STORAGE DEVELOPER CONFERENCE

SDC21

BY Developers FOR Developers

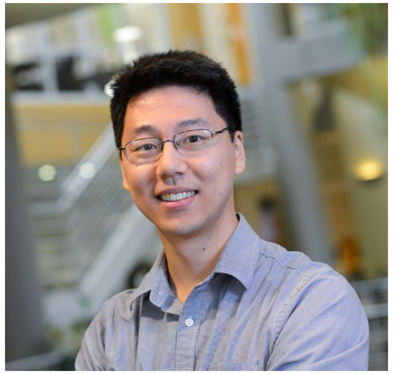
Virtual Conference
September 28-29, 2021

Scalable and Dynamic File Operations for DNA-based Data Storage

James Tuck, Professor, NC State University / Co-Founder DNAli Data Technologies

Acknowledgement

Thanks to this amazing team and support from our sponsors.



Albert Keung, PhD
Prof. at NC State
and DNAli Co-Founder

Kyle Tomek, PhD
DNAli Co-Founder

DNALI
A MOUNTAIN OF DATA IN A DROP

Funding from:

NSF

NC STATE UNIVERSITY

NC Biotech Center

GAAN Fellowship

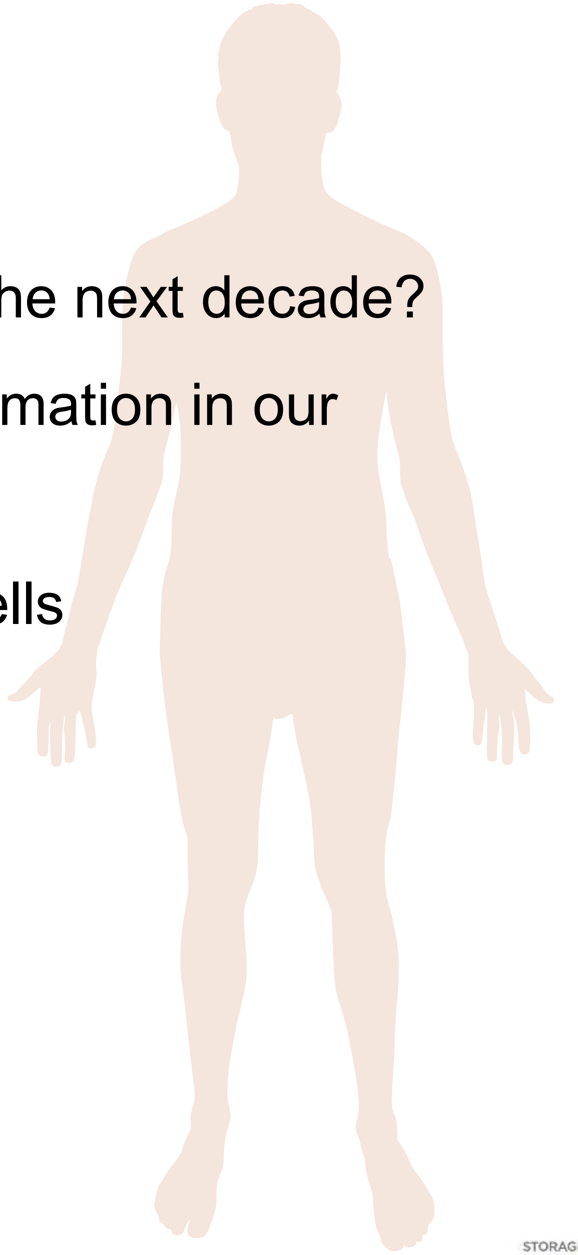
Team Members in Keung Lab and Tuck Group:

Grad Students: Kevin Lin, Kevin Volkel, Karishma Matange, Doug Townsend, Magdelene Lee, Alexander Simpson

Undergraduates: Elaine Indermaur, Austin Hass, Zach McCracken, Sarah Orr, Connor Boyce, Sam Crochet, Kathy Tran, Noah Eggenschwiler

The Extreme Density of DNA

- Where will we store the > 100 zetabytes (10^{21}) projected over the next decade?
- Each of us has the equivalent of the world's current digital information in our body (and that's just the DNA)
- Human body = 12 zeta bytes (10^{21}) = 1.2 GB/cell x 10 trillion cells
  - Where 0's and 1's can be converted into the A-G-C-T's of DNA
- Also, it has the potential to offer:
  - Century-scale stability for archival storage
  - Easily transported (clandestinely)
  - Rapid copying through molecular biology processes



Extreme densities pose extreme challenges

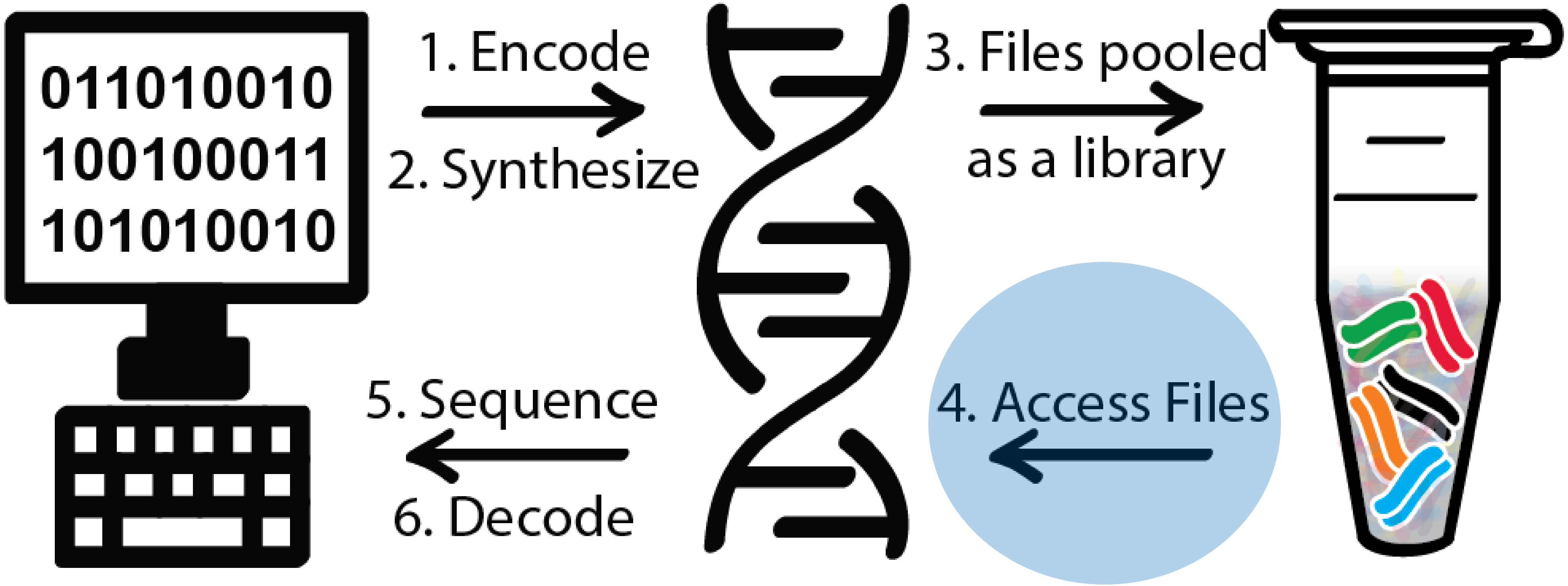
- Synthesis, sequencing, and physical manipulation of trillions upon trillions of DNA molecules
  - These costs are improving exponentially over time

- Molecular crowding (a LOT of diverse DNA in a small volume)

- Data organization and retrieval from crowded, complex mixtures

- Useful file and data functions

A DNA storage system in a nutshell



Data stored in a million trillion strands of DNA

Some important background work

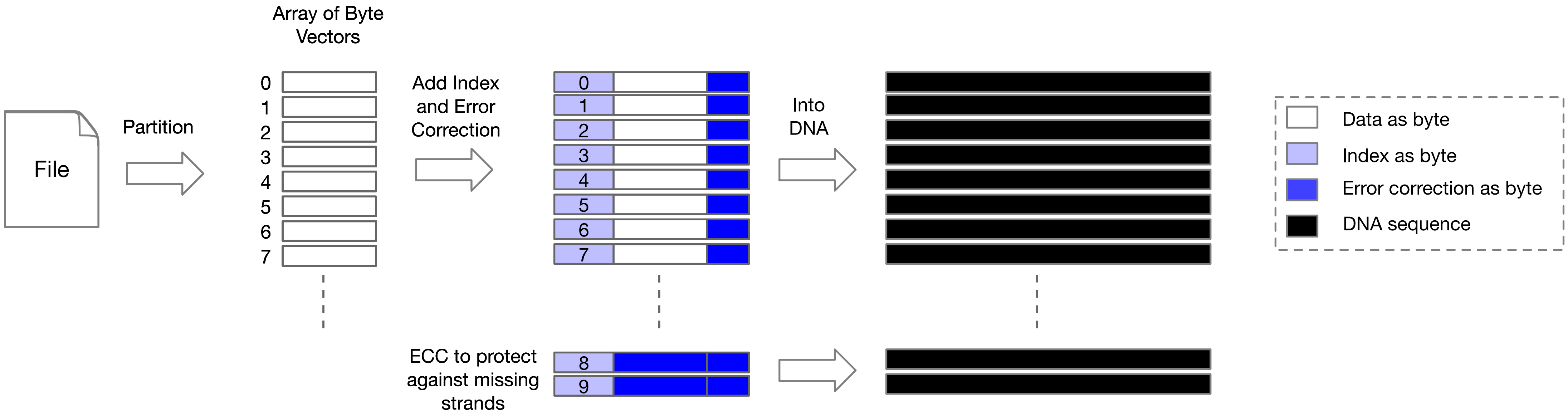
- Clelland et al. *Hiding messages in DNA microdots*, Nature,1999.
  - A hidden message was flanked with primers and obscured with genomic DNA.
- Bancroft et al*. Long-Term Storage of Information in DNA*, Science, 2001.
  - Footnote 9 observes that PCR plus sequencing is essentially random access memory.
- Church et al. *Next Generation Information Storage in DNA*, Science, 2012.
- Goldman et al. *Towards practical, high-capacity, low-maintenance information storage in synthesized DNA*. Nature, 2013.
- Grass et al. *Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes,* Angewandte Chemie, 2015.
  - First deep look at the long term stability of DNA for holding information.
- Yazdi et al. *A Rewritable, Random-Access DNA-Based Storage System*, Scientific Reports, 2015.
  - Imagines a way of using PCR to rewrite data stored in DNA.
- Organick et al. *Random access in large-scale DNA data storage*, Nature Biotech, 2018.
  - Deep look at scaling up random access in DNA using PCR.

A brief review of access

- Strand design
  - Each DNA strand is synthesized as a linear data structure that enables access and decoding
- Separation
  - Removal of strands from the library for sequencing or manipulation
- Polymerase Chain Reaction (PCR)
  - Common technique for amplifying strands of interest to prepare them for sequencing

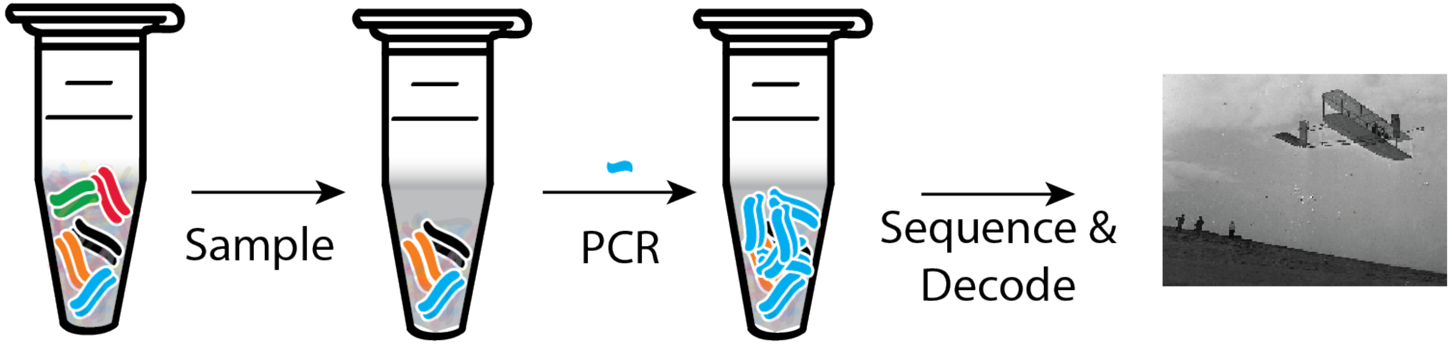
Basic encoding strategy for a single file/object

- Relatively short DNA strands < 200 nt.
- Includes the index as part of the strand.
- Error correction codes are added to ensure successful retrieval.



Access a file from a DNA library/database

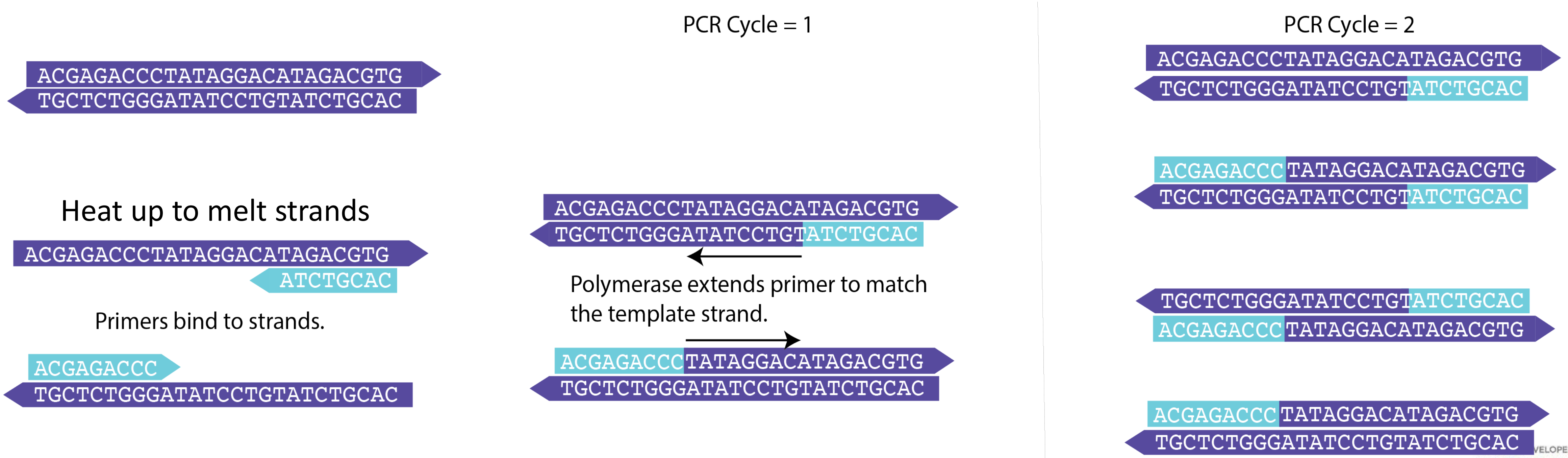
- Sample, amplify with PCR, sequence, decode



- PCR makes many copies of only the desired (blue) strands so that they are overwhelmingly represented.
  - PCR is like a memcpy that starts copying at a specified (blue) sub-sequence
- The removed sample is destroyed in this process

Using PCR for random access

- Polymerase extends new strand from primer along template
- Primer becomes part of the new strand
- Repeated cycles grow the copies exponentially



ACGAGACCCTATAGGACATAGACGTG
TGCTCTGGGATATCCTGTATCTGCAC

Heat up to melt strands

ACGAGACCCTATAGGACATAGACGTG
ATCTGCAC

Primers bind to strands.

ACGAGACCC
TGCTCTGGGATATCCTGTATCTGCAC

PCR Cycle = 1

ACGAGACCCTATAGGACATAGACGTG
TGCTCTGGGATATCCTGTATCTGCAC

Polymerase extends primer to match the template strand.

ACGAGACCCTATAGGACATAGACGTG
TGCTCTGGGATATCCTGTATCTGCAC

PCR Cycle = 2

ACGAGACCCTATAGGACATAGACGTG
TGCTCTGGGATATCCTGTATCTGCAC

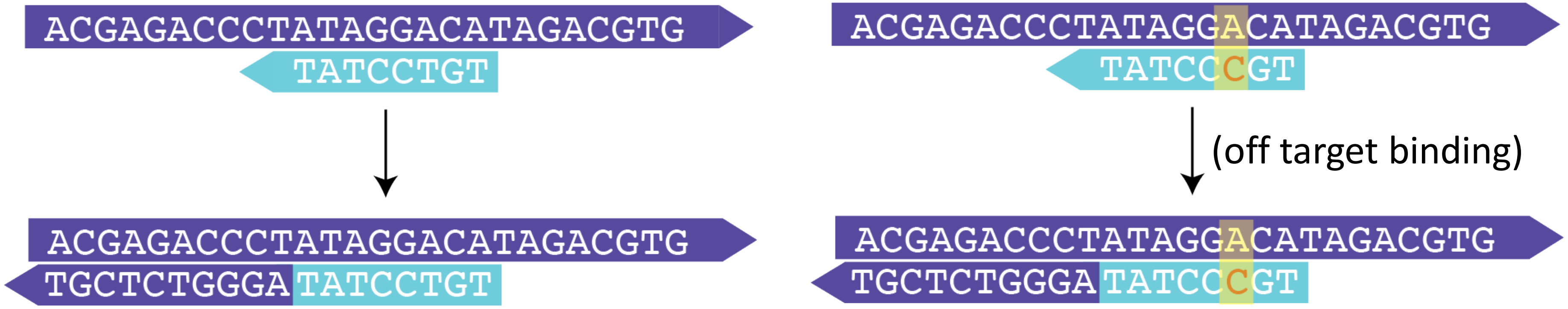
ACGAGACCCTATAGGACATAGACGTG
TGCTCTGGGATATCCTGTATCTGCAC

TGCTCTGGGATATCCTGTATCTGCAC
ACGAGACCCTATAGGACATAGACGTG

ACGAGACCCTATAGGACATAGACGTG
TGCTCTGGGATATCCTGTATCTGCAC

PCR caveats

- Primer can bind anywhere favorable (few mismatches)



ACGAGACCCTATAGGACATAGACGTG
TATCCTGT

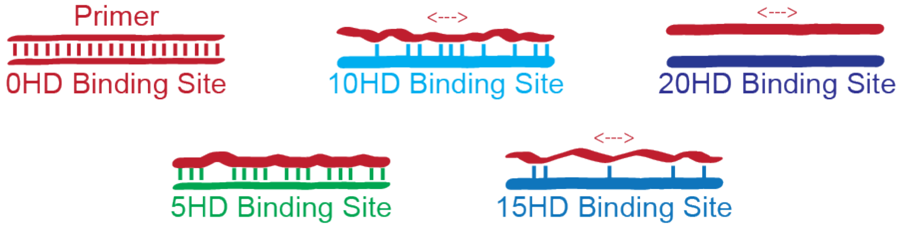
ACGAGACCCTATAGGACATAGACGTG
TGCTCTGGGA TATCCTGT

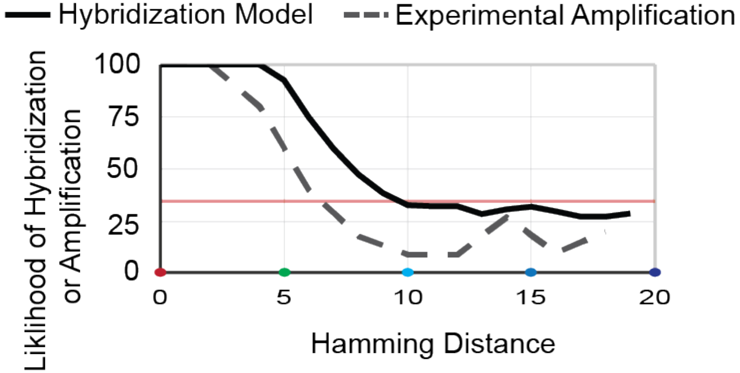
ACGAGACCCTATAGGACATAGACGTG
TATCCCGT

(off target binding)

ACGAGACCCTATAGGACATAGACGTG
TGCTCTGGGA TATCCCGT

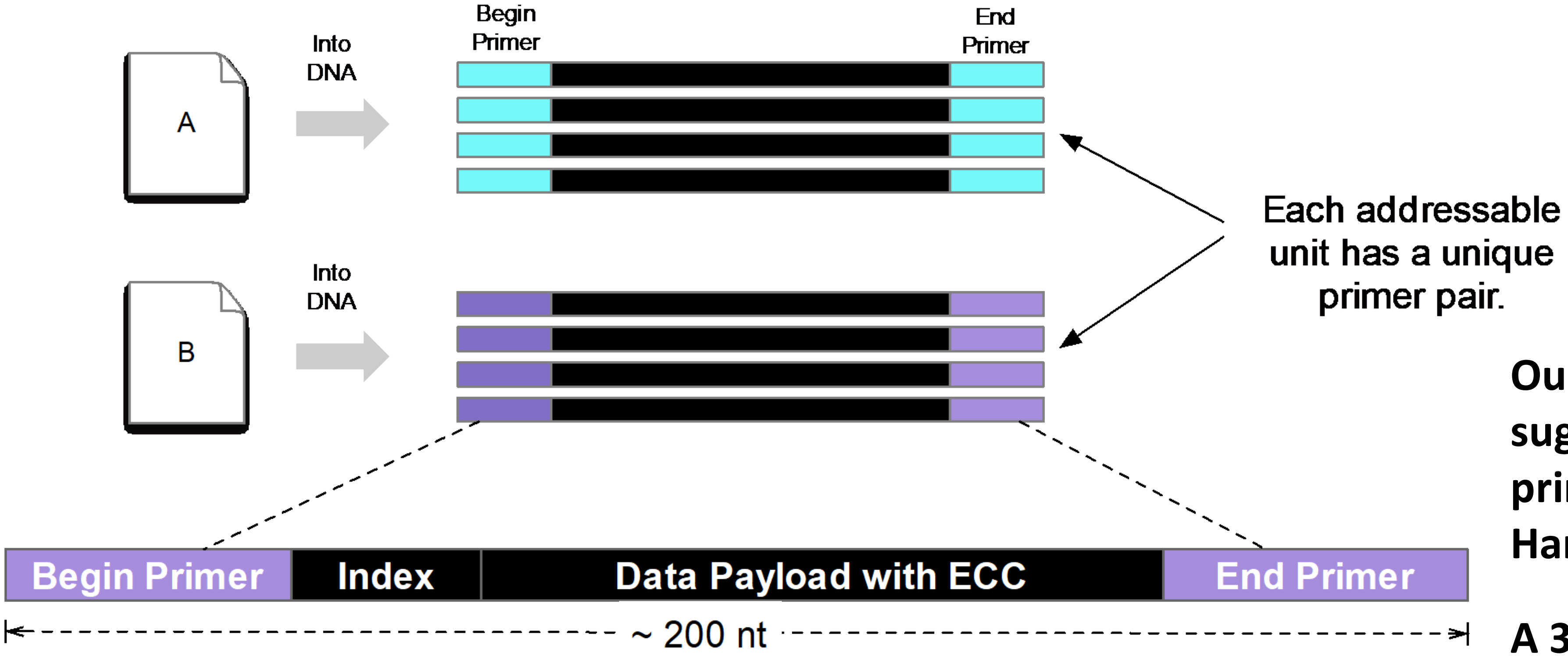
- Primers should only bind to the strands we want to amplify
- Primers for random access are picked at high Hamming distances





Random access

- Each file or “addressable unit” gets a primer sequence tuned to bind only to its own strands



Each addressable unit has a unique primer pair.

Our analysis and past work suggests there are ~3000 primers of length 20 nt and Hamming distance=10.

A 3 GB block/file size implies 9 TB pool of DNA.

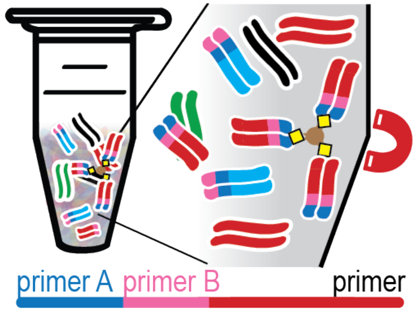
Observations

- Strand structure and access mechanisms are co-designed
- Pool capacity is limited by the number of primers
- Primers must avoid unintentional interactions with other primers and data payloads leading to few ”good ones”
- Destructive access will limit the number of reads before losing or rewriting data

Questions – Can we …

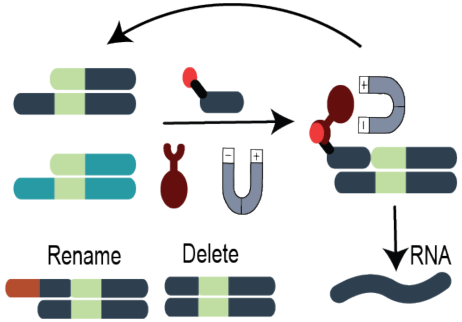
- Scale to much higher capacities and still access our data?
- Provide less destructive access modes?
- Offer additional in-storage functionalities?

Our focus on access has revealed several interesting system designs



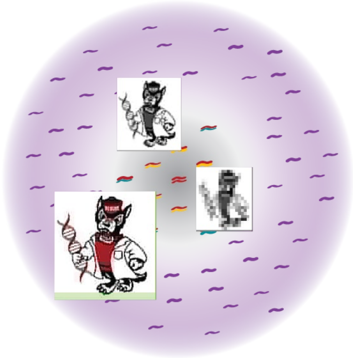
DENSE: Nested primers with DNA enrichment enable efficient access in high capacity pools.

Tomek et al. ACS SynBio 2019.



DORIS: repeatable information access with dynamic in-storage file operations.

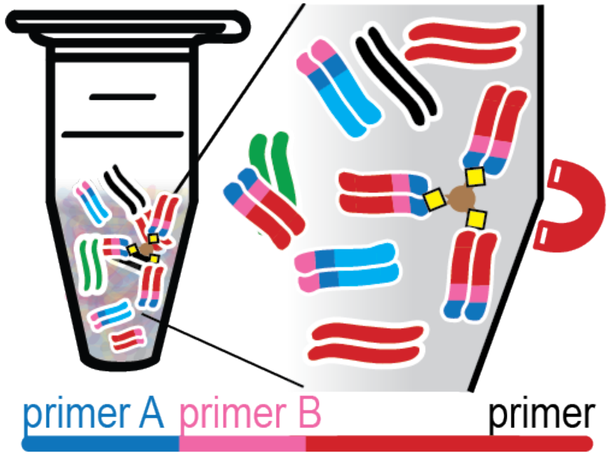
Lin et al, Nature Comm. 2020.



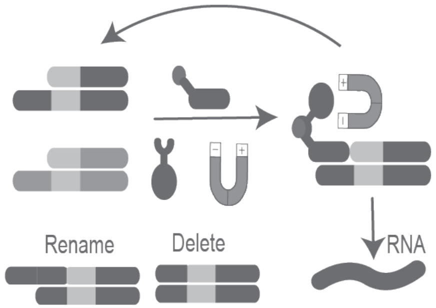
File Preview: exploit primer promiscuity to access a file partially or fully.

Tomek et al, Nature Comm. 2021.

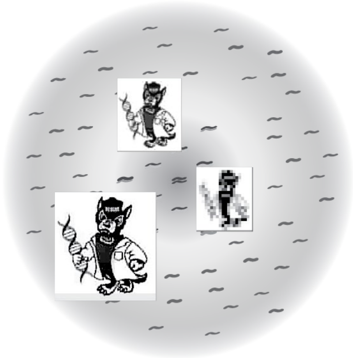
Outline



DENSE: Nested primers with DNA enrichment enable efficient access in high capacity pools.



DORIS: repeatable information access with dynamic in-storage file operations.

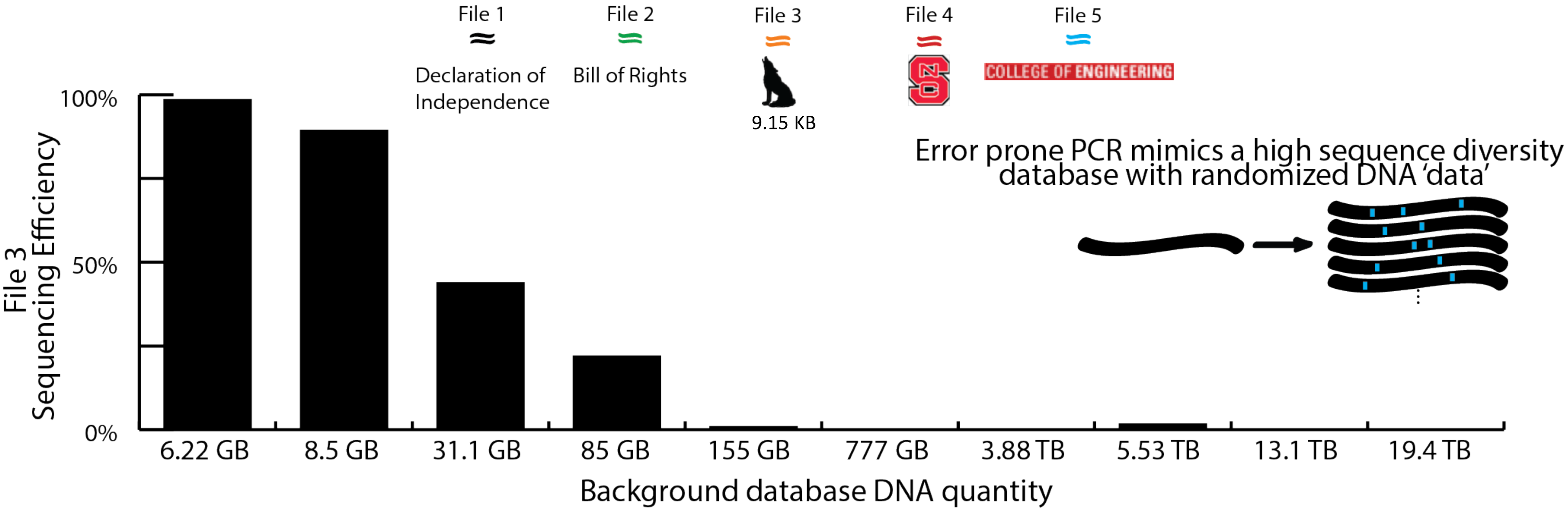


File Preview: exploit primer promiscuity to access a file partially or fully.

Motivation for DENSE

- Scale-up capacity to higher than 9 TB = 3000 x 3 GB?
- Access data efficiently in a scaled-up capacity system?
- Retain the library on an access?

PCR access in scaled-up system is inefficient



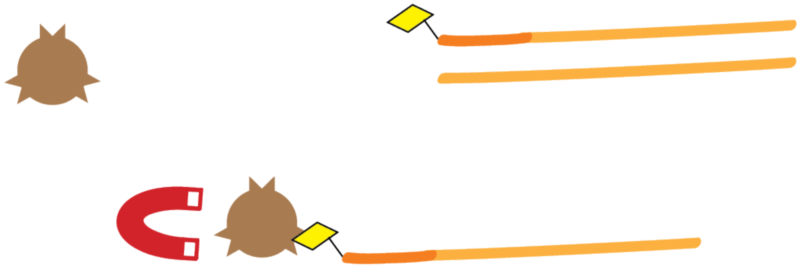
Error prone PCR mimics a high sequence diversity database with randomized DNA 'data'

There are so many library strands in a scaled-up system that they still overwhelm sequencing even after 30 cycles of ePCR.

DNA enrichment rescues efficiency

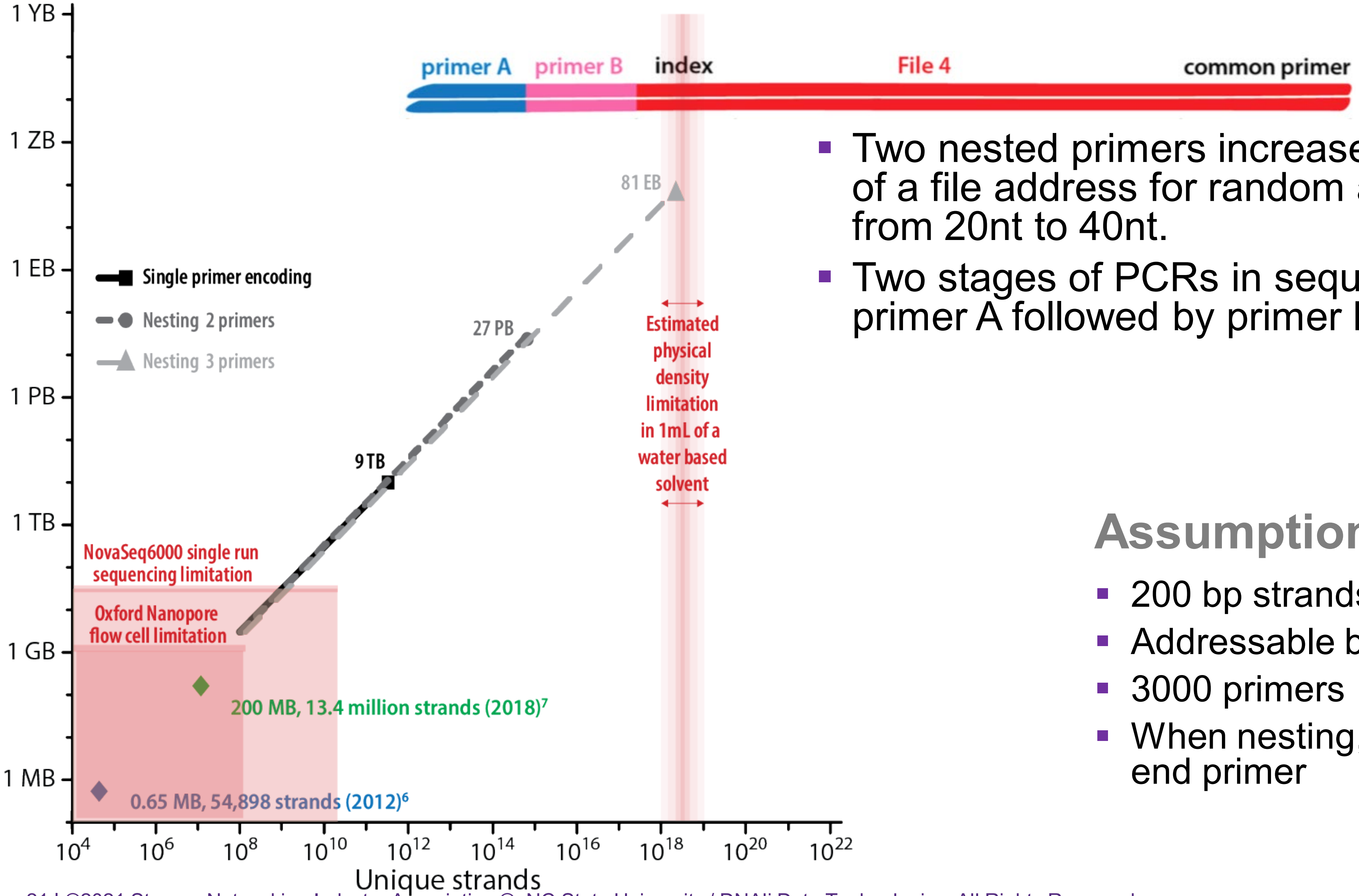
Streptavidin coated magnetic bead

PCR with Biotin labeled primer



Biotin has strong attraction to streptavidin

Boost capacity by nesting primers



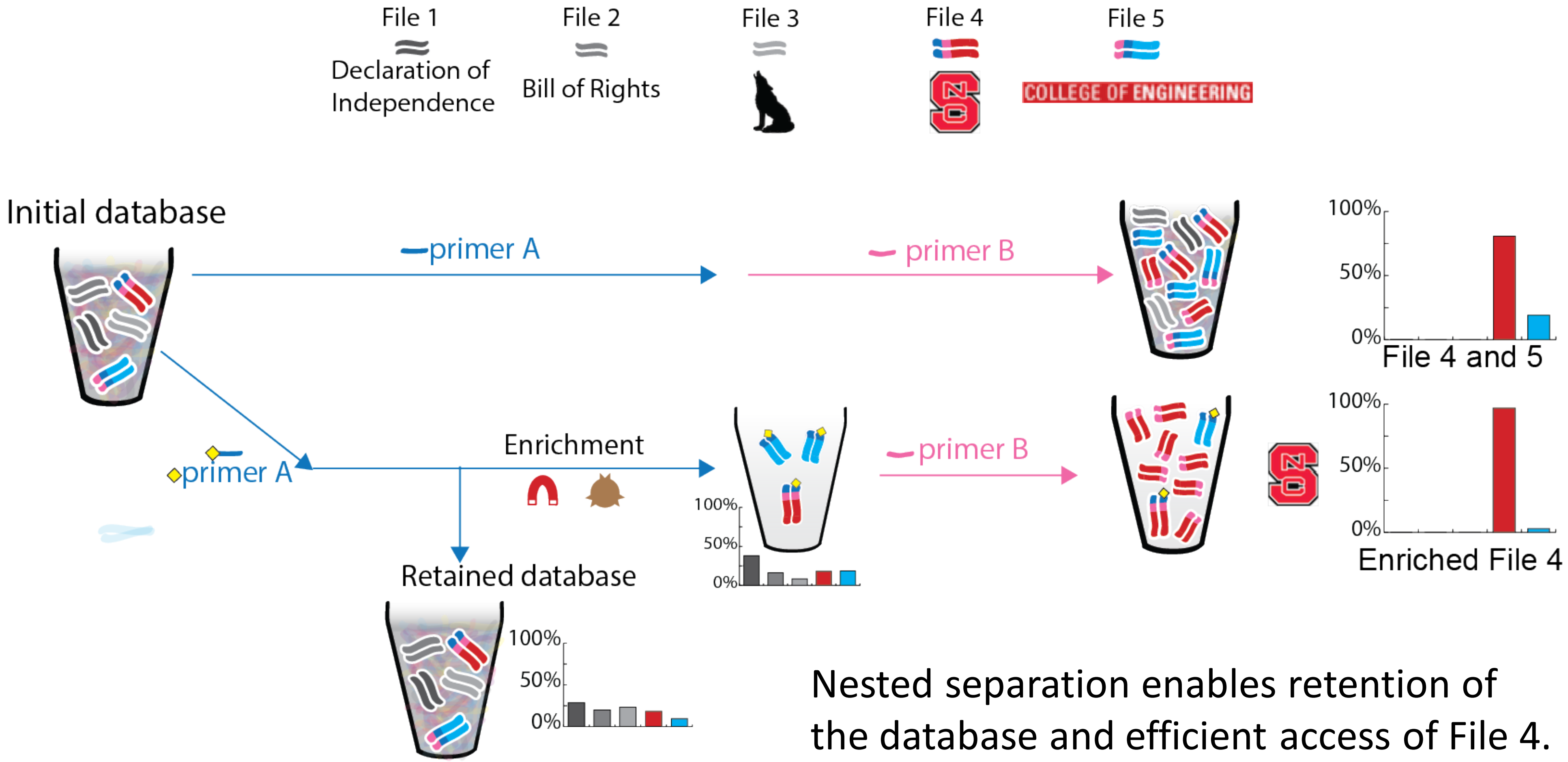
- Two nested primers increase the width of a file address for random access from 20nt to 40nt.
- Two stages of PCRs in sequence on primer A followed by primer B

Assumptions

- 200 bp strands
- Addressable block size of 3 GB
- 3000 primers
- When nesting, leave out one common end primer

Nested primers with enrichment is more efficient

- Use nested PCRs to access a file with nested primers.

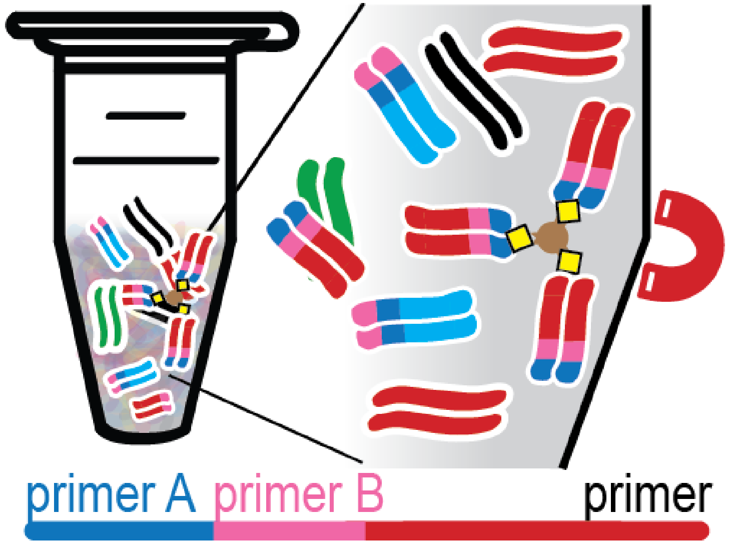


Nested separation enables retention of the database and efficient access of File 4.

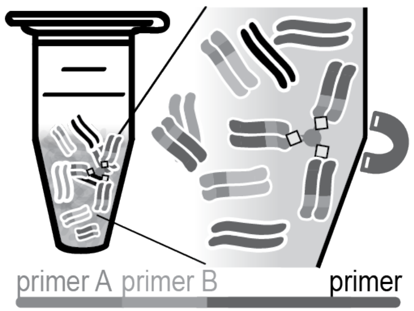
DENSE: DNA Enrichment with Nested Separation

Key take-away ideas

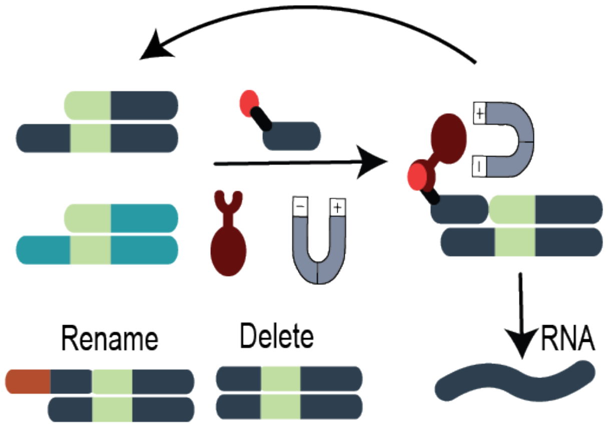
- DNA enrichment with magnetic beads enables specific access in dense backgrounds and retention of original data strands.
- Nest primers to increase address space and boost capacity.
- Nested primers benefit from enrichment for efficient access.



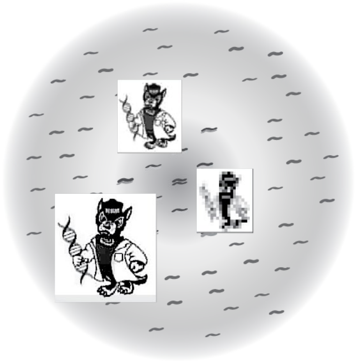
Outline



DENSE: Nested primers with DNA enrichment enable efficient access in high capacity pools.



DORIS: repeatable information access with dynamic in-storage file operations.



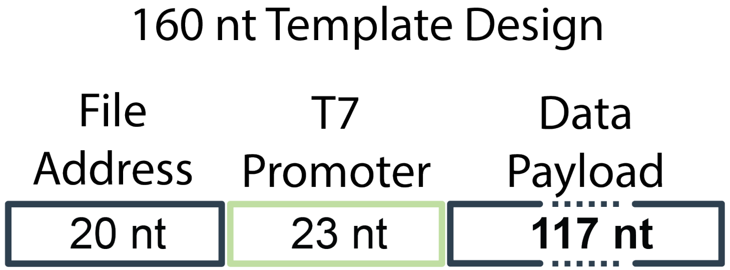
File Preview: exploit primer promiscuity to access a file partially or fully.

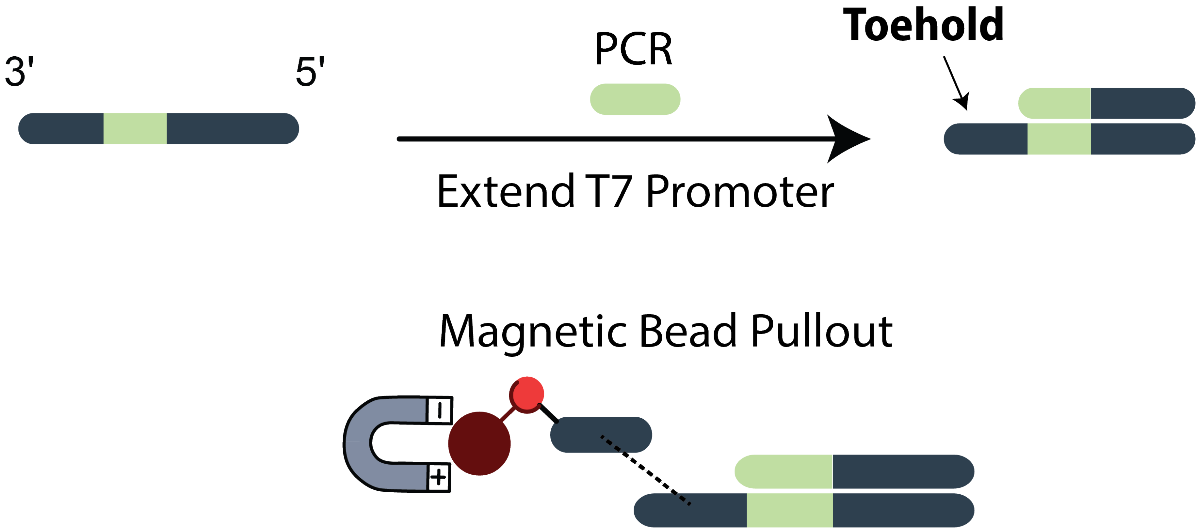
Motivation for DORIS

- Can we retain the library on repeated accesses in a similar way that eukaryotic cells use RNA?

- Can we offer additional in-storage functionalities?

A toehold structure offers extra functionality

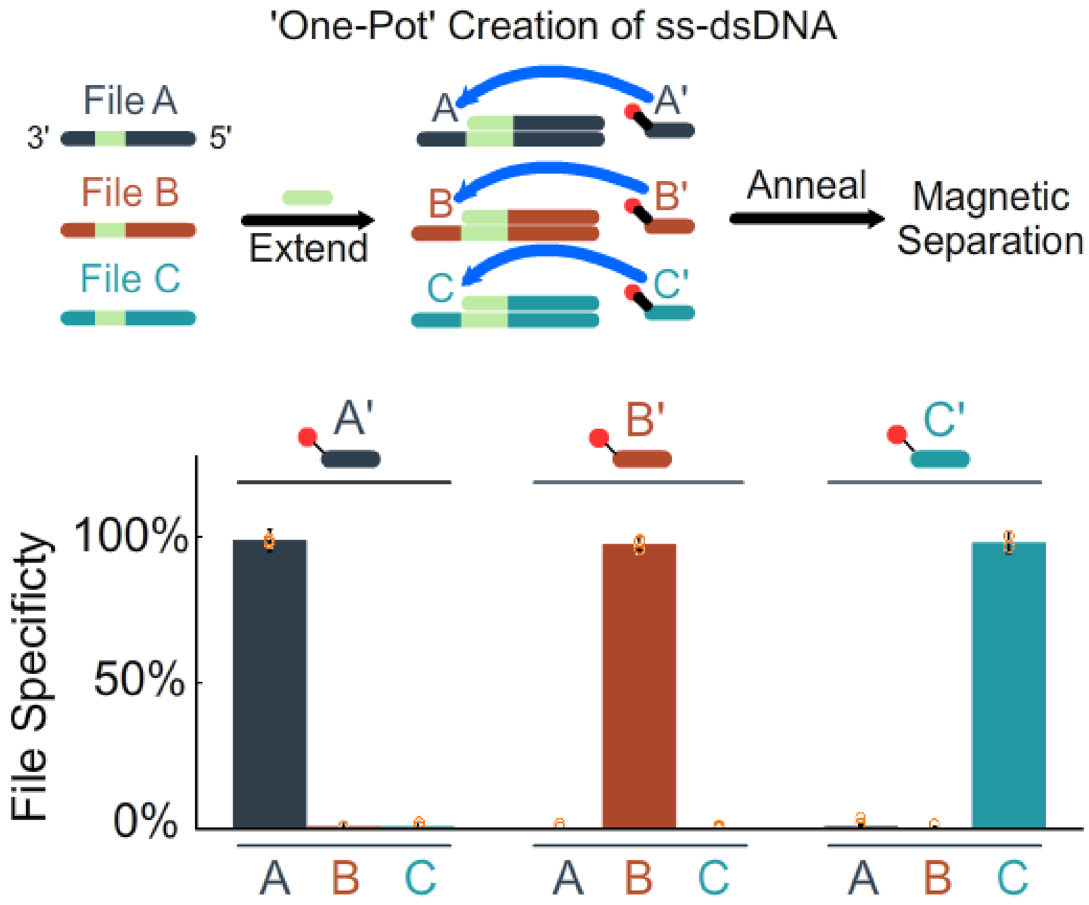




T7 promoter is a special sequence recognized by the T7 RNA polymerase.

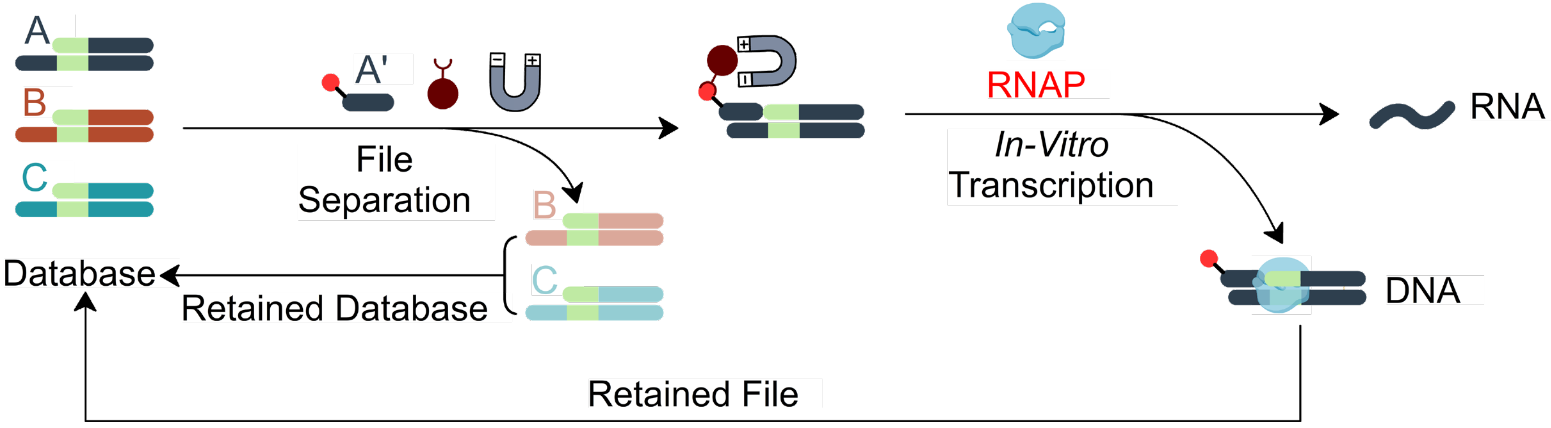
Toeholds are created in parallel

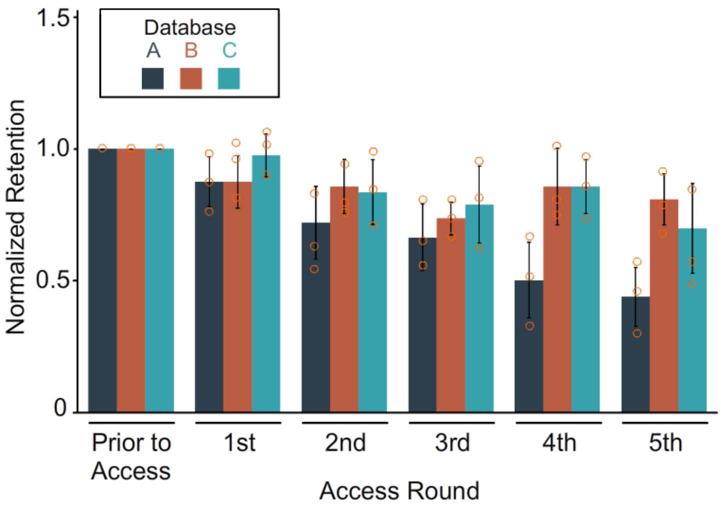
- Toehold created in parallel through PCR on T7
- One-pot creation observed to yield good file specificity



File retained after repeated accesses

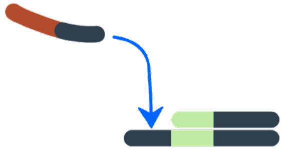
- Transcription to RNA occurs on bead in *near-room-temperature* conditions
- DNA bound to bead is returned to the database, preserving it
- Repeated accesses to File A demonstrate good retention
  - File B and C are roughly stable, File A loses 50% over 5 accesses
  - Better than destroying a fraction of the library



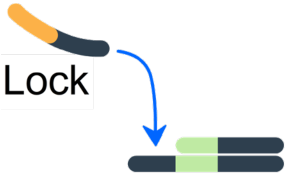


Toehold structure enables in-storage operations

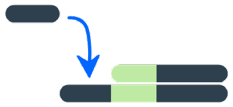
Rename

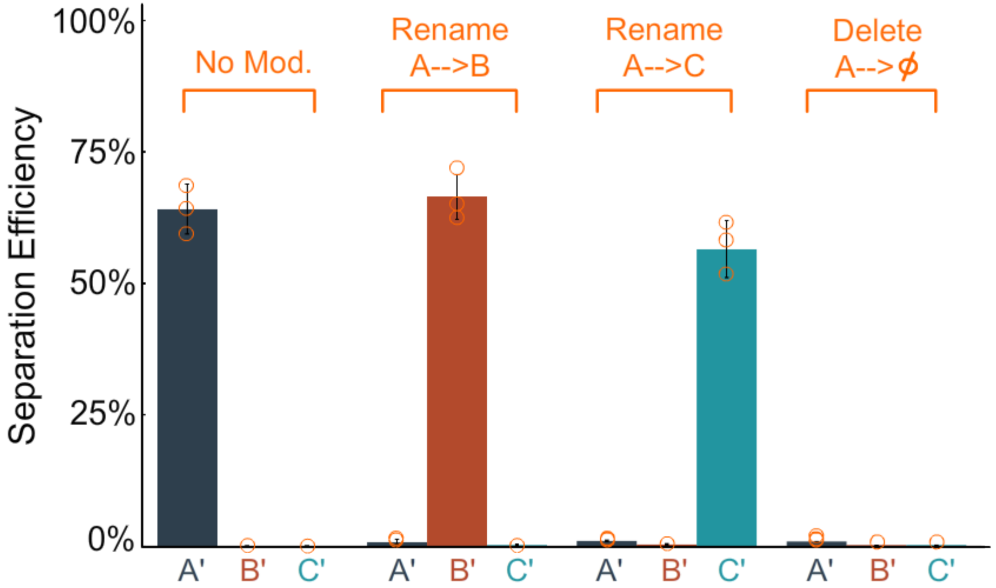


Lock/Unlock



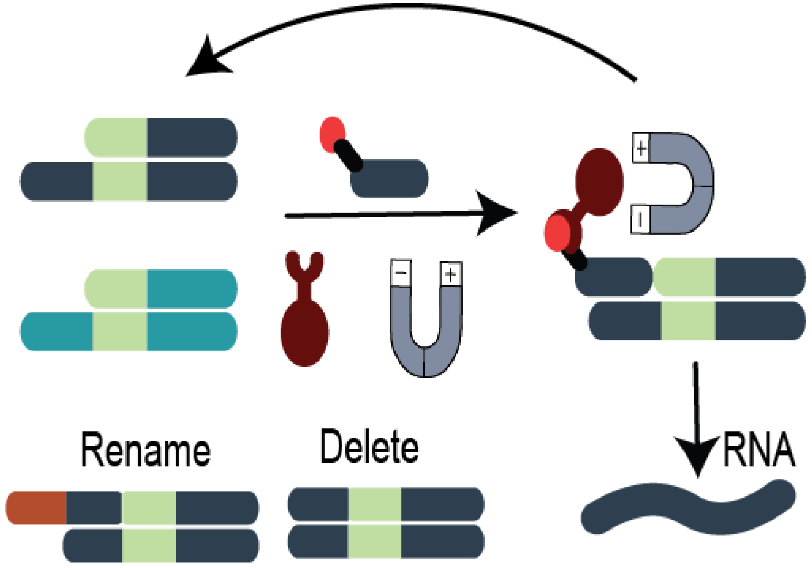
Delete



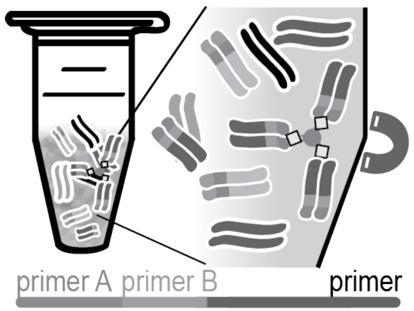


DORIS offers repeatable information access with dynamic file operations

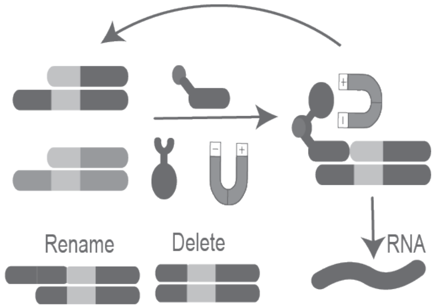
- **Key take-away ideas**
  - DNA strands are constructed with a single-stranded toehold "file address"
  - Transcription to RNA while on the bead allows retention of original library strands
  - Toehold enables in-storage file operations like renaming, deletion, lock/unlock



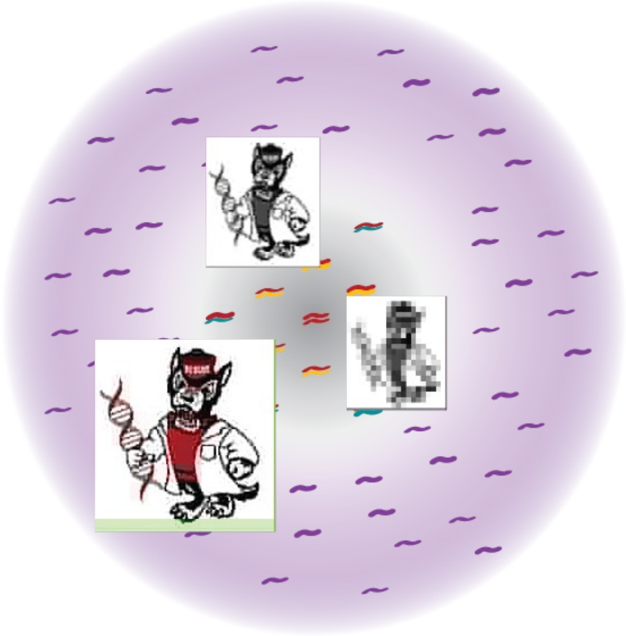
Outline



DENSE: Nested primers with DNA enrichment enable efficient access in high capacity pools.



DORIS: repeatable information access with dynamic in-storage file operations.

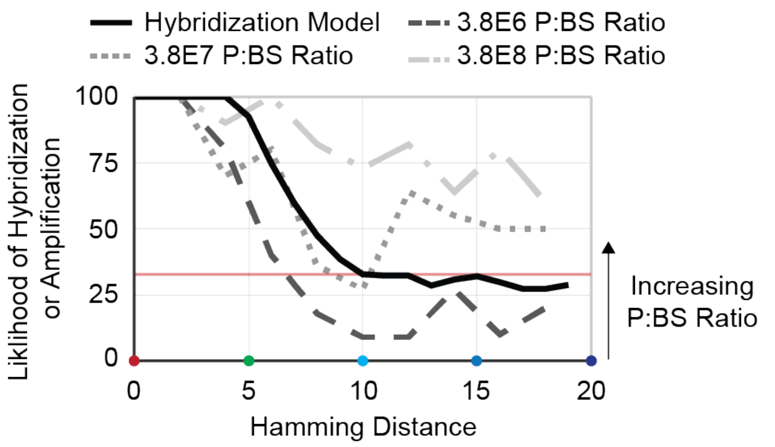


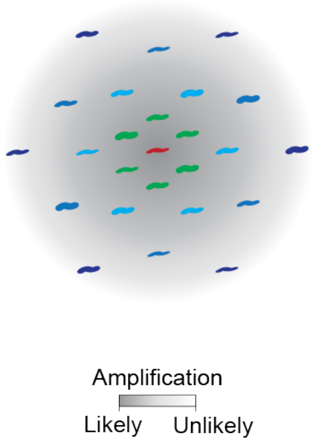
File Preview: exploit primer promiscuity to access a file partially or fully.

Motivation for File Preview

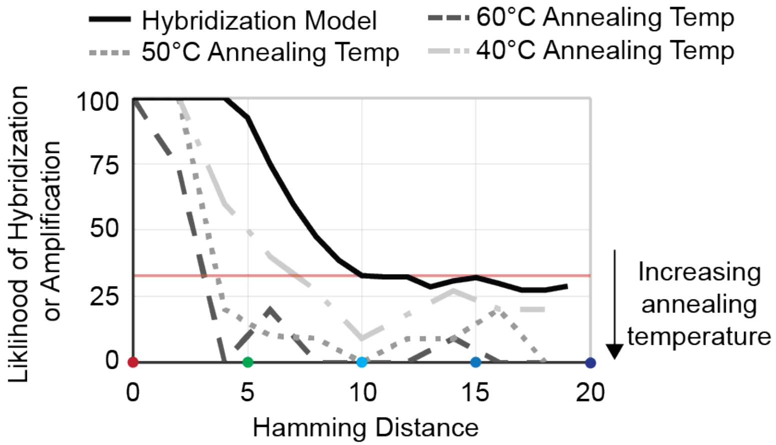
- Can we offer additional in-storage functionalities?
- Can we take advantage of primer’s willingness to bind off-target?

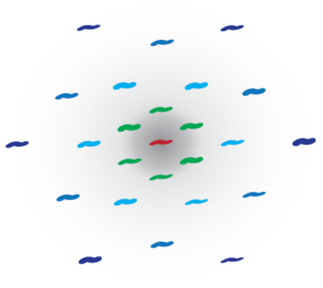
Primers offer tunable specificity





Higher primer concentration lowers specificity.
The amplification radius widens, access more data.

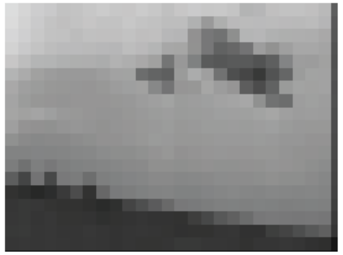




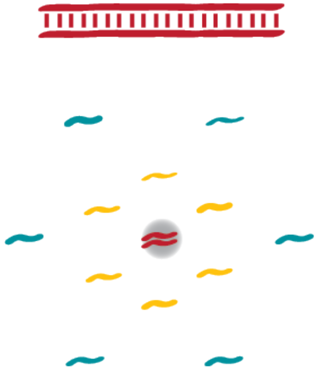
Higher temperature increases specificity.
The amplification radius narrows, access less data.

File encoding and access for “preview”

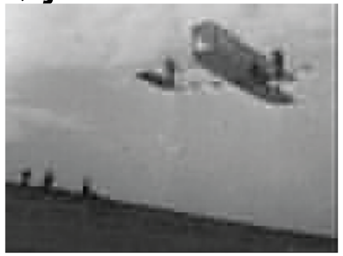
- Use JPEG Progressive Encoding to partition strands among ”preview” and full
- Always access with the same primer, just alter access conditions



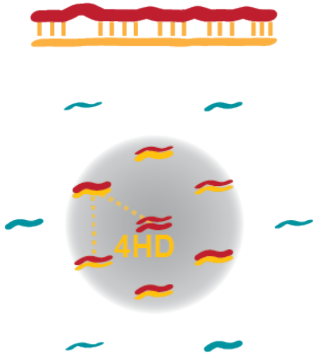
2% of file, 0 HD



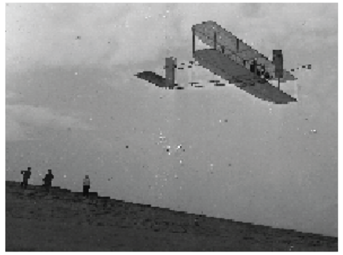
60°C, 250nM primer,
0.75mM MgCl2,
50mM KCl,
20s anneal and
20s extension



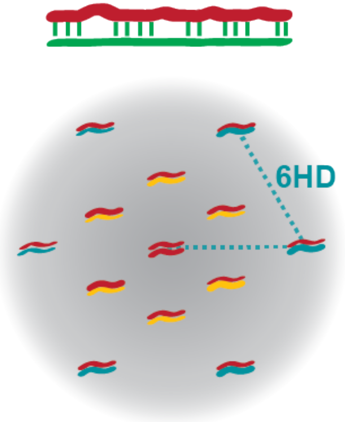
10% of file, 0 to 4 HD



40°C, 1000nM primer,
3mM MgCl2,
200mM KCl,
0.1% Triton X-100
90s anneal and
90s extension

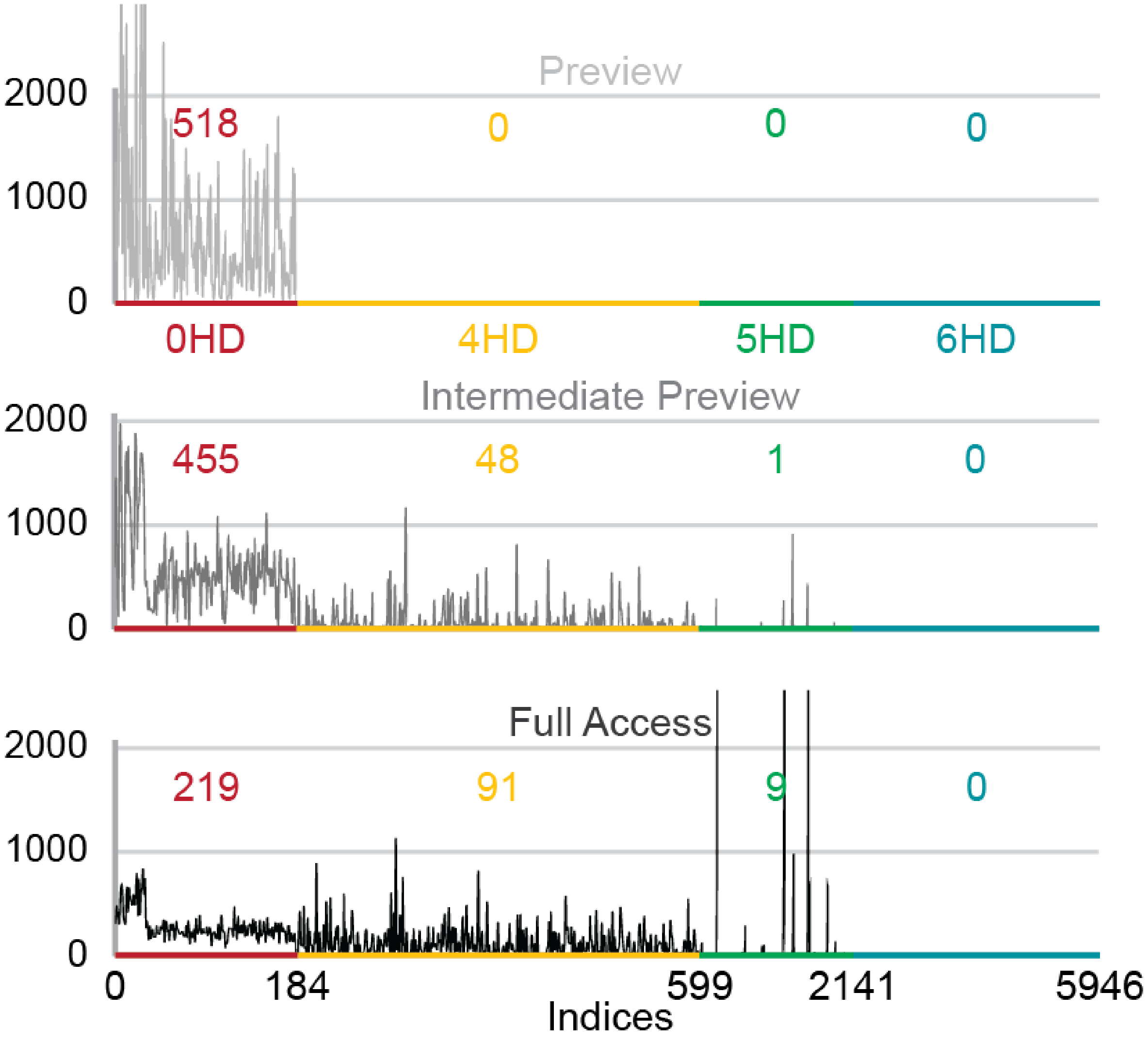


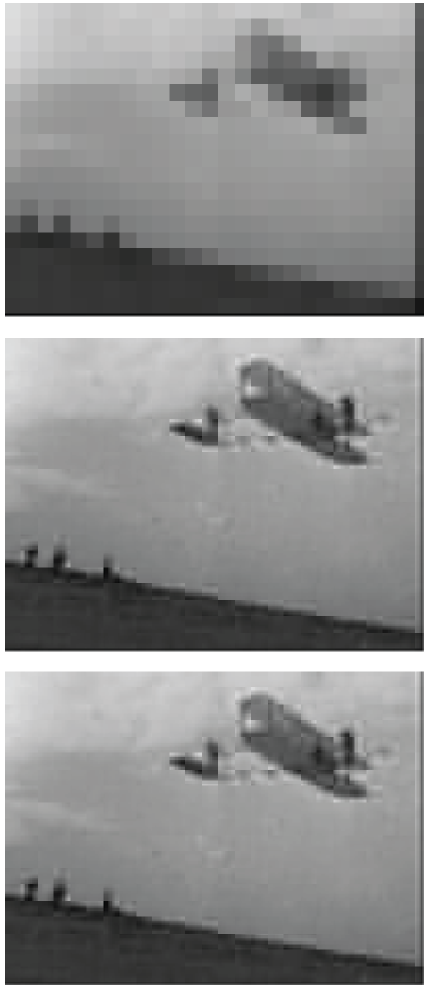
100% of file, 0 to 6 HD



45°C, 500nM primer,
1.5mM MgCl2,
50mM KCl,
60s anneal and
60s extention

Preview, intermediate, and full access results



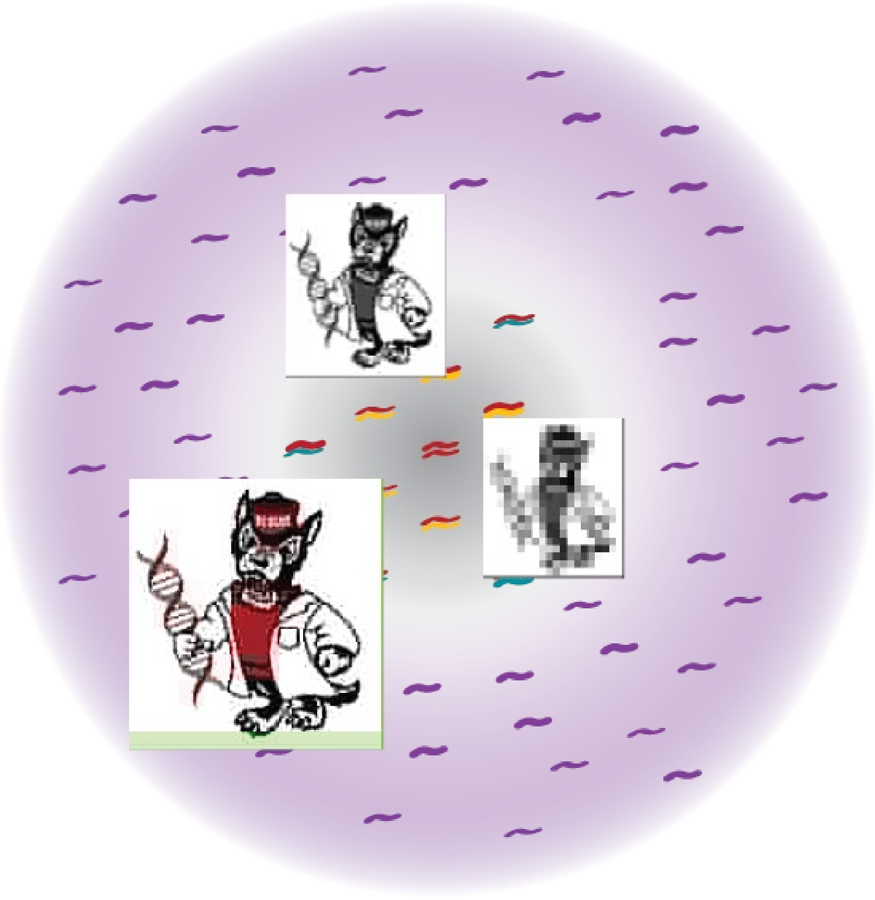


- ~50 previews can be sequenced and decoded for the same cost as full access
- We get this by trading-off density—more copies of high HD strands are needed.

File Preview

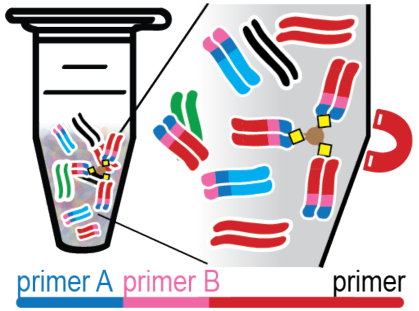
Key take-away Ideas

- Specificity of primer access is tunable based on temperature and concentration
- We designed access protocols and encodings for full file access versus partial file access using these knobs
  - Fuzzy image vs. full detail



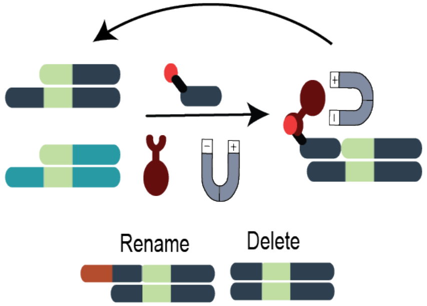
Outline

You can find additional interesting results in our papers!



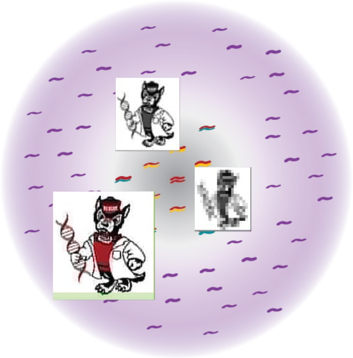
DENSE: Nested primers with DNA enrichment enable efficient access in high capacity pools.

Tomek et al. ACS SynBio 2019.



DORIS: repeatable information access with dynamic in-storage file operations.

Lin et al, Nature Comm. 2020.



File Preview: exploit primer promiscuity to access a file partially or fully.

Tomek et al, Nature Comm. 2021.

Summary

- Scale to much higher capacities and still access our data?
  - Nested primers enable a larger address space
  - Magnetic bead pullouts using biotin labeled primers enable efficient access
- Provide less destructive access methods?
  - Biotin labeling enables separation of copied strands from original library
  - RNA transcription directly from linked beads enables separation of RNA from DNA and retention of the library
- Offer additional in-storage functionalities?
  - Tuning access conditions enables preview versus full access of a file
  - Toehold strand design enables in-storage operations like delete and rename

Questions?

jtuck@ncsu.edu

info@dnalidata.com