Shifted Hamming Distance (SHD): A Fast and Accurate SIMD-Friendly Filter for Local Alignment in Read Mapping

Hongyi Xin[1], John Greth[1], John Emmons[1], Gennady Pekhimenko[1], Carl Kingsford[1], Can Alkan[2], Onur Mutlu[1]

[1] Departments of Computer Science and Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

[2] Dept. of Computer Engineering, Bilkent University, Ankara, Turkey

Problem:

- NGS mappers can be divided into two categories: **Suffix-array based** and **seed-and-extend based**
  1. Suffix-array based mappers (i.e. bwa, bowtie2, SOAP3) find **the best mappings fast** but **lose high-error mappings**
  2. Seed-and-extend based mappers (i.e., mrfast, shrimp, RazerS3) finds **all mappings** but **waste resources** on rejecting **incorrect mappings**
- **Our goal:** Provide an effective filter to **efficiently** filter out incorrect mappings

Shifted Hamming Distance (SHD):

- **Key idea:** use simple **bit-parallel** and **SIMD operations** to quickly filter out incorrect mappings
- **Key observations:**
  1. If two strings differ by ≤*e* errors, then every non-erroneous bp can be aligned in **at most *e* shifts**
  2. If two strings differ by ≤*e* errors, then they share **at most (*e*+1) identical sections** (Pigeonhole Principle)
- SHD consist of **two parts**: **shifted Hamming mask-set (SHM)** and **speculative removal of short-matches (SRS)**

1 2 3 4

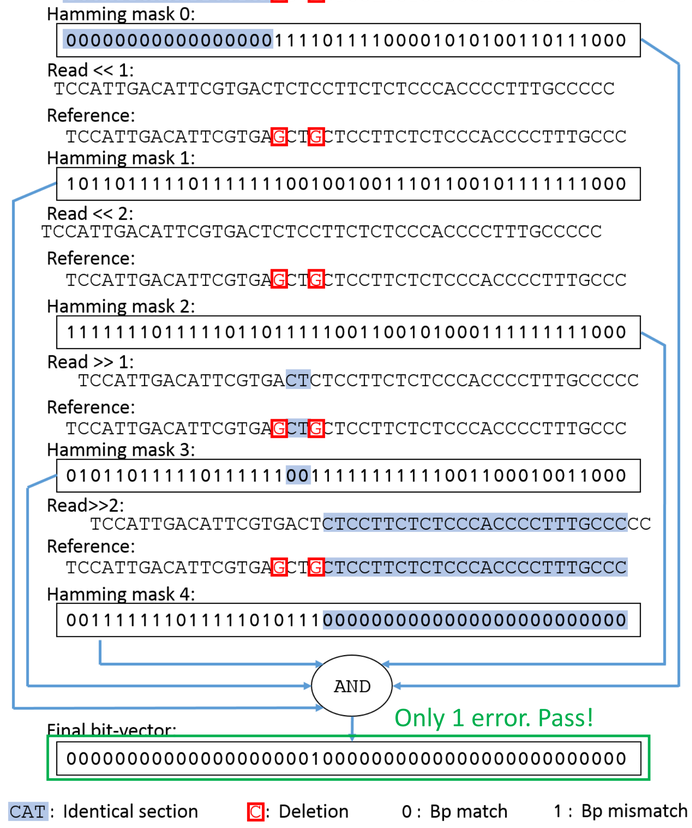
Shifted Hamming Mask-set (SHM):

- **Key idea: SHM** identifies matching by **incrementally shifting** the read against the reference
- **Mechanism:** Use bit-wise XOR to **find all matching bps**. Then use bit-wise AND to **merge them together** (Fig 1)
- Mappings that contain more than *e* '1's in the final bit-vector **must contain more than *e* errors**
- **Cons:** SHD may let incorrect mappings pass through (Fig 2) because **all '0's "survive"** the AND operations

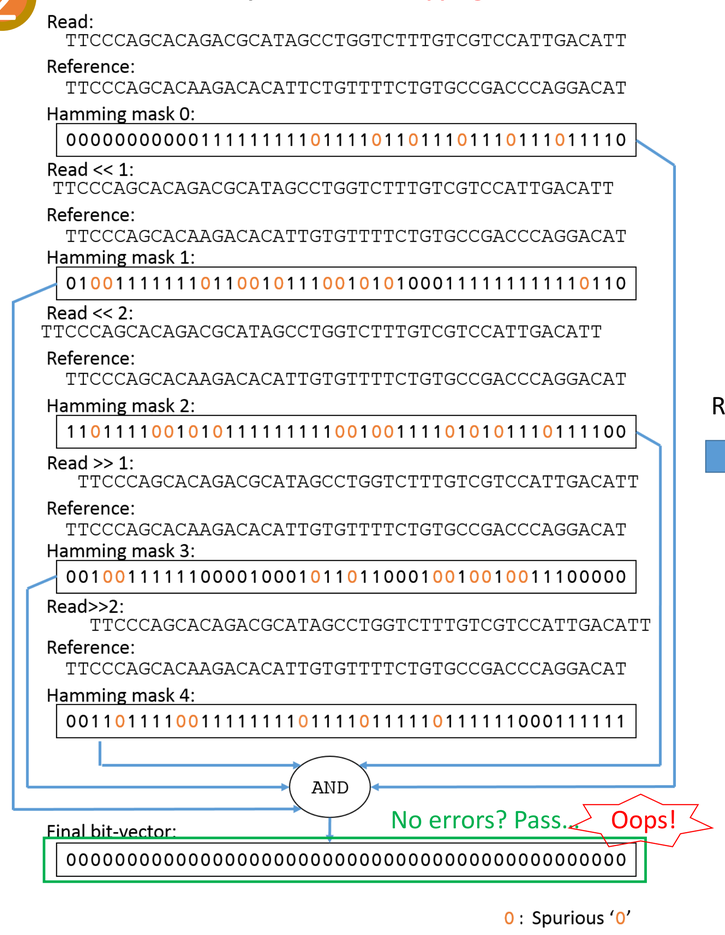
Speculative Removal of Short-matches (SRS):

- **Key idea: SRS** refines SHM by **removing short stretches of matches (<3 bps)** identified in the Hamming masks (Fig 3)
- **Key observations**:
  1. Identical sections tend to be **long (≥ 3 bps)**
  2. Short stretches of matches (streaks of '0' <3 bps) are likely to be **random matches** of bps (generate spurious '0's)
- **Mechanism:** Amend **short streak** of '0's into '1's while count errors **conservatively** in the final bit-vector (Fig 4)

1 Identifying all matching bps of a correct mapping with SHM (e=2)



2 SHM fails to identify an incorrect mapping due to random matches (e=2)



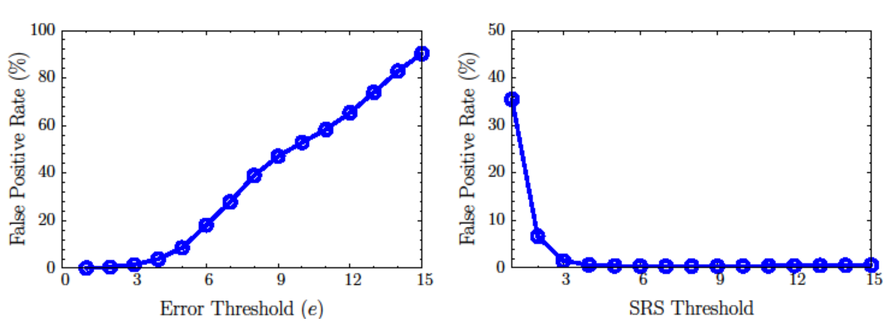
Refined by **SRS**

3 SRS removes short random matches from the Hamming masks (e=2)

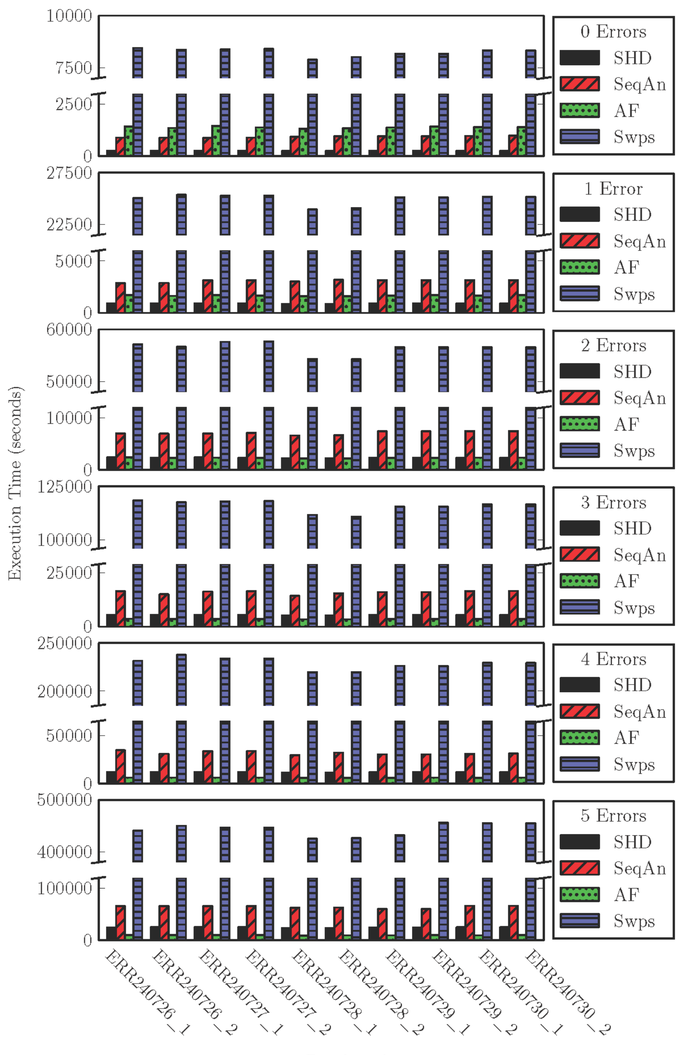
4 SRS counts errors conservatively to preserve correctness

5 Results and Conclusion:

- **SHM** and **SRS** are implemented using bit-parallel and SIMD operations (with **Intel SSE,** details in upcoming paper)
- The threshold of SRS is **platform dependent** (3 bps at maximum on Intel platforms)
- We compare SHD against:
  - **SeqAn:** Gene Myers' bit-vector algorithm
  - **Swps:** A SIMD implementation of Smith-Waterman algorithm
  - **AF:** A k-mer locality based filter, FastHASH
- We used **mrFast** to retrieve all potential mappings (read-reference pairs) from **ten real read sets** from 1000 Genomes Project
- The false positive rate of SHD **increases with larger error thresholds**. SHD is effective with up to 5% error rate
- **Key Conclusion:** SHD is **3x faster** than the best previous implementation of edit-distance calculation, while having a false positive rate of only **7%** (e = 5)



Performance results:



False positive results: