How is Digital Sequence Information (DSI) produced ?

Dr Wael Houssen
EPSRC Fellow
Institute of Medical Sciences,
University of Aberdeen.
https://www.abdn.ac.uk/people/w.houssen

UNEP

CBD

Convention on Biological Diversity

Distr.
GENERAL

CBD/DSI/AHTEG/2020/1/3
29 January 2020

ENGLISH ONLY

AD HOC TECHNICAL EXPERT GROUP ON
DIGITAL SEQUENCE INFORMATION ON
GENETIC RESOURCES
Montreal, Canada, 17-20 March 2020

Annex

Digital Sequence Information on Genetic Resources: Concept, Scope and Current Use

Wael Houssen, Rodrigo Sara, Marcel Jaspars

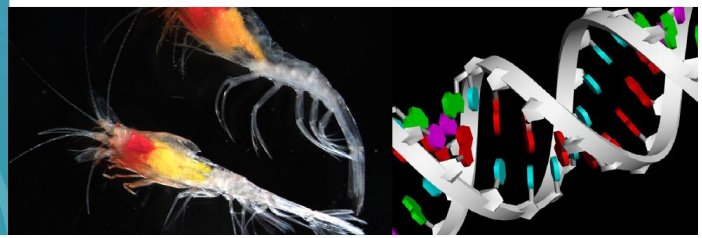
Links:

- https://www.cbd.int/doc/c/fef9/2f90/70f037ccc5da885dfb293e88/dsi-ahteg-2020-01-03-en.pdf
- https://www.dosi-project.org/wp-content/uploads/070-DSI-Policy-brief-V4-WEB.pdf

POLICYBRIEF
MARCH 2020

DOSI
DEEP-OCEAN STEWARDSHIP INITIATIVE

Digital Sequence Information – Clarifying Concepts



This Policy Brief is Based on:

"Digital Sequence Information on Genetic Resources: Concept, Scope and Current Use". Authors:

Wael Houssen, Rodrigo Sara, Marcel Jaspars

The Central Dogma of Molecular Biology

DNA

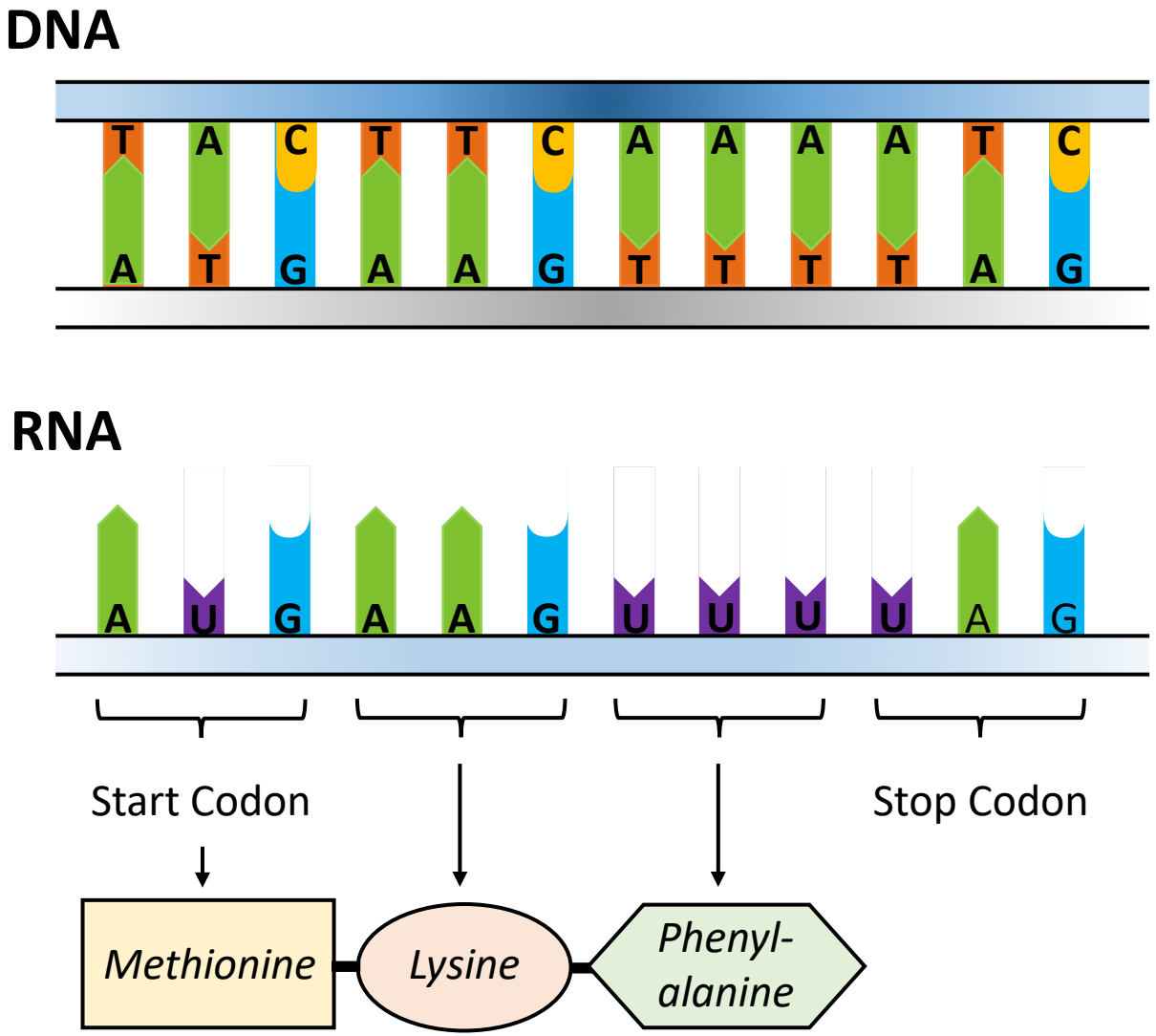
Transcription

RNA

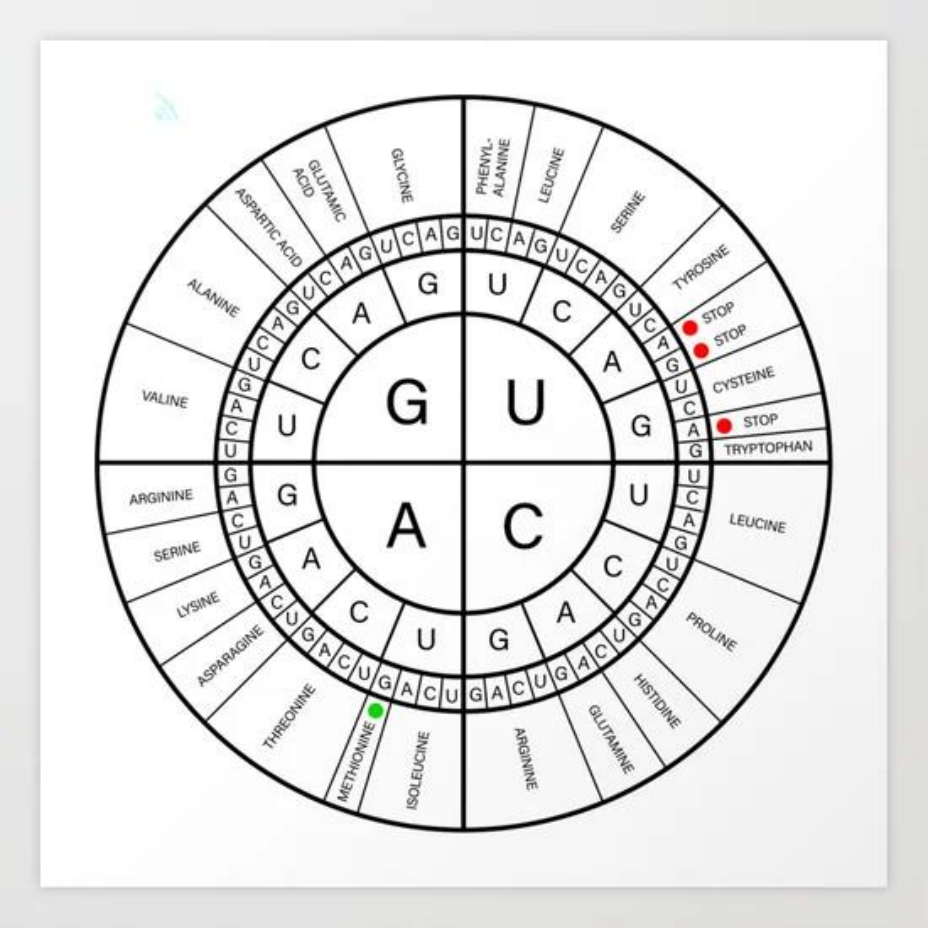
Translation

Protein

Metabolites



The Codon wheel



- Sixty-four codons encode for the twenty natural amino acids.
- 'AUG' encodes the amino acid methionine, and activates the ribosome to start the process of making a protein and is thus known as "start codon".
- "Stop codons" signal the termination of translation into proteins.
- Genetic code is degenerate is that many amino acids can be encoded by different codons.
- Each organism has a 'preference' for the use of a particular codon for a specific amino acid. This is called "Codon bias".

Origin of Digital Sequence Information

Applications

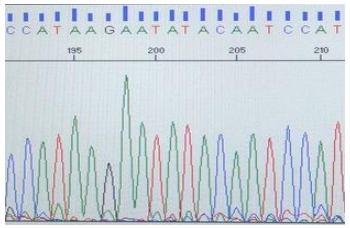
Metadata

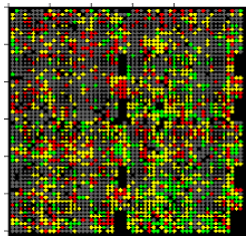
Functional annotations

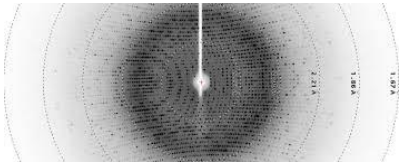
Gene expression profiling

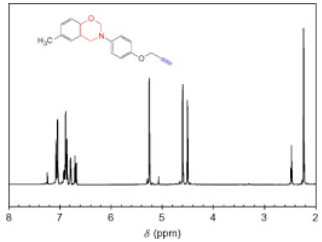
Protein structural and functional characterisation

Metabolite structure and bioactivity









Genomics

Transcriptomics

Proteomics

Metabolomics

Genetic Resource → DNA ↔ RNA → Protein → Metabolite

Raw sequence vs. value-added DSI

DNA sequence alone does not provide information on:

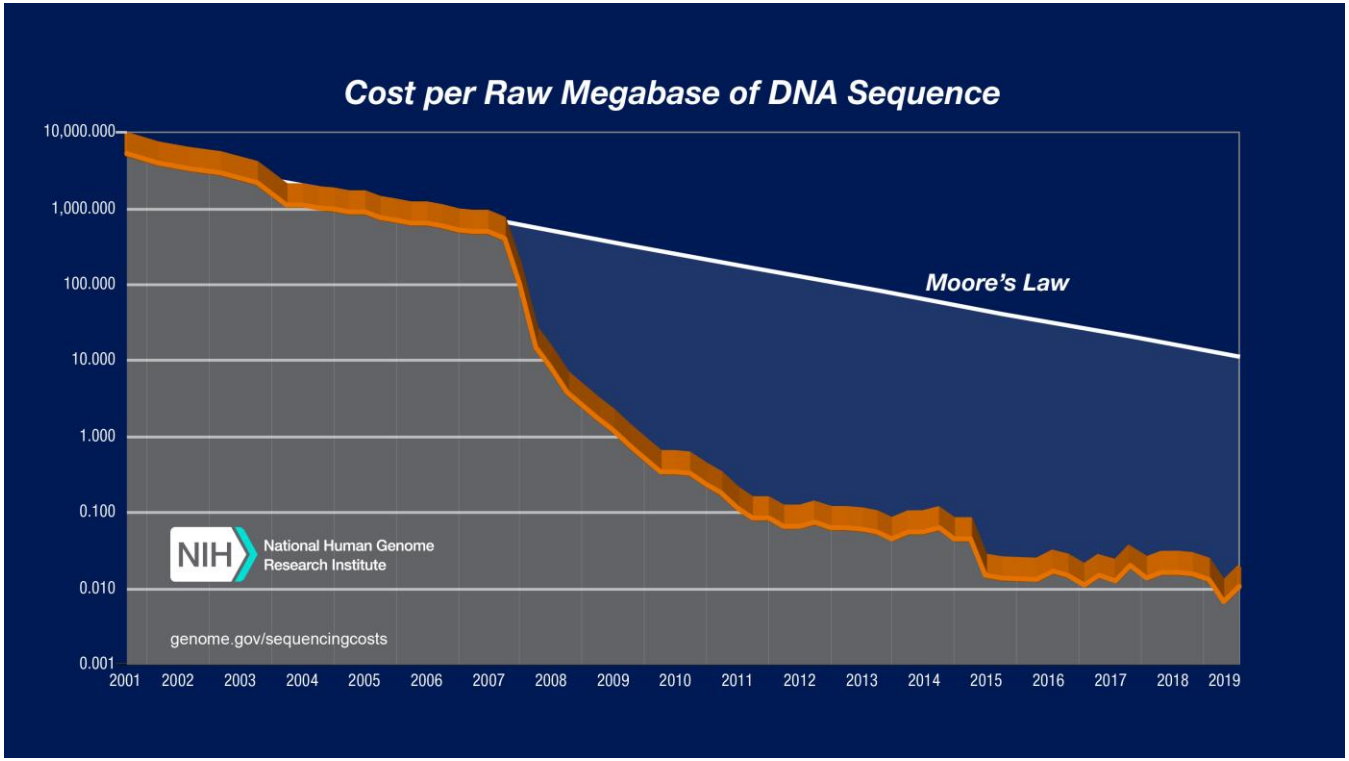
- The function of each gene.
- The level of expression of each gene.
- How the encoded protein sequence folds. Misfolded proteins are known to underlie serious diseases.
- The posttranslational modifications in proteins e.g. protein glycosylation.

Next Generation Sequencing Technologies

Platform / Instrument	Throughput range (Gb)	Read length (bp)	Strength	Weakness
Sanger sequencing				
ABI 3500/3730	0.0003	Up to 1kb	Read accuracy and length	Cost and throughput
Illumina				
MiniSeq	1.7–7.5	1×75 to 2×150	Low initial investment	Run and read length
MiSeq	0.3–15	1×36 to 2×300	Read length, scalability	Run length
NextSeq	10–120	1×75 to 2×150	Throughput	Run and read length
HiSeq (2500)	10–1000	1×50 to 2×250	Read accuracy, throughput, low per sample cost	High initial investment, run length
NovaSeq 5000/6000	2000–6000	2×50 to 2×150	Read accuracy, throughput Low per sample cost	High initial investment, run and read length
Ion Torrent				
PGM	0.08–2	Up to 400	Read length, speed	Throughput, homopolymers
S5	0.6–15	Up to 400	Read length, speed, scalability	Homopolymers
Proton	10–15	Up to 200	Speed, throughput	Homopolymers
Pacific BioSciences				
PacBio RSII	0.5–1	Up to 60 kb (Average 10 kb, N50 20 kb)	Read length, speed	High error rate and initial investment, low throughput
Sequel	5–10	Up to 60 kb (Average 10 kb, N50 20 k)	Read length, speed	High error rate
Oxford Nanopore				
MInION	0.1–1	Up to 100 kb	Read length, portability	High error rate, run length, low throughput

Clin Microbiol Infect 24: 335

Sequencing costs



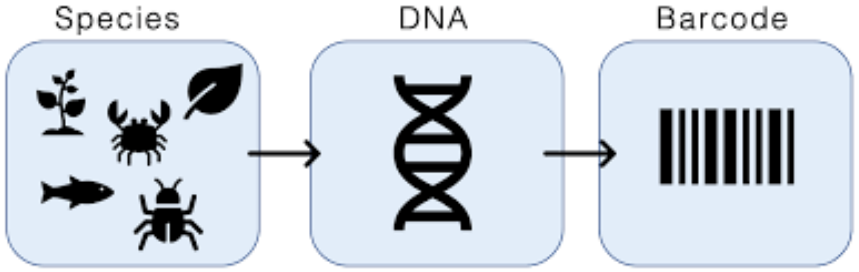
- Sequencing is getting **faster** – **cheaper**
- Number of sequences continues to increase.

What is being sequenced?

- Whole genome.
- Metagenome.
- 16S/18S/ITS/CO1 amplicon for molecular identification of microorganisms.
- Environmental DNA.

Examples of the applications of DSI

- Taxonomical identification of organisms: DNA barcoding



- Biodiversity conservation
- Agriculture and food security: Genetic markers to assist selective breeding; Development of GMO
- Drug discovery: Identification of new drug targets
- Medicine: Detection of infectious pathogens
- SynBio and biotechnology: Recombinant production of proteins e.g. insulin and enzymes for laundry detergents.