ADS Codex

Adaptive Codec for Organic Molecular Archives

Latchesar Ionkov

Bradley Settlemyer

Dominic Manno

NNSA National Nuclear Security Administration

Managed by Triad National Security, LLC for the U.S. Department of Energy's NNSA

LA-UR-20-21586

Goals

- Provide high bit density
- Adapt to oligos of different lengths
- Adapt to different errors
  - Error types
  - Error rates for different types
  - Spatial distribution of errors
  - Missing whole oligos (erasures)

DNA 101

- Structure
  - Complex organic polymer
  - Consists of sequences of **A**, **C**, **T**, and **G** *nucleotides* (nts)
  - Can be single or double stranded
  - Very stable if lyophilized (dried)
  - Not all nts are equal (Gs)
- Synthesis
  - Uncommon in nature
  - Append a single nucleotide at the end of the molecule
  - Can take minutes per nt
- Amplification (PCR)
  - Common in nature
  - Creates copies of an existing molecule
  - Extremely cheap to create copies of existing data
- Sequencing
  - Most techniques use a form of PCR
  - Each sequence is sequenced more than one time (read depth)

Processes are very slow and need to be done at extreme scale (> million molecules) in order to be practical for data storage.

DNA Codec Challenges

- Short sequences (100-300 nts)
- Need for oligo identification
- Very high error rates (0.3-20% per nucleotide)
- “Structural” oligo restrictions
  - Homopolymer length
  - CG Content
- Sources of errors
  - Synthesis
  - Amplification
  - Sequencing
- Types of errors
  - Substitutions
  - Deletions
  - Insertions

Bit Packing

- How to encode as many bits as possible per nucleotide (nt)?
- Deal with “structural” errors only
- Theoretical maximum: 2 b/nt
- Existing methods:
  - Goldman et al.: 1.58 b/nt
  - Grass et al.: 1.78 b/nt
  - Erlich et al.: 1.98 b/nt

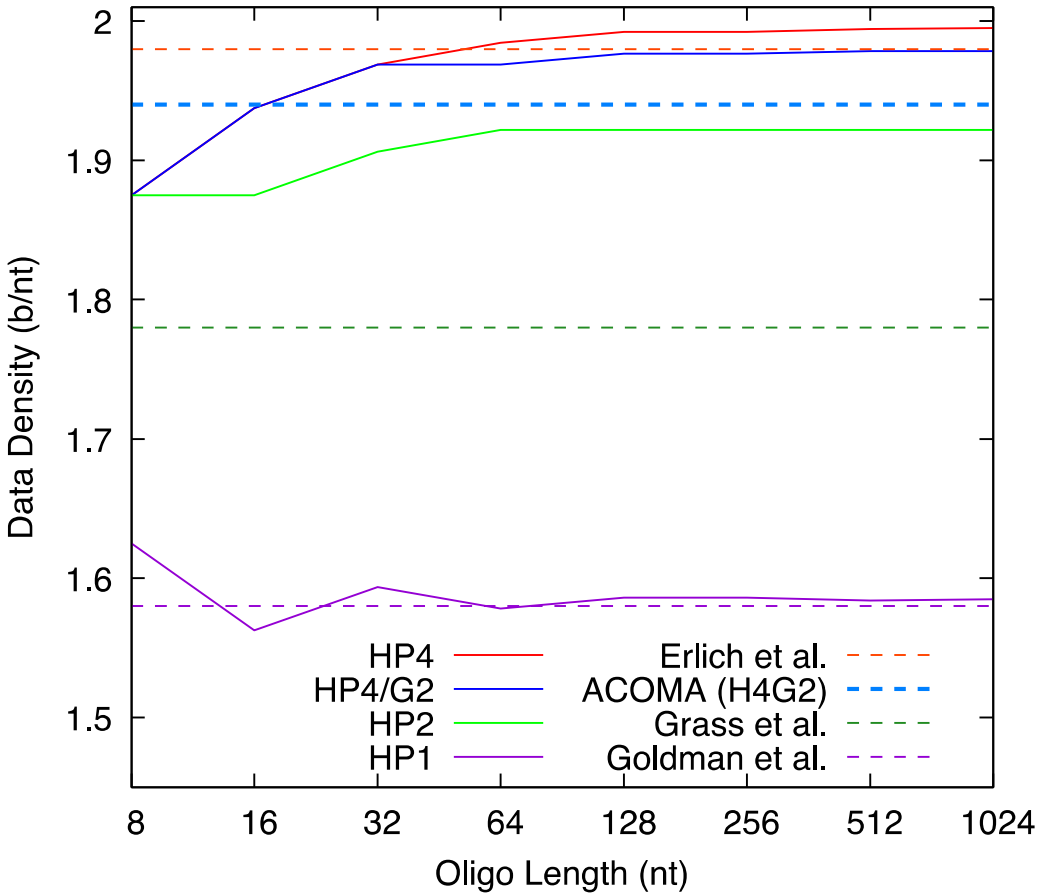
Bit Packing

Oligo counting

- Order oligos
- Reject the ones that fail criteria

~~*AAAAA*~~	~~*X*~~
AAAAT	*0*
AAAAC	*1*
…	
AAGGC	*61*
~~*AAGGG*~~	~~*XX*~~
ATAAA	*62*
…	
GGCGG	*980*
~~*GGGAA*~~	~~*XXX*~~
…	
~~*GGGGG*~~	~~*XXX*~~

Total: 981 values (1.8 b/nt)



ADS Codex

- **Level 0**: bit packing
  - How to convert between bits and nts
- **Level 1**: single oligo layout
  - Oligo identification (+ metadata error correction)
  - Data
- **Level 2**: error correction with multiple oligos
  - Error correction for data

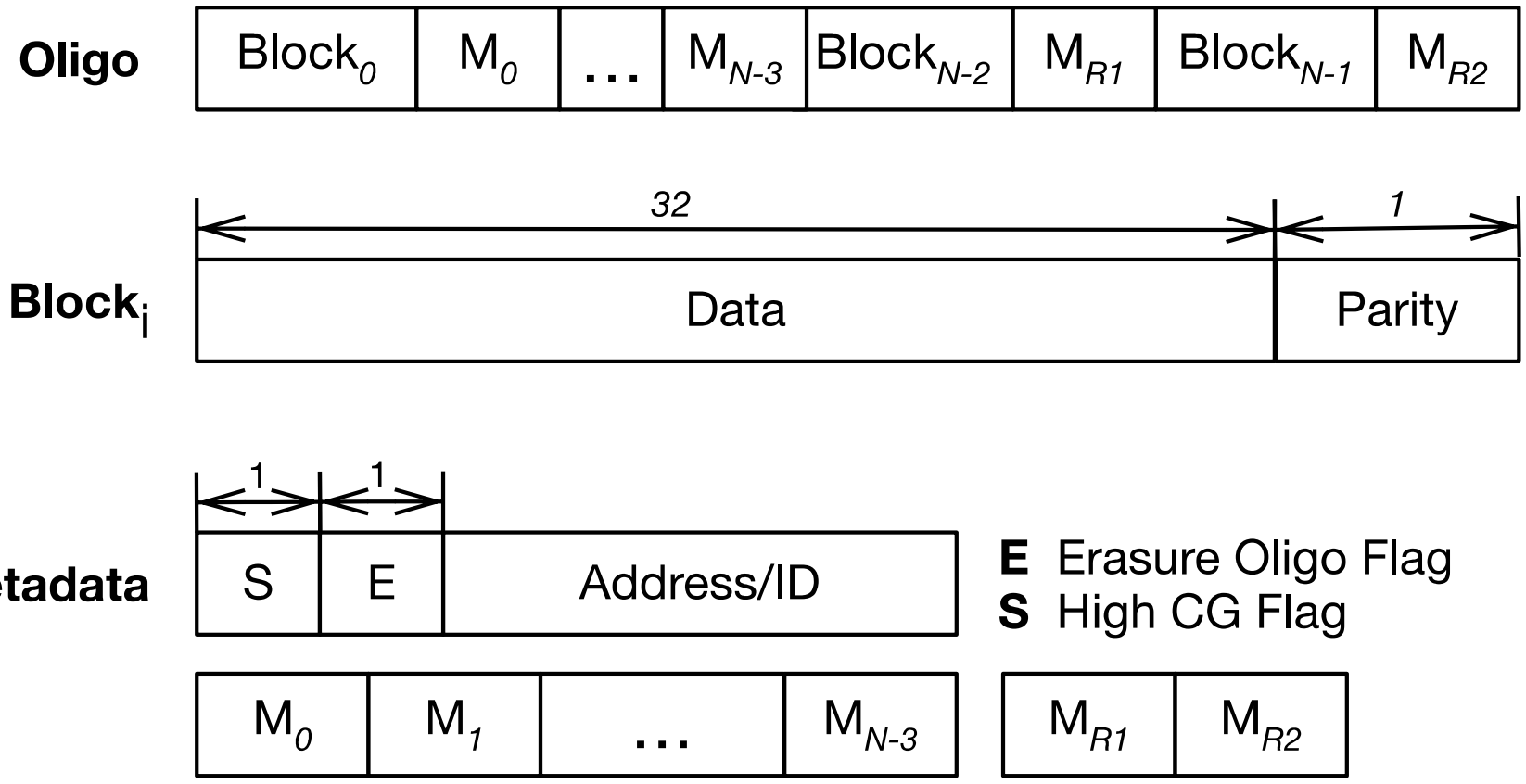
ADS Codex Bit Packing (Level 0)

- Same as oligo counting
- Homopolymer restrictions (H4G2):
  - Maximum 4 for A, T, and C
  - Maximum 2 for G
- High CG content: handled at a higher codec level
- It takes exponentially longer time to encode/decode
- Limit up to 17 nts (32 bits data + 1 parity bit)
  - Minimum 10,315,700,031 values (> 2^{33})
  - Data density: 33/17 = **1.94 b/nt**
- Oligo construction:
  - Allows adding up to 17 nts fragments at the end
  - Preserves the homopolymer length restrictions

ADS Codex Oligo Layout (Level 1)

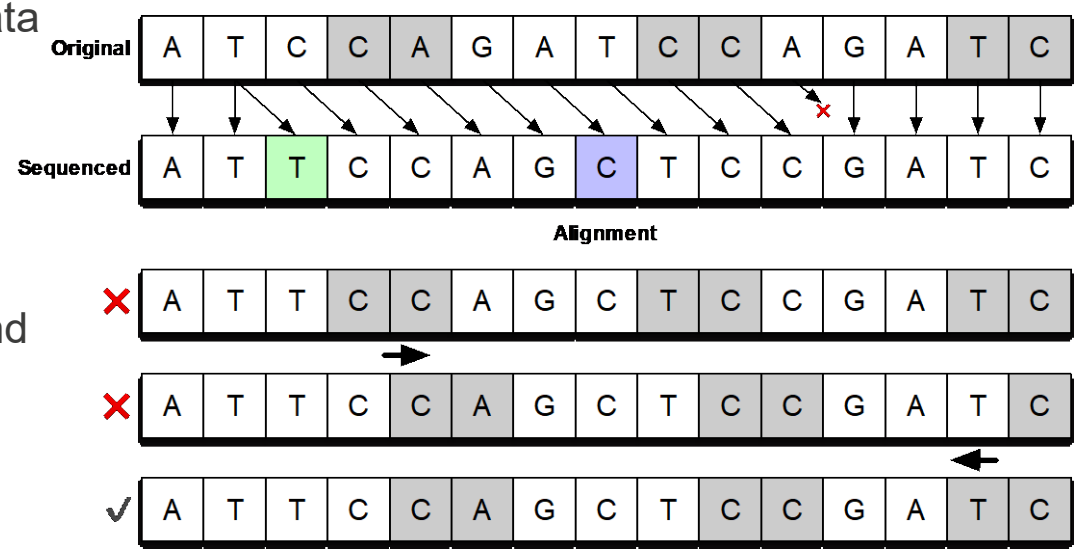
- Data blocks
  - ***N*** 33 bit blocks (32 bit + parity)
- Metadata
  - High CG flag (1 bit)
  - Erasure code oligo flag for Level 2 (1 bit)
  - Address/ID
- Distributed metadata
  - Split into ***N*** blocks and place between the data blocks
  - Some of the metadata blocks contain CRC or RS error detection codes
  - Used to discover deletions/insertions and align data blocks

ADS Codex Oligo Layout (Level 1)



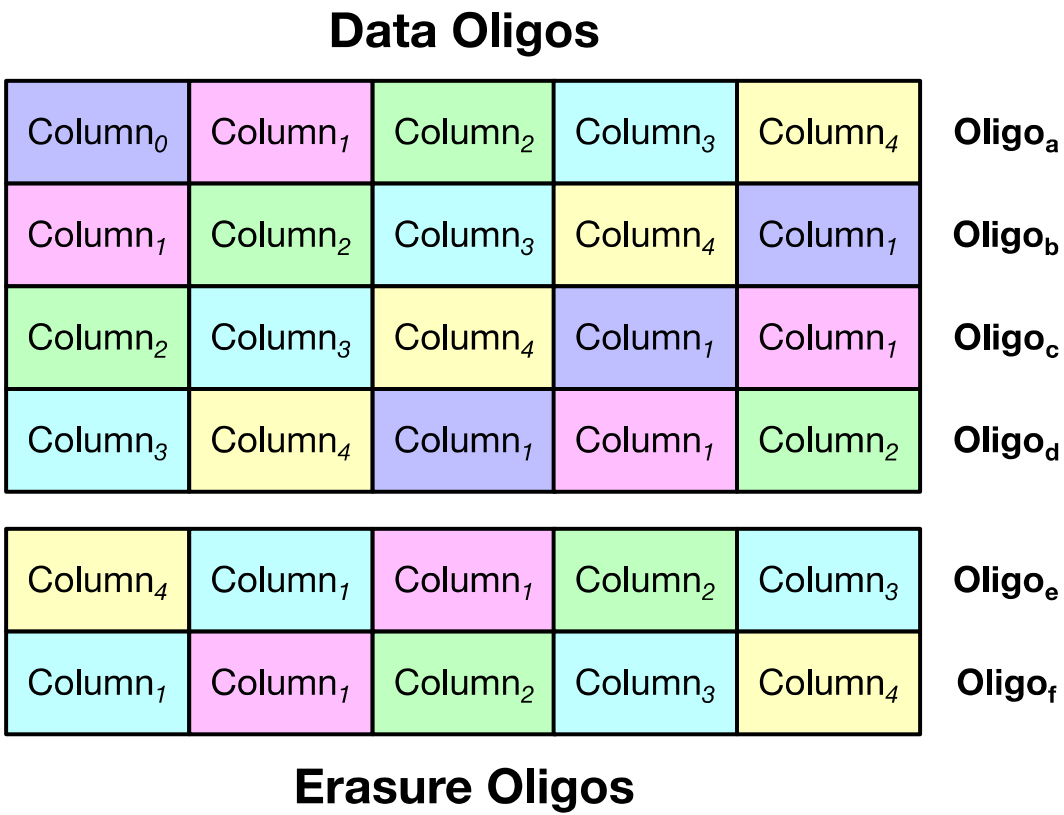
Block Alignment

- Try to detect insertions and deletions by shifting the metadata blocks
- Shift by 1 left/right and try to match the checksum
- Multiple levels of complexity (and speed):
  - Level 0: do nothing
  - Level 1: Assume 1-2 errors in data blocks
  - Level 2: Assume insertion errors in metadata blocks



ADS Codex Erasure Coding (Level 2)

- 2D erasure encoding
- Group *M* oligos together
- Calculate erasure codes for data blocks from different oligos
- Columns consist of blocks from different positions (to correct for spatial bias of errors)
- Reed-Solomon erasure coding
- Parity bit in data blocks used to detect errors



ADS Codex Examples

ADS Codex Parameters for 107 nt oligos

- 5 data blocks, 17 nts each = 85 nts: **1.94 b/nt**
- 3*4 nts metadata blocks + 2*5 nts RS blocks = 22 nts
- Metadata
  - 4 nts -> 186 values
  - 186^3 = 6,434,856 values
  - 2 bits for High-CG and Erasure-Oligo flags
  - 1608714 addresses/IDs left (30.7 MB)
- Level 1: **1.50 b/nt**
- Erasure Coding
  - 2 (1) Erasure Oligos for every 4 data oligos
  - 4*5*32 bits = 640 data bits per group for 606 (505) nts
- Level 2: **0.996 (1.196) b/nt**

ADS Codex Parameters for 110 nt oligos

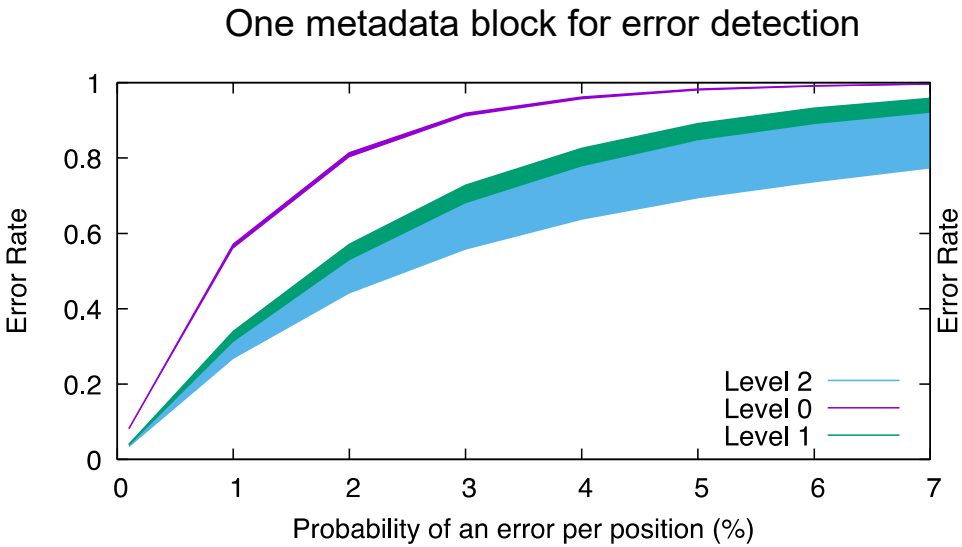
- 5 data blocks, 17 nts each = 85 nts: **1.94 b/nt**
- 3*5 nts metadata blocks + 2*5 nts RS block = 25 nts
- Metadata
  - 5 nts -> 733 values
  - 733^3 = 393,832,837 values
  - 2 bits for High-CG and Erasure-Oligo flags
  - 98,458,209 addresses/IDs left (1.8 GB)
- Level 1: **1.45 b/nt**
- Erasure Coding
  - 2 (1) Erasure Oligos for every 4 data oligos
  - 4*5*32 bits = 640 data bits per group for 630 (525) nts
  - Level 2: **0.969 (1.163) b/nt**

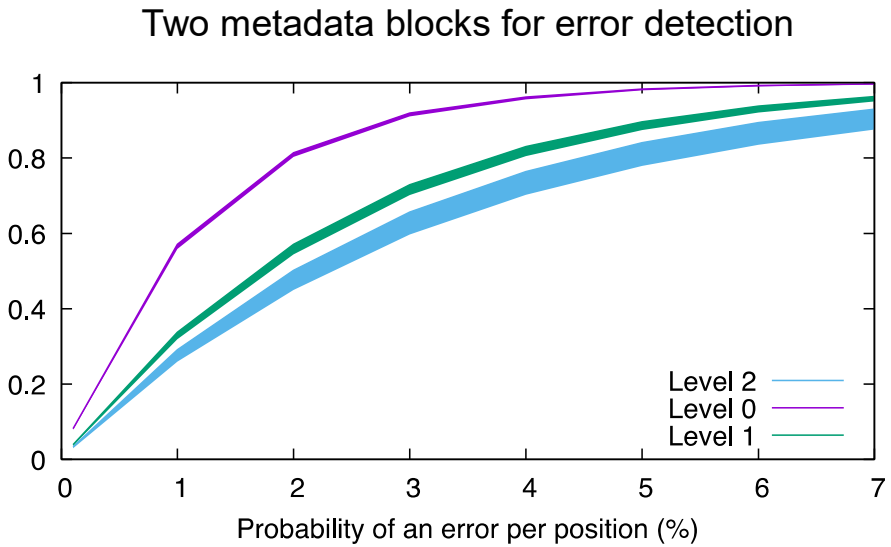
ADS Codex Parameters for 204 nt oligos

- 10 data blocks, 17 nts each = 170 nts: **1.94 b/nt**
- 8*3 nts metadata blocks + 2*5 nts RS blocks = 34 nts
- Metadata
  - 3 nts -> 47 values
  - 47^8 = 23,811,286,661,761 values
  - 2 bits for High-CG and Erasure-Oligo flags
  - 5,952,821,665,440 addresses/IDs left (216 TB)
- Level 1: **1.57 b/nt**
- Erasure Coding
  - 2 (1) Erasure Oligos for every 4 data oligos
  - 4*10*32 bits = 1280 data bits per group for 1182 (985) nts
  - Level 2: **1.045 (1.254) b/nt**

Error Resiliency of Level 1

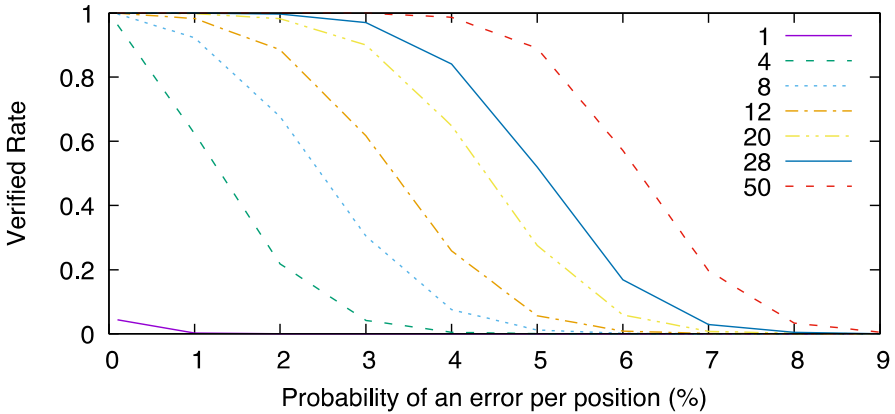
- Vary the probability of errors per position
- Line thickness shows the false positive error rate



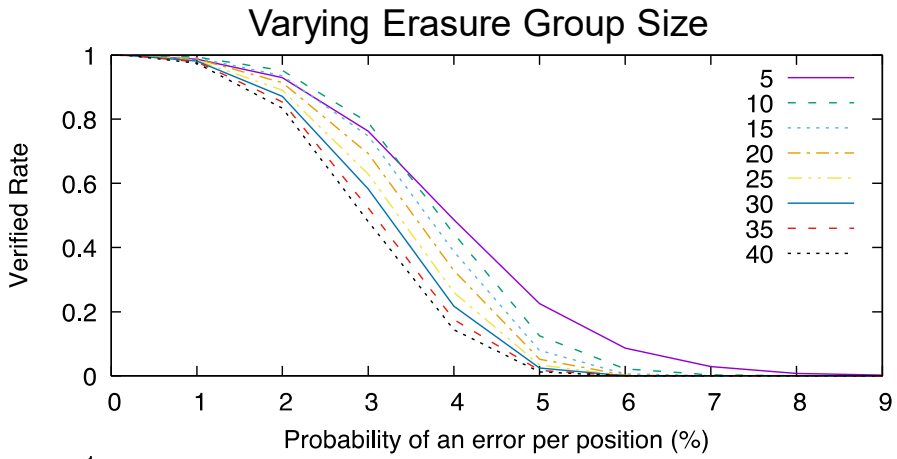


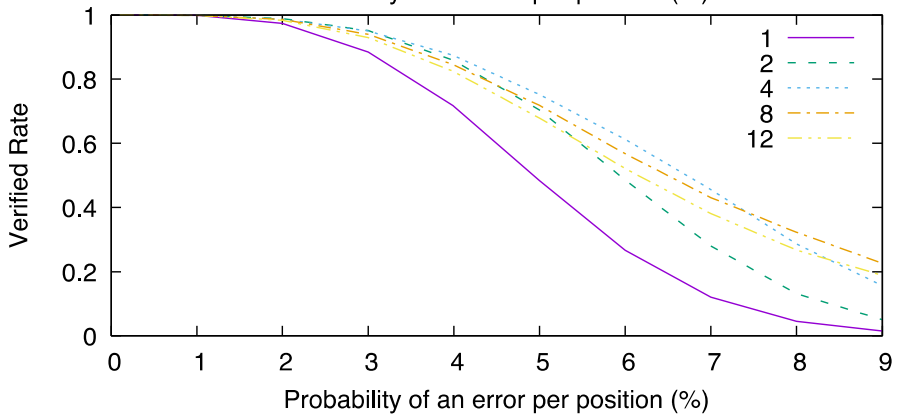
Error Resiliency of Level 2

- Vary the probability of errors per position
- Graphs show the rate of recovered data for different experiments



Varying Read Depth

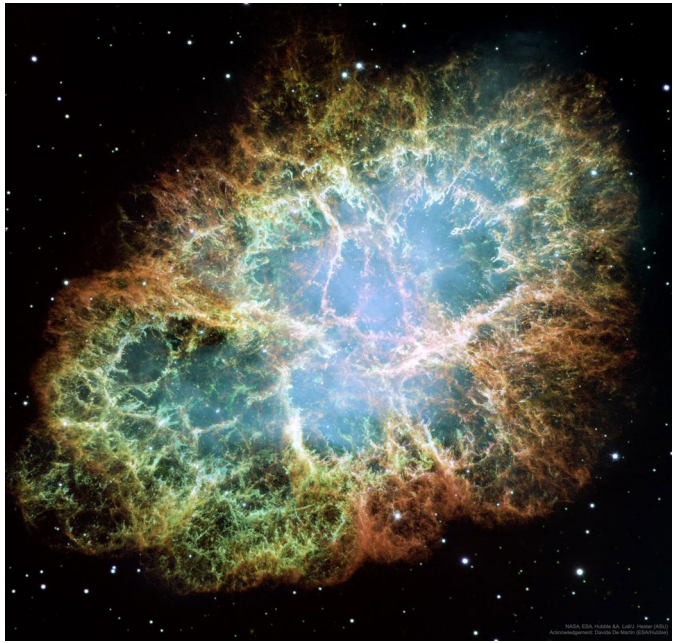




Varying Number of Erasure Oligos

ADS Codex Digital Data Recovery

- 823,630 bytes image
- 99,856 oligos
- Parameters
  - 3 data blocks
  - 5 nts per metadata block
  - Erasure group: 11 data/5 erasure



Syn ID	Payload	5' primer	3' primer	Total	Process
163	66	33	34	133	Regular
164	66	58	61	185	Regular
165	66	33	34	133	High Fidelity

Decoding results from Downsampling: Minimum Depth that Succeeds

Sequence ID	Total Reads	Depth	Unique Reads	Matched
163R1	500,000	5	234,146	96,446
163R2	750,000	8	292,718	99,060
164R1	1,725,000	7	671,785	96,242
164R2	1,725,000	7	637,359	97,294
165R1	500,000	5	178,973	98,073
165R2	400,000	4	160,217	95,837

Decoding results from Downsampling: Max Depth that Failed

Sequence ID	Total Reads	Depth	Unique Reads	Matched	Unverified (bytes)	Holes (bytes)
163R1	400,000	4	203,669	93,348	21,296	88
163R2	600,000	6	253,082	97,923	484	0
164R1	1,380,000	6	575,072	93,059	8252	8
164R2	1,380,000	6	547,527	94,695	8740	4
165R1	400,000	4	160,555	95,969	6336	0
165R2	300,000	3	139,373	91,023	87956	748

Conclusions

- Adaptive data packing
  - Can use any oligo acceptance criteria, as long is it has reasonable locality
  - Demonstrated good data density with homopolymer restrictions
- Works with different oligo lengths
- Detects deletions and insertions and can align data and metadata blocks
- 2D erasure coding that corrects for spatial error bias and allows multiple erasure oligos