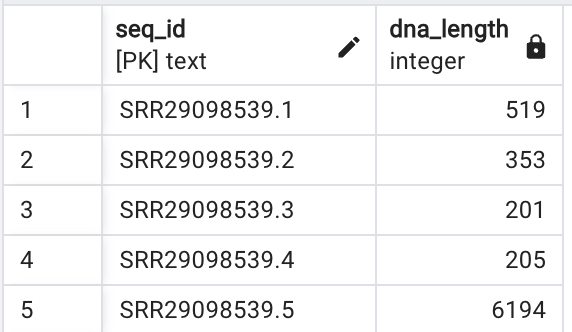
Test outputs

Real-world data (17020 dna sequences, 30032947 kmers).

**-- Test length functions**

SELECT seq\_id, length(seq) AS dna\_length

FROM dna\_sequences;



"QUERY PLAN"

"Seq Scan on dna\_sequences (cost=0.00..1143.38 rows=10190 width=21) (actual time=0.011..4.577 rows=10190 loops=1)"

"Planning Time: 0.049 ms"

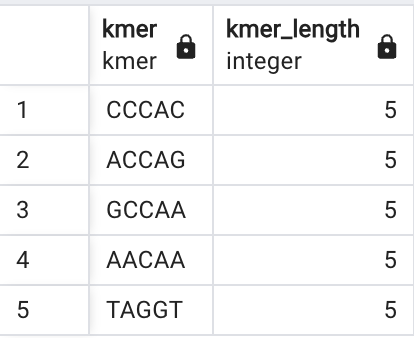
"Execution Time: 5.202 ms"

—------------------------------------

SELECT DISTINCT kmer, length(kmer) AS kmer\_length

FROM kmers

LIMIT 5;



"QUERY PLAN"

"Limit (cost=746399.89..746399.95 rows=5 width=14) (actual time=3906.110..3906.111 rows=5 loops=1)"

" -> HashAggregate (cost=746399.89..746850.82 rows=36074 width=14) (actual time=3906.109..3906.109 rows=5 loops=1)"

" Group Key: kmer, length(kmer)"

" Batches: 1 Memory Usage: 1617kB"

" -> Seq Scan on kmers (cost=0.00..596237.35 rows=30032508 width=14) (actual time=0.086..1462.618 rows=30032947 loops=1)"

"Planning Time: 0.081 ms"

"Execution Time: 3906.382 ms"

**-- Test k-mer counting**

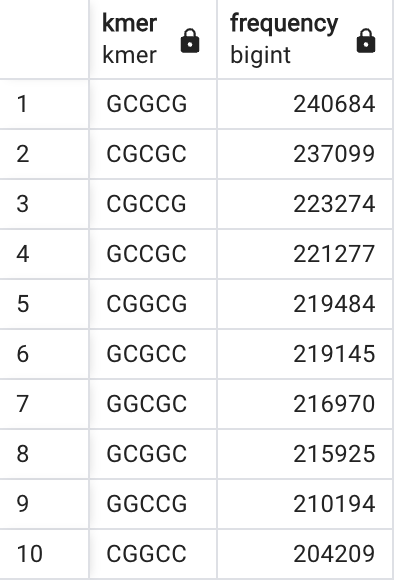
SELECT kmer, COUNT(\*) AS frequency

FROM kmers

GROUP BY kmer

ORDER BY frequency DESC

LIMIT 10;



"QUERY PLAN"

"Limit (cost=418610.74..418610.77 rows=10 width=18) (actual time=1175.630..1176.433 rows=10 loops=1)"

" -> Sort (cost=418610.74..418700.93 rows=36074 width=18) (actual time=1175.629..1176.431 rows=10 loops=1)"

" Sort Key: (count(\*)) DESC"

" Sort Method: top-N heapsort Memory: 26kB"

" -> Finalize HashAggregate (cost=417470.45..417831.19 rows=36074 width=18) (actual time=1175.453..1176.374 rows=1024 loops=1)"

" Group Key: kmer"

" Batches: 1 Memory Usage: 913kB"

" -> Gather (cost=409534.17..417109.71 rows=72148 width=18) (actual time=1174.724..1175.925 rows=3072 loops=1)"

" Workers Planned: 2"

" Workers Launched: 2"

" -> Partial HashAggregate (cost=408534.17..408894.91 rows=36074 width=18) (actual time=1169.714..1169.910 rows=1024 loops=3)"

" Group Key: kmer"

" Batches: 1 Memory Usage: 1681kB"

" Worker 0: Batches: 1 Memory Usage: 1681kB"

" Worker 1: Batches: 1 Memory Usage: 1681kB"

" -> Parallel Seq Scan on kmers (cost=0.00..345966.45 rows=12513545 width=10) (actual time=0.095..377.429 rows=10010982 loops=3)"

"Planning Time: 0.104 ms"

"Execution Time: 1176.860 ms"

**-- Most frequent k-mers**

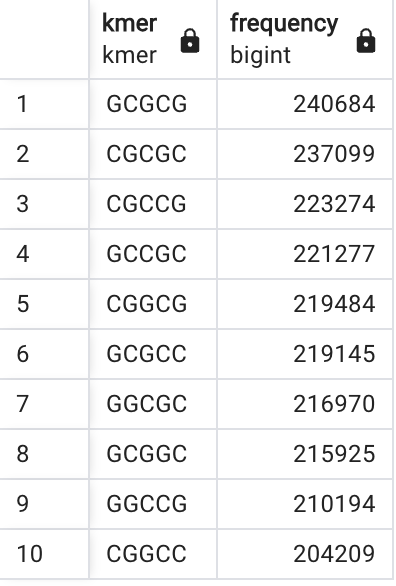
SELECT kmer, COUNT(\*) AS frequency

FROM kmers

GROUP BY kmer

ORDER BY frequency DESC

LIMIT 10;



"QUERY PLAN"

"Limit (cost=418610.74..418610.77 rows=10 width=18) (actual time=1182.487..1183.118 rows=10 loops=1)"

" -> Sort (cost=418610.74..418700.93 rows=36074 width=18) (actual time=1182.484..1183.115 rows=10 loops=1)"

" Sort Key: (count(\*)) DESC"

" Sort Method: top-N heapsort Memory: 26kB"

" -> Finalize HashAggregate (cost=417470.45..417831.19 rows=36074 width=18) (actual time=1182.323..1183.062 rows=1024 loops=1)"

" Group Key: kmer"

" Batches: 1 Memory Usage: 913kB"

" -> Gather (cost=409534.17..417109.71 rows=72148 width=18) (actual time=1181.573..1182.584 rows=3072 loops=1)"

" Workers Planned: 2"

" Workers Launched: 2"

" -> Partial HashAggregate (cost=408534.17..408894.91 rows=36074 width=18) (actual time=1176.983..1177.174 rows=1024 loops=3)"

" Group Key: kmer"

" Batches: 1 Memory Usage: 1681kB"

" Worker 0: Batches: 1 Memory Usage: 1681kB"

" Worker 1: Batches: 1 Memory Usage: 1681kB"

" -> Parallel Seq Scan on kmers (cost=0.00..345966.45 rows=12513545 width=10) (actual time=0.091..378.269 rows=10010982 loops=3)"

"Planning Time: 0.168 ms"

"Execution Time: 1183.590 ms"

**-- K-mer frequency per sequence**

SELECT s.seq\_id, k.kmer, COUNT(\*) AS frequency

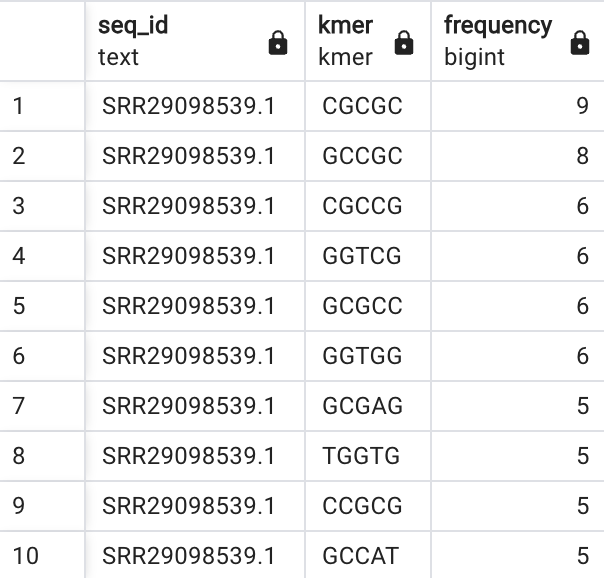
FROM kmers k

JOIN dna\_sequences s ON k.seq\_id = s.seq\_id

GROUP BY s.seq\_id, k.kmer

ORDER BY s.seq\_id, frequency DESC

LIMIT 10;

****

"QUERY PLAN"

"Limit (cost=7082516.57..7082516.60 rows=10 width=35) (actual time=13200.455..13200.457 rows=10 loops=1)"

" -> Sort (cost=7082516.57..7157597.84 rows=30032508 width=35) (actual time=13200.454..13200.455 rows=10 loops=1)"

" Sort Key: s.seq\_id, (count(\*)) DESC"

" Sort Method: top-N heapsort Memory: 26kB"

" -> HashAggregate (cost=5311998.40..6433524.87 rows=30032508 width=35) (actual time=7383.887..12799.815 rows=6405026 loops=1)"

" Group Key: s.seq\_id, k.kmer"

" Planned Partitions: 128 Batches: 641 Memory Usage: 4409kB Disk Usage: 1550344kB"

" -> Hash Join (cost=624.51..600648.71 rows=30032508 width=27) (actual time=4.026..3231.557 rows=30032947 loops=1)"

" Hash Cond: (k.seq\_id = s.seq\_id)"

" -> Seq Scan on kmers k (cost=0.00..521156.08 rows=30032508 width=27) (actual time=0.057..1033.756 rows=30032947 loops=1)"

" -> Hash (cost=497.14..497.14 rows=10190 width=17) (actual time=3.944..3.944 rows=10190 loops=1)"

" Buckets: 16384 Batches: 1 Memory Usage: 620kB"

" -> Index Only Scan using dna\_sequences\_pkey on dna\_sequences s (cost=0.29..497.14 rows=10190 width=17) (actual time=0.009..1.717 rows=10190 loops=1)"

" Heap Fetches: 8"

"Planning Time: 0.370 ms"

"Execution Time: 13219.133 ms"

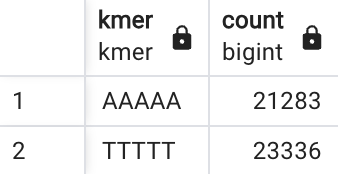
**-- Canonicalize**

select kmer, count(\*)

from kmers

WHERE kmer ='AAAAA' or kmer ='TTTTT'

group by kmer;



"QUERY PLAN"

"Finalize HashAggregate (cost=409637.00..409647.81 rows=1081 width=18) (actual time=539.426..540.725 rows=2 loops=1)"

" Group Key: kmer"

" Batches: 1 Memory Usage: 73kB"

" -> Gather (cost=409536.46..409632.43 rows=914 width=18) (actual time=539.376..540.717 rows=6 loops=1)"

" Workers Planned: 2"

" Workers Launched: 2"

" -> Partial HashAggregate (cost=408536.46..408541.03 rows=457 width=18) (actual time=534.547..534.549 rows=2 loops=3)"

" Group Key: kmer"

" Batches: 1 Memory Usage: 37kB"

" Worker 0: Batches: 1 Memory Usage: 37kB"

" Worker 1: Batches: 1 Memory Usage: 37kB"

" -> Parallel Seq Scan on kmers (cost=0.00..408534.18 rows=457 width=10) (actual time=0.369..532.731 rows=14873 loops=3)"

" Filter: ((kmer = 'AAAAA'::kmer) OR (kmer = 'TTTTT'::kmer))"

" Rows Removed by Filter: 9996109"

"Planning Time: 0.113 ms"

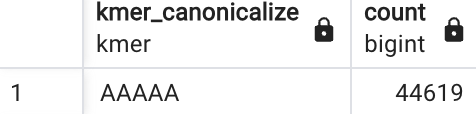
"Execution Time: 540.778 ms"

—--------------------------------

select kmer\_canonicalize(kmer), count(\*)

from kmers

group by kmer\_canonicalize(kmer);



"QUERY PLAN"

"Finalize HashAggregate (cost=1560485.73..1562063.97 rows=36074 width=40) (actual time=1538.047..1539.155 rows=512 loops=1)"

" Group Key: (kmer\_canonicalize(kmer))"

" Batches: 1 Memory Usage: 849kB"

" -> Gather (cost=1375423.43..1554172.78 rows=72148 width=40) (actual time=1537.652..1538.878 rows=1536 loops=1)"

" Workers Planned: 2"

" Workers Launched: 2"

" -> Partial HashAggregate (cost=1374423.43..1545957.98 rows=36074 width=40) (actual time=1534.420..1534.519 rows=512 loops=3)"

" Group Key: kmer\_canonicalize(kmer)"

" Batches: 1 Memory Usage: 849kB"

" Worker 0: Batches: 1 Memory Usage: 849kB"

" Worker 1: Batches: 1 Memory Usage: 849kB"

" -> Parallel Seq Scan on kmers (cost=0.00..377250.31 rows=12513545 width=32) (actual time=0.061..791.829 rows=10010982 loops=3)"

"Planning Time: 0.100 ms"

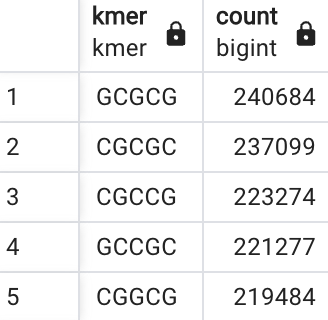
"Execution Time: 1539.335 ms"

-- Count all 5-mers in a DNA sequence

SELECT k.kmer, count(\*)

FROM kmers k GROUP BY k.kmer

ORDER BY count(\*) DESC;



"QUERY PLAN"

"Sort (cost=420561.76..420651.94 rows=36074 width=18) (actual time=1176.248..1176.886 rows=1024 loops=1)"

" Sort Key: (count(\*)) DESC"

" Sort Method: quicksort Memory: 129kB"

" -> Finalize HashAggregate (cost=417470.45..417831.19 rows=36074 width=18) (actual time=1176.031..1176.773 rows=1024 loops=1)"

" Group Key: kmer"

" Batches: 1 Memory Usage: 913kB"

" -> Gather (cost=409534.17..417109.71 rows=72148 width=18) (actual time=1175.252..1176.316 rows=3072 loops=1)"

" Workers Planned: 2"

" Workers Launched: 2"

" -> Partial HashAggregate (cost=408534.17..408894.91 rows=36074 width=18) (actual time=1170.557..1170.773 rows=1024 loops=3)"

" Group Key: kmer"

" Batches: 1 Memory Usage: 1681kB"

" Worker 0: Batches: 1 Memory Usage: 1681kB"

" Worker 1: Batches: 1 Memory Usage: 1681kB"

" -> Parallel Seq Scan on kmers k (cost=0.00..345966.45 rows=12513545 width=10) (actual time=0.095..377.816 rows=10010982 loops=3)"

"Planning Time: 0.061 ms"

"Execution Time: 1177.312 ms"

-- Return the total, distinct and unique count of 5-mers in a DNA sequence

WITH kmers AS (

SELECT k.kmer, count(\*)

FROM kmers k GROUP BY k.kmer

)

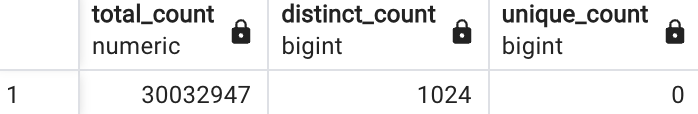
SELECT sum(count) AS total\_count,

count(\*) AS distinct\_count,

count(\*) FILTER (WHERE count = 1) AS unique\_count FROM kmers;

CREATE INDEX kmer\_spgist\_idx ON kmers USING spgist (kmer);

set enable\_seqscan = off;



"QUERY PLAN"

"Aggregate (cost=418552.68..418552.69 rows=1 width=48) (actual time=1185.856..1186.765 rows=1 loops=1)"

" -> Finalize HashAggregate (cost=417470.45..417831.19 rows=36074 width=18) (actual time=1185.710..1186.737 rows=1024 loops=1)"

" Group Key: k.kmer"

" Batches: 1 Memory Usage: 913kB"

" -> Gather (cost=409534.17..417109.71 rows=72148 width=18) (actual time=1184.999..1186.288 rows=3072 loops=1)"

" Workers Planned: 2"

" Workers Launched: 2"

" -> Partial HashAggregate (cost=408534.17..408894.91 rows=36074 width=18) (actual time=1181.565..1181.766 rows=1024 loops=3)"

" Group Key: k.kmer"

" Batches: 1 Memory Usage: 1681kB"

" Worker 0: Batches: 1 Memory Usage: 1681kB"

" Worker 1: Batches: 1 Memory Usage: 1681kB"

" -> Parallel Seq Scan on kmers k (cost=0.00..345966.45 rows=12513545 width=10) (actual time=0.109..388.540 rows=10010982 loops=3)"

"Planning Time: 0.072 ms"

"Execution Time: 1186.861 ms"

—------------------------------------

**Performance tests:**

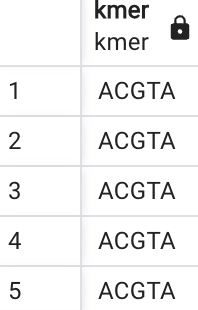
No index (sequential scan):

**-- Test equality operator**

SELECT kmer

FROM kmers

WHERE kmer = 'ACGTA'::kmer;



(5644 rows)

"QUERY PLAN"

"Seq Scan on kmers (cost=10000000000.00..10000596243.70 rows=572 width=10) (actual time=0.891..1076.090 rows=5644 loops=1)"

" Filter: (kmer = 'ACGTA'::kmer)"

" Rows Removed by Filter: 30027303"

"Planning Time: 0.068 ms"

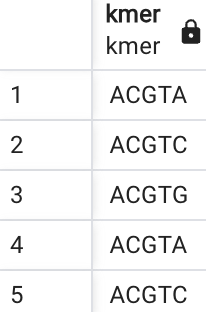
"Execution Time: 1076.220 ms"

**-- Test starts\_with with kmer**

SELECT kmer

FROM kmers

WHERE kmer ^@ 'ACGT';



55257 rows

"QUERY PLAN"

"Seq Scan on kmers (cost=10000000000.00..10000596243.70 rows=15016508 width=10) (actual time=0.404..1110.929 rows=55257 loops=1)"

" Filter: (kmer ^@ 'ACGT'::kmer)"

" Rows Removed by Filter: 29977690"

"Planning Time: 0.035 ms"

"Execution Time: 1111.875 ms"

**-- Test contains**

SELECT kmer

FROM kmers

WHERE 'ACG' @> kmer;

"QUERY PLAN" "Seq Scan on kmers (cost=10000000000.00..10000596242.85 rows=15016474 width=10) (actual time=1039.801..1039.801 rows=0 loops=1)"

" Filter: ('ACG'::qkmer @> kmer)"

" Rows Removed by Filter: 30032947"

"Planning Time: 0.405 ms"

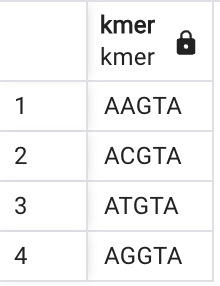
"Execution Time: 1039.815 ms"

**-- Test contains with qkmer**

SELECT DISTINCT kmer

FROM kmers

WHERE 'ANGTA'::qkmer @> kmer;



"QUERY PLAN"

"HashAggregate (cost=633777.99..634138.73 rows=36074 width=10) (actual time=1116.981..1117.120 rows=4 loops=1)"

" Group Key: kmer"

" Batches: 1 Memory Usage: 1561kB"

" -> Seq Scan on kmers (cost=0.00..596237.35 rows=15016254 width=10) (actual time=0.356..1113.237 rows=28750 loops=1)"

" Filter: ('ANGTA'::qkmer @> kmer)"

" Rows Removed by Filter: 30004197"

"Planning Time: 0.063 ms"

"Execution Time: 1117.199 ms"

Spgist index

**-- Test equality operator**

"QUERY PLAN"

"Index Only Scan using kmer\_spgist\_idx on kmers (cost=0.42..22.43 rows=572 width=10) (actual time=0.913..3.154 rows=5644 loops=1)"

" Index Cond: (kmer = 'ACGTA'::kmer)"

" Heap Fetches: 0"

"Planning Time: 2.064 ms"

"Execution Time: 3.396 ms"

**-- Test starts\_with with kmer**

"QUERY PLAN"

"Index Only Scan using kmer\_spgist\_idx on kmers (cost=0.42..506344.71 rows=15016474 width=10) (actual time=0.149..17.509 rows=55257 loops=1)"

" Index Cond: (kmer ^@ 'ACGT'::kmer)"

" Heap Fetches: 0"

"Planning Time: 0.306 ms"

"Execution Time: 19.516 ms"

**-- Test contains**

"QUERY PLAN"

"Index Only Scan using kmer\_spgist\_idx on kmers (cost=0.42..506344.71 rows=15016474 width=10) (actual time=0.017..0.017 rows=0 loops=1)"

" Index Cond: (kmer <@ 'ACG'::qkmer)"

" Heap Fetches: 0"

"Planning Time: 0.124 ms"

"Execution Time: 0.082 ms"

**-- Test contains with qkmer**

"QUERY PLAN"

"Index Only Scan using kmer\_spgist\_idx on kmers (cost=0.42..506468.71 rows=15016474 width=10) (actual time=0.118..10.296 rows=28750 loops=1)"

" Index Cond: (kmer <@ 'ANGTA'::qkmer)"

" Heap Fetches: 0"

"Planning Time: 0.089 ms"

"Execution Time: 11.440 ms"

Hash Index

**-- Test equality operator**

"QUERY PLAN"

"Bitmap Heap Scan on kmers (cost=20.43..2228.25 rows=572 width=10) (actual time=1.641..36.521 rows=5644 loops=1)"

" Recheck Cond: (kmer = 'ACGTA'::kmer)"

" Heap Blocks: exact=5511"

" -> Bitmap Index Scan on kmer\_hash\_idx (cost=0.00..20.29 rows=572 width=0) (actual time=0.900..0.901 rows=5644 loops=1)"

" Index Cond: (kmer = 'ACGTA'::kmer)"

"Planning Time: 0.791 ms"

"Execution Time: 36.817 ms"