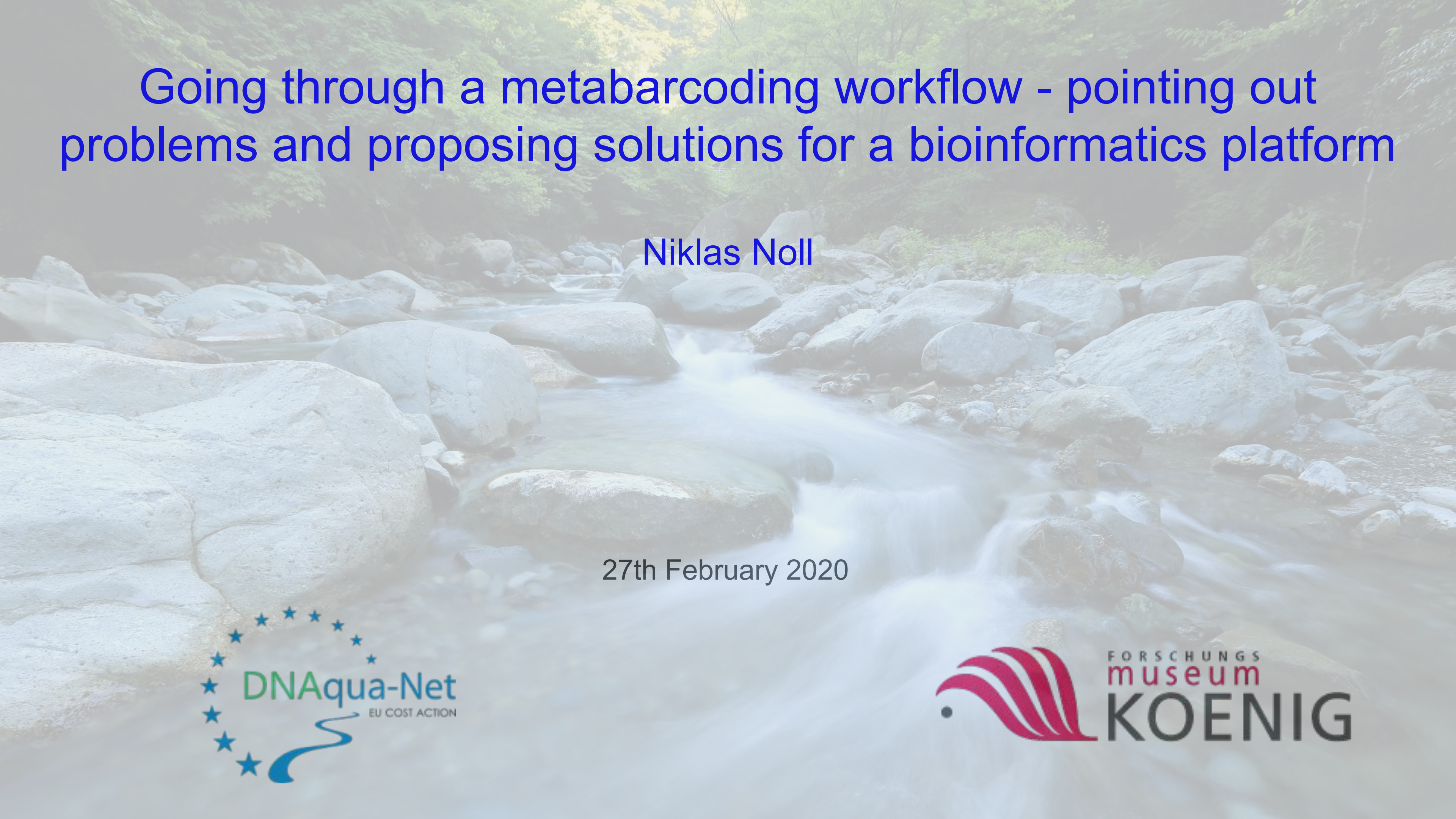


Going through a metabarcoding workflow - pointing out problems and proposing solutions for a bioinformatics platform

Niklas Noll

27th February 2020

DNAqua-Net
EU COST ACTION

FORSCHUNGS museum KOENIG

Purpose

Showing problems of current metabarcoding analysis

Propose suggestions for an online metabarcoding platform

Summary

- The platform needs to be flexible to keep up with changes
- The fast pace of technological innovations complicate standardization
- An online platform should include best practices and recommendations
- Databases should have well curated and public records
- Analysis workflow needs to be transparent and reproducible
- It should be possible to repeat the analysis at any time

Agenda

Why flexibility is important

sequencing

P0 P1 P2 P3 P4

Agenda

Why flexibility is important

Things to consider during read preparation

read preparation

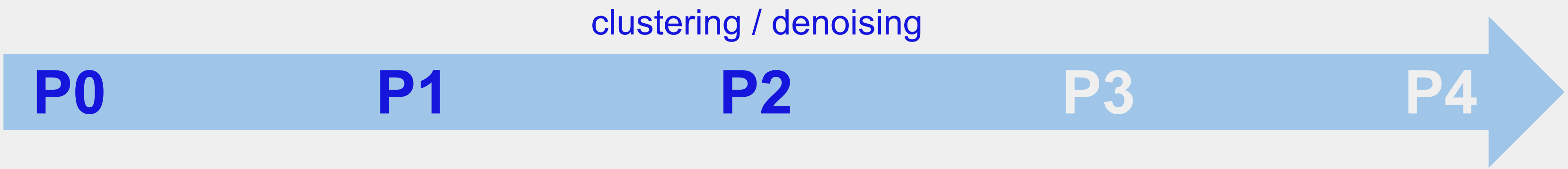
P0 P1 P2 P3 P4

Agenda

Why flexibility is important

Things to consider during read preparation

How to handle sequencing errors



clustering / denoising

P0 P1 P2 P3 P4

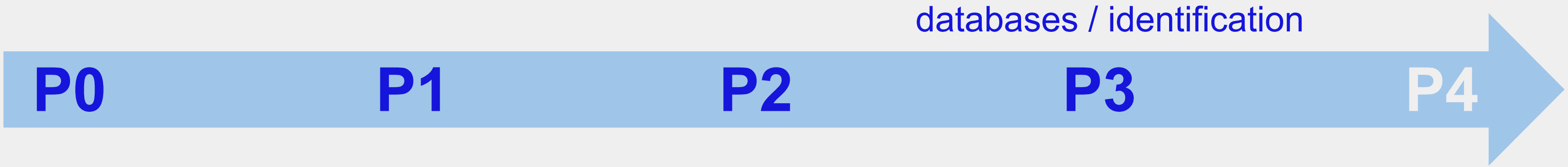
Agenda

Why flexibility is important

Things to consider during read preparation

How to handle sequencing errors

The need for well curated reference databases


databases / identification

P0

P1

P2

P3

P4

Agenda

Why flexibility is important

Things to consider during read preparation

How to handle sequencing errors

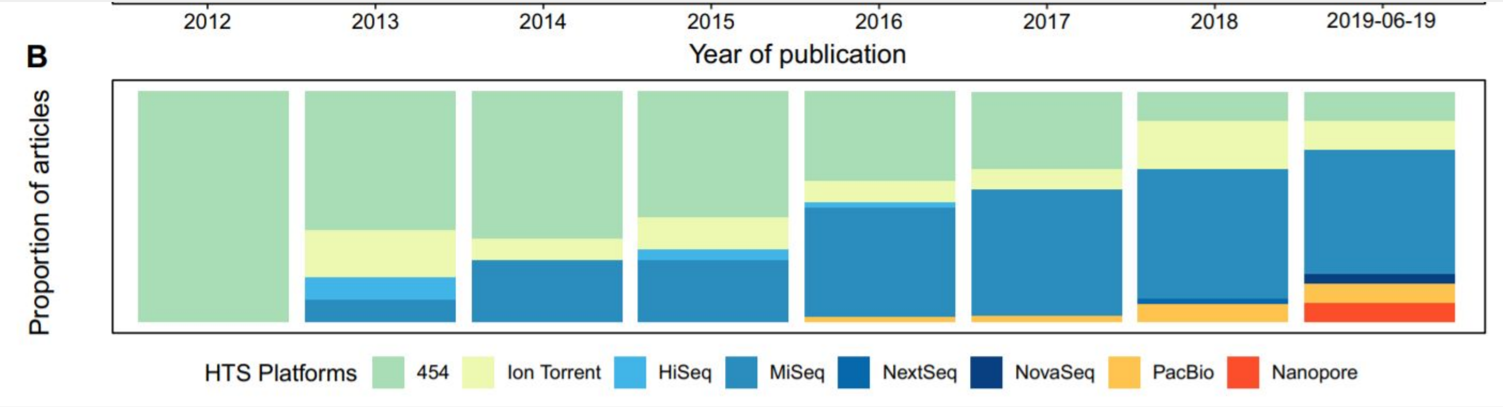
The need for well curated reference databases

How to deal with the final results

taxa list / bioassessment

P0 P1 P2 P3 P4

Sequencing technology and it’s usage changes fast



(Piper et al., 2019)

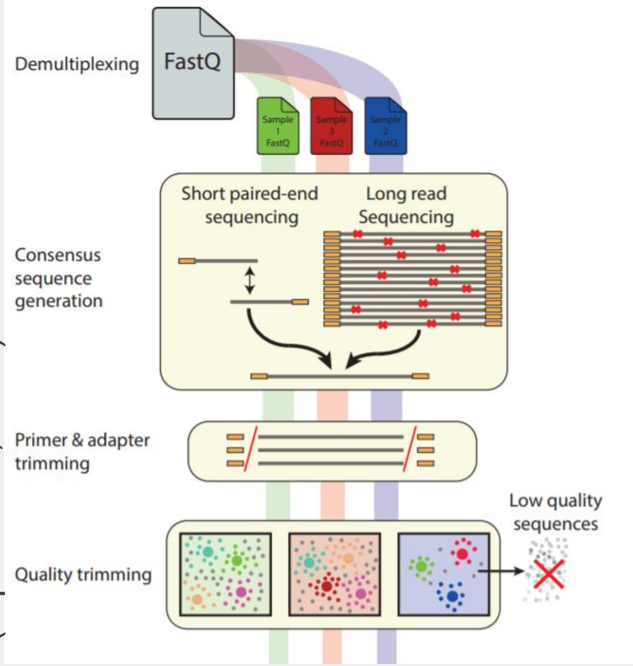
Therefore the platform should constantly expande

- ...to be compatible with new sequencing technologies
- But should also support older output formats

sequencing

P0 P1 P2 P3 P4

Read preparation needs to be flexible



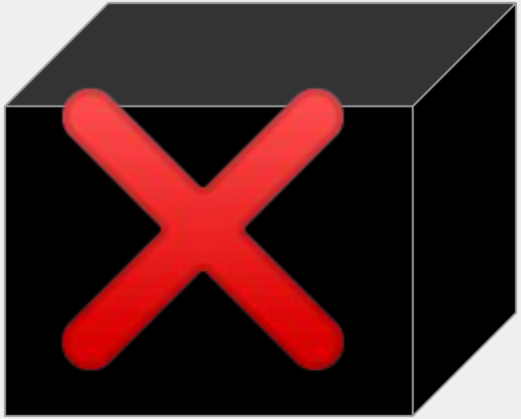
(Piper et al., 2019)

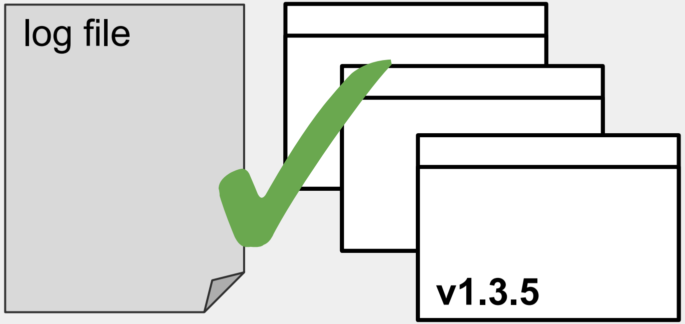
- Trade-off between sensitivity to rare taxa, errors and computing time: priority depends on goal
- Every tool and setting should be adjustable and explained
- Recommendations and guidance need to be dataset-specific

read preparation

P0 **P1** **P2** **P3** **P4**

A online platform should have a logging and versioning system



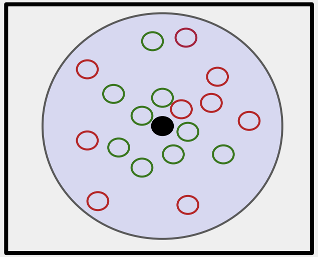


read preparation

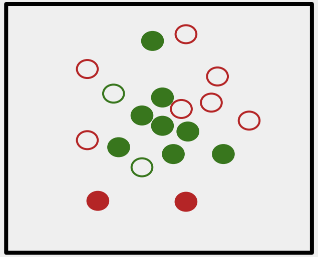
P0 P1 P2 P3 P4

OTU clustering and denoising should be included

OTU clustering

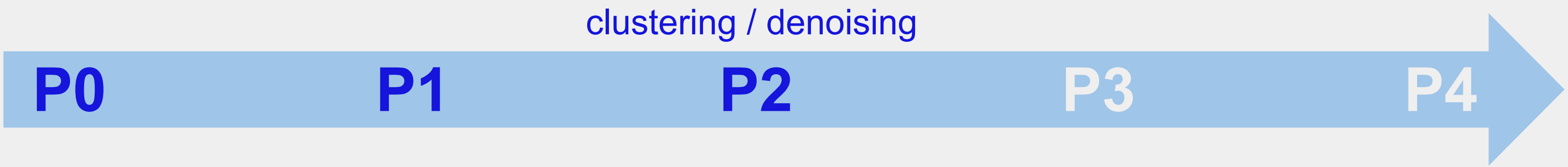


denoising



- Both methods are frequently used and need to be implemented

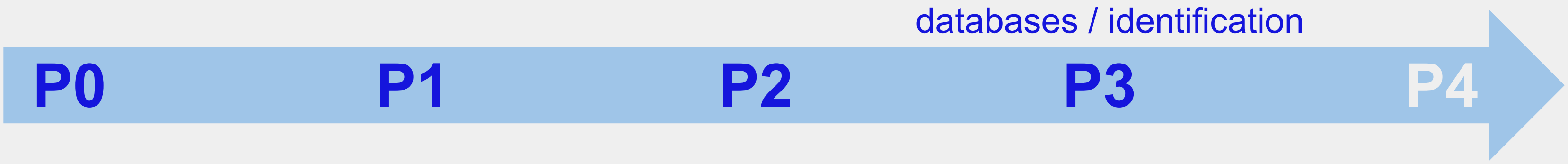
- In the long run we should switch to denoising
    - improved taxonomic resolution
    - with OTU clustering the analysis has to be rerun if new samples are added
    - OTU clustering becomes computationally extensive



clustering / denoising

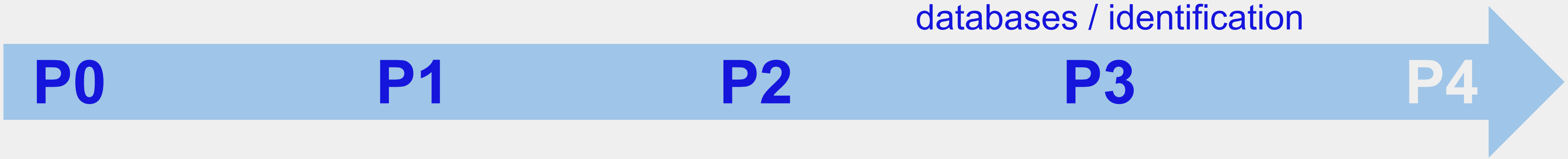
Downloading many sequences from BOLD is slow and unreliable

- Download speed is about 100 kb/s for Arthropoda dataset
- Download crashes regularly



The BOLD online identification system is very restricted

- 1 sequence per query for not-registered users, 100 sequences for registered users
- Only 4 pre-defined current reference databases are usable for CO1
- Search result per query has no meta data and is only available as web page
- Search Parameters are fixed

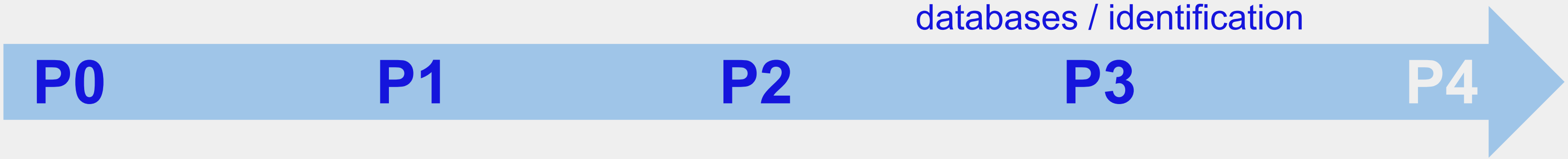


databases / identification

P0 P1 P2 P3 P4

Meta and taxon information are sometimes wrong

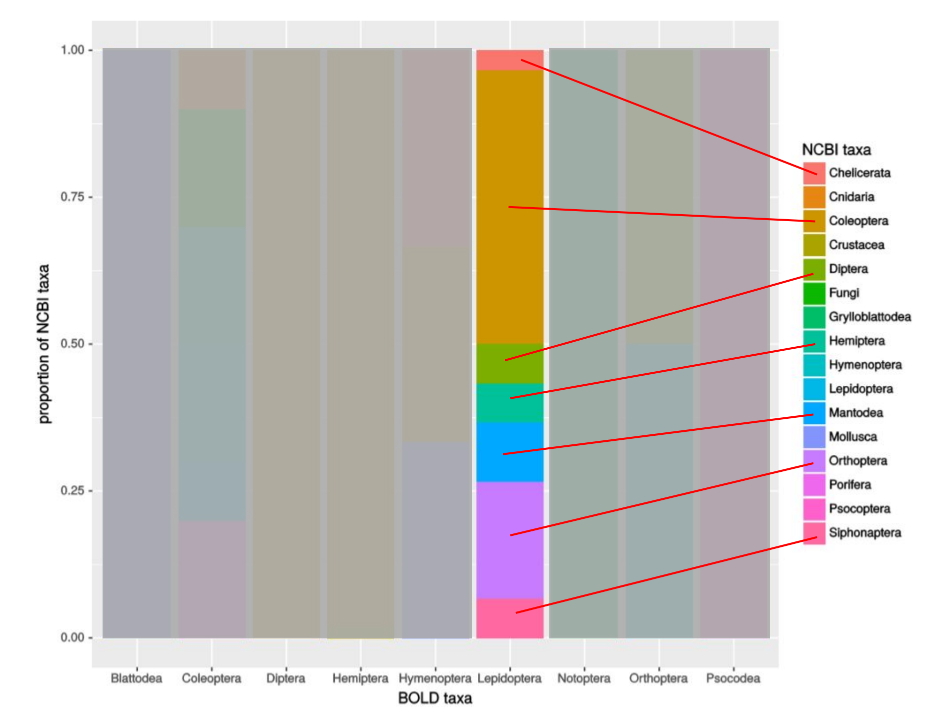
- Incorrect species identification
- Insufficiently annotated sequences
- Mining Errors



databases / identification

P0 P1 P2 P3 P4

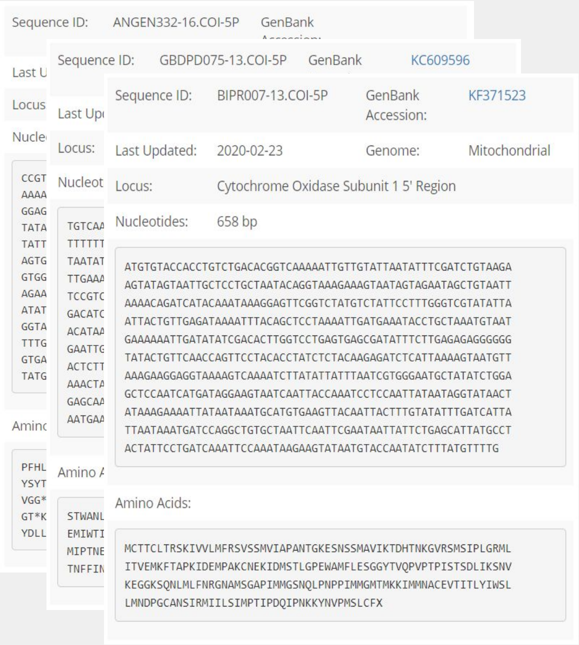
BOLD assigns different taxa than GenBank



databases / identification

P0 **P1** **P2** **P3** **P4**

Nucleotide sequences and their translations have errors



Sequence ID: ANGEN332-16.COI-5P GenBank

Sequence ID: GBDPD075-13.COI-5P GenBank KC609596

Sequence ID: BIPR007-13.COI-5P GenBank Accession: KF371523

Last Updated: 2020-02-23 Genome: Mitochondrial

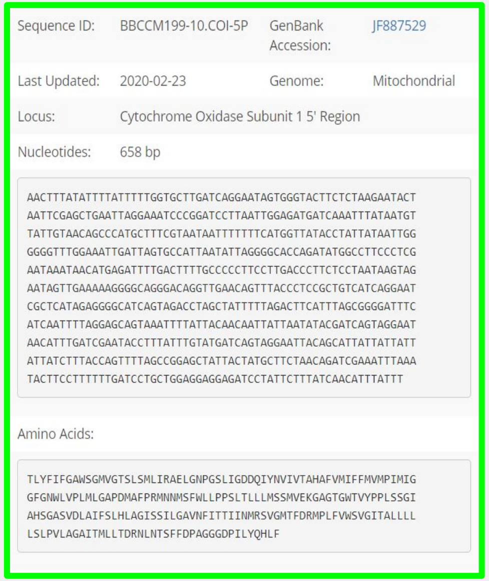
Locus: Cytochrome Oxidase Subunit 1 5' Region

Nucleotides: 658 bp

```
ATGTGTACCACCTGTCTGACACGGTCAAAAATTGTTGTATTAATATTTCGATCTGTAAGA
AGTATAGTAATTGCTCCTGCTAATACAGGTAAAGAAAGTAATAGTAGAATAGCTGTAATT
AAAACAGATCATACAAATAAAGGAGTTCGGTCTATGTCTATTCCTTTGGGTCGTATATTA
ATTACTGTTGAGATAAAATTTACAGCTCCTAAAATTGATGAAATACCTGCTAAATGTAAT
GAAAAAATTGATATATCGACACTTGGTCCTGAGTGAGCGATATTTCTTGAGAGAGGGGGG
TATACTGTTCAACCAGTTCCTACACCTATCTCTACAAGAGATCTCATTAAAAGTAATGTT
AAAGAAGGAGGTAAAAGTCAAAATCTTATATTATTTAATCGTGGGAATGCTATATCTGGA
GCTCCAATCATGATAGGAAGTAATCAATTACCAAATCCTCCAATTATAATAGGTATAACT
ATAAAGAAAATTATAATAAATGCATGTGAAGTTACAATTACTTTGTATATTTGATCATTA
TTAATAAATGATCCAGGCTGTGCTAATTCAATTCGAATAATTATTCTGAGCATTATGCCT
ACTATTCCTGATCAAATTCCAAATAAGAAGTATAATGTACCAATATCTTTATGTTTTG
```

Amino Acids:

```
MCTTCLTRSKIVVLMFRSVSSMVIAPANTGKESNSSMAVIKTDHTNKGVRSMSIPLGRML
ITVEMKFTAPKIDEMPAKCNEKIDMSTLGPEWAMFLESGGYTVQPVPTPISTSDLIKSNV
KEGGKSQNLMLFNRGNAMSGAPIMMGSNQLPNPPIMMGMTMKKIMMNACEVTITLYIWSL
LMNDPGCANSIRMIILSIMPTIPDQIPNKKYNVPMSLCFX
```



Sequence ID: BBCCM199-10.COI-5P GenBank Accession: JF887529

Last Updated: 2020-02-23 Genome: Mitochondrial

Locus: Cytochrome Oxidase Subunit 1 5' Region

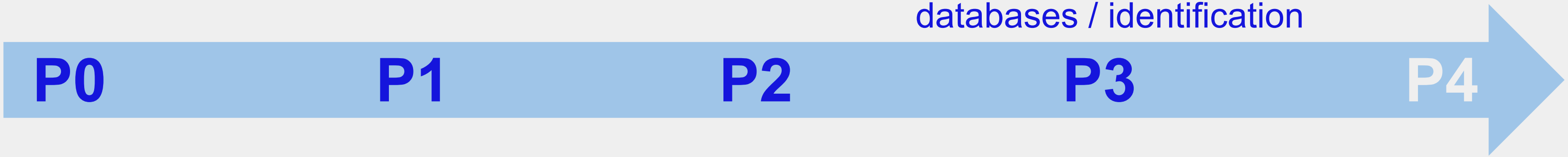
Nucleotides: 658 bp

```
AACTTTATATTTTATTTTTGGTGCTTGATCAGGAATAGTGGGTACTTCTCTAAGAATACT
AATTCGAGCTGAATTAGGAAATCCCGGATCCTTAATTGGAGATGATCAAATTTATAATGT
TATTGTAACAGCCCATGCTTTCGTAATAATTTTTTTCATGGTTATACCTATTATAATTGG
GGGGTTTGGAAATTGATTAGTGCCATTAATATTAGGGGCACCAGATATGGCCTTCCCTCG
AATAAATAACATGAGATTTTGACTTTTGCCCCCTTCCTTGACCCTTCTCCTAATAAGTAG
AATAGTTGAAAAAGGGGCAGGGACAGGTTGAACAGTTTACCCTCCGCTGTCATCAGGAAT
CGCTCATAGAGGGGCATCAGTAGACCTAGCTATTTTTAGACTTCATTTAGCGGGGATTTC
ATCAATTTTAGGAGCAGTAAATTTTATTACAACAATTATTAATATACGATCAGTAGGAAT
AACATTTGATCGAATACCTTTATTTGTATGATCAGTAGGAATTACAGCATTATTATTATT
ATTATCTTTACCAGTTTTAGCCGGAGCTATTACTATGCTTCTAACAGATCGAAATTTAAA
TACTTCCTTTTTTGATCCTGCTGGAGGAGGAGATCCTATTCTTTATCAACATTTATTT
```

Amino Acids:

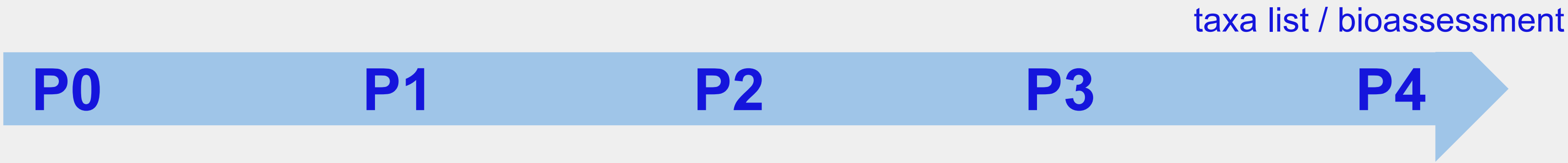
```
TLYFIFGAWSGMVGTSLSMLIRAELGNPGSLIGDDQIYNVIVTAHAFVMIFFMVMPIMIG
GFGNWLVPLMLGAPDMAFPRMNNMSFWLLPPSLTLLLMSSMVEKGAGTGWTVYPPLSSGI
AHSGASVDLAIFSLHLAGISSILGAVNFITTIINMRSVGMTFDRMPLFVWSVGITALLLL
LSLPVLAGAITMLLTDRNLNTSFFDPAGGGDPILYQHLF
```

databases / identification



The whole analysis should be easy to repeat

- Analysis workflow should be transparent and reproducible
- All users should be able to rerun a analysis, especially if a new database version, or a new software version is released
- Taxa lists, ASVs, OTUs and bioassessment outcomes are available for other users



taxa list / bioassessment

P0 P1 P2 P3 P4

Summary

- The platform needs to be flexible to keep up with changes
- The fast pace of technological innovations complicate standardization
- An online platform should include best practices and recommendations
- Databases should have well curated and public records
- Analysis workflow needs to be transparent and reproducible
- It should be possible to repeat the analysis at any time

Citations

- Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C., & Blacket, M. J. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. GigaScience, 8(8). https://doi.org/10.1093/gigascience/giz092

- Title photo by kazuend on Unsplash