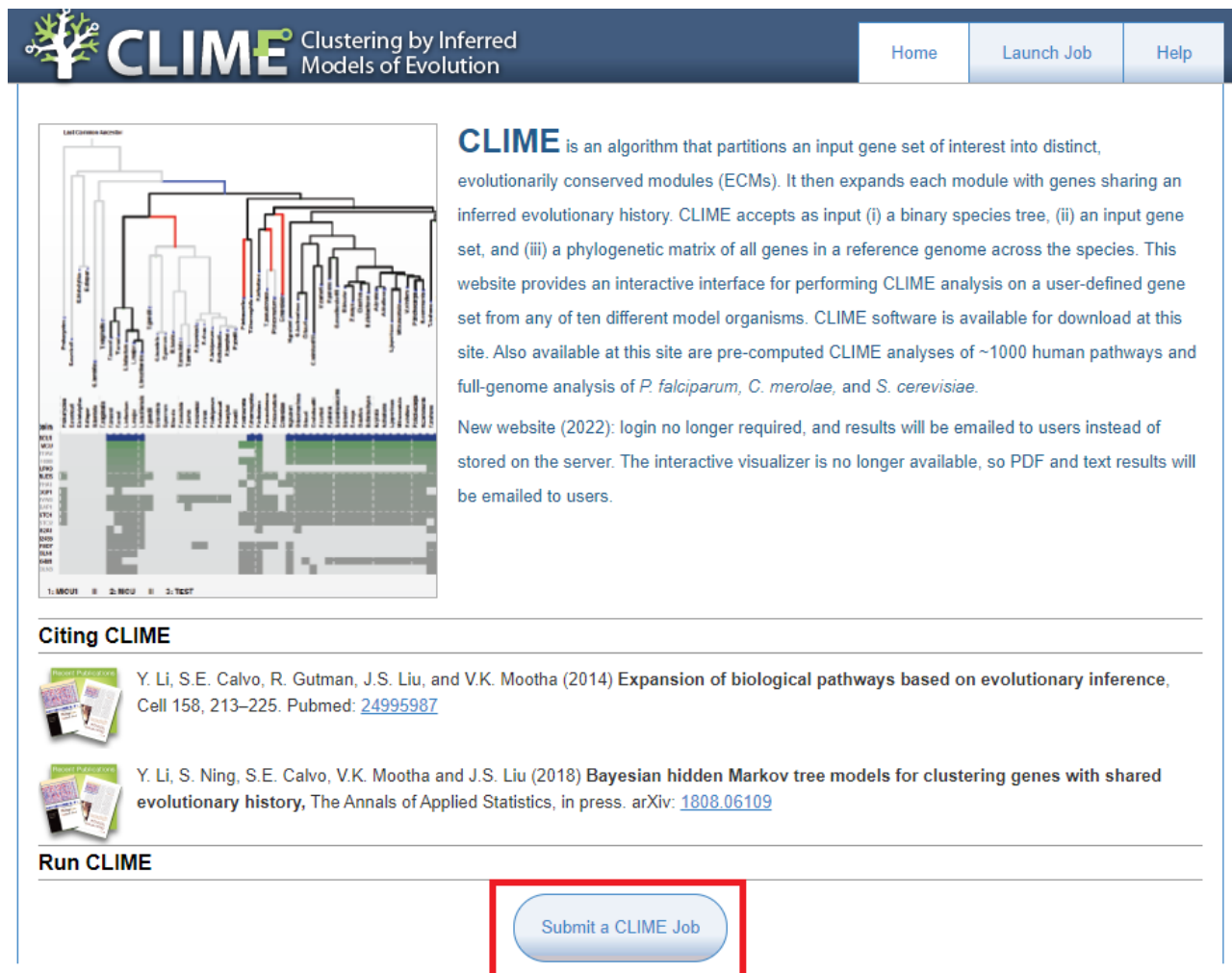


Tutorial of the CLIME analysis portal

CLIME (CLustering by Inferred Models of Evolution) is a statistical algorithm that partitions an input gene set of interest into distinct, evolutionarily conserved modules (ECMs), and then expands each module with genes sharing an inferred evolutionary history. This is a step-by-step tutorial that introduces the users how to use the online CLIME analysis portal. For demonstration, we apply CLIME to analysis of 7 genes in the human WASH complex as an example.

1. First we visit the CLIME portal in a web browser (www.gene-clime.org). The front page is shown below. It contains a brief description of CLIME algorithm and provides the links for submitting online CLIME jobs and downloading pre-computed CLIME outputs for about 1025 classically defined human pathways. To submit a CLIME job, we click the “Submit a CLIME job” button.




CLIME Clustering by Inferred Models of Evolution


Home Launch Job Help

CLIME is an algorithm that partitions an input gene set of interest into distinct, evolutionarily conserved modules (ECMs). It then expands each module with genes sharing an inferred evolutionary history. CLIME accepts as input (i) a binary species tree, (ii) an input gene set, and (iii) a phylogenetic matrix of all genes in a reference genome across the species. This website provides an interactive interface for performing CLIME analysis on a user-defined gene set from any of ten different model organisms. CLIME software is available for download at this site. Also available at this site are pre-computed CLIME analyses of ~1000 human pathways and full-genome analysis of *P. falciparum*, *C. merolae*, and *S. cerevisiae*.

New website (2022): login no longer required, and results will be emailed to users instead of stored on the server. The interactive visualizer is no longer available, so PDF and text results will be emailed to users.

Citing CLIME

 Y. Li, S.E. Calvo, R. Gutman, J.S. Liu, and V.K. Mootha (2014) **Expansion of biological pathways based on evolutionary inference**, Cell 158, 213–225. Pubmed: [24995987](https://pubmed.ncbi.nlm.nih.gov/24995987/)

 Y. Li, S. Ning, S.E. Calvo, V.K. Mootha and J.S. Liu (2018) **Bayesian hidden Markov tree models for clustering genes with shared evolutionary history**, The Annals of Applied Statistics, in press. arXiv: [1808.06109](https://arxiv.org/abs/1808.06109)


Run CLIME

Submit a CLIME Job

2. Results of the CLIME job will be emailed to the user. Here we enter:

- 1) Job name: used for future reference to this job. **Enter: WASH complex**
- 2) Organisms: use dropdown menu to select which of 10 precomputed phylogenetic profile matrices to input. **Select H. sapiens**
- 3) Gene list: list of input genes, with one gene symbol/Entrez ID per line. Entered names will be mapped to standard symbols using a static lookup table of gene aliases (NCBI Entrez ID, official gene symbol, other symbols listed in NCBI). Maximum 200 genes can be entered (users can analyze larger gene sets by downloading CLIME software available on the main web page).
Enter the gene symbols: CCDC53 WASH1 KIAA0196 23325 FAM21A FAM21B FAM21C
- 4) Notes: optional annotation to a CLIME job. **Leave blank**

Next click box "I'm not a robot" and then "Submit CLIME job".

Home Launch Job Help

Submit a new CLIME Job

Instructions

Please fill out the form below to submit your CLIME job for analysis. You may enter up to 200 genes for analysis. Runtimes of submitted jobs depend on the current load in our compute cluster. When the job is completed, a link with the results will be emailed to you and the files must be downloaded within 7 days. If you want to run CLIME with more than 200 genes or get your results more immediately, please download and use the command-line CLIME software package from the home page.

Job Parameters

Email Address

Job description (short identifier for this gene set)

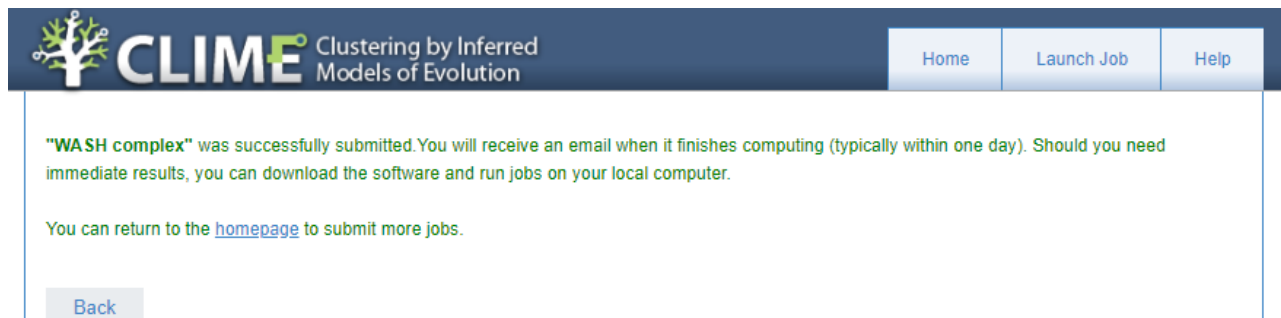
Organism

Predictions per ECM:

Gene List: One gene per line, Entrez gene IDs or symbols only

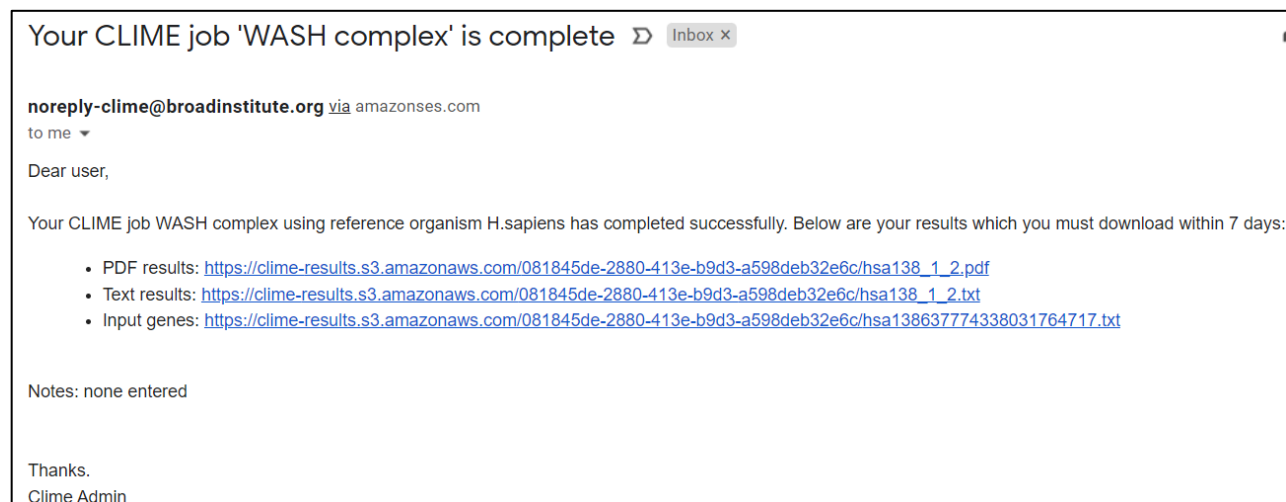
Gene List: One gene per line, Entrez gene IDs or symbols only	Notes
CCDC53 WASH1 KIAA0196 23325 FAM21A FAM21B FAM21C	

3. The CLIME portal will validate the input gene symbols/IDs. If all genes exist in the selected organism, a CLIME job will be submitted to the cloud compute farm, and the browser will be directed to the page shown below. An email will be received upon job completion – typically within an hour.



The screenshot shows the CLIME portal interface. The header includes the CLIME logo (a tree with a DNA helix) and the text "CLIME Clustering by Inferred Models of Evolution". Navigation buttons for "Home", "Launch Job", and "Help" are visible. The main content area displays a green message: "'WASH complex' was successfully submitted. You will receive an email when it finishes computing (typically within one day). Should you need immediate results, you can download the software and run jobs on your local computer." Below this, it says "You can return to the [homepage](#) to submit more jobs." and a "Back" button is at the bottom.

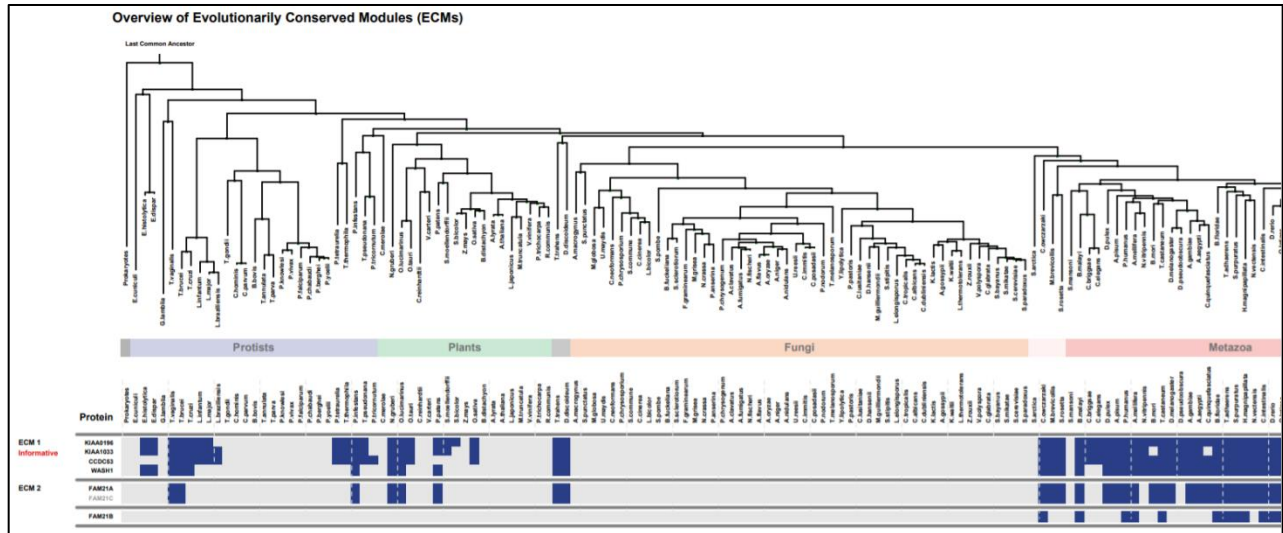
4. When the job is done, an email containing the link to the results will be sent to the user, shown below.



The screenshot shows an email notification. The subject line is "Your CLIME job 'WASH complex' is complete" with an "Inbox X" icon. The sender is "noreply-clime@broadinstitute.org via amazonses.com" and it is addressed "to me". The body of the email says "Dear user," followed by "Your CLIME job WASH complex using reference organism H.sapiens has completed successfully. Below are your results which you must download within 7 days:". A bulleted list provides three links: "PDF results: https://clime-results.s3.amazonaws.com/081845de-2880-413e-b9d3-a598deb32e6c/hsa138_1_2.pdf", "Text results: https://clime-results.s3.amazonaws.com/081845de-2880-413e-b9d3-a598deb32e6c/hsa138_1_2.txt", and "Input genes: <https://clime-results.s3.amazonaws.com/081845de-2880-413e-b9d3-a598deb32e6c/hsa138637774338031764717.txt>". Below the list, it says "Notes: none entered" and "Thanks. Clime Admin".

5. The **PDF** file contains the visual representation of partitioning of the gene sets into ECMs, each ECM's evolutionary model, and the symbols and phylogenetic profiles of the genes in the ECM. The **text** file contains the same information as the PDF file, and can be useful to view in Excel and to link gene symbols to other annotations (see Help). The detailed description of PDF output can be found in the firstpage on PDF file and in the Help page.

For our example gene set, the **PDF results** show the input genes have been clustered into two Evolutionary Conserved Modules (ECMs), with one gene unclustered.

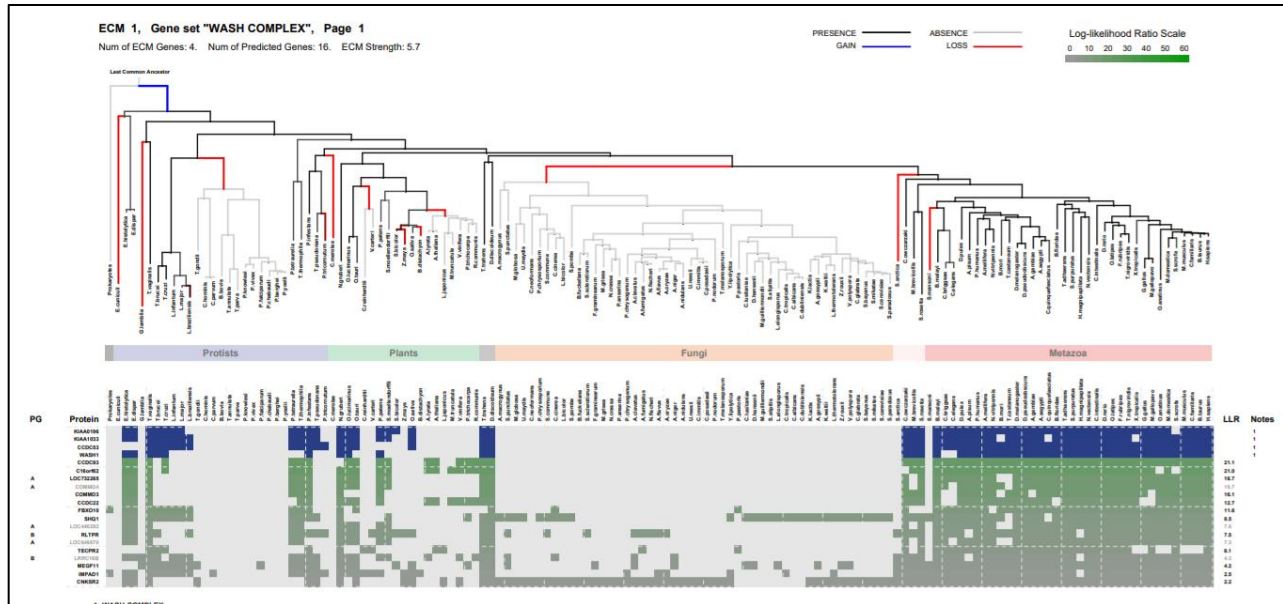


Zooming in on the gene names, we see:

Protein	
ECM 1 Informative	KIAA0196
	KIAA1033
	CCDC53
	WASH1
ECM 2	FAM21A
	FAM21C
	FAM21B

This shows that ECM1 contains four genes (KIAA0196, KIAA1033, CCDC53, WASH1), ECM2 contains 2 genes that are paralogs to each other (FAM21A, FAM21C, note gray color indicates a paralog to another human gene), and one gene that is not clustered (FAM21B).

The following pages expand on each ECM. Below we show the visualization of ECM1. The ECM genes are shown in blue rows and genes with similar evolutionary histories (ECM+ genes) are shown in green/gray rows. The gene symbols are shown at left.



At the top of this page above is the pre-computed phylogenetic tree of species. Red branches indicate places in evolution where the ECM genes have been lost. The heatmap below shows one row per gene within the input gene list (ECM genes in blue, ECM+ predictions in green), and one column per species. A blue or green or dark gray square indicates the gene's homolog is present in that species based on BlastP (Expect < 1e-3), whereas light gray indicates no homolog exists. This example shows that fungi have lost these WASH complex genes, along with several lineages of plants and protists. The gene symbols are shown on the left. Genes with paralogs in the reference species (human in this example) are indicated by the Paralog Group (PG) column on the left, and the second paralog of a paralog group have gene symbols shown in gray. The green-to-gray color scale (legend top right) indicates the Log-likelihood ratio score, which measures the likelihood that the given gene shares the same evolutionary history as the blue ECM genes.