NHCHB 2020 - Glossary of Terms and Abbreviations

* Definitions are mostly taken verbatim from deWaard et al. (2019; doi: 10.1038/s41597-019-0320-2), Cristescu (2014; doi: 10.1016/j.tree.2014.08.001), Eberle et al. (2020; doi: 10.1016/j.tree.2019.12.003), Porter & Hajibabaei (2018; doi: 10.1111/mec.14478), and Young & Gillung (2019; doi: 10.1111/syen.12406).

Adapters: Short, synthetic, single-stranded DNA molecules that are ligated to the ends of DNA fragments of interest. These allow target DNA fragments to bind to a flow cell for amplification in a HTS platform. Adapters are indexed or 'barcoded' with a short, unique sequence usually six to 10 bases long, which enables multiple samples to be sequenced simultaneously (multiplexing). After sequencing, these indexes are used to associate individual reads back to their correct sample (demultiplexing).

Amplicon: The short DNA sequence products of polymerase chain reaction (PCR) amplification using taxon- or gene-specific primers to target a particular region of the genome.

BIN system: The Barcode Index Number system provides an objective approach for OTU delineation of animals that is coupled with a persistent registry. Since BINs correspond well with Linnaean species in many animal groups, BIN-based biodiversity assessments can be implemented for groups that lack well-developed taxonomy. BINs are assigned by the Refined Single Linkage (RESL) algorithm implemented on BOLD. Individual records are either assigned to an existing BIN or found a new BIN, but they only enter the RESL analysis if they meet the following criteria: greater than 300 bp coverage of the barcode region, less than 1% ambiguous bases, and no stop codon or contamination of the sequence. For inclusion into an existing BIN, sequence records must include >300 bp of the barcode region (between positions 70 and 700 of the BOLD alignment) while records that establish a new BIN must include >500 bp of the barcode region. The RESL algorithm runs monthly on all qualifying barcode sequences in BOLD – which currently contains 7.9 million animal specimen records and 0.66 million BINs (as of January 2020). BIN designations and assignments generated by RESL on BOLD are accessible for independent validation through the 'BIN pages' that aggregate the specimen and sequence information of its members (e.g. the eastern yellowjacket wasp, *Vespula maculifrons* (Buysson): https://doi.org/10.5883/BOLD:AAD5593).

Biodiversity genomics: Biodiversity assessed using high-throughput DNA-based methods or data from whole genomes integrated with a broad array of metadata describing biological and environmental indicators.

Biodiversity: The diversity of life, their relationships and their functions within ecosystems.

Biomonitoring: Biodiversity analysis that is repeated across space and time that may focus on a target organism such as invasive or at-risk species, an assemblage such as the bioindicator groups (amphibians, birds, macroinvertebrates) as an indicator of ecosystem status.

BIOSCAN: Current research program of iBOL, initiated in June 2019 (www.ibol.org/programs/bioscan/). BIOSCAN's three research themes employ DNA barcodes to speed *species discovery*, to probe *species interactions*, and to track *species dynamics*.

BMTA: Biological Material Transfer Agreement.

BOLD: Barcode of Life Data System (www.boldsystems.org), the central informatics platform for iBOL. BOLD is the online workbench and database that supports the assembly, analysis, and publication of DNA barcode data.

CBG: Centre for Biodiversity Genomics (University of Guelph, Canada).

CCDB: Canadian Centre for DNA Barcoding (also known as CBG Genomics).

CITES: The Convention on International Trade in Endangered Species of Wild Fauna and Flora is an international agreement between governments aimed to ensure the international trade of wild fauna and flora does not threaten their survival.

COI: Cytochrome *c* oxidase subunit 1.

Contig: A contiguous sequence of DNA constructed by merging (or assembly) at least two individual DNA sequence reads. With traditional Sanger sequencing, contigs are typically constructed from a forward and reverse read of the same gene region, whereas in NGS methods, contigs are typically assembled using large numbers of partially overlapping reads in both the forward and reverse directions.

CTAB: Cetyl trimethylammonium bromide is a detergent used in DNA extraction to facilitate the separation of polysaccharides.

De novo assembly: The processes of constructing contigs and scaffolds without the use of a pre-existing reference genome from a related organism. The methods used by de novo assembly software are varied, but the most common type for short reads is assembly by de Bruijn graphs.

DNA barcoding: The identification of species using standardized DNA fragments. The ideal DNA barcoding procedure starts with well-curated voucher specimens deposited in natural history collections and ends with a unique sequence deposited in a public reference library of species identifiers that could be used to assign unknown sequences to known species. The standardized barcode for most animals is a fragment of the mitochondrial COI gene, the standardized barcode for plants is a fragment of the plastid gene ribulose 1,5-bisphosphate carboxylase gene (rbcL) combined with a fragment of the maturase (matK) gene, whereas the barcode for fungi is the nuclear internal transcribed spacer (ITS) of the ribosomal DNA.

DOI: A Digital Object Identifier is a registered, accessible, and persistent identifier used on digital networks.

eDNA: Environmental DNA comprised of free degraded DNAs in the environment as well as DNA co-extracted from whole organisms such as microscopic organisms, arthropods, nematodes; shed cells; faeces; as well as the DNA contained within dead or dormant cells such as seeds or spores.

Exons: Parts of a gene that become part of the mature mRNA; can contain untranslated regions and coding sequences that are translated into amino acids.

Gene tree: A phylogenetic tree based on data (or information) from a single locus.

Genome skimming: Shallow shotgun sequencing of total genomic DNA of an organism.

Genome: The complete set of genetic data contained in an organism including organellar DNA.

Genomics: The sequencing and analysis of the genetic material of an organism.

GGBN: Global Genome Biodiversity Network (www.ggbn.org).

GuSCN: Guanidine thiocyanate, a chaotropic agent commonly used in DNA and RNA extraction.

Homology: similarity due to shared ancestry. Opposed to analogous characters, homologous characters can be compared across organisms to infer, for example, phylogenetic relatedness or species boundaries.

HPC: High-performance computing, computer clusters can be used to run the same analysis for many samples in parallel, or splitting large jobs into many smaller ones for a quicker overall runtime. Available through private clusters or third-party cloud computing services.

HTS: High-throughput sequencing, sometimes referred to as next-generation sequencing or second-generation sequencing. Distinguished by the high number of sequencing reactions that occur in parallel.

iBOL: The International Barcode of Life Consortium (www.ibol.org)

Incomplete lineage sorting (ILS): The random inheritance of only some gene variants (or alleles) by a new species from a founding population during speciation events. ILS can cause mismatches between the species tree and individual gene trees.

Introns: transcribed non-coding parts of a gene that are removed by RNA splicing during mRNA maturation.

ITS: Nuclear internal transcribed spacer.

KOAc: Potassium acetate, used as a buffer in the isolation of DNA.

Marker: A gene or signature region of DNA with a known location in the genome and can be used to identify individuals or species.

Metadata: Supplementary data linked to DNA sequences that provide information in a standard and searchable way such as organismal or bulk environmental sample description.

matK: Maturase gene.

Metabarcoding: a rapid method of high-throughput, DNA-based identification of multiple species from a complex and possibly degraded sample of eDNA or from mass collection of specimens. The metabarcoding approach is often applied to microbial communities, but can be also applied to meiofauna or even megafauna.

Metagenomics: The study of genetic material isolated directly from environmental samples, such as water, soil or sediments, may also be referred to as environmental genomics, ecogenomics or community genomics.

Mitochondrial metagenomics: The assembly of whole mitochondrial DNA sequences from eDNA samples.

Multiplexing: A procedure that allows large numbers of DNA libraries to be pooled and sequenced simultaneously during a single run on a high-throughput instrument (e.g. Illumina). Individual barcode sequences are added to each library so that reads can later be identified and associated with their respective species.

NGS: Next generation sequencing – see HTS.

NHC: Natural history collection.

NCBI: National Center for Biotechnology Information is a resource for biomedical and genomic information that houses GenBank and other repositories.

Oligonucleotides: Relatively short nucleotide molecules used as primers for PCR, as probes on microarrays, or baits during target enrichment.

Orthologs: genes that arose from a single ancestral gene in an organismic group of interest by speciation. By contrast, paralogs are genes that arose in an organismic lineage of interest from an ancestral gene by gene duplication within a genome.

OTU: Operational taxonomic unit, a group of similar DNA sequences sometimes used as a proxy for “species” in diversity measures.

Primers: Short oligonucleotides that are complementary to a particular region of the genome and are a starting point for DNA replication by DNA polymerase during PCR.

PVA: Polyvinyl alcohol, a polymer used for DNA preservation.

rbcL: Ribulose 1,5-bisphosphate carboxylase gene.

Read: A single-stranded DNA sequence that has been 'read' by a DNA sequencer.

Reference-guided assembly: Also called referenced assembly, this procedure aligns individual reads to a pre-existing reference sequence or genome from a related organism to construct contigs and scaffolds.

Refined Single Linkage (RESL) algorithm: Implemented on BOLD to assign BINs to individual records (see BIN system).

rRNA genes: genes that code structural components of ribosomes. In metazoans, the corresponding genes of the large subunit (60S) are 5S, 5.8S, and 28S, and that of the small subunit (40S) is 18S. These genes are arranged in tandem repeats and present in large numbers.
Sequencing depth. The number of times individual bases are sequenced. To assemble contigs of single-copy loci in a genome, greater sequencing depth is required, as multiple-copy loci will be sequenced repeatedly before sufficient reads from single-copy areas of the genome are obtained.

SDS: Sodium dodecyl sulfate is a detergent used in DNA extraction protocols to disrupt the cell membrane of cells to expose the genomic DNA.

Single-copy gene: gene that is present in a single copy in all genomes of the respective organismic lineage. Species delimitation: recognition of boundaries between species.

Species identification: Assigning a taxonomic name to a species. Species boundaries must be known in advance.

Super-barcoding: The use of whole (or near whole) organellar DNA sequences for species identification.

Taxon: An organism identified to any taxonomic rank (e.g., species to kingdom); plural taxa.

Taxonomy: The science of discovering, describing, classifying, and naming organisms.

TE: Target enrichment focusses HTS resources on subsets of the genome that are of interest (generally through hybrid capture), leading to reduced costs and simplified analyses.

Ultra-conserved elements (UCEs): DNA segments of about 200 base pairs that are highly conserved across a large phylogenetic range. They may be present in non-coding regions or overlap with genes (exons and introns).

UMI: Unique molecular identifiers (or MID, multiplex identifier tags) are short sequences added to DNA fragments in some HTS library preparation protocols that function to identify the input fragment following sequencing.

Universal single-copy orthologs (USCOs): Sets of protein protein-coding genes that were introduced in an attempt to provide a benchmarking for assessing the completeness and quality of genome and transcriptome assemblies.

Whole-genome sequencing (WGS): Whole-genome sequencing involves determining the complete DNA sequence of an organism's genome, also known as complete genome sequencing, full-genome sequencing, entire genome sequencing.

16S: Ribosomal RNA gene of the small subunit (30S) in prokaryotes **or** the mitochondrially encoded ribosomal RNA gene in eukaryotes.

18S: Ribosomal RNA gene of the small subunit (40S) in metazoans.