INSTITUTO DE PESQUISAS TECNOLÓGICAS

COMUNICAÇÃO TÉCNICA

Nº 179057

Challenges of encoding digital data into DNA sequences and decording it back to digital

Andre Guilherme da Costa Martins

Palestra apresentada na disciplina de Biologia Molecular do Programa de Pós-graduação Multicêntrico em Bioquímica e Biologia Molecular, do Campus Centro-Oeste Dona Lindu-UFSJ

Instituto de Pesquisas Tecnológicas do Estado de São Paulo S/A - IPT
Av. Prof. Almeida Prado, 532 | Cidade Universitária ou
Caixa Postal 0141 | CEP 01064-970
São Paulo | SP | Brasil | CEP 05508-901
Tel 11 3767 4374/4000 | Fax 11 3767-4099

www.ipt.br

Challenges of encoding digital data into DNA sequences and decoding it back to digital

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

PROMETHEUS

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

What is a CODEC?

PROMETHEUS

Storaged DNA molecules

DNA Synthesis

DNA Sequencing

Original Binary data

ATGCTAGCA
AGGCGTGCA
Encoded DNA

ATGCTAGCA
AGGCGTGCA
Sequenced DNA

Recovered Binary data

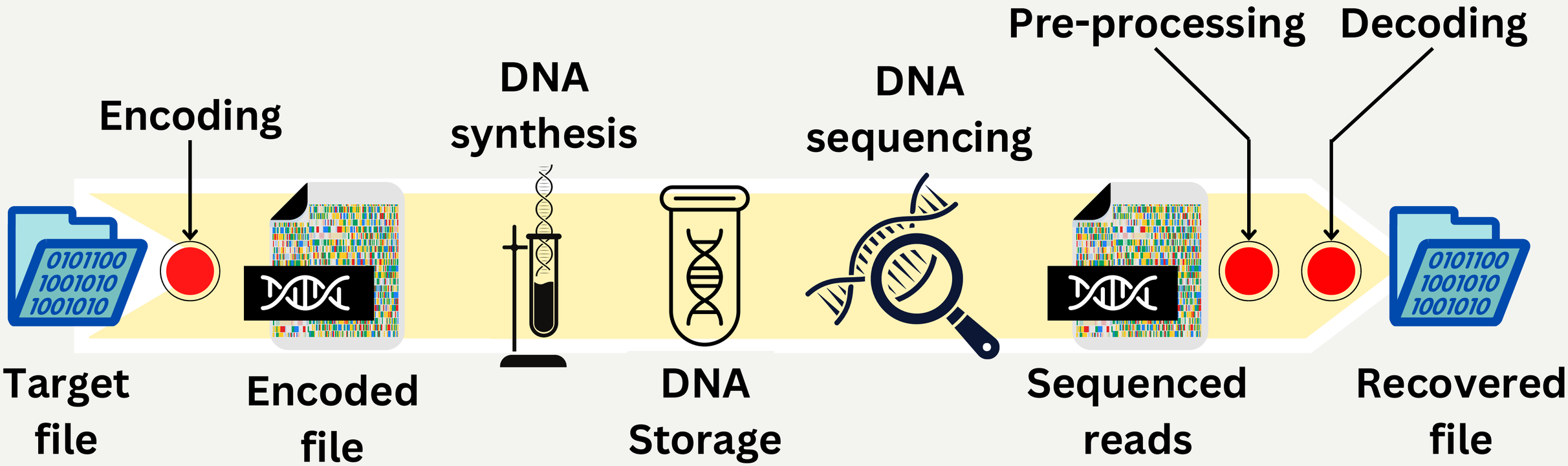
ENCODING

DECODING

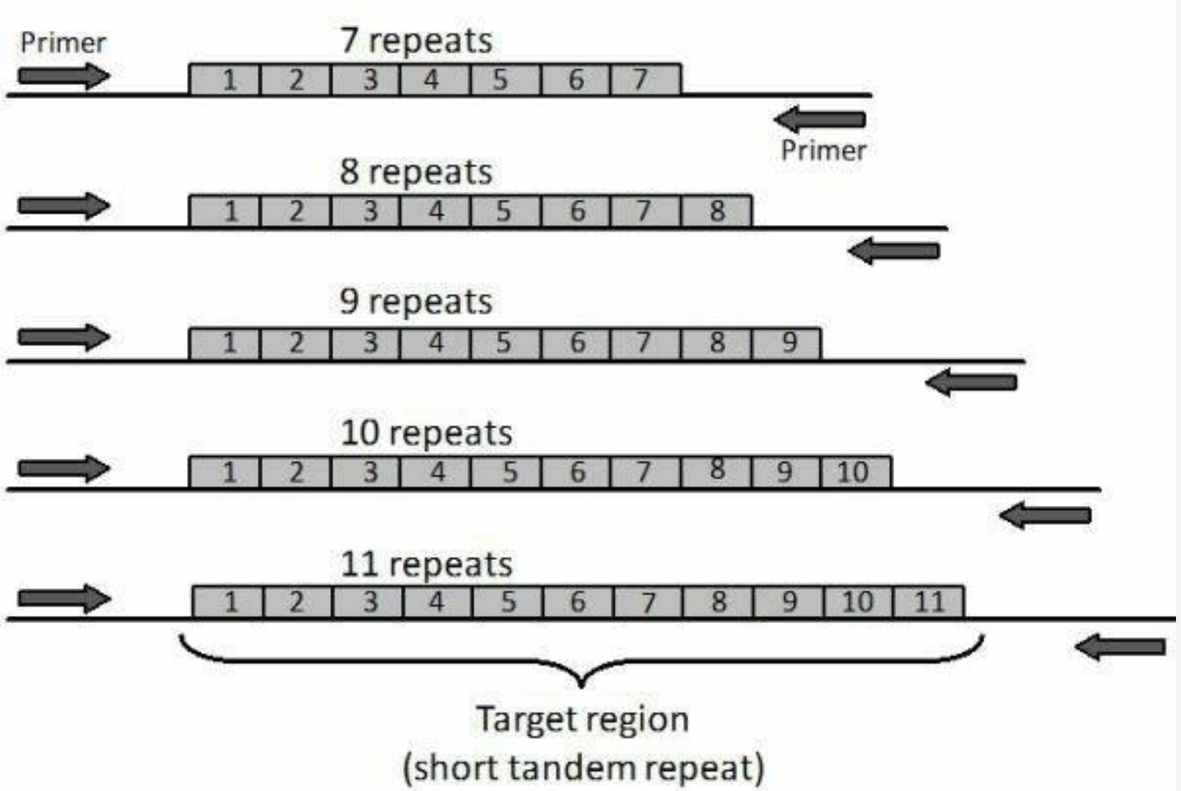
ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

Lenovo

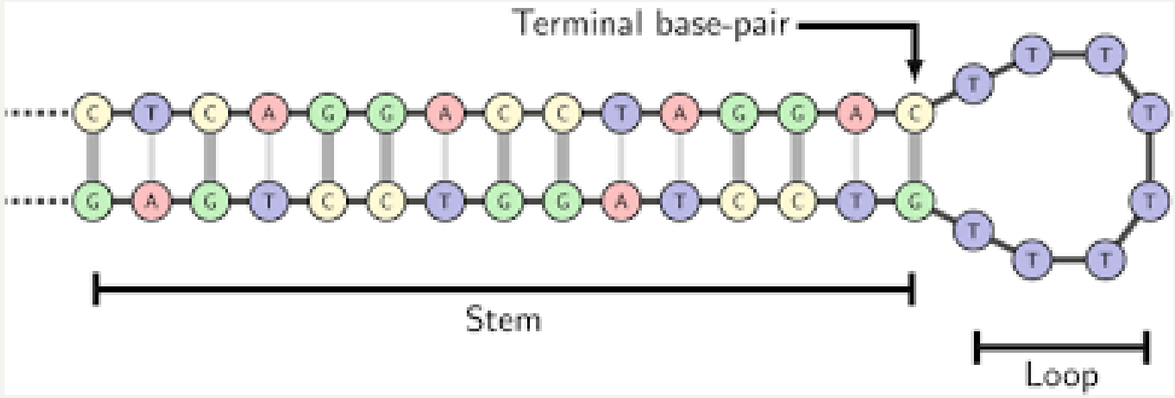
The workflow



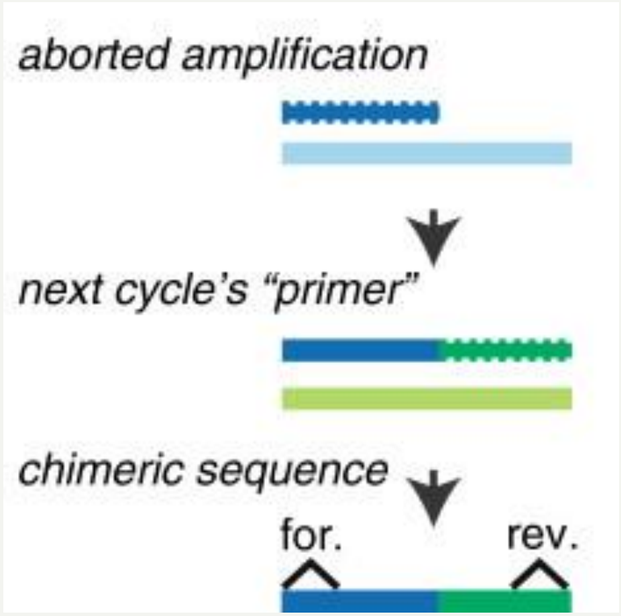
Challenges of encoding



Repetitive segments



DNA hairpins



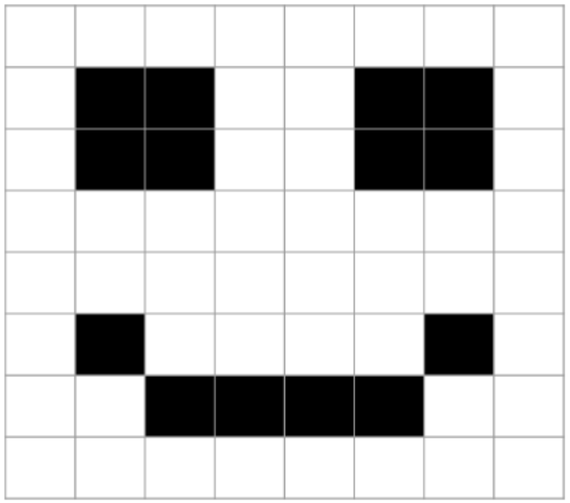
Interspecific similarity

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

Challenges of encoding

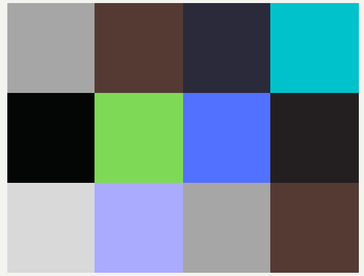
IPT - Main gate

Bitmap image



0	0	0	0	0	0	0	0
0	1	1	0	0	1	1	0
0	1	1	0	0	1	1	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	1	0	0	0	0	1	0
0	0	1	1	1	1	0	0
0	0	0	0	0	0	0	0

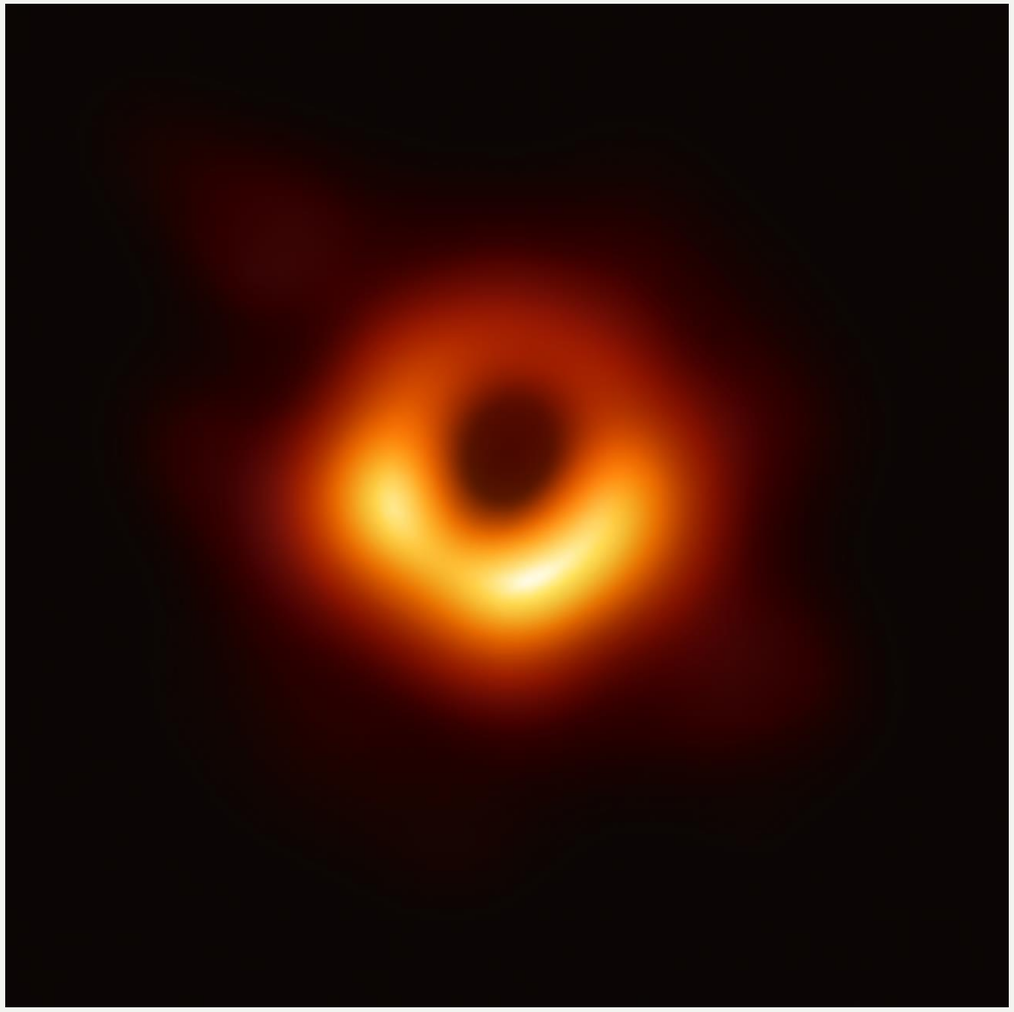
High complexity regions



Low complexity regions

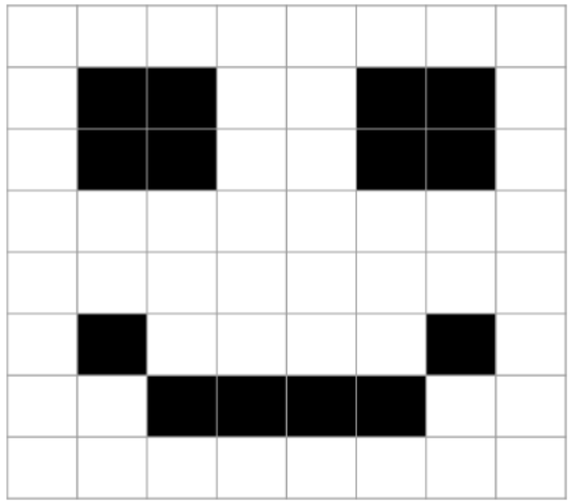
ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

Challenges of encoding



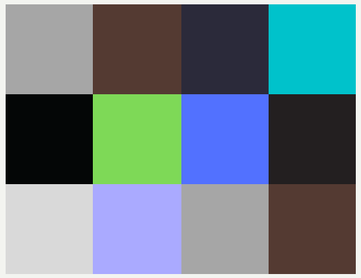
M87* black hole, taken on 11 April 2017

Bitmap image



0	0	0	0	0	0	0	0
0	1	1	0	0	1	1	0
0	1	1	0	0	1	1	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	1	0	0	0	0	1	0
0	0	1	1	1	1	0	0
0	0	0	0	0	0	0	0

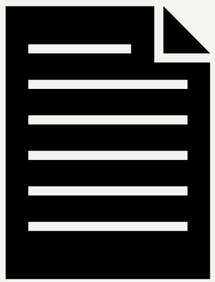
High complexity regions



Low complexity regions

Challenges of processing NGS reads

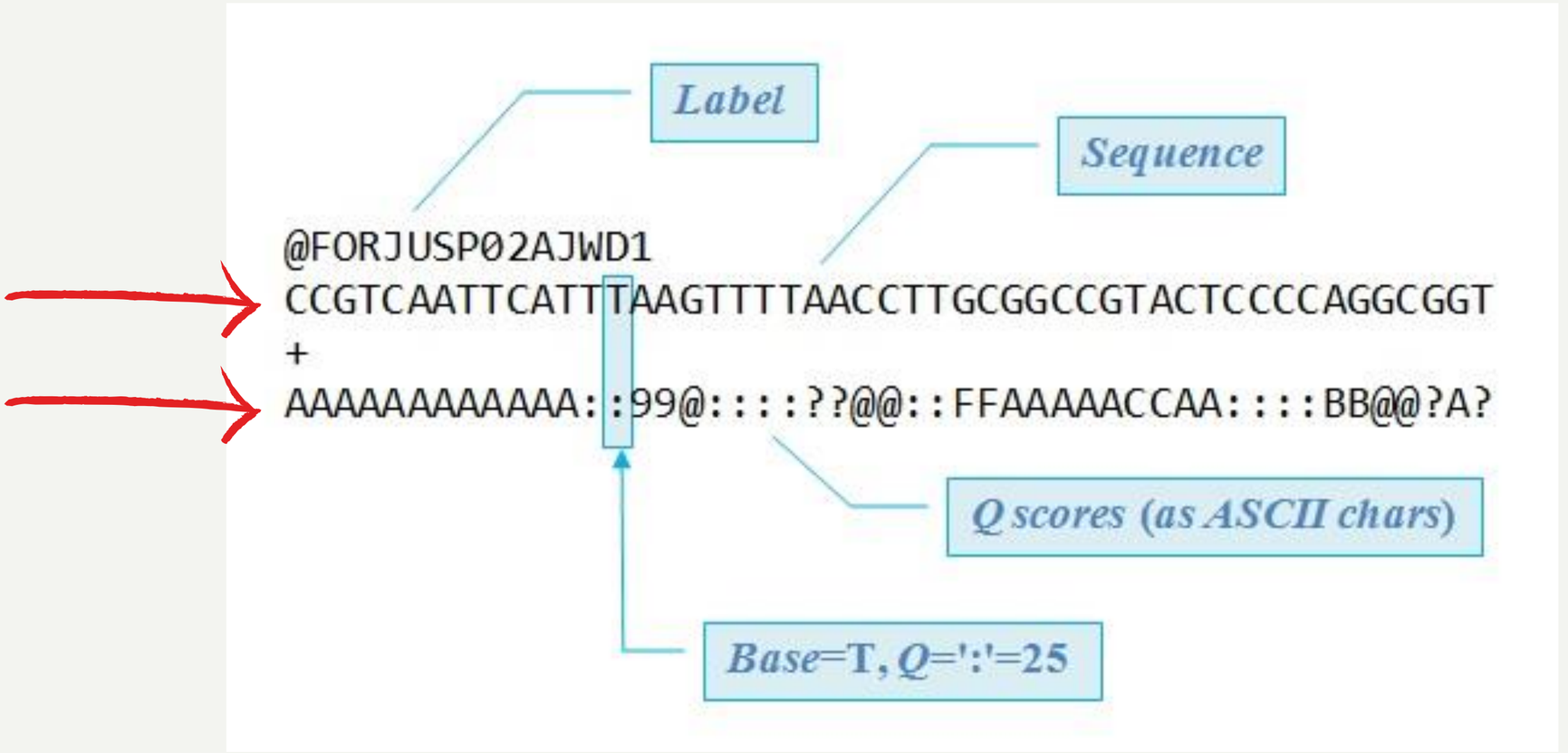
Fastq



DNA

~~Pasequence~~

quality



Challenges of processing NGS reads

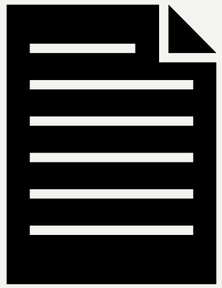
Analogy!

Encoding complexity

Fasta

Decoding complexity

Fastq

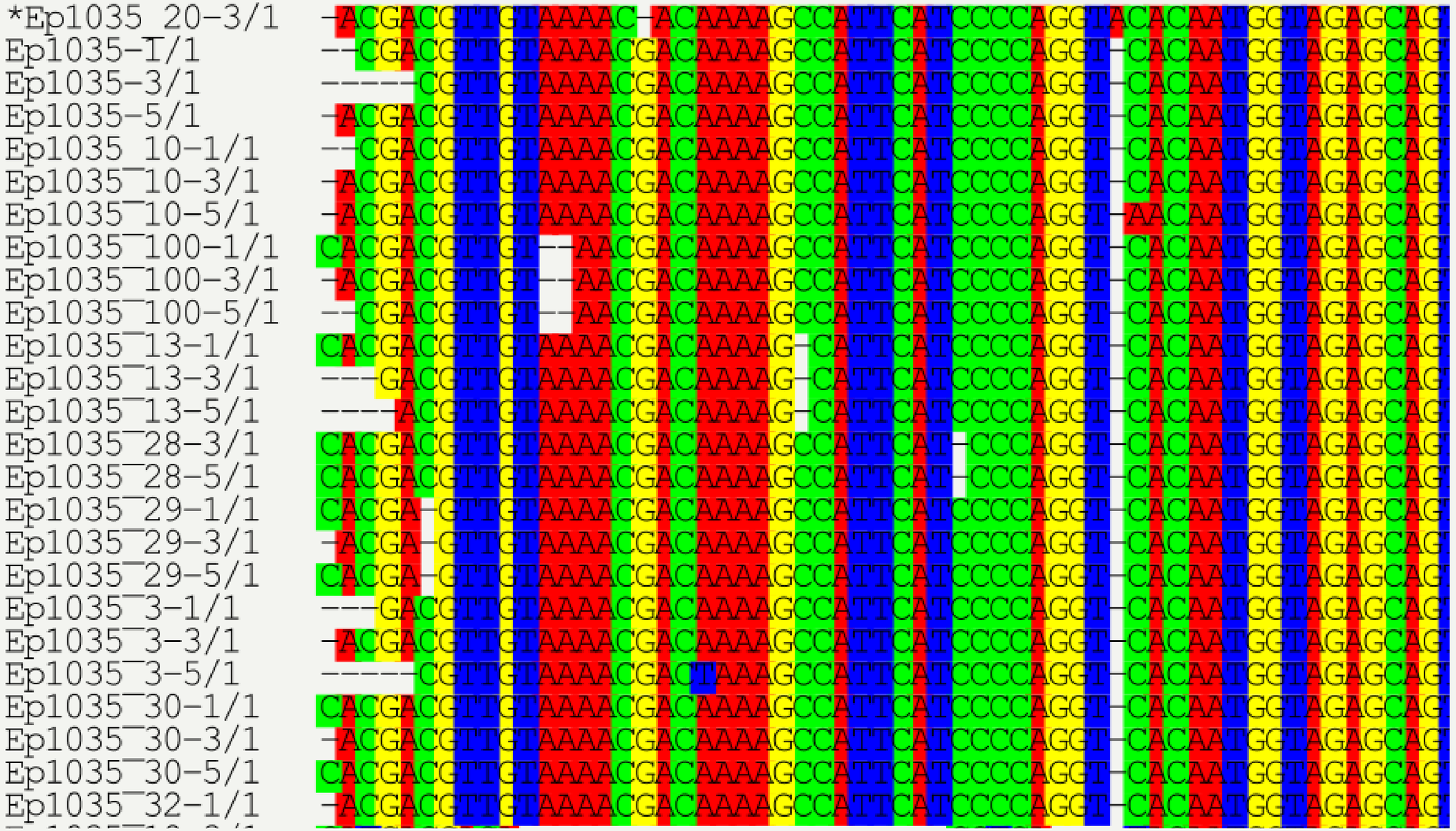


ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

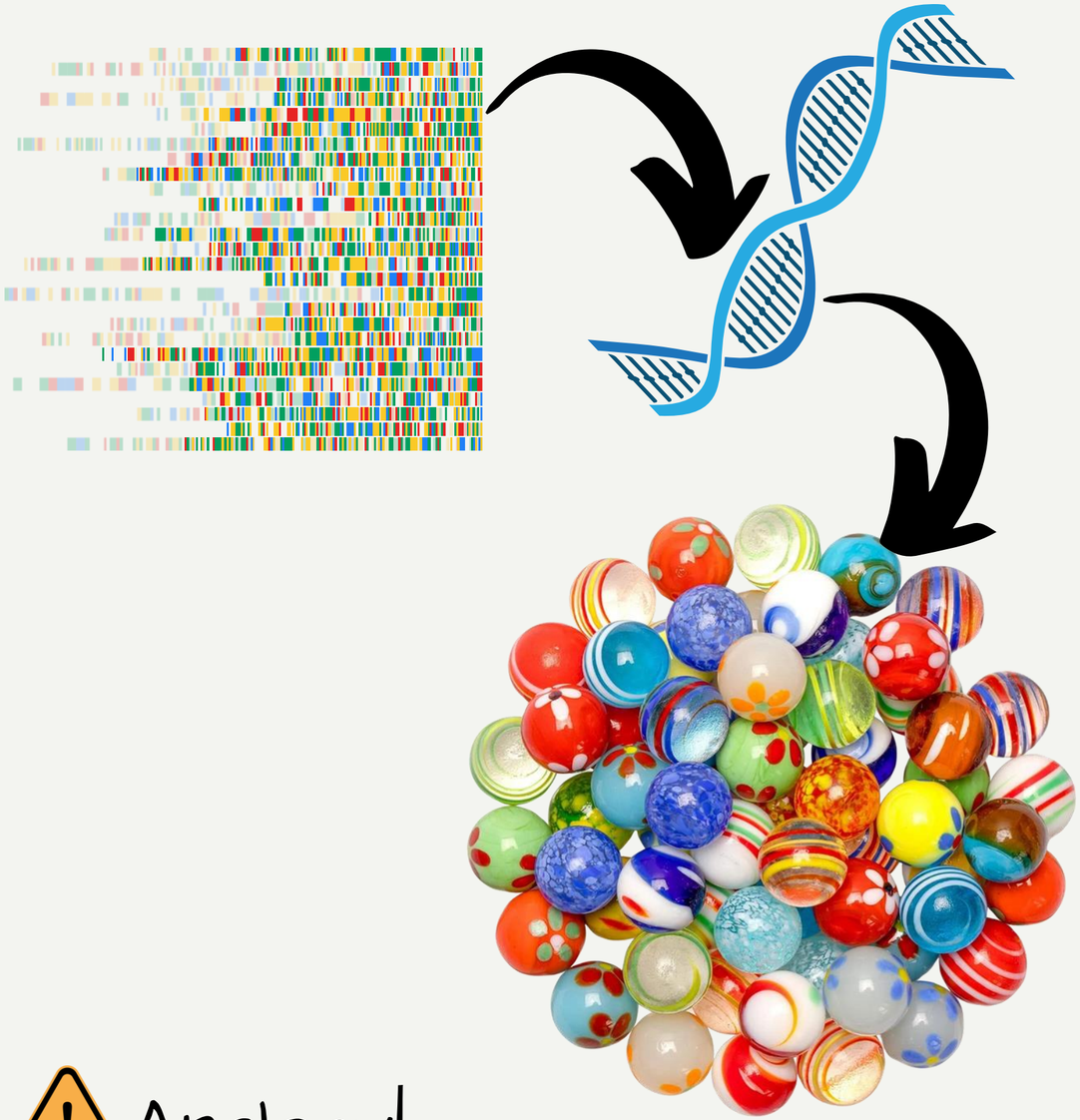
Challenges of processing NGS reads

Analogy!

Challenges of processing NGS reads

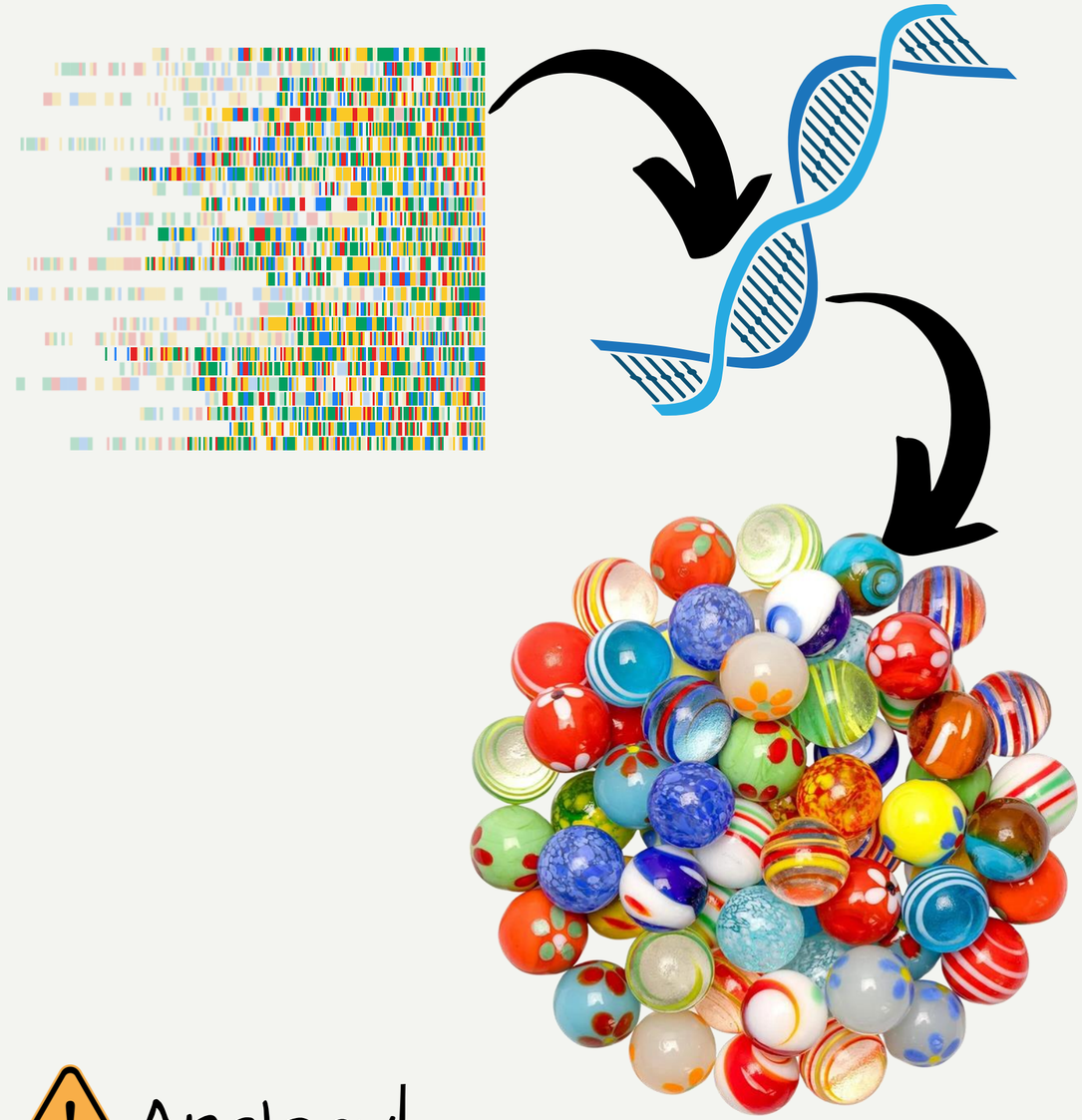


Challenges of decoding

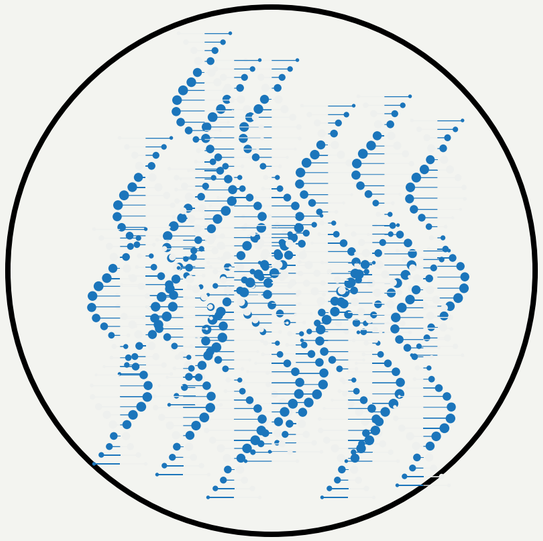


Analogy!

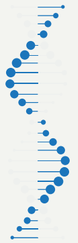
Challenges of decoding



Each DNA synthesis site



10^n copies of same ssDNA



PCR amplification

dsDNA

Analogy!

Challenges of decoding

Analogy!

Challenges of decoding

Library preparation for NGS

Illumina Nextseq 2000

Analogy!

Challenges of decoding

Expectations... :)

Challenges of decoding

Reality... :(

Developed tools

PROMETHEUS

apollo

artemis

chiron

hermes

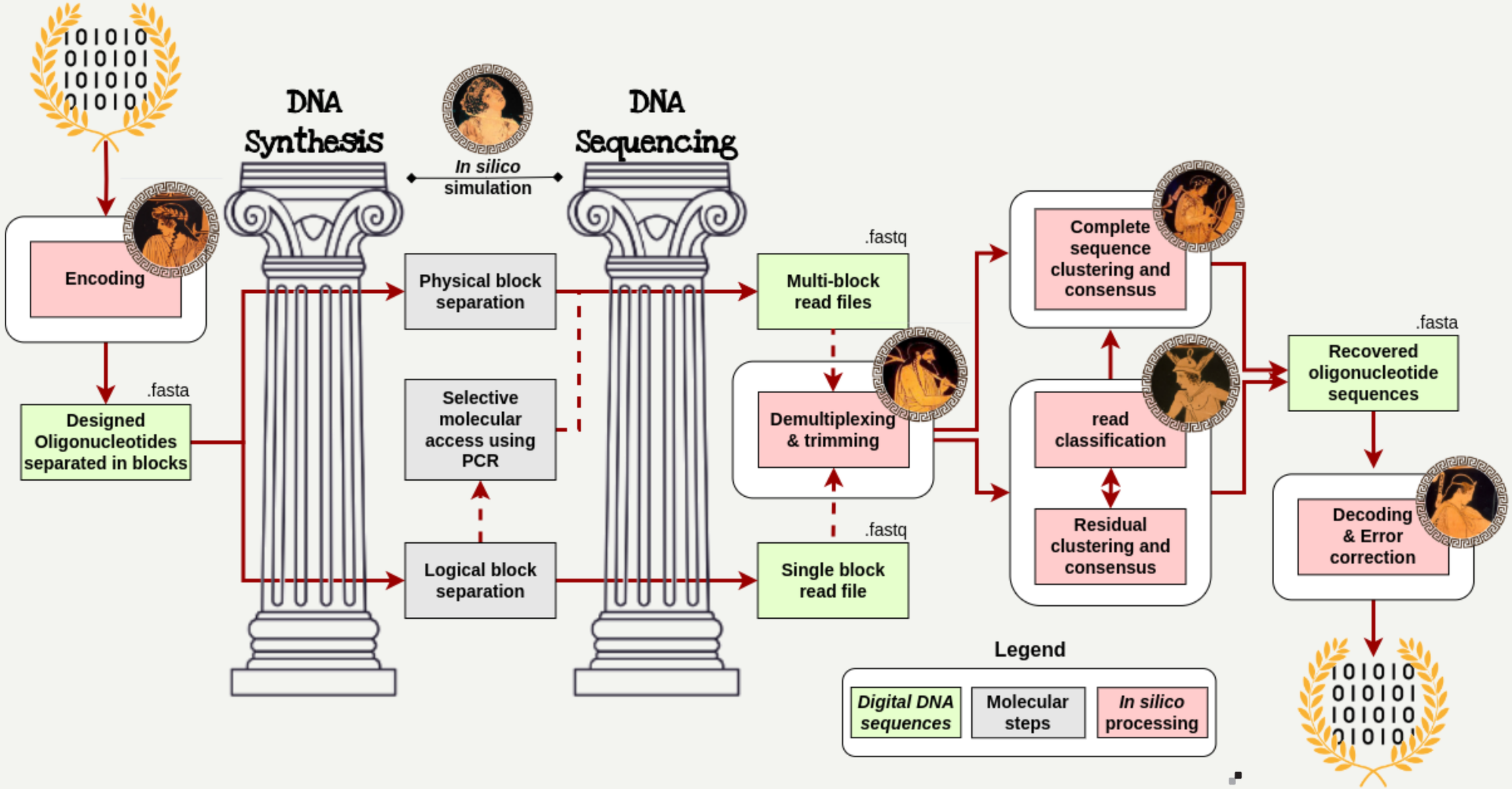
hephaestus

gaia

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

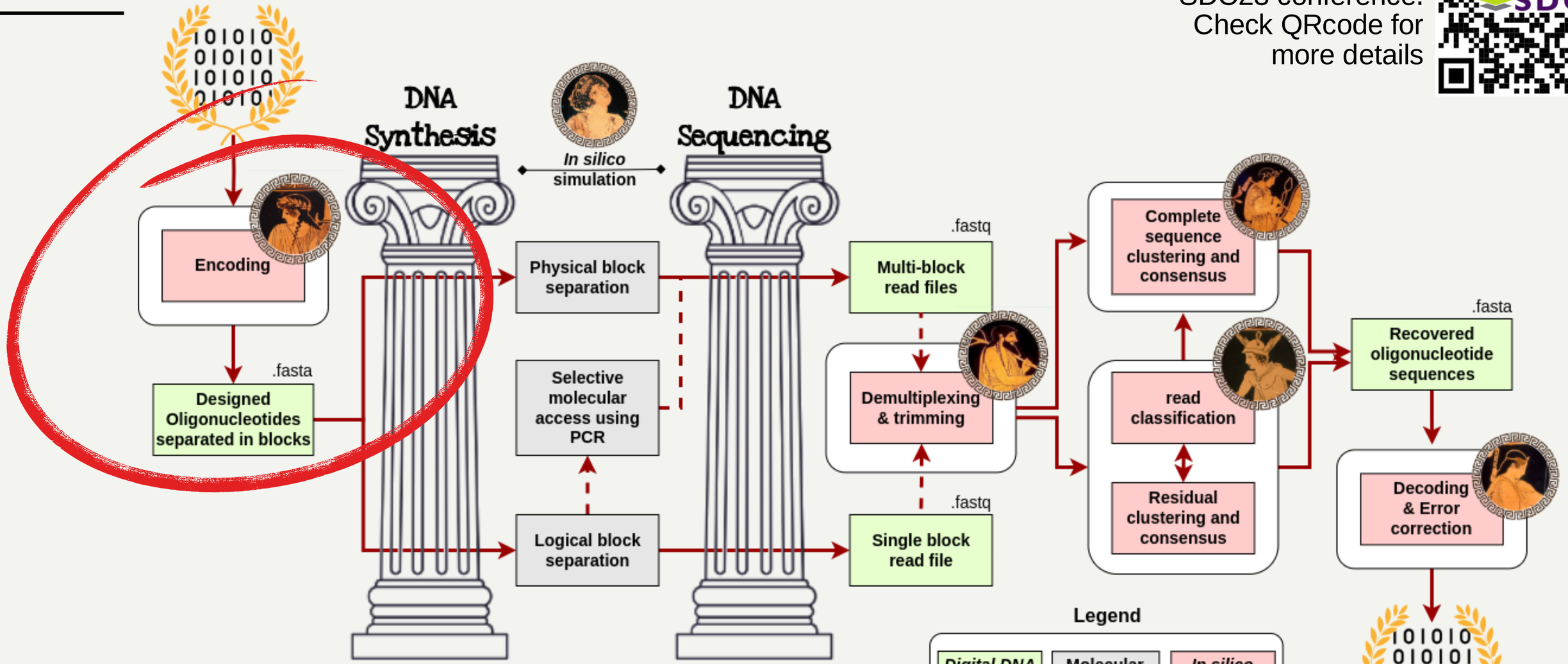
Lenovo™

DDS Workflow



DDS Workflow

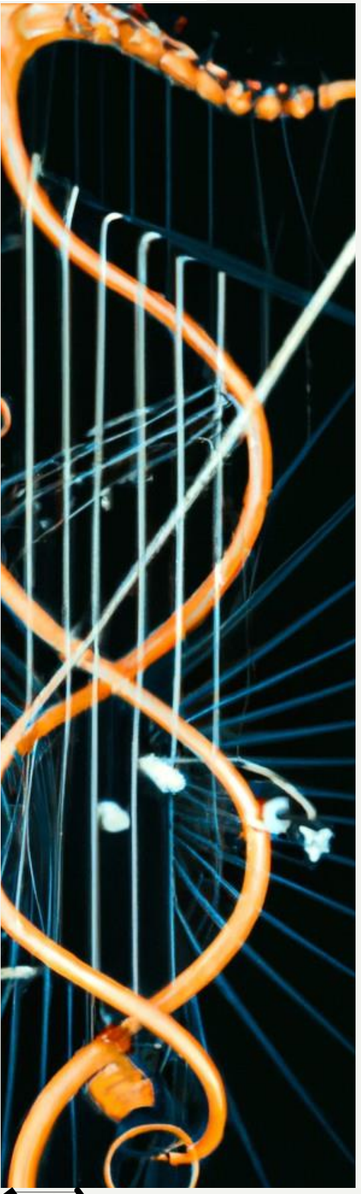
We have presented all Pantheon DNA Gods at SDC23 conference. Check QRcode for more details

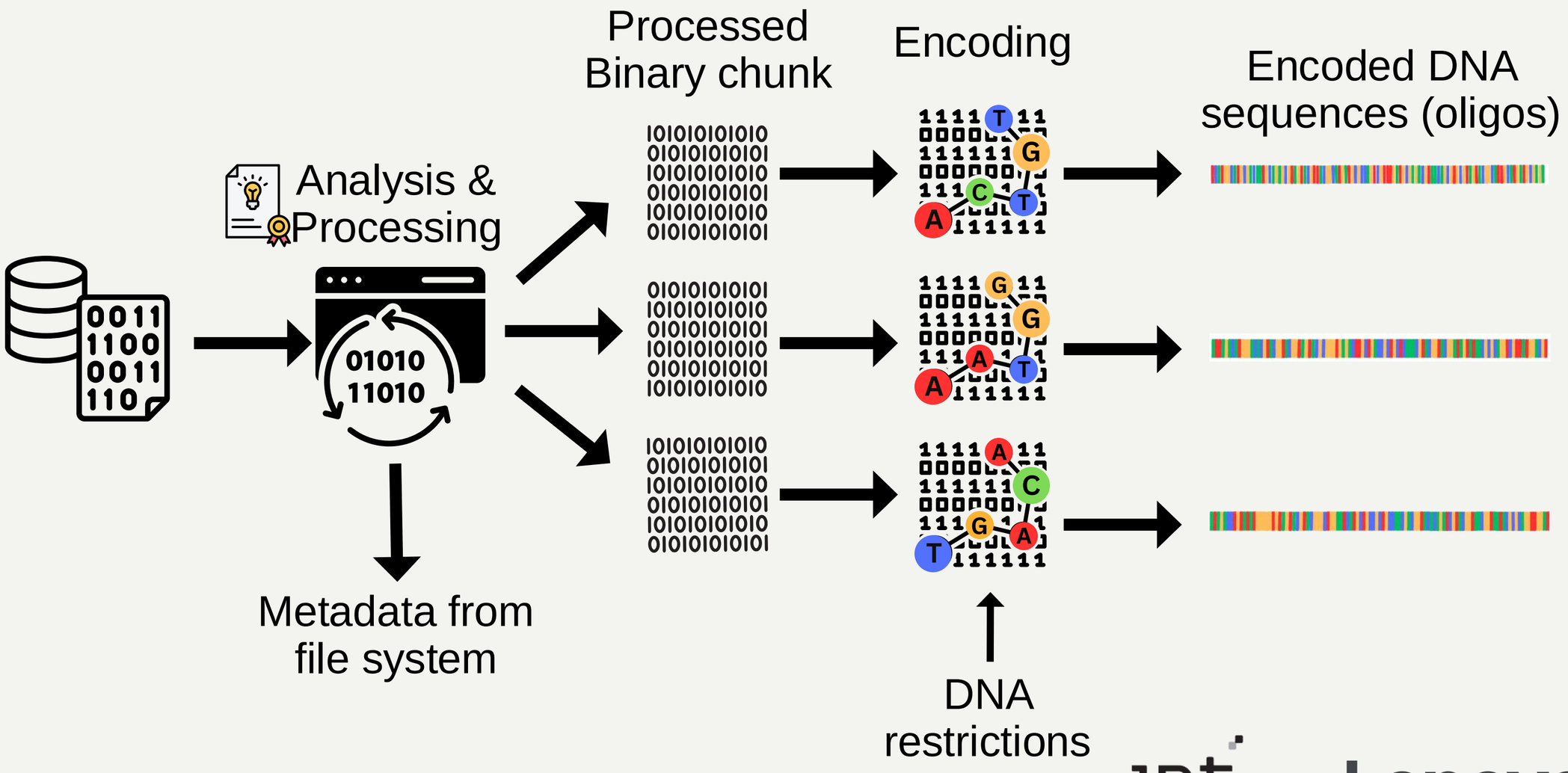


ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

Lenovo

How data is organized into DNA?





PROMETHEUS

Apollo

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

Lenovo

Oligonucleotide data structure

Apollo

PROMETHEUS

Cytosine
Guanine
Adenine
Thymine

Analogy!

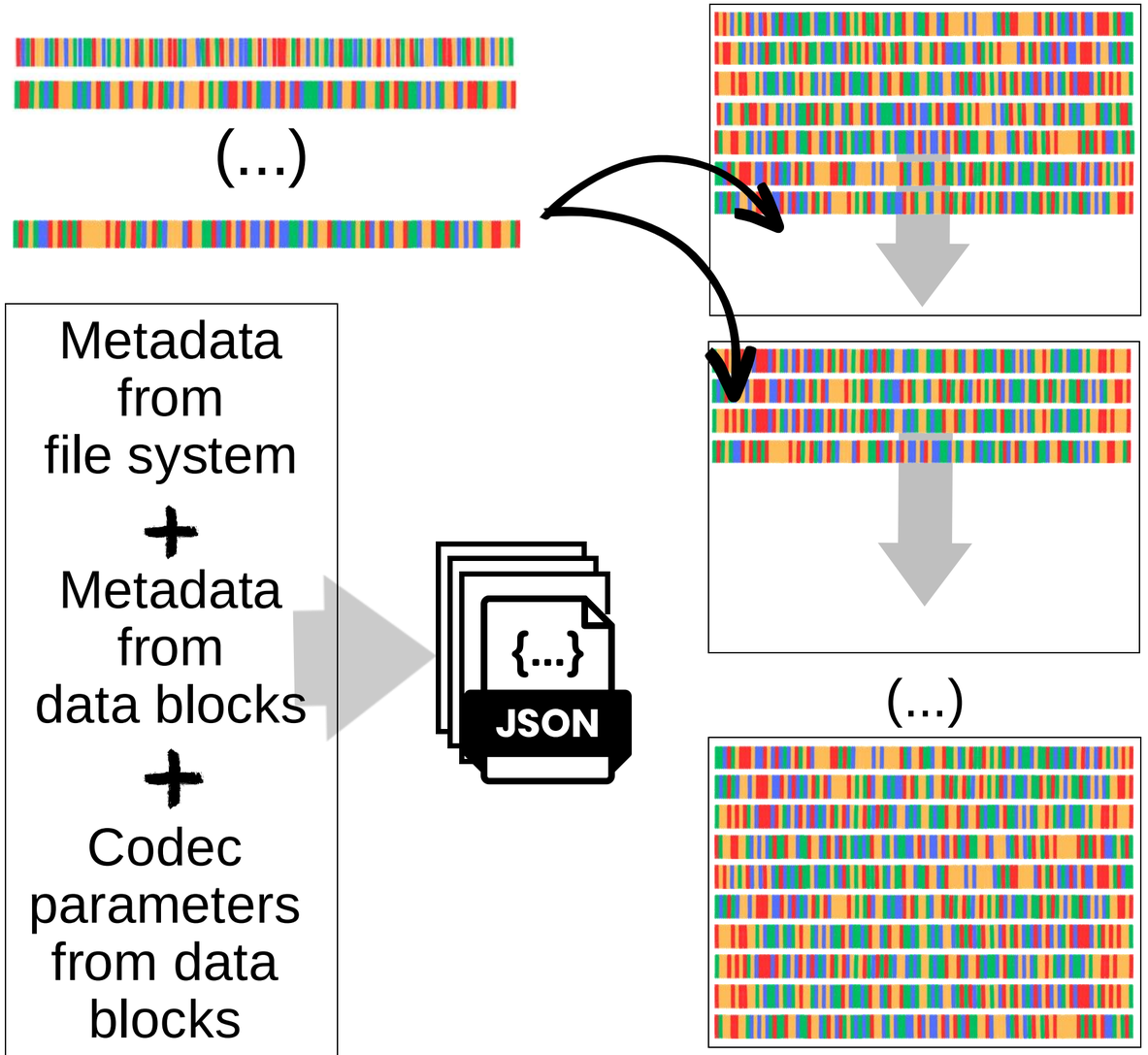
Floppy disk

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

Lenovo

Data Block structure

PROMETHEUS



Apollo

Cytosine
Guanine
Adenine
Thymine

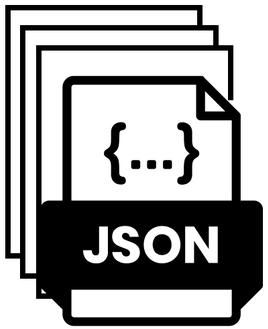
“Disk sectors“

Analogy!

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

Lenovo

Data Block structure



Archive Metadata Block (AMB)

(...)

Data blocks

Outer ECC

Outer ECC

Outer ECC

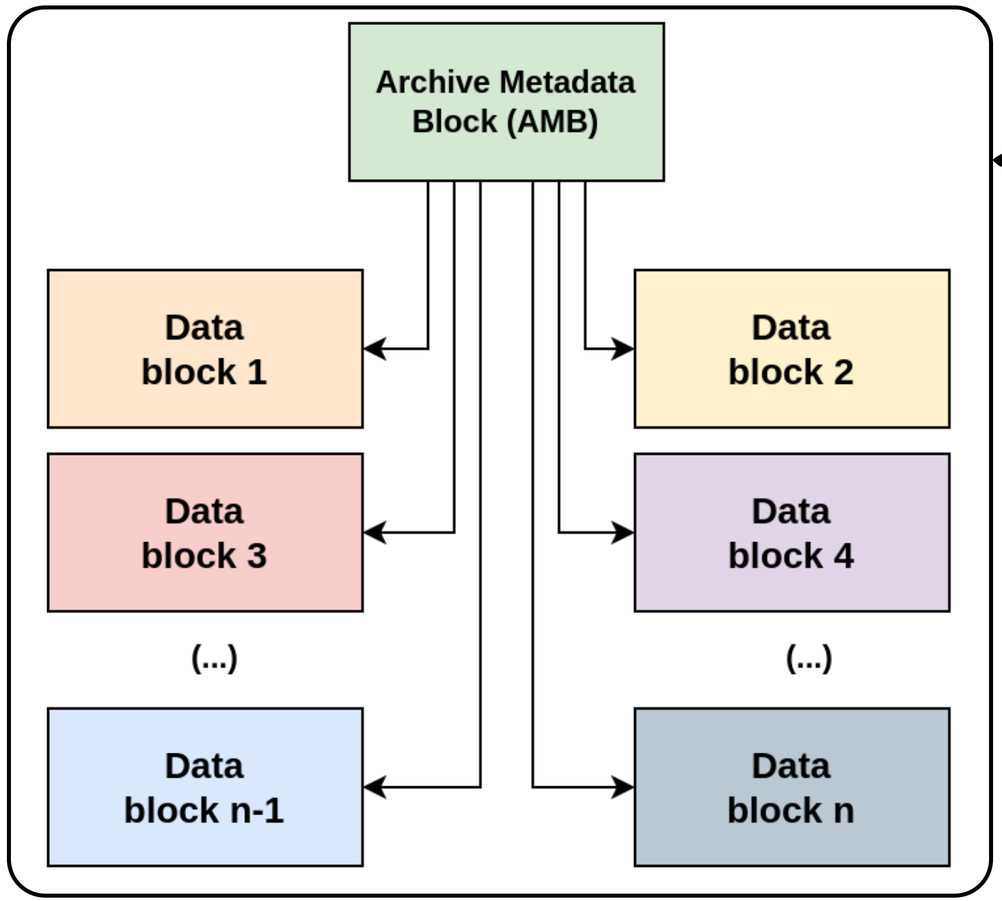
PROMETHEUS

Apollo

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

Lenovo

How data are organized into DNA?



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

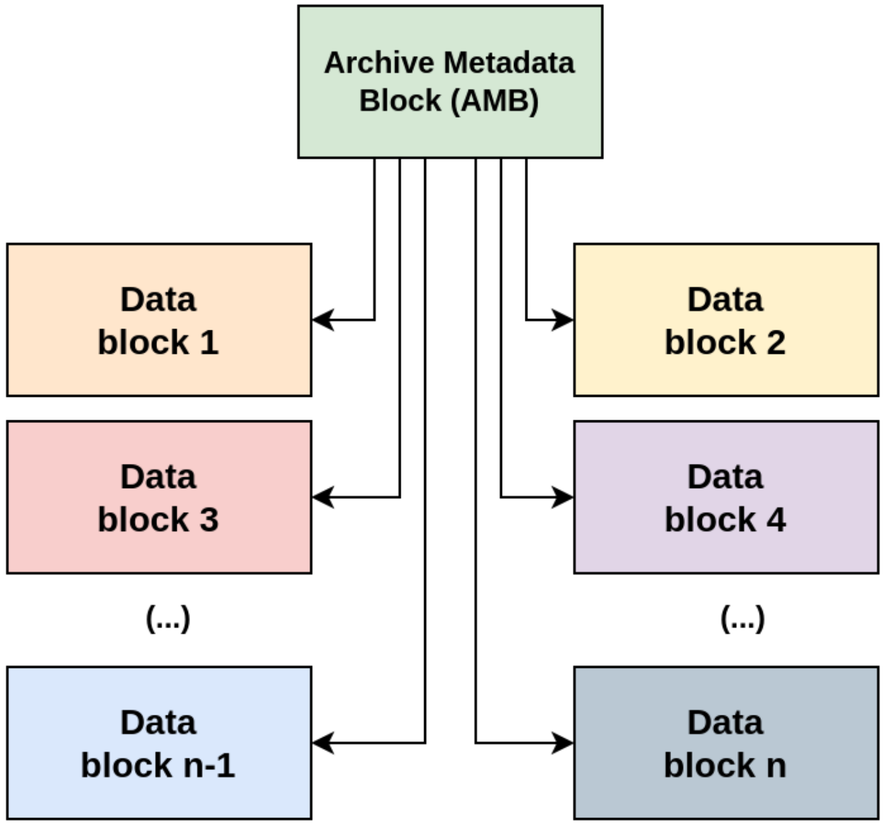
File 1 (...)
File 16 File n
File 3 File n-1

File 12
File 15

File 8 File 8

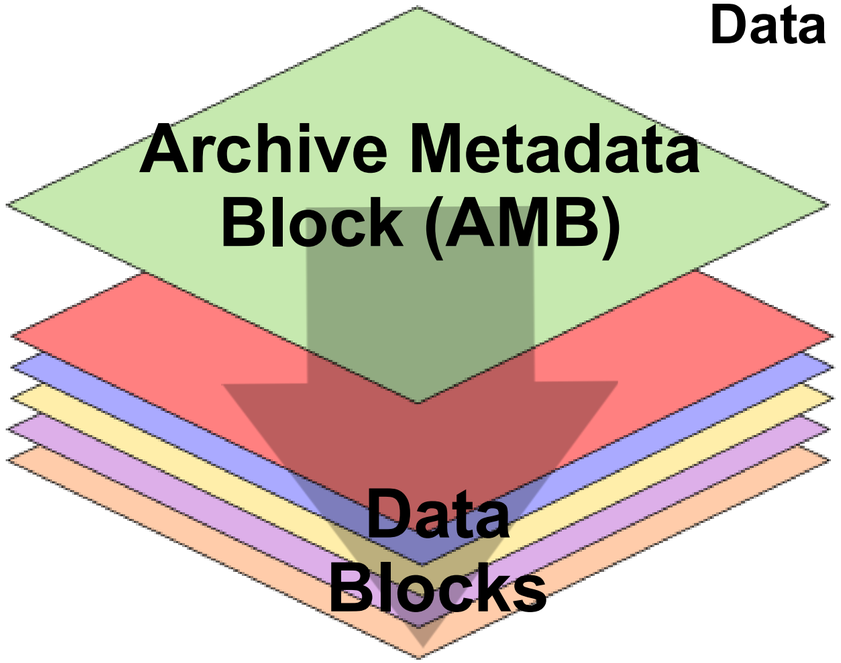
Filling the block is like a Tetris game...

How to read it back?



Chiron

Demultiplexing Data Blocks



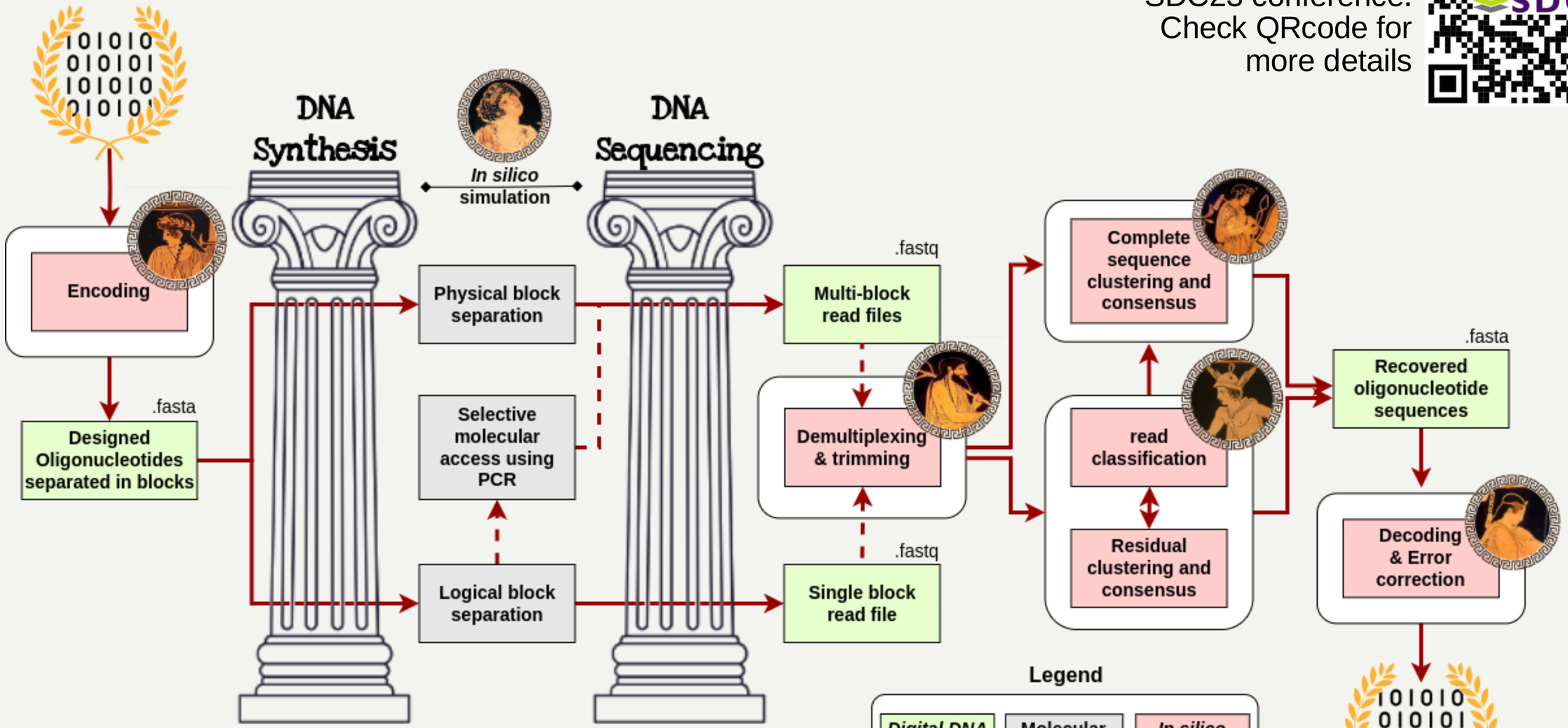
PROMETHEUS

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

Lenovo

DDS Workflow

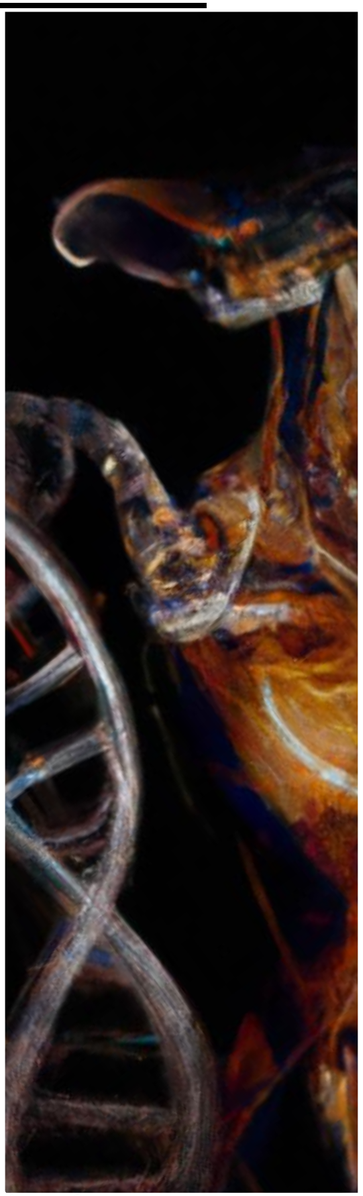
We have presented all Pantheon DNA Gods at SDC23 conference. Check QRcode for more details



ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

Lenovo

Pre-processing NGS reads



Pre-processing steps:

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

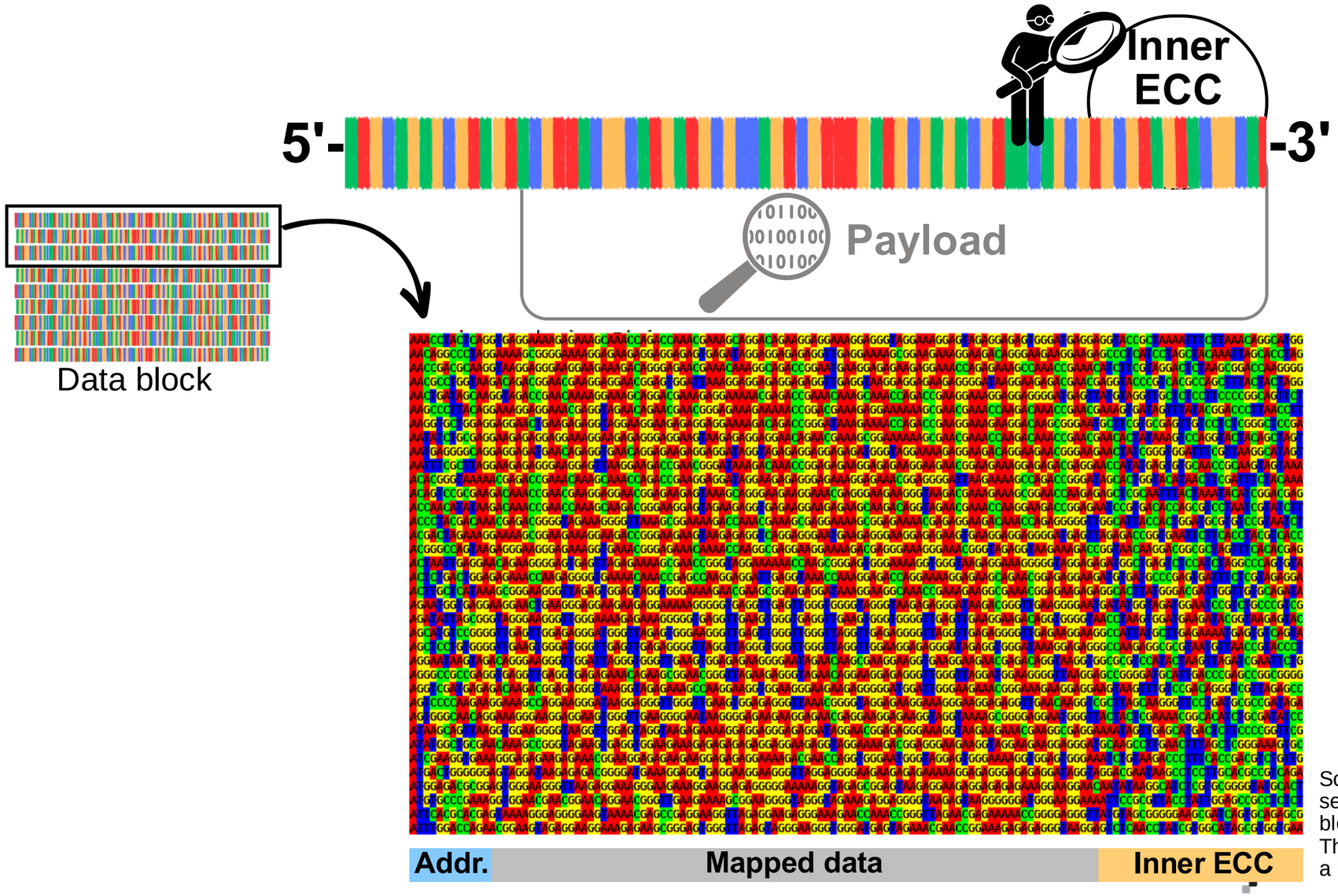
Block specific prime pairs

Planned data blocks

Sequenced DNA reads

Sequenced & Demultiplexed into data blocks

Pre-processing NGS reads

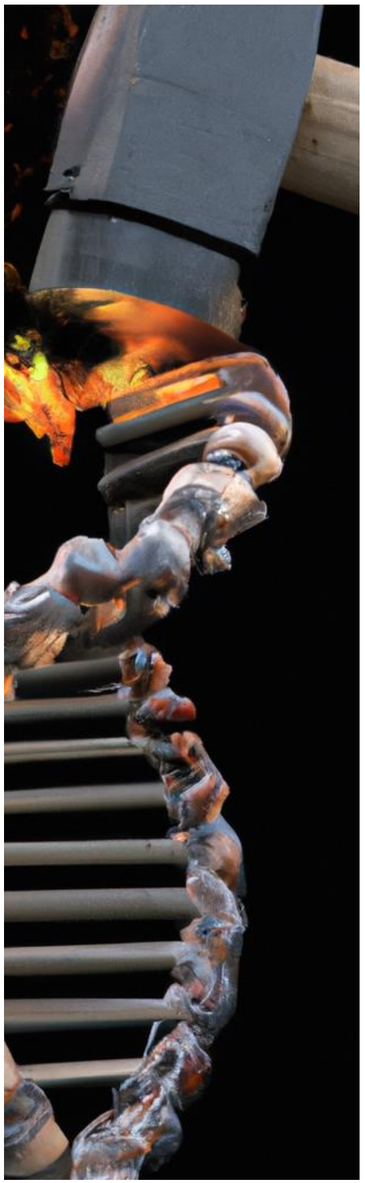


Cytosine
Guanine
Adenine
Thymine

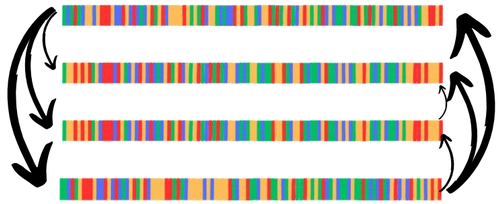
Screenshot of the first 35 DNA sequences from a planned data block in a multi-fasta format. The payload was encoded using a G&A rich scheme.

PROMETHEUS

Pre-processing NGS reads



Demultiplexed block.fastq



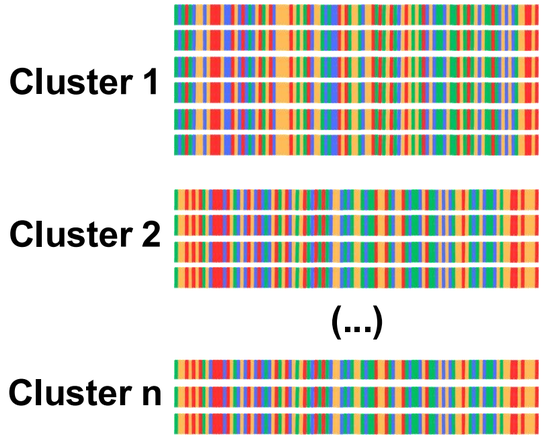
Demultiplexing speeds up clustering

5'-ATCGGCAT-3'

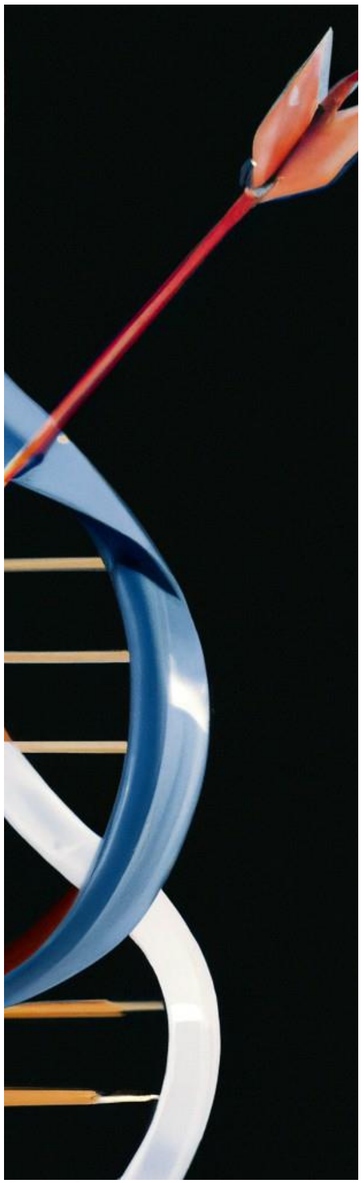
ATCG
TCGG
CGGC
GGCA
GCAT

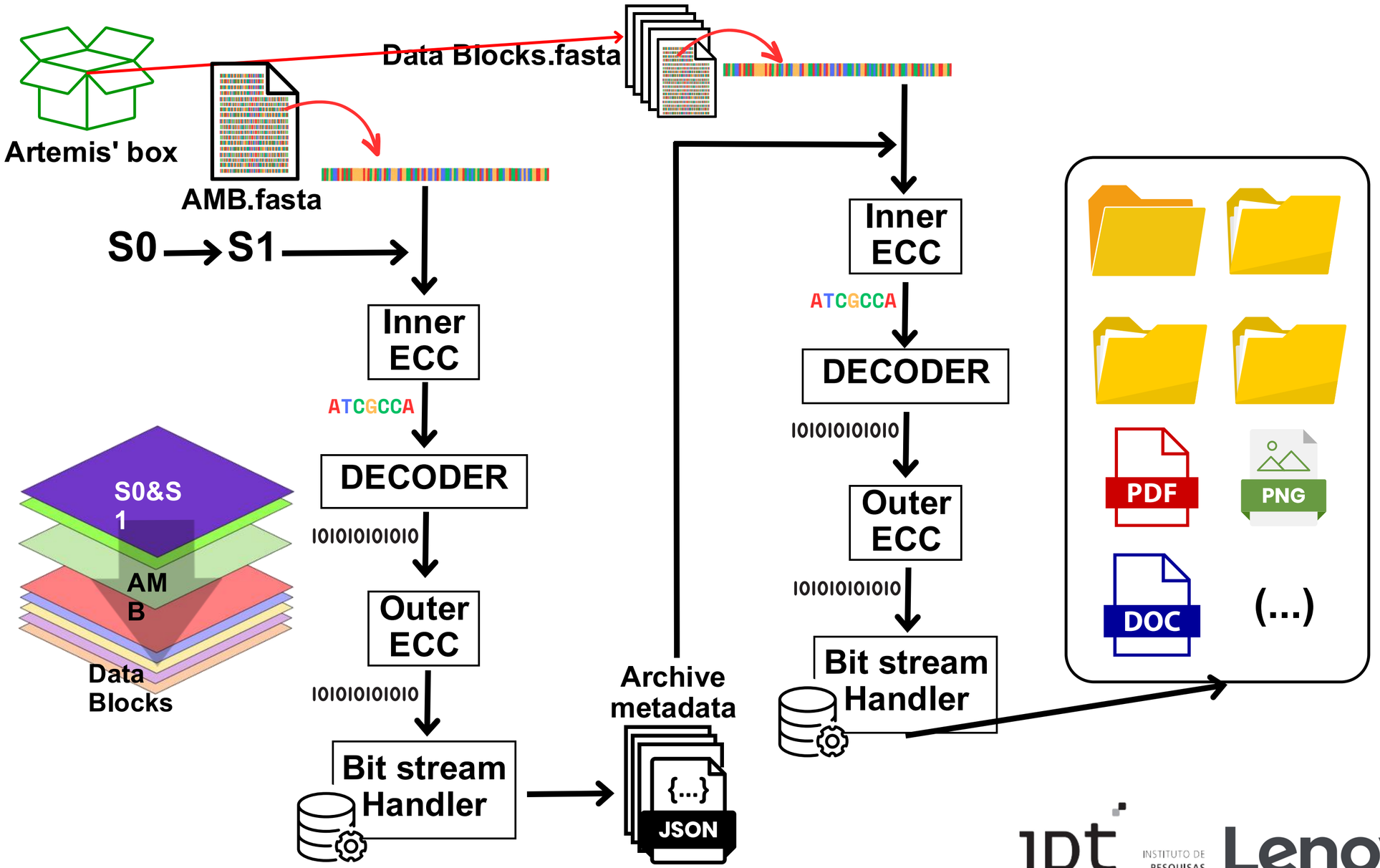
k=4

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



Decoding





PROMETHEUS

artemis

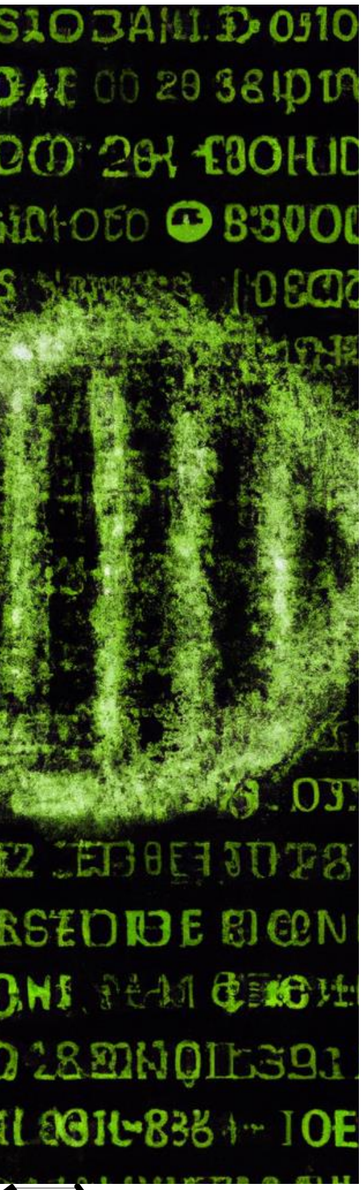
ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

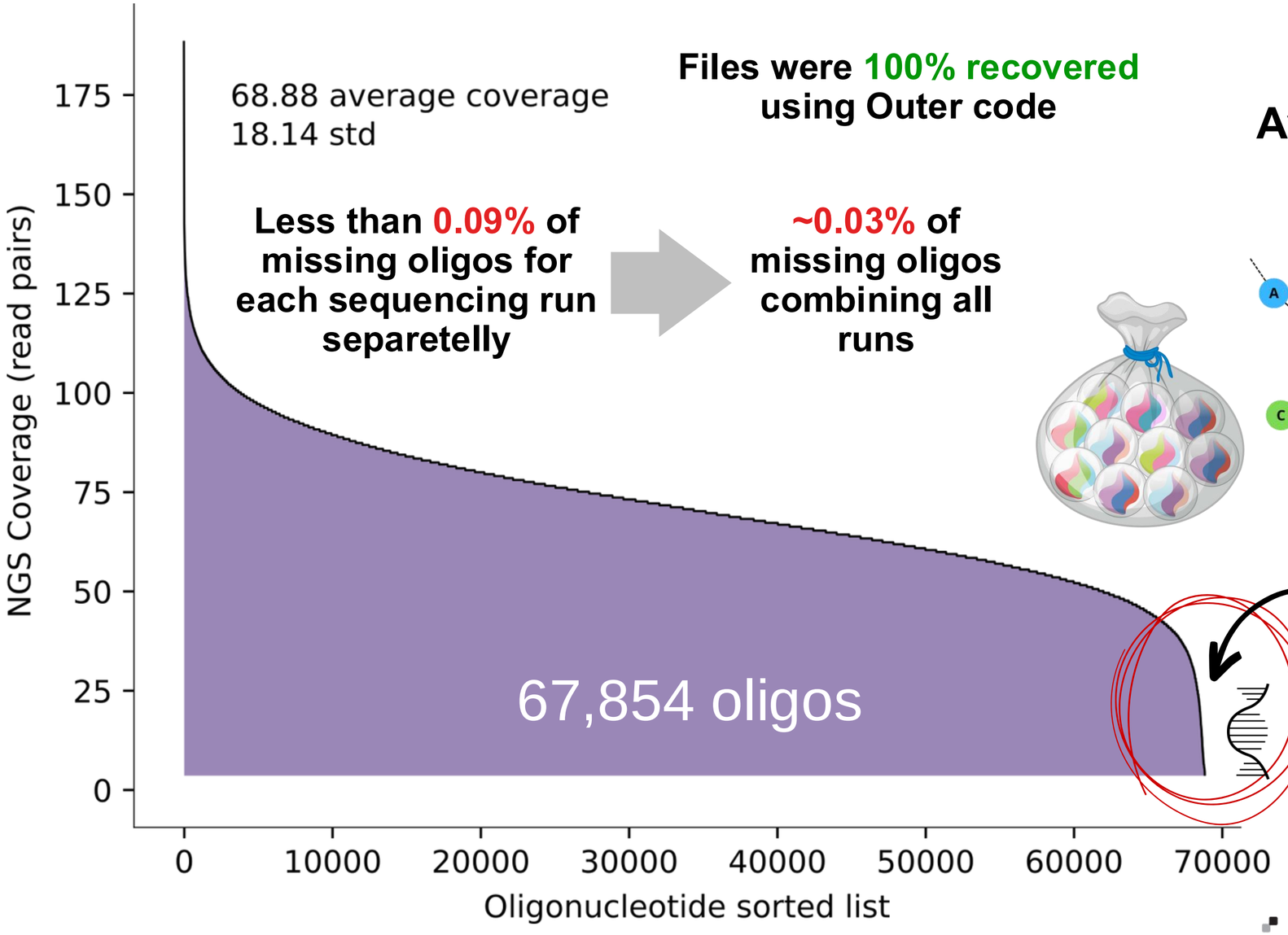
Lenovo

Simulation tool

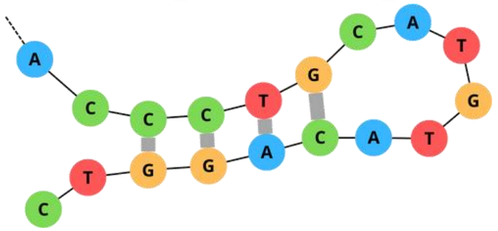
- What Gaia does:
  - Simulate different sequencing strategies
    - Single or Paired-ends
    - Library preparation
    - Coverage variation
    - Sequencing platforms
  - Simulate different synthesis strategies and biases
    - Pandora's box of bias models

Testing CODEC with real data





Avoiding unwanted DNA patterns in designed oligos



"The pit of Tartarus for DNA data storage"

PROMETHEUS

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

Lenovo

Thank you!

Bionanomanufacturing Unity-
IPT, São Paulo - SP, Brazil

bionano@ipt.br

+55 11 3767 4100