meta-CATS

Metadata-driven Comparative Analysis Tool for Sequences (meta-CATS): an Automated Process for Identifying Significant Sequence Variations that Correlate with Virus Attributes

1. Purpose:

The meta-CATS metadata genome comparison tool takes sequence data and determines the aligned positions that significantly differ between two (or more) user-specified groups.

2. Method description:

The method is invoked in a web application. Once an analysis is started, a multiple sequence alignment is performed if the input was unaligned (such as from a database query). A chi-square test of independence is then performed on each non-conserved column of the alignment, to identify those that have a non-random distribution of bases. A quantitative statistical value of variation is computed for all positions. Columns that are perfectly conserved will not be identified as statistically significant. All other non-conserved columns will be evaluated to determine whether the p-value is lower than the specified threshold value. Terminal gaps flanking the aligned sequences will not be taken into account for the analysis.

For nucleotide sequences, the degrees of freedom used to calculate the statistical value is statically set at [(number of possible residues-1)*(number of groups being calculated-1)], where the "number of possible residues" is consistently 4, to account for the possibility of any nucleotide being incorporated at any position. Whether all possible nucleotide changes result in viable virus being produced during infection is another matter entirely.

For protein sequences, the degrees of freedom used to calculate the statistical value is dynamically set at [(number of observed residues at each aligned position-1)*(number of groups being calculated-1)]. Using the observed amino acid residues instead of all 20 possible amino acids accounts for the fact that not all amino acid substitutions are equally possible at any position, and the residues observed in the sample are thus more likely to represent those found in the population. If gaps are present in the aligned column, the "number of residues" is increased by 1, to account for any and all bases that may not be represented due to sample size and/or sampling bias.

Each column that is found to be significantly non-random is then subjected to a Pearson's chi-square calculation, to determine the pairs that contribute to skewing of the data. Logically, the second chi-square calculation is helpful if sequences are assigned to three or more groups since positions found as significantly varying between sequences assigned to only two groups will completely overlap with the results from the chi-square test of independence.

After all of the computations are completed, the results from the various statistical tests performed in this workflow are then summarized in tabular form within a webpage and are made available for download in comma-separated value (csv) format.

3. Input Data Preparation:

Sequence data can be input from: 1) a ViPR database query, 2) an existing ViPR working set of sequences, or by 3) uploading a personal file containing other sequences to the ViPR system, or by 4) a combination of these options.

4. Parameters:

Sequences of the selected strains, regardless of their source, are extracted from ViPR data warehouse and formatted appropriately. The default is set, such that if the resulting p-value from any column > 0.05, then that column will not be displayed in the results. The user specifies the number of groups (default = 2), as well as the sequences belonging to each group.

5. Output data post-processing and display:

The Calculated ("c") value reported by this tool is equal to the p-value calculated during the chi-square test of independence. The different name is necessary to acknowledge that several underlying statistical assumptions are not met for the chi-square test. These assumptions are not met for the majority of sequence analyses regardless of the statistical method and include: independence of positions (i.e. no covariance, codon bias, wobble positions, etc.), normal distribution of residues in each aligned position, and multiple hypothesis testing.

As C becomes <~ 0.01, its significance becomes nearly identical to the actual p-value. Although the aforementioned statistical assumptions are not met, the output from the Metadata Genome Compare tool is still biologically relevant and is useful in generating hypotheses that can be experimentally validated in the laboratory.

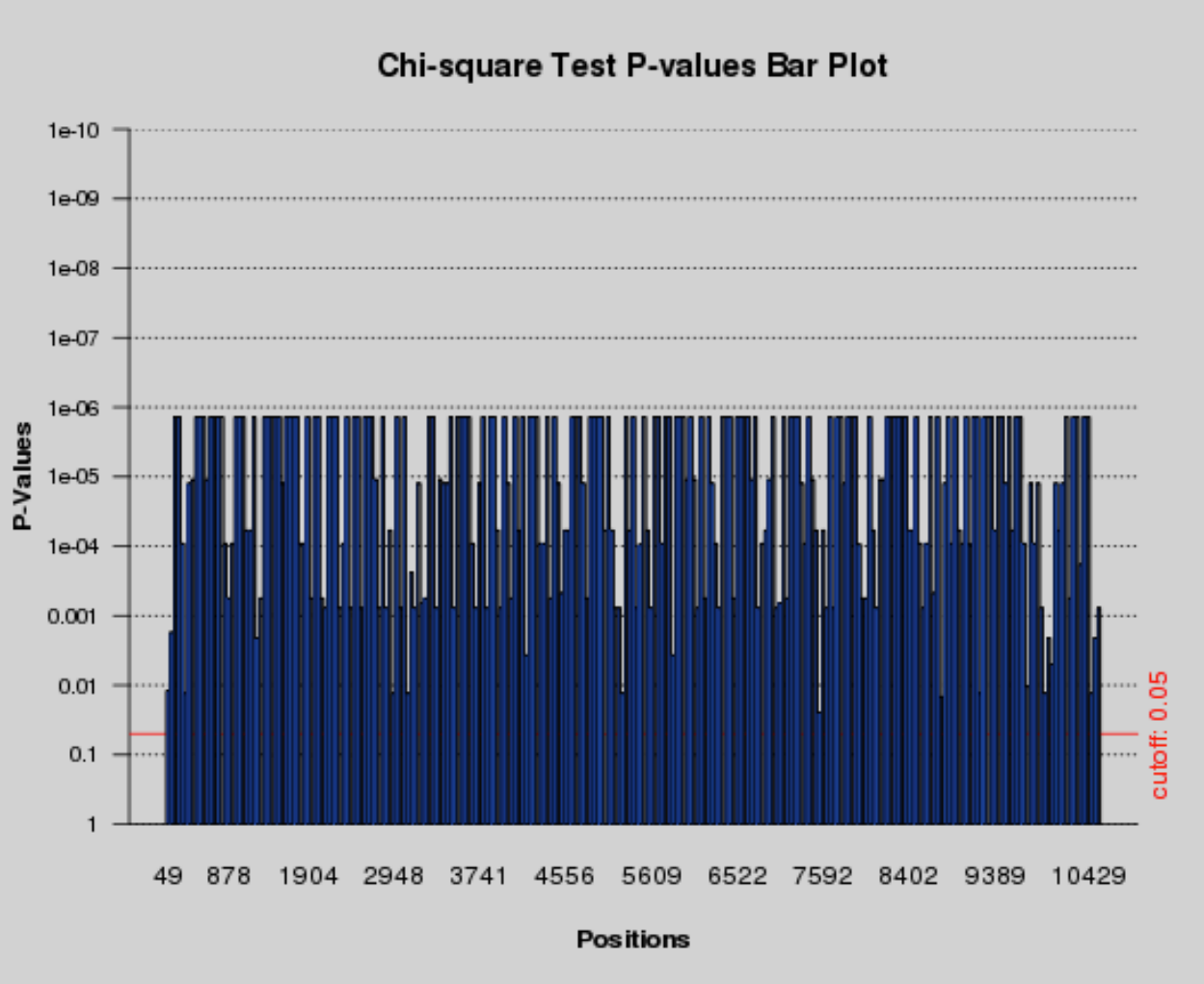
Results from the analysis are reported in tabular form, as in the example found below.

Chi-square Goodness of Fit Test Result

There are 261 positions that have a significant non-random distribution between the specified groups.

Position	Chi-square Value	C-value	Degree Freedom
49	12.855	0.01201	4
89	15.097	0.001736	3
197	29.991	1.386E-6	3
272	29.994	1.384E-6	3
308	21.292	9.156E-5	3
314	10.766	0.01306	3
326	25.449	1.244E-5	3
390	25.709	1.097E-5	3
410	29.994	1.384E-6	3
434	29.994	1.384E-6	3
467	29.994	1.384E-6	3
518	25.709	1.097E-5	3
578	29.994	1.384E-6	3
587	29.994	1.384E-6	3
698	29.994	1.384E-6	3
716	29.994	1.384E-6	3
873	21.3	9.122E-5	3
878	17.496	5.587E-4	3
884	21.3	9.122E-5	3
896	29.991	1.386E-6	3
920	29.991	1.386E-6	3
971	29.994	1.384E-6	3
1025	22.204	5.916E-5	3
1115	22.204	5.916E-5	3
1130	29.991	1.386E-6	3
1244	14.691	0.0021	3
1256	17.496	5.587E-4	3
1271	29.994	1.384E-6	3
1306	29.994	1.384E-6	3

This example analysis compared the polyprotein (amino acid) sequences of 100 HCV subtype 1a strains with 100 HCV subtype 1b strains. Results can also be displayed graphically by clicking on the "Show P-values Bar Plot" button at the top of the results page, to give the following plot.



6. References

Pickett BE, Liu M, Sadat EL, Squires RB, Noronha JM, He S, Jen W, Zaremba S, Gu Z, Zhou L, Larsen CN, Bosch I, Gehrke L, McGee M, Klem EB, and Scheuermann RH. (2013) Metadata-driven comparative analysis tool for sequences (meta-CATS): an automated process for identifying significant sequence variations that correlate with virus attributes. *Virology. Dec;447(1-2):45-51*. doi: 10.1016/j.virol.2013.08.021. PMID: 24210098; PMCID: PMC5040469.

The full text version of this publication is available at:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5040469/

The open source code repository is available at:

https://github.com/JCVenterInstitute/Meta-CATS