Minimizers and MinHashing

Adam M. Phillippy

CMSC701: April 23, 2019

@aphillippy

National Human Genome Research Institute

The Forefront of Genomics®

The minimizer: a simple idea

- The lexicographically smallest *k*-mer
  - Motivation: genome assembly

```
ATGATCGTGATGTCGTAGTATCGTGCAA
ATGAT GTGAT TCGTA TATCG
 TGATC TGATG CGTAG ATCGT
  GATCG GATGT GTAGT TCGTG
   ATCGT ATGTC TAGTA CGTGC
    TCGTG TGTCG AGTAT GTGCA
     CGTGA GTCGT GTATC TGCAA
```

NIH NHGRI

Reducing storage requirements for biological sequence comparison. Roberts ... Yorke (2004)

(*w*,*k*)-minimzers

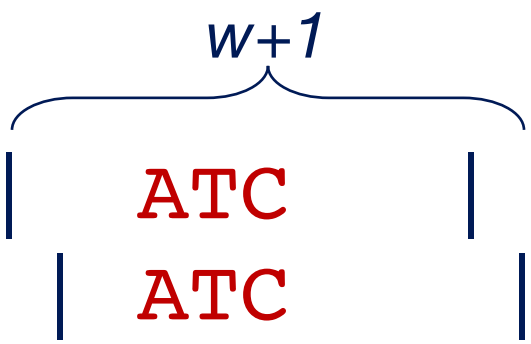
- Window size *w*, k-mer size *k*
  - What are the (10, 3)-minimizers?

```
ATGATCGTGATGTCGTAGTATCGTGCAA
|   ATC     |
 |          |
  |          |
   |          |
    |          |
     |   ATG  |
      |        |
       |     AGT|
        |  ATC    |
```

NIH
NHGRI

Winnowing

- Non-lexicographic permutation π
  - A hash function h: K $\rightarrow$ {0, 1, 2, 3, …, 2^b-1}
  - e.g. Rabin fingerprints
  - Motivating: duplicate document detection
- What's the expected density of random minimizers?
  - $d = 2 / (w + 1)$



Winnowing: Local Algorithms for Document Fingerprinting. Schleimer, Wilkerson, Aiken (2003)

Expected minimizer density

Ordering	Density factor
DOCKS	1.737
Random	1.999
Lexicographic	2.236

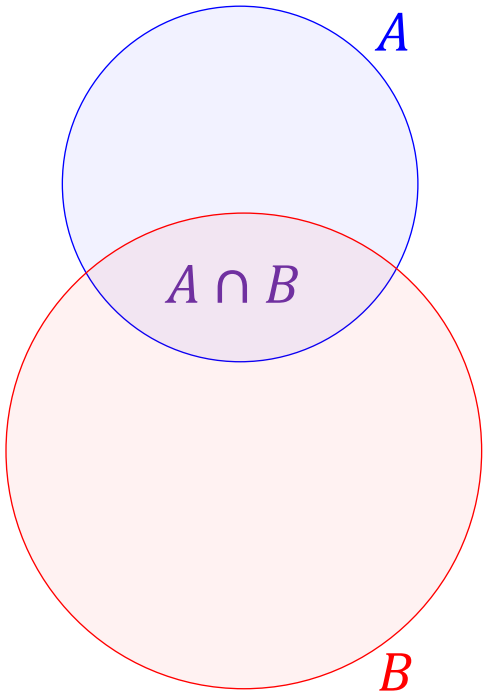
For $k = 10$, $w = 10$ on a binary alphabet

- Universal hitting set (DOCKS)
  - Given integers k and L, find a smallest set U_{kL} of k-mers such that any string of length L or longer must contain at least one k-mer from U_{kL}

Improving the performance of minimizers and winnowing schemes. Marçais *et al.* (2017)

MinHash and the WWW

- The smallest k-mer in *t*
  - *s=1 : a minimizer*
  - Match two similar strings



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- The smallest *s* k-mers in *t*
  - *s>1 : a sketch*
  - Estimate the Jaccard index
  - Motivation: duplicate document detection

NIH NHGRI

On the resemblance and containment of documents. Broder (1997)

Mini probability review

- Sample space (Ω) is set of all possible outcomes
  - Ω = { all possible rolls of 2 dice }

- An event (A, B, ...) is a subset of Ω
  - A = { rolls where first die is odd }
  - B = { rolls where second die is even }

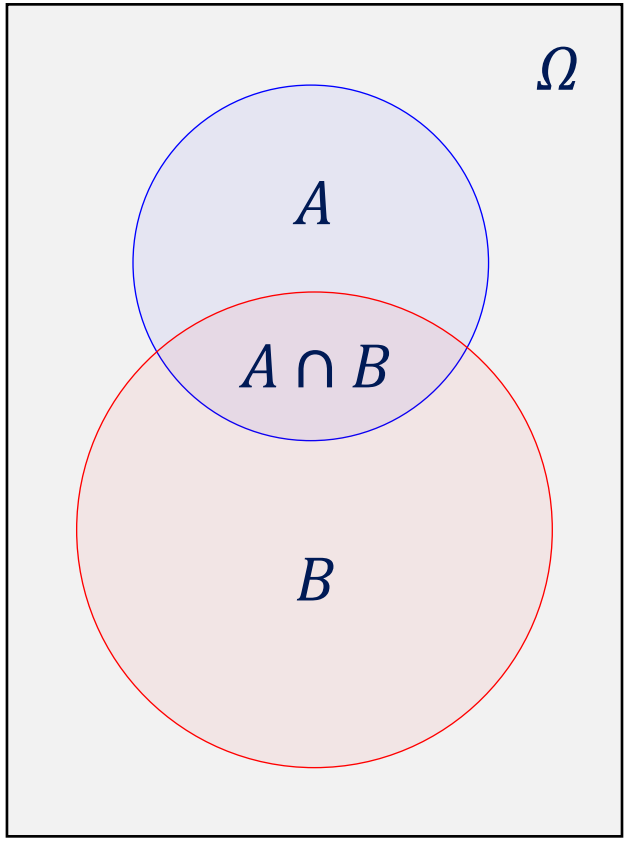
- P(A): fraction of all possible outcomes that are in A
  - $P(A) = |A| / |\Omega| = 18 / 36 = 0.5$

Mini probability review

- If A and B are independent:

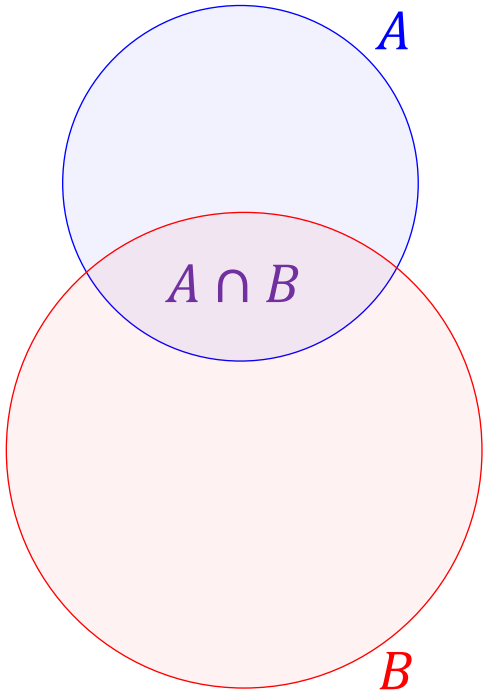
- $P(A \cap B)$: fraction of all possible outcomes in both A *and* B
  - $P(A \cap B) = |A \cap B| / |\Omega| = 9 / 36 = 0.25$
  - $P(A \cap B) = P(A) * P(B)$

- $P(A \cup B)$: fraction of all possible outcomes in either A *or* B
  - $P(A \cup B) = |A \cup B| / |\Omega| = 27 / 36 = 0.75$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



MinHash, the WWW, and assembly

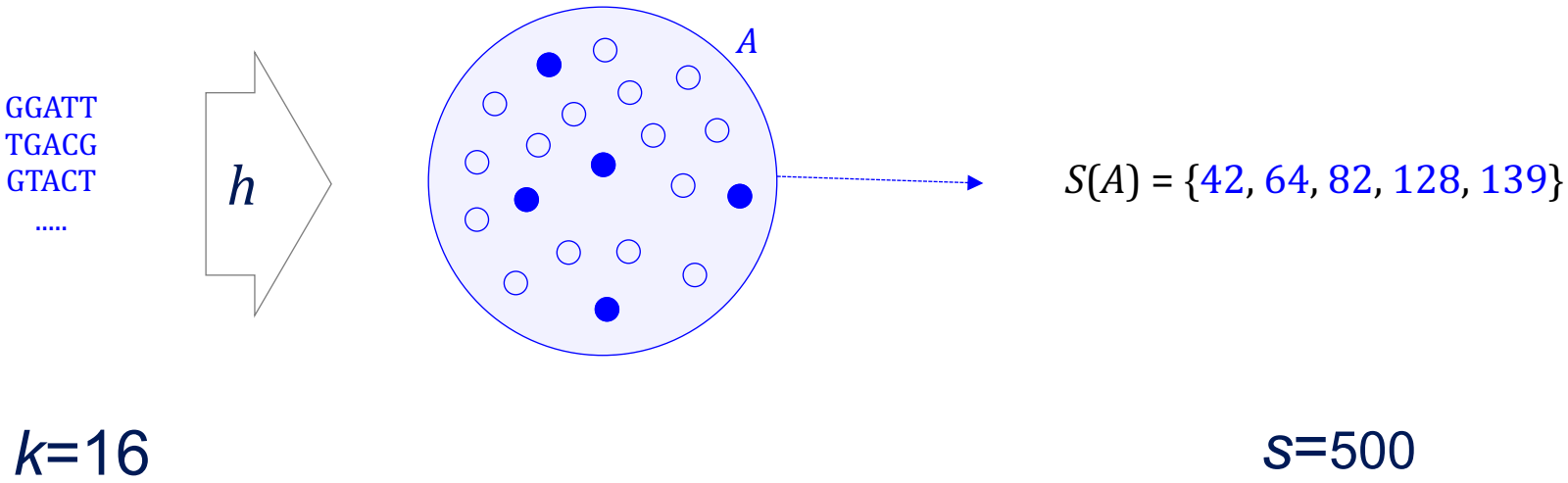
- Define a random permutation and take the *s* smallest *k*-mers
  - *s=1 : a minimizer*
    - $P[h_{min}(A) = h_{min}(B)] = J(A, B)$
  - *s>1 : a sketch*
    - $P[h1_{min}(A) = h1_{min}(B)] = J(A, B)$
    - $P[h2_{min}(A) = h2_{min}(B)] = J(A, B)$
    - …
- Better estimate with larger *s*



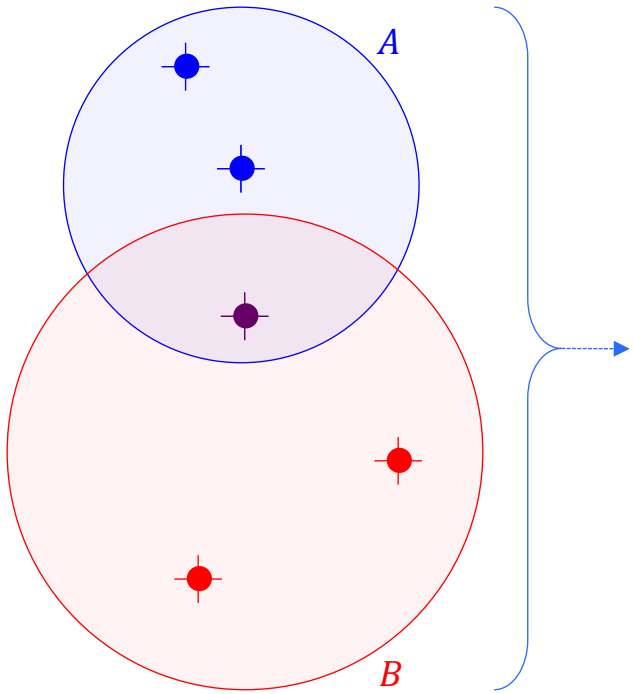
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

On the resemblance and containment of documents. Broder (1997)

MinHash “bottom sketch”



Testing resemblance with sketches

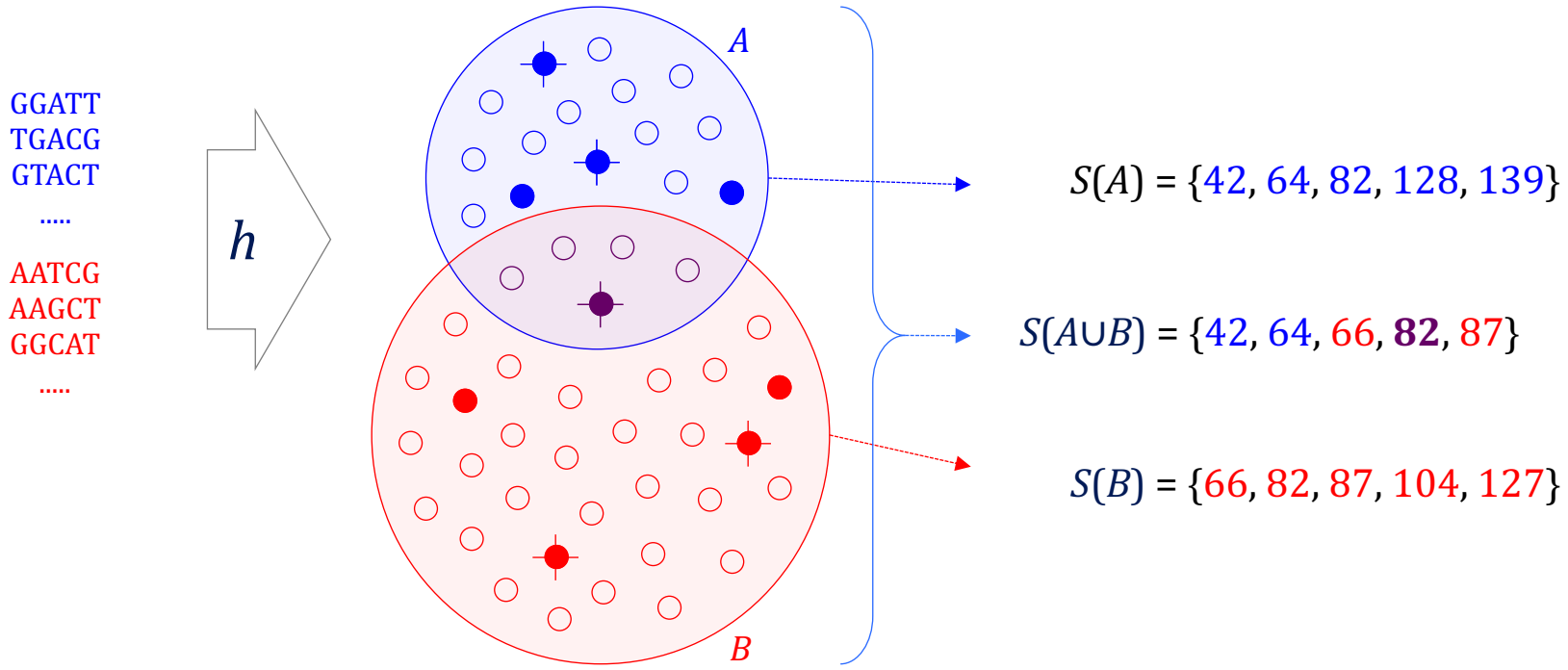


$S(A \cup B) = \{42, 64, 66, \mathbf{82}, 87\}$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

NIH
NHGRI

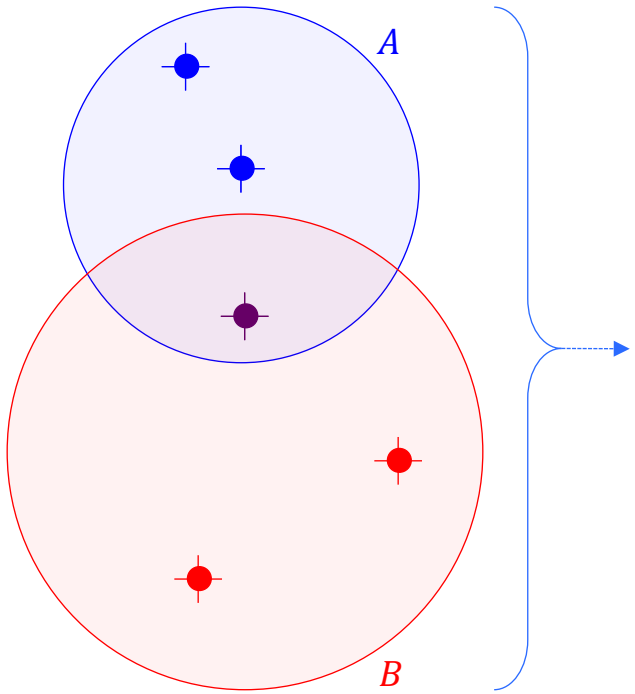
Testing resemblance with sketches



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

NIH
NHGRI

Testing resemblance with sketches

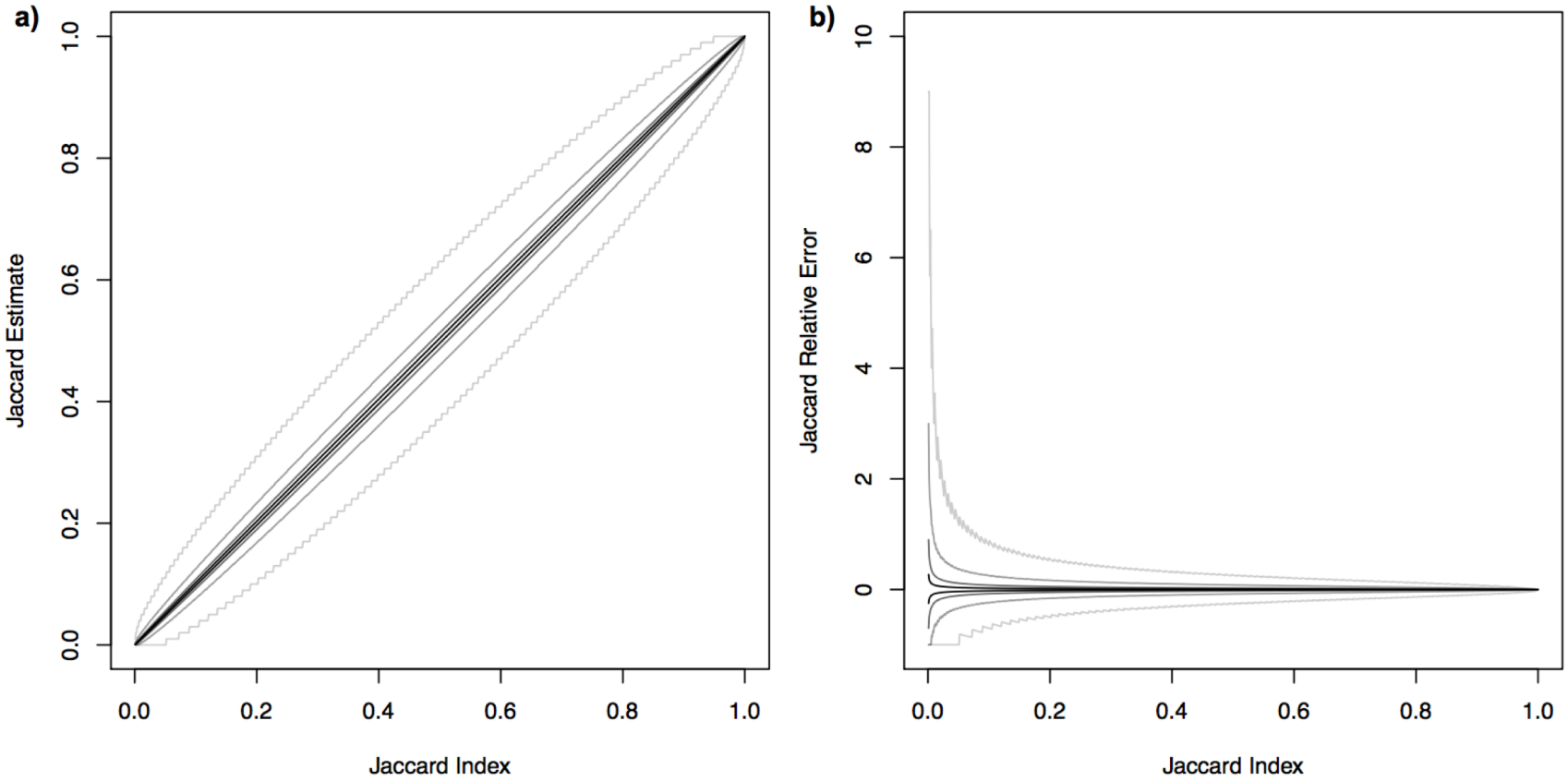


$S(A \cup B) = \{42, 64, 66, \mathbf{82}, 87\}$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

NIH
NHGRI

MinHash error bounds O(1/√s)



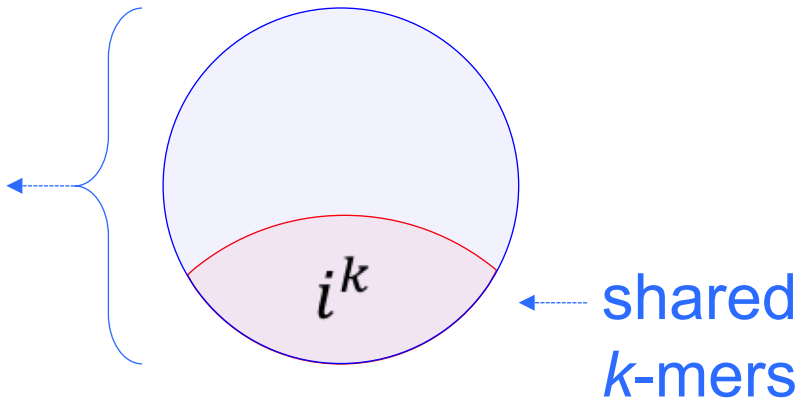
Supplementary Figure 1. Absolute and relative error bounds for Mash Jaccard estimates given various sketch sizes. Increasing sketch sizes are progressively shaded from *s*=100 (light gray), *s*=1,000, *s*=10,000, and *s*=100,000 (black). Upper and lower bounds are drawn using the binomial inverse cumulative distribution function, with the same parameters from equation 8, such that for a given Jaccard index there is a 0.99 probability that the corresponding Jaccard estimate **(a)** or relative error **(b)** will fall within the bounds. These plots illustrate that relative error can grow quite large when estimating small

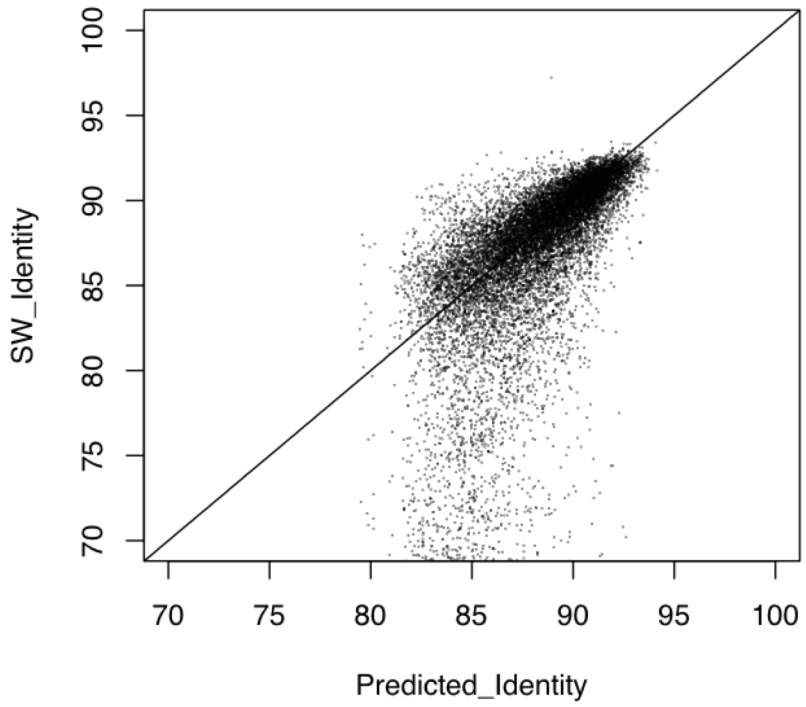
NIH
NHGRI

Identity estimation

- Jaccard relies on *k* and is not linear w.r.t. identity
- Expected *k*-mer survival rate $(1\text{-}e)^k$

$$i = \left(\frac{2j}{1+j}\right)^{1/k}$$



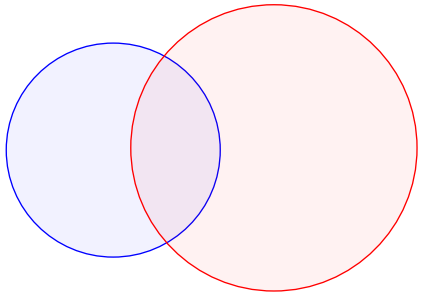


Mash: fast genome and metagenome distance estimation using MinHash.
Ondov et al. *Genome Biology* (2016)

NIH NHGRI

Mash P-value

- Probability of seeing *x* or more matches given two random genomes
  - Hypergeometric approx. by binomial distribution
  - *r* = expected Jaccard of two random genomes



$$r = \frac{P(K \in X)\,P(K \in Y)}{P(K \in X) + P(K \in Y) - P(K \in X)\,P(K \in Y)}$$

$$P(K \in X) = 1 - \left(1 - |\Sigma|^{-k}\right)^n$$

$$p(x; s; r) = 1 - \sum_{i=0}^{x-1} \binom{s}{i} r^i (1-r)^{s-i}$$

Mash: fast genome and metagenome distance estimation using MinHash.
Ondov et al. *Genome Biology* (2016)

NIH NHGRI

Advantages of MinHash?

- **Comparisons are instantaneous**
  - Primary overhead is sketching
  - O(n) sketch, $O(n^2)$ comparisons
- **Sketches are very small**
  - 3 Gbp primate genome
    - 8 kB vs. 750 MB
  - 10 Tbp of samples
    - 71 MB vs. 2.5 TB

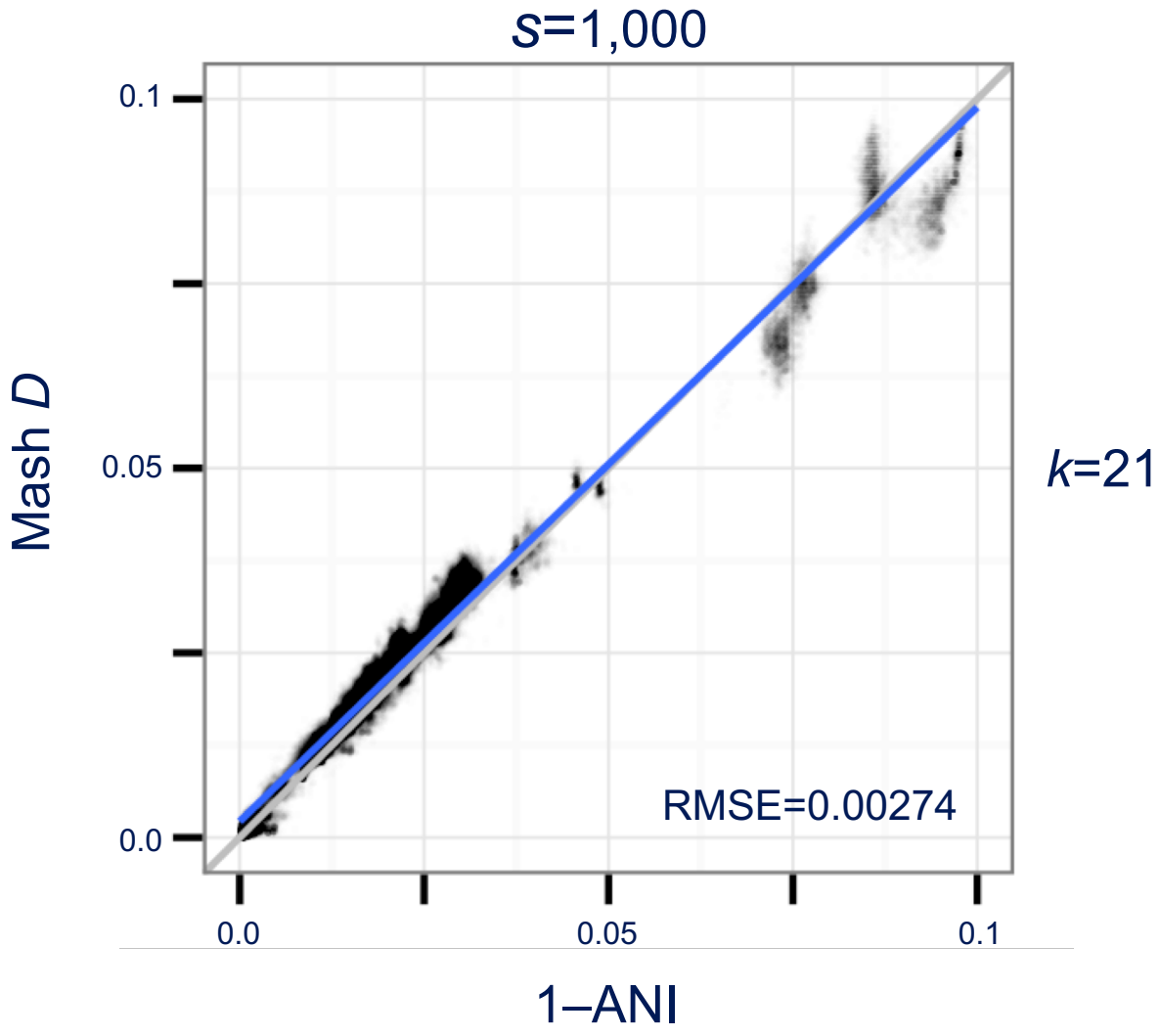
NIH
NHGRI

Example applications in bioinformatics

- **Minimizers**
  - UMD overlapper
  - KMC k-mer counting
  - Mashmap and Minimap long-read alignment
- **MinHash**
  - MHAP overlapper
  - Mash genomic distance

Mash distance correlates with ANI

- **Average Nucleotide Identity**
  - Avg. alignment identity of orthologous genes
  - Defined by a “core genome alignment”, i.e. reciprocally best
  - Scaling up to more genomes
    - BLAST 2005
    - MUMmer 2009
    - Subset of genes 2013
    - Mash 2016
    - FastANI 2018



All-pairs comparison of 500 Escherichia genomes

NIH NHGRI

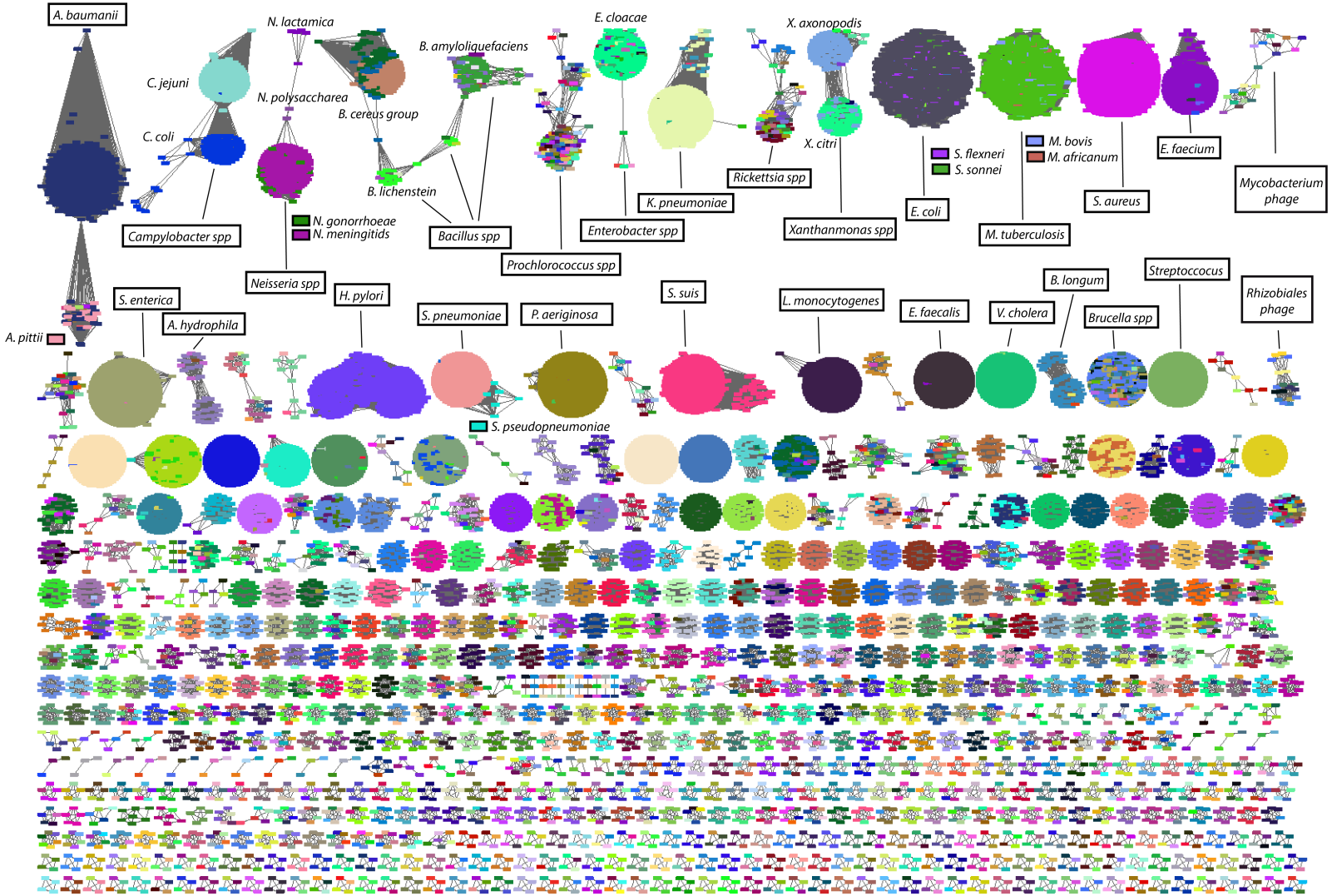
MHAP assembly performance

- **629,000 CPU hours** (December 2014)
  - BLASR overlapping + CA PBcR
  - 98% of runtime in overlapping
- **1,086 CPU hours** (April 2014)
  - MHAP overlapping + CA PBcR
  - k=16, sketch=512, minmatch=3
  - Better assembly

NIH NHGRI

Assembling large genomes with single-molecule sequencing and locality-sensitive hashing.
Berlin *et al.* (2015)

Clustering all of NCBI RefSeq



Sketch DB <100 MB, ~1.5 billion distances computed in 46 CPU h, linear genome search <1 CPU s

NIH NHGRI