Session 8: Intro to Sequence Similarity - BLAST, MASH, ANI, and Other Applications in Public Health Bioinformatics

Joel R. Sevinsky, Ph.D.
AMD Academy June 2020

Overview

- BLAST - Overview and brief description of algorithm
- BLAST exercise using NCBI interface
- Microbial identification - BLAST in ANI
- Microbial identification - beyond BLAST → MASH
- Other areas of sequence similarity searching

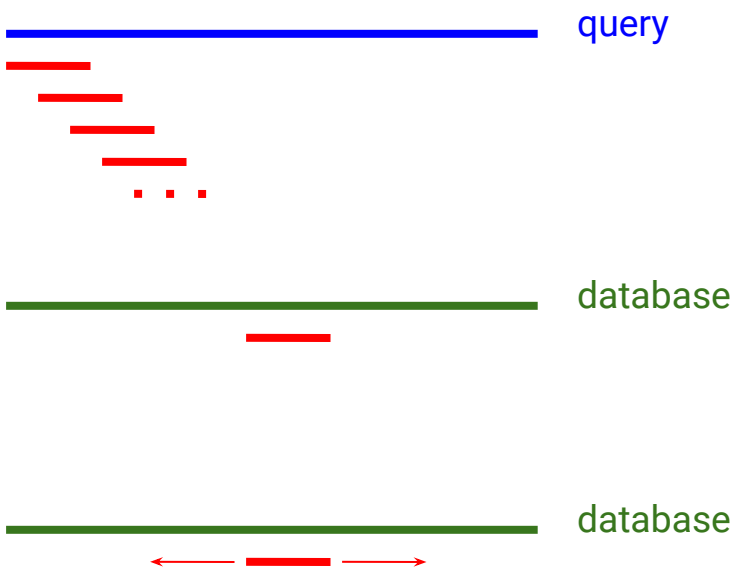
BLAST: Basic Local Alignment Search Tool

- Suite of tools for comparing a query sequence to a database of sequences
- You can compare DNA:DNA, DNA:protein, protein:DNA, and protein:protein
- “Seed and extend” heuristic (indexing based local alignment)
  - Seed
    - make starting words (k-mers) in assembly from your query sequence
    - compare query words to an indexed word list from your database
    - identify exact matches
  - Extend
    - once exact matches are found, try and extend the alignments with a scoring algorithm
    - different algorithms available for DNA vs protein
    - keep score as you extend beyond exact match, and maximize the score

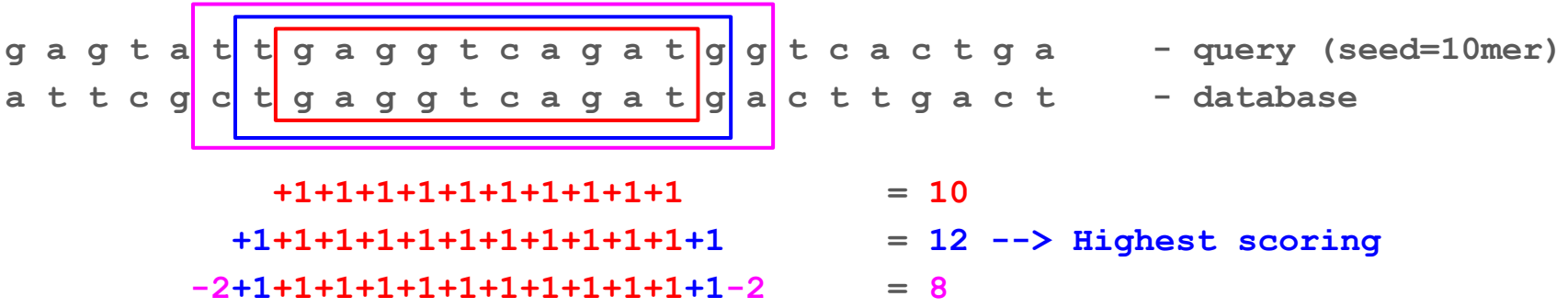
BLAST: Basic Local Alignment Search Tool

Seed and extend process

- Create query dictionary
- Align/search database dictionary
- Extend query along target



BLAST: Basic Local Alignment Search Tool



- Initial, or seed match, is extended on each side until the aggregate alignment score falls below a predetermined threshold.

BLAST: Basic Local Alignment Search Tool

https://blast.ncbi.nlm.nih.gov/Blast.cgi

What Tredegar steps use sequence similarity?

```
(base) amd-academy@blast:~/data$ staphb-wf tredegar -o blast illumina/
/home/amd-academy/anaconda3/lib/python3.7/site-packages/staphb_toolkit/lib
3.7/site-packages/staphb_toolkit/workflows/tredegar//tredegar.nf -profile
blast/logs/20_06_14_17_40_41_Tredegar_trace.txt -with-report blast/logs/20
st/logs/work
Starting the Tredegar pipeline:
N E X T F L O W  ~  version 20.01.0
Launching `/home/amd-academy/anaconda3/lib/python3.7/site-packages/staphb_
 - revision: 679f403fdf
executor >  local (8)
[ea/c591c8] process > preProcess     [100%] 6 of 6 ✔
[dc/dec1a6] process > mash_dist      [ 17%] 1 of 6
[-          ] process > mash_species  -
[88/cacb86] process > trim           [  0%] 0 of 6
[-          ] process > cleanreads    -
[-          ] process > shovill       -
[-          ] process > quast         -
[-          ] process > cg_pipeline   -
[-          ] process > emmtype_finder -
[-          ] process > seqsero       -
[-          ] process > serotypefinder -
[-          ] process > results       -
```

1. mash
2. shovill
3. cg_pipeline
4. emmtype_finder
5. seqsero
6. serotypefinder

What Tredegar steps use sequence similarity?

- Assembly
  - genome assembly, reference based mapping
- Identity
  - species identification
- Annotation
  - gene prediction, serotype, AMR
- Genetic Characterization
  - phylogenetic trees, SNPs

What Tredegar steps use sequence similarity?

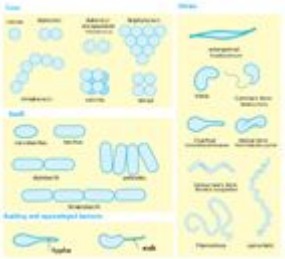
- Assembly
  - genome assembly, reference based mapping
- Identity
  - species identification
- Annotation
  - gene prediction, serotype, AMR
- Genetic Characterization
  - phylogenetic trees, SNPs

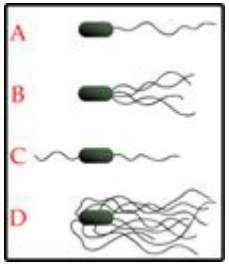
Microbial Identification Using Sequence Similarity

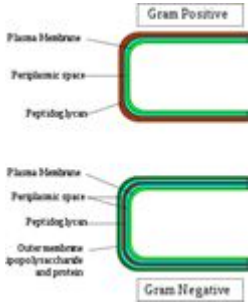
- Traditional microbial species identification
- Transition to genotyping methods
- Using NGS data to genotype pathogens
  - Average nucleotide identity (ANI)
  - MinHash dimensionality-reduction (Mash)

Phenotypic Characteristics of Taxonomic Value

- Morphology
- Motility
- Metabolism
- Physiology and Biochemical Data
- Cell Chemistry
- Others







Phenotypic Approach – Disadvantages

- Need experienced staff
  - Lots of validations, competencies, etc
- Can be a complicated process
  - Multiple tests and results necessary for interpretation
- Labor consuming
  - Hands on process for most tests
- Time consuming
  - Some testing needs to be sequential, often growth required

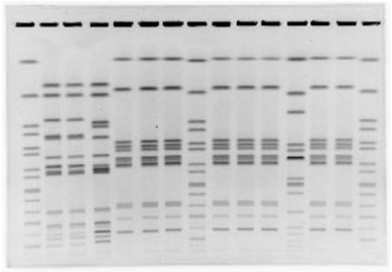
Genotypic Approach

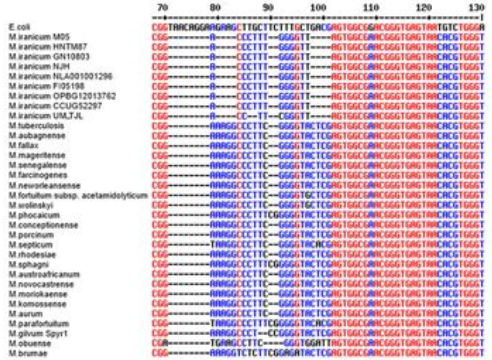
Same genotypes, different phenotypes

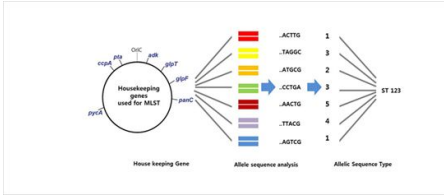


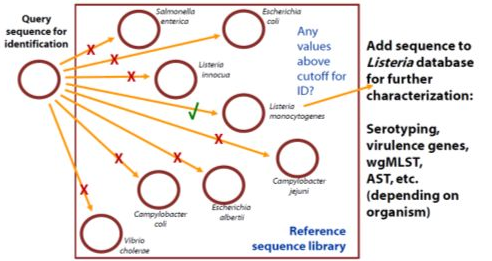

How to Use Sequence Similarity for Species ID

- PFGE
  - global examination of a subset of gDNA from a high level.
- 16S rDNA
  - specific examination of a small, highly discriminatory region of gDNA.
- MLST
  - specific examination of several small discriminatory regions of gDNA.
- NGS
  - global yet specific examination of almost all gDNA.









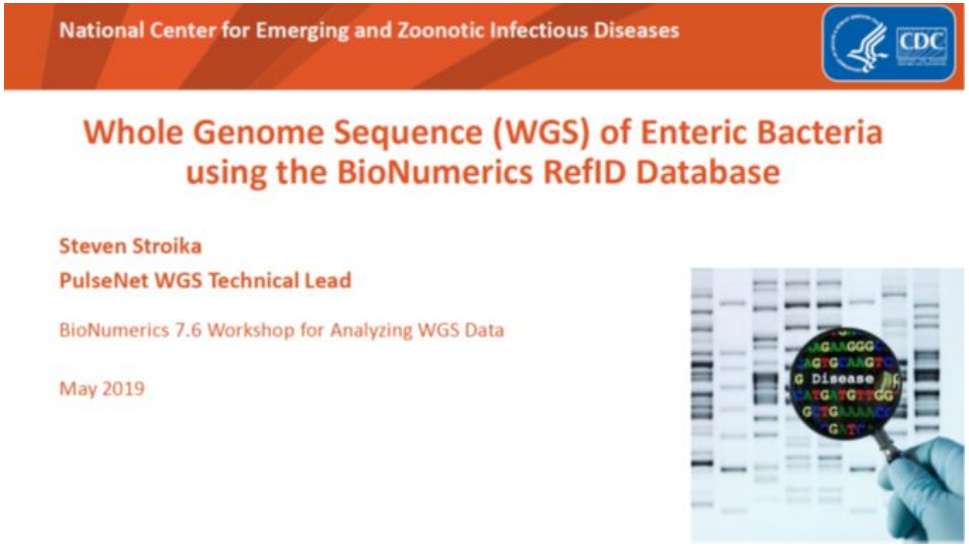
https://www.macrogen.com/en/business/ces_service4.php

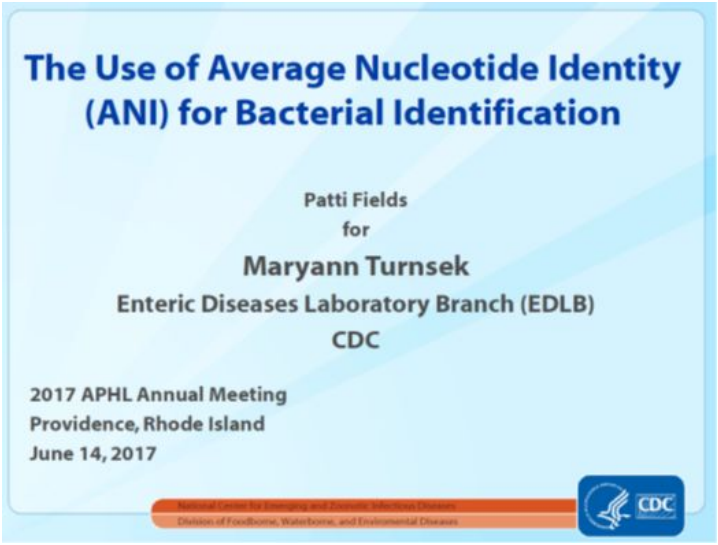
Big Picture

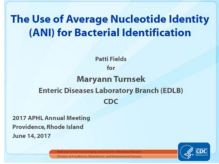
- NGS genotyping methods attempt to match (sequence similarity) experimentally acquired DNA sequences with reference DNA sequences for microbial identification.
  - No reference sequence, no identification.
- Methods vary by:
  - How are the query and reference sequence represented?
    - Are the sequences converted to a new data structure or kept as a string?
    - How are the strings broken up?
    - Are the reads used natively or are they assembled first?
  - What is the algorithm for comparison?
    - Is BLAST used?
    - Is an algorithm optimized for the new data structure used?
    - Does the algorithm use any approximations?

Genotyping - Average Nucleotide Identity (ANI)

The next few slides borrow heavily from the presentations:







What is Average Nucleotide Identity (ANI)

- A computation method to compare two genomes
  - Compares and unknown query sequence to a well-characterized reference genome.
  - Two calculations:
    - Compares the genetic similarity of shared sequences.
    - Determines the proportion of bases aligned.
- Closely mirrors comparisons by DNA-DNA hybridization
  - The traditional gold standard method for determining species boundaries.

How ANI Works

- Aligns shared sequences and calculates percent identical nucleotides
- Answers the question: Are these two genomes the same taxon? Yes or No
- In this example, 53/55 aligned bases = 96.4% identity
- The ANI “cutoff” value for % identity and % bases aligned is determined empirically for each taxon
    - Published values are on the order of 95% identity.

ANI vs DNA-DNA Hybridization

DNA–DNA hybridization values and their relationship to whole-genome sequence similarities

Johan Goris,[1]† Konstantinos T. Konstantinidis,[1]‡ Joel A. Klappenbach,[1] Tom Coenye,[2] Peter Vandamme[2] and James M. Tiedje[1]

[1]Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824, USA

[2]Laboratory for Microbiology, Gent University, K. L. Ledeganckstraat 35, B-9000 Gent, Belgium

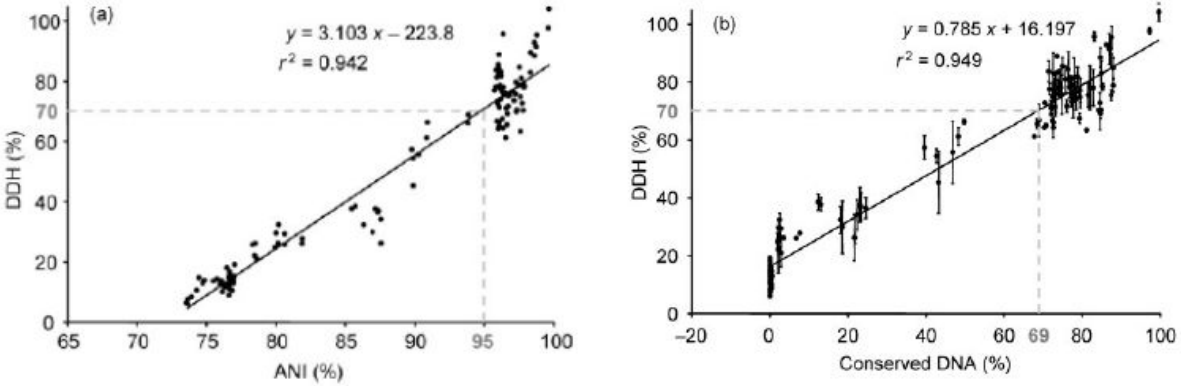


Fig. 1. Relationship between DDH values and genomic sequence identity and conservation. Each filled circle represents the value for DDH between two strains (*y*-axis), plotted against the ANI of the conserved genes between the strains (a) and the percentage of conserved DNA between the strains (b). The standard deviations for the DDH values, omitted from (a) for simplicity, are shown in (b). A linear trend line is shown, but other regression models were evaluated as well (see text). The horizontal broken lines denote the 70% DDH recommendation for species delineation, while the vertical broken lines denote the corresponding ANI (a) and percentage of conserved DNA (b) values for linear regression.

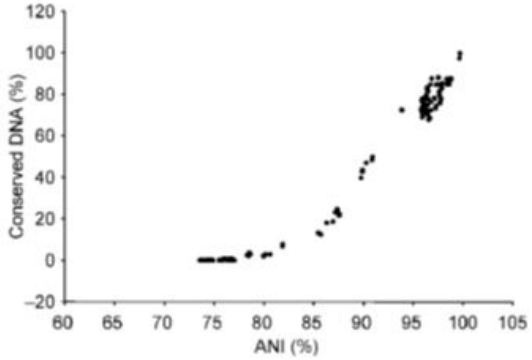


Fig. 2. Relationship between genomic sequence identity and conservation. Each filled circle represents the percentage of conserved DNA shared between two strains (determined at 90% nucleotide identity), plotted against the ANIs of their common genes.

ANI Algorithms

- ANI BLAST used originally (ANIb)
- ANI MUMer used in BN (ANIm)

Shifting the genomic gold standard for the prokaryotic species definition

Michael Richter and Ramon Rosselló-Móra[1]

Marine Microbiology Group, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), E-07190 Esporles, Spain

Edited by James M. Tiedje, Center for Microbial Ecology, East Lansing, MI, and approved September 16, 2009 (received for review June 11, 2009)

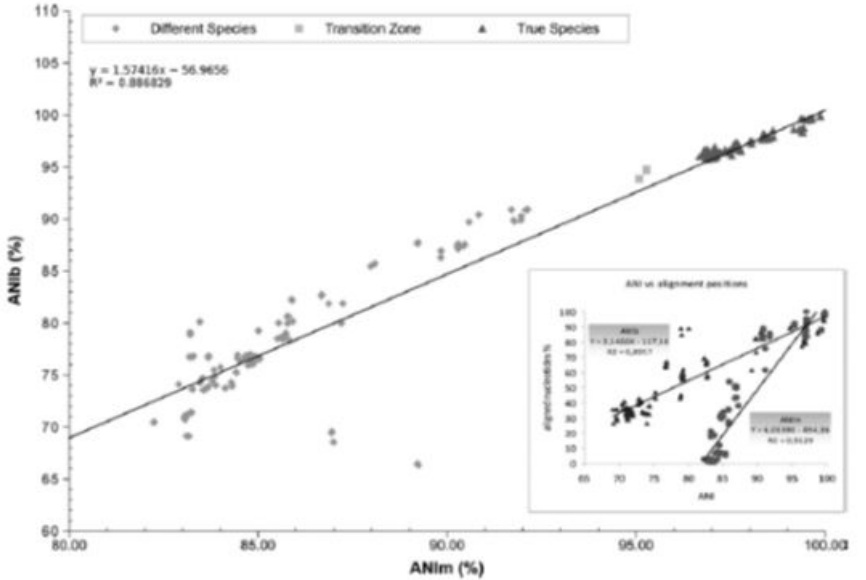


Fig. 1. Plotted results of ANIb versus ANIm. The triangles show those values that correspond to what taxonomists consider as "true" species according to the DDH values traditionally applied and that have previously been classified. *Inset* shows the regression lines of the pairwise comparisons of ANIb or ANIm values with their corresponding percentage of aligned stretches (percentage of nucleotides included in the study).

ANI in Public Health Bioinformatics

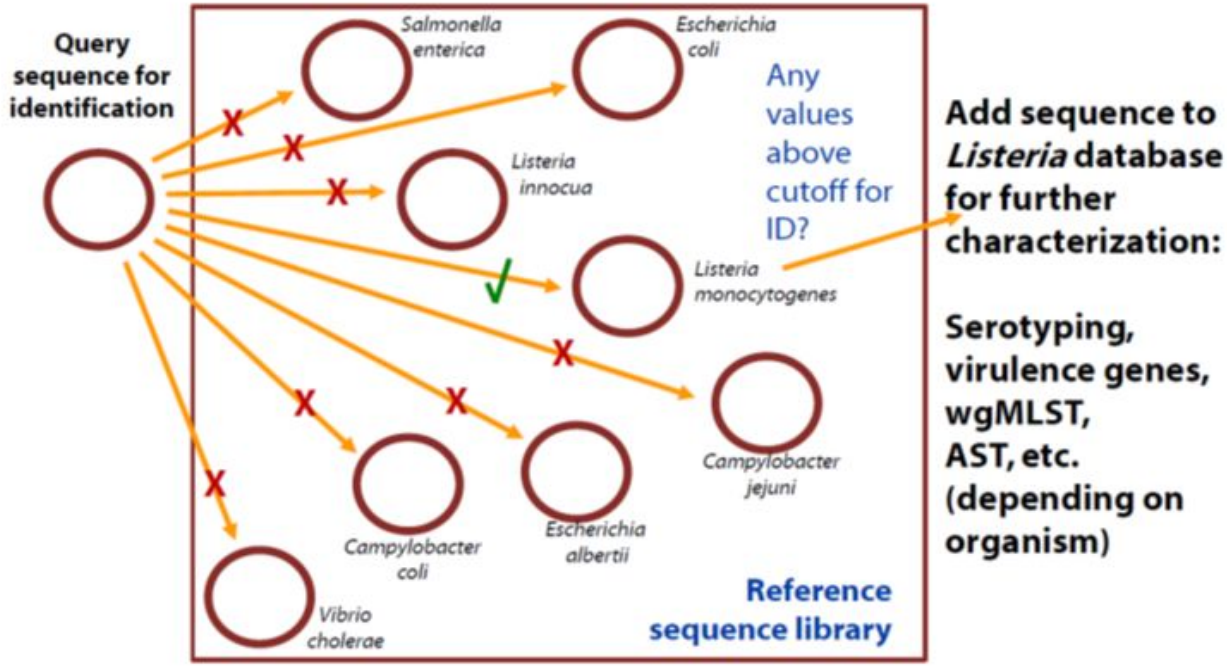
WGS data

Reference Identification Database
(ANI: Genus. Species)

Organism-specific Databases

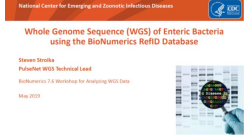
Further characterization:
MLST, serotype, lineage, AST, virulence, plasmids, etc…

How searching works



“…determined empirically for each taxon.”

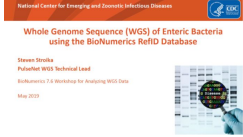
Tableau visualization of ANI values between and within species.



Empirical Values Used in BN

Genera	Species	ANI value (%)	Genome size (MB)
Campylobacter	*coli* *fetus* *jejuni* *lari* *upsaliensis* *hyointestinalis**	≥92	1.4-2.2
Escherichia	*albertii** *coli* and *Shigella* *fergusonii**	≥95	4.5-5.5
Listeria	*innocua** *ivanovii** *marthii** *monocytogenes* *seeligeri** *welshimeri**	≥92	2.7-3.2

Genera	Species	ANI value (%)	Genome size (MB)
Salmonella	*bongori* *enterica*	≥93	4.5-5.0
Vibrio	*cholerae* *Parahaemolyticus* *vulnificus* *alginolyticus** *cidicii** *cincinnatiensis** *fluvialis** *furnissii** *garveyi** *metoecus** *metschnikovii** *mimicus** *navarrensis**	≥95	4.0-5.0



Pros and Cons for ANI

- Pros
  - Replicates species determinations by DNA-DNA hybridization
  - Very rapid: Compare two genomes in seconds
  - Very robust: Reliable answer with 5X sequence coverage (based on down-sampling experiment)
  - Relatively easy to interpret with clear cut off values
- Cons
  - Definitive identification requires representative genome is in the Reference Sequence Library
    - New or unrepresented species cannot be identified
  - Useful for comparing closely related bacteria only
    - Distantly related => No Match
  - As reference library gets bigger, computation time gets longer

Genotyping – Mash (MinHash)

Ondov et al. Genome Biology (2016) 17:132
DOI 10.1186/s13059-016-0997-x

Genome Biology

SOFTWARE Open Access

Mash: fast genome and metagenome distance estimation using MinHash

Brian D. Ondov[1], Todd J. Treangen[1], Páll Melsted[2], Adam B. Mallonee[1], Nicholas H. Bergman[1], Sergey Koren[3] and Adam M. Phillippy[3*]

- Current ANI database for Bionumerics contains ~40 reference genomes.
- Current NCBI Pathogen Detection Browser (January 2020) contains ~500,000 isolates.

If you want to start dramatically increasing your reference database size, use reads rather than assemblies, and keep your search times short, you will need to reduce the dimensionality of your data and make some assumptions.

K-mers and Hash Tables

```
5′-AGGGCGGTTTAATAATCTACGGCTTATTGTTGAACGA-3′

   AGGGCGGTTTAATAATCTACG
    GGGCGGTTTAATAATCTACGG
     GGCGGTTTAATAATCTACGGC
      GCGGTTTAATAATCTACGGCT
       CGGTTTAATAATCTACGGCTT
        GGTTTAATAATCTACGGCTTA
         GTTTAATAATCTACGGCTTAT
          TTTAATAATCTACGGCTTATT
           TTAATAATCTACGGCTTATTG
            TAATAATCTACGGCTTATTGT
             AATAATCTACGGCTTATTGTT
              ATAATCTACGGCTTATTGTTG
               TAATCTACGGCTTATTGTTGA
                AATCTACGGCTTATTGTTGAA
                 ATCTACGGCTTATTGTTGAAC
                  TCTACGGCTTATTGTTGAACG
                   CTACGGCTTATTGTTGAACGA
```

DNA sequence L=37

k-mer size k=21

#k-mers = L – k + 1

17 k-mers

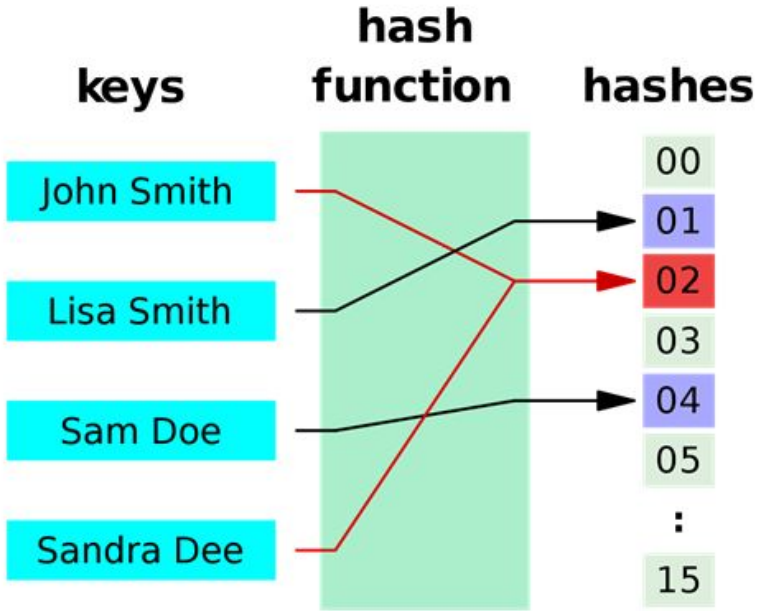
nk possible k-mers

~17+ billion unique k-mers

K-mers and Hash Tables

Convert a string into a number in a reproducible way.

Numbers are faster to compare than text.



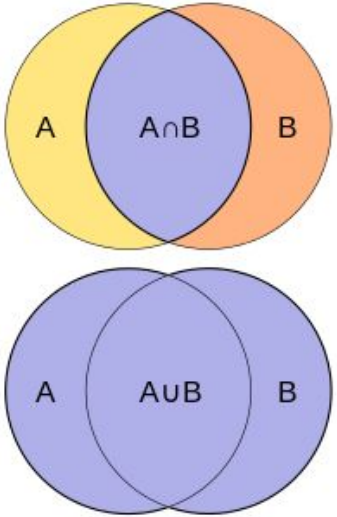
Mash

Jaccard Index

- Compute the ratio of the shared elements over all elements.

Mash

- Uses subsampling

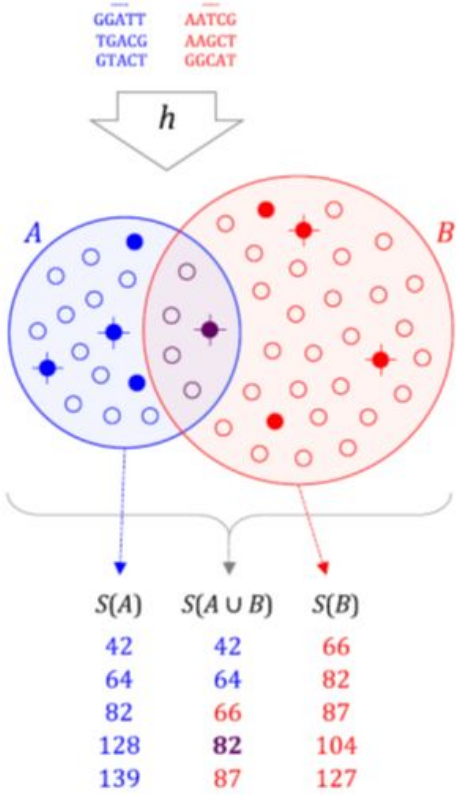


$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

Mash (MinHash)

Mash will:

1. Create a hash sketch from k-mers of user defined size (15, 17, 19, 21, 23, …)
2. Grab the X smallest hash values, where X is user defined (usually 500-1,000)
3. Compare these subsets as an estimate of similarity/dissimilarity and produce a Mash distance.



Mash

The Mash distance correlates well with ANI (or correctly 1-ANI), especially at high levels of similarity.

Not so good for distantly related species.

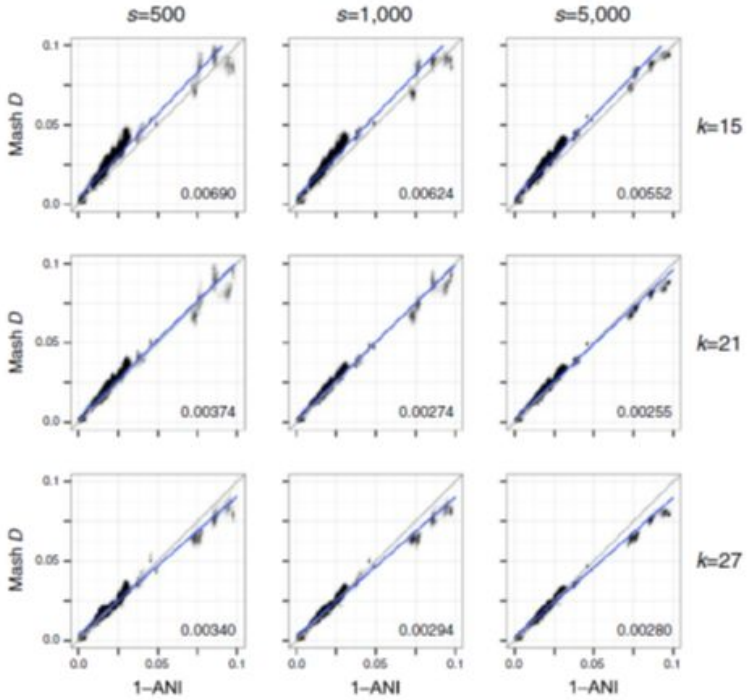


Tableau Analysis of Mash

1. Mash output – command line

Why is all this important?

1. These sequence similarity tools are becoming embedded in our workflows as NGS adoption continues.
2. These tools will need to be validated, and a deeper understanding of how they work, along with parameter optimization, is needed.
3. As NGS data increases, algorithms that use data reduction, subsampling, approximation, etc, to accelerate sequence similarity searching will become more and more necessary in order to take advantage of the wealth of data available.

Additional Resources

http://www.staphb.org/training/pairwise/