# ByteWise Data Engineering
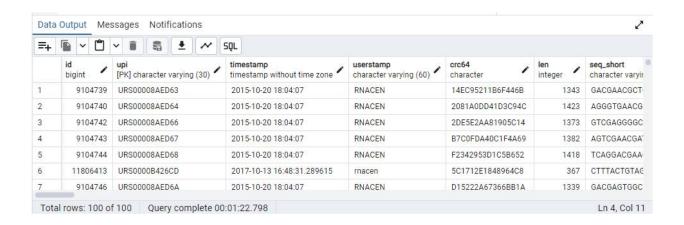
## Task 2

## Query 1

Write a query to get data having length of Rna structures more than 12 with them being added after 2008

```
SELECT *
        FROM rnacen.rna
        where (len > 12) AND (Extract(year from "timestamp") > 2008)
        limit 100
```



| | id<br>bigint | upi<br>[PK] character varying (30) | timestamp<br>timestamp without time zone | userstamp<br>character varying (60) | crc64<br>character | len<br>integer | seq_short<br>character varyi |
|---|---|---|---|---|---|---|---|
| 1 | 9104739 | URS00008AED63 | 2015-10-20 18:04:07 | RNACEN | 14EC95211B6F446B | 1343 | GACGAACGCT |
| 2 | 9104740 | URS00008AED64 | 2015-10-20 18:04:07 | RNACEN | 2081A0DD41D3C94C | 1423 | AGGGTGAACG |
| 3 | 9104742 | URS00008AED66 | 2015-10-20 18:04:07 | RNACEN | 2DE5E2AA81905C14 | 1373 | GTCGAGGGGC |
| 4 | 9104743 | URS00008AED67 | 2015-10-20 18:04:07 | RNACEN | B7C0FDA40C1F4A69 | 1382 | AGTCGAACGA |
| 5 | 9104744 | URS00008AED68 | 2015-10-20 18:04:07 | RNACEN | F2342953D1C5B652 | 1418 | TCAGGACGAA |
| 6 | 11806413 | URS0000B426CD | 2017-10-13 16:48:31.289615 | rnacen | 5C1712E1848964C8 | 367 | CTTTACTGTAG |
| 7 | 9104746 | URS00008AED6A | 2015-10-20 18:04:07 | RNACEN | D15222A67366BB1A | 1339 | GACGAGTGGC |

Total rows: 100 of 100    Query complete 00:01:22.798    Ln 4, Col 11
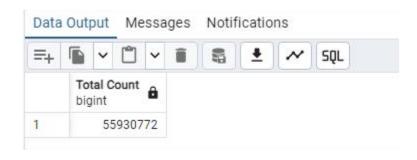
## Query 2:

How many pre computed RNA are present that are still active and got their last release update before 2022

SELECT count(*) AS "Total Count"

    from rnacen.rnc_rna_precomputed

    where is_active = true and EXTRACT (year from update_date) < 2022

Data Output   Messages   Notifications

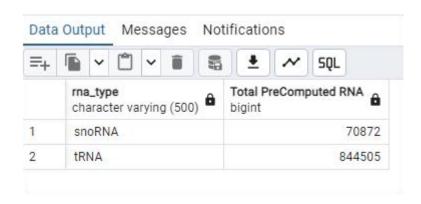| | Total Count 🔒 bigint |
|---|---|
| 1 | 55930772 |

## Query 3

How many total pre computed RNA records for snoRNA and tRNA were recorded in 2011, 2016, 2014, and 2020

SELECT rna_type, COUNT(*) AS "Total PreComputed RNA"

FROM rnacen.rnc_rna_precomputed

WHERE rna_type IN ('snoRNA', 'tRNA')

AND EXTRACT(year FROM update_date) IN (2011, 2016, 2014, 2020)

GROUP BY rna_type;

Data Output   Messages   Notifications

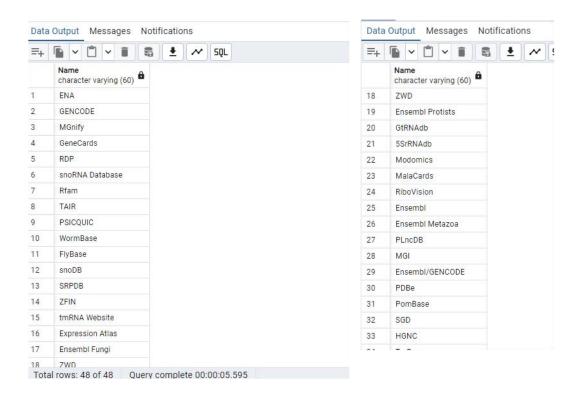| | rna_type<br>character varying (500) | Total PreComputed RNA<br>bigint |
|---|---|---|
| 1 | snoRNA | 70872 |
| 2 | tRNA | 844505 |

## Query 4

Can you give me the names of all databases built for RNA with minimum length other than 100, 200, 300, 400 and 15

SELECT display_name AS "Name"

    FROM rnacen.rnc_database

    WHERE min_length not in (100, 200, 300, 400, 15)

Data Output | Messages | Notifications

| | Name character varying (60) |
|---|---|
| 1 | ENA |
| 2 | GENCODE |
| 3 | MGnify |
| 4 | GeneCards |
| 5 | RDP |
| 6 | snoRNA Database |
| 7 | Rfam |
| 8 | TAIR |
| 9 | PSICQUIC |
| 10 | WormBase |
| 11 | FlyBase |
| 12 | snoDB |
| 13 | SRPDB |
| 14 | ZFIN |
| 15 | tmRNA Website |
| 16 | Expression Atlas |
| 17 | Ensembl Fungi |
| 18 | ZWD |

Total rows: 48 of 48    Query complete 00:00:05.595

Data Output | Messages | Notifications

| | Name character varying (60) |
|---|---|
| 18 | ZWD |
| 19 | Ensembl Protists |
| 20 | GtRNAdb |
| 21 | 5SrRNAdb |
| 22 | Modomics |
| 23 | MalaCards |
| 24 | RiboVision |
| 25 | Ensembl |
| 26 | Ensembl Metazoa |
| 27 | PLncDB |
| 28 | MGI |
| 29 | Ensembl/GENCODE |
| 30 | PDBe |
| 31 | PomBase |
| 32 | SGD |
| 33 | HGNC |

Data Output | Messages | Notifications

| | Name character varying (60) |
|---|---|
| 33 | HGNC |
| 34 | TarBase |
| 35 | Ribocentre |
| 36 | NONCODE |
| 37 | REDIPortal |
| 38 | CRW |
| 39 | lncRNAdb |
| 40 | EVLncRNAs |
| 41 | snOPY |
| 42 | MirGeneDB |
| 43 | RGD |
| 44 | Greengenes |
| 45 | dictyBase |
| 46 | LncBook |
| 47 | LncBase |
| 48 | IntAct |

## Query 5

Can you get complete 500 records of sequences for active regions and name your column as myregions in which you are getting the region name column value. Then tell me what different chromosomes with exon_count we have for regions including center, east and north using the name you set for your column

SELECT

    sr.region_name AS myregions, sr.chromosome, sr.exon_count

    FROM rnacen.rnc_sequence_regions sr

    LEFT JOIN rnacen.rnc_rna_precomputed rp

    ON sr.urs_taxid = rp.id

    WHERE rp.is_active = true AND sr.region_name in ("center", "east", "north")

    limit 100