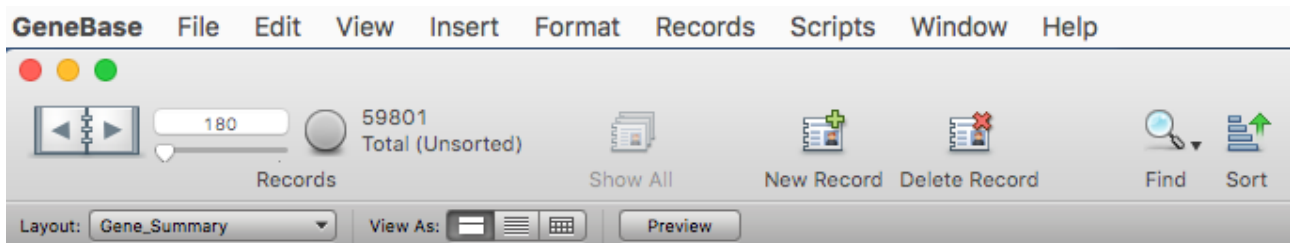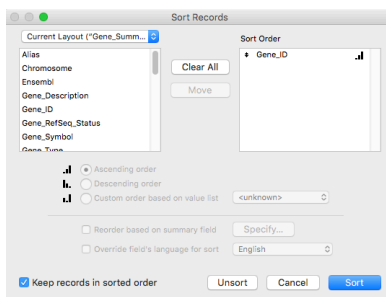**GeneBase 1.1 searches useful to reproduce count and statistic numbers present in Figures, Tables and Supplementary material.**

Introduction

The following picture shows the GeneBase 1.1 Status Toolbar:

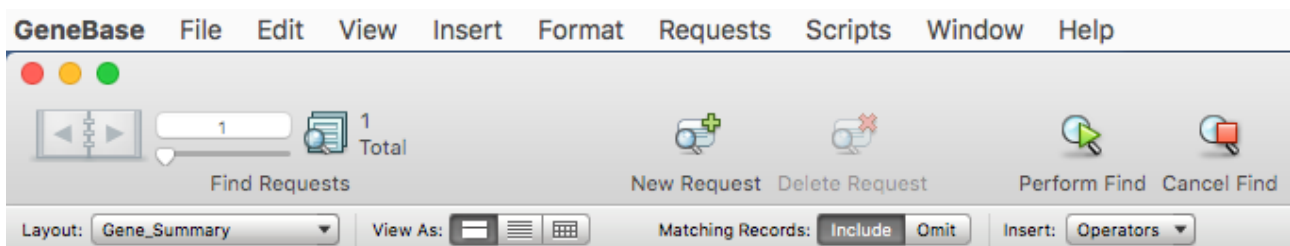
"**Layout:**" displays the name of the current table.

By clicking on the "**Sort**" button, a window will allow the user to choose the sorting criteria:

from the the left square select the desired table field available for sorting and click on the "Move" button. Clicking on the "Clear" button will delete previously chosen fields. Clicking on the "Sort" button will perform the sorting.

By clicking on the "**Find**" button (the magnifying glass) of Status Toolbar shown above, you will open the "Find Mode" of the current table useful to perform searches in GeneBase 1.1 (the "Find Mode" can also be chosen from the "View" menu).

The following picture shows the Status Toolbar seen in the "Find mode":

The most useful buttons (and cited throughout the text) are the "**New Request**" button, which creates a new find request, giving the possibility to combine more requests at a time; the "**Omit**" button which allows finding records except those matching specified criteria; the "**Perform Find**" button will start the search.

1. Gene types (Figures 1 A and B)

In order to retrieve the GeneBase 1.1 gene type composition as annotated in NCBI Gene from the Gene_Summary table, perform the search choosing in the Gene_Type field the keywords from the drop-down list: protein-coding, pseudo (pseudogenes), ncRNA (non-coding RNA), snoRNA (small nucleolar RNA), snRNA (small nuclear RNA), rRNA (ribosomal RNA), tRNA (transfer RNA), other and unknown.

Select "Gene_Summary" from the Layout drop-down menu or click on the violet button "Gene_Summary";

Click on the "Find" button;

      Gene_Type: Select "protein-coding" from the drop-down list;

Click on the "Perform Find" button.

Then perform the same steps selecting the other keywords in the Gene_Type field (one different for each research).

In order to retrieve the GeneBase 1.1 gene type composition for the subset of "REVIEWED" and "VALIDATED" genes with at least one "REVIEWED" or "VALIDATED" transcript excluding genes not in current annotation release, perform the same search also choosing in the RNA_RefSeq_Status field the keywords from the drop-down list and exclude records not in current annotation release.

Select "Gene_Summary" from the Layout drop-down menu or click on the violet button "Gene_Summary";

Click on the "Find" button;

      Gene_Type: Select "protein-coding" from the drop-down list;

      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

Click on the "New Request" button;

      Gene_Type: Select "protein-coding" from the drop-down list;

      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

Click on the "New Request" button;

      Gene_Type: Select "protein-coding" from the drop-down list;

      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

Click on the "New Request" button;

      Gene_Type: Select "protein-coding" from the drop-down list;

      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

Click on the "New Request" button and then on the "Omit" button;

      Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

This search will find protein-coding genes with "REVIEWED" or "VALIDATED" status with at least one "REVIEWED" or "VALIDATED" transcript. Then perform the same steps selecting the other keywords in the Gene_Type field (one different gene type for each search).

2. RefSeq status (Supplementary Table S2)

In order to retrieve the number of records with a specific RefSeq status assigned to genes as annotated in NCBI Gene from the Gene_Summary table, perform the search choosing in the Gene_RefSeq_Status field the keywords from the drop-down list: "REVIEWED" or "VALIDATED" or "PROVISIONAL" or "INFERRED" or "PREDICTED". Perform the search typing the symbol = in the Gene_RefSeq_Status field in order to retrieve records with this record empty.

Select "Gene_Summary" from the Layout drop-down menu or click on the violet button "Gene_Summary";

Click on the "Find" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

Click on the "New Request" button and then on the "Omit" button;

Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

This search will find genes with the "REVIEWED" status with at least one "REVIEWED" transcript. Then perform the same steps selecting and combining the other keywords in the Gene_RefSeq_Status and in the RNA_RefSeq_Status fields.

In order to retrieve the number of records with a specific RefSeq status assigned at transcripts and genes, from the Transcripts table, perform the search choosing in the Gene_RefSeq_Status and in the RNA_RefSeq_Status fields the combination of keywords from the drop-down lists: "REVIEWED" or "VALIDATED" or "PROVISIONAL" or "INFERRED" or "PREDICTED". Perform the search typing the symbol = in order to retrieve records with this record empty. In the Genome_Annotation_Status field choose not in current annotation release from the drop-down list in order to retrieve records with this characteristic.

Select "Transcripts" from the Layout drop-down menu or click on the violet button "Transcripts";

Click on the "Find" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

Click on the "New Request" button and then on the "Omit" button;

Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

This search will find transcripts with "REVIEWED" status whose corresponding genes have a "REVIEWED" RefSeq status. Then perform the same steps selecting and combining the other keywords in the Gene_RefSeq_Status and in the RNA_RefSeq_Status fields.

3. Number of genes per chromosome (Table 1 and Supplementary Table S3)

In order to retrieve the gene number for each chromosome, from the Gene_Summary table perform the search choosing numbers from 1 to 22 for autosomes or the letters X or Y for sex chromosome in the Chromosome field. In particular, for numbers 1 and 2, type in the Chromosome field "==1" and "==2" (in order to exclude 11, 12, 13, and so forth from the retrieved records). In order to retrieve the gene number per chromosome for the subset of "REVIEWED" and "VALIDATED" transcripts, perform the same search also choosing in the RNA_RefSeq_Status field the keywords from the drop-down list and exclude records not in current annotation release (as detailed below).

Select "Gene_Summary" from the Layout drop-down menu or click on the violet button "Gene_Summary";

Click on the "Find" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

Chromosome: type (without quotes): "=21" (for chromosome 21);

Click on the "New Request" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

Chromosome: type (without quotes): "=21" (for chromosome 21);

Click on the "New Request" button;

       Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       Chromosome: type (without quotes): "=21" (for chromosome 21);

Click on the "New Request" button;

       Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       Chromosome: type (without quotes): "=21" (for chromosome 21);

Click on the "New Request" button and then on the "Omit" button;

       Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

This search will find genes on chromosome 21 with "REVIEWED" or "VALIDATED" status with at least one "REVIEWED" or "VALIDATED" transcript. Then perform the same steps typing the other autosome numbers or the letters X or Y for sex chromosome in Chromosome field (one different for each search).

4. Number of genes per chromosome and gene type (Supplementary Table S3)

In order to retrieve the gene type composition for each chromosome, combine search numbers 1 and 3.

Select "Gene_Summary" from the Layout drop-down menu or click on the violet button "Gene_Summary";

Click on the "Find" button;

       Gene_Type: Select "protein-coding" from the drop-down list;

       Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       Chromosome: type (without quotes): "=21" (for chromosome 21);

Click on the "New Request" button;

       Gene_Type: Select "protein-coding" from the drop-down list;

       Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       Chromosome: type (without quotes): "=21" (for chromosome 21);

Click on the "New Request" button;

       Gene_Type: Select "protein-coding" from the drop-down list;

       Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       Chromosome: type (without quotes): "=21" (for chromosome 21);

Click on the "New Request" button;

       Gene_Type: Select "protein-coding" from the drop-down list;

       Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       Chromosome: type (without quotes): "=21" (for chromosome 21);

Click on the "New Request" button and then on the "Omit" button;

       Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

This search will find protein-coding genes on chromosome 21 with "REVIEWED" or "VALIDATED" status with at least one "REVIEWED" or "VALIDATED" transcript. Then perform the same steps typing the other autosome numbers or the letters X or Y for sex chromosome in the Chromosome field (one different for each search) and selecting the other keywords in the Gene_Type field (one different for each search).

5. Human gene statistics (Table 2)

Total, median, mean, standard deviation, minimum and maximum <u>number of genes</u> are derived from Supplementary Table S3. In order to retrieve median, mean, standard deviation and total <u>lengths for protein-coding genes</u> from Genes table, choose protein-coding keyword from the drop-down list of Gene_Type field and combine more requests choosing also "REVIEWED" and "VALIDATED" keywords from the drop-down list of Gene_RefSeq_Status and RNA_RefSeq_Status fields, excluding genes not in current annotation release as detailed below.

Select "Genes" from the Layout drop-down menu or click on the violet button "Genes";
Click on the "Find" button;

  Gene_Type: Select "protein-coding" from the drop-down list;

  Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

  RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

Click on the "New Request" button;

  Gene_Type: Select "protein-coding" from the drop-down list;

  Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

  RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

Click on the "New Request" button;

  Gene_Type: Select "protein-coding" from the drop-down list;

  Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

  RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

Click on the "New Request" button;

  Gene_Type: Select "protein-coding" from the drop-down list;

  Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

  RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

Click on the "New Request" button and then on the "Omit" button;

  Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

In order to retrieve minimum and maximum lengths, sort the retrieved records by Gene_Length_bp field, by clicking on the "Sort" button as shown in the Introduction section. The first and the last records will be the minimum and the maximum gene lengths, respectively.

This search is also useful in order to retrieve median, mean, standard deviation, minimum and maximum <u>number of transcripts</u> per gene, sorting the same retrieved records by Transcripts_per_Gene field (useful for search number 6) by clicking on the "Sort" button as shown in the Introduction section. The first and the last records will be the minimum and the maximum number of transcripts per gene, respectively.

In order to retrieve median, mean, standard deviation and total <u>lengths for non-coding genes</u> from Genes table, combine more requests choosing also "REVIEWED" and "VALIDATED" keyword from the drop-down list of Gene_RefSeq_Status and RNA_RefSeq_Status fields, excluding protein-coding keyword from

the drop-down list of Gene_Type field and excluding genes not in current annotation release as detailed below. In order to retrieve minimum and maximum lengths, sort the retrieved records by Gene_Length_bp field, by clicking on the "Sort" button as shown in the Introduction section. The first and the last records will be the minimum and the maximum gene lengths, respectively. This search is also useful in order to retrieve median, mean, standard deviation, minimum and maximum number of transcripts per gene, sorting the same retrieved records by Transcripts_per_Gene field (useful for search number 7) by clicking on the "Sort" button as shown in the Introduction section. The first and the last records will be the minimum and the maximum number of transcripts per gene, respectively.

Select "Genes" from the Layout drop-down menu or click on the violet button "Genes";
Click on the "Find" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;
Click on the "New Request" button and then on the "Omit" button;
      Gene_Type: Select "protein-coding" from the drop-down list;
Click on the "New Request" button and then on the "Omit" button;
      Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;
Click on the "Perform Find" button.

6. Protein-coding transcript statistics (Table 3)

Search number 5 (length of protein-coding genes) is also useful in order to retrieve median, mean, standard deviation, minimum and maximum number of transcripts per gene, sorting the same retrieved records by Transcripts_per_Gene field by clicking on the "Sort" button as shown in the Introduction section. The first and the last records will be the minimum and the maximum number of transcripts per gene, respectively.

In order to retrieve median, mean, standard deviation and total lengths for transcripts belonging to protein-coding genes from the Transcripts table, type NM_ in the RNA_Accession field (thus restricting the search to protein-coding transcripts) and combine more requests choosing also "REVIEWED" and "VALIDATED" keywords from the drop-down list of Gene_RefSeq_Status and RNA_RefSeq_Status fields, excluding genes not in current annotation release as detailed below.
In order to retrieve minimum and maximum lengths, sort the retrieved records by Transcript_Length_bp field by clicking on the "Sort" button as shown in the Introduction section. The first and the last records will be the minimum and the maximum transcript lengths, respectively. This search is also useful in order to

retrieve median, mean, standard deviation, minimum and maximum <u>number of exons per transcript</u> (and the number of introns per transcript subtracting one), sorting the same retrieved records by Exons_per_RNA by clicking on the "Sort" button as shown in the Introduction section. The first and the last records will be the minimum and the maximum number of exons per RNA, respectively.

While the "NM" or "NR" prefix used to label protein-coding or non-coding transcript repectively is implicit when searching for RefSeq annotated records, we use queries explicitly searching for transcripts with the "NM" or "NR" type of RefSeq RNA accession number which are useful to select for protein-coding or non-coding transcripts, respectively.

Select "Transcripts" from the Layout drop-down menu or click on the violet button "Transcripts";
Click on the "Find" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";
Click on the "New Request" button and then on the "Omit" button;
      Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;
Click on the "Perform Find" button.

The number of transcripts with <u>one exon</u> is given performing this search and typing "==1" in the Exons_per_RNA field. Sorting the retrieved records by Gene_ID and exporting this field from the sorted records will give the number of genes with at least one transcript with one exon. The same search typing "==2" in Exons_per_RNA field will retrieve transcripts with <u>one intron</u>.

Select "Transcripts" from the Layout drop-down menu or click on the violet button "Transcripts";
Click on the "Find" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";
      Exons_per_RNA: type (without quotes) "==1";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Exons_per_RNA: type (without quotes) "==1";

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Exons_per_RNA: type (without quotes) "==1";

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Exons_per_RNA: type (without quotes) "==1";

Click on the "New Request" button and then on the "Omit" button;

Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

This search will find intronless protein-coding transcripts with "REVIEWED" or "VALIDATED" status, with an "NM_" RNA accession number type, whose corresponding genes have a "REVIEWED" or "VALIDATED" RefSeq status. Then perform the same steps typing without quotes "==2" in the Exons_per_RNA field in order to find transcripts with one intron.

In order to retrieve the total number of exons belonging to protein-coding genes from the Gene_Table table, type NM_ in the RNA_Accession field and combine more requests choosing also "REVIEWED" and "VALIDATED" keywords from the drop-down list of Gene_RefSeq_Status (at the bottom-right corner of the window) and RNA_RefSeq_Status fields, excluding genes not in current annotation release as detailed below and excluding records with the Exon_Start field empty. Clicking on the violet Reports button will show the numbers related to median, mean, standard deviation and total lengths for exons for the subset of records retrieved with the search performed from the Gene_Table table. In order to exclude last exons, perform the search adding a "New Request" and then choosing the "Omit" button and choosing "Yes" from the drop-down list of Last_Exon field.

Select "Gene_Table" from the Layout drop-down menu or click on the violet button "Gene_Table";

Click on the "Find" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Click on the "New Request" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Click on the "New Request" button and then on the "Omit" button;

Exon_Start: type (without quotes) "=";

Click on the "New Request" button and then on the "Omit" button;

Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

Perform the same steps selecting "Yes" in the Last_Exons field in order to find exons belonging to protein-coding "REVIEWED" or "VALIDATED" transcripts with an "NM_" RNA accession number type, excluding last exons.

In order to retrieve all these numbers for the subset of non-redundant exons, repeat the same search choosing also "Yes" from the drop-down list of the Non_Redundant_Exon field. In order to exclude last exons, perform the search adding a "New Request" and then choosing the "Omit" button and choosing "Yes" from the drop-down list of Last_Exon field.

Select "Gene_Table" from the Layout drop-down menu or click on the violet button "Gene_Table";

Click on the "Find" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Non_Redundant_Exon: Select "Yes" from the drop-down list;

Click on the "New Request" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Non_Redundant_Exon: Select "Yes" from the drop-down list;

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Non_Redundant_Exon: Select "Yes" from the drop-down list;

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Non_Redundant_Exon: Select "Yes" from the drop-down list;

Click on the "New Request" button and then on the "Omit" button;

Exon_Start: type (without quotes) "=";

Click on the "New Request" button and then on the "Omit" button;

Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

Perform the same steps selecting "Yes" in the Last_Exons field in order to find non-redundant exons belonging to protein-coding "REVIEWED" or "VALIDATED" transcripts with an "NM_" RNA accession number type, excluding last exons.

Performing the searches excluding records with the Downstream_Intron_Start field empty (instead of the Exon_Start field empty) will retrieve the total number of introns belonging to protein-coding genes.

The difference between the total number of exons and the total number of intron is, as expected, the sum of the monoexonic genes and of the last exons which both are not followed by an intron.

Select "Gene_Table" from the Layout drop-down menu or click on the violet button "Gene_Table";
Click on the "Find" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";
Click on the "New Request" button and then on the "Omit" button;
      Downstream_Intron_Start: type (without quotes) "=";
Click on the "New Request" button and then on the "Omit" button;
      Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;
Click on the "Perform Find" button.

Performing the searches excluding records with the Coding_Exon_Start field empty (instead of the Exon_Start field empty) will retrieve the total number of coding exons belonging to protein-coding genes.

Select "Gene_Table" from the Layout drop-down menu or click on the violet button "Gene_Table";
Click on the "Find" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_Accession: type (without quotes) "NM_";

Click on the "New Request" button and then on the "Omit" button;

       Coding_Exon_Start: type (without quotes) "=";

Click on the "New Request" button and then on the "Omit" button;

       Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

7. Non-coding transcript statistics (Table 4)

Search number 5 (length of non-coding genes) is also useful in order to retrieve median, mean, standard deviation, minimum and maximum <u>number of transcripts</u> per gene, sorting the same retrieved records by Transcripts_per_Gene field by clicking on the "Sort" button as shown in the Introduction section. The first and the last records will be the minimum and the maximum number of transcripts per gene, respectively.

In order to retrieve median, mean, standard deviation and total <u>lengths for transcripts</u> belonging to non-coding genes from the Transcripts table, type NR_ in the RNA_Accession field (thus restricting the search to non-coding transcripts) and combine more requests choosing also "REVIEWED" and "VALIDATED" keywords from the drop-down lists of the Gene_RefSeq_Status and RNA_RefSeq_Status fields, excluding genes not in current annotation release as detailed below. In order to retrieve minimum and maximum lengths, sort the retrieved records using the Transcript_Length field by clicking on the "Sort" button as shown in the Introduction section. The first and the last records will be the minimum and the maximum transcript lengths, respectively.

 This search is also useful in order to retrieve median, mean, standard deviation, minimum and maximum <u>number of exons per transcript</u> (and the number of introns per transcript subtracting one), sorting the same retrieved records by Exons_per_RNA by clicking on the "Sort" button as shown in the Introduction section. The first and the last records will be the minimum and the maximum number of exons per RNA, respectively.

While the "NM" or "NR" prefix used to label protein-coding or non-coding transcript repectively is implicit when searching for RefSeq annotated records, we use queries explicitly searching for transcripts with the "NM" or "NR" type of RefSeq RNA accession number which are useful to select for protein-coding or non-coding transcripts, respectively.

Select "Transcripts" from the Layout drop-down menu or click on the violet button "Transcripts";

Click on the "Find" button;

       Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       RNA_Accession: type (without quotes) "NR_";

Click on the "New Request" button;

       Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       RNA_Accession: type (without quotes) "NR_";

Click on the "New Request" button;

       Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NR_";

Click on the "New Request" button;

      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

      RNA_Accession: type (without quotes) "NR_";

Click on the "New Request" button and then on the "Omit" button;

      Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

The number of transcripts with <u>one exon</u> is given performing this search and typing "==1" in Exons_per_RNA field. Sorting the retrieved records by Gene_ID and exporting this field from the sorted records will give the number of genes with at least one transcript with one exon. The same search typing "==2" in Exons_per_RNA field will retrieve transcripts with <u>one intron</u>.

Select "Transcripts" from the Layout drop-down menu or click on the violet button "Transcripts";

Click on the "Find" button;

      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

      RNA_Accession: type (without quotes) "NR_";

      Exons_per_RNA: type (without quotes) "==1";

Click on the "New Request" button;

      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

      RNA_Accession: type (without quotes) "NR_";

      Exons_per_RNA: type (without quotes) "==1";

Click on the "New Request" button;

      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

      RNA_Accession: type (without quotes) "NR_";

      Exons_per_RNA: type (without quotes) "==1";

Click on the "New Request" button;

      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

      RNA_Accession: type (without quotes) "NR_";

      Exons_per_RNA: type (without quotes) "==1";

Click on the "New Request" button and then on the "Omit" button;

      Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

This search will find intronless non-coding transcripts with "REVIEWED" or "VALIDATED" status, with an "NR_" RNA accession number type, whose corresponding genes have a "REVIEWED" or "VALIDATED" RefSeq status. Then perform the same steps typing without quotes "==2" in the Exons_per_RNA field in order to find transcripts with one intron.

In order to retrieve the total <u>number of exons</u> belonging to non-coding genes from the Gene_Table table, type NR_ in the RNA_Accession field and combine more requests choosing also "REVIEWED" and "VALIDATED" keywords from the drop-down list of Gene_RefSeq_Status and RNA_RefSeq_Status fields,

excluding genes not in current annotation release as detailed below and excluding records with the Exon_Start field empty. Clicking on the violet Reports button will show the numbers related to median, mean, standard deviation and total <u>lengths for exons</u> for the subset of records retrieved with the search performed from the Gene_Table table. In order to exclude last exons, perform the search adding a "New Request" and then choosing the "Omit" button and choosing "Yes" from the drop-down list of Last_Exon field.

Select "Gene_Table" from the Layout drop-down menu or click on the violet button "Gene_Table";
Click on the "Find" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_Accession: type (without quotes) "NR_";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_Accession: type (without quotes) "NR_";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_Accession: type (without quotes) "NR_";
Click on the "New Request" button;
      Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;
      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_Accession: type (without quotes) "NR_";
Click on the "New Request" button and then on the "Omit" button;
      Exon_Start: type (without quotes) "=";
Click on the "New Request" button and then on the "Omit" button;
      Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;
Click on the "Perform Find" button.
Perform the same steps selecting "Yes" in the Last_Exons field in order to find exons belonging to non-coding "REVIEWED" or "VALIDATED" transcripts with an "NR_" RNA accession number type, excluding last exons.

In order to retrieve all these numbers for the subset of non-redundant exons, repeat the same search also choosing Yes from the drop-down list of Non_Redundant_Exon field. In order to exclude last exons, perform the search adding a "New Request" and then choosing the "Omit" button and choosing "Yes" from the drop-down list of Last_Exon field.

Select "Gene_Table" from the Layout drop-down menu or click on the violet button "Gene_Table";
Click on the "Find" button;
      Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;
      RNA_Accession: type (without quotes) "NR_";
      Non_Redundant_Exon: Select "Yes" from the drop-down list;
Click on the "New Request" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NR_";

Non_Redundant_Exon: Select "Yes" from the drop-down list;

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NR_";

Non_Redundant_Exon: Select "Yes" from the drop-down list;

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_Accession: type (without quotes) "NR_";

Non_Redundant_Exon: Select "Yes" from the drop-down list;

Click on the "New Request" button and then on the "Omit" button;

Exon_Start: type (without quotes) "=";

Click on the "New Request" button and then on the "Omit" button;

Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

Perform the same steps selecting "Yes" in the Last_Exons field in order to find non-redundant exons belonging to non-coding "REVIEWED" or "VALIDATED" transcripts with an "NR_" RNA accession number type, excluding last exons.

Performing the searches excluding records with the Downstream_Intron_Start field empty (instead of empty Exon_Start field) will retrieve the total number of introns belonging to non-coding genes.

The difference between the total number of exons and the total number of intron is, as expected, the sum of the monoexonic genes and of the last exons which both are not followed by an intron.

Select "Gene_Table" from the Layout drop-down menu or click on the violet button "Gene_Table";

Click on the "Find" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_Accession: type (without quotes) "NR_";

Click on the "New Request" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NR_";

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NR_";

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_Accession: type (without quotes) "NR_";

Click on the "New Request" button and then on the "Omit" button;

Downstream_Intron_Start: type (without quotes) "=";

Click on the "New Request" button and then on the "Omit" button;

       Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

8. Protein-coding and non-coding transcript comprehensive analysis (Supplementary Table S4)

The searches detailed below can be performed without selecting protein-coding in the Gene_Type field and without typing NM_ or NR_ in the RNA_Accession field thus giving a comprehensive analysis of all transcripts present in GeneBase 1.1, both protein-coding and non-coding.

For example:

Select "Transcripts" from the Layout drop-down menu or click on the violet button "Transcripts";

Click on the "Find" button;

       Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       Exons_per_RNA: type (without quotes) "==1";

Click on the "New Request" button;

       Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       Exons_per_RNA: type (without quotes) "==1";

Click on the "New Request" button;

       Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       Exons_per_RNA: type (without quotes) "==1";

Click on the "New Request" button;

       Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       Exons_per_RNA: type (without quotes) "==1";

Click on the "New Request" button and then on the "Omit" button;

       Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

This search will find intronless transcripts with "REVIEWED" or "VALIDATED" status whose corresponding genes have a "REVIEWED" or "VALIDATED" RefSeq status. Then perform the same steps typing without quotes "==2" in the Exons_per_RNA field in order to find transcripts with one intron.

Another example:

Select "Gene_Table" from the Layout drop-down menu or click on the violet button "Gene_Table";

Click on the "Find" button;

       Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       Non_Redundant_Exon: Select "Yes" from the drop-down list;

Click on the "New Request" button;

       Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

       RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

       Non_Redundant_Exon: Select "Yes" from the drop-down list;

Click on the "New Request" button;

       Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

Non_Redundant_Exon: Select "Yes" from the drop-down list;

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

Non_Redundant_Exon: Select "Yes" from the drop-down list;

Click on the "New Request" button and then on the "Omit" button;

Exon_Start: type (without quotes) "=";

Click on the "New Request" button and then on the "Omit" button;

Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

This search will find non-redundant exons belonging to transcripts with "REVIEWED" or "VALIDATED" status whose corresponding genes have a "REVIEWED" or "VALIDATED" RefSeq status. Perform the same steps selecting "Yes" in the Last_Exons field in order to find non-redundant exons belonging to non-coding "REVIEWED" or "VALIDATED" transcripts with an "NR_" RNA accession number type, excluding last exons.

9. Protein-coding transcript regions (Table 5)

Search number 6 performed in order to retrieve protein-coding lengths is also useful for length information about 5´ and 3´ UTRs and CDS regions (untranslated and coding DNA sequence regions, respectively). Minimum and maximum lengths can be found sorting the retrieved records by UTR5'_Length_bp, UTR3'_Length_bp and CDS_Length fields, respectively.

Select "Transcripts" from the Layout drop-down menu or click on the violet button "Transcripts";

Click on the "Find" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Click on the "New Request" button;

Gene_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Click on the "New Request" button;

Gene_RefSeq_Status: Select "VALIDATED" from the drop-down list;

RNA_RefSeq_Status: Select "REVIEWED" from the drop-down list;

RNA_Accession: type (without quotes) "NM_";

Click on the "New Request" button and then on the "Omit" button;

Genome_Annotation_Status: Select "not in current annotation release" from the drop-down list;

Click on the "Perform Find" button.

10. Non-coding RNAs (Supplementary Table S6)

In order to retrieve the main non-coding RNA species (on the left column) from Gene_Summary table type the terms specified in the column on the right (field: words) excluding records not in current annotation release (Click on the "New Request" button and then on the "Omit" button, select "not in current annotation release" from the drop-down list in the Genome_Annotation_Status field).

RNA species	Field: words to be typed in
rRNA 28S	Official_Full_Name: RNA, 28S
rRNA 18S	Official_Full_Name: RNA, 18S
rRNA 5.8S	Official_Full_Name: RNA, 5.8S
rRNA 5S	Official_Full_Name: RNA, 5S, Gene_Type: rRNA
tRNA	Gene_Type: tRNA
snRNA U1	Gene_Type: snRNA, Gene_Description: "RNA, U1 small nuclear"
snRNA U2	Gene_Type: snRNA, Gene_Description: "RNA, U2 small nuclear"
snRNA U6	Gene_Type: snRNA, Gene_Description: "RNA, U6 small nuclear"
snRNA 7SK	Gene_Type: snRNA, Gene_Symbol: RN7SK
snoRNA C/D box	Gene_Type: snoRNA, Gene_Description: C/D box
snoRNA H/ACA box	Gene_Type: snoRNA, Gene_Description: H/ACA box
TERC	Gene_Symbol: =TERC
HBII-5	Gene_Symbol: =SNORD115-1
XIST	Gene_Symbol: =XIST
antisense RNAs	Gene_Description: antisense, excluding Gene_Type: protein-coding
TSIX	Gene_Symbol: =TSIX
NUDT6	Gene_Symbol: =NUDT6
7SL	Gene_Symbol: =RN7SL1
microRNA	Gene_Type: ncRNA, Gene_Description: microRNA