, then building to similar levels from 2017 forward. Data collection rates have benefitted from many factors. These include: improvement in digital cameras, the growing popularity of whale watching, the 2015 establishment of the Happywhale platform, increased effort by many NPPID collaborators to capture fluke photos within existing field efforts, and the 2020 establishment and NOAA Fisheries funding of the SPLASH-2 program. The latter helped fund data collection efforts in poorly sampled areas, and infrastructure to support submissions to Happywhale. Our peak sample year was 2019, with 6,384 (21%) of 30,100 known North Pacific humpback whales documented. The COVID-19 pandemic interrupted both field research efforts and tourism in 2020 and 2021 (Fig. 4), though we believe sampling will recover and continue to increase.

Building a successful collaboration. The NPPID study benefitted from the largely successful precedent of the SPLASH study both in providing a foundation of data (Fig. 4) and as a collaborative framework. The current study began at a time when new methods were needed to efficiently manage large volumes of post-SPLASH data, where separate research efforts were constrained by time-intensive visual matching of photo-ID datasets.

Although the SPLASH study produced notable insight and remains frequently cited, and the catalog was made available online, the study was not intended to continue beyond 2006, and the online dataset was not built to facilitate photo-ID matching. The role of the NPPID collaboration agreement was to establish clear expectations and create an environment of openness, trust, transparency, and consistency. This context was necessary for research collaborators to feel comfortable sharing images and data that were products of many thousands of person-hours and costs in the field. Positive and useful feedback delivered by rapid results from image recognition efforts was also necessary. Researchers were enticed to join the collection in part by the instant gratification when most of their flukes immediately matched to known individuals; this was a welcome change from years of toil over visually matching isolated photo-ID datasets. Success was crafted by a combination of a high-quality product supported by solid guiding principles of Transparency, Responsibility, User focus, Sustainability and Technology (TRUST), to promote digital repository trustworthiness[66]. The idea behind these principles is that as a data repository, we must earn the trust of the community we serve and demonstrate that we are reliable and capable of appropriately managing the data we curate. Empowerment comes through this intentional framework, with a feeling of collective ownership rather than isolated possession. This then supports sustainable collaboration by creating active participation of research users.

As an ongoing, living dataset, the NPPID developed active, increasingly decentralized participation in ongoing data management with an intent to serve diverse needs in the research community. System development remains ongoing, with a focus on providing research collaborators with tools to become more directly involved with data management. This development reduces centralized data management costs, serves the real-time needs of collaborators, and benefits the dataset with local expertise, potentially detecting data issues that would not be recognized by remote data managers.

Conclusion: sustainability and maximizing future value

The NPPID effort has established a single unified repository. This has been accomplished by reconciling all available research catalogs and ID nomenclature, and by aggregating all individual identities and encounter data into a state of data readiness unprecedented on a long-term and ocean-basin-scale. The first benefits are cost savings and organizational effectiveness. Particularly in well-sampled areas, data processing is revolutionized by immediate access to a fast and reliable photo-ID system. Collaborators reported that this “saves countless hours of manual visual matching, allowing us to get our data out in products, papers, and outreach more quickly” (JN) and “reduces lab time by 90%” (AS). However, collaborators face the challenge of how to maximize the present and future value of the NPPID dataset. A primary outstanding need is to create clarity for how researchers efficiently access, establish permission, and create sub-collaborations to develop further studies beyond the term of the NPPID collaboration.

System functionality was developed in a constant feedback cycle to accommodate progressively larger datasets through the study. This dataset appears to document most living humpback whales across the North Pacific Ocean basin, creating an abundance of data and inspiring an ambition to monitor populations in near-real time. With heterogeneous sampling effort over the study area, critical data gaps can be identified for understanding abundance and population structure. In addition, minimum sample sizes for reliable, robust population models can be established. Given the low cost of data storage, and if the incremental cost of each additional data point is driven to near zero, there is very little cost to overshooting a threshold of “enough” data.

Having now acquired sufficient baseline data for North Pacific populations in the face of a changing ocean, we aim for data readiness to understand the implications of ecosystem events on a timescale that benefits resource management. This study concentrates on humpback whales of the North Pacific, but the concept and methods can be extended to many species. Baleen whales are recognized to influence marine ecosystems on a massive scale[67]. In recognition of the concept of essential biological variables[68,69], there is a need for marine observation and data at an ocean-basin-wide scale[70–72]. This dataset, the collaboration agreement, and the system established to create and maintain it can contribute to our understanding of essential ocean variables.

This study established an extremely cost effective and utilitarian information architecture, delivering an essential service for ongoing studies. If investment in collaborator engagement, upkeep, development, and data management continue, the future of this collaborative system promises great contributions to the understanding of North Pacific humpback whale populations. Sustainability will require a transition from the centralized efforts of a multi-year study to an established project at a stable institution with community ownership, oversight, and funding. We see this effort not as collecting and possessing a dataset, but as curating a public good for the betterment of science, education, and marine conservation. The FAIR and TRUST principles are central to guiding development, recognizing that accessibility requires more than just a data search feature via a web browser. To truly achieve full potential will require decentralizing data management to research collaborators, a shift that requires further system development, funding, user training, and commitment. Involving scientists in data management has evolved through time from a widespread disconnect[73] to a current trend of ecological “big data” where data management is a necessary skill for ecologists, as has already happened with statistics and GIS[74]. We believe that establishing this scale-shifting dataset, given continued investment, will continue to improve understanding, awareness, stewardship, and respect for the North Pacific marine ecosystem.

Data availability

The publicly viewable 66% of the full dataset used in this study, with ongoing additions and updates is available for exploration at www.Happywhale.com. All data are available with collaborator agreement to explore at Happywhale and in spreadsheet format. Please contact the corresponding author for discussion and permission. Approximately one-third of the dataset is public domain, but the collaborators believe that providing this partial dataset for open access download would be a disservice to the integrity of the full dataset.

Received: 21 November 2022; Accepted: 12 June 2023
Published online: 23 June 2023

References

1. Karczmarski, L., Chan, S. C. Y., Rubenstein, D. I., Chui, S. Y. S. & Cameron, E. Z. Individual identification and photographic techniques in mammalian ecological and behavioural research—Part 1: Methods and concepts. *Mamm. Biol.* **102**, 545–549 (2022).
2. Clapham, P. J. & Mead, J. G. *Megaptera novaeangliae*. *Mamm. Species* **3**, 1–9 (1999).
3. Mackintosh, N. A. The natural history of whalebone whales. *Biol. Rev.* **21**, 60–74 (1946).
4. Calambokidis, J. *et al.* Movements and population structure of humpback whales in the North Pacific. *Mar. Mammal Sci.* **17**, 769–794 (2001).
5. Chittleborough, R. G. Dynamics of two populations of the humpback whale, *Megaptera novaeangliae* (Borowski). *Mar. Freshw. Res.* **16**, 33–128 (1965).
6. Dawbin, W. H. The Seasonal Migratory Cycle of Humpback Whales. in *Whales, Dolphins, and Porpoises* 145–170 https://doi.org/10.1525/9780520321373-011/HTML (2021).
7. Baker, C. *et al.* Migratory movement and population structure of humpback whales (*Megaptera novaeanglieae*) in the central and eastern North Pacific. *Mar. Ecol. Prog. Ser.* **31**, 105–119 (1986).
8. Calambokidis, J. *et al.* in *SPLASH: Structure of Populations, Levels of Abundance and Status of Humpback Whales in the North Pacific*. Final Report Contract AB133F-03-RP-00078. (United States Department of Commerce West. Adm. Cent., 2008).
9. Johnson, C. M. *et al.* Protecting Blue Corridors - Challenges and solutions for migratory whales navigating national and international seas. https://doi.org/10.5281/ZENODO.6196131 (2022).
10. Di Lorenzo, E. & Mantua, N. Multi-year persistence of the 2014/15 North Pacific marine heatwave. *Nat. Clim. Chang.* **6**, 1042–1047 (2016).
11. Hu, Z. Z., Kumar, A., Jha, B., Zhu, J. & Huang, B. Persistence and predictions of the remarkable warm anomaly in the Northeastern Pacific ocean during 2014–16. *J. Clim.* **30**, 689–702 (2017).
12. Arimitsu, M. L. *et al.* Heatwave-induced synchrony within forage fish portfolio disrupts energy flow to top pelagic predators. *Glob. Chang. Biol.* **27**, 1859–1878 (2021).
13. Gabriele, C. M. *et al.* Sharp decline in humpback whale (*Megaptera novaeangliae*) survival and reproductive success in southeastern Alaska during and after the 2014–2016 Northeast Pacific marine heatwave. *Mamm. Biol.* https://doi.org/10.1007/s42991-021-00187-2 (2022).
14. Cartwright, R. *et al.* Fluctuating reproductive rates in Hawaii's humpback whales, *Megaptera novaeangliae*, reflect recent climate anomalies in the North Pacific. *R. Soc. Open Sci.* **6**, 181463 (2019).
15. Frankel, A. S., Gabriele, C. M., Yin, S. & Rickards, S. H. Humpback whale abundance in Hawai'i: Temporal trends and response to climatic drivers. *Mar. Mammal Sci.* **38**(1), 118–138. https://doi.org/10.1111/mms.12856 (2021).
16. Kügler, A., Lammers, M., Zang, E., Kaplan, M. & Mooney, T. Fluctuations in Hawaii's humpback whale *Megaptera novaeangliae* population inferred from male song chorusing off Maui. *Endanger. Species Res.* **43**, 421–434 (2020).
17. von Biela, V. R. *et al.* Premature mortality observations among Alaska's pacific salmon during record heat and drought in 2019. *Fisheries* **47**, 157–168 (2022).
18. Rocha, R. C., Clapham, P. J. & Ivashchenko, Y. V. Emptying the oceans: A summary of industrial whaling catches in the twentith century. *Mar. Fish. Rev.* **76**, 37–48 (2014).
19. Ivashchenko, Y. V. & Clapham, P. J. Too much is never enough: The cautionary tale of soviet illegal whaling. *Mar. Fish. Rev.* **76**, 1–21 (2014).
20. Johnson, J. H. & Wolman, A. A. The Humpback Whale, *Megaptera novaeangliae*. *Mar. Fish. Rev.* **46**, 30–37 (1984).
21. Barlow, J. *et al.* Humpback whale abundance in the North Pacific estimated by photographic capture-recapture with bias correction from simulation studies. *Mar. Mammal Sci.* **27**, 793–818 (2011).
22. Katona, S. K. & Whitehead, H. P. Identifying Humpback Whales using their natural markings. *Polar Rec. (Gr. Brit.)* **20**, 439–444 (1981).
23. Jurasz, C. & Jurasz, V. in *Results of 1977 Studies on Humpback Whales in Glacier Bay National Monument*. Final Rept. (1978).
24. Bettridge, S. *et al.* in *Status Review of the Humpback Whale (Megaptera novaeangliae) Under the Endangered Species act*. (NOAA-TM-NMFS-SWFSC-540, 2015).
25. Herman, L. M. *et al.* Resightings of humpback whales in Hawaiian waters over spans of 10–32 years: Site fidelity, sex ratios, calving rates, female demographics, and the dynamics of social and behavioral roles of individuals. *Mar. Mammal Sci.* **27**, 736–768 (2011).
26. Smith, T. D. *et al.* An ocean-basin-wide mark-recapture study of the north Atlantic humpback whale (*Megaptera novaeangliae*). *Mar. Mammal Sci.* **15**, 1–32 (1999).
27. Federal Register. Endangered and threatened species. in *Identification of 14 Distinct Population Segments of the Humpback Whale (Megaptera novaeangliae) and Revision of Species-Wide Listing* 62260–62319. (2016). Available at: https://www.federalregister.gov/documents/2016/09/08/2016-21276/endangered-and-threatened-species-identification-of-14-distinct-population-segments-of-the-humpback. (Accessed: 21st Mar 2023)
28. Mobley, J. R. Jr., Deakos, M. H., Pack, A. A., Bortolotto, G. A. & Joseph Mobley, C. R. Aerial survey perspectives on humpback whale resiliency in Maui Nui, Hawai'i, in the face of an unprecedented North Pacific marine warming event. *Mar. Mammal Sci.* https://doi.org/10.1111/MMS.13018 (2023).
29. Cates, K. A. *et al.* Corticosterone in central North Pacific male humpback whales (*Megaptera novaeangliae*): Pairing sighting histories with endocrine markers to assess stress. *Gen. Comp. Endocrinol.* **296**, 113540 (2020).
30. Cates, K. A. *et al.* Testosterone trends within and across seasons in male humpback whales (*Megaptera novaeangliae*) from Hawaii and Alaska. *Gen. Comp. Endocrinol.* **279**, 164–173 (2019).
31. Pack, A. A. *et al.* Comparing depth and seabed terrain preferences of individually identified female humpback whales (*Megaptera novaeangliae*), with and without calf, off Maui. *Hawaii. Mar. Mammal Sci.* **34**, 1097–1110 (2018).
32. Pack, A. A. *et al.* Habitat preferences by individual humpback whale mothers in the Hawaiian breeding grounds vary with the age and size of their calves. *Anim. Behav.* **133**, 131–144 (2017).
33. Pack, A. A. *et al.* Size-assortative pairing and discrimination of potential mates by humpback whales in the Hawaiian breeding grounds. *Anim. Behav.* **84**, 983–993 (2012).
34. Herman, L. M. *et al.* Humpback whale song: Who sings?. *Behav. Ecol. Sociobiol.* **67**, 1653–1663 (2013).
35. Hill, M. *et al.* Found: a missing breeding ground for endangered western North Pacific humpback whales in the Mariana Archipelago. *Endanger. Species Res.* **41**, 91–103 (2020).
36. Espinoza Rodríguez, I. J., Frisch Jordán, A. & Noriega Betancourt, F. Humpback whales in Banderas Bay, Mexico: Relative abundance and temporal patterns between 2004 and 2017. *Lat. Am. J. Aquat. Mamm.* **16**, 33–39 (2021).
37. Martien, K. K. *et al.* NOAA technical memorandum NMFS-SWFSC-658 evaluation of Mexico distinct population segment of humpback whales as units under the marine mammal protection act. https://doi.org/10.25923/nvw1-mz45 (2021).
38. Taylor, B. L. *et al.* Evaluation of humpback whales wintering in Central America and southern Mexico as a demographically independent population. *US Dep. Commer. Natl. Ocean. Atmos. Adm. Natl. Mar. Fish. Serv. Southwest Fish. Sci. Cent.* **3**, 103–111 (2021).

39. Wade, P. R. *et al.* Estimates of abundance and migratory destination for North Pacific humpback whales in both summer feeding areas and winter mating and calving areas. in *Paper SC/66b/IA21 submitted to the Scientific Committee of the International Whaling Commission, June 2016, Bled, Slovenia.* Available at https://archive.iwc.int/. (2016).
40. Hendrix, A. N., Straley, J., Gabriele, C. M. & Gende, S. M. Bayesian estimation of humpback whale (*Megaptera novaeangliae*) population abundance and movement patterns in southeastern Alaska. *Can. J. Fish. Aquat. Sci.* **69**, 1783–1797 (2012).
41. Gabriele, C. M. *et al.* Natural history, population dynamics, and habitat use of humpback whales over 30 years on an Alaska feeding ground. *Ecosphere* **8**, e01641 (2017).
42. Cheeseman, T. *et al.* Advanced image recognition: A fully automated, high-accuracy photo-identification matching system for humpback whales. *Mamm. Biol.* **2021**, 1–15. https://doi.org/10.1007/S42991-021-00180-9 (2021).
43. Stevick, P. *et al.* North Atlantic humpback whale abundance and rate of increase four decades after protection from whaling. *Mar. Ecol. Prog. Ser.* **258**, 263–273 (2003).
44. Wieczorek, J. *et al.* Darwin core: An evolving community-developed biodiversity data standard. *PLoS One* **7**, e29715 (2012).
45. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).
46. Lammers, M. O. *et al.* The occurrence of humpback whales across the Hawaiian archipelago revealed by fixed and mobile acoustic monitoring. *Front. Mar. Sci.* **10**, 49 (2023).
47. Lammers, M. O. *et al.* Humpback whale *Megaptera novaeangliae* song reveals wintering activity in the Northwestern Hawaiian Islands. *Mar. Ecol. Prog. Ser.* **423**, 261–268 (2011).
48. Johnston, D. W., Chapla, M. E., Williams, L. E. & Mattila, D. K. Identification of humpback whale *Megaptera novaeangliae* wintering habitat in the Northwestern Hawaiian Islands using spatial habitat modeling. *Endanger. Species Res.* **3**, 249–257 (2007).
49. Ten Michener, W. K. Simple rules for creating a good data management plan. *PLoS Comput. Biol.* **11**, e1004525 (2015).
50. Darling, J. D. *et al.* Humpback whales (*Megaptera novaeangliae*) attend both Mexico and Hawaii breeding grounds in the same winter: Mixing in the northeast Pacific. *Letters* https://doi.org/10.1098/rsbl.2021.0547 (2022).
51. Tackaberry, J., Dobson, E., Flynn, K., Cheeseman, T. & Calambokidis, J. Low resighting rate of entangled humpback whales within the California, Oregon, and Washington region based on photo-identification and long-term life history data. *Front. Mar. Sci.* **8**, 2092 (2022).
52. Curtis, K. A. *et al.* NOAA technical memorandum NMFS Abundance of humpback whales (*Megaptera Novaeangliae*) wintering in Central America and Southern Mexico from a one-dimensional spatial capture-recapture model https://doi.org/10.25923/9cq1-rx80 (2022).
53. Patton, D. & Lawless, S. Surface and underwater observation of a humpback whale (*Megaptera novaeangliae*) birth in progress off Lahaina, Maui, and subsequent encounter of the female with a healthy calf. *Aquat. Mammals* https://doi.org/10.1578/AM.47.6.2021.550 (2021).
54. Henderson, E. E., Deakos, M. & Engelhaupt, D. Dive and movement behavior of a humpback whale competitive group and a multiday association between a primary escort and female in Hawai‘i. *Mar. Mammal Sci.* https://doi.org/10.1111/MMS.12891 (2021).
55. Lowe, C. L. *et al.* Patterns of cortisol and corticosterone concentrations in humpback whale (*Megaptera novaeangliae*) baleen are associated with different causes of death. *Conserv. Physiol.* https://doi.org/10.1093/conphys/coab096 (2021).
56. Lowe, C. L. *et al.* Case studies on longitudinal mercury content in humpback whale (*Megaptera novaeangliae*) baleen. *Heliyon* **8**, e08681 (2022).
57. Gewin, V. Data sharing: An open mind on open data. *Nat.* **529**, 117–119 (2016).
58. Mills, J. A. *et al.* Archiving primary data: Solutions for long-term studies. *Trends Ecol. Evol.* **30**, 581–589 (2015).
59. Urbano, F., Cagnacci, F. & Initiative, E. C. Data management and sharing for collaborative science: Lessons learnt from the euromammals initiative. *Front. Ecol. Evol.* **9**, 727023 (2021).
60. Suryan, R. M. *et al.* Ecosystem response persists after a prolonged marine heatwave. *Sci. Rep.* **11**, 1–17 (2021).
61. Samhouri, J. F. *et al.* Defining ecosystem thresholds for human activities and environmental pressures in the California current. *Ecosphere* **8**, e01860 (2017).
62. Samhouri, J. F. *et al.* Marine heatwave challenges solutions to human–wildlife conflict. *Proc. R. Soc. B* **288**, 1964 (2021).
63. Santora, J. A. *et al.* Habitat compression and ecosystem shifts as potential links between marine heatwave and record whale entanglements. *Nat. Commun.* **11**, 1–12 (2020).
64. Costello, M. J., Horton, T. & Kroh, A. Sustainable biodiversity databasing: International, collaborative, dynamic. *Centralised. Trends Ecol. Evol.* **33**, 803–805 (2018).
65. Hackett, R. A. *et al.* A data management workflow of biodiversity data from the field to data users. *Appl. Plant Sci.* **7**, e11310 (2019).
66. Lin, D. *et al.* The TRUST Principles for digital repositories. *Sci. Data* **7**, 1–5 (2020).
67. Savoca, M. S. *et al.* Baleen whale prey consumption based on high-resolution foraging measurements. *Nature* **599**, 85–90 (2021).
68. Jetz, W. *et al.* Essential biodiversity variables for mapping and monitoring species populations. *Nat. Ecol. Evol.* **3**, 539–551 (2019).
69. Kissling, W. D. *et al.* Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biol. Rev.* **93**, 600–625 (2018).
70. Bax, N. J. *et al.* A response to scientific and societal needs for marine biological observations. *Front. Mar. Sci.* **6**, 395 (2019).
71. Muller-Karger, F. E. *et al.* Advancing marine biological observations and data requirements of the complementary Essential Ocean Variables (EOVs) and Essential Biodiversity Variables (EBVs) frameworks. *Front. Mar. Sci.* **5**, 211 (2018).
72. Miloslavich, P. *et al.* Essential ocean variables for global sustained observations of biodiversity and ecosystem changes. *Glob. Chang. Biol.* **24**, 2416–2433 (2018).
73. Lynch, C. How do your data grow?. *Nature* **455**, 28–29 (2008).
74. Urbano, F. *et al.* Wildlife tracking data management: a new vision. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 2177–2185 (2010).
75. Titova, O. V. *et al.* Photo-identification matches of humpback whales (*Megaptera novaeangliae*) from feeding areas in Russian Far East seas and breeding grounds in the North Pacific. *Mar. Mammal Sci.* **34**, 100–112 (2018).
76. Calambokidis, J. & Barlow, J. Abundance of blue and humpback whales in the Eastern North Pacific estimated by capture-recapture and line-transect methods. *Mar. Mammal Sci.* **20**, 63–85 (2004).
77. Calambokidis, J. *et al.* Interchange and isolation of humpback whales off California and other North Pacific feeding grounds. *Mar. Mammal Sci.* **12**, 215–226 (1996).
78. Calambokidis, J. & Barlow, J. Trends in the abundance of Humpback Whales in the North Pacific Ocean from 1980 to 2006. in *WC Report SC/A17/NP/10 for the Workshop on the Comprehensive Assessment of North Pacific Humpback Whales. 18–21 April 2017.* 16 (2017).
79. Calambokidis, J. *et al.* Migratory destinations of humpback whales that feed off California, Oregon and Washington. *Mar. Ecol. Prog. Ser.* **192**, 295–304 (2000).
80. Palacios, D. M. *et al.* Humpback Whale Tagging in Support of Marine Mammal Monitoring Across Multiple Navy Training Areas in the Pacific Ocean: Final Report for the Pacific Northwest Feeding Area in Summer/Fall 2019, Including Historical Data from Previous Tagging Efforts off the US West Coast. Prepared for Commander, U.S. Pacific Fleet. Submitted to Naval Facilities Engineering Command Southwest, under Cooperative Ecosystem Studies Unit, Department of the Navy Cooperative Agreement No. N62473-19-2-0002. Oregon State University, Newport, Oregon. pp. 153 (2020).

81. Burdin, A. M., Titova, O. & Hoyt, E. in *Humpback Whales of Russian Far East Seas. Photo-ID Catalog 2004–2014*. (Publisher: Russian Geographical Society, 2014). Available at: https://www.researchgate.net/publication/272493253_Humpback_Whales_of_Russian_Far_East_Seas_Photo-ID_Catalog_2004-2014. (Accessed: 22nd Aug 2022)

Acknowledgements

The authors thank María González for creating manuscript figures, and Kaitlin Palmer for early review. The Happywhale dataset would not exist without the dedication, support, and trust of a very large community; please see Supplementary Information II for full acknowledgements.

Author contributions

T.C., J.C., P.C. and K.S. conceived of the study, T.C., J.C., K.A., J.E.M., J.S. and S.T. secured funding, T.C. and K.S. developed the Happywhale system and web platform, T.C., J.C., K.A., J.S., S.T., P.C., J.M.A., L.B., C.B., A.L.B., J.K.B., R.C., J.J.C., A.J.G.C., J.C., J.D.W., N.D., T.D.V., K.D., O.F., R.F., K.F., J.F., A.F.J., C.G., B.G., C.H., J.H., M.C.H., J.J., M.J., N.K., E.L., M.M., E.M., P.M.L., C.M., C.M.M., J.R.M., J.N., H.R.N., H.N., H.O., M.O., A.P., D.P., H.P., E.Q.R., R.F.R.B., N.R., F.S., T.S,. S.S., I.S., A.S, O.T., J.U., M.V.A., O.V.Z., B.W., J.W., K.M.Y. and D.Z. contributed data, T.C., J.C., N.D., K.F., P.M.L., E.J.L., H.N., M.O., R.F.R.B., D.Z. and A.M. managed data, T.C., J.J.C., K.F. and D.P. crafted the manuscript including table and figure development, T.C., K.A., J.B., A.B., P.C., J.J.C., J.D.W., T.D.V., K.D., K.F., J.F., C.G., C.H., M.C., E.L., C.M.M., J.E.M., J.R.M., J.N., A.P., D.P., H.P., H.R.N., E.Q.R., N.R., J.S., O.T,. and M.V.A. edited the manuscript, and all authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-36928-1.

Correspondence and requests for materials should be addressed to T.C.

Reprints and permissions information is available at www.nature.com/reprints.

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