# Tutorial 9: Taxonomic Classification of Sequences using DIAMOND

Christopher Uzokwe and Willow Livengood

# Taxonomic Classification Background

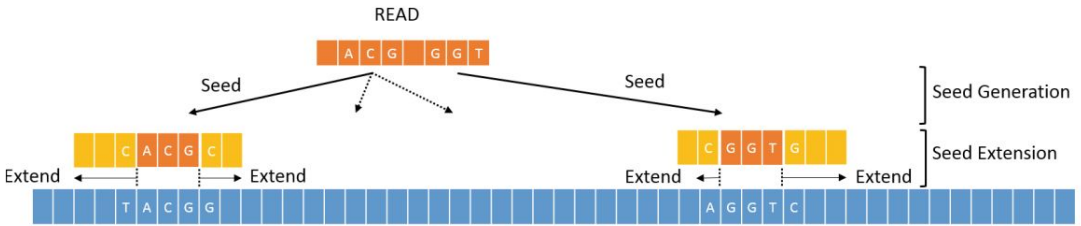
- Goal of taxonomic classification is to analyze sequences to determine the functional or taxonomic content of microbial samples from the environment
- Aligning translated DNA sequences against a reference database of protein sequences, NCBI non-redundant or KEGG
- Millions of sequence reads available
- Alignment of sequencing reads against a protein reference database is a major bottleneck in metagenomics and data-intensive evolutionary projects
- Gold standard is BLASTX, alternate tools USEARCH, BLAT, RAPSearch2 offer only modest speedup or low sensitivity

# DIAMOND Background

- DIAMOND (double index alignment of next-generation sequencing data) [1]
- Benjamin Buchfink, Chao Xie & Daniel H. Huson in 2015
- Goal to replace BLASTX in a high-throughput setting
  - 4 orders of magnitude faster on short DNA reads against the NCBI-nr database
  - Comparable level of sensitivity on alignments with an e-value < $10^{-3}$
- Open source software, implemented in C++
- Designed to run on modern computer architectures that have large memory capacity and many cores
  - High memory server for maximum performance, but can be efficiently handled by a machine with 16GB of memory at about half the speed
  - 16GB RAM is readily available at a price of $160 on a standard desktop computer

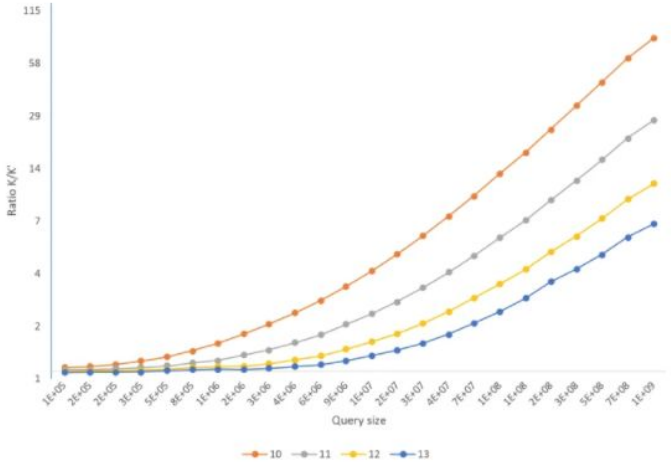
# DIAMOND Steps

1. Seed and Extend
   a. Exact occurrences of short words of fixed length located within reference sequence
   b. Seed matches extended to full alignments
   c. Substantial impact on performance (short seeds increase sensitivity, long increase speed)
2. Reduced Alphabet
   a. Increase speed, without losing sensitivity
   b. Published reductions of 4, 8, 10 letters
   c. Developed alphabet size 11 for best sensitivity: [KREDQN] [C] [G] [H] [ILV] [M] [F] [Y] [W] [P] [STA]
3. Spaced Seeds
   a. Longer seeds with only subset positions used equals better performance, increased sensitivity
   b. 4 shapes of length 15-24, weight 12
4. Seed Index
   a. Decompose the problem
   b. Leverage cache hierarchy



# DIAMOND Steps

1. Double Indexing
    a. Index both the queries and the references
    b. Ability to run in parallel to get all matched seeds
    c. Linear approach for memory access, good for performance
    d. Well-known database sort-merge join algorithm
2. Memory Efficiency
    a. Multiple spaced seeds typically take up to 1.6TB of energy for 16 shapes
    b. Indexes shapes one at a time, sensitive config 16-shape takes maximum energy of 1 shape
    c. Radix cluster technique, seed space is decomposed into 1,024 disjoint partitions
    d. Build and process indexes for only a subset, memory usage limited to size of subset index
3. Seed Extension
    a. Determine if seed match can be extended into 10+ amino acids, Smith-Waterman alignment
    b. Extension only if left-most seed match in corresponding ungapped alignment




# Memory calls for Standard Index / Double Index vs. Query Size

# DIAMOND vs BLAST

| DIAMOND | BLASTX |
|---|---|
| Aligns short sequence reads up to 20,000x faster | Baseline for short sequence reads |
| Similar Sensitivity | Slightly higher sensitivity |
| All-mapper | All-mapper |
| Seed-and-Extend + Double indexing | Seed-and-Extend paradigm |
| Spaced Seeds | Single Consecutive Seeds |
| 2011 study on 12 permafrost samples against KEGG reference database [2] | |
| 246 million reads | 176 million reads |
| 2.3 h on a single workstation | 800,000 CPU hours, supercomputing center |

# Dataset

Diamond creates databases in reference to the input fasta files. We need to use convert the fastq mappings used on Kraken2 to make  the database:

```
sed -n '1~4s/^@/>/p;2~4p' INFILE.fastq > OUTFILE.fasta
```

With the correct fasta files we can make the database for each.

```
[cnu25@node009 eces450650Grp]$ singularity run --bind /ifs/groups/eces450650Grp/ containers/diamond_latest.sif makedb --in ./ECES450650_SP21/Tutorial9/R1.fasta
 -d ./ECES450650_SP21/Tutorial9/R1db.dmnd
```

Need picotte permissions -- or do locally. See text tutorial

# DIAMOND with Picotte/Singularity

Reserve a node from picotte --

```
[cnu25@node009 eces450650Grp]$ srun --nodes=1 --ntasks=20 --cpus-per-task=1 --mem=120GB --time=48:00:00 --pty /bin/bash
```

Run diamond docker on singularity --

```
biobakery_workflows.sif  diamond_latest.sif  kraken2_latest.sif  metaspades_latest.sif  modulefiles  nfcore-magbusco-1.2.0.img  qiime2_latest.sif
conda-qiime2_latest.sif  edirect_latest.sif  metabat_latest.sif  miniconda3             nextflow     qiime                      README
[cnu25@node009 containers]$ pwd
/ifs/groups/eces450650Grp/containers
```

*No parameter run brings up usage notes*

```
Online documentation at http://www.diamondsearch.org
[cnu25@node009 eces450650Grp]$ singularity run --bind /ifs/groups/eces450650Grp/ containers/diamond_latest.sif
```

# DIAMOND with Picotte/Singularity - R1

```
[cnu25@node009 eces450650Grp]$ singularity run --bind /ifs/groups/eces450650Grp/ containers/diamond_latest.sif blastp -q ./ECES450650_SP21/Tutorial9/R1.fasta -d ./ECES450650_SP21/Tutorial9/R1db.dmnd -o ./ECES450650_SP21/Tutorial9/outR2.tsv --very-sensitive
```

singularity run --bind /ifs/groups/eces450650Grp/ containers/diamond_latest.sif blastp -q ./ECES450650_SP21/Tutorial9/R1.fasta -d ./ECES450650_SP21/Tutorial9/R1db.dmnd -o ./ECES450650_SP21/Tutorial9/out.tsv

```
head out.tsv
```

https://github.com/bbuchfink/diamond/wiki/1.-Tutorial <- field descriptions available

```
NS500207:12:H04WYAFXX:2:11209:25191:12566	NS500207:12:H04WYAFXX:2:11209:25191:12566	100.0	18	0	0	1	18	1	18	6.5e-06	32.3
NS500207:12:H04WYAFXX:4:21510:17309:16845	NS500207:12:H04WYAFXX:4:21510:17309:16845	100.0	23	0	0	1	23	1	23	3.6e-07	35.8
NS500207:12:H04WYAFXX:3:21502:15852:7106	NS500207:12:H04WYAFXX:3:21502:15852:7106	100.0	15	0	0	1	15	1	15	3.3e-05	30.4
NS500207:12:H04WYAFXX:2:21203:4549:13503	NS500207:12:H04WYAFXX:2:21203:4549:13503	100.0	15	0	0	1	15	1	15	5.7e-06	32.3
NS500207:12:H04WYAFXX:1:21309:5037:19218	NS500207:12:H04WYAFXX:1:21309:5037:19218	100.0	17	0	0	1	17	1	17	1.1e-06	34.3
NS500207:12:H04WYAFXX:2:11301:10362:4540	NS500207:12:H04WYAFXX:2:11301:10362:4540	100.0	17	0	0	1	17	1	17	2.2e-06	33.5
NS500207:12:H04WYAFXX:2:11110:17918:5260	NS500207:12:H04WYAFXX:2:11110:17918:5260	100.0	15	0	0	1	15	1	15	2.8e-06	33.1
NS500207:12:H04WYAFXX:2:11110:17918:5260	NS500207:12:H04WYAFXX:3:11411:20707:13587	80.0	15	3	0	1	15	3	17	5.2e-04	27.3
NS500207:12:H04WYAFXX:4:11501:4052:3384	NS500207:12:H04WYAFXX:4:11501:4052:3384	100.0	17	0	0	1	17	1	17	1.1e-06	34.3
NS500207:12:H04WYAFXX:3:11505:2023:4796	NS500207:12:H04WYAFXX:3:11505:2023:4796	100.0	19	0	0	1	19	1	19	3.4e-06	33.1
NS500207:12:H04WYAFXX:1:21101:25252:12389	NS500207:12:H04WYAFXX:1:21101:25252:12389	100.0	18	0	0	1	18	1	18	6.9e-08	37.4
NS500207:12:H04WYAFXX:4:11512:19612:10619	NS500207:12:H04WYAFXX:4:11512:19612:10619	100.0	16	0	0	1	16	1	16	7.3e-07	34.7
NS500207:12:H04WYAFXX:4:11512:19612:10619	NS500207:12:H04WYAFXX:3:11601:23999:14913	85.7	14	2	0	3	16	4	17	7.7e-04	26.9
NS500207:12:H04WYAFXX:4:21608:23571:16090	NS500207:12:H04WYAFXX:4:21608:23571:16090	100.0	39	0	0	1	39	1	39	3.5e-11	47.0
NS500207:12:H04WYAFXX:3:11404:6524:14416	NS500207:12:H04WYAFXX:3:11404:6524:14416	100.0	16	0	0	1	16	1	16	1.2e-05	31.6
NS500207:12:H04WYAFXX:2:21110:13416:3566	NS500207:12:H04WYAFXX:2:21110:13416:3566	100.0	22	0	0	1	22	1	22	8.4e-08	37.4
NS500207:12:H04WYAFXX:2:21205:15790:2428	NS500207:12:H04WYAFXX:2:21205:15790:2428	100.0	17	0	0	1	17	1	17	3.1e-06	33.1
```

# DIAMOND with Picotte/Singularity - R2

```
[cnu25@node001 eces450650Grp]$ singularity run --bind /ifs/groups/eces450650Grp/ containers/diamond_latest.sif blastp -q ./ECES450650_SP21/Tutorial9/R2.fasta -d ./ECES450650_SP21/Tutorial9/R2db.dmnd
-o ./ECES450650_SP21/Tutorial9/R2out.tsv --very-sensitive
```

singularity run --bind /ifs/groups/eces450650Grp/ containers/diamond_latest.sif blastp -q ./ECES450650_SP21/Tutorial9/R2.fasta -d ./ECES450650_SP21/Tutorial9/R2db.dmnd -o ./ECES450650_SP21/Tutorial9/R2out.tsv

https://github.com/bbuchfink/diamond/wiki/1.-Tutorial <- field descriptions available

```
[cnu25@node001 Tutorial9]$ head R2out.tsv
NS500207:12:H04WYAFXX:4:21604:20798:5389        NS500207:12:H04WYAFXX:4:21604:20798:5389        100.0   25      0       0       1       25      1       25      4.0e-09 40.8
NS500207:12:H04WYAFXX:4:21404:17427:17699       NS500207:12:H04WYAFXX:4:21404:17427:17699       100.0   18      0       0       1       18      1       18      1.1e-06 34.3
NS500207:12:H04WYAFXX:3:11602:24840:19766       NS500207:12:H04WYAFXX:3:11602:24840:19766       100.0   16      0       0       1       16      1       16      1.6e-05 31.2
NS500207:12:H04WYAFXX:3:11505:24990:12808       NS500207:12:H04WYAFXX:3:11505:24990:12808       100.0   15      0       0       1       15      1       15      5.4e-06 32.3
NS500207:12:H04WYAFXX:1:21101:21945:8146        NS500207:12:H04WYAFXX:1:21101:21945:8146        100.0   17      0       0       1       17      1       17      5.1e-07 35.0
NS500207:12:H04WYAFXX:3:11612:15415:11197       NS500207:12:H04WYAFXX:3:11612:15415:11197       100.0   17      0       0       1       17      1       17      1.7e-05 31.2
NS500207:12:H04WYAFXX:3:11402:16135:10210       NS500207:12:H04WYAFXX:3:11402:16135:10210       100.0   18      0       0       1       18      1       18      2.2e-06 33.5
NS500207:12:H04WYAFXX:4:11410:13187:13268       NS500207:12:H04WYAFXX:4:11410:13187:13268       100.0   19      0       0       1       19      1       19      6.8e-08 37.4
NS500207:12:H04WYAFXX:1:11109:23245:5947        NS500207:12:H04WYAFXX:1:11109:23245:5947        100.0   16      0       0       1       16      1       16      2.0e-06 33.5
NS500207:12:H04WYAFXX:3:21408:12189:14108       NS500207:12:H04WYAFXX:3:21408:12189:14108       100.0   16      0       0       1       16      1       16      4.0e-06 32.7
```

# DIAMOND vs Kraken2 Results

| | DIAMOND | Kraken2 |
|---|---|---|
| Unclassified (%) | 15.41% | 95.04% |
| Classified (%) | 84.59% | 4.96% |
| Runtime | ~2.56s | ~0.039s |

```
Total time = 2.557s
Reported 12573 pairwise alignments, 12573 HSPs.
10635 queries aligned.
```

# Q&A

[1] Buchfink, Benjamin, Chao Xie, and Daniel H Huson. “Fast and sensitive protein alignment using DIAMOND.” *Nature Methods* 12 (2015): 59-60. https://doi.org/10.1038/nmeth.3176

[2] Mackelprang, Rachel, Mark P. Waldrop, Kristen M. DeAngelis, Maude M. David, Krystle L. Chavarria, Steven J. Blazewicz, Edward M. Rubin, and Janet K. Jansson. “Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw.” *Nature* 480 (2011): 368-371. https://doi.org/10.1038/nature10576

[3] Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. “Basic local alignment search tool.” *Journal of Molecular Biology* 215, is. 3 (1990): 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2