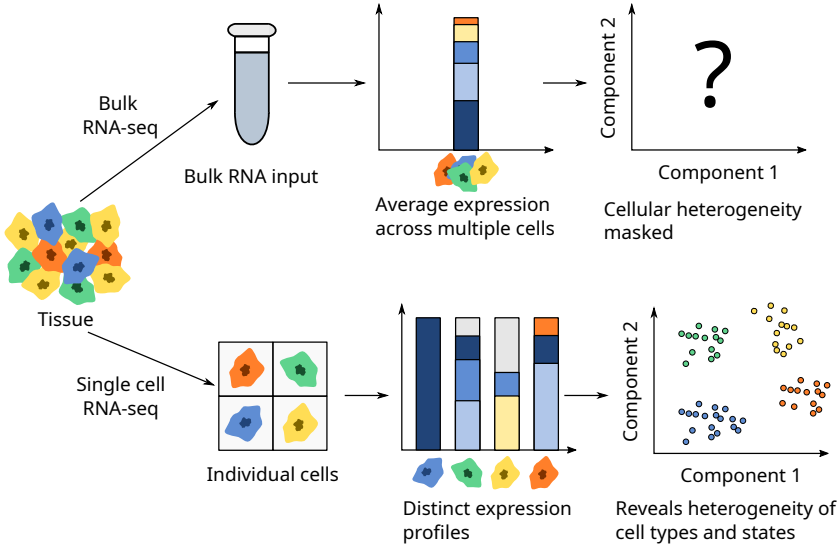
UNIVERSITÄT DES SAARLANDES

Error Correction Algorithms for Efficient Gene Expression Quantification in Single Cell Transcriptomics

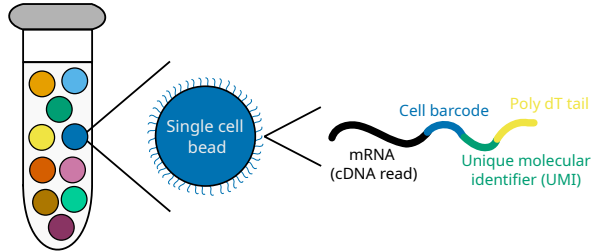
Jens Zentgraf, Johanna E. Schmitz, Andreas Keller, Sven Rahmann

Motivation: RNAseq

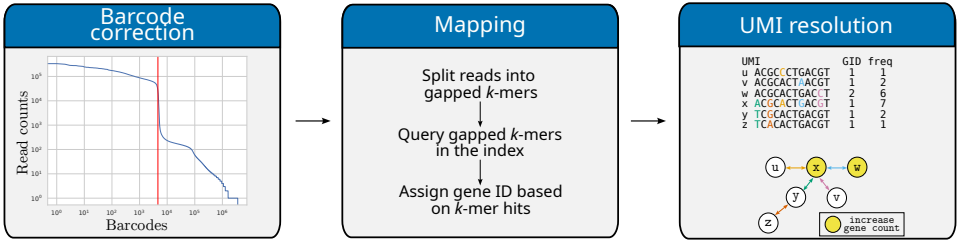


Motivation

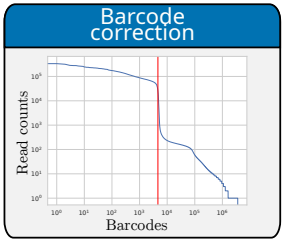
- Massive increase in data
- Possible to sequence thousands of cells in parallel
- 1st step:
  Create cell x gene count matrix
- Faster and scalable methods needed



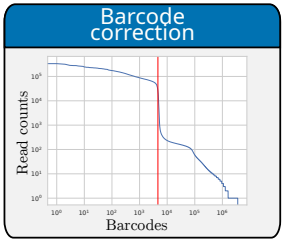
Problem


Barcode correction
Read counts
Barcodes
Mapping
Split reads into gapped k-mers
Query gapped k-mers in the index
Assign gene ID based on k-mer hits
UMI resolution
UMI GID freq
u ACGCCTGACGT 1 1
v ACGCACTAACGT 1 2
w ACGCACTGACCT 2 6
x ACGCACTCACGT 1 7
y TCGCACTGACGT 1 2
z TCACACTGACGT 1 1
u
x
w
y
v
z
increase gene count

Problem



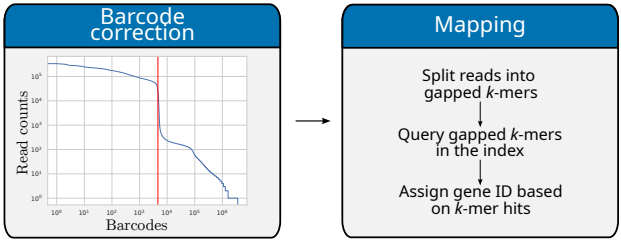
Problem



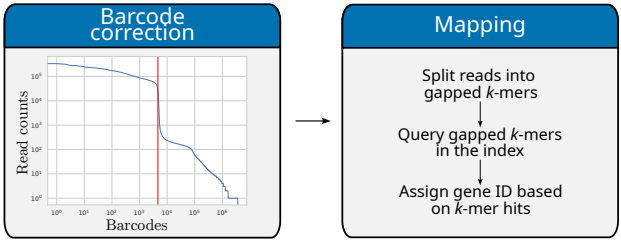
Barcode correction

- Each cell has a unique barcode from an inclusion list
- Some droplets contain no cell or a corrupted cell
- Sequencing errors lead to wrong barcodes
- Goal:
  - Discard empty droplets
  - Rescue barcodes with sequencing errors

Problem


Barcode correction
Read counts
Barcodes
Mapping
Split reads into gapped k-mers
Query gapped k-mers in the index
Assign gene ID based on k-mer hits

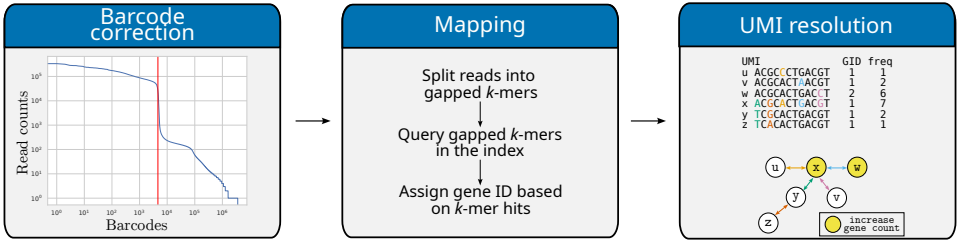
Problem



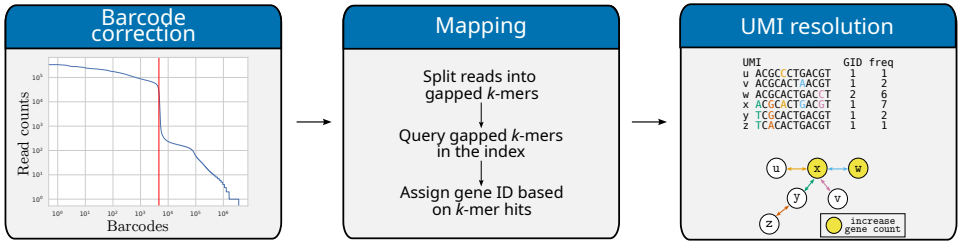
Mapping

- Map reads to the correct gene
- Similar or overlapping genes
- Avoid alignment (slow)
- Alignment-free approach

Problem


Barcode correction
Read counts
Barcodes
Mapping
Split reads into gapped k-mers
Query gapped k-mers in the index
Assign gene ID based on k-mer hits
UMI resolution
UMI GID freq
u ACGCCTGACGT 1 1
v ACGCACTAACGT 1 2
w ACGCACTGACCT 2 6
x ACGCACTCACGT 1 7
y TCGCACTGACGT 1 2
z TCACACTGACGT 1 1
u
x
w
y
v
z
increase gene count

Problem



UMI resolution

- Each UMI corresponds in principle to a unique molecule
- Same UMIs should be counted only once
- Sequencing errors can inflate the number of distinct UMIs
- UMI collision can cause the same UMI to map to different genes
- Differentiate between true UMI collisions and mapping errors

Background

Definitions

- *k*-mers
  - Substrings of length *k*

```
ACTGAAGCTAGCT
ACTGA
 CTGAA
  TGAAG
   GAAGC
    AAGCT
     AGCTA
      GCTAG
       CTAGC
        TAGCT
```

Zentgraf, J., Rahmann, S.: Swiftly identifying strongly unique *k*-mers. Algorithms for Molecular Biology 20(1), 13 (2025). https://doi.org/10.1186/s13015-025-00286-6

Background

Definitions

- k-mers
  - Substrings of length k
- Gapped k-mers (spaced seeds)
  - k significant positions (#)
  - Window size w
  - $w - k$ insignificant positions (_)
  - ##_#_##

```
ACTGAAGCTAGCT
AC  G  AG
 CT  A  GC
  TG  A  CT
   GA  G  TA
    AA  C  AG
     AG  T  GC
      GC  A  CT
```

Zentgraf, J., Rahmann, S.: Swiftly identifying strongly unique k-mers. Algorithms for Molecular Biology 20(1), 13 (2025). https://doi.org/10.1186/s13015-025-00286-6

Background

Definitions

- k-mers
  - Substrings of length k
- Gapped k-mers (spaced seeds)
  - k significant positions (#)
  - Window size w
  - $w - k$ insignificant positions (_)
  - ##_#_##
- A k-mer x is unique,
  if it occurs in exactly one gene.

```
ACTGAAGCTAGCT
AC  G  AG
 CT  A  GC
  TG  A  CT
   GA  G  TA
    AA  C  AG
     AG  T  GC
      GC  A  CT
```

Zentgraf, J., Rahmann, S.: Swiftly identifying strongly unique k-mers. Algorithms for Molecular Biology 20(1), 13 (2025). https://doi.org/10.1186/s13015-025-00286-6

Background

Definitions

- k-mers
  - Substrings of length k
- Gapped k-mers (spaced seeds)
  - k significant positions (#)
  - Window size w
  - $w - k$ insignificant positions (_)
  - ##_#_##
- A k-mer x is unique, if it occurs in exactly one gene.
- A k-mer $x \in K$ is weak in K iff a k-mer $y \in K$ with a Hamming distance of $H(x, y) = 1$ exists in K. The other k-mers are strong in K.

```
TTACTGATTAGTT
ACTGAAGCTAGCT
GTAGTACGTAGGA
ACTGAGGCTAGCT
CCTACACCTTGCA
```

Zentgraf, J., Rahmann, S.: Swiftly identifying strongly unique k-mers. Algorithms for Molecular Biology 20(1), 13 (2025). https://doi.org/10.1186/s13015-025-00286-6

Background

Definitions

- k-mers
  - Substrings of length k
- Gapped k-mers (spaced seeds)
  - k significant positions (#)
  - Window size w
  - $w - k$ insignificant positions (_)
  - ##_#_##
- A k-mer x is unique, if it occurs in exactly one gene.
- A k-mer $x \in K$ is weak in K iff a k-mer $y \in K$ with a Hamming distance of $H(x, y) = 1$ exists in K. The other k-mers are strong in K.

```
TTACTGATTAGTT
ACTGAAGCTAGCT
GTAGTACGTAGGA
ACTGAGGCTAGCT
CCTACACCTTGCA
```

- Strongly unique k-mers are robust identifiers.

Zentgraf, J., Rahmann, S.: Swiftly identifying strongly unique k-mers. Algorithms for Molecular Biology 20(1), 13 (2025). https://doi.org/10.1186/s13015-025-00286-6

State of the art

CellRanger

- Commercial solution from 10X Genomics
- Splicing-aware alignment to the reference genome
- Uses STAR aligner
- Slow

State of the art

CellRanger

- Commercial solution from 10X Genomics
- Splicing-aware alignment to the reference genome
- Uses STAR aligner
- Slow

Kallisto|Bustools[1]

- Alignment-free
- Pseudoalignment using a color-compacted de Bruijn graph

[1] Melsted, P., Booeshaghi, A.S., Liu, L., Gao, F., Lu, L., Min, K.H.J., da Veiga Beltrame, E., Hjörleifsson, K.E., Gehring, J., Pachter, L.: Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nature Biotechnology 39(7), 813–818 (2021). https://doi.org/10.1038/s41587-021-00870-2

State of the art

CellRanger

- Commercial solution from 10X Genomics
- Splicing-aware alignment to the reference genome
- Uses STAR aligner
- Slow

Kallisto|Bustools[1]

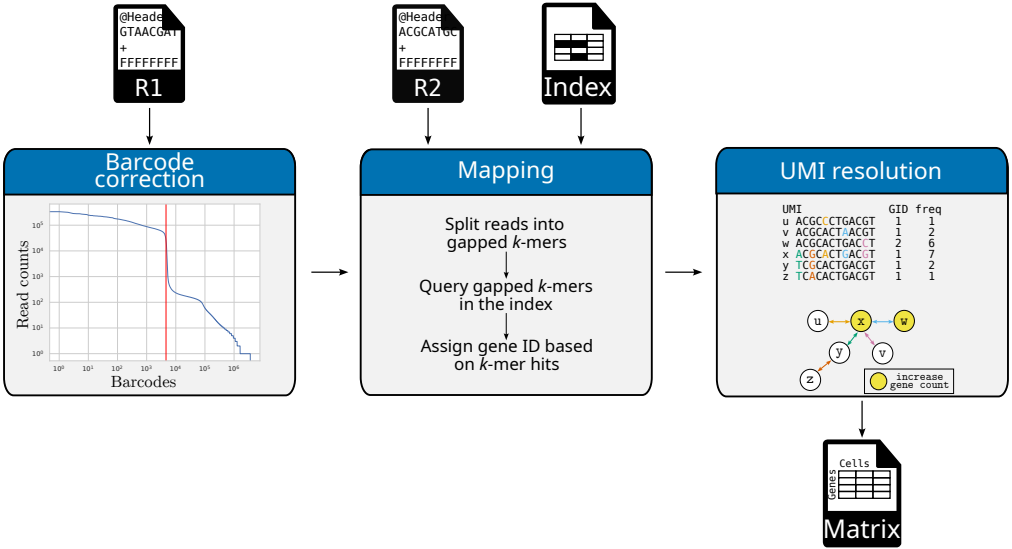
- Alignment-free
- Pseudoalignment using a color-compacted de Bruijn graph

Alevin-Fry[2]

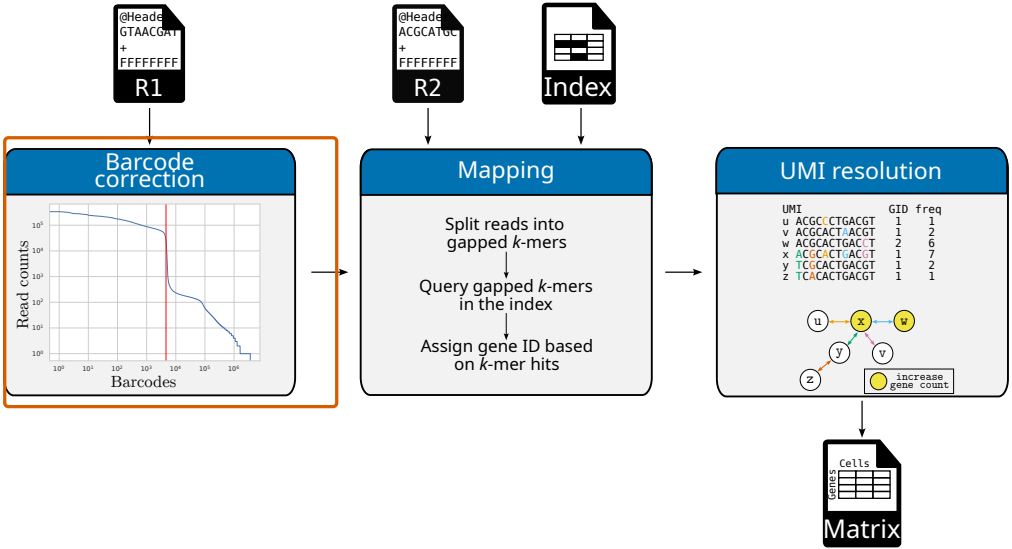
- Alignment-free
- Pseudoalignment using a color-compacted de Bruijn graph
- CR-like or custom barcode correction and UMI resolution

[1] Melsted, P., Booeshaghi, A.S., Liu, L., Gao, F., Lu, L., Min, K.H.J., da Veiga Beltrame, E., Hjörleifsson, K.E., Gehring, J., Pachter, L.: Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nature Biotechnology 39(7), 813–818 (2021). https://doi.org/10.1038/s41587-021-00870-2
[2] He, D., Zakeri, M., Sarkar, H., Soneson, C., Srivastava, A., Patro, R.: Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data. Nature Methods 19(3), 316–322 (2022). https://doi.org/10.1038/s41592-022-01408-3

Overview of arcane



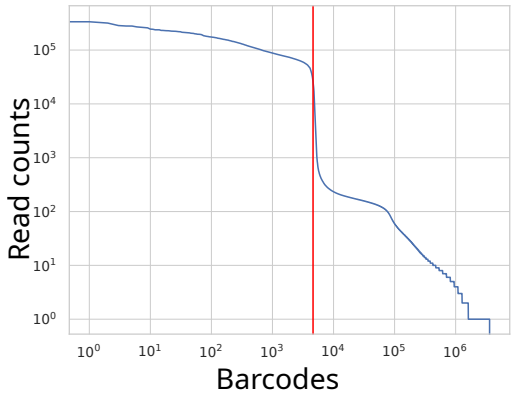
Overview of arcane



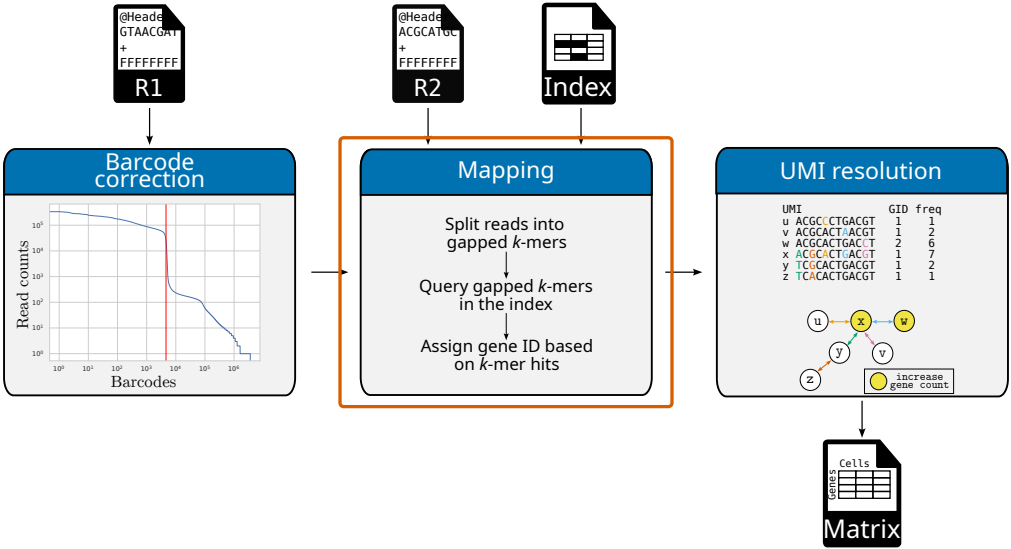
Barcode correction

Solution

- Correct invalid barcodes that have a Hamming distance 1 neighbor on the inclusion list
- Discard all barcodes with a count below the *Knee* (red line)



Overview of arcane

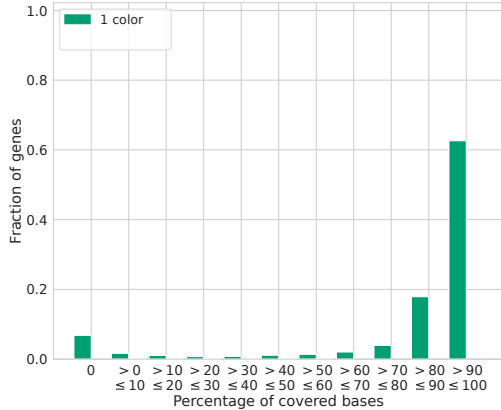


Arcane index

- Other tools use color-compacted de Bruijn graphs
- Do we need all colors?
  - More colors, more contradictory data

Arcane index

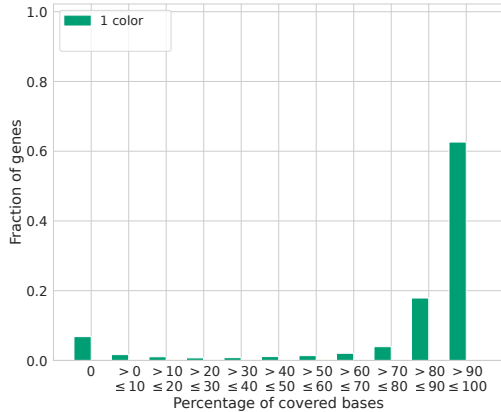
- Other tools use color-compacted de Bruijn graphs
- Do we need all colors?
  - More colors, more contradictory data



Arcane index

Observation

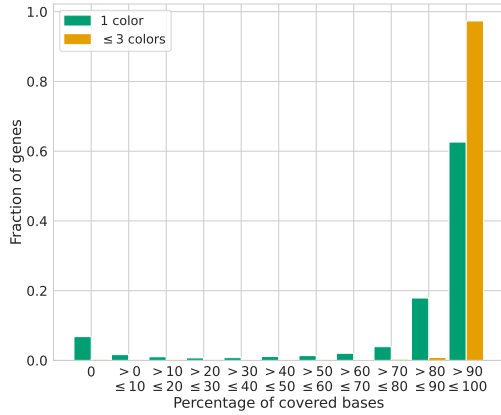
- Already high coverage using one color



Arcane index

Observation

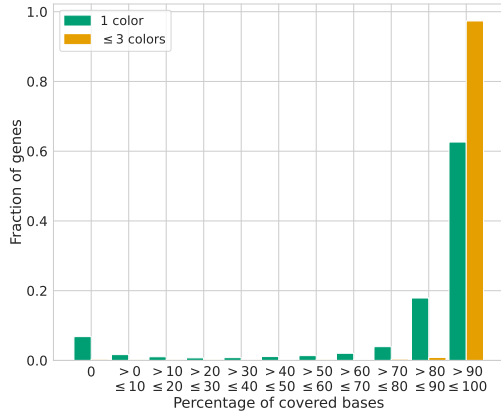
- Already high coverage using one color



Arcane index

Observation

- Already high coverage using one color
- Most genes are covered using up to 3 colors



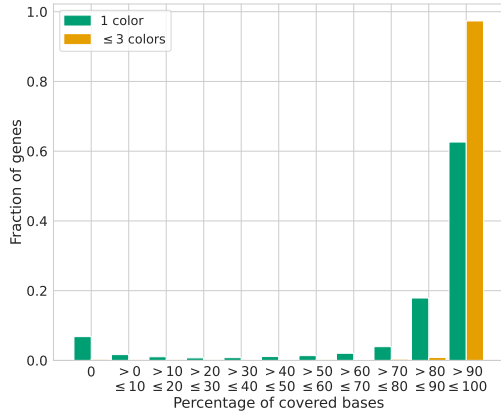
Arcane index

Observation

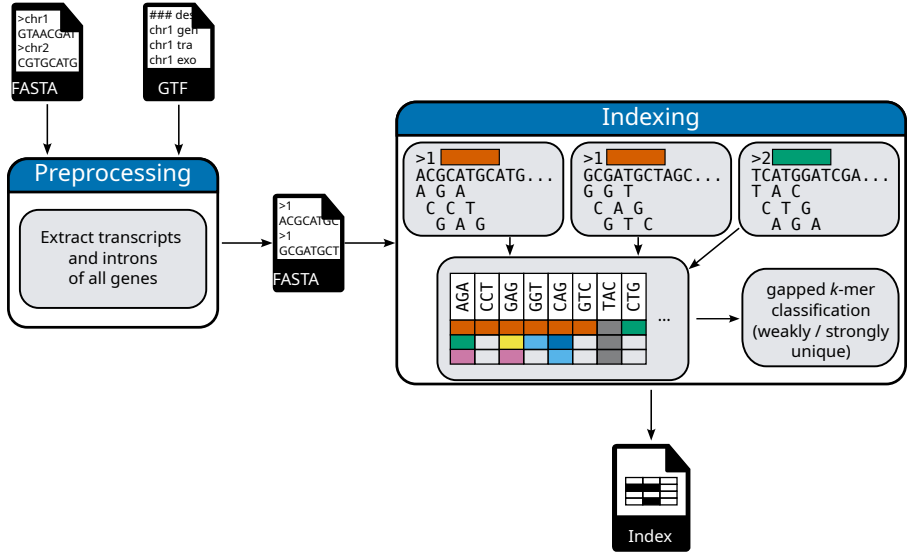
- Already high coverage using one color
- Most genes are covered using up to 3 colors

Idea

- Simplify colored de Bruijn graph by storing a limited number of colors per k-mer
- Depending on index: fewer cache misses per query



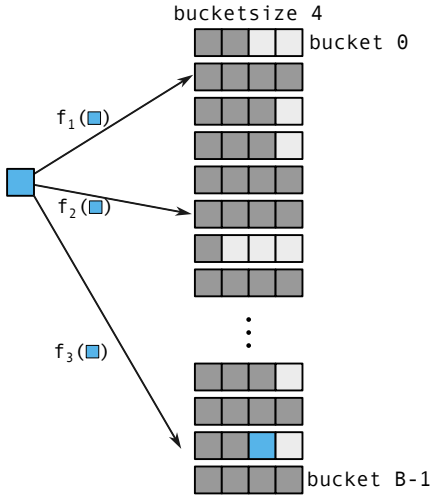
Arcane index



Arcane index

(3, 4)-Cuckoo Hashing

- We use $h = 3$ bucket hash functions $f_i : \mathcal{U} \to [B] := \{0, 1, \ldots, B-1\}$
- Each bucket has 4 slots
- To query a key, check if it is stored in one of the $h \cdot l$ slots
- To insert a key, check if one of $h \cdot l$ slots is empty, if not, perform a random walk
- Maximum achievable load for $(3, 4)$-Cuckoo hashing is ≈ 0.99

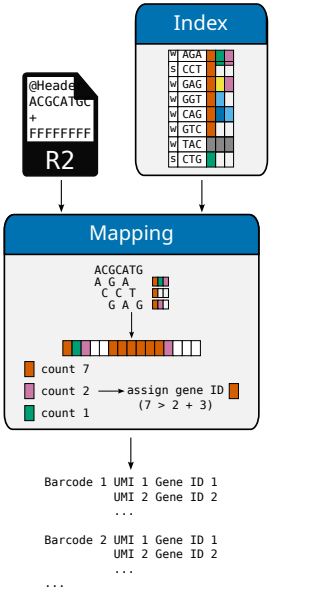


Walzer. Load Thresholds for Cuckoo Hashing with Overlapping Blocks. ACM Trans. Algorithms, 19(3):24:1–24:22, 2023. doi: 10.1145/3589558.

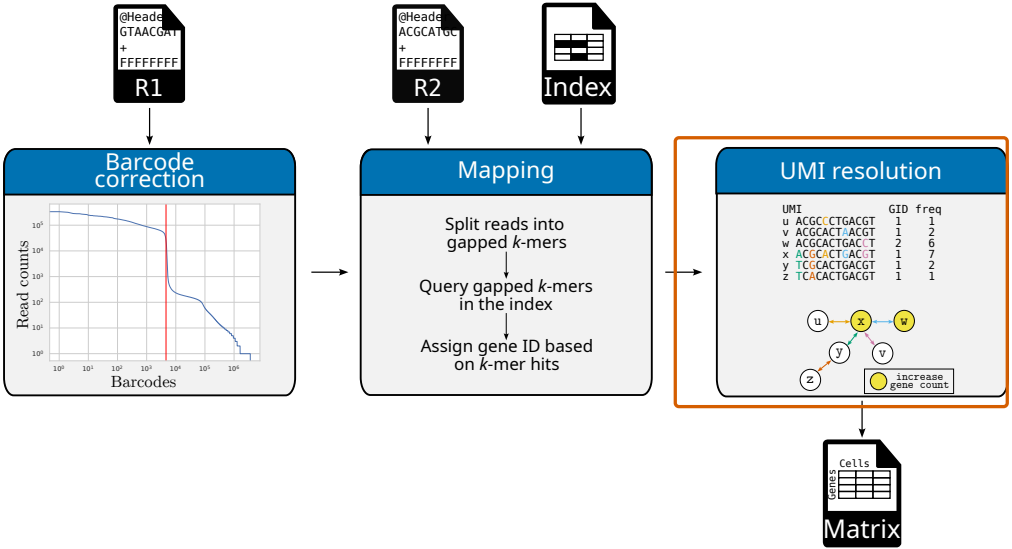
J. Zentgraf and S. Rahmann. Fast Gapped k-mer Counting with Subdivided Multi-Way Bucketed Cuckoo Hash Tables. WABI 2022, volume 242, pages 12:1–12:20, 2022. doi: 10.4230/LIPIcs.WABI.2022.12

Gene mapping

- For each read, query every gapped k-mer
- Count for each gapped k-mer its colors
- Gene weighting for more robust mapping:
  - Strongly unique k-mer: 5
  - Weakly unique k-mer: 3
  - Non-unique k-mer: 1
- Assign gene with highest count if its count is at least 3 larger than the next highest count



Overview of arcane

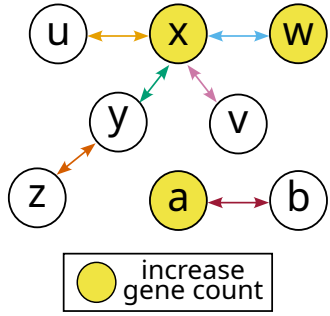


UMI resolution / de-duplication

Solution

For all UMIs with the same barcode

- Build a graph for all UMIs that are connected to their Hamming distance 1 neighbors
- Compute connected components in the graph
- For a connected component, count only UMI-gene combinations with a count $\geq$ the expected count



UMI		GID	freq
u	ACGCCTGACGT	1	1
v	ACGCACTAACGT	1	2
w	ACGCACTGACCT	2	6
x	ACGCACTGACGT	1	7
y	TCGCACTGACGT	1	2
z	TCACACTGACGT	1	1
a	ACTGAAACTGAG	7	5
b	ACTGTAACTGAG	7	2

Results

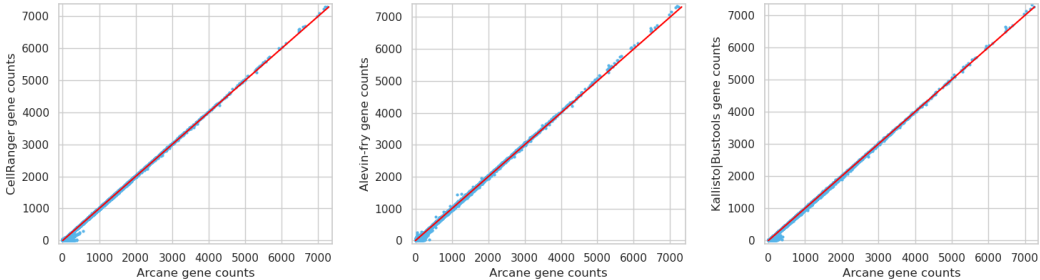
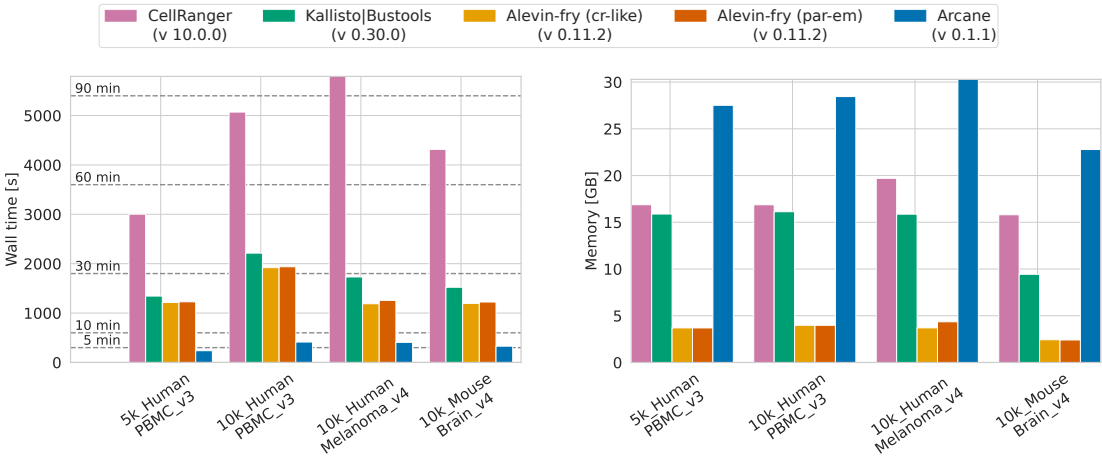


Figure: Comparison on the mouse brain dataset (10 000 cells, v4 chemistry).

Arcane index uses (31, 43) mask: ####_#_##_###_#_###_###_###_#_###_##_#_####

Results



All benchmarks were run on an Ubuntu (24.04 LTS) server with two AMD EPYC 9534 64-Core Processors, 1.5TB of 4800-MHz DDR5 memory and a KIOXIA CD8P SSD. Running time and maximum memory usage were measured using /usr/bin/time -v (using 16 threads).

Summary

Arcane

- Fast
- Memory footprint < 32 GB
- Supports shared memory for the index
- Yields similar results as Alevin-fry, Kallisto|Bustools and CellRanger

Future and ongoing work

- More pipelining (improved in new version)
- Try out new index data structures to save memory