Breakout groups
Feedback

Should the SeqCode be restricted to uncultured organisms?

Question 1:

- NO, if a high quality genome is available and whatever else criteria settled by the SeqCode are fulfilled, a species should be covered by the SeqCode, no matter whether cultured or not.
- The info that a microorganism has been cultured (at least once) is nice to have and flags could be used to indicate such cases, with detailed information on the availability of the corresponding strains.
- If we have two codes in parallel, most microbiologists will ask and go for the easier procedure.
  So, if the SeqCode is easier to validate a new name, microbiologists with isolates will be annoyed if the SeqCode is NOT accepting their data.
  The discussion panel points out that this may lower the interest in depositing isolates in international collections.
  However, “nomenclature” and “deposition and sharing of isolates” are different topics.
- This may not be the major issue to be addressed by the SeqCode!? Creating and maintaining a stable system that takes care solely of the uncultured majority is already a tremendous task!
- The priority is to avoid chaos due to a surge of submissions by setting clear and transparent rules. The rest will follow.
  The SeqCode should not antagonize the current code and both sides should work together on a harmonized system once the SeqCode has been launched and works well.
- Besides fastidious organisms that fail to be deposited in international collections, strains from some countries cannot be deposited due to Nagoya regulations. The SeqCode will not be an alternative, as genome information must follow the same rules
- For a new code system to be adopted it needs to be simple & not confusing.
  Costs & long-term data sharing need... even more... discussion :)

- Mixed opinion- some in group yes, others “Having two systems with different standards doesn’t make sense. Genomes are a common currency”

- Nagoya- productive and cautionary discussion, digital sequence information/genetic resource https://www.youtube.com/watch?v=rO_pAQDDHII, https://www.dsmz.de/collection/nagoya-protocol/digital-sequence-information/dsi-policy-options-webinar-2020)

- “by getting all this data, people will want to grow again to understand the organisms. All genomes named and put into databases will encourage isolation”

- “We need cultures to learn about the biology, but not for naming”.

- “The genome alone is enough to provide a stable name”.

Should the SeqCode allow replacement of sequences as types with strains as types [should they become cultivated]?

Replacement of sequences as types: NO!

Cons: Not allowing replacement will: (i) ensure stability, (ii) avoid confusion (short-lived, only until approved lists are published), (iii) not impact motivation to cultivate

Concerns: (i) stringent genome descriptions and cut-offs are required, (ii) single database for comparative analysis with Type Sequences, (iii) Two nomenclature systems, (iv) enormity of unexplored majority

Possible Solution: (i) ICNP should recognize priority of names proposed under the SeqCode, (ii) publish periodic approved lists validating and establishing synonymy between Type Sequence and Type Strain, (iii) ICNP and SeqCode must work together, (iv) save type sequences in database with version numbers

Outcome: The Code/ ICNP is strictly dedicated to the description of taxa that can be cultivated. SeqCode will be more inclusive allowing naming of uncultivated taxa. For nomenclature stability and to avoid duplication of names and/or type material, type sequences should not be replaced and the ICNP should recognize priority of names from the SeqCode and their type material.

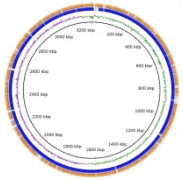
- Participants largely agreed that a new stain should supercede the SeqCode genomic sequence as type material, once an ICNP-conform description is available. It conforms with a higher standard of scientific rigor and reproducibility, and makes the taxon available for physiological and biochemical characterization.

- The naming was always an honor and reward for culturing something new. When switching type material, the authorship on the taxon could potentially be changed as well. A replacement of sequences as types with strains as types could, therefore, involve a replacement of the authors of the corresponding taxon description.

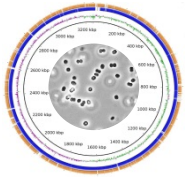
- Implementing this practice would fuel cultivation research.

SeqCode

Type material is a sequence



⟹



Code

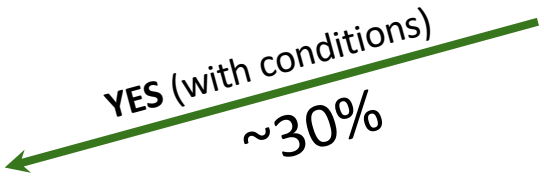
Type material is a culture

***Name* X^{Ts}, Authors (of the sequence-based taxon description)**

***Name* XX^{T}, Authors (of the culture-based taxon description)**

Should the SeqCode specify that priority of higher taxa depend on the priority of the genus and species

Priority of higher taxa depend on the priority of the genus and species



This would ensure **stability in the SeqCode**, but not necessarily stability in the names.

Priority = date < representation < cultivation

For this reason, only ultra-high quality genomes should be named (e.g., assembled to replicon).

Pros:

- Automate changes in taxonomy
- Trace taxa with priority & type
- Only high-quality MAGs named

Cons:

- Stability in names not guaranteed
- Not all known diversity encompassed

This would ensure **stability in names**, but not necessarily stability in the SeqCode.

Priority should only be used to determine type when taxa are transferred to a yet unnamed taxon/lineage.

Types not necessarily those with priority.

Pros:

- More stability in the names
- Trace taxa that are type not priority
- Better coverage of known diversity

Cons:

- Manual curation vs. automation
- Lower quality MAGs named

Should the SeqCode include a proposal for an Approved List to incorporate the current *Candidatus* taxa if they include a DNA sequence as type and are otherwise compliant with the SeqCode?

Yes.

- **Start with the Oren et al. *Candidatus* taxa**
  - Ensure they follow SeqCode requirements
- **Add in high quality MAGs and SAGs** (cutoffs to be determined)
  - Comb literature similar to Oren et al.
  - Identify key taxa in existing databases not represented in ICNP
  - Does this require a new database?
- This initial list may need to be followed up with more after requesting community input and/or submissions (easiest with new SeqCode database/search platform)

Other Considerations:

1. Many existing *Candidatus* taxa do not have a type designated; does the committee give power to assign types?
2. Some validly published taxa do not yet have a genome sequence; this may require consolidation of SeqCode-named taxa once matched to newly-sequenced ICNP taxa
3. Submission of data for SeqCode approval should include a classification search to determine relatedness to existing taxa (with suggested cutoff for level of novelty), and metadata (MIMAG/MISAG) including environmental characteristics,

- **Yes!**
- Starting point: Oren's Candidatus list
  - Must be checked for: (i) unambiguous reference to the type sequence (may be designated based on communication with authors), (ii) synonymy/redundancy with valid names under ICNP, other codes, and other *Candidatus* names, (iii) agreed standards, and (iv) comprehensiveness of the list.
  - Should we transfer a *Candidatus* name from a 16S rRNA gene sequence (no genome) to a MAG or SAG if the identity is high? (No?)
  - Should taxa on the Approved Lists get a distinct superscript (e.g., similar to ^{AL}) or should this just be noted in a metadata column of the SeqCode Registry? (No?)
- Standards: Should be separate from SeqCode but must be decided by one or more international committees and be flexible.
  - Initial standards could: (i) follow MiMAG/MiSAG criteria, requiring high-quality category, (ii) allow exceptions when a case is clearly made (e.g., degenerate symbiont genomes), (iii) be checked during registration of a genome, (iv) be customized for certain taxonomic groups if endorsed by a committee.

Should the SeqCode include a digital protologue as part of the naming process?

Yes!

But should avoid reinventing the wheel. For example:

Don’t reproduce genome sequence or raw data, link to INSDC database entry

Don’t introduce new unique identifier, adopt an existing one, e.g. genome assembly identifier.

Other resources with overlapping information that could be leveraged: LPSN, GOLD. Both have ongoing support, could a digital protologue become part of one of these resources?

Include a front end ANI-based classification of submitted genomes

To avoid proposing names for already named taxa

Requires sufficient financial support and Society endorsement to be successful

Exhibit A: failed first attempt due to lack of $$$ and IJSEM buy in

The team considered that the DP is a need, should capture all features relevant for the taxon description and should have all links to the metadata and relevant data in other repositories to make it interactive

If so, what information fields would be included in the digital protologue?

- Metadata included should reflect the existing standards as e.g. the provided by the Genomics Standards Consortium, and other databases as the former Digital Protologue Database and BacDive
- There should be some flexibility with some metadata fields compulsory and other optional
- May have as many auto-fillable fields as possible to avoid inconsistencies (e.g. orthography, metrics,…)
- Images of the organism could be optional in the DP
- Protologues could be autogenerated for genomic features using programs as the protologger.de, yet to be improved, but can capture the major diagnostic traits
- Non-type MAGs may be also included to complement not binned features in the type or enlarge the knowledge on the pangenome. Direct link to the raw reads of the metagenome for future reassemblies
- The tools used (assemblers, binning, …) should be specified to track possible problems & solutions
- A dedicated person may be needed for manual curation and not rely only on automatization

If a digital protologue were integrated with the SeqCode, presumably this could not be integrated with data for microbial isolates. Is there any path to do this?

- Protologues should be as broad as possible capturing the relevant features and metadata for cultured in case of merge.
- Automatization of protologues could be used for type strains for which genomes are available.
- The DP should be able to check whether not only the name, but the species exists in the database to avoid synonymy

Would a doi or some other numerical or alphanumerical code be useful for each name and associated digital protologue in the SeqCode? What is the importance of such a code and how should it best be implemented (e.g., for record keeping or bioinformatic analyses)?

- Yes, should have a DOI to be citable & also a unique identifier (Taxonumber). DOI cannot substitute Taxonumbers
- DOIs can avoid one MAG – one Species – One publication if are citable and serve as for appropriate credit attribution.

Should the SeqCode require ‘perfect Latin’ ?

- Silly question?!
  - should a scientific paper or talk require “perfect English”?
- No consensus in our group on whether there should be a SeqCode
  - Most of the fuss seems to centre on priority of names, which is already enforceable by norms/editors/reviewers without a new code
  - But if there is, it should not diverge in linguistic use from ICNP
- Only minor linguistic skills needed to create names
  - Certainly easier than mastering Python or R or the command line
  - No need to read Homer in the original Greek
  - More resources than ever to help (many online dictionaries)
  - And many more user-friendly resources to come (e.g. Gan 2.0)
- NB. the current code (Rule 61 of the ICNP) already allows linguistic errors to be corrected.

Perfect Latin

YES 80%

Implications:
There is the eventual opportunity for a unified Code that encompass cultivated and uncultivated prokaryotes.

Pros:

- Potential for unified Code
- People know how the binomial system works
- GAN allows naming at scale

Cons:

- Less appealing or accessible to non-biologist
- Difficult to apply in current state
- Potentially slower rate of naming

NO 20%

Implications:
Names formed under the SeqCode would likely be illegitimate under the ICNP = no chances for merging of the two Codes in unified Code.

Caveat: some rules still required, e.g., suffixes

Pros:

- Would greatly ease in naming process
- Overcome equity issues in process
- More accessible to non-biologists
- Increased rate of naming (anyone can)

Cons: - Difficulties in reaching a unified Code

How should the SeqCode be managed and revised?

Before establishing the administrative structure for the SeqCode, the **organizational structures of the nomenclature committees for plants, animals and viruses should be examined.** Personal inquiries should be made to determine the advantages and disadvantages these structures.

If these investigations do not result in alternative structures, **in principle the SeqCode could be managed and revised by two committees similar in structure to the ICSP and its Judicial Commission.** The ICSP-like committee would be responsible for administration of the validation process and updating the SeqCode. The Judicial-like Commission would be responsible for resolving disputes concerning the interpretation of the SeqCode.

In either case, **the size of the committees should balance the need for diversity while avoiding the inefficiencies of large committees. Diversity: Committees should be international and representative of as many scientific disciplines as possible**, such as clinical, symbionts, archaea, bacteria, marine, terrestrial, ecology, and extremophiles. **Size: Committees can form subcommittees to efficiently examine particular issues as they arise.** It is recommended that the committees have a small executive committee to deal with routine administration. **Of high priority will be the proper management of the validation system: transparent, efficient and of as little effort and cost to its users as possible.** One possibility would be to include the accession number of the sequence in the publication and include the validation number in the database accession. **The committees should be independent of the IUMS**. Membership could be determined by invitation from the current members as well as periodic nomination by societies.

Also discussed the quality of MAGs and SAGs necessary for validation. Some criteria included 90 % completeness, <1 % contamination; requirement for 16S rRNA gene sequence or a certain fraction of other marker genes or ribosomal proteins; of sufficient quality to generate e.g. probes for transcription and FISH (CARD-FISH) analysis and to interpret proteomic analysis to allow establishing the morphology and the potentially active metabolisms in order to acquire phenotypic description.

Comment was made that the ICSP is outdated and many people don't really care about the bureaucracy – are there sufficient people interested enough to run these organisations?

- Is there an opportunity for a more radical departure from the ICNP model? How do we avoid perpetuating systems where an unrepresentative 'club' has a stranglehold?
- Are there new social platforms that can replace old committee ways of working?
- Is it really progress to have a clone of the ICSP generating redundant names?
- What about a wikipedia "swarm intelligence" approach, but the question then is how is that managed and moderated - would maybe a simple registration be enough to get this organized?

Are there pipelines and database platforms that can help scale up naming? Do we have to leave it to machines and let humans deal with data quality, QC etc

- Would there need to be a system for keeping up with data as a 'moving target' e.g. genome data that are reassembled? Do raw reads from a type genome need to be released and stored, with the associated meta-data.

There was cautious optimism that improved stability of classification from phylogenomic methods might minimise the need for conflict resolution by a JC type body (which does not mean one isn't needed!)

How is naming (or even the desirability of naming) to be approached without first a consensus on what is being named at a given rank? The "I will know a species when I see it" question - or are we being narrow minded when biological reality is so much more complex?

Should the SeqCode include the rank of subspecies?

Summary: A mixed response from participants with the majority opposed to the idea of including the rank of subspecies in the SeqCode.

Support for the rank of subspecies

1. Genomic matrices (such as ANI and AAI) lack linearly and show bimodal/multimodal distribution that may be the result of major extinction events. These groups are subspecies within the species boundaries.
2. These natural discontinuities in genomic similarities are defined as subspecies using genome-based matrices (GGDC and TYGS).
3. Groups/individuals with clear differences in phenotypes or genome characteristics within species boundaries.

Arguments against the rank of subspecies

1. In evolutionary terms, these may be protospecies that will split into separate species.
2. Variations in the genomes cannot be linked to specific features to define subspecies.
3. No support for subspecies based on metagenomics analyses.
4. Species are easy to remember, subspecies will make it complicated.
5. Use of variety, ecotypes and pathovars over subspecies (has advantages and disadvantages).

A need of clear definition and criteria to make the rank acceptable

1. Genome-based thresholds (e.g., ANI, AAI, GGDC) with clear guidance on how seriously those cut-off values to be considered.
2. A clear genetic basis to define subspecies, i.e., presence of specific gene clusters that may be linked to phenotypic variations.

Other suggestions:

1. SeqCode may interact with phylocode for circumscription of (higher) taxa that should be defined based on monophyly.
2. Gaps in pairwise genomic matrices warrant the need of a new rank between genus and species, **subgenus** .