Digital sequence information on genetic resources: Concept, scope and current use.

Marcel Jaspars & Wael Houssen, University of Aberdeen, Scotland, UK

Rodrigo Sara, Consultant to the Secretariat of the Convention on Biological Diversity

Outline

- Key issues to be addressed
- Clarifying concept & scope
  - AHTEG list
  - Flow of data and information from a genetic resource
  - Options for DSI subject matter and proposed groupings
  - Implications of the proposed groupings on life sciences sectors
  - Additional challenges
- Clarifications

Key Issues to be Addressed

- Distinguishing between different types of data/information: extent of biological processing and proximity to the underlying genetic resource
- How can we describe how far DSI could/may extend - DNA, RNA, protein sequences and metabolites…… metadata, traditional knowledge?
- Mapping existing terminology to replace DSI against the different options proposed

Clarifying Concept & Scope:

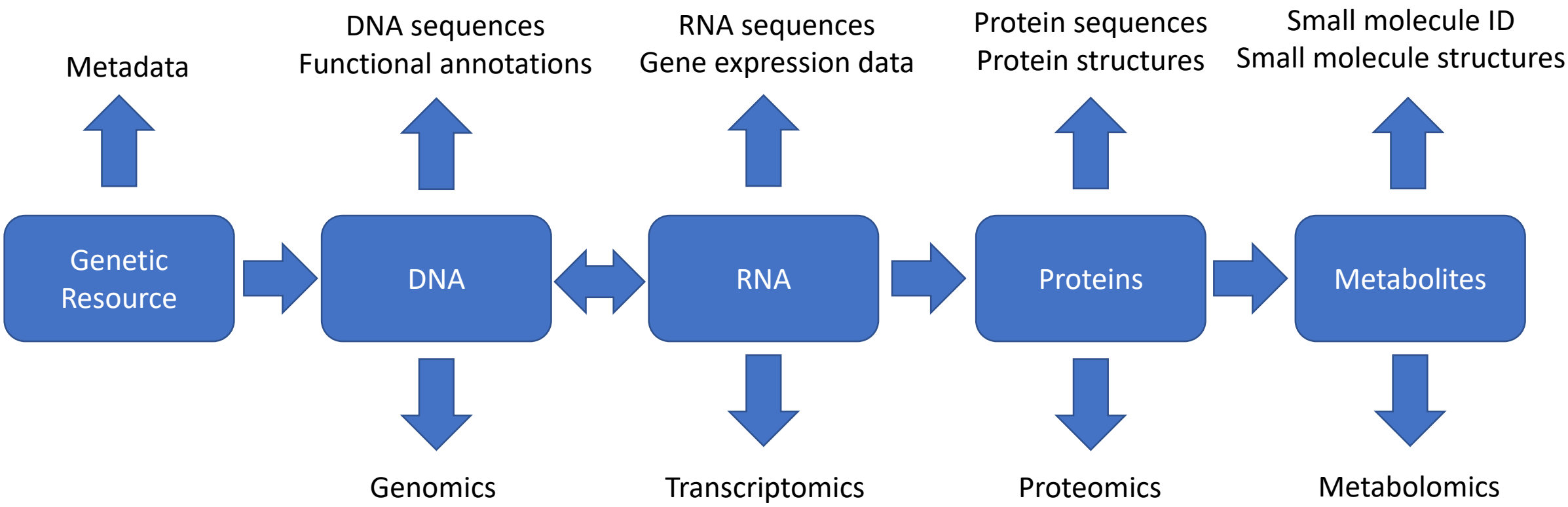
The AHTEG List

Molecular Data

a) The nucleic acid sequence reads and the associated data
b) Information on the sequence assembly, its annotation and genetic mapping. This information may describe whole genomes, individual genes or fragments thereof, barcodes, organelle genomes or single nucleotide polymorphisms.
c) Information on gene expression
d) Data on macromolecules and cellular metabolites
e) Information on ecological relationships, and abiotic factors of the environment
f) Function, such as behavioral data
g) Structure, including morphological data and phenotype
h) Information related to taxonomy
i) Modalities of use

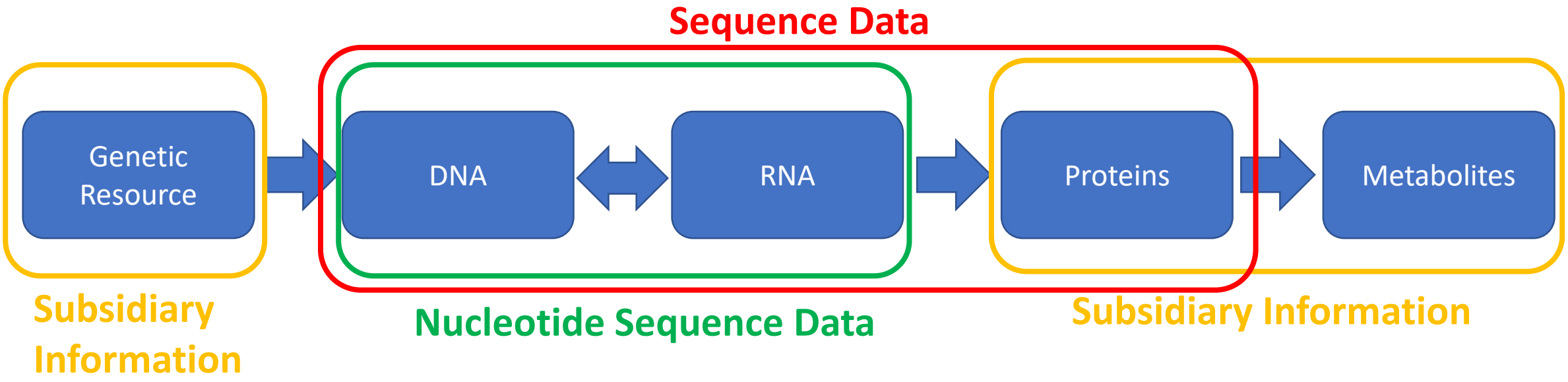
Clarifying Concept & Scope:

Flow of materials and information



Clarifying Concept & Scope:

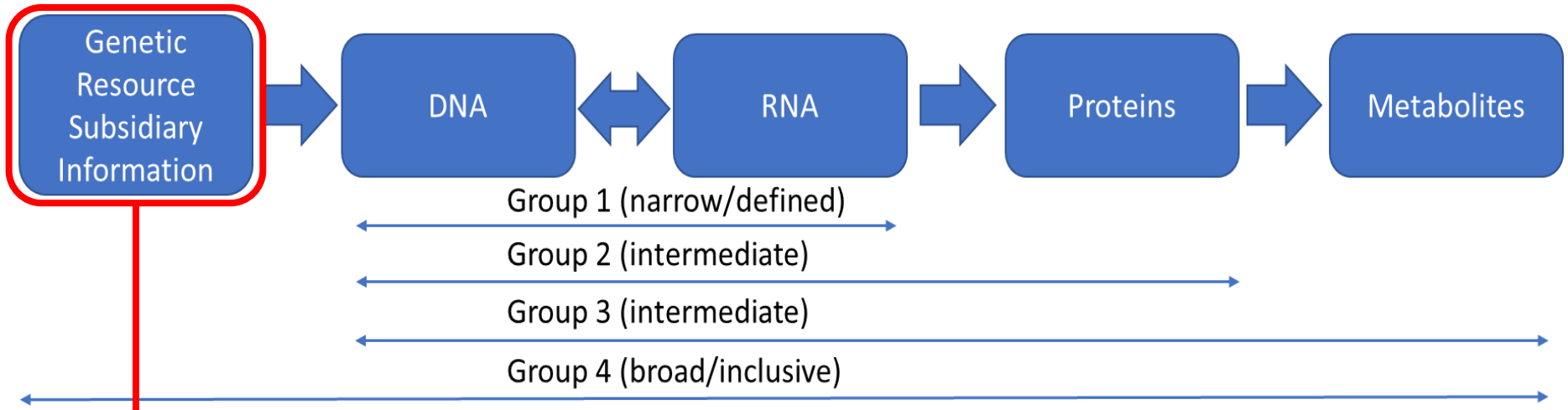
Flow of materials and information



Proximity to genetic resource + extent of processing can be used to propose subject matter groupings

Clarifying Concept & Scope:

DSI Subject matter groupings

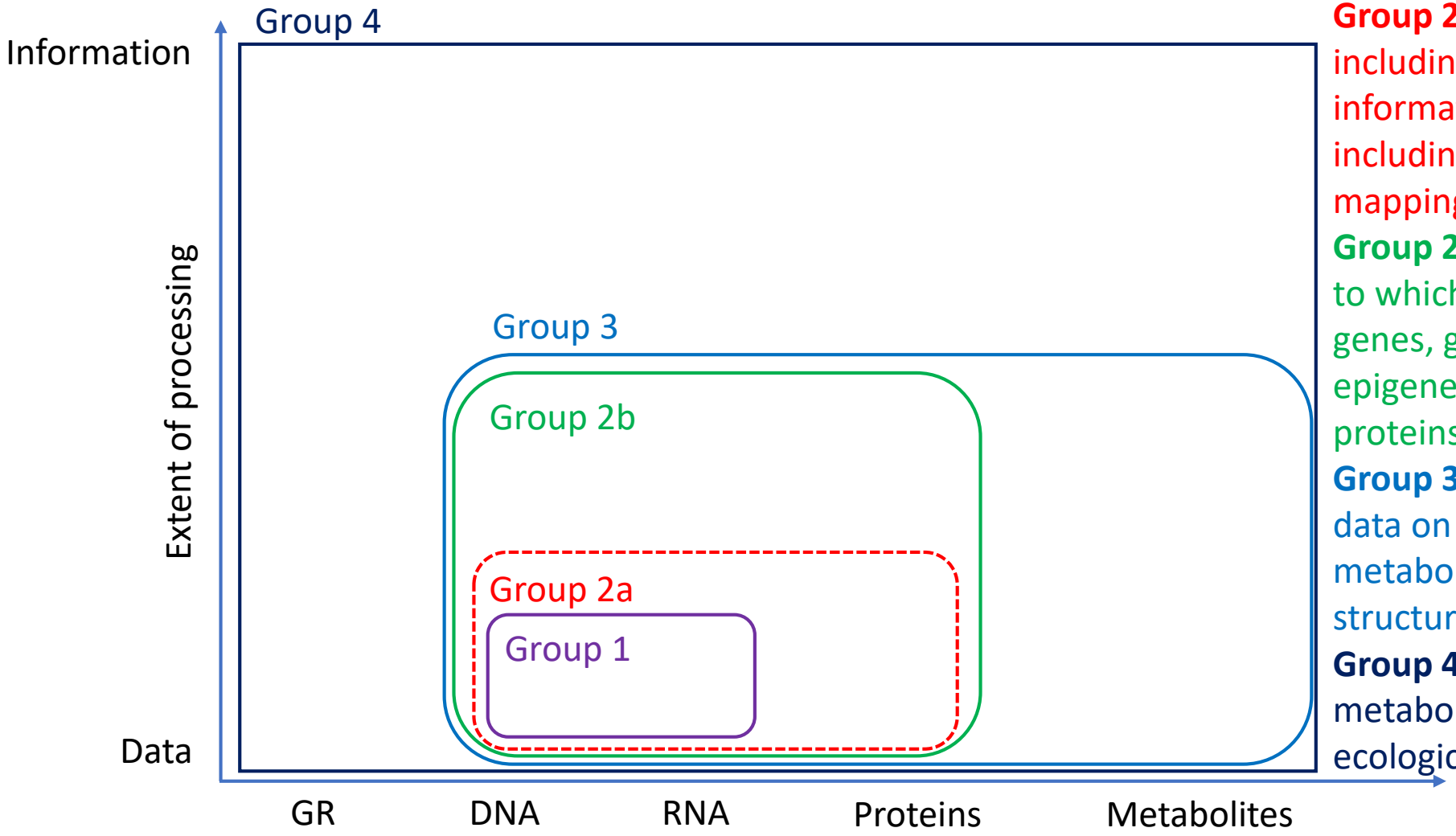


Group 1 – DNA and RNA sequence data
Group 2 – DNA and RNA sequence data + data/information concerning proteins
Group 3 – DNA and RNA sequence data + data/information concerning proteins + metabolites
Group 4 – DNA and RNA sequence data + data/information concerning proteins + metabolites + other data/information

Separate from other subsidiary information
(traditional knowledge, ecological interactions etc.)

Clarifying Concept & Scope:

Granular Options for Subject Matter Groupings



Group 1 - Narrow: DNA and RNA

Group 2a includes DNA/RNA sequence data including non-coding sequences, and information on the sequence assembly, including structural annotation and genetic mapping, as well as protein sequence data.

Group 2b is the same as group 2a in addition to which it includes functional annotation of genes, gene expression information, epigenetic data, and molecular structures of proteins.

Group 3 is the same as group 2b, but adds data on other macromolecules and metabolites, including their molecular structures

Group 4 – Broad: DNA, RNA, protein, metabolites + traditional knowledge, ecological interactions, etc.

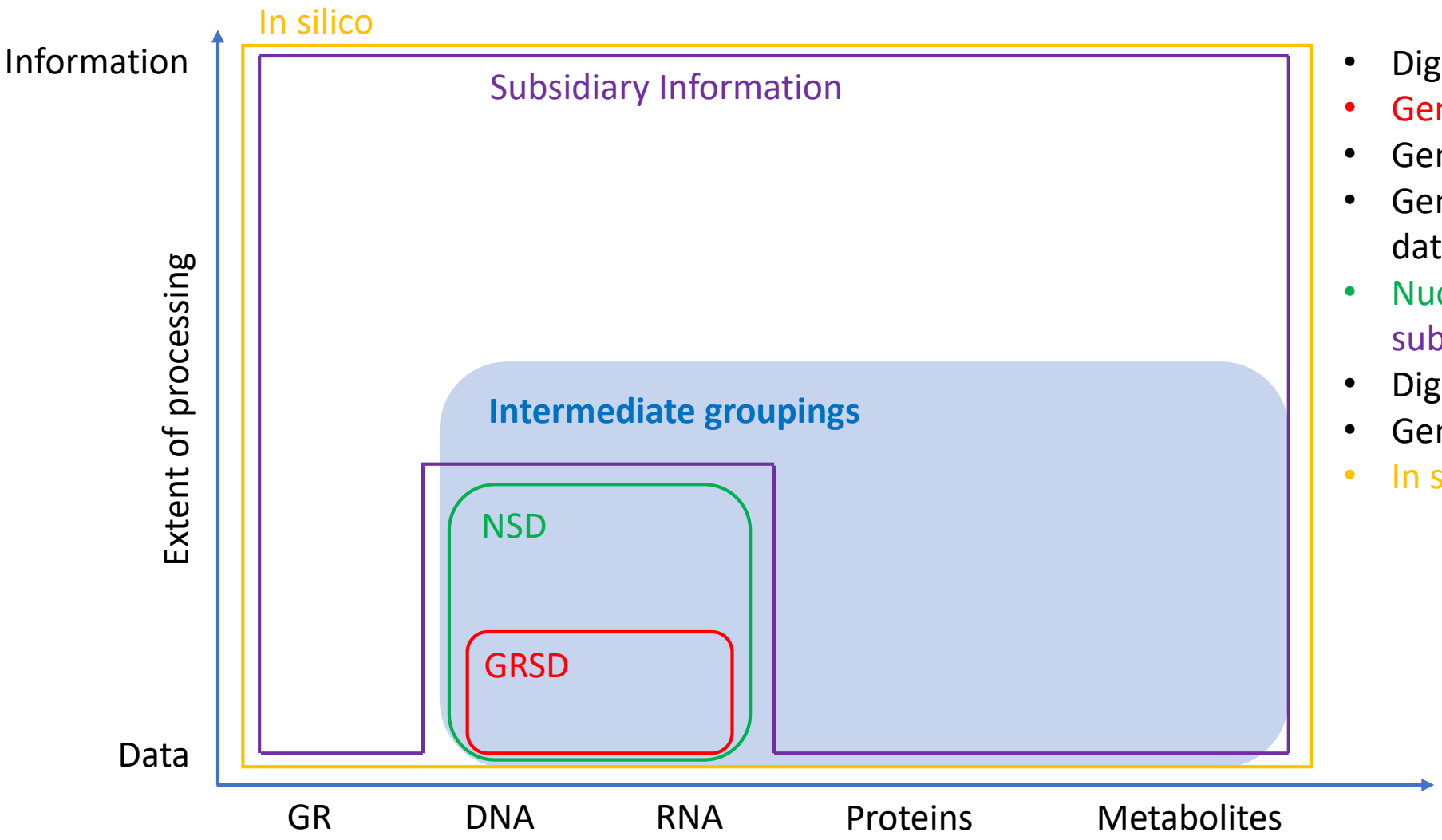
Clarifying Concept & Scope:

Scope of existing terminologies

		Narrow/Defined (Group 1)					Intermediate (Groups 2 & 3)			Broad/Inclusive (Group 4)			
AHTEG Category	**Component**	**DSD**	**GRSD**	**GS**	**GSD GSI**	**NSD**	**2a**	**2b**	**3**	**In silico**	**DGR**	**GI**	**SI**
a1	Nucleic acid sequence reads	+	+	+	+	+	+	+	+	+	+	+	
a2	Associated data to nucleic acid reads (technical aspects of sequencing experiments: the sequencing libraries, preparation techniques and data files).		+	+	+	+	+	+	+	+	+	+	
b1	Information on the sequence assembly, including structural annotation and genetic mapping. (This information may describe whole genomes, individual genes or fragments thereof, barcodes, organelle genomes or single nucleotide polymorphisms).			+	+	+	+	+	+	+	+	+	
b2	Non-coding nucleic acid sequences		+	?	?	+	+	+	+	+	+	+	
b3	Functional annotation of genes					?		+	+	+	?	+	
c1	Information on gene expression							+	+	+	?	+	+
c2	Epigenetic heritable elements (e.g. methylation patterns).							+	+	+	?	+	+
d1	Amino-acid sequence of proteins produced by gene expression.						+	+	+	+	?	+	+
d2	Molecular structures of proteins.							+	+	+	?	+	+
d3	Data on other macromolecules (not DNA, RNA or proteins) and cellular metabolites. (Molecular structures).								+	+	?	+	+
e	Information on ecological relationships, and abiotic factors of the environment.									+	?	+	+
f	Function, such as behavioural data (this would include environmental influences).									+	?	+	+
g	Structure, including morphological data and phenotype (this would include environmental influences).									+	?	+	+
h	Information related to taxonomy.									+	?	+	+
i	Modalities of use.									+	?	+	+
	Additional undefined elements.									+	?	+	+

Clarifying Concept & Scope:

Terminology Used for DSI



- Digital sequence data
- Genetic resource sequence data
- Genetic sequences
- Genetic sequence data/information
- Nucleotide sequence data & subsidiary information
- Digital genetic resources
- Genetic information
- In silico

Clarifying Concept & Scope:

Applying the proposed DSI subject matter groupings to different life-sciences sectors.

Grouping	Taxonomy & Conservation	Agriculture & Food Security	Industrial & Synthetic Biology	Healthcare & Pharmaceuticals

Critical

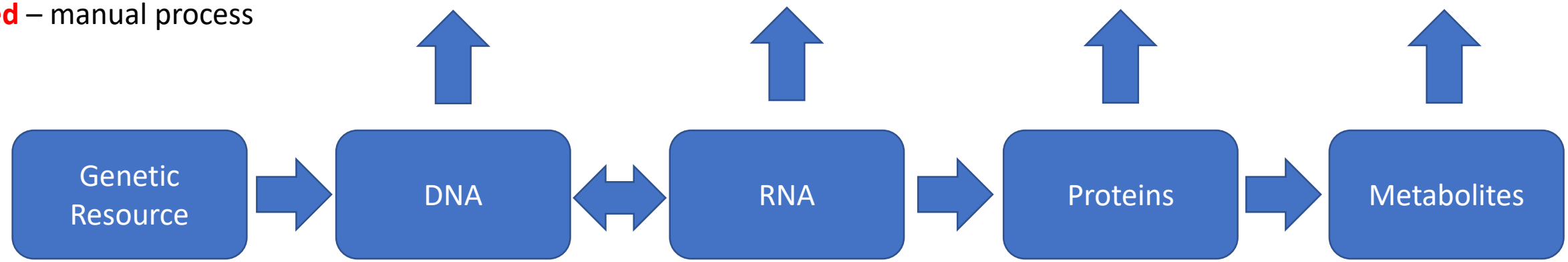
Important

Useful

Clarifications

How the subject matter groupings may accommodate scientific advances

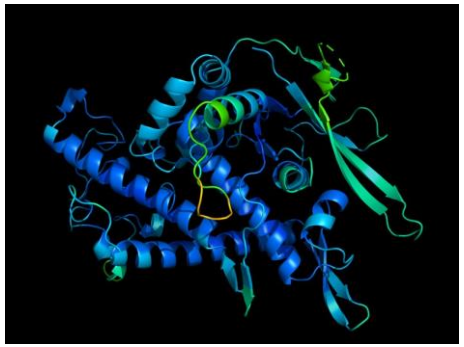
Green – automatic processing
Orange – partially automated
Red – manual process



NEWS · 30 NOVEMBER 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

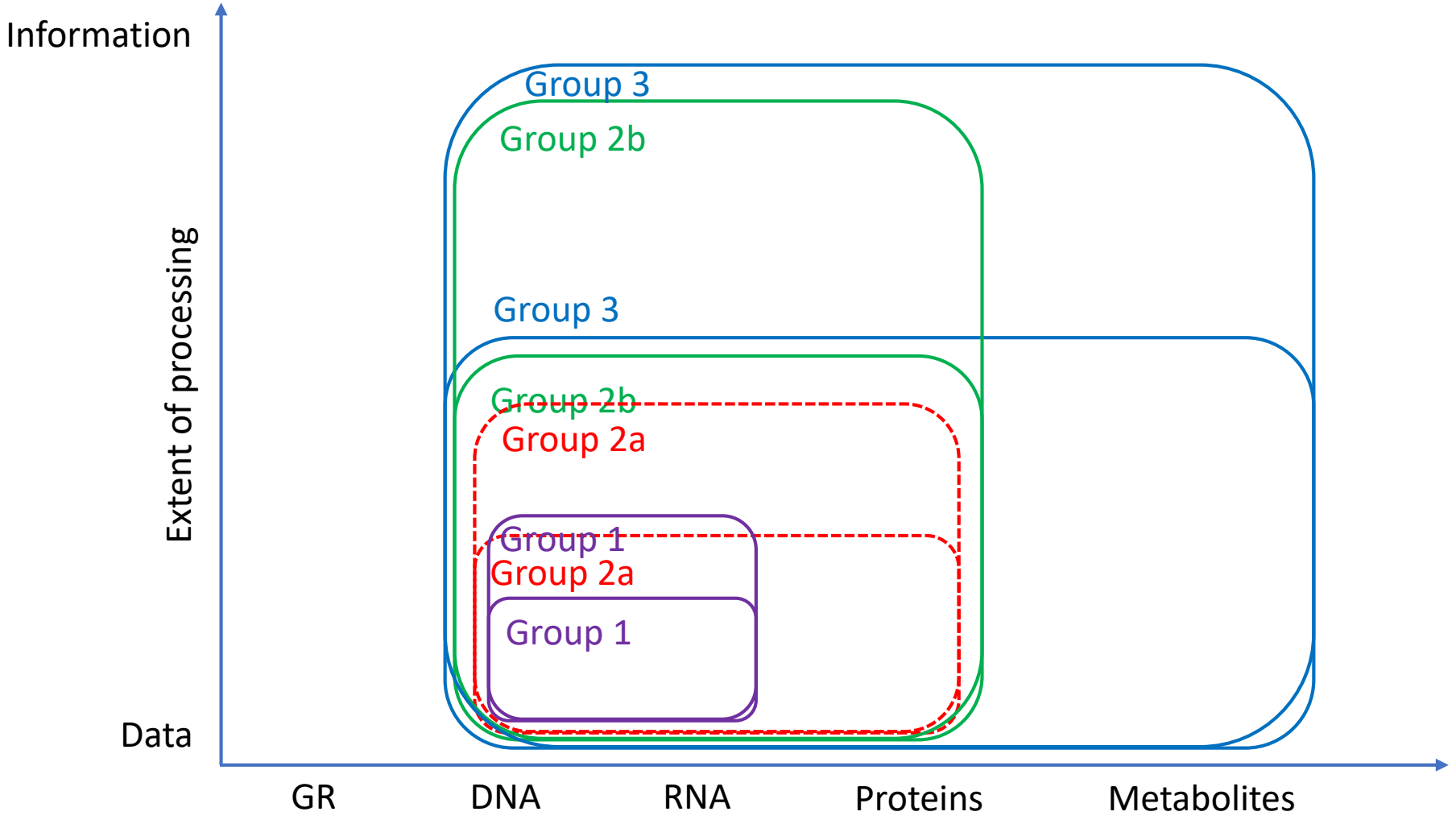
Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.



3D Protein structures

https://www.nature.com/articles/d41586-020-03348-4

How the subject matter groupings may accommodate scientific advances



What length of sequence can still be considered as a ‘sequence’?

Theoretical probabilities of randomly having two identical sequences of the same length within a given set of sequences

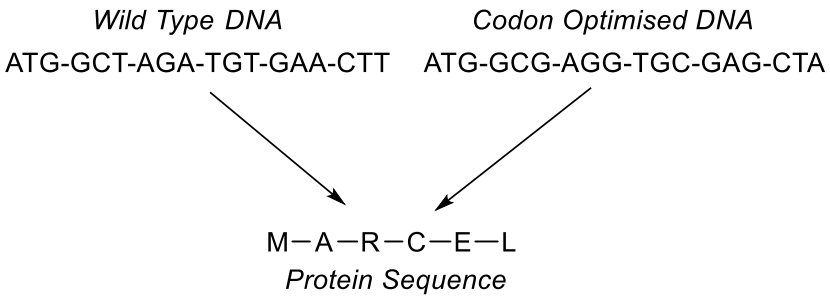
Data set	10 bp sequence	20 bp sequence	25 bp sequence	30 bp sequence
Human Genome (3×10^9 bp)	**100%**	**0.3%**	**~0%**	**~0%**
GenBank (1.65×10^{12} bp)	**100%**	**77.7%**	**0.15%**	**~0%**
GenBank x 10	**100%**	**99.9%**	**1.5%**	**~0%**
GenBank x 100	**100%**	**100%**	**13.6%**	**~0%**

Not all sequence variation is governed only by random factors, but it is governed by selection that could lead to convergence for some DNA sequences, meaning that the same sequence, longer than 30 residues, could occur in multiple species.

https://www.cbd.int/dsi-gr/2019-2020/studies/#tab=1

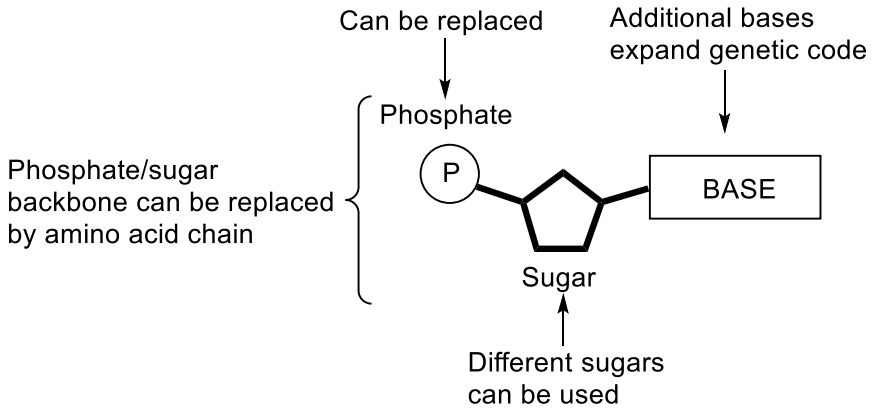
What should be included under DSI?

a.) Codon Optimisation

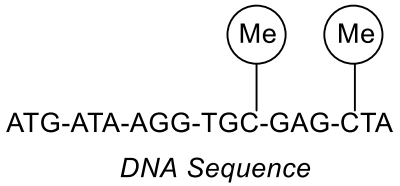


Different DNA sequences be translated to give the same protein as there is degeneracy in the genetic code.

b.) Nucleotide Modifications



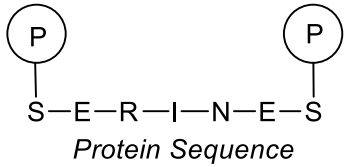
c.) Epigenetic modifications



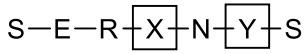
Nucleotide bases can be methylated to allow heritable changes to be made without changing the DNA sequence

e.) Non-coding DNA

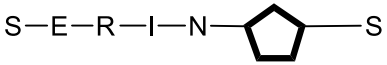
d.) Protein Modifications



Protein sequences can be phosporylated through post-translational modifications



Non-canonical amino acids can be included (non-ribosomal peptide synthesis)



Post translational modifications can occur within the peptide chain. (RiPP metabolites)

Clarifying Concept & Scope: Additional challenges

Key challenges if options/groups not adopted

Q: How far along the flow from genetic resource onwards to DNA, RNA, protein sequences and metabolites DSI can be considered to extend. Specifically: whether macromolecules (e.g. proteins, polysaccharides) are included under DSI and whether small molecules (metabolites) are included under DSI

A: This can be resolved by utilizing the four groups proposed to clarify the scope of DSI subject matter, in which case all macromolecules (non DNA/RNA) and metabolites would be excluded under Group 1 or 2, whereas they would be included under Groups 3 or 4.

Clarifying Concept & Scope: Additional challenges

Key challenges if options/groups not adopted

Q: The distinction between data and information and how this is stored and processed, including the extent to which data has been processed before it can be considered information

A: This can be resolved by utilizing the four groups proposed to clarify the scope of DSI subject matter as these have clear subject matter boundaries and so an approach, criteria or definition for distinguishing between data and information is not necessary.

Clarifying Concept & Scope: Additional challenges

Length of sequence & non-coding DNA

Q: What length of sequence can still be considered as a ‘sequence’

A: Sequences below 30 nucleotides may not be unique and so this may provide a logical threshold below which information should be excluded from DSI subject matter.

Q: Whether non-coding DNA should be included under ‘DSI’

A: Genetic elements which do not encode proteins (such as promotors) may have a natural functional role in transcription, translation or biosynthesis and on this basis it may be considered an inherent part of the underlying genetic resource, such that it would be illogical to distinguish between coding and non-coding sequences.

Clarifying Concept & Scope: Additional challenges

Epigenetic heritable factors

Q: Whether epigenetic heritable factors should be included under DSI.

A: Epigenetic heritable factors may have a natural functional role in transcription, translation or biosynthesis and therefore it may be logical to exclude it from DSI subject matter (assuming the rationale for non-coding DNA is also accepted).

Clarifying Concept & Scope: Additional challenges

Modifications to DNA, RNA (and proteins)

Q: Whether modified DNA, RNA (and proteins) should be included under DSI

A: Naturally modified DNA, RNA or proteins may nevertheless have a natural functional role in transcription, translation or biosynthesis and on this basis these may be considered an inherent part of the underlying genetic resource such that it would be illogical to exclude from DSI subject matter, at least to the same extent that DSI subject matter includes DNA, RNA and/or proteins. Conversely synthetically modified DNA, RNA or proteins cannot be said to have a natural functional role and so on this basis could be considered not to be an inherent part of the underlying genetic resource.

Clarifying Concept & Scope: Additional challenges

Further Considerations

- What is the threshold level for natural variation of DNA, RNA (and protein) sequences – what is substantially similar or identical?
- E.g. Peer review comments by Canada
  - Are sequences ‘forensic’ to identify origin needed for traceability?
  - How can we distinguish between engineered sequences and those resulting from natural variation?
  - To identify origin requires fuller sequence comparisons across multiple regions of genome.
  - For microorganisms sequencing more than one colony may be necessary.
- Is such sequence variation better discussed on a case-by-case basis as it is under the IP system?