cubap Documentation

Release latest

Jan 31, 2023

Contents:

CHAPTER 1

Codon Frequency

1.1 Graphs:

Average Codon Frequency For the selected gene(s)/isoform(s), the average number of times each codon occurs across all samples.

Standard Deviation of Frequency For the selected gene(s)/isoform(s), the standard deviation of the number of times each codon occurs across all samples.

Average Codon Frequency by Superpopulation This violin plot shows the frequency of a certain codon, in the selected gene(s)/isoform(s), across all populations. Thicker sections represent a greater number of samples that contain that number of codon occurances.

Average Frequency by Population This graph shows the same information as the violin plot except individual subpopulation averages are shown. Each point represents the average frequency of the selected codon for all samples in that subpopulation.

1.2 Options:

Select Gene Use the search and dropdown box features to find your gene(s) of interest. Click on the gene name to query it. You can query multiple genes by holding either the 'CTRL' or 'command' key. Next to the genes are the isoform numbers; the longest ones are marked. You must next select an isoform for every gene you've selected in order to view the data in the graphs.

View Options If you have queried multiple genes or isoforms, you can view the data in two different ways. You can select the 'Compare Genes' button in order to view codon frequencies of different genes side by side. You can also select the "Average of Genes" button in order to view the average codon frequencies across all selected genes/isoforms. By default, the 'Average of Genes' view option is selected. These view options only affect the *Average Codon Frequency* and *Standard Deviation of Frequency* graphs.

Select Population Choose a super- or subpopulation to view only the codon frequency data from those populations. This will filter every graph.

Select Codon Choose a codon of interest to view how the frequency of this codon, in the selected gene(s)/isoform(s), across all populations. This only alters the *Average Codon Frequency by Superpopulation* and *Average Frequency by Population* graphs.

CHAPTER 2

Identical Codon Pairing & Co-tRNA Codon Pairing

These visuals are set up identically to the *Codon Frequency* visual except that instead of showing the frequency of each codon, it shows the frequency of each codon pair. For co-tRNA codon pairing, synonymous (but not identical!) pairs are shown by their common amino acid.

CHAPTER 3

Codon Aversion

3.1 Graphs:

Total Codon Aversion Across all Genes by Superpopulation This graph shows how often each codon is averted. Specifically, it is the total number of alleles, summed for each gene, in which the codon is not present.

Total Number of Alleles per Superpopulation with Codon Across the x-axis the codon aversion motif (all the codons that are missing in the selected gene(s)/isform(s)) and the number of alleles, per superpopulation, that are missing that codon.

3.2 Options:

Select Gene Use the search and dropdown box features to find your gene(s) of interest. Click on the gene name to query it. You can query multiple genes by holding either the 'CTRL' or 'command' key. Next to the genes are the isoform numbers; the longest ones are marked. You must next select an isoform for every gene you've selected in order to view the data in the graphs.

Compare Subpopulations Click this button to view the number of alleles that are missing codons by subpopulation instead of superpopulation. This only affects the *Total Number of Alleles per Superpopulation with Codon* graph. Click the *Reset* button to view superpopulation data again.

Select Subpopulations This filter only applies to the *Total Number of Alleles per Subpopulation with Codon* graph. It allows you to ivew the number of alleles of only certain subpopulations.

CHAPTER 4

Ramp Sequences

4.1 Graphs:

Ramp Harmonic Mean RSCU by Subpopulation For all selected gene(s)/isoform(s), the harmonic mean of all RSCU values for each codon in the ramp sequence. This is plotted in a box and whiskers plot by subpopulation.

Ramp Harmonic Mean RSCU by Superpopulation For all selected gene(s)/isoform(s), the harmonic mean of all RSCU values for each codon in the ramp sequence. This is plotted in a box and whiskers plot by superpopulation.

Gene Harmonic Mean RSCU by Subpopulation For all selected gene(s)/isoform(s), the harmonic mean of all RSCU values for each codon in the entire gene sequence. This is plotted in a box and whiskers plot by subpopulation.

Gene Harmonic Mean RSCU by Superpopulation For all selected gene(s)/isoform(s), the harmonic mean of all RSCU values for each codon in the entire gene sequence. This is plotted in a box and whiskers plot by superpopulation.

Percent Samples With Ramp A pie chart that shows what percentage of individuals in the selected superpopulation(s) or subpopulation(s) have a ramp sequence in the selected gene. Only populations that have at least one individual with a ramp sequence are shown.

4.2 Averages:

Ramp Sequence

- RSCU: the average RSCU value for all codons in the ramp sequence, from all populations
- Length: the average length of the ramp sequence, in number of codons, from all populations

Entire Gene

- RSCU: the average RSCU value for all codons in the entire gene sequence, from all populations
- Length: the average length of the entire gene sequence, in number of codons, from all populations

If multiple genes/isoforms are selected, these values will also be the averages of all those.

4.3 Options:

Select Gene Use the search and dropdown box features to find your gene(s) of interest. Click on the gene name to query it. You can query multiple genes by holding either the 'CTRL' or 'command' key. Next to the genes are the isoform numbers; the longest ones are marked. You must next select an isoform for every gene you've selected in order to view the data in the graphs.

Show Superpopulation Switch the box plots to group RSCU values by superpopulation.

Show Subpopulation Switch the box plots to group RSCU values by subpopulation.

Pie Chart Show the frequency of individuals/samples in a population that have a ramp sequence in a pie chart.

Table Show the frequency of individuals/samples in a population that have a ramp sequence in a table. All populations are shown.

CHAPTER 5

Nucleotide Composition

5.1 Graphs:

Average Nulceptide Frequency The average number of occurances of each nucleotide in the selected gene(s)/isoform(s).

Standard Deviation of Nucleotide Frequency The standard deviation of the number of occurances of each nucleotide in the selected gene(s)/isoform(s).

GC Content Across Populations A violin plot of GC content compared across Superpopulations.

Nucleotide Frequency by Superpopulation The frequency of the selected nucleotide for each superpopulation.

Nucleotide Frequency by Superpopulation and Subpopulation The frequency of the selected nulceotide for each subpopulation, ordered by superpopulation. The *Average GC Content %* is also shown. This is computed from all selected genes/isoforms and from all populations.

5.2 Options:

Select Gene Use the search and dropdown box features to find your gene(s) of interest. Click on the gene name to query it. You can query multiple genes by holding either the 'CTRL' or 'command' key. Next to the genes are the isoform numbers; the longest ones are marked. You must next select an isoform for every gene you've selected in order to view the data in the graphs.

GC Content This is selected by default. It shows the *GC Content Across Populations* graph.

Nucleotide Frequency Click this button to view the *Nucleotide Frequency by Superpopulation* and *Nucleotide Frequency by Superpopulation and Subpopulation* graphs.

CHAPTER 6

Frequently Asked Questions

- *Why are some of the graphs blank?* Because these graphs compare population differences, but the selected gene(s)/isoform(s) only have one allele in the entire 1000 Genome Sample and thus do not differ at all between populations.
- *Why do some of the isoforms say longest?* These are the longest isoform for that particular gene. If several different isoforms of a gene all were all equally long, then the first isoform was indicated as the longest. If the gene only has one isoform then it was marked as the longest.
- *How can I upload my own data to CUBAP?* This feature is not available due to contraints on the free license of Power BI Desktop and how Power BI connects data to its visuals. To perform your own analyses, please see our github for scripts that will allow you to compute codon usage biases: https://github.com/kauwelab/cubap