STORAGE DEVELOPER CONFERENCE

SDC 23

BY Developers FOR Developers

Pantheon DNA Data Storage CODEC

Experiences, Challenges, and Innovations

André Guilherme da Costa Martins, PhD Biomed. Sci.
Bioinformatics researcher
Institute for Technological Research - IPT, Brazil
andremartins@ipt.br

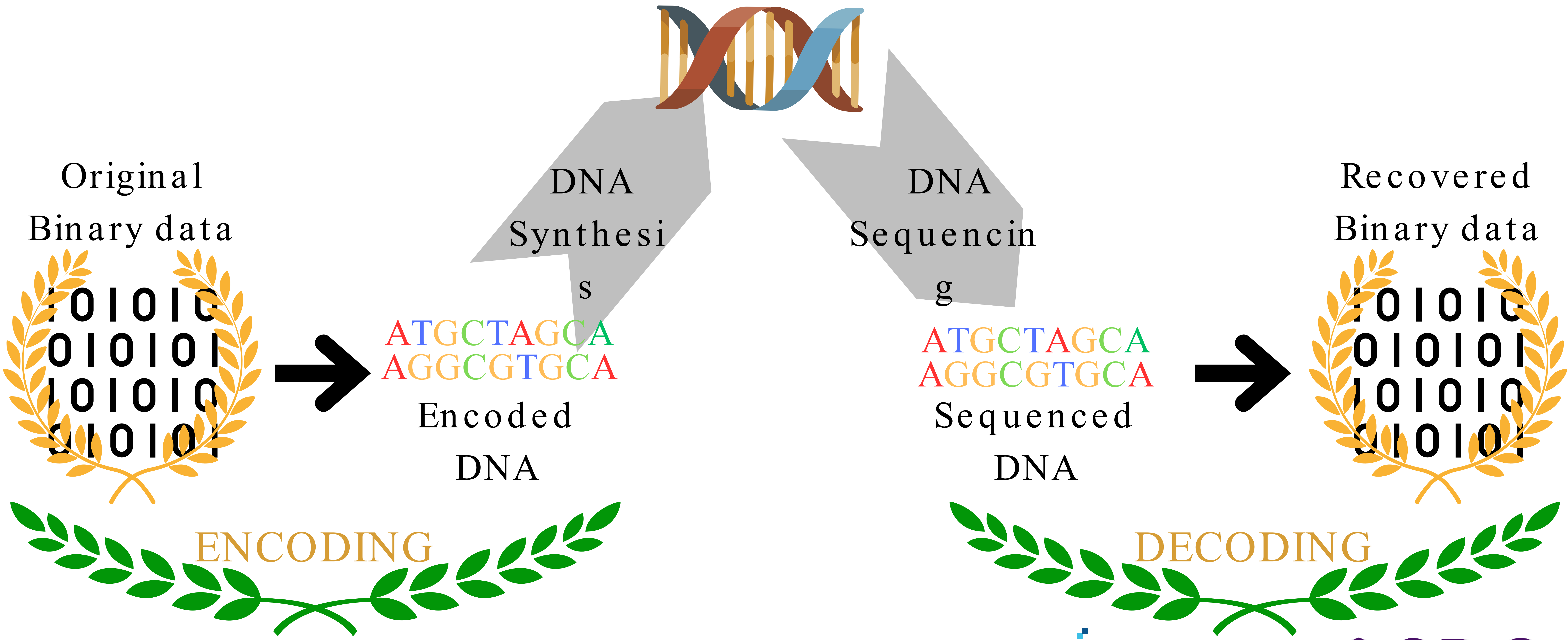
Who are we?

The Institute for Technological Research - IPT has been contributing actively for 124 years to science and technical advances.

We provide technical solutions for industry, governments, and society, enabling them to overcome the challenges of our time.

PROMETHEUS

Storaged DNA molecules



Original Binary data

ATGCTAGCA
AGGCGTGCA

Encoded DNA

DNA Synthesis

DNA Sequencing

ATGCTAGCA
AGGCGTGCA

Sequenced DNA

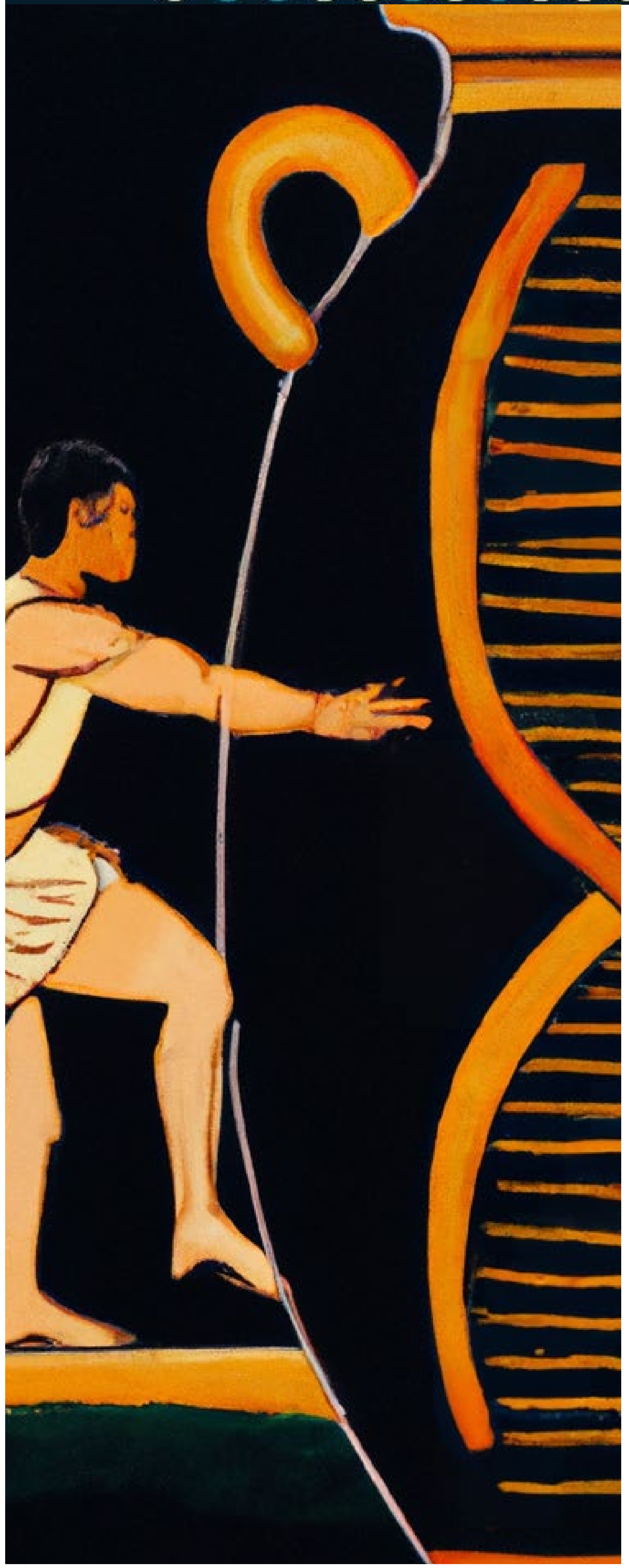
Recovered Binary data

ENCODING

DECODING

The Pantheon CODEC

- **A versatile CODEC:**
  - Robust DNA data architecture
    - Binary data pre-processing
    - Multiple choices for mapping algorithms
    - Multi-layer ECC strategy
  - Supports SNIA's sectors (S0 & S1)
  - Includes NGS processing algorithms
    - Compatible with multiple sequencing and storage strategies



PROMETHEUS

Apollo

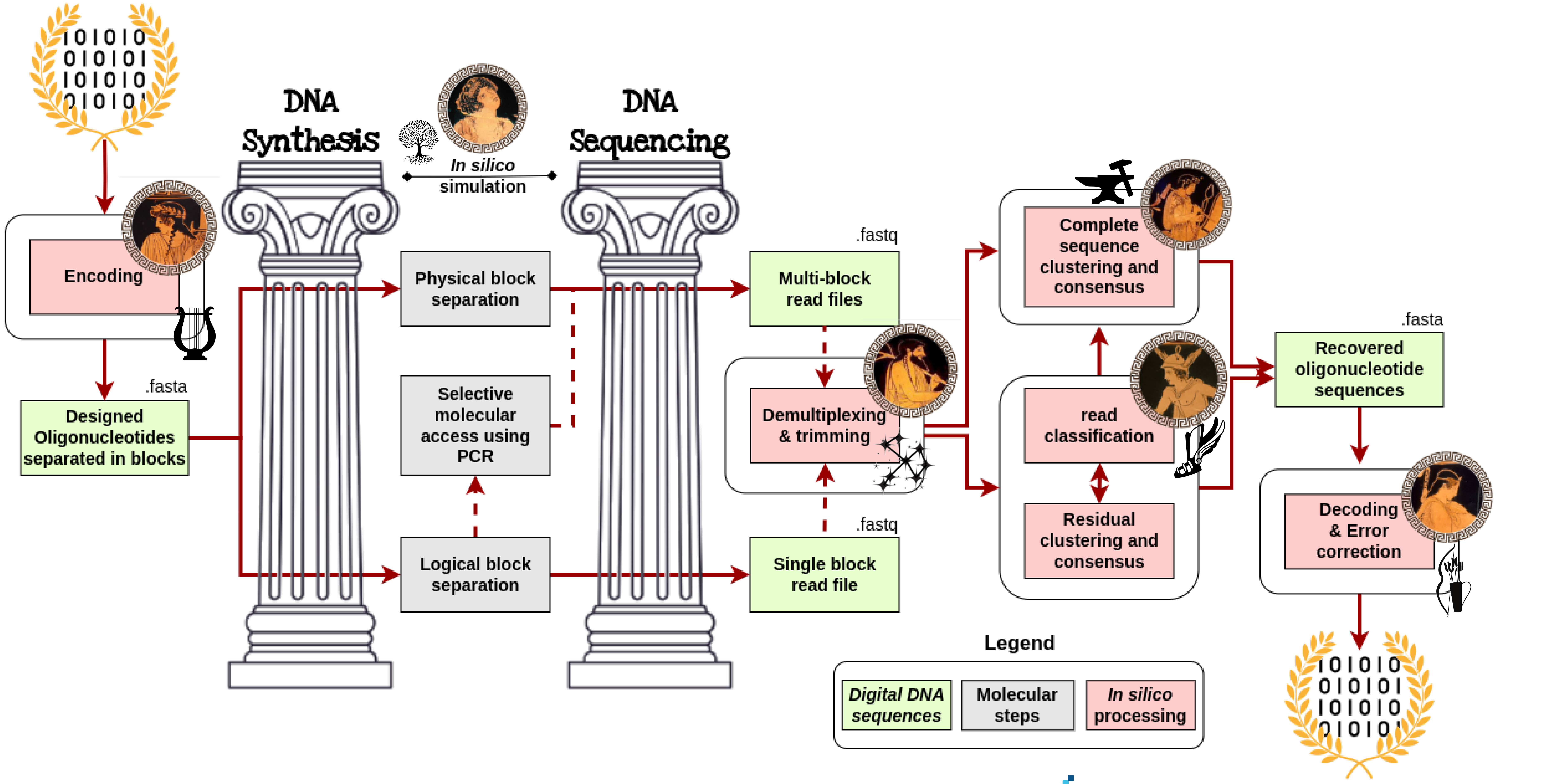
Artemis

Chiron

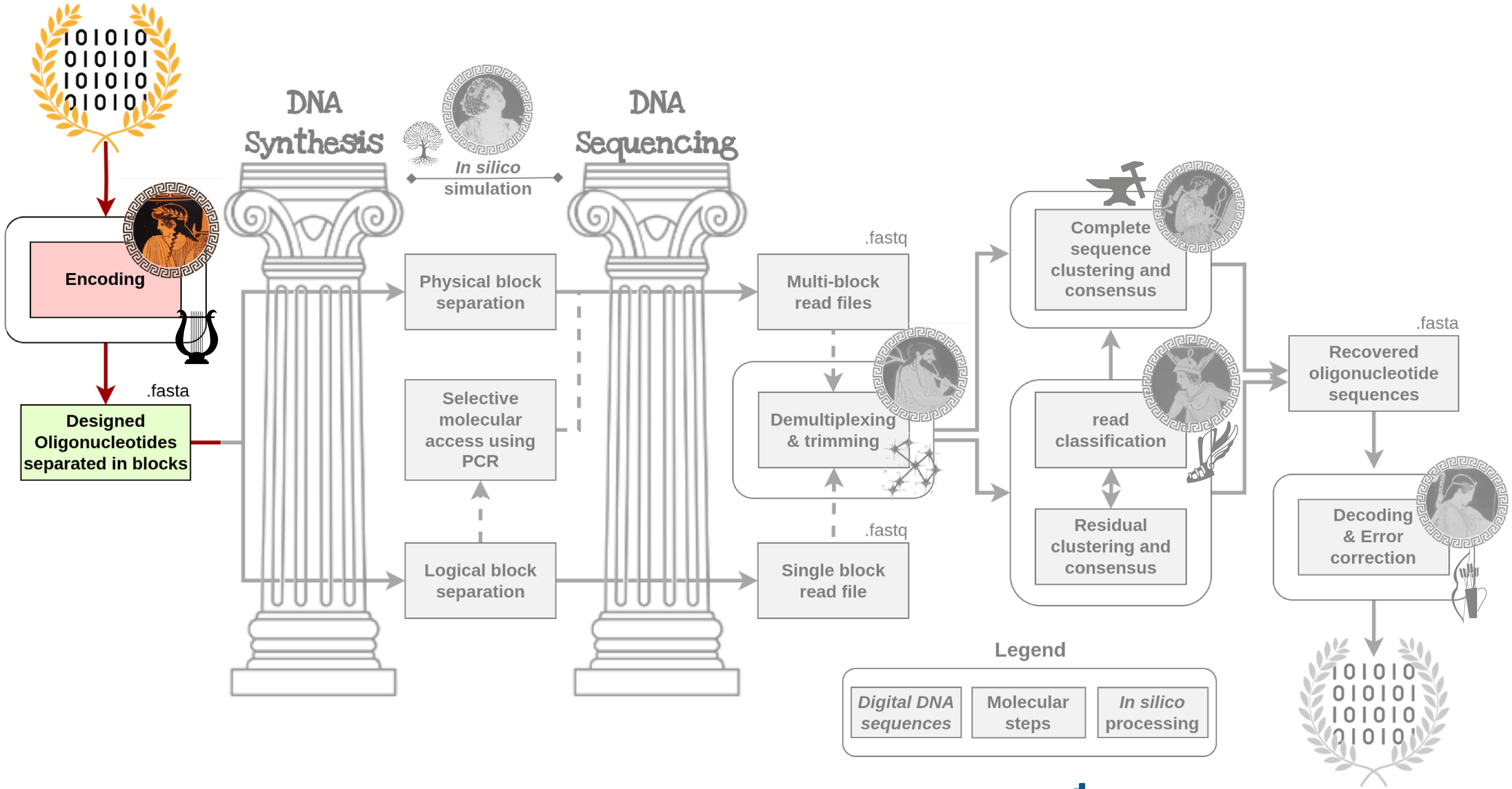
Hermes

Hephaestus

Gaia


DNA Synthesis
In silico simulation
DNA Sequencing
Encoding
.fasta
Designed Oligonucleotides separated in blocks
Physical block separation
Selective molecular access using PCR
Logical block separation
.fastq
Multi-block read files
Demultiplexing & trimming
.fastq
Single block read file
Complete sequence clustering and consensus
read classification
Residual clustering and consensus
.fasta
Recovered oligonucleotide sequences
Decoding & Error correction
Legend
Digital DNA sequences
Molecular steps
In silico processing

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS Lenovo SDC 23



Encoding

.fasta

Designed Oligonucleotides separated in blocks

DNA Synthesis

In silico simulation

DNA Sequencing

Physical block separation

Selective molecular access using PCR

Logical block separation

.fastq

Multi-block read files

Demultiplexing & trimming

.fastq

Single block read file

Complete sequence clustering and consensus

read classification

Residual clustering and consensus

.fasta

Recovered oligonucleotide sequences

Decoding & Error correction

Legend

Digital DNA sequences

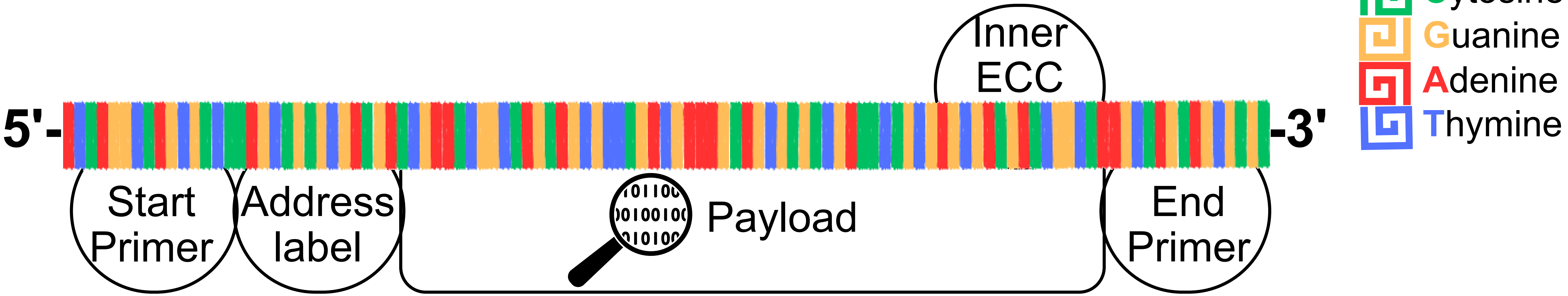
Molecular steps

In silico processing

Apollo, the encoding module

Apollo and the Muses - Michel Dorigny, early 1640s

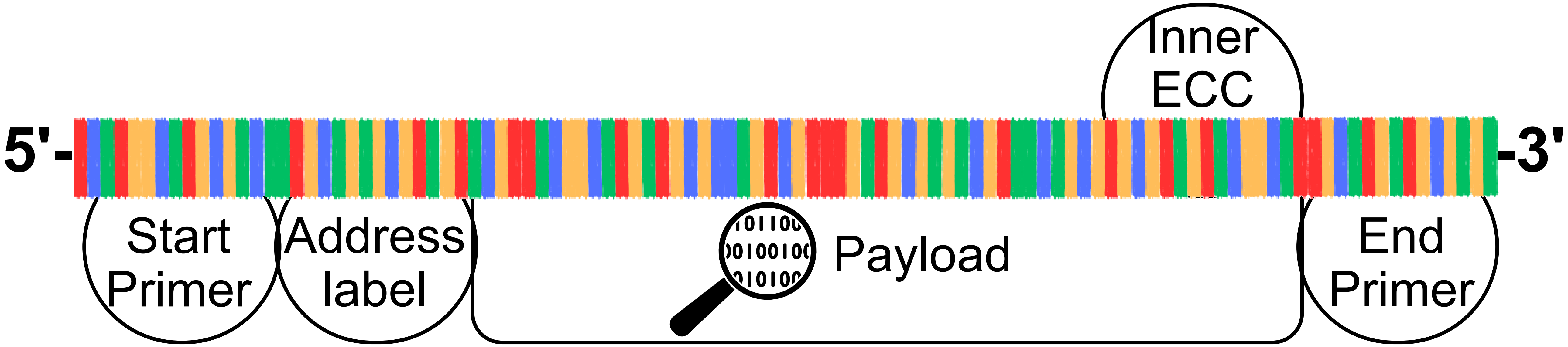
Apollo: oligonucleotide structure



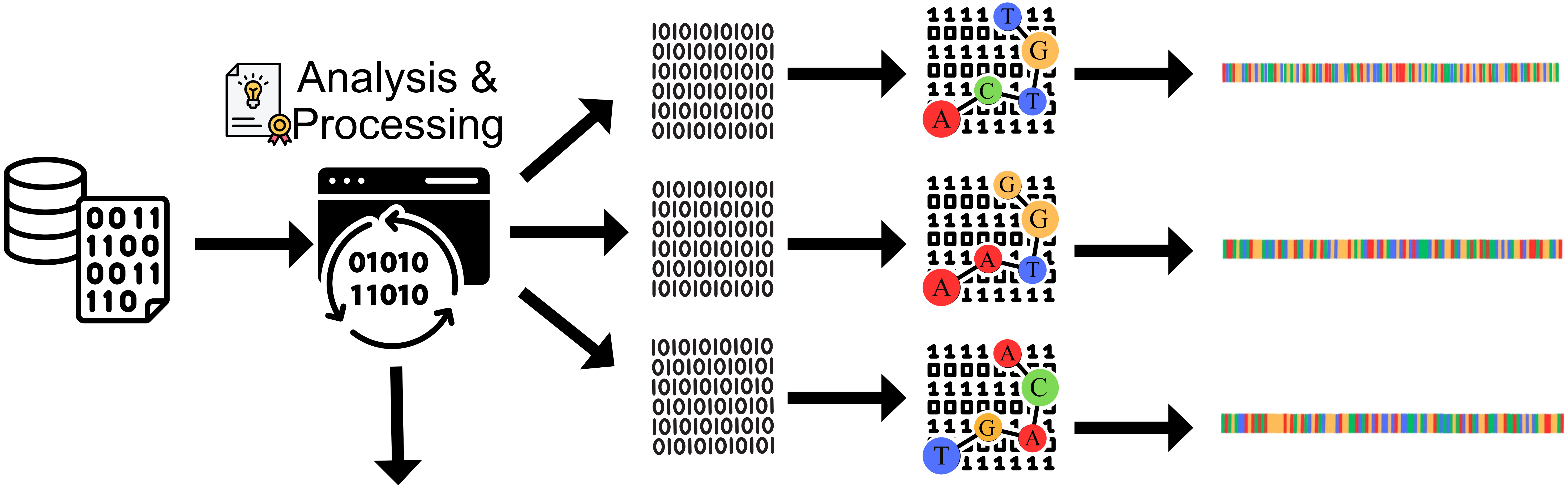
PROMETHEUS

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS Lenovo SDC 23

Apollo: encoding process



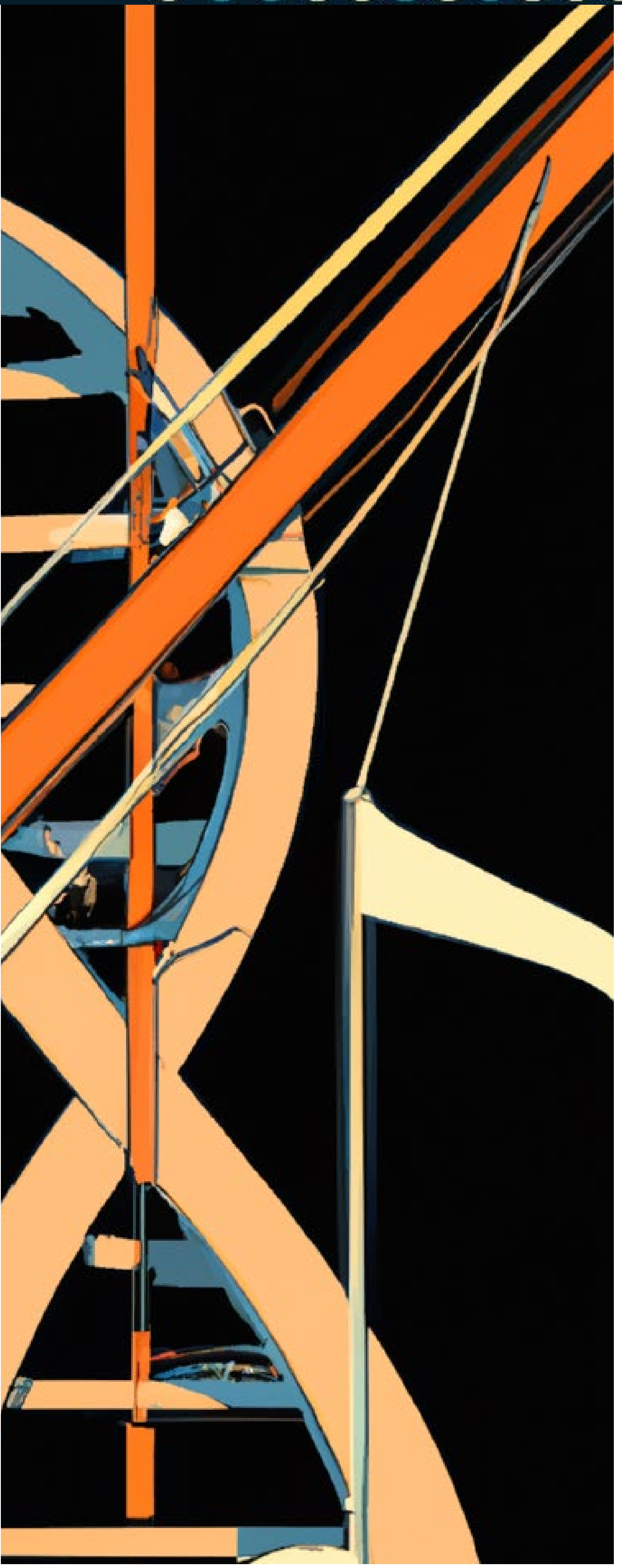
🟩 Cytosine
🟨 Guanine
🟥 Adenine
🟦 Thymine

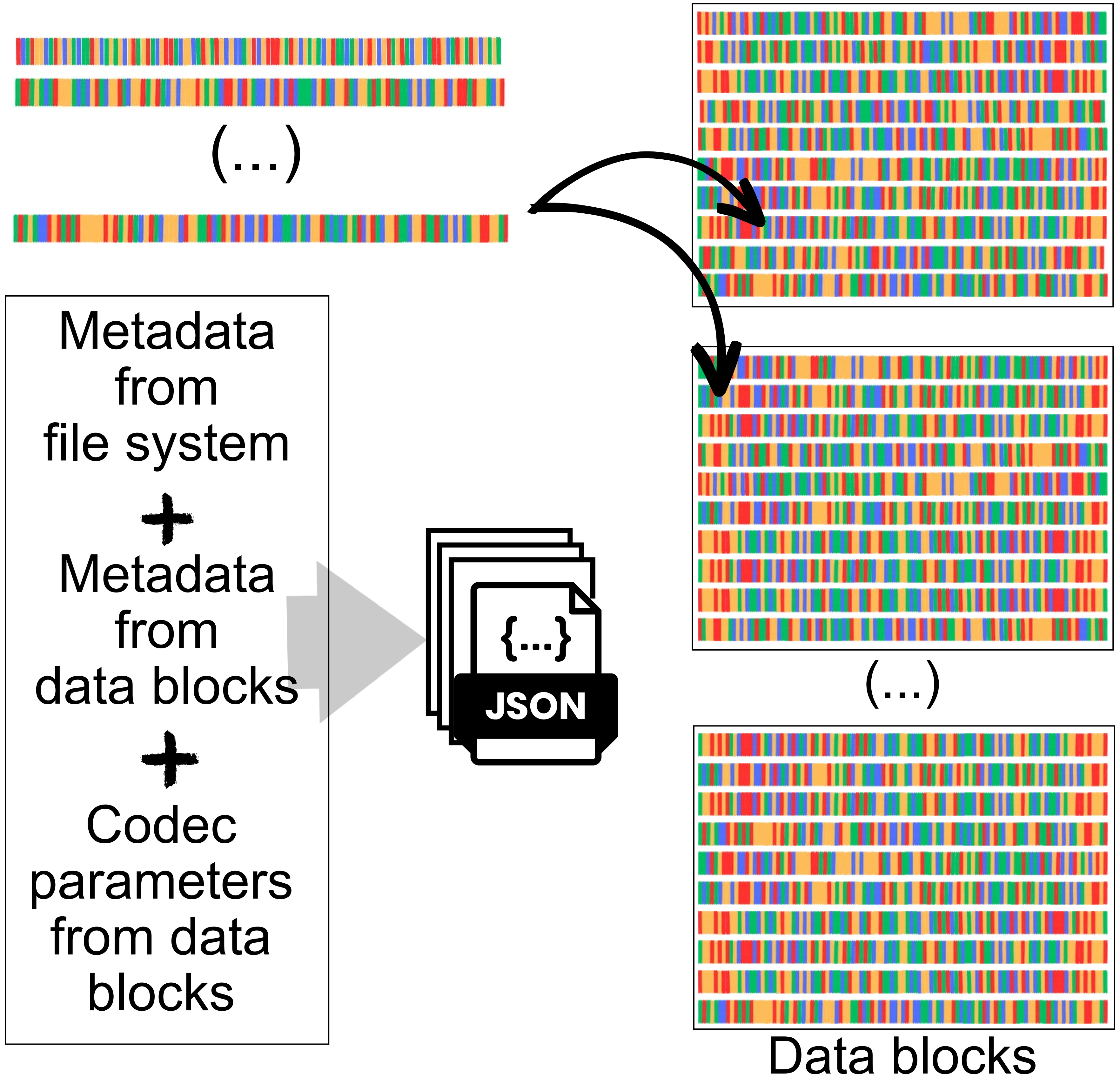


Analysis & Processing

Metadata from file system

Apollo: metadata & Outer ECC





Metadata from file system

+

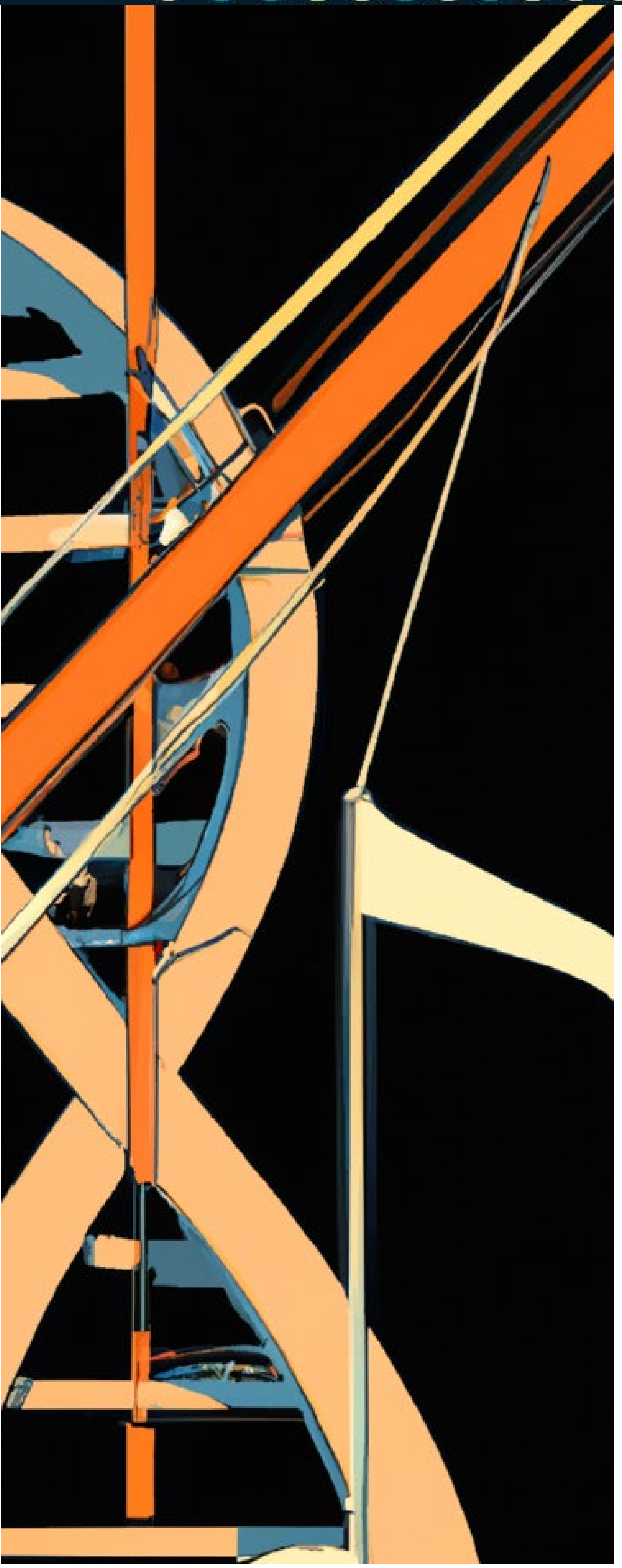
Metadata from data blocks

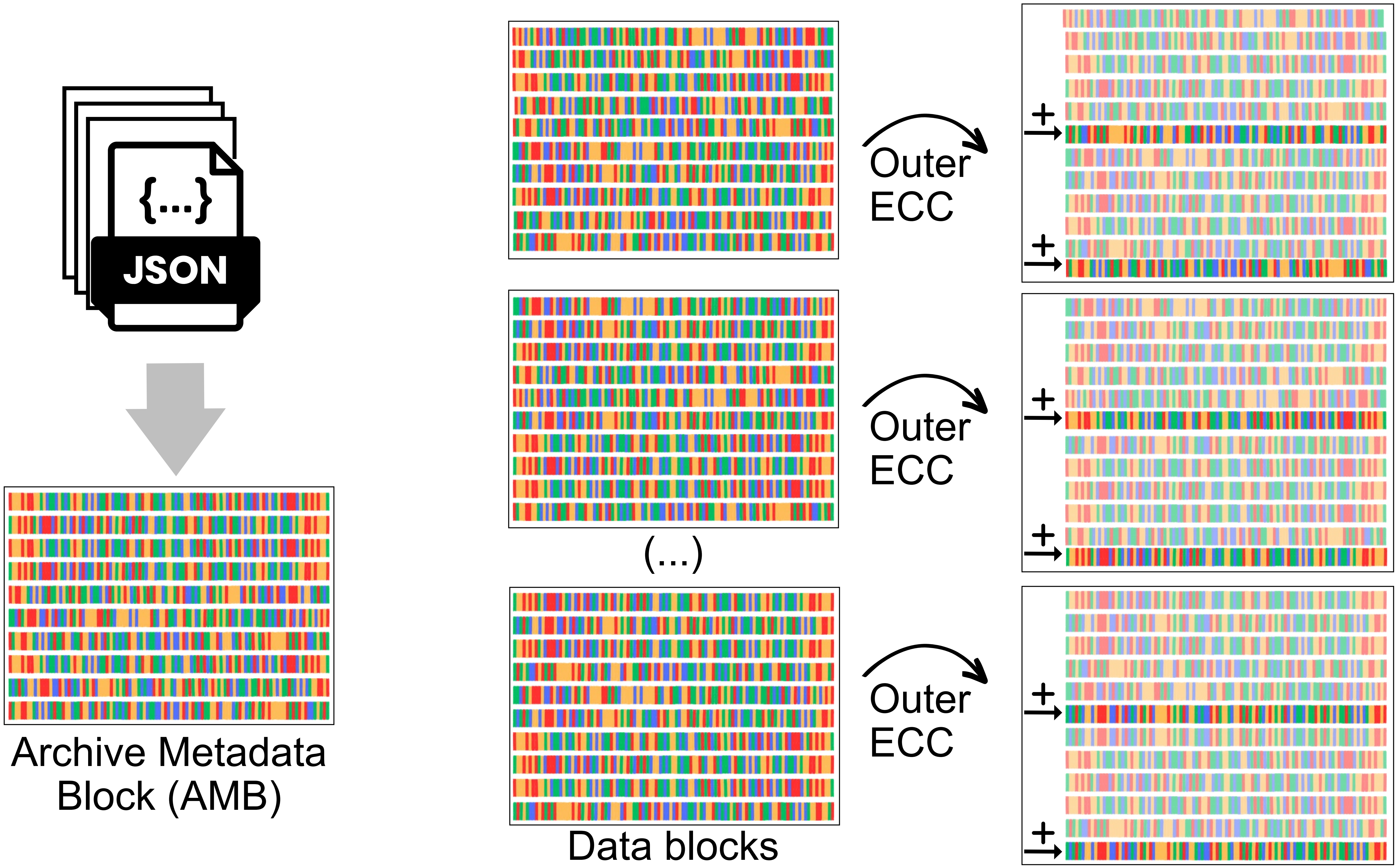
+

Codec parameters from data blocks

Data blocks

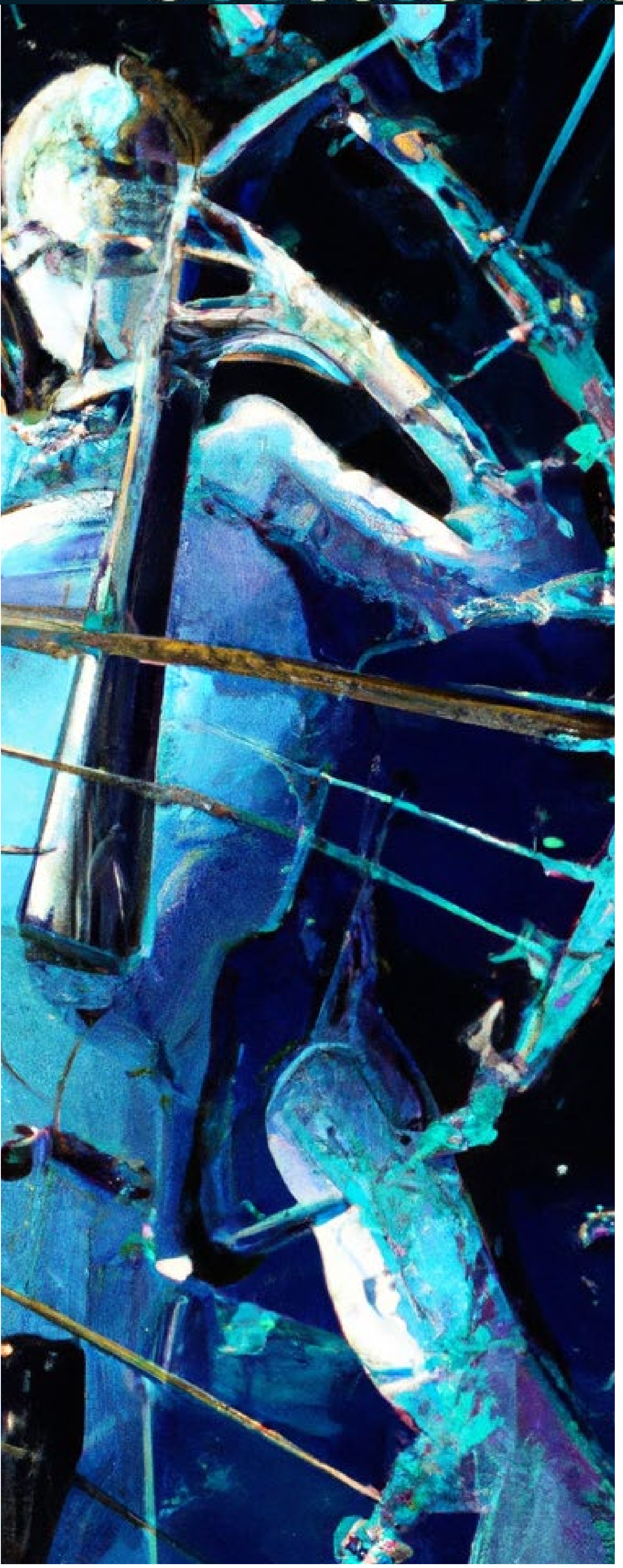
Apollo: metadata & Outer ECC



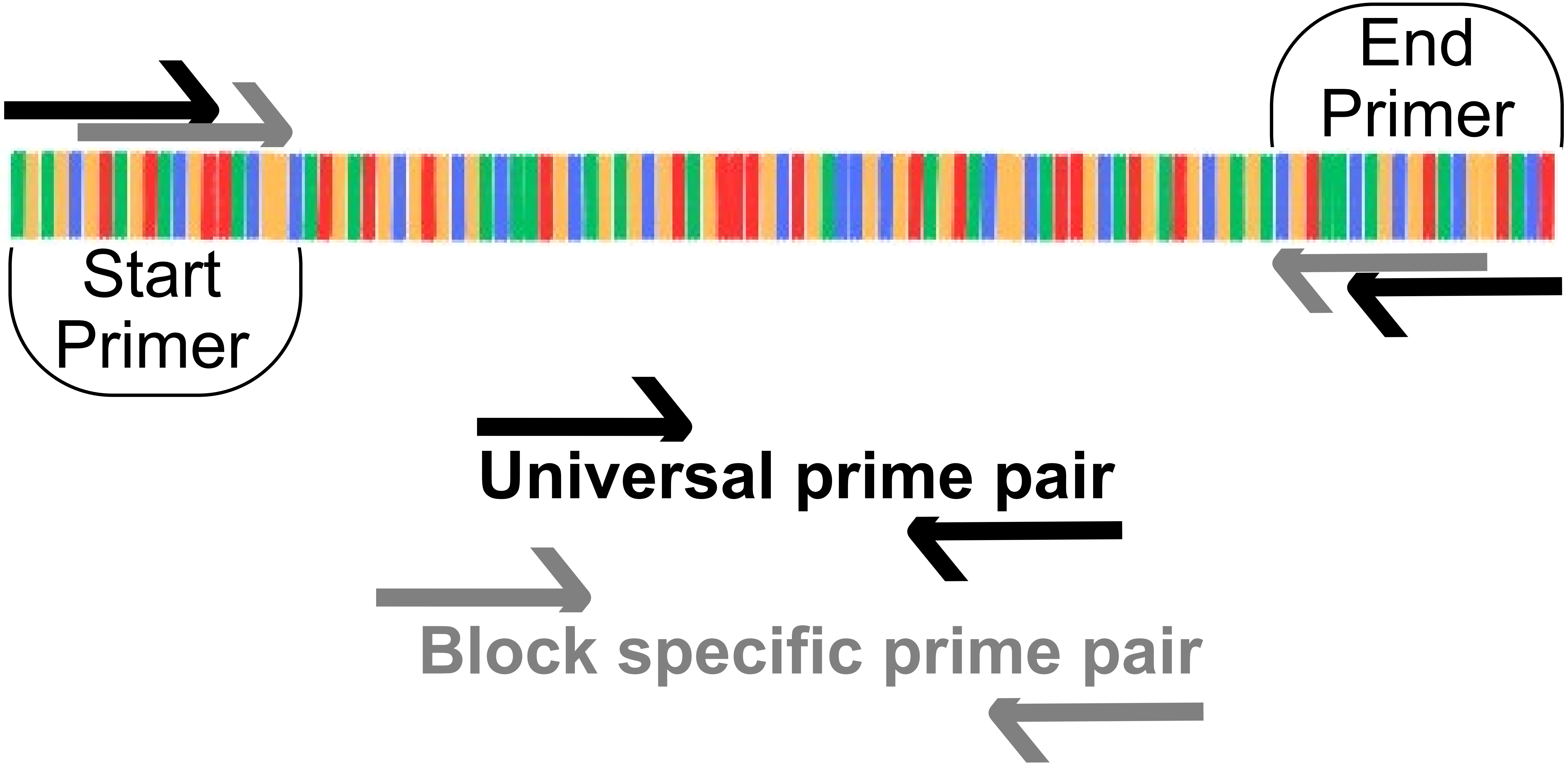


Archive Metadata Block (AMB)

Data blocks



Apollo: primers architecture

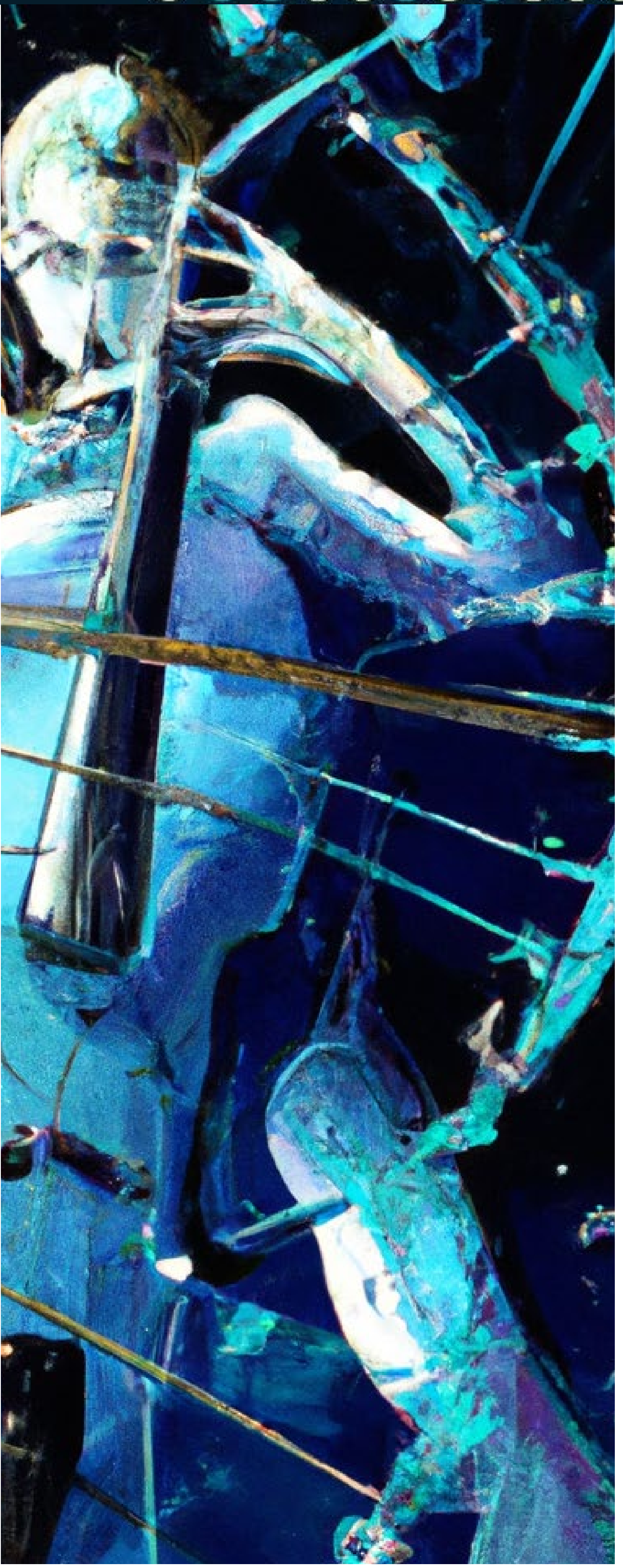


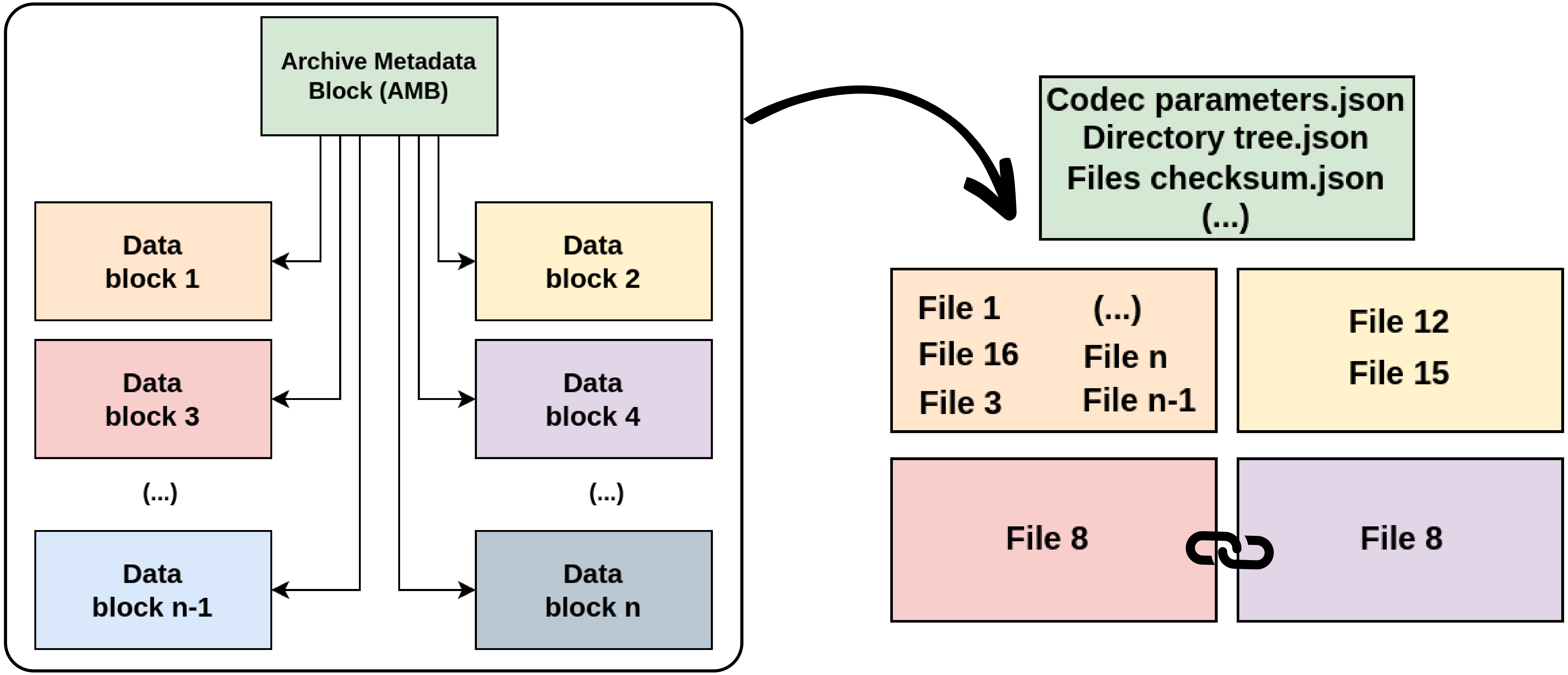
Universal prime pair

Block specific prime pair

PROMETHEUS

Apollo: DNA blocks architecture





Archive Metadata Block (AMB)

Data block 1

Data block 2

Data block 3

Data block 4

(...)

(...)

Data block n-1

Data block n

Codec parameters.json
Directory tree.json
Files checksum.json
(...)

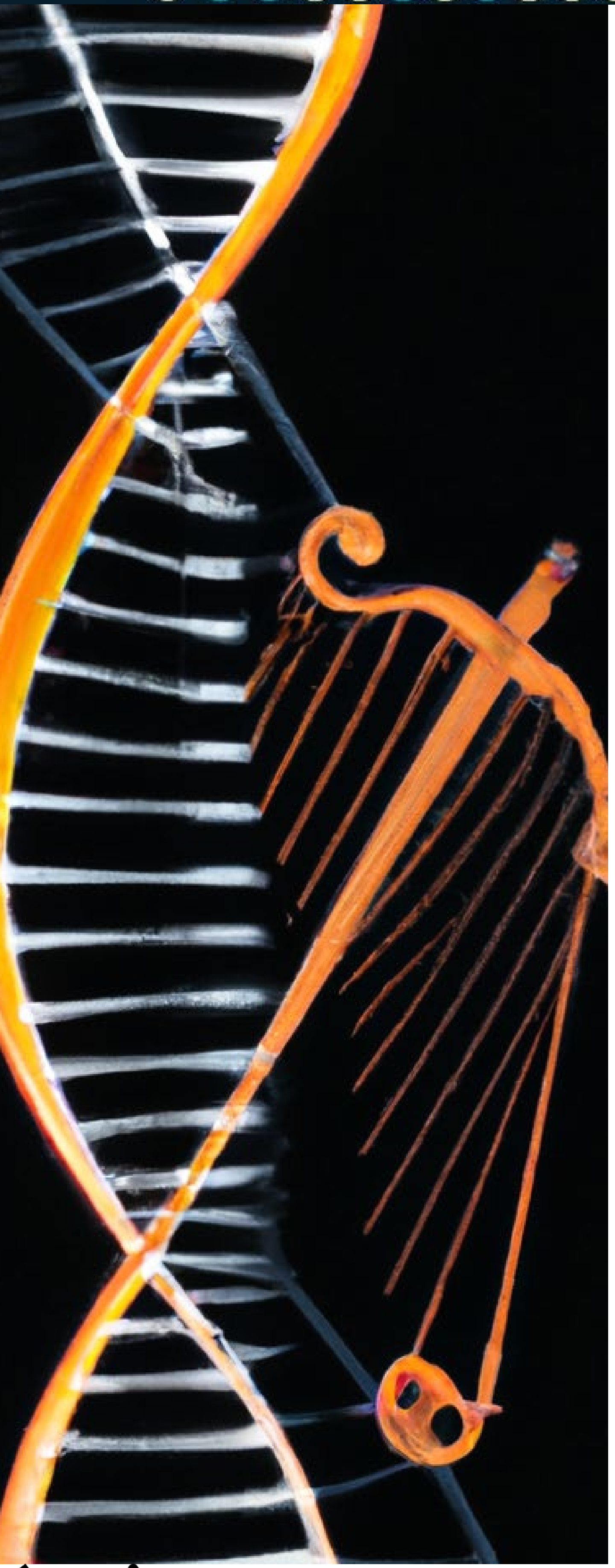
File 1
File 16
File 3

(...)
File n
File n-1

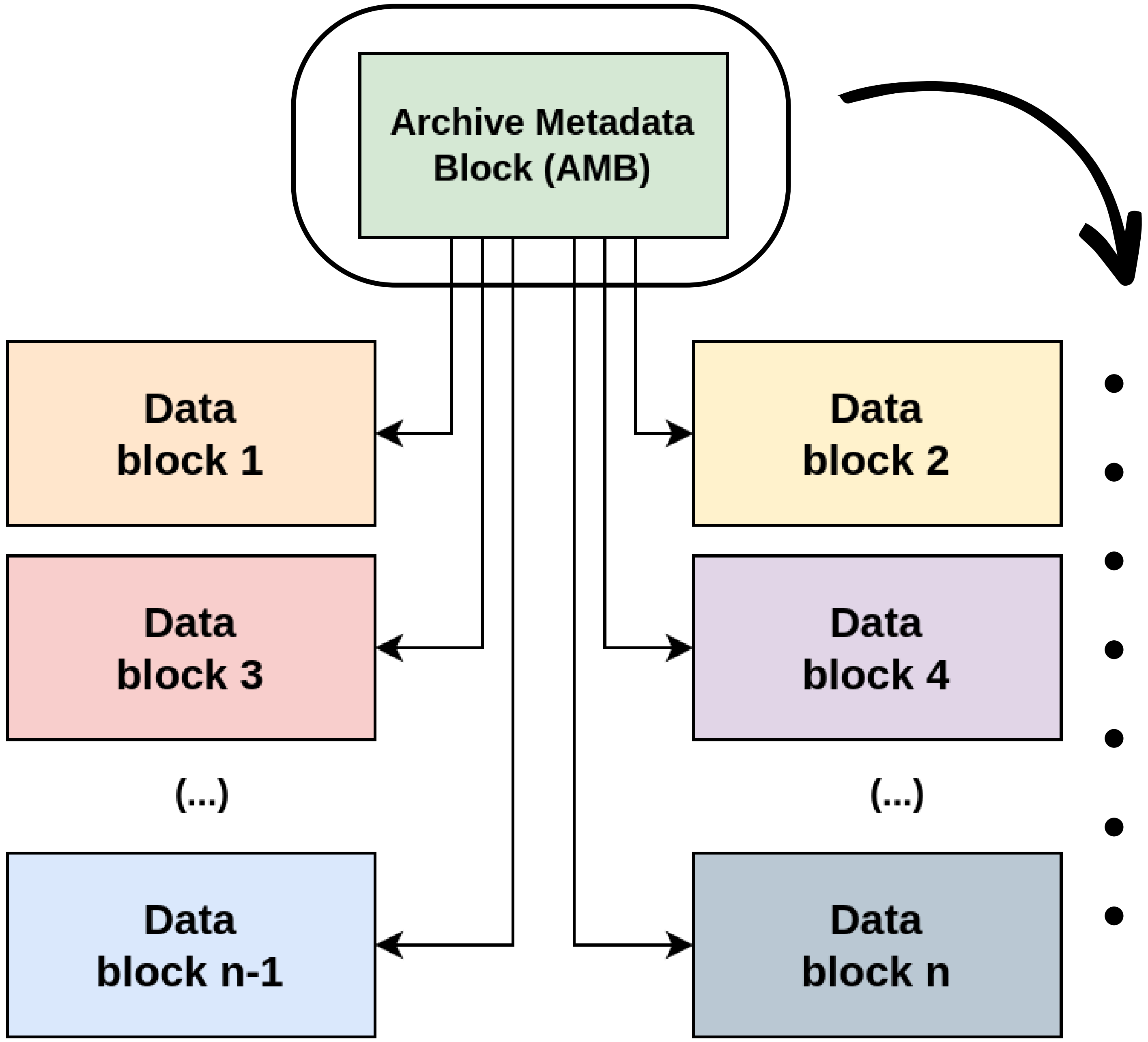
File 12
File 15

File 8

File 8

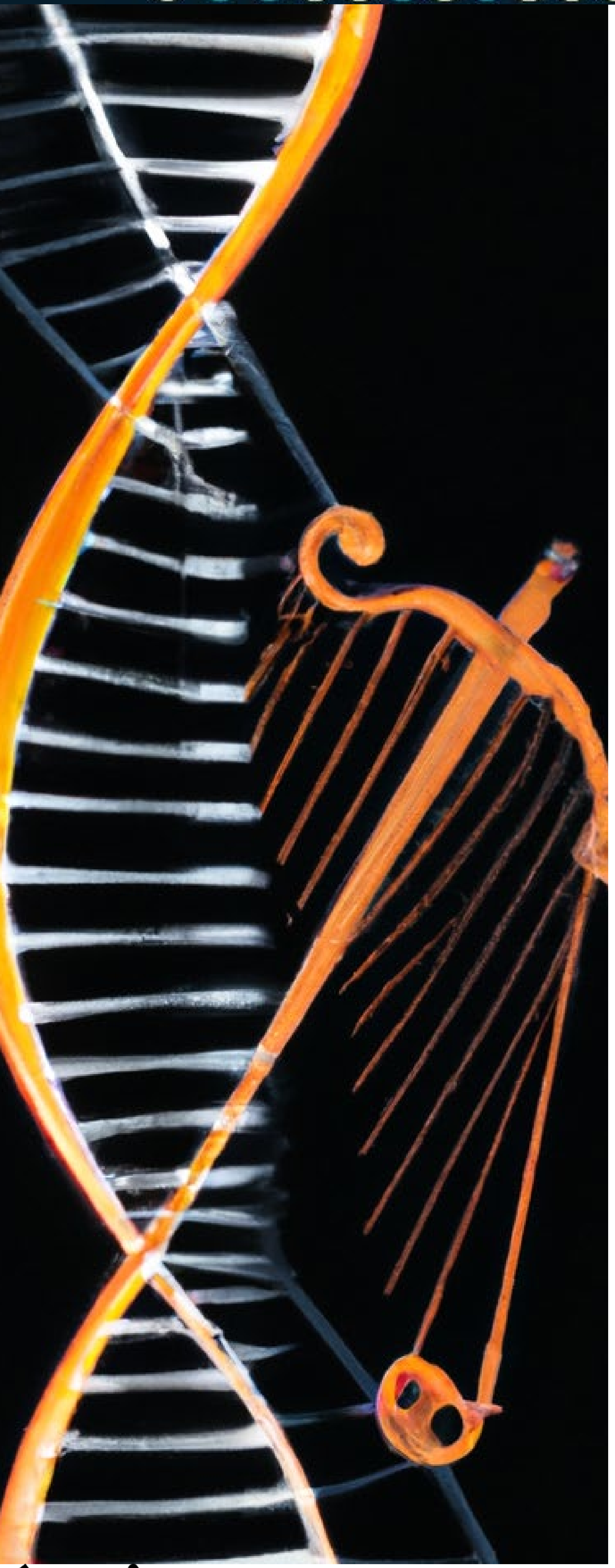


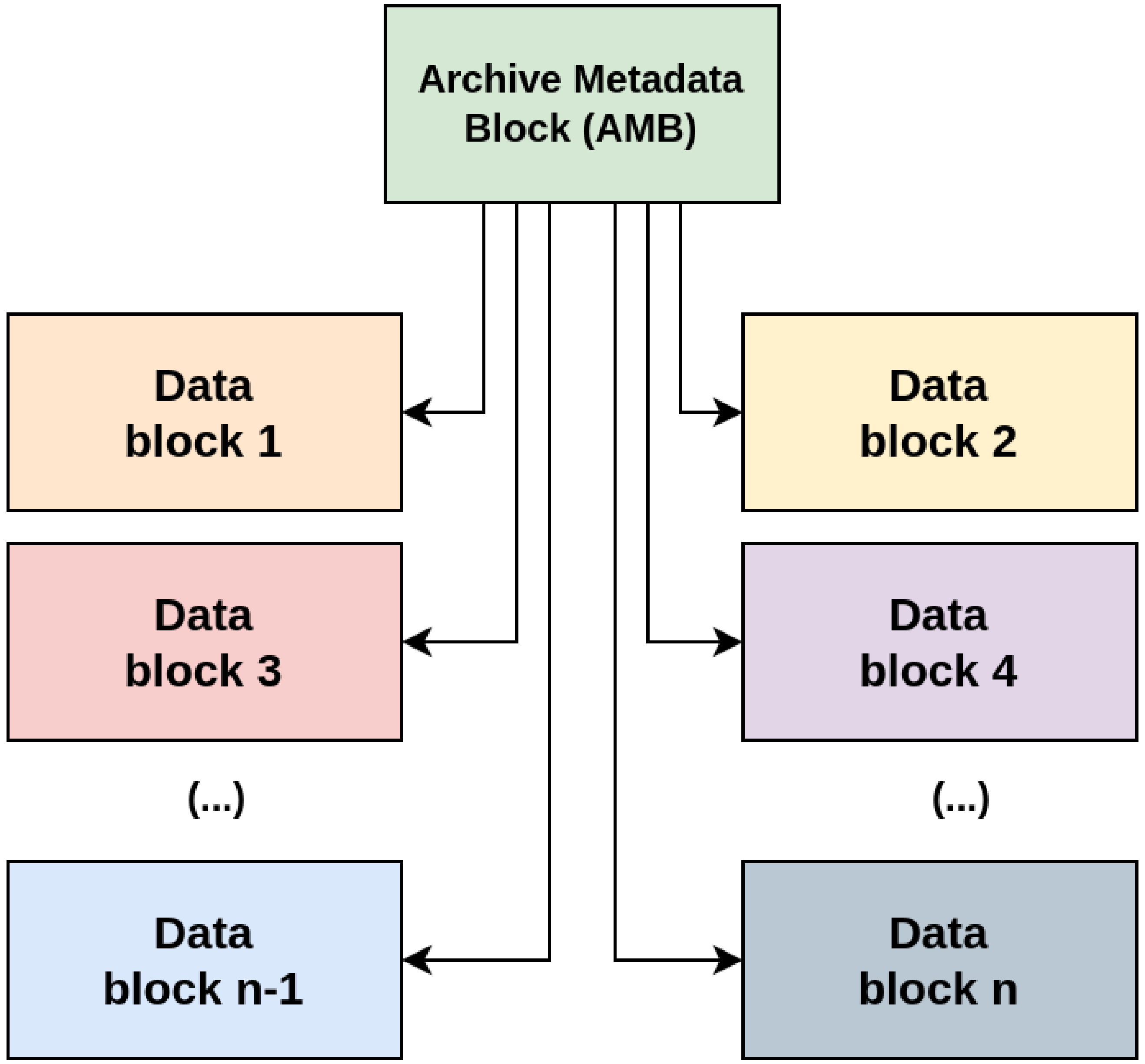
Apollo: DNA blocks architecture

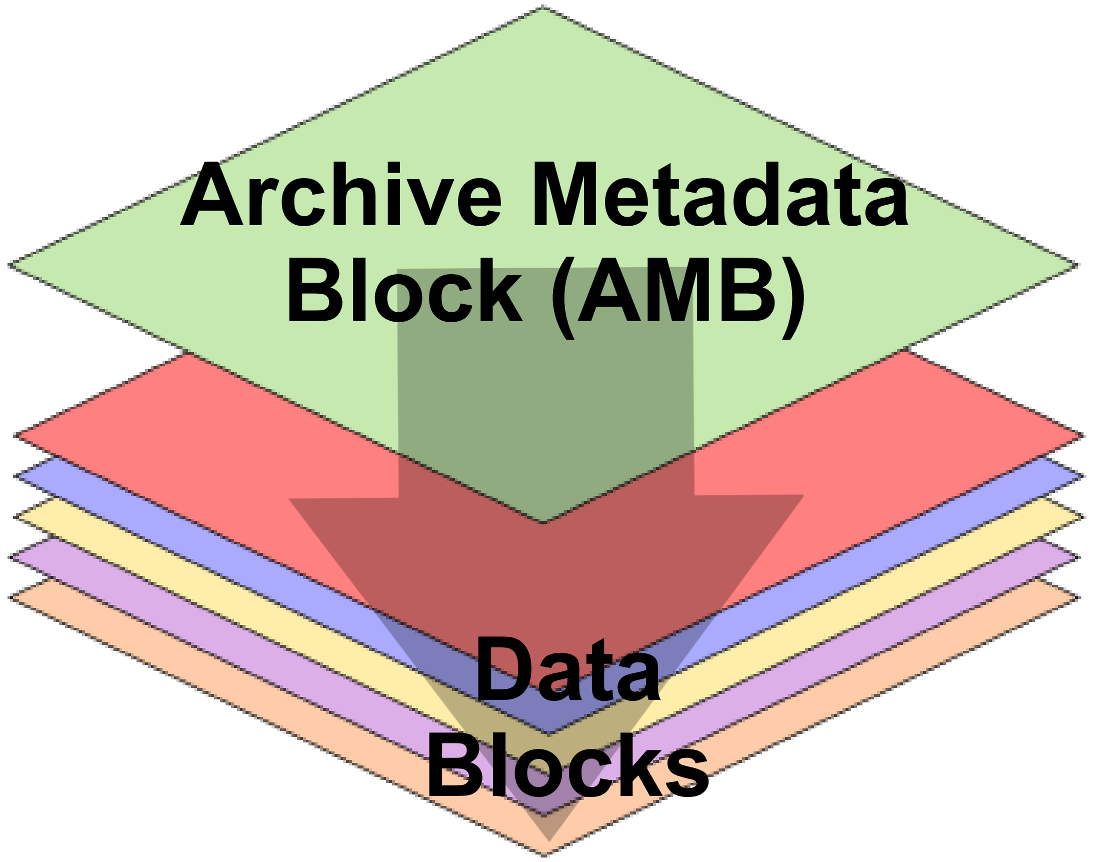


- **System directory tree**
- **Files checksum**
- **Files coordenates within blocks**
- **CODEC parameters for blocks**
- **Primers data**
- CODEC manual
- Other types of data to assist data recover

Apollo: DNA blocks architecture

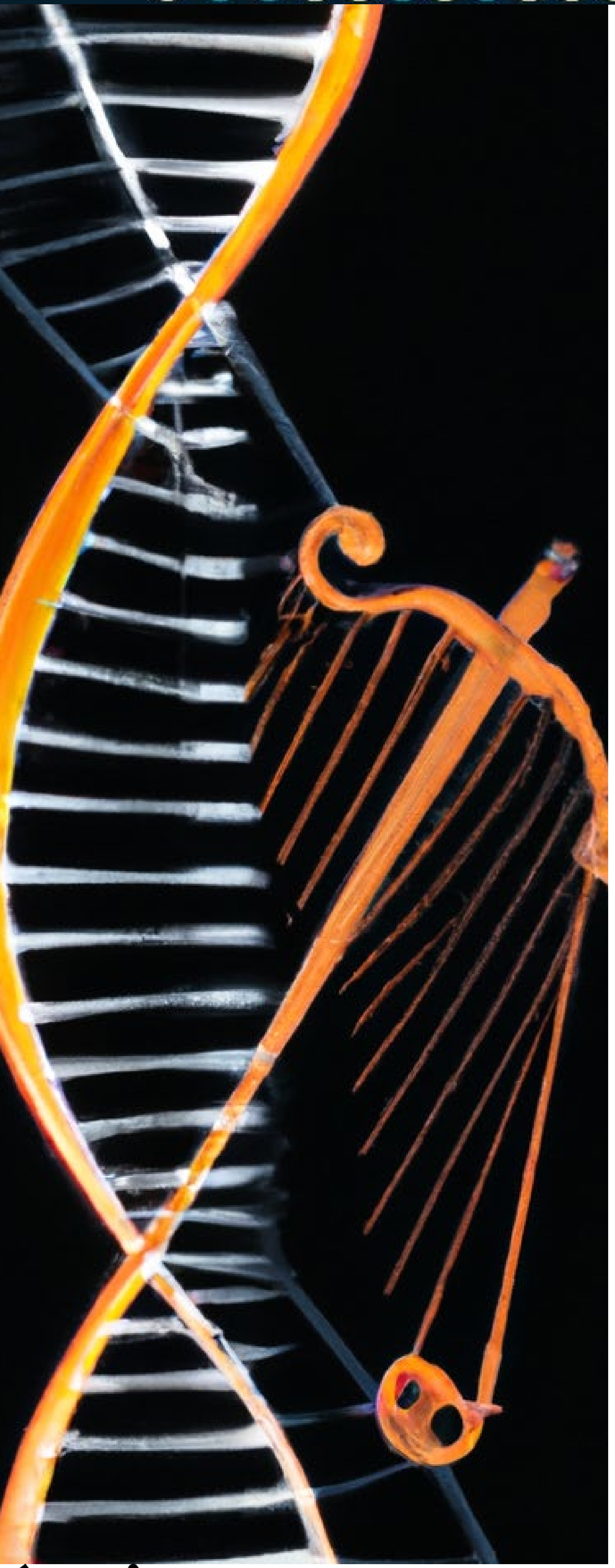






PROMETHEUS

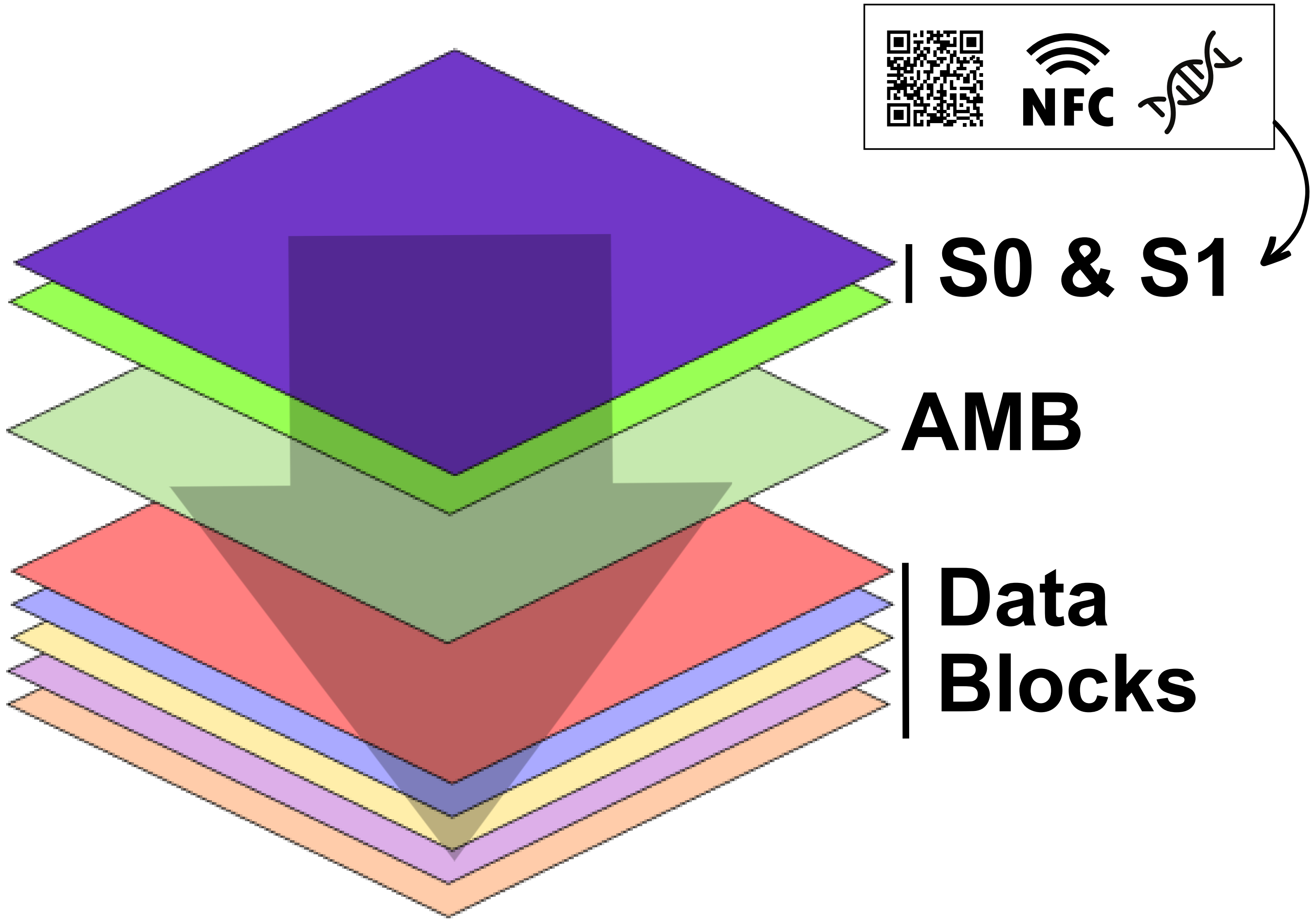
©2023 SNIA. All Rights Reserved.

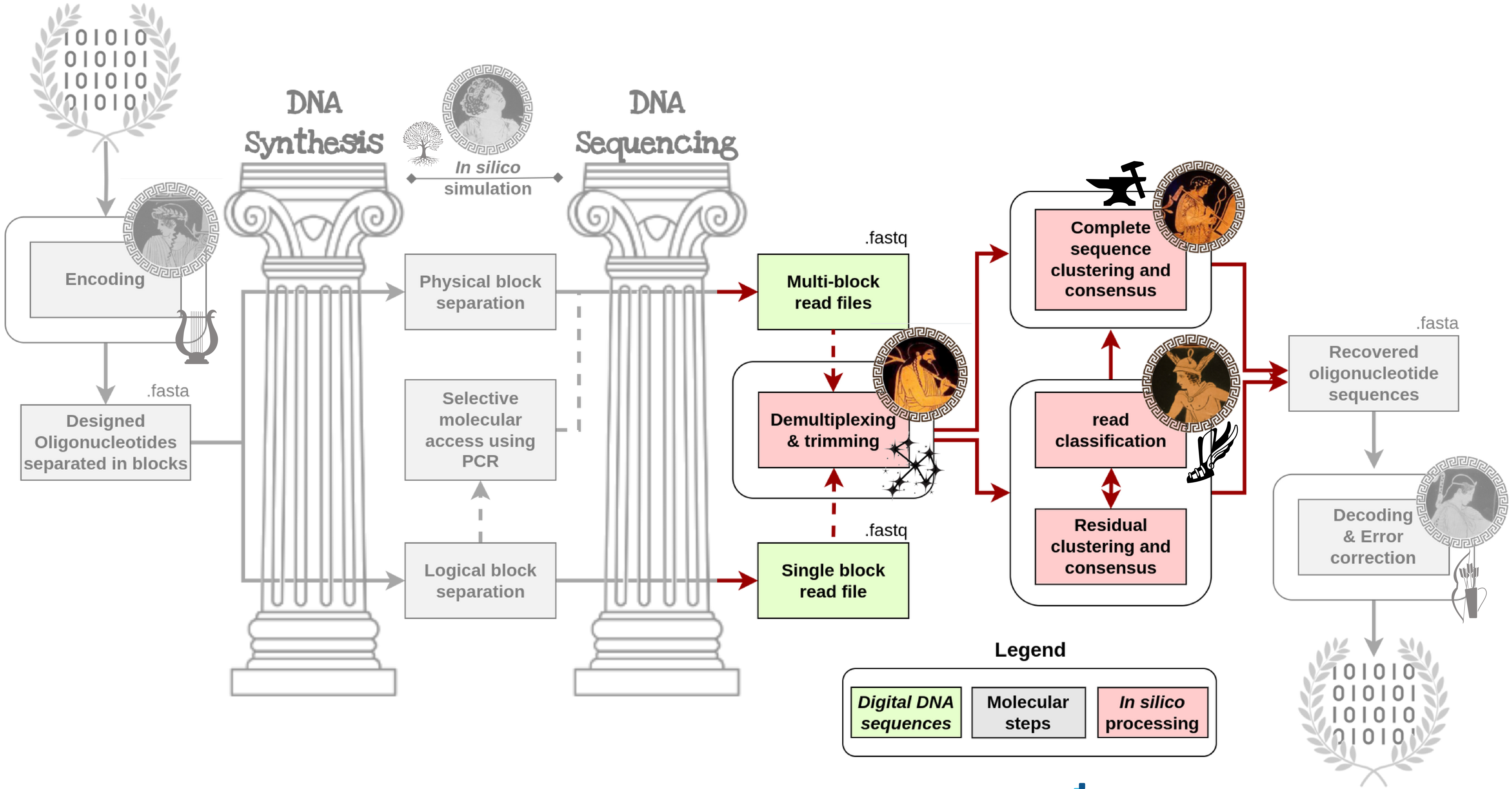


Apollo: Integration with SNIA's specifications for sectors S0 & S1

S1 - CODEC parameters

- General CODEC & ECC parameters
- AMB especific ECC & CODEC parameters
- AMB files checksum





DNA Synthesis

In silico simulation

DNA Sequencing

Encoding

.fasta

Designed Oligonucleotides separated in blocks

Physical block separation

Selective molecular access using PCR

Logical block separation

.fastq

Multi-block read files

Demultiplexing & trimming

.fastq

Single block read file

Complete sequence clustering and consensus

read classification

Residual clustering and consensus

.fasta

Recovered oligonucleotide sequences

Decoding & Error correction

Legend

Digital DNA sequences	Molecular steps	In silico processing

Chiron, the NGS pre-processing module

The Education of Achilles - Bénigne Gagneraux, 1785

Chiron: Pre-processing NGS reads

Pre-processing steps:

- Merge read pair (paired-end strategy)
- Adapters/Primer trimming
- Demultiplexing coding blocks
- Reorient DNA sequences
- Discard low-quality reads

PROMETHEUS

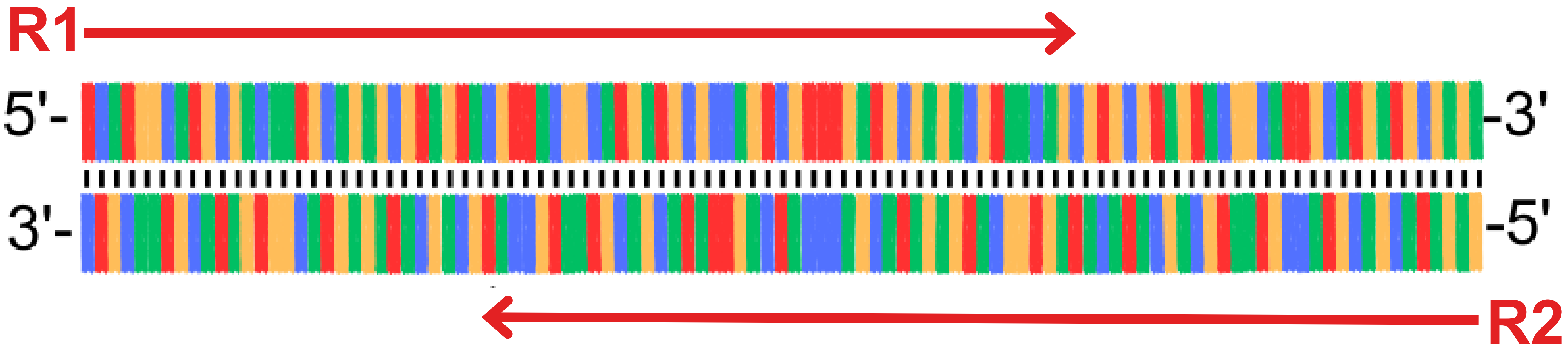
ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS Lenovo SDC 23

Chiron: Pre-processing NGS reads

Pre-processing steps:

- Merge read pair (paired-end strategy)
- Adapters/Primer trimming
- Demultiplexing coding blocks
- Reorient DNA sequences
- Discard low-quality reads

🟩 Cytosine
🟨 Guanine
🟥 Adenine
🟦 Thymine



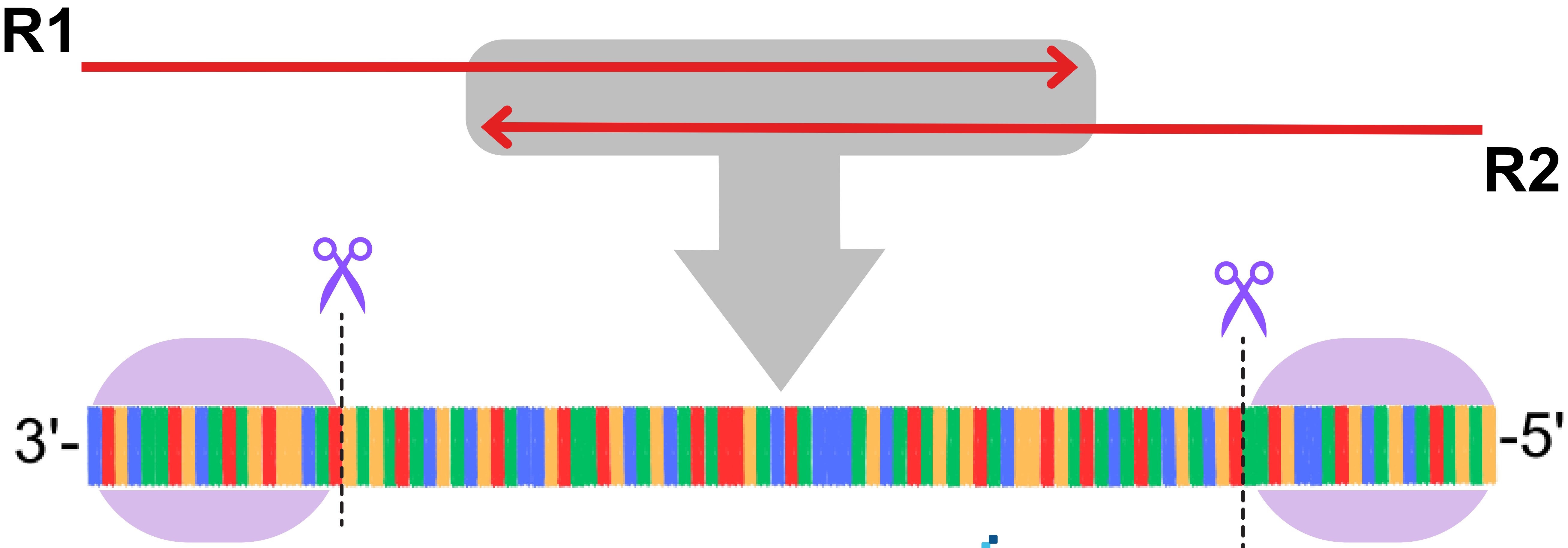
Chiron: Pre-processing NGS reads

PROMETHEUS

🟩 Cytosine
🟨 Guanine
🟥 Adenine
🟦 Thymine

Pre-processing steps:

- Merge read pair (paired-end strategy)
- Adapters/Primer trimming
- Demultiplexing coding blocks
- Reorient DNA sequences
- Discard low-quality reads



PROMETHEUS

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS Lenovo SDC 23

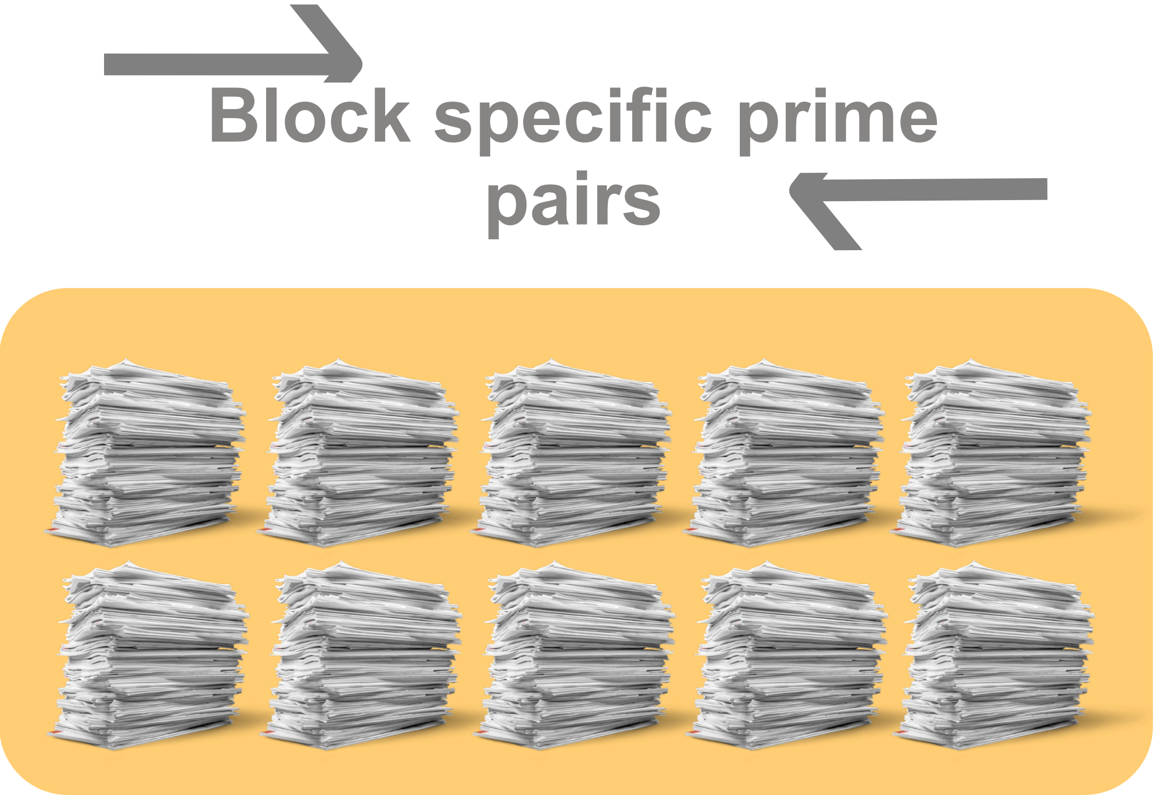
Chiron: Pre-processing NGS reads

Pre-processing steps:

- Merge read pair (paired-end strategy)
- Adapters/Primer trimming
- Demultiplexing coding blocks
- Reorient DNA sequences
- Discard low-quality reads

Planned data blocks

Sequenced DNA reads

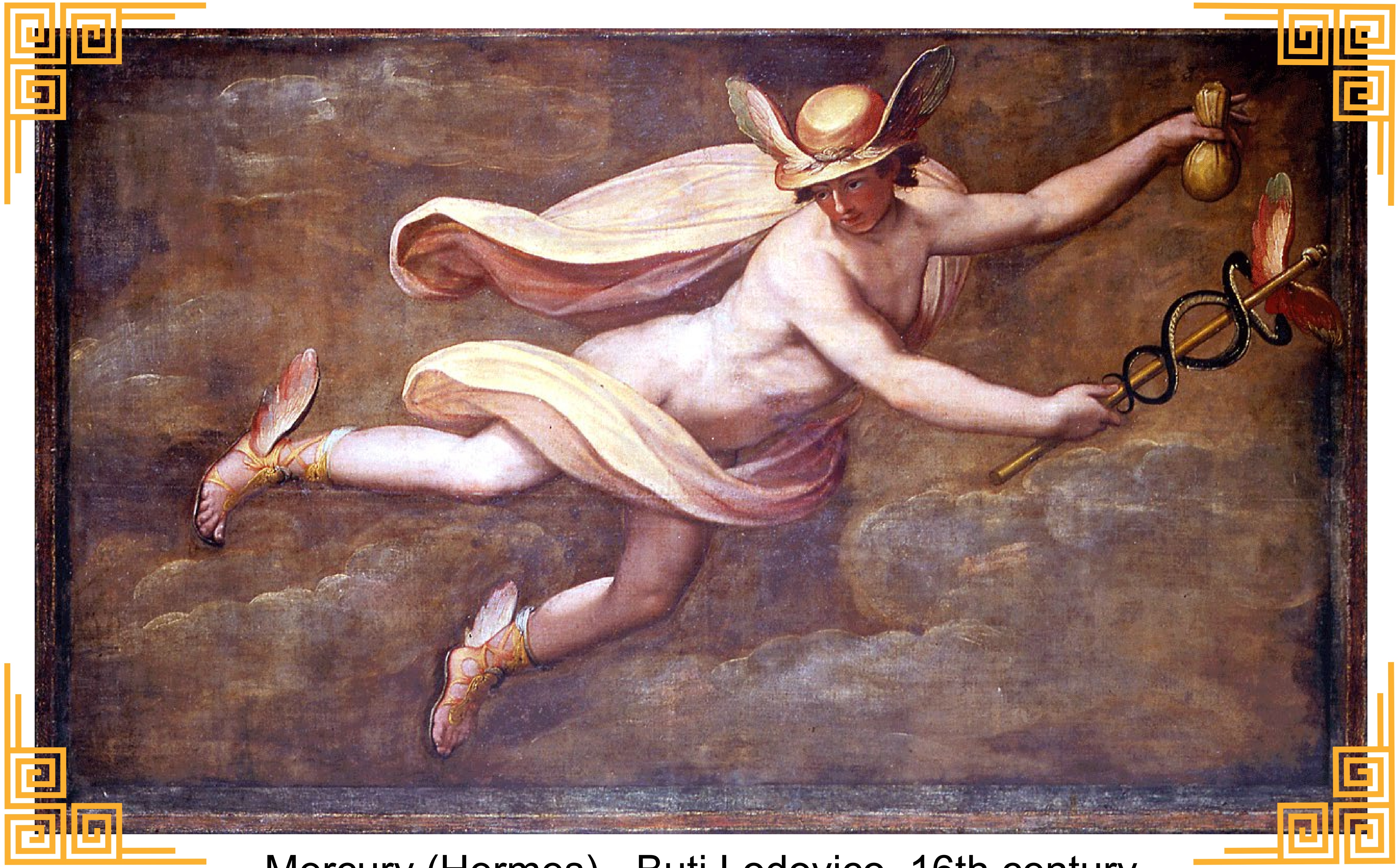


Sequenced & Demultiplexed into data blocks

PROMETHEUS

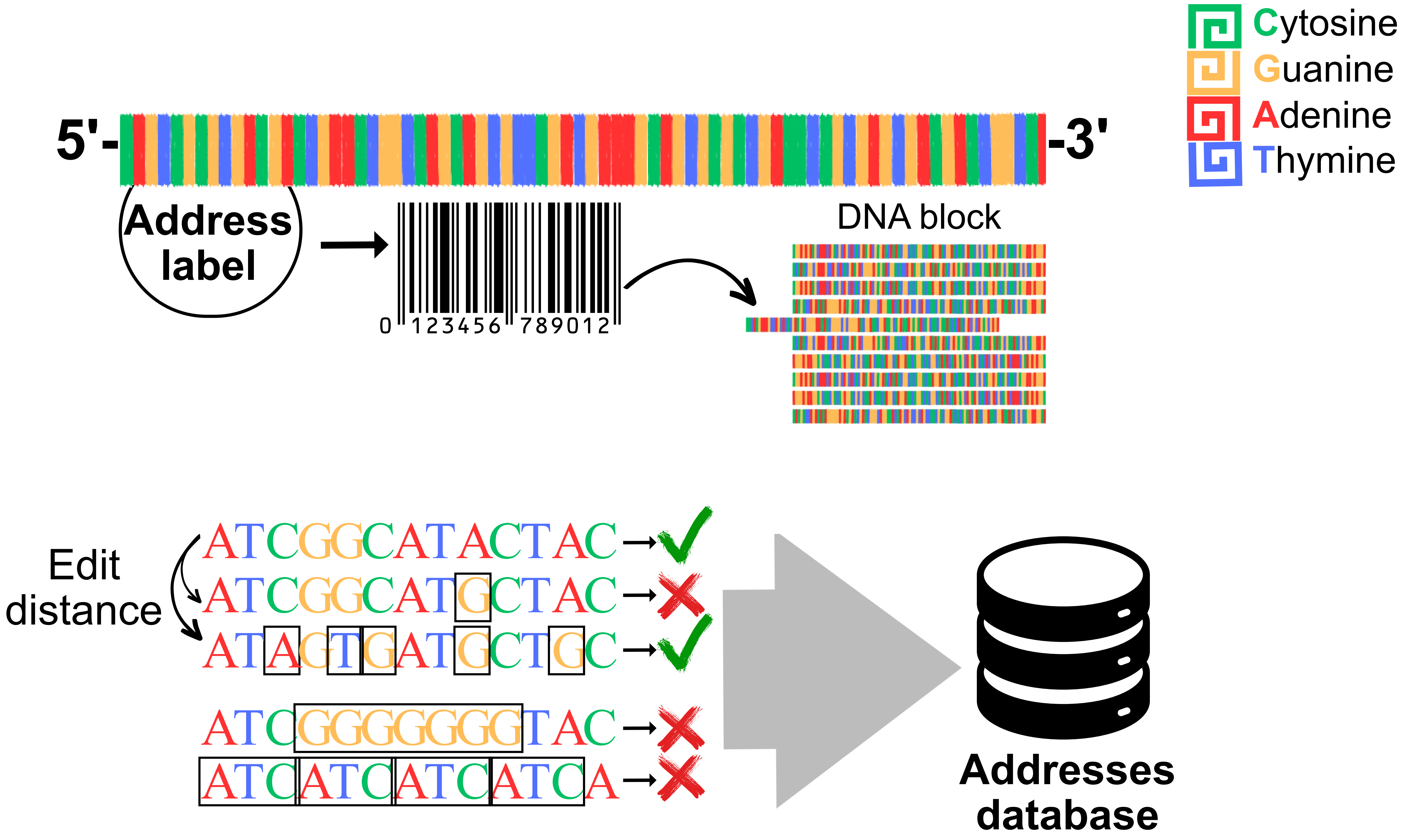
ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS Lenovo SDC 23

Hermes, address-oriented module

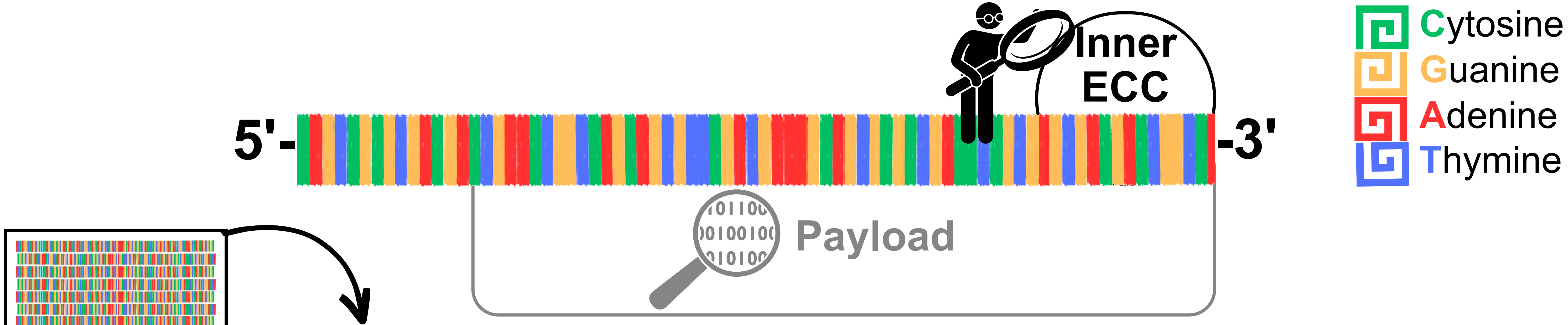


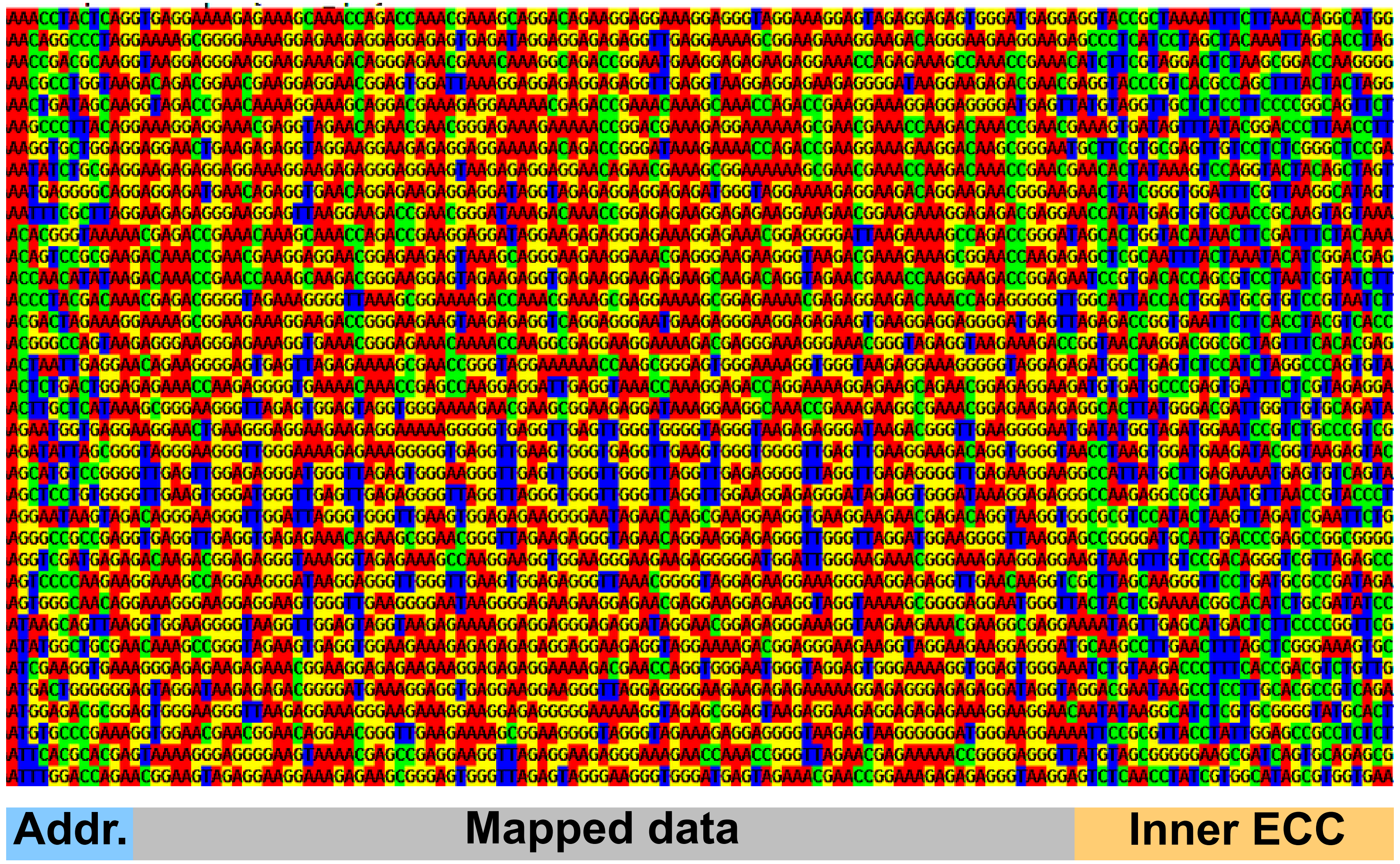
Mercury (Hermes) - Buti Lodovico, 16th century

Hermes: address library



Hermes: parity check

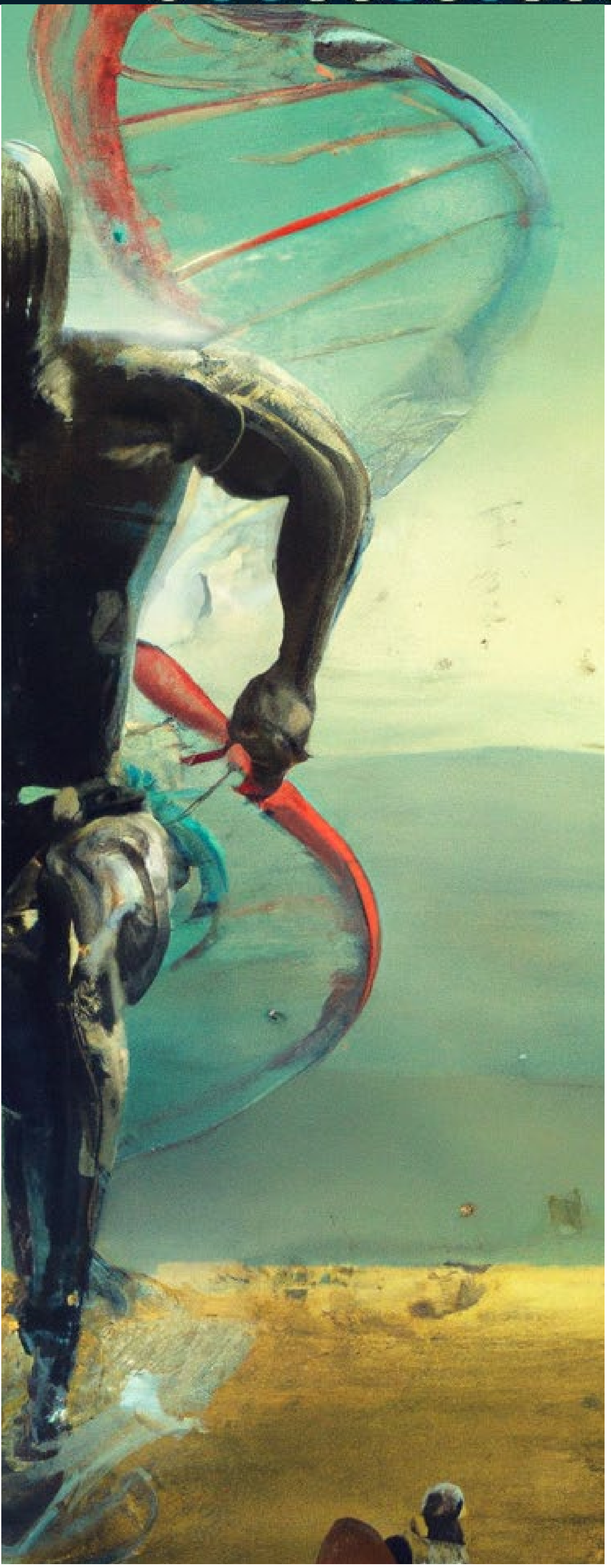


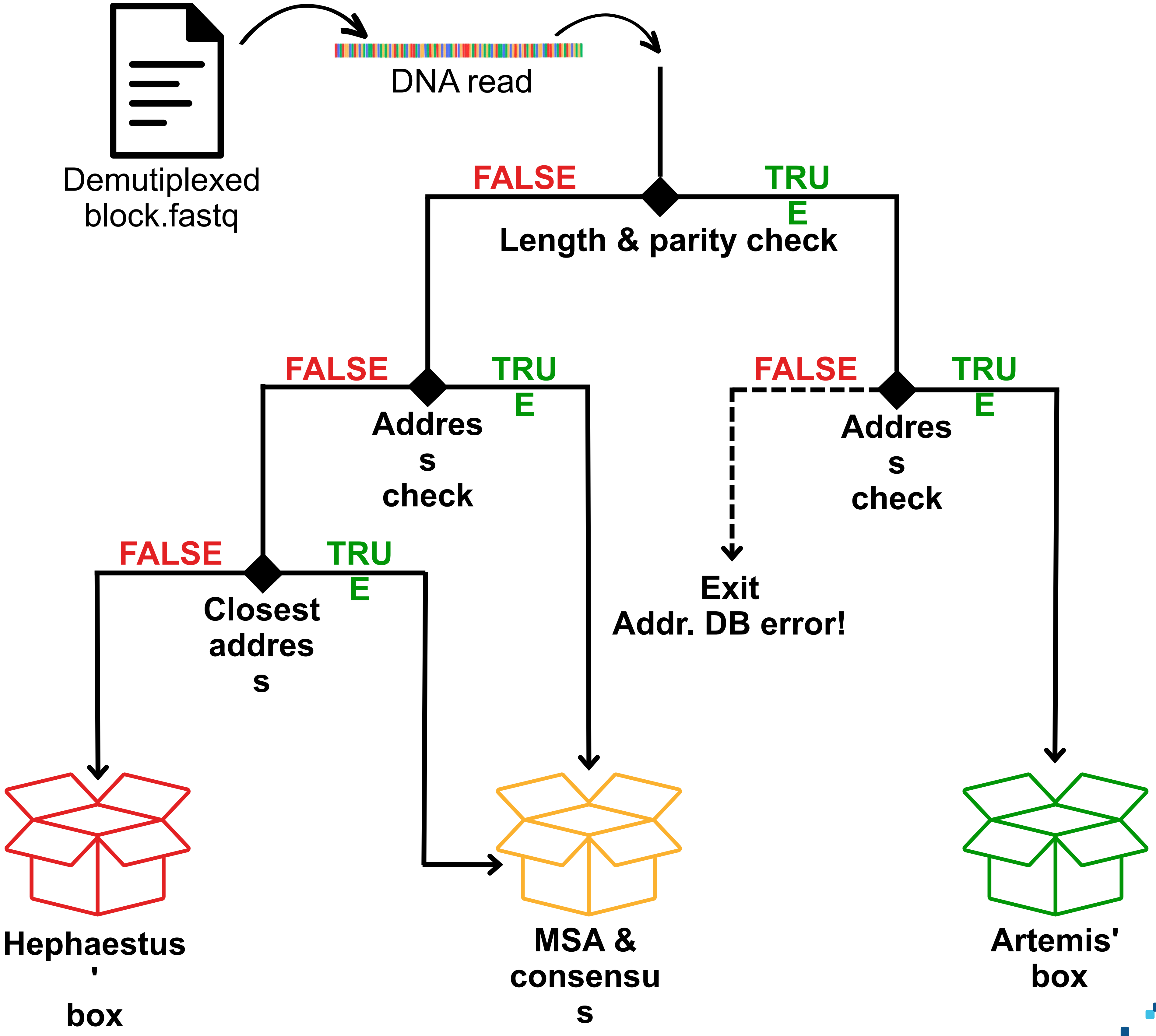


Screenshot of the first 35 DNA sequences from a planned data block in a multi-fasta format. The payload was mapped using a G&A rich scheme.

PROMETHEUS

Hermes: basic algorithm, part 1



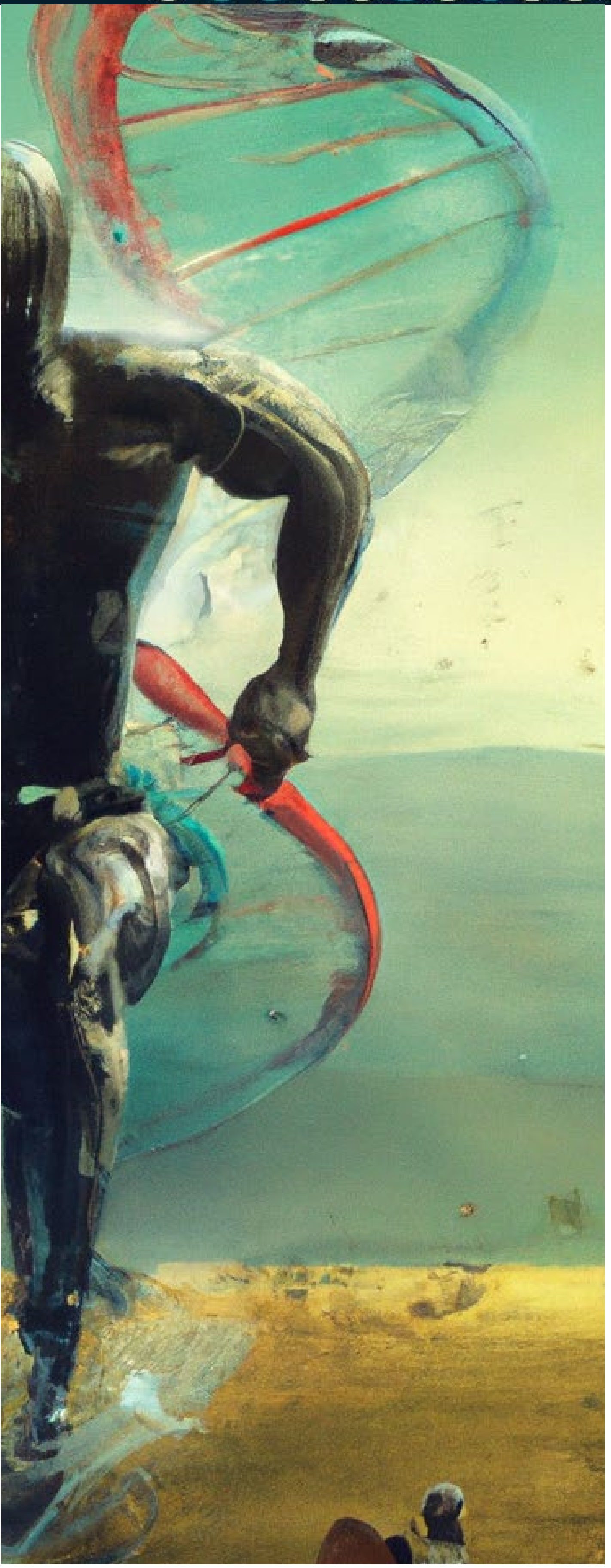


Hermes inputs

Demplx.
block.fastq + Addr. DB

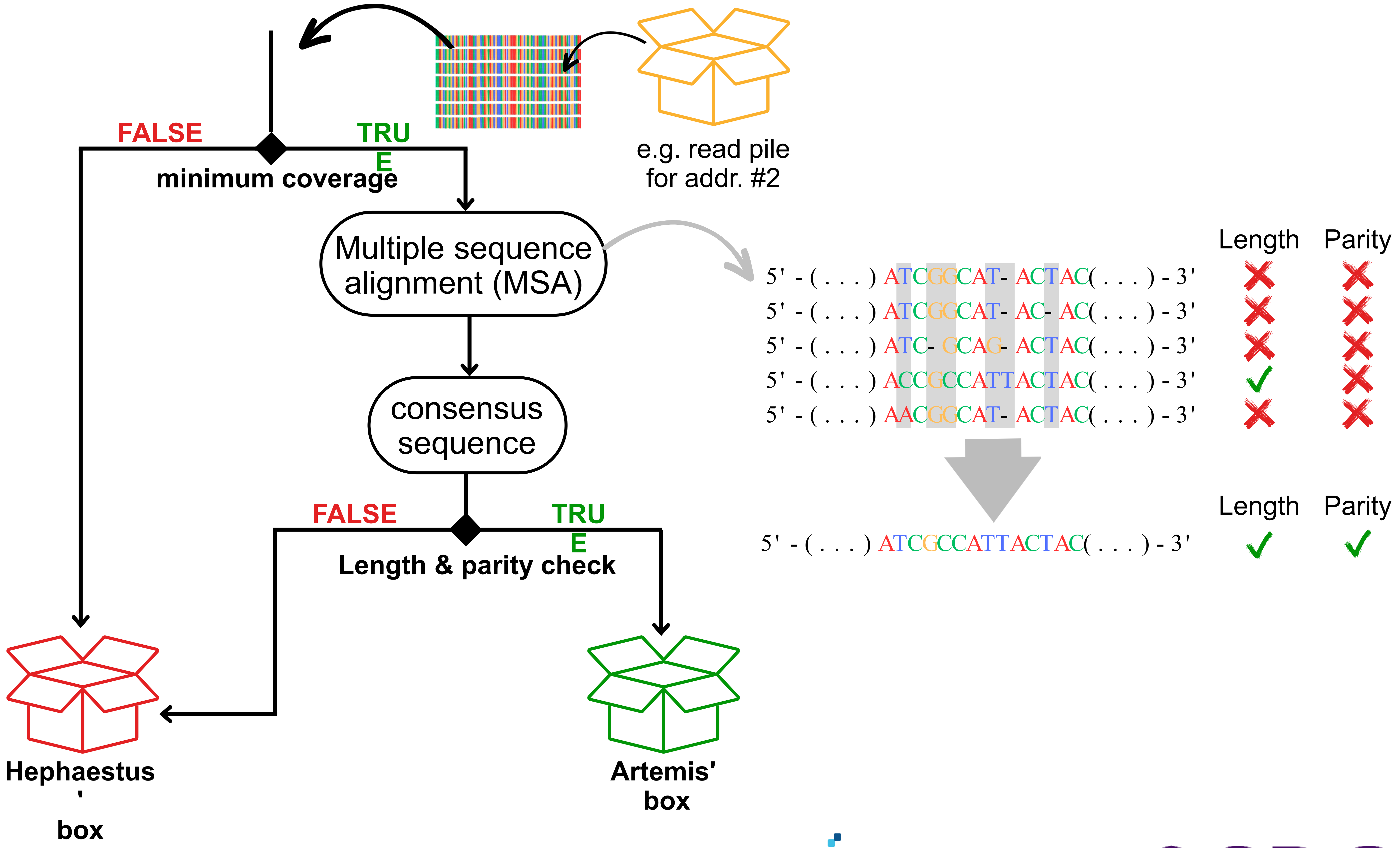
Address table

Addr.	Seq. pile	Status
#1	5 3	Close
#2	1	Filled
#3	2	Open
...	...	...
#n	8 7	Close



PROMETHEUS

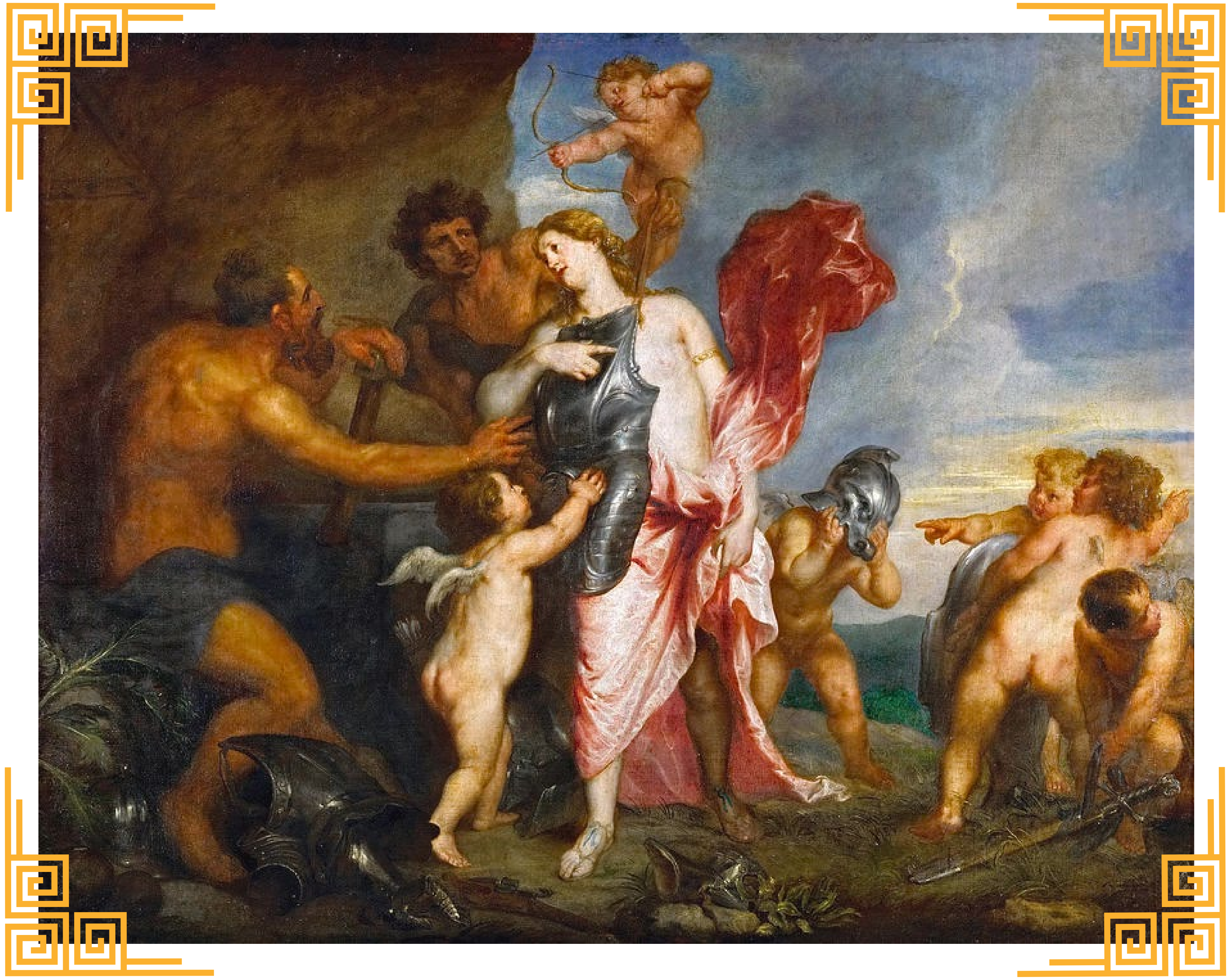
Hermes: basic algorithm, part 2



	Length	Parity
5' - (. . .) ATCGGCAT- ACTAC(. . .) - 3'	✗	✗
5' - (. . .) ATCGGCAT- AC- AC(. . .) - 3'	✗	✗
5' - (. . .) ATC- GCAG- ACTAC(. . .) - 3'	✗	✗
5' - (. . .) ACCGCCATTACTAC(. . .) - 3'	✓	✗
5' - (. . .) AACGGCAT- ACTAC(. . .) - 3'	✗	✗

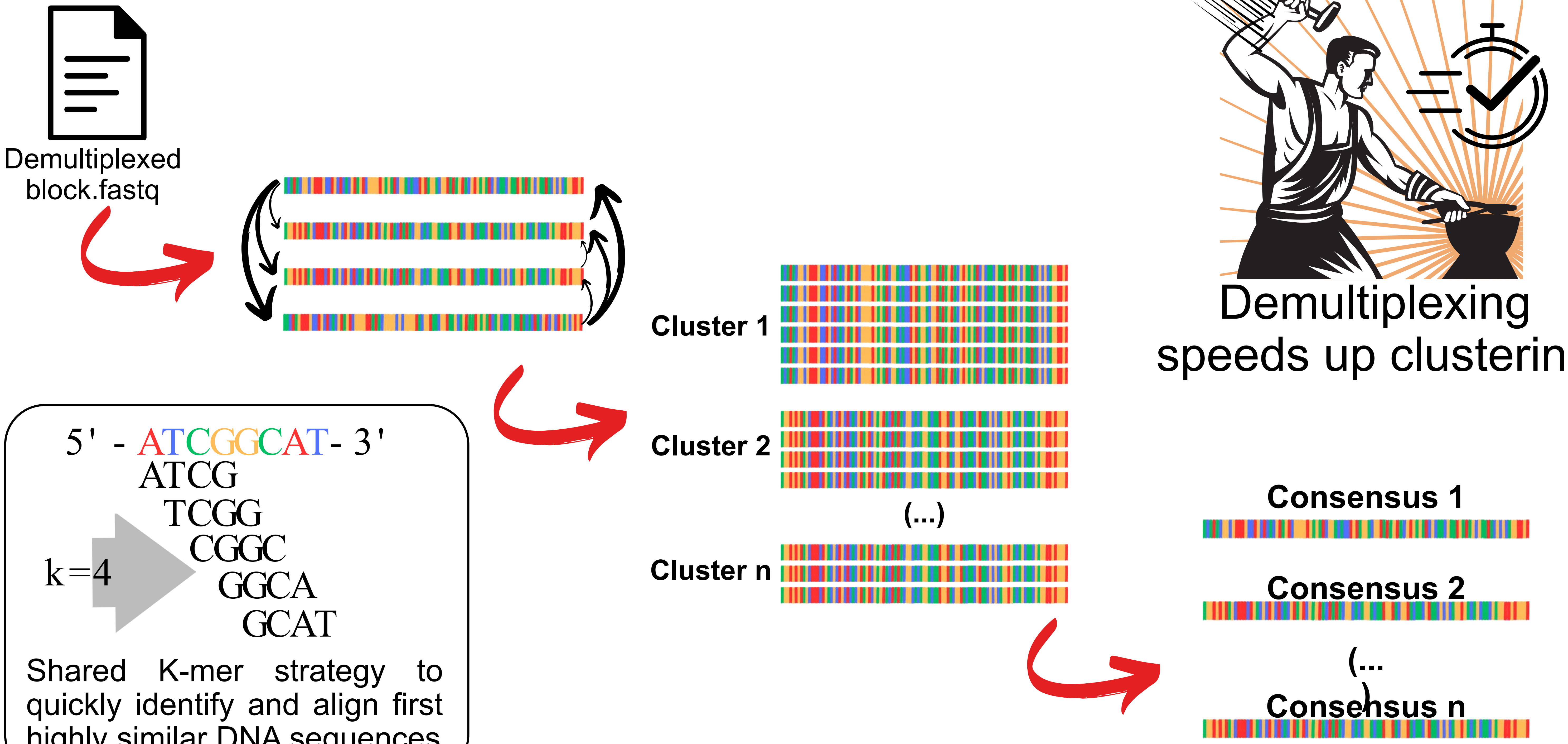
	Length	Parity
5' - (. . .) ATCGCCATTACTAC(. . .) - 3'	✓	✓

Hephaestus, full-length clustering module



Thetis Receiving the Weapons of Achilles from Hephaestus
Anthony van Dyck, 1632

Hephaestus: pair-wise clustering and consensus

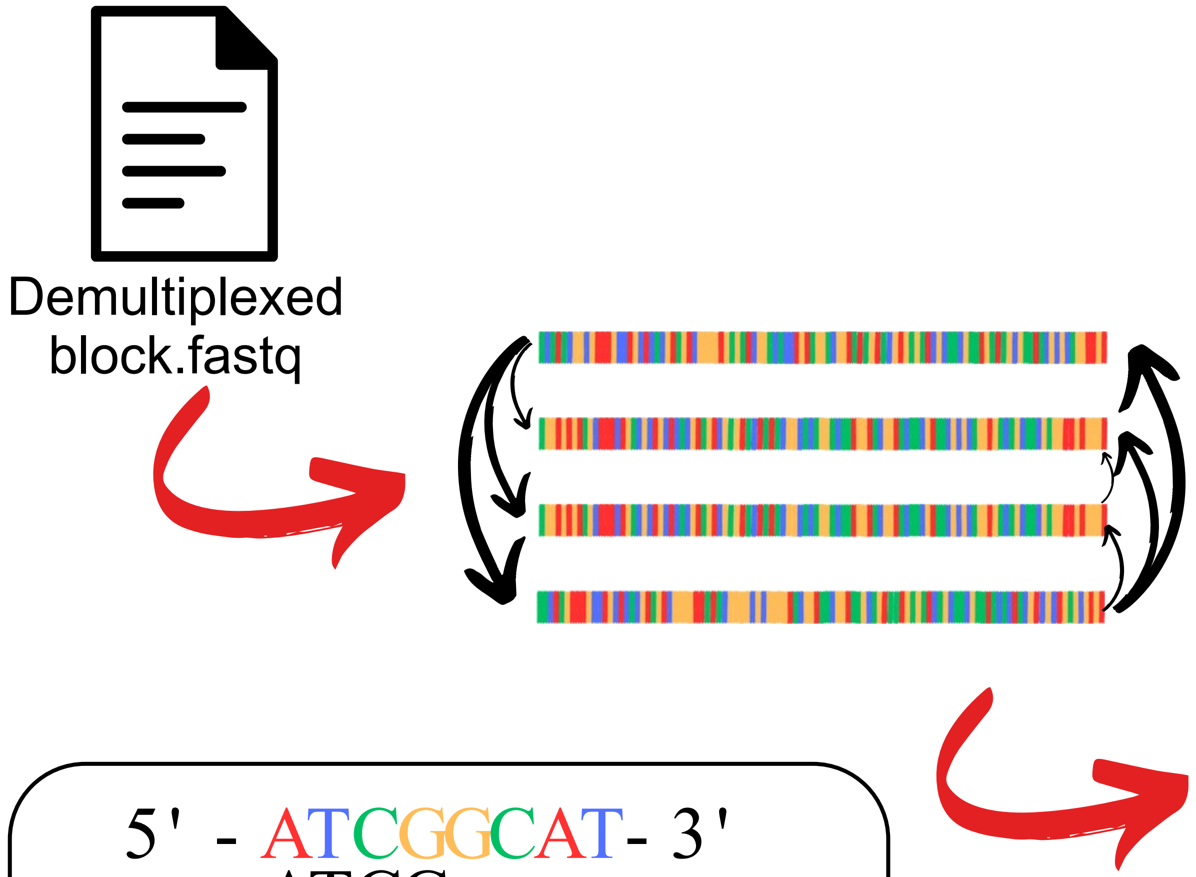


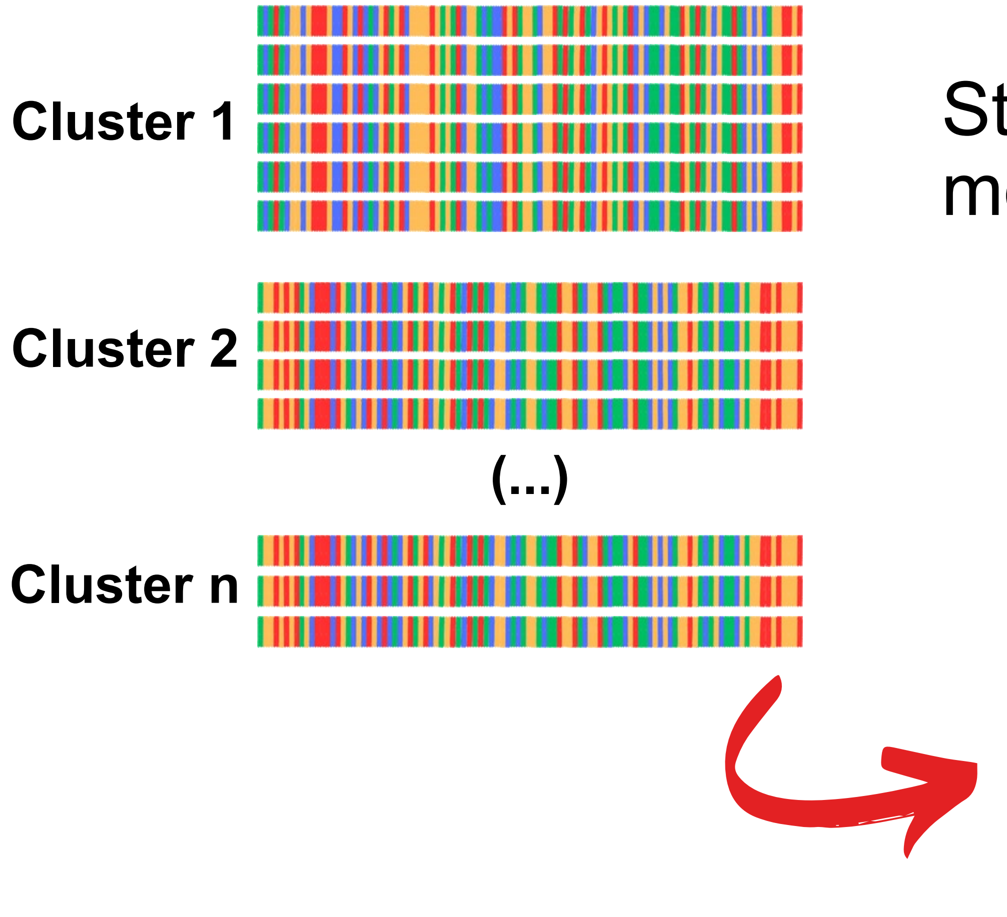
Shared K-mer strategy to quickly identify and align first highly similar DNA sequences

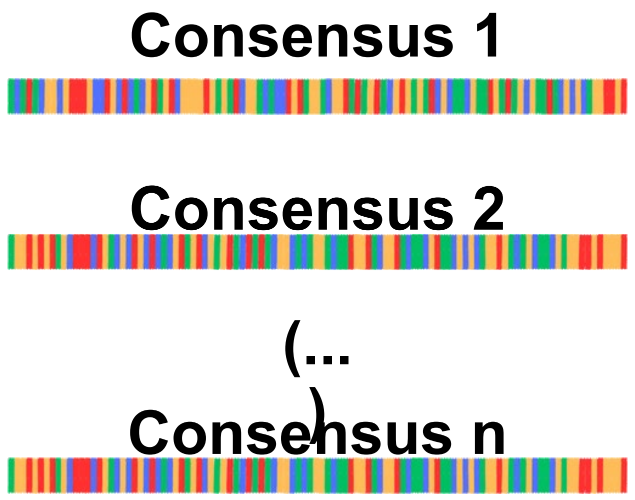
Demultiplexing speeds up clustering

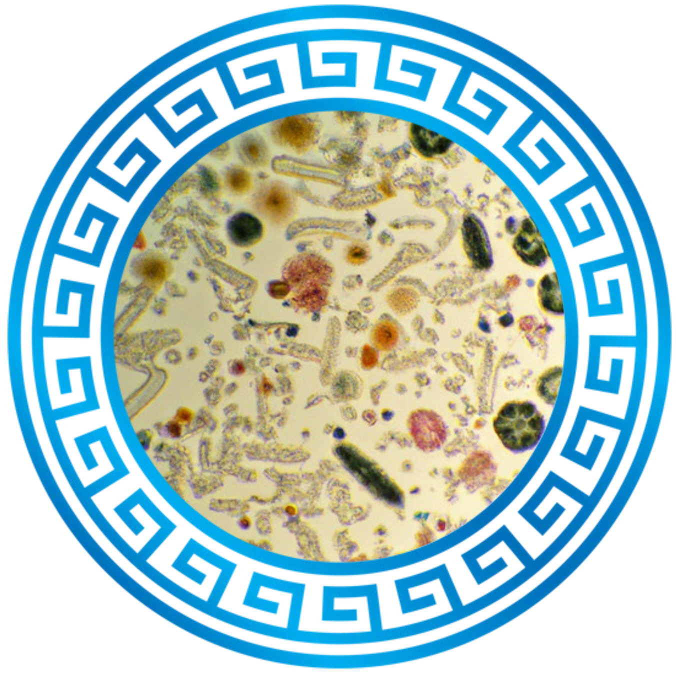
PROMETHEUS

Hephaestus: pair-wise clustering and consensus







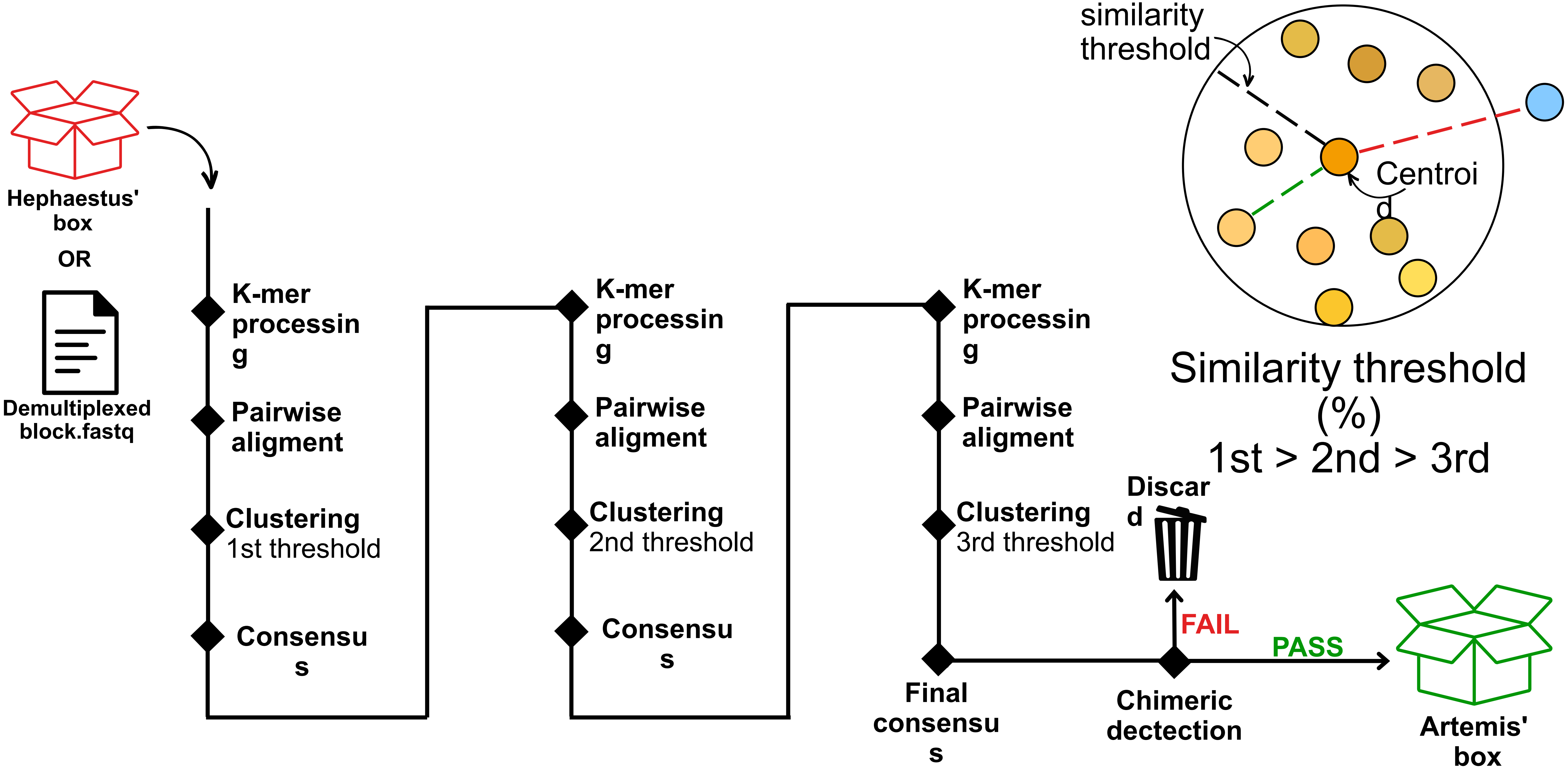


Strategy adapted from metagenomics studies

5' - ATCGGCAT- 3'
ATCG
TCGG
CGGC
GGCA
GCAT
k=4

Shared K-mer strategy to quickly identify and align first highly similar DNA sequences

Hephaestus: basic clustering algorithm

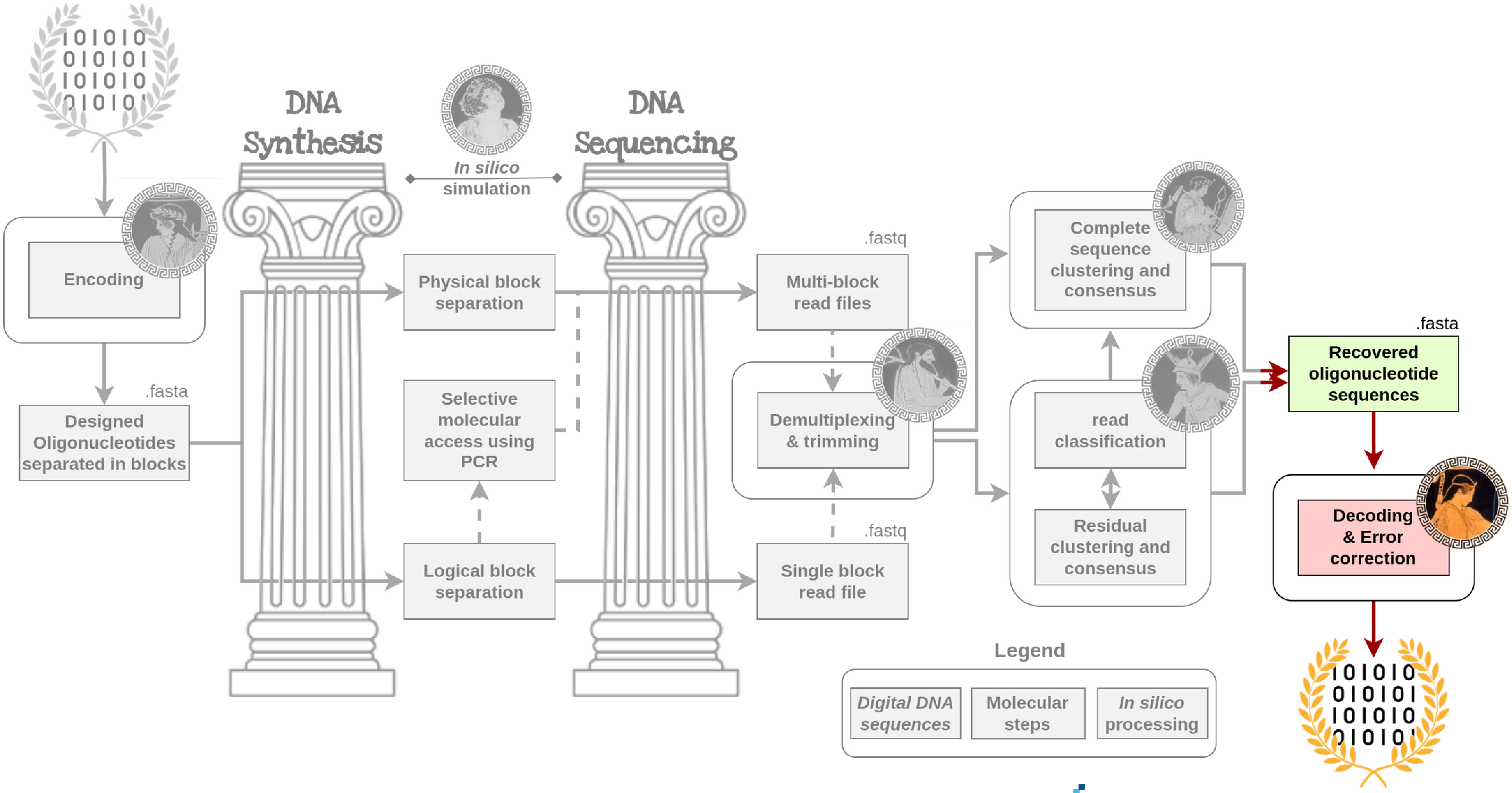


Similarity threshold (%)

1st > 2nd > 3rd

PROMETHEUS

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS Lenovo SDC23

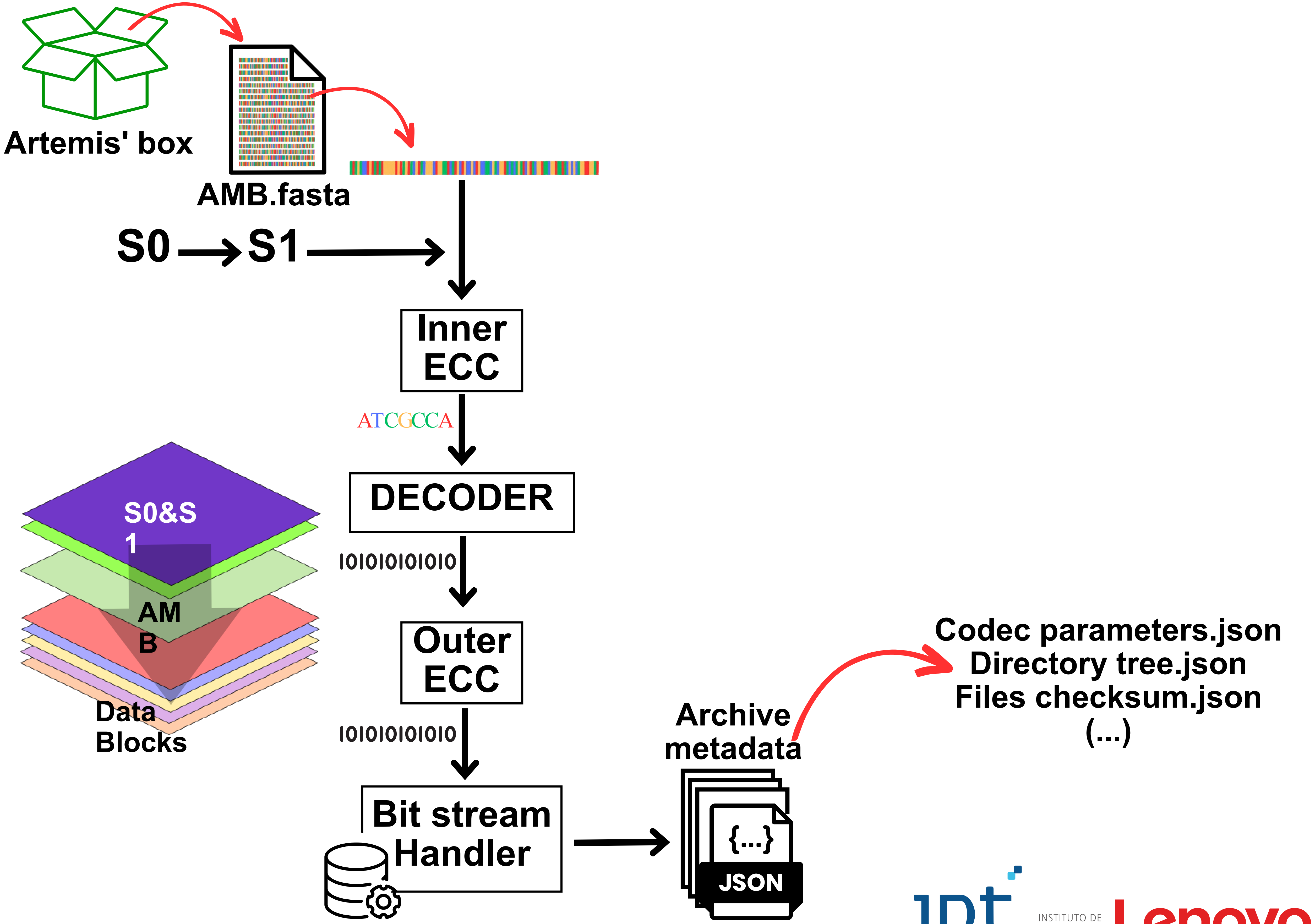


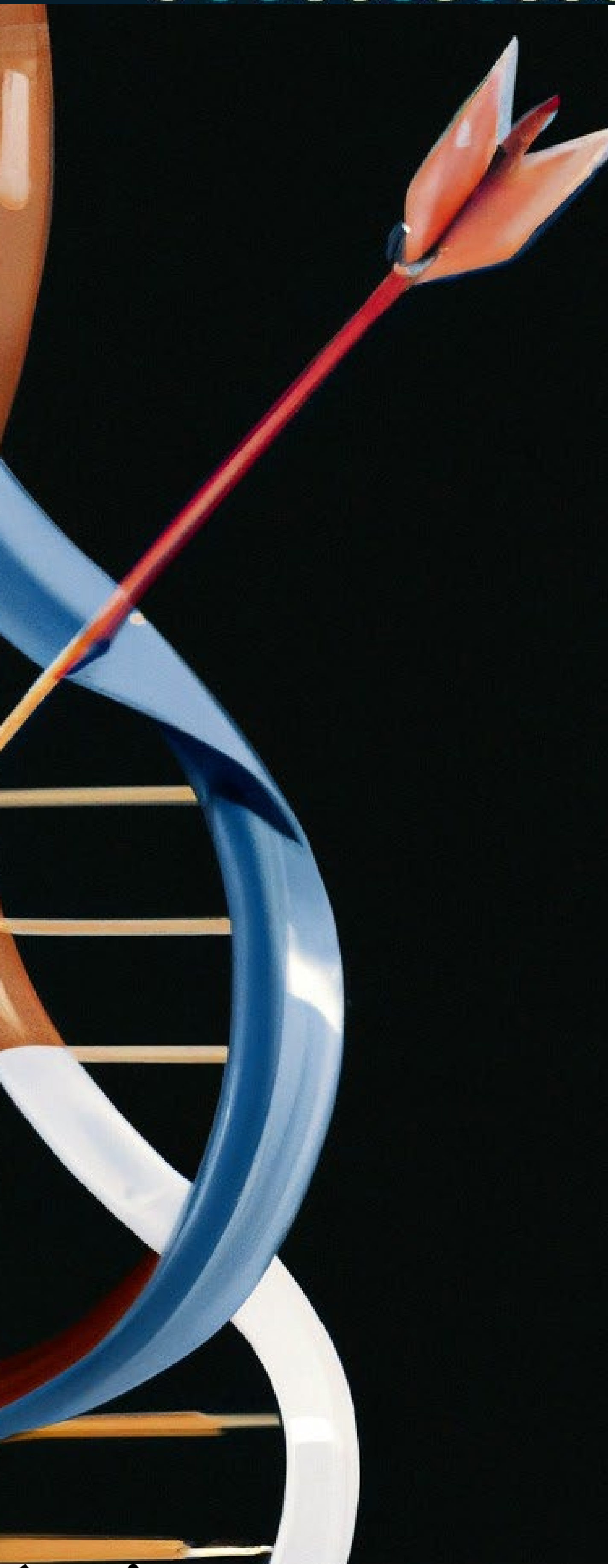
DNA Synthesis
In silico simulation
DNA Sequencing
Encoding
.fasta
Designed Oligonucleotides separated in blocks
Physical block separation
Selective molecular access using PCR
Logical block separation
.fastq
Multi-block read files
Demultiplexing & trimming
.fastq
Single block read file
Complete sequence clustering and consensus
read classification
Residual clustering and consensus
.fasta
Recovered oligonucleotide sequences
Decoding & Error correction
Legend
Digital DNA sequences
Molecular steps
In silico processing

Artemis, the decoding module

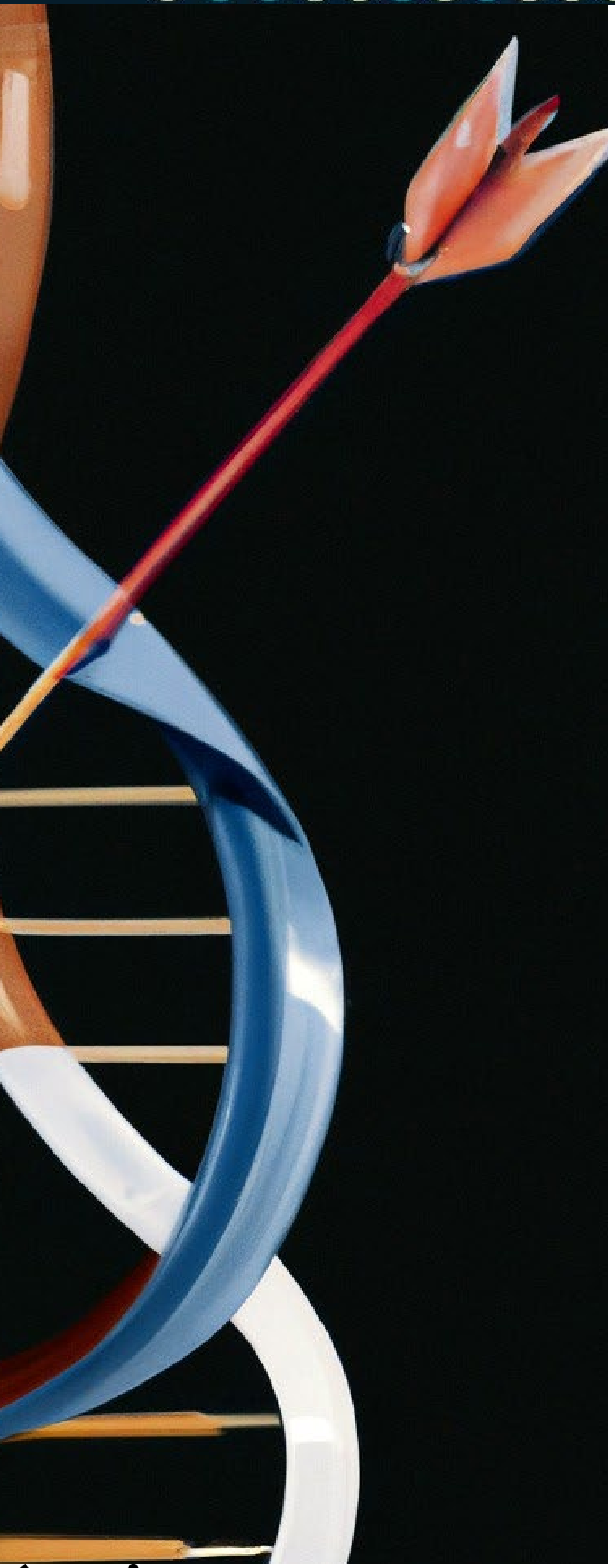
Artemis returning from the hunt - Colombel, 1697

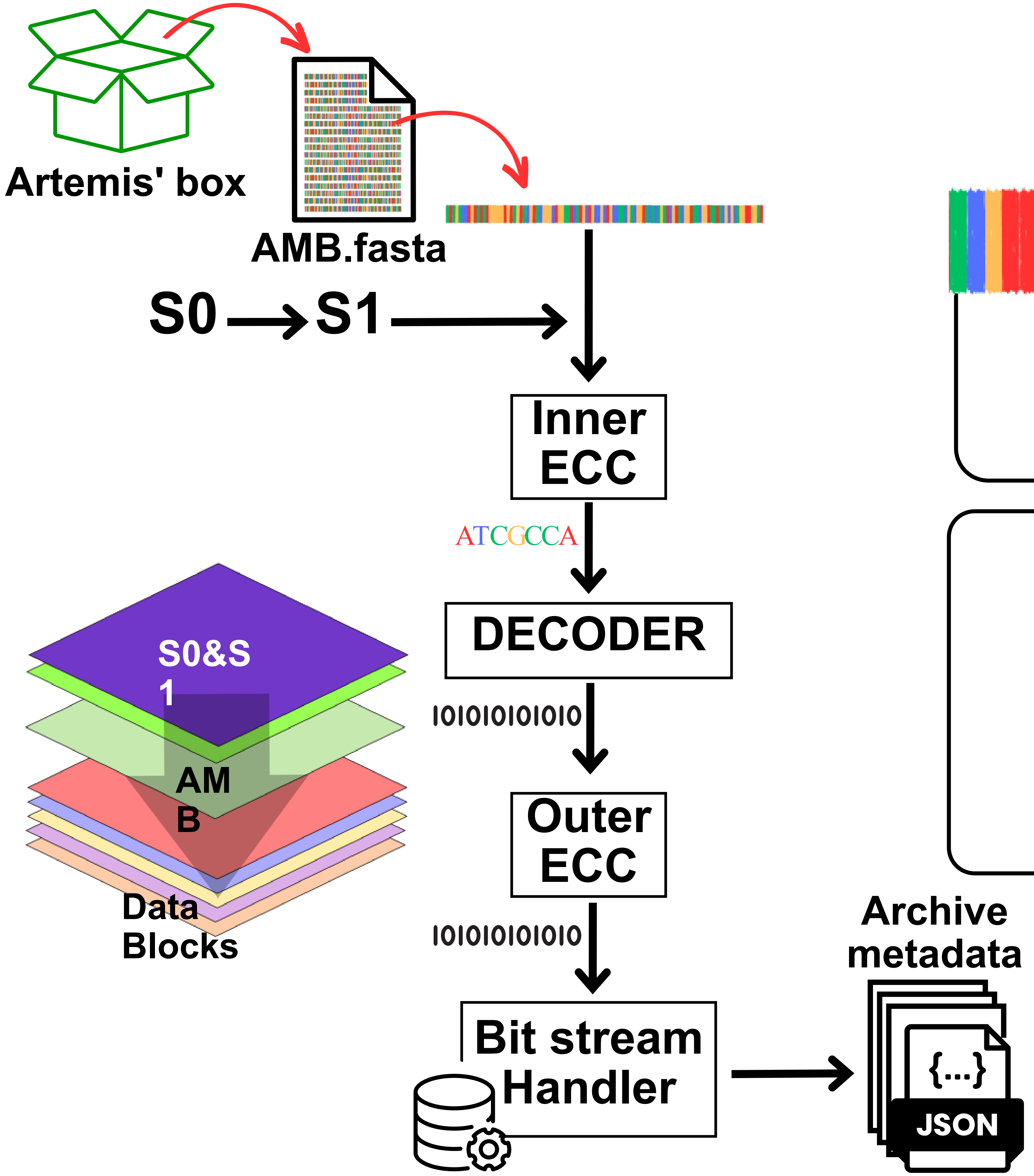
Artemis: decoding the archive metadata block

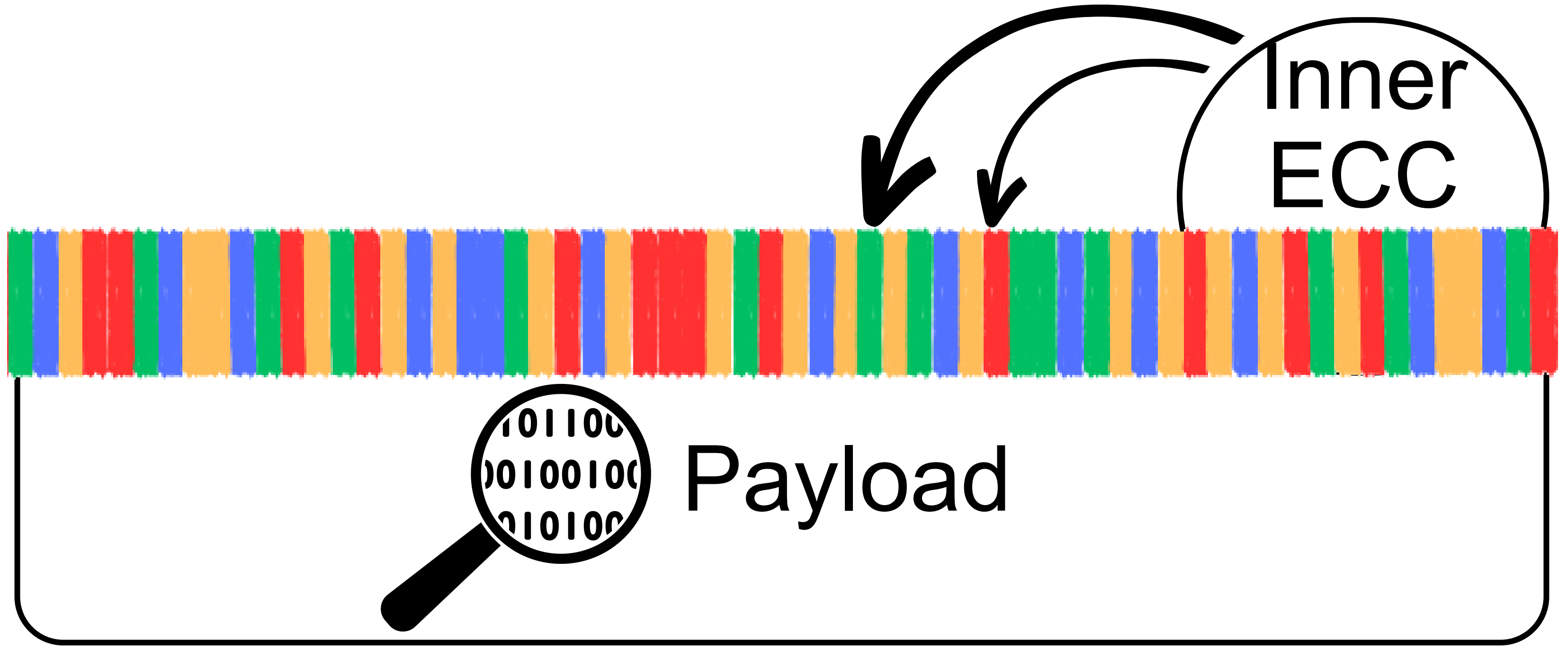


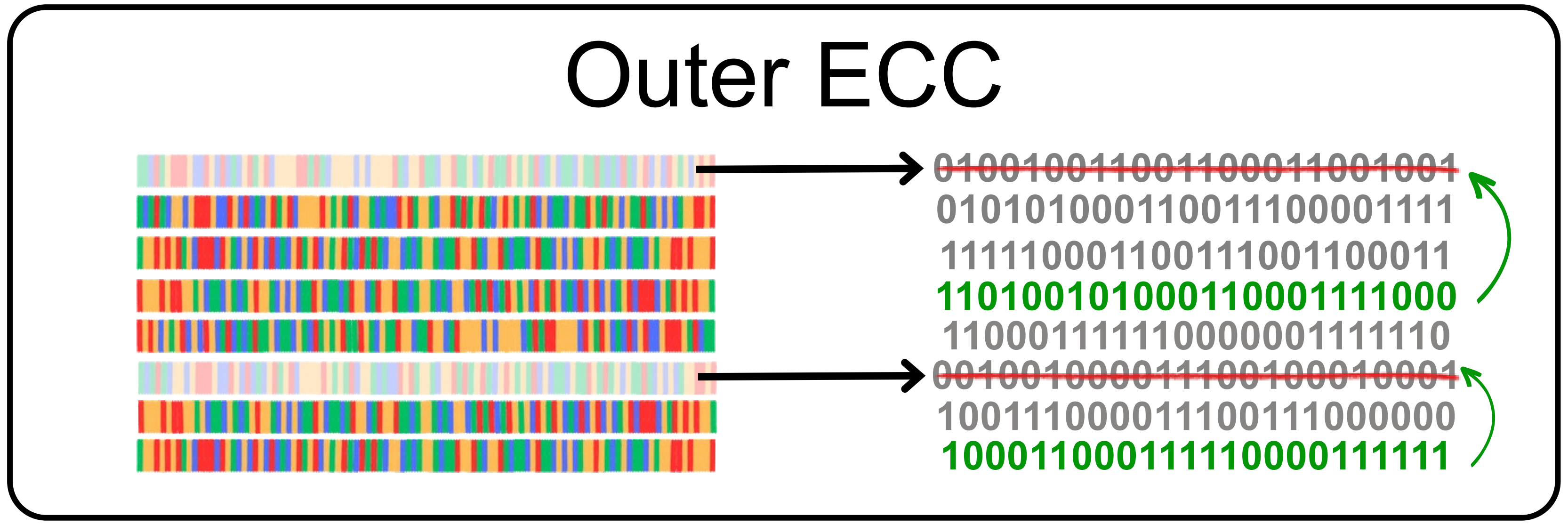


Artemis: applying ECC to correct errors





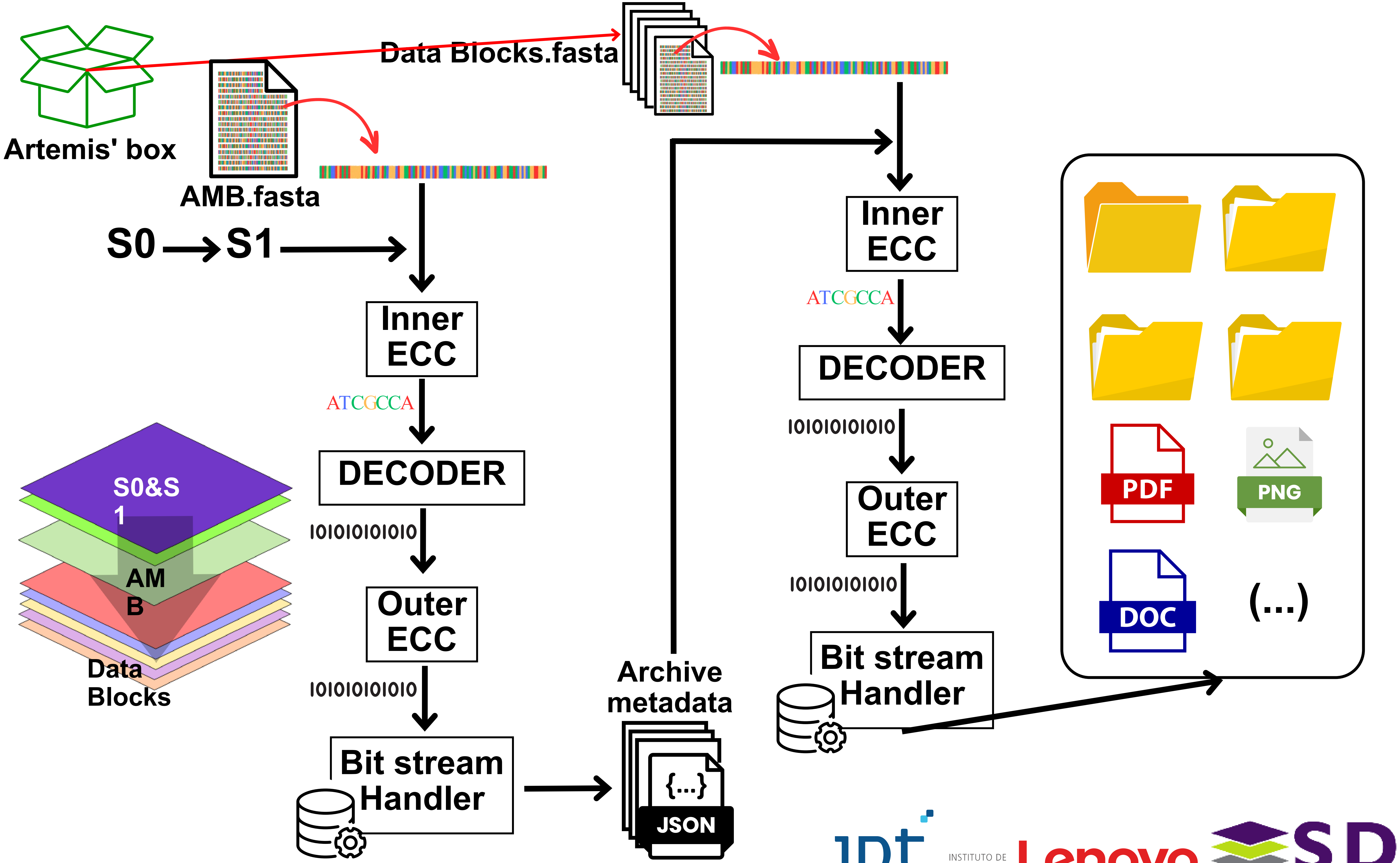




PROMETHEUS

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS Lenovo SDC 23

Artemis: Decoding the data blocks



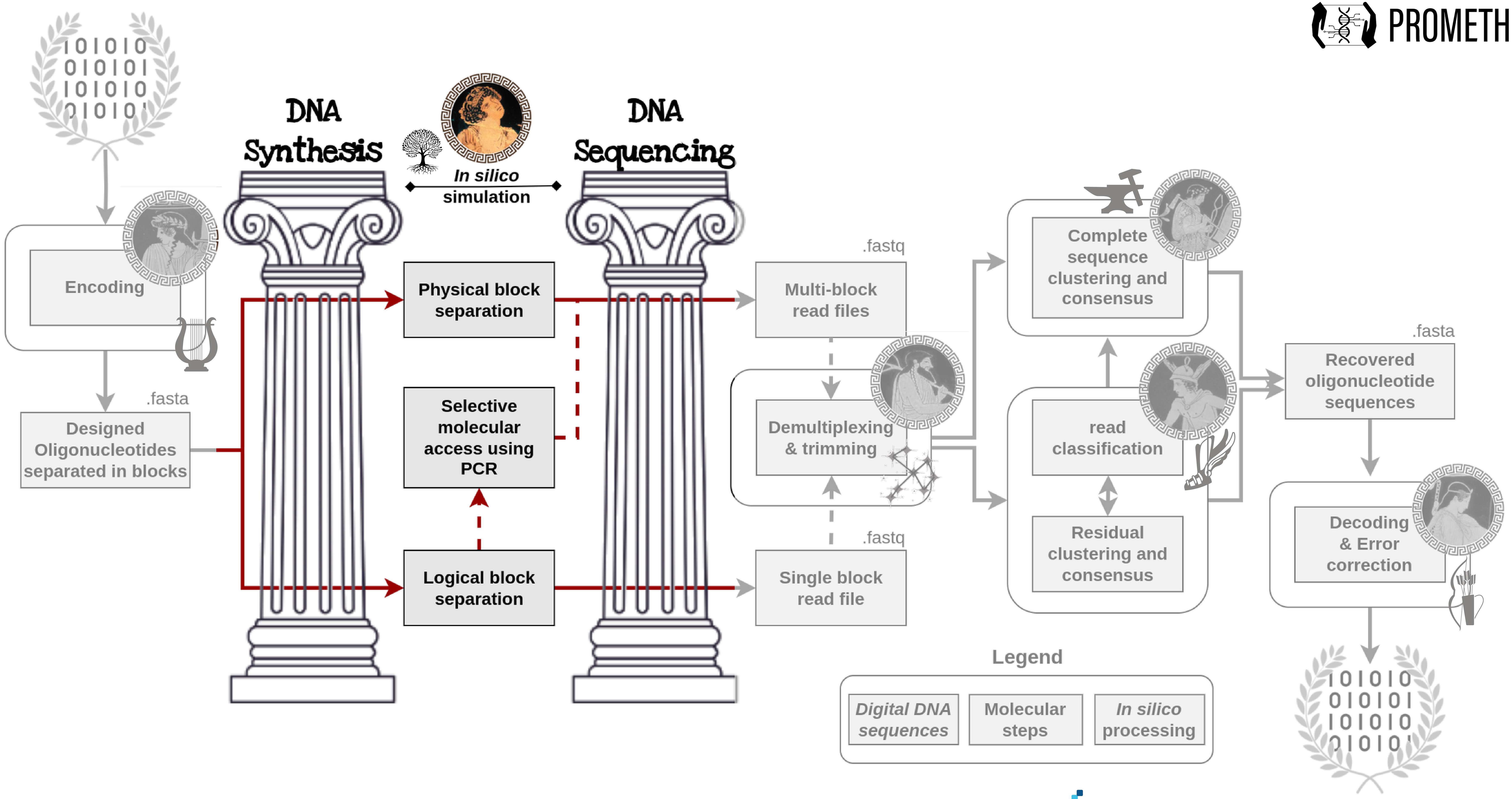
PROMETHEUS

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

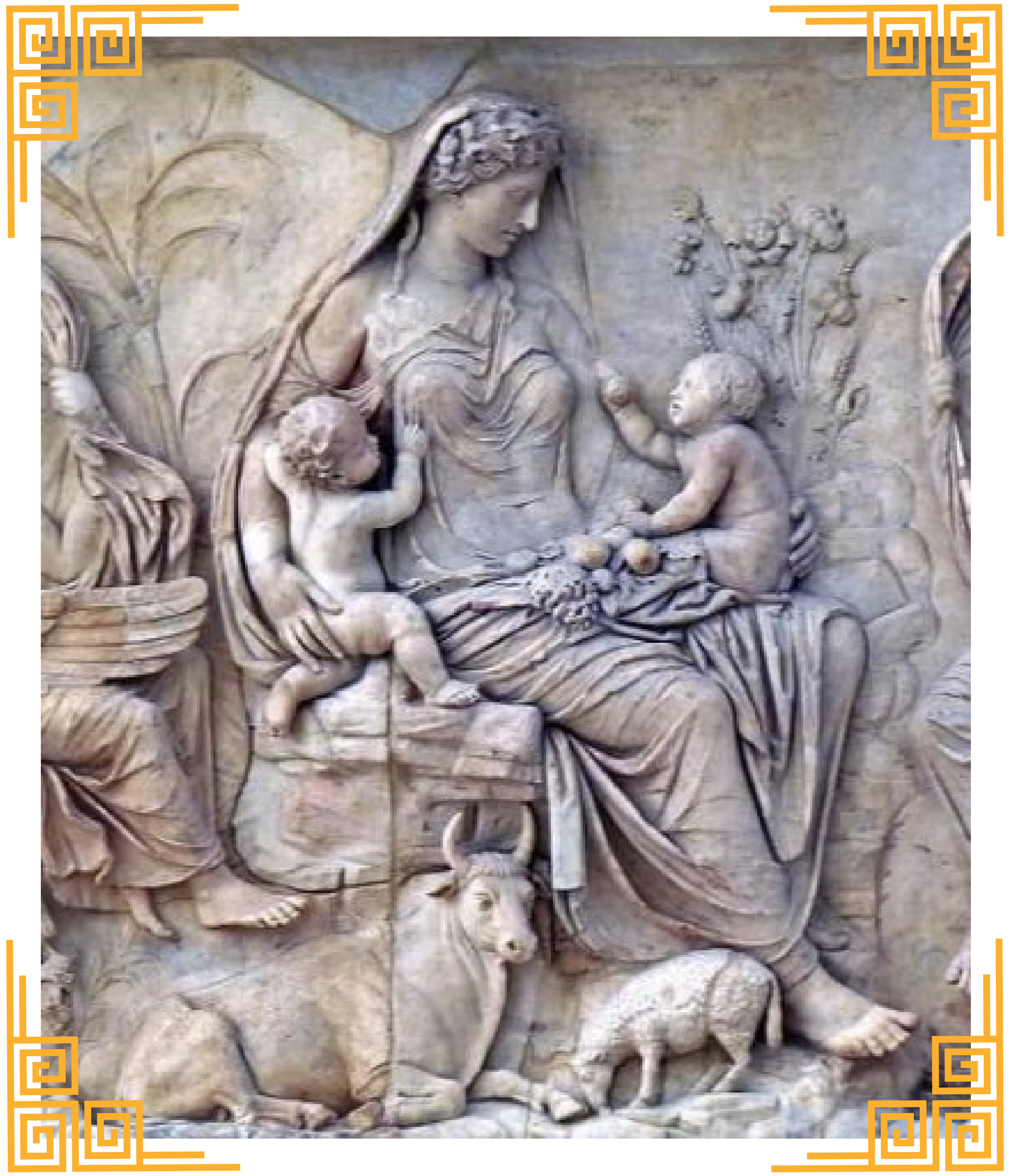
Lenovo

SDC 23

| ©2023 SNIA. All Rights Reserved.


DNA Synthesis
In silico simulation
DNA Sequencing
Encoding
.fasta
Designed Oligonucleotides separated in blocks
Physical block separation
Selective molecular access using PCR
Logical block separation
.fastq
Multi-block read files
Demultiplexing & trimming
.fastq
Single block read file
Complete sequence clustering and consensus
read classification
Residual clustering and consensus
.fasta
Recovered oligonucleotide sequences
Decoding & Error correction
Legend
Digital DNA sequences
Molecular steps
In silico processing

Gaia, the development module



Tellus Mater (Mother Gaia) panel,
Ara Pacis Augustae - Rome, Italy

Gaia: a sandbox module to support development

- What Gaia does:
  - Simulate different sequencing strategies
    - Single or Paired-ends
    - Library preparation
    - Coverage variation
    - Sequencing platforms
  - Simulate different synthesis strategies and biases
    - Pandora's box of bias models

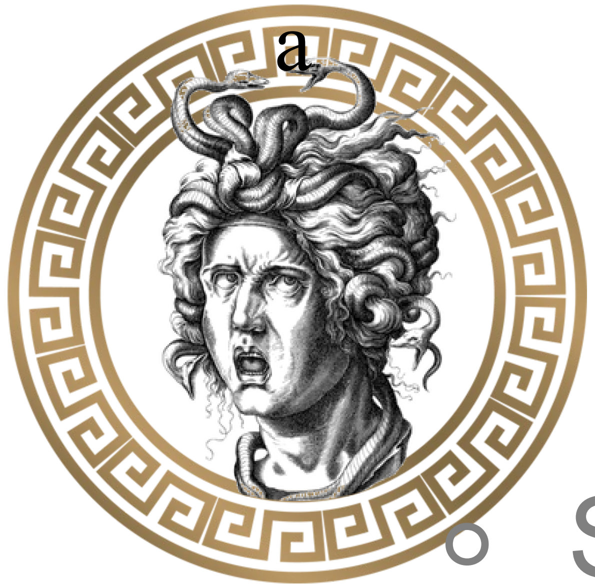
PROMETHEUS

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS Lenovo SDC 23

Gaia: a sandbox module to support development

- What Gaia does:
  - Simulate different sequencing strategies
    - Single or Paired-ends
    - Library preparation
    - Coverage variation
    - Sequencing platforms
  - Simulate different synthesis strategies and biases
    - Pandora's box of bias models

Medusa



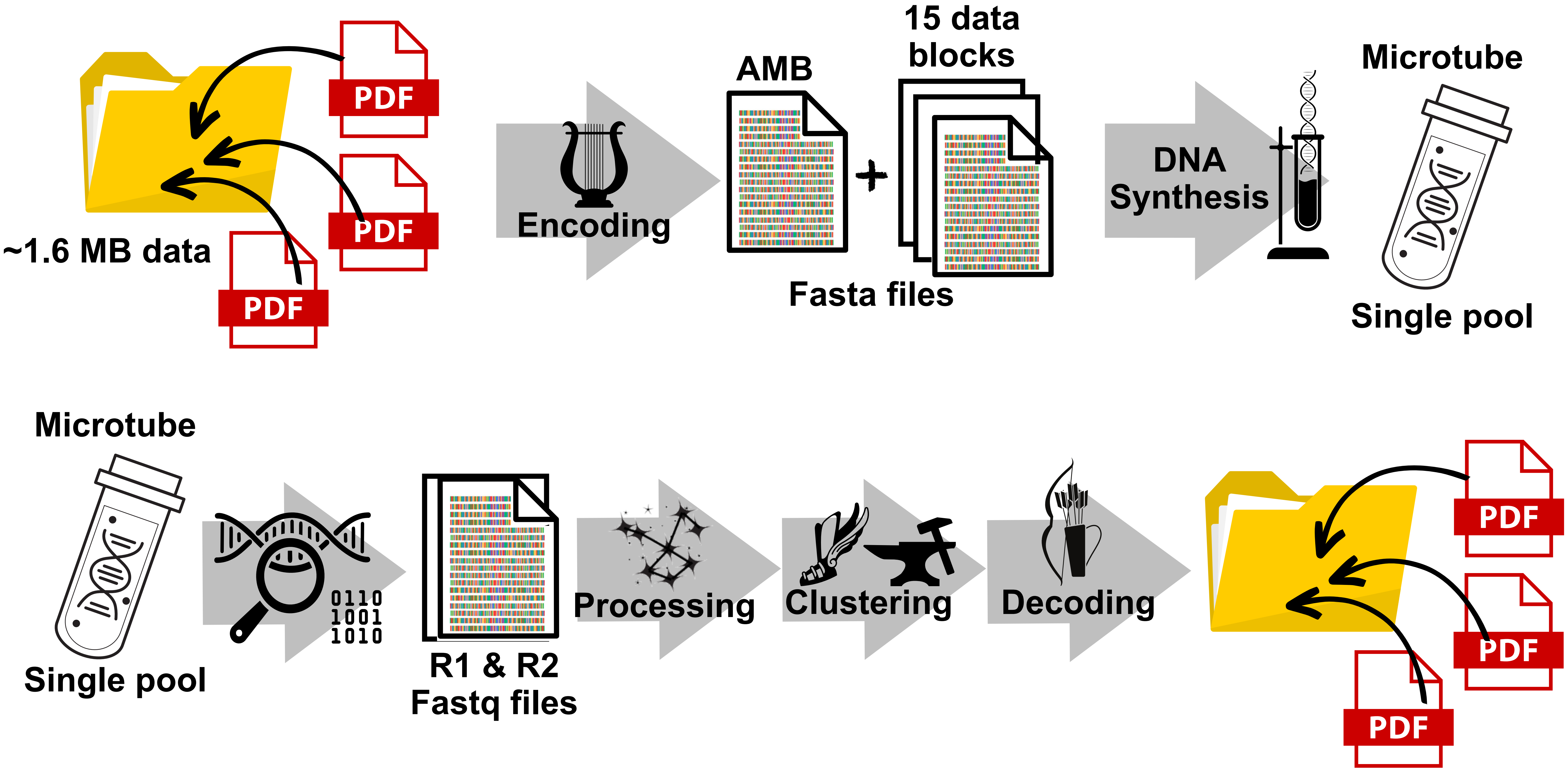
Manticore

PROMETHEUS

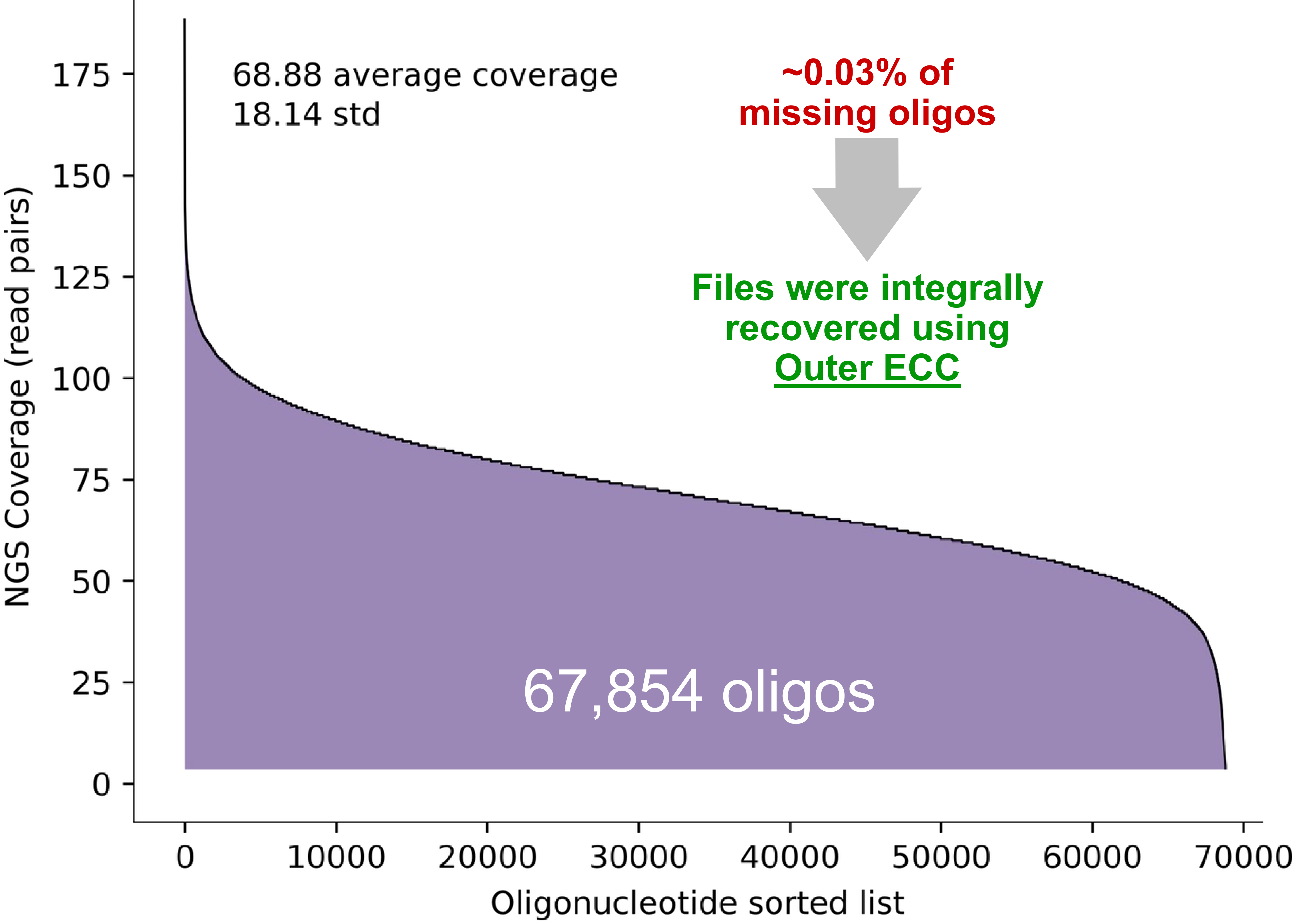
ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS Lenovo SDC 23

Testing Phanteon CODEC with real data

Experiment workflow



Sequencing results



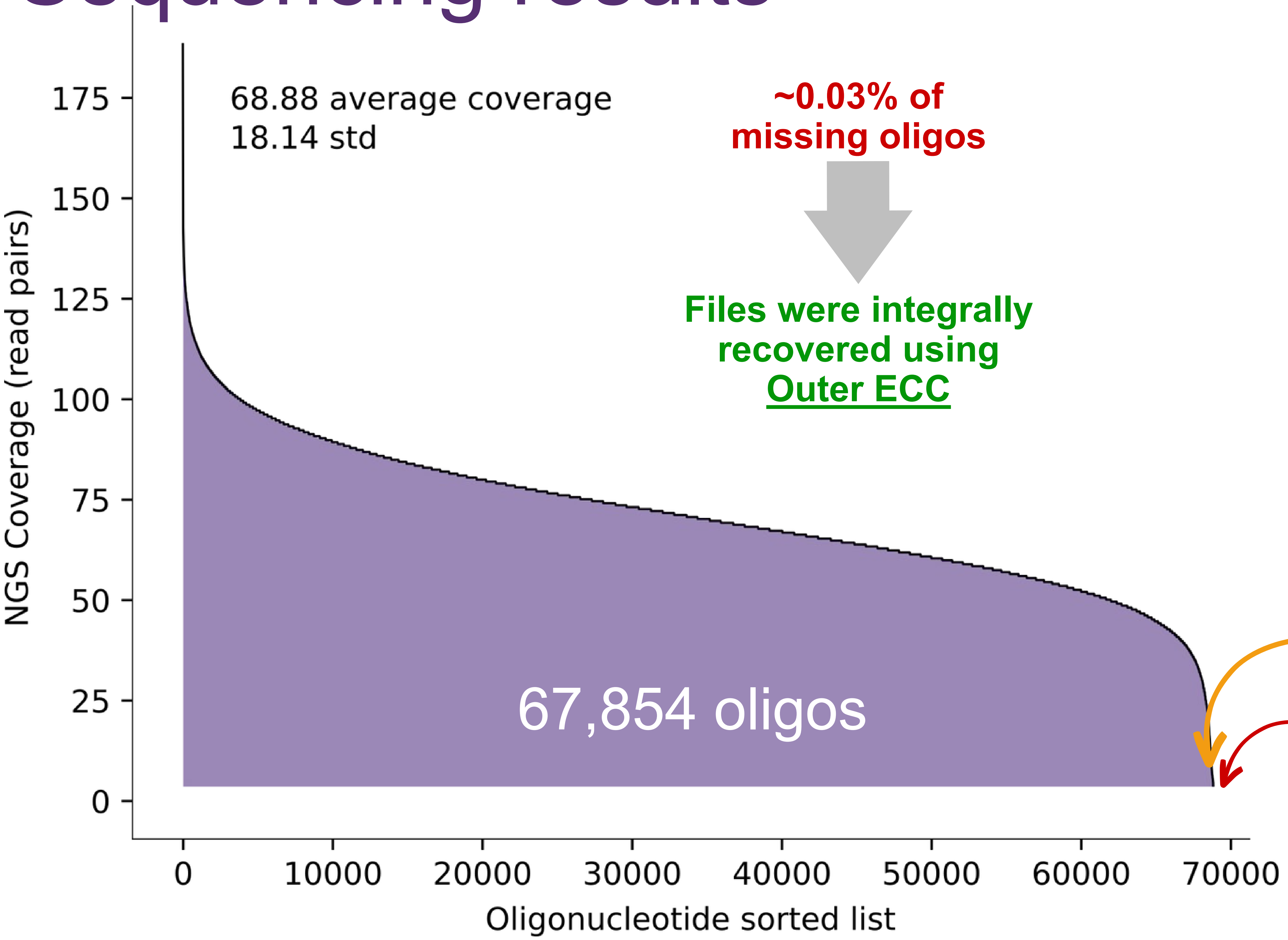
68.88 average coverage
18.14 std

~0.03% of missing oligos

Files were integrally recovered using <u>Outer ECC</u>

67,854 oligos

Sequencing results



~0.03% of missing oligos

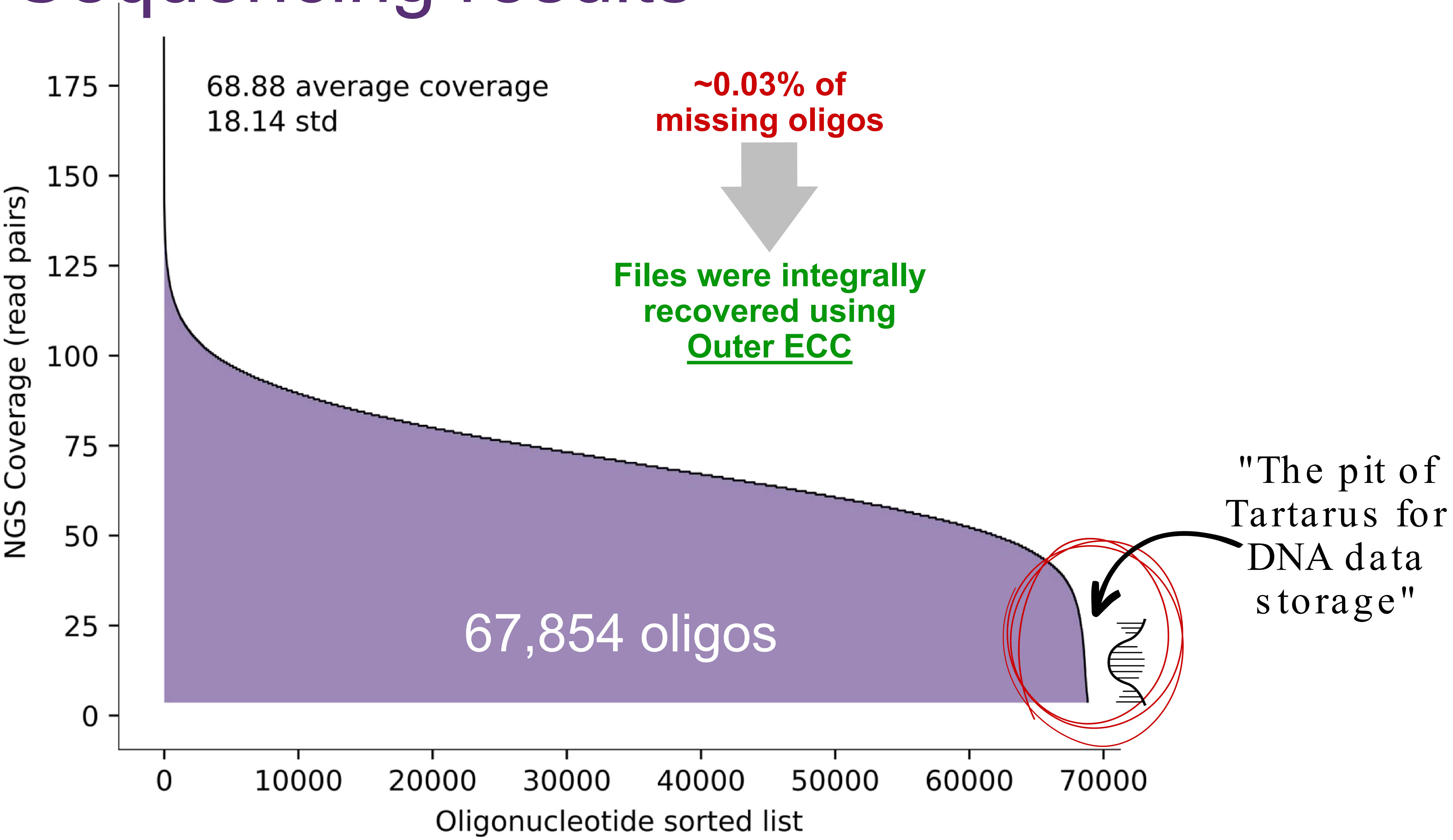
Files were integrally recovered using Outer ECC

162 oligos were low coveraged (≤ 5)

22 oligos were missing (coverage = 0)

PROMETHEUS

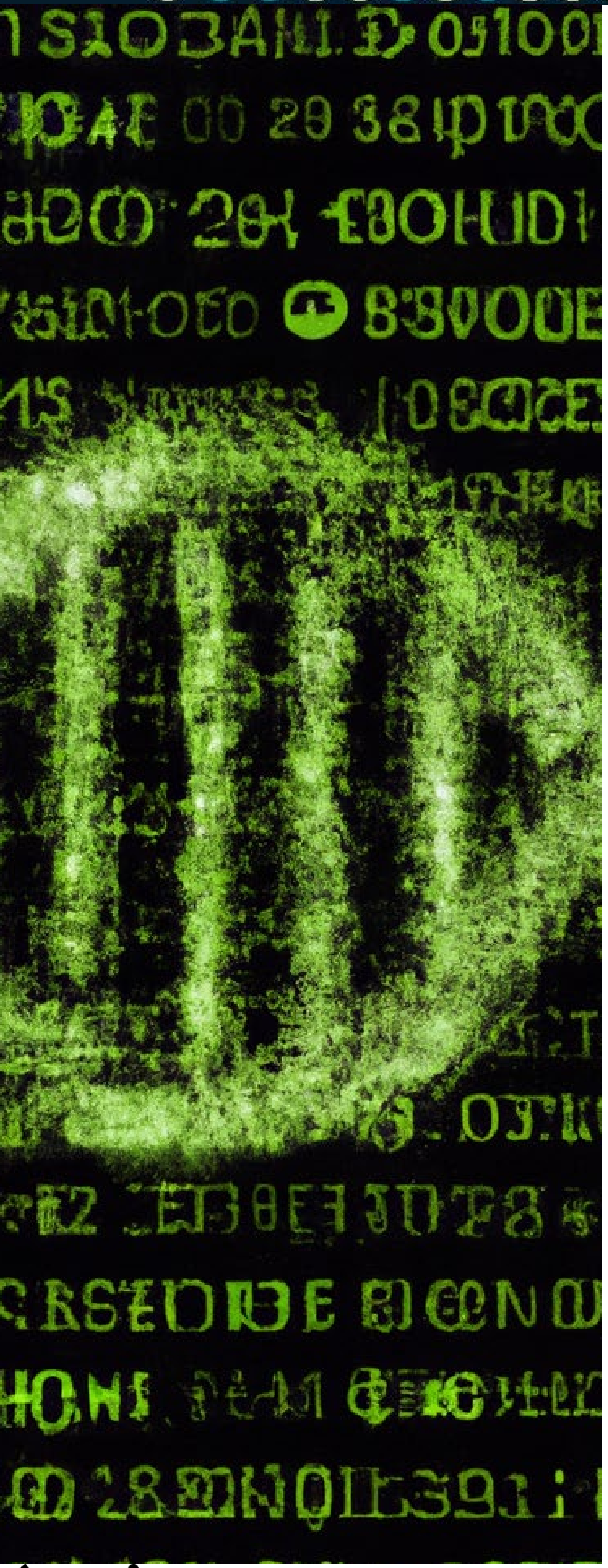
Sequencing results



~0.03% of missing oligos

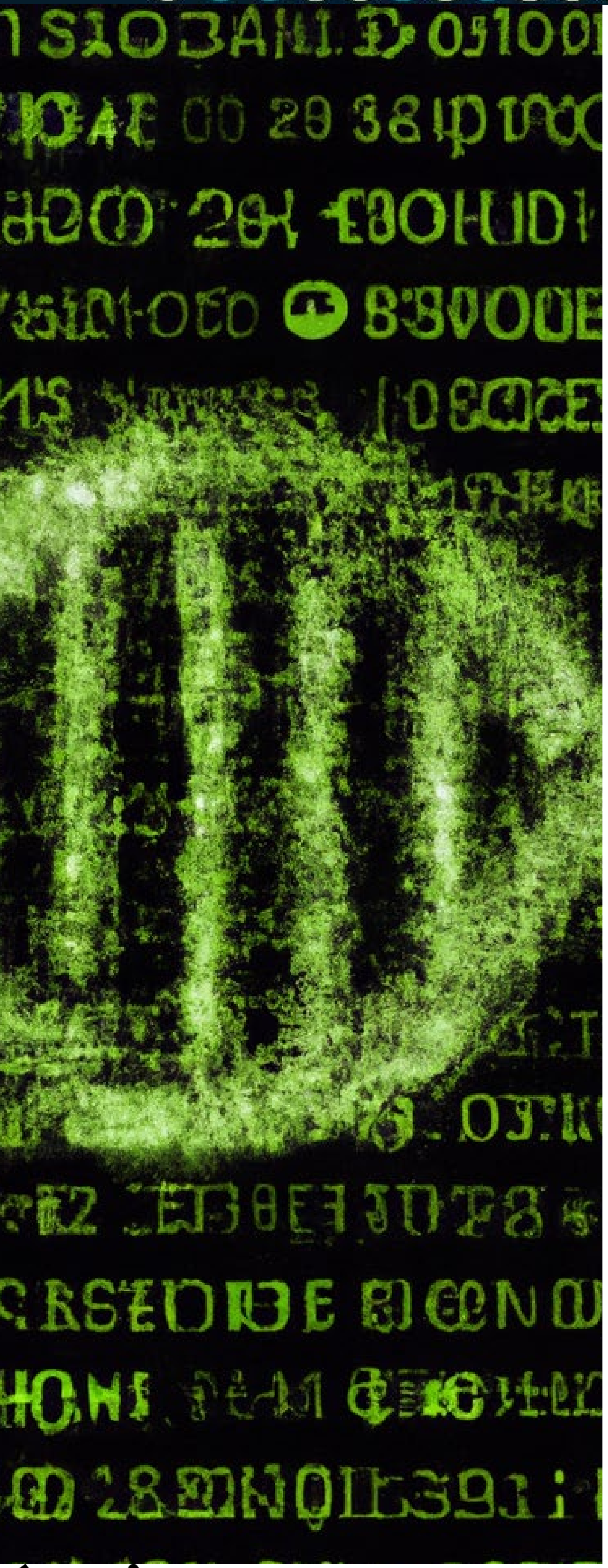
Files were integrally recovered using Outer ECC

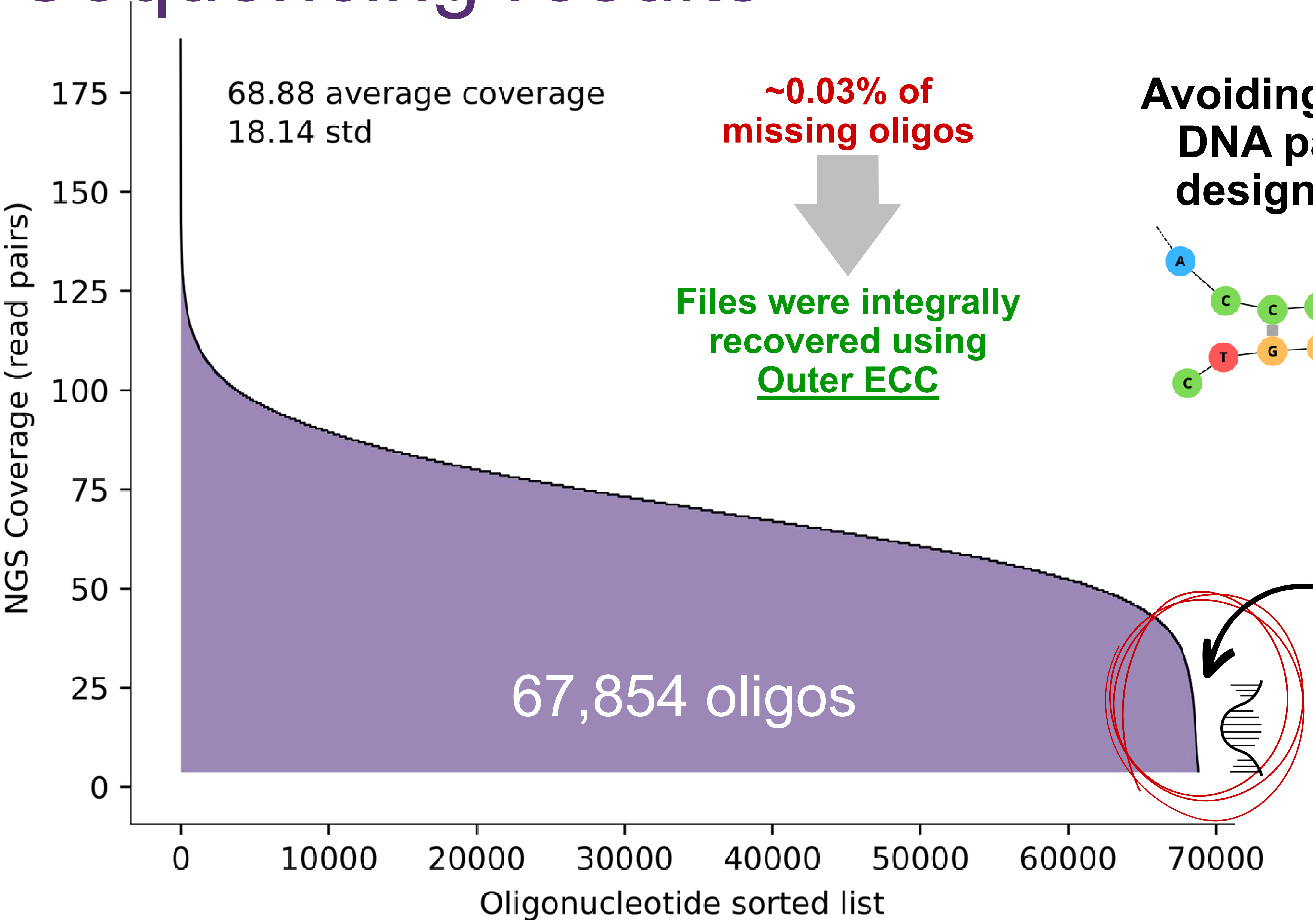
"The pit of Tartarus for DNA data storage"



PROMETHEUS

Sequencing results

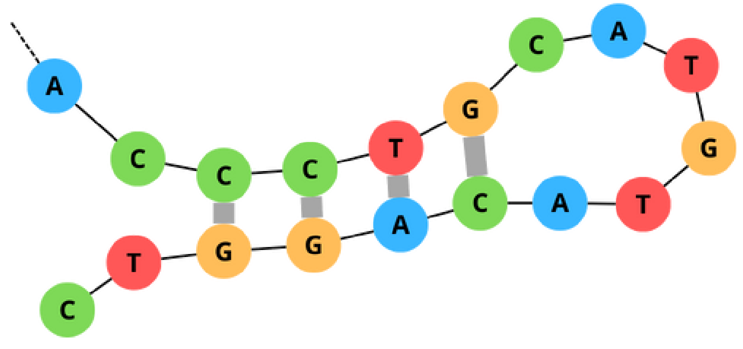




~0.03% of missing oligos

Files were integrally recovered using Outer ECC

Avoiding unwanted DNA patterns in designed oligos



"The pit of Tartarus for DNA data storage"

CODEC performance

HH:MM:SS.SSS

00:00:31.490 (AMB + DB)

Apollo

00:00:00.200 AMB

Hermes

00:01:30.730 AMB

Chiron

00:00:07.800 AMB

Artemis

00:00:01.700 AMB

Hephaestus

4 fastq files with 9.3M reads in total
A single CPU core was used
AMB (Archive metadata block)

PROMETHEUS

CODEC performance

HH:MM:SS.SSS

Apollo

00:00:31.490 (AMB + DB)

Hermes

00:00:00.590 DB

Chiron

00:15:42.000 DB

00:00:58.350 DB (fast mode)*

*Increases the sequence loss

Artemis

00:00:21.040 DB

Hephaestus

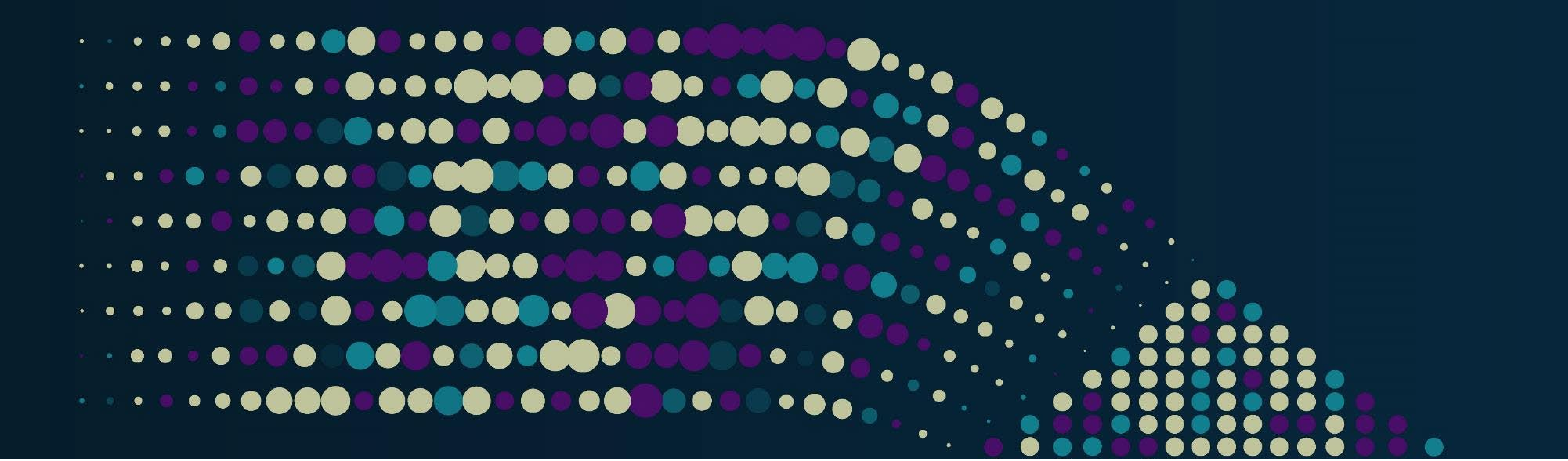
00:01:00.000 DB

09:11:28.000 DB**

**no prior demultiplexing step

4 fastq files with 9.3M reads in total
A single CPU core was used
DB (Data Blocks)

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS Lenovo SDC 23



Please take a moment to rate this session.

Your feedback is important to us.

PROMETHEUS

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS Lenovo SDC 23