Robust image coding on synthetic DNA: Reducing sequencing noise with inpainting

Eva Gil San Antonio, Mattia Piretti, Melpomeni Dimopoulou, Marc Antonini

Université Côte d'Azur

Laboratoire I3S / CNRS

This project has received funding from the European Union's Horizon2020 research and innovation program under grant agreement No 863320

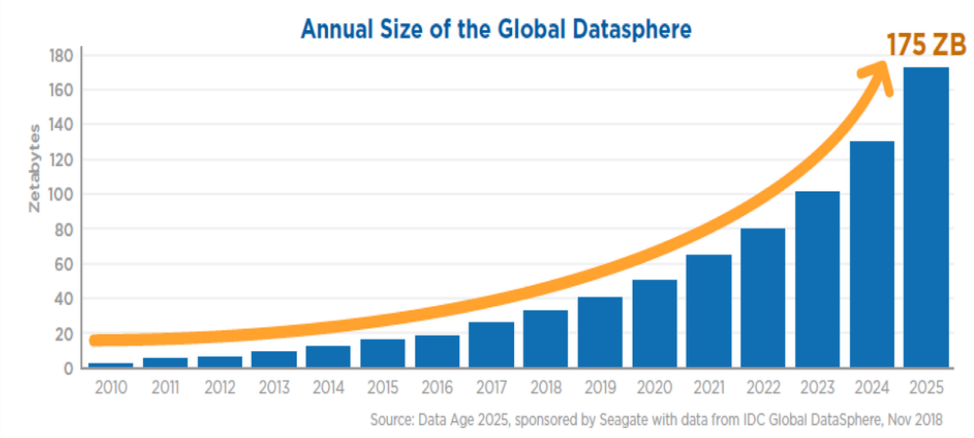
DNA data storage

Problem definition

Improvement in storage density ~20% per year

vs.

Cold data growth ~60% per year



SOLUTION: Storage into DNA (quaternary genetic code)

High storage capacity

(455 exabytes in 1 gram)

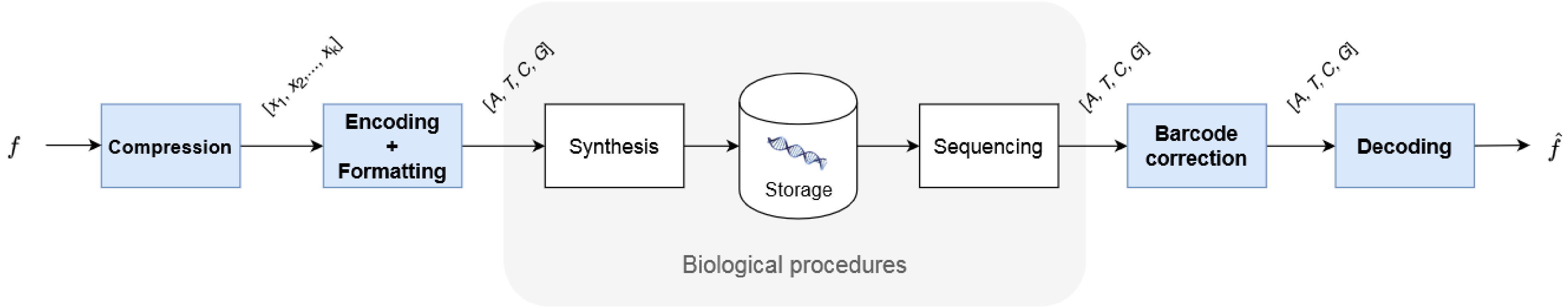
Data longevity

(for centuries even in harsh environments)

Easy, quick and chip *in-vitro* replication

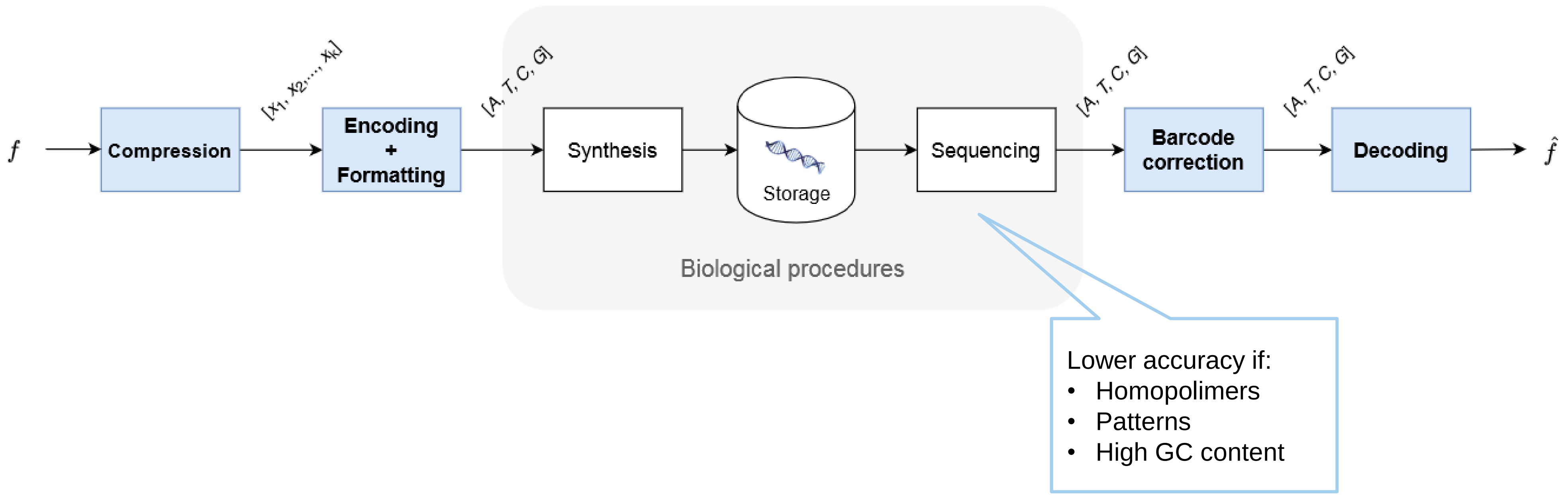
DNA data storage

General Workflow



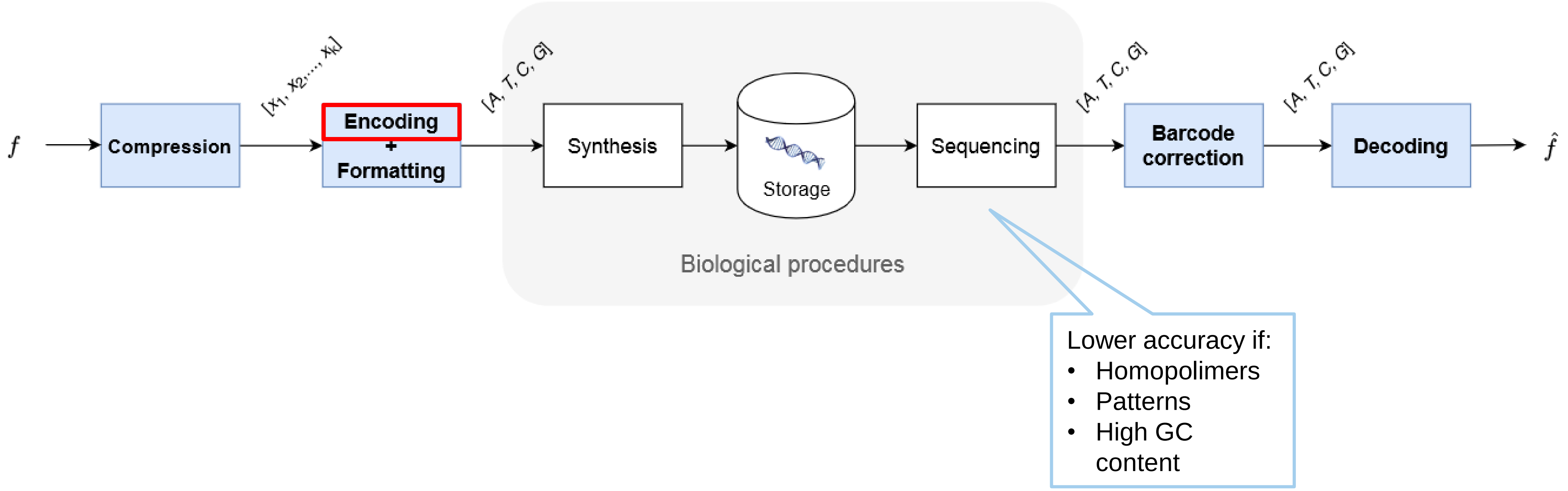
DNA data storage

General Workflow



DNA data storage

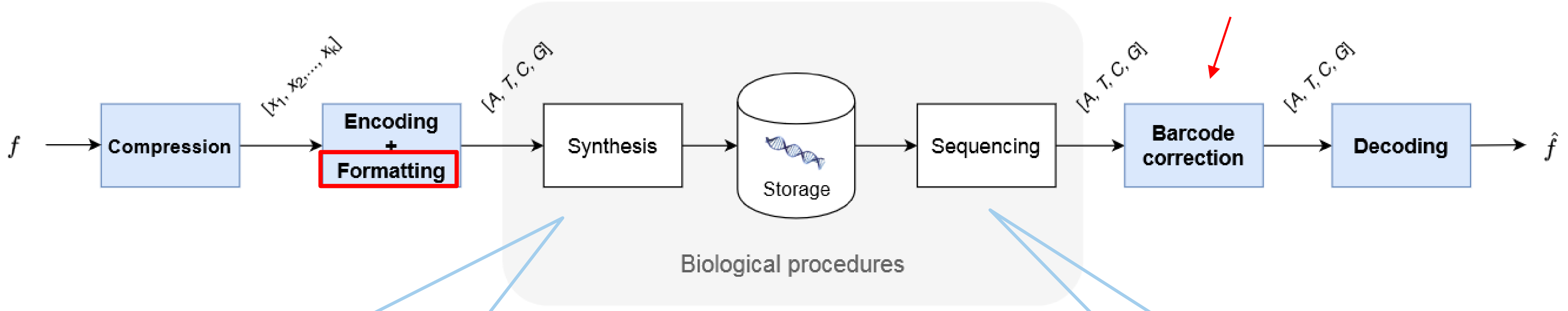
General Workflow



M. Dimopoulou, M. Antonini, P. Barbry, R. Appuswamy, “A biologically constrained solution for long-term storage of images onto synthetic”, EUSIPCO, 2019.

DNA data storage

General Workflow



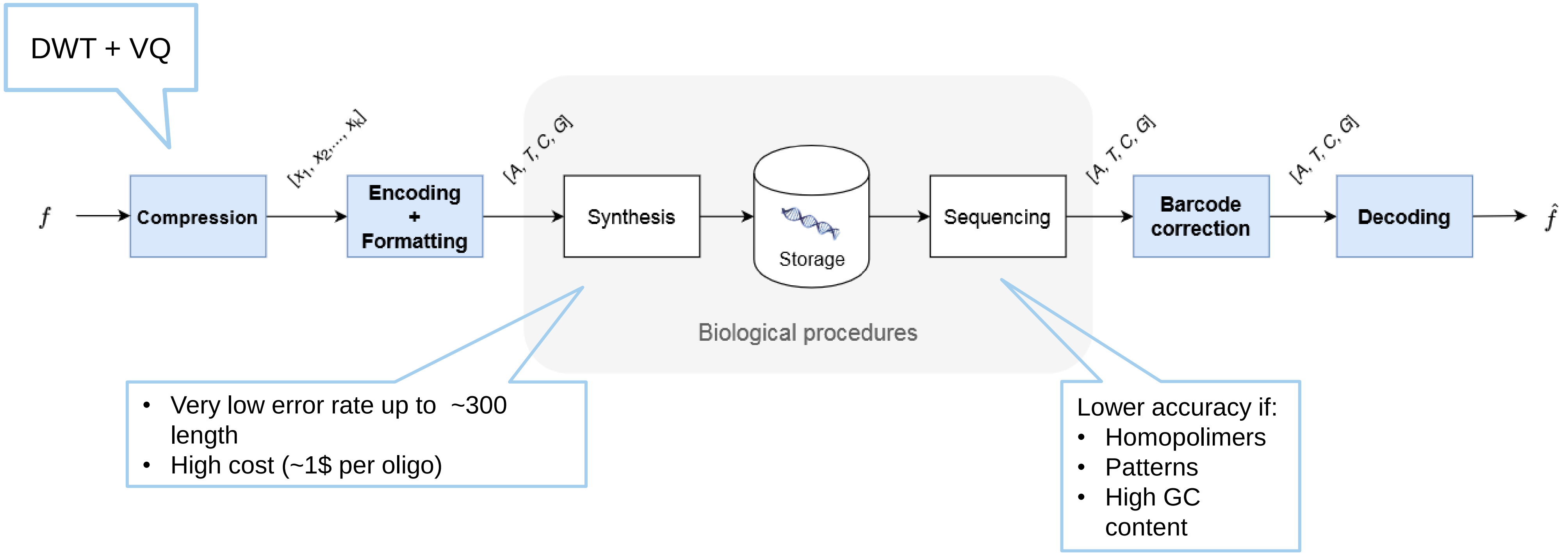
- Very low error rate up to ~300 length
- High cost (~1$ per oligo)

Lower accuracy if:

- Homopolimers
- Patterns
- High GC content

DNA data storage

General Workflow

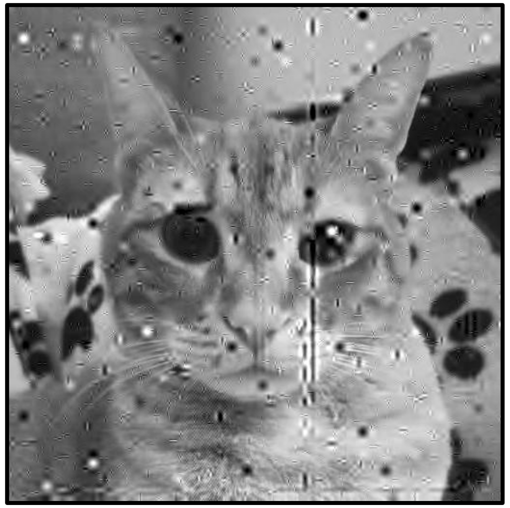


M. Dimopoulou, E. Gil San Antonio, M. Antonini, “An efficient sequencing noise resistant mapping for the encoding of images onto synthetic DNA”, MMSP, 2020

Inpainting

TS on decoded image vs. TS on wavelet subbands

Original image



Noisy image

Inpainting on decoded image
PSNR = 35.01 dB

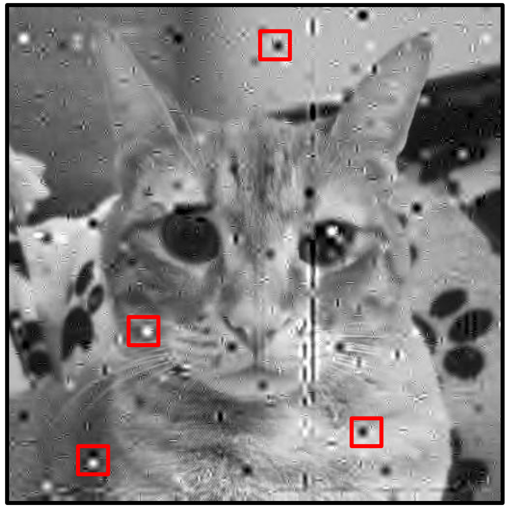
Inpainting on decoded wavelet subbands
PSNR = 36.38 dB

Inpainting

TS on decoded image vs. TS on wavelet subbands

Original image

Noise in low-frequency subband



Noisy image

Inpainting on decoded image
PSNR = 35.01 dB

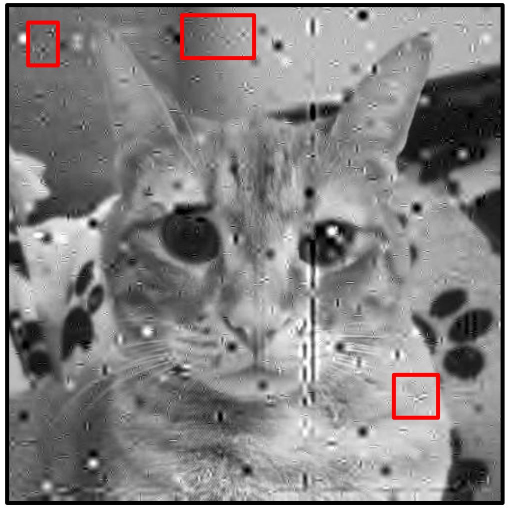
Inpainting on decoded wavelet subbands
PSNR = 36.38 dB

Inpainting

TS on decoded image vs. TS on wavelet subbands

Original image

Damage in high-frequency subbands



Noisy image

Inpainting on decoded image
PSNR = 35.01 dB

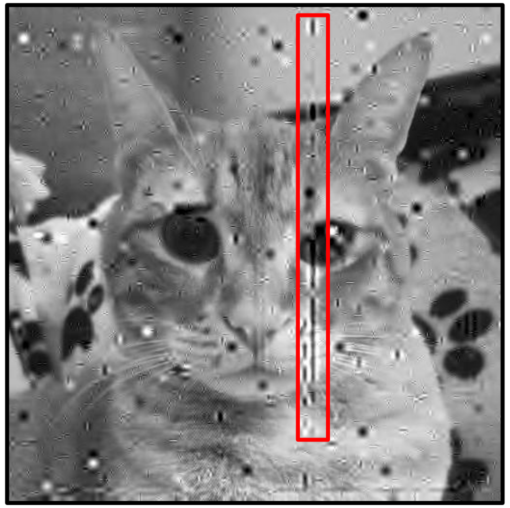
Inpainting on decoded wavelet subbands
PSNR = 36.38 dB

Inpainting

TS on decoded image vs. TS on wavelet subbands

Original image

Missing oligos



Noisy image

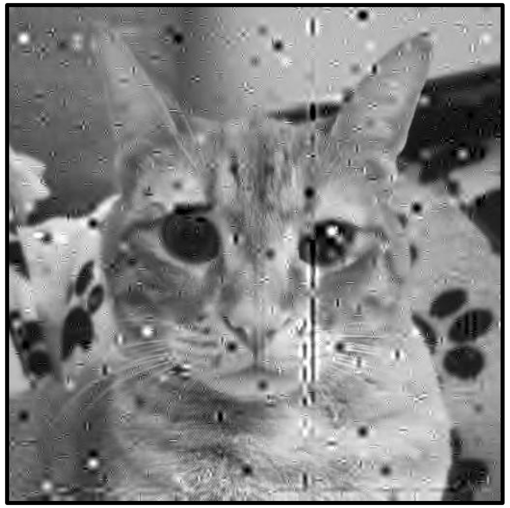
Inpainting on decoded image
PSNR = 35.01 dB

Inpainting on decoded wavelet subbands
PSNR = 36.38 dB

Inpainting

TS on decoded image vs. TS on wavelet subbands

Original image



Noisy image

Inpainting on decoded image
PSNR = 35.01 dB

Inpainting on decoded wavelet subbands
PSNR = 36.38 dB

Inpainting

Automatic detection of the damaged areas

2-step algorithm (performed in each subband separately):

1. Detection of errors in single pixels (substitutions) → *Deviation of each pixel and its neighbors*

τ_1 : phase 1 threshold
N_p : neighborhood of the pixel p.
I : damaged image.
O : mask

$$S, M : \forall p_{(x,y)} \in I, S_p = \sigma(N_p) \text{ and } M_p = \overline{N_p}$$

$$\forall p_{(x,y)} \in I, \text{ if } \frac{\sqrt{(p_{(x,y)} - M_p)^2}}{S_p} >= \tau_1 : \text{ then } O_{(x,y)} = true$$

2. Detection of damaged neighborhoods (indels) → *Internal variance of the neighborhoods*

τ_2 : phase 2 threshold.
$S_{mean} = \overline{S}$

$$\forall p_{(x,y)} \in I, \text{ if } \frac{S_p}{S_{mean}} >= \tau_2 : \text{ then } M_{(x,y)} = true$$

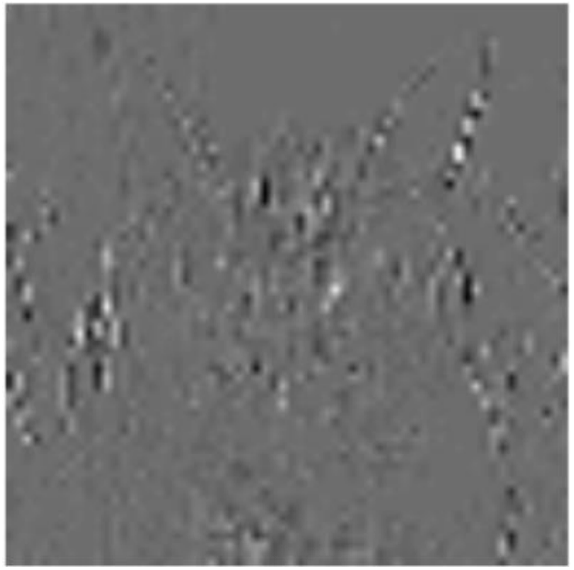
Experiment set-up

Texture Synthesis

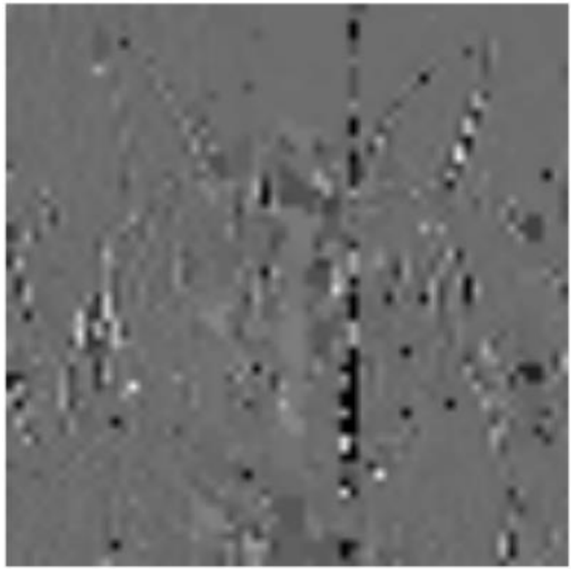
- **Encoding: constrained codebook [1], Vector Quantization [2] and controlled mapping resistant to noise [3]**
- **Simulation of sequencing noise:**
    - Substitution and indel rates adapted from [4]:
      2.3% deletions
      1.01% insertions
      1.5% substitutions
    - 80% of the noise concentrated in the first and last 20nt of each oligo [5]
    - 200 noisy copies of each input oligo
- **Headers encoded using barcodes**
- **Oligos are clustered after barcode correction and filtered by length.**
- **Consensus based on majority voting**

Inpainting

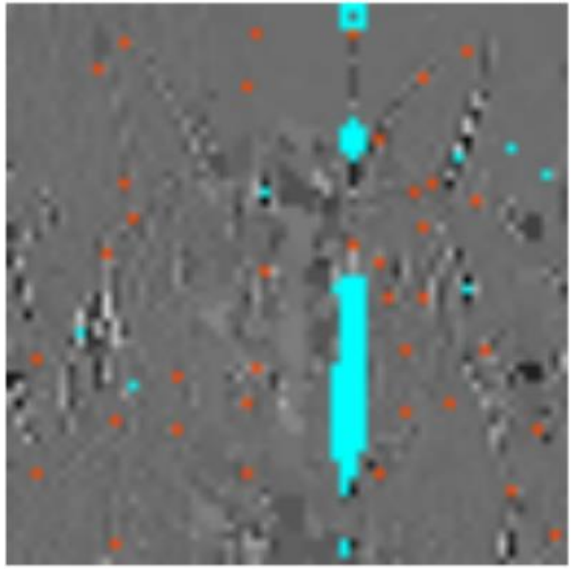
Automatic detection of the damaged areas



Quantized subband without sequencing noise



Quantized subband with sequencing noise



Detected damaged areas (Red: 1st step red; blue: 2nd step)

Inpainting

Texture synthesis on Wavelet subbands. Results.

Quantized image without sequencing noise
Compression ratio = 4.9708 bits/nt
PSNR = 48.12 dB, SSIM = 0.991

Visual impact of sequencing noise in the image encoded using the controlled mapping (after barcode correction)
PSNR = 36,2 dB, SSIM = 0.92

Post-processed image using inpainting
PSNR = 38.7 dB, SSIM = 0.94

Thank you for your attention!