EXTENDED EXPERIMENTAL PROCEDURES

Statistical Framework of CLIME

Summary

CLIME (CLustering by Inferred Models of Evolution) is a Bayesian statistical method that is based on a mixture of hidden Markov models (HMMs) with Dirichlet process prior to cluster input genes into modules by virtue of shared evolutionary history. CLIME then uses the inferred evolutionary HMMs to identify other genes that were not part of the original input gene set but may also share close evolutionary history with each of the clusters.

Notations

Let symbol G denote the input gene set with n genes, i.e., $|G| = n$, and let X denote the input phylogenetic profile matrix for all N genes in the reference genome. For example, G could be 44 subunit genes of human mitochondrial complex I, and X could be the phylogenetic profile matrix for all 22,000 human genes. The input phylogenetic tree has S extant species indexed by $1, \ldots, S$, and $S-1$ ancestral species indexed as $S+1, \ldots, 2S-1$. The $2S-1$ extinct and extant species are connected by the $2S-2$ branches on the tree. For each gene $g \in G$, let $X_g = (X_{g,1}, \ldots, X_{g,S})$ with $X_{g,j} = 1$ *or* 0 denote its phylogenetic profile of presence/absence across the S extant species, and let $H_g = (H_{g,1}, \ldots, H_{g,2S-1})$ denote its ancestral and extant presence/absence states of $2S-1$ species. We refer to the gene clusters with shared evolutionary history as evolutionarily conserved modules (ECM). Let $I = (I_1, \ldots, I_n)$ denote the ECM assignment indicators of genes, where $I_g = k$ indicates gene g is assigned to ECM k. We assume each gene can only be gained once throughout the entire evolutionary history, which happened at an unobserved branch λ_g. In our model, we condition on the observed data, X_g, and infer both latent variables λ_g and H_g using CLIME.

Generative Model for Phylogenetic Profiles

We use a tree-structured HMM to model the evolution history of genes. For each gene g, its complete evolutionary history $H_g = (H_{g,1}, \ldots, H_{g,2S-1})$ is partially observed because the phylogenetic profile vector $X_g = (X_{g,1}, \ldots, X_{g,S})$ is the observed (with error) presence/absence states for extant species $H_{g,1}, \ldots, H_{g,S}$. We assume that genes in one ECM share the same set of branch-specific probabilities of gene loss for the $2S-2$ branches, which is denoted by $\theta_k = (\theta_{k,1}, \ldots, \theta_{k,2S-2})$. For genes in ECM k, the transition of presence/absence states from its direct ancestor to species s is characterized by transition matrix $Q_{k,s}$,

$$Q_{k,s} = \begin{bmatrix} 1 & 0 \\ \theta_{k,s} & 1-\theta_{k,s} \end{bmatrix}. \qquad \text{(Equation S1)}$$

Because of our single gain branch assumption, the first row of $Q_{k,s}$ indicates that the transition probability from absence to absence is 1, and from absence to presence (re-gain) is 0. The second row shows our parameterization that the transition probability from presence to absence (gene loss) is $\theta_{k,s}$, and presence to presence is $1-\theta_{k,s}$.

Let $\sigma(s)$ denote the direct ancestor species of s, and let set $T(\lambda_g)$ contain all of the species in the sub-tree of a gain branch λ_g. Obviously $H_{g,s} = 0$ if species s is not in $T(\lambda_g)$. The likelihood function of evolutionary history H_g conditional on gene g being in ECM k is

$$P(H_g | \lambda_g, \theta_k, I_g = k) = \prod_{s \in T(\lambda_g)} Q_{k,s}(H_{g,\sigma(s)}, H_{g,s}), \qquad \text{(Equation S2)}$$

The phylogenetic profile vector X_g for each gene g represents the presence/absence of homologs of g across the S species. To account for potential errors in the presence/absence matrix, we allow each component of the observed phylogenetic profile, X_g, to have an independent probability ε to be erroneous (i.e., different from the true state $H_{g,s}$). ε is assumed to be low (default 0.01). For each gene g, the likelihood function of X_g given H_g is

$$P(X_g | H_g) = \prod_{s=1}^{S} P(X_{g,s} | H_g) = \prod_{s=1}^{S} (1-\varepsilon)^{\mathbb{I}\{X_{g,s} = H_{g,s}\}} (\varepsilon)^{\mathbb{I}\{X_{g,s} \neq H_{g,s}\}}, \qquad \text{(Equation S3)}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, that is equal to 1 if the statement in the parentheses is true, and 0 otherwise. Conditional on gene g being in ECM k, the complete likelihood function for gene g is

$$P(X_g, H_g | \lambda_g, \theta_k, I_g = k) = \left[\prod_{s \in T(\lambda_g)} Q_{k,s}(H_{g,\sigma(s)}, H_{g,s}) \right] \left[\prod_{s=1}^{S} (1-\varepsilon)^{\mathbb{I}\{X_{g,s} = H_{g,s}\}} (\varepsilon)^{\mathbb{I}\{X_{g,s} \neq H_{g,s}\}} \right]. \qquad \text{(Equation S4)}$$

Integrating out the hidden evolutionary history H_g, we are able to calculate the observed likelihood of X_g conditional on gene g being in ECM k,

$$P(X_g|\lambda_g,\theta_k,l_g=k)=\sum_{H_g}P(X_g,H_g|\lambda_g,\theta_k,l_g=k). \quad \text{(Equation S5)}$$

CLIME uses dynamic programming to calculate the summation above to avoid exponential amount of enumeration over all possible evolutionary histories. This scheme is called the backward (or peeling) procedure (Felsenstein, 1981; Durbin, 1998) and is widely used in computations for linear or tree-structure models such as HMMs.

Preprocessing Step

In the "Preprocessing" step (Figure 2C), CLIME infers the gain branch λ_g for each gene g, and then estimates the background null model for gene loss events from phylogenetic profiles of all genes in the input matrix. The null model is an ECM-independent HMM whose branch-specific loss probabilities are averaged over all genes in the genome.

For each gene g, CLIME assigns a single low loss probability $\tilde{\theta}$ (default 0.03) to all branches and calculates the likelihood of X_g conditional on each branch being its gain branch, then selects the gain branch that maximizes the marginal likelihood, $P(X_g|\lambda_g,\tilde{\theta})$, i.e.,

$$\widehat{\lambda}_g=\underset{\lambda_g=1,\ldots,2S-1}{\operatorname{argmax}}\ P(X_g|\lambda_g,\tilde{\theta}), \quad \text{(Equation S6)}$$

where $P(X_g|\lambda_g,\tilde{\theta})$ is computed by integrating out H_g from $P(X_g,H_g|\lambda_g,\tilde{\theta})$ using the backward procedure as described for Equation (S5). In all future steps of CLIME, the gain branch for each gene will be defined as $\widehat{\lambda}_g$ and treated as known. An alternative way for estimating the gain branch for each gene in G is to update λ_g in each MCMC iteration of the partitioning step and then calculate the posterior distribution of λ_g. There are two reasons why we chose to estimate the gain branch for each gene in preprocessing step and keep it fixed in later two steps. First, the gain branches usually have very small uncertainty and can be confidently estimated in the preprocessing step. Therefore, it is reasonable to keep them as known and fixed in later steps. Second, by estimating the gain branches at the preprocessing step and setting them as fixed in the later steps, we substantially reduce the computation time compared to updating the gain branches at each MCMC iteration.

CLIME then estimates the background null model θ_0, which is defined as the average probabilities of gene loss for all genes in reference genome X. CLIME first imputes the missing evolutionary histories for all genes in X by forward-summation-backward-sampling method (Liu, 2008) with a single initial low loss rate (default 0.03), and then computes background loss probability for each tree branch s as the fraction of genes lost on s. To account for the uncertainty in gene's evolutionary history, for each gene g we impute its evolutionary history by drawing 100 samples $H_g^{(1)},\ldots,H_g^{(100)}$ from the conditional distribution $P(H_g|X_g,\tilde{\theta})$ with forward-summation-backward-sampling method, and then estimate the background null model, $\widehat{\theta}_0=(\widehat{\theta}_{0,1},\ldots,\widehat{\theta}_{0,2S-2})$, as

$$\widehat{\theta}_{0,s}=\frac{\sum_{g=1}^{N}\sum_{j=1}^{100}\mathbb{I}\left\{H_{g,\sigma(s)}^{(j)}=1,H_{g,s}^{(j)}=0\right\}}{\sum_{g=1}^{N}\sum_{j=1}^{100}\mathbb{I}\left\{H_{g,\sigma(s)}^{(j)}=1\right\}},\quad \text{for } s=1,\ldots,2S-2. \quad \text{(Equation S7)}$$

Partitioning Input Gene Set with Bayesian Mixture Model

In the "Partition" step (Figure 2C), CLIME uses a Bayesian mixture model (Gelman et al., 2013) with Dirichlet process (or Chinese-restaurant process) prior (Ferguson, 1973) to partition the input gene set G into ECMs. We let K denote the number of ECMs in input gene set G. The application of Dirichlet process mixture model enables CLIME to automatically determine K in the partitioning step.

Let $X=(X_1,\ldots,X_n)$, $H=(H_1,\ldots,H_n)$ and $\theta=(\theta_1,\ldots,\theta_K)$, where each θ_j is a vector of length *2S-1*, corresponding to probabilities of genes loss on all the branches of the evolutionary tree for the K ECMs. The complete likelihood function for X and H given θ and I is

$$P(X,H|\theta,I)=\prod_{g=1}^{n}P\left(X_g,H_g|\widehat{\lambda}_g,\theta_{I_g}\right), \quad \text{(Equation S8)}$$

where θ_{I_g} denotes the probabilities of gene loss for the ECM that gene g belongs to. We use Dirichlet process to model the prior distribution of the probabilities of gene loss for unknown number of ECMs. In particular, for each gene g we let the prior distribution of θ_g follow Dirichlet process with concentration parameter α and base distribution $\mathcal{D}_0$. This derives the following Bayesian hierarchical model:

$$\begin{aligned}
X_g|H_g &\sim P(X_g|H_g),\\
H_g|\theta_g &\sim P\left(H_g|\widehat{\lambda}_g,\theta_g\right),\\
\theta_g|D &\sim D,\\
D &\sim \mathrm{DP}(D_0,\alpha),\\
D_0 &= \prod_{s=1}^{2S-2}\mathrm{Beta}(a,b),
\end{aligned} \quad \text{(Equation S9)}$$

where $DP(\mathcal{D}_0, \alpha)$ stands for the Dirichlet process with base distribution D_0 and scaling parameter α, and the base distribution D_0 is formularized as a product of the Beta conjugate prior distributions for branch-specific gene loss probabilities.

The partitioning based on Dirichlet process mixture model can be naturally implemented by Markov chain Monte Carlo (MCMC) sampling algorithm (Liu, 2008; Neal, 2000). In particular, we use Chinese restaurant process representation (Aldous, 1985; Pitman, 1996) of a Dirichlet process and use the Gibbs sampler (Liu, 2008; Gelfand and Smith, 1990) algorithm to sample from the joint posterior distribution of ECM assignments $I = (I_1, \ldots, I_n)$, branch specific probabilities of gene loss $\theta = (\theta_1, \ldots, \theta_K)$ and evolutionary histories $H = (H_1, \ldots, H_n)$. The Chinese restaurant process representation of CLIME's hierarchical Bayesian model can be formulated as,

$$\begin{aligned} X_g|H_g &\sim P(X_g|H_g), \\ H_g|\theta_{I_g} &\sim P\left(H_g|\widehat{\lambda}_g, \theta_{I_g}\right), \\ \theta_k &\sim \prod_{s=1}^{2S-2} \text{Beta}(a,b), \quad k = 1,2,\ldots,K \\ P(I_g = k|I_1,\ldots,I_{g-1}) &= n_{g,k}/(g-1+\alpha), \quad g = 1,2,\ldots,n \\ P(I_g \neq I_j|I_1,\ldots,I_{g-1}) &= \alpha/(g-1+\alpha), \quad g = 1,2,\ldots,n \end{aligned} \qquad \text{(Equation S10)}$$

where $n_{g,k}$ is the number of I_j for $j<g$ that are equal to k. The Chinese restaurant process prior for clusters assignments is exchangeable (Aldous, 1985), therefore the prior distribution for I is invariant to the order of genes in G. We designed a Gibbs sampler for exploring the posterior distribution space, in which each iteration has the following three steps (Figure 2C):

a) update each $\boldsymbol{H_g} = (\boldsymbol{H_{g,1}}, \ldots, \boldsymbol{H_{k,2S-1}})$, $g = 1, \ldots, G$, by drawing from distribution $\boldsymbol{P}(\boldsymbol{H_g}|\widehat{\lambda}_g, \boldsymbol{X_g}, \theta_{I_g})$
b) update each $\theta_k = (\theta_{k,1}, \ldots, \theta_{k,2S-2})$, $k = 1, \ldots, K$, by drawing from distribution $\boldsymbol{P}(\theta_k|\widehat{\lambda}_g, \boldsymbol{H_k})$, where $\boldsymbol{H_k}$ denote all $\boldsymbol{H_g}$ for that $I_g = k$, i.e., $\boldsymbol{H_k} = \{\boldsymbol{H_g} : \boldsymbol{I_g} = \boldsymbol{k}\}$
c) update each $\boldsymbol{I_g}$, $g = 1, \ldots, G$, by re-assigning gene $\boldsymbol{g}$ to an existing ECM $\boldsymbol{k}$ with probability $\boldsymbol{P}(\boldsymbol{I_g} = \boldsymbol{k}|\boldsymbol{X_g}, \widehat{\lambda}_g, \theta_k)$ or forming a new ECM with probability $\boldsymbol{P}(\boldsymbol{I_g} = \boldsymbol{K}+\boldsymbol{1}|\boldsymbol{X_g}, \widehat{\lambda}_g)$.

For step (a), it is straightforward to draw H_g by forward-summation-backward-sampling method (Scott, 2002). Note that conditional on the ECM assignment of gene g, the distribution of H_g can be written in the factorized form,

$$P(H_g|X_g, \theta_{I_g}) = \prod\nolimits_{s \in T(\lambda_g)} P\left(H_{g,s}|H_{g,\sigma(s)}, X_g, \widehat{\lambda}_g, \theta_{I_g}\right), \qquad \text{(Equation S11)}$$

which suggests a sequential sampling scheme (Liu, 2008). In the sampling procedure, we first set $H_{g,s} = 0$ for all $s \notin T(\lambda_g)$ and $H_{g,\lambda_g} = 1$, then starting from the gain branch λ_g we iteratively draw each $H_{g,s}$ on tree from distribution $P(H_{g,s}|H_{g,\sigma(s)}, X_g, \theta_{I_g})$ conditional on the drawn state of its direct ancestral species $H_{g,\sigma(s)}$. The probability distributions $P(H_{g,s}|H_{g,\sigma(s)}, X_g, Q_{I_g})$ are calculated by the backward procedure (Durbin, 1998; Scott, 2002) in time complexity linear to S with the dynamic programming scheme.

For step (b), since we adopt a conjugate Beta(a,b) prior distribution for each $\theta_{k,s}$, the conditional distribution $P(\theta_k|H_k)$ is simply the product of Beta posterior distributions,

$$P\left(\theta_k|\widehat{\lambda}_g, H_k\right) = \prod_{s \in T_k} \text{Beta}\left(\theta_{k,s}|a + \sum_{I_g = k} \mathbb{I}\left(H_{g,\sigma(s)} = 1, H_{g,s} = 0\right), b + \sum_{I_g = k} \mathbb{I}\left(H_{g,\sigma(s)} = 1, H_{g,s} = 1\right)\right), \qquad \text{(Equation S12)}$$

where T_k is the minimal sub-tree that contains all λ_g's for $I_g = k$. For the Beta(a,b) prior distributions, we select small a and b so that the prior has little effect to the posterior distribution. Specifically, we let $a = 0.0045$ and $a+b = 0.15$ so that the prior mean of probabilities of gene loss is $a/(a+b) = 0.03$, which is observed to be the average loss probability of all 20,000 human genes on the 138 eukaryotic species tree. Note that in the implementation of the algorithm, we integrated θ s out of the model and applied collapsed Gibbs sampler scheme (Liu, 1994) so that this step (b) can be skipped without changing the target distribution of partitioning I for the Gibbs sampler.

For step (c), we calculate the probabilities of gene g joining any existing ECM k or forming a new ECM by multiplying the data likelihood function with the prior distribution. By the property of Dirichlet process, the prior probabilities for each gene to join the existing ECMs are proportional to the sizes of the ECMs, and the prior probability for each gene to form a new ECM is proportional to the α, which is set to be 1 by default and can be adjusted freely by the user. The prior distribution for ECM assignment I_g conditional on $I_{[-g]}$ is formulated as

$$P(I_g = k|I_{[-g]}) = \begin{cases} n_{k,[-g]}/(n-1+\alpha), & \text{for } k = 1,\ldots,K \\ \alpha/(n-1+\alpha), & \text{for } k = K+1, \end{cases} \qquad \text{(Equation S13)}$$

where $n_{k,[-g]}$ is the size of ECM k excluding gene g, and $K+1$ indicates forming a new ECM. Multiplying the prior distribution with the likelihood function, we get the posterior conditional distribution of I_g (up to a normalizing constant),

$$P\left(l_g = k | l_{[-g]}, X_g, \widehat{\lambda}_g, \theta\right) \propto \begin{cases} n_{k,[-g]} \cdot P\left(X_g | \widehat{\lambda}_g, \theta_k\right), & \text{for } k = 1, \ldots, K, \\ \alpha \cdot P\left(X_g | \widehat{\lambda}_g\right), & \text{for } k = K+1, \end{cases} \quad \text{(Equation S14)}$$

where $P(X_g|\widehat{\lambda}_g, \theta_k)$ is the likelihood of phylogenetic profile of gene g if it belongs to ECM k, integrating over all possible evolutionary histories,

$$P\left(X_g | \widehat{\lambda}_g, \theta_k\right) = \sum_{H_g} P\left(X_g, H_g | \widehat{\lambda}_g, \theta_k\right). \quad \text{(Equation S15)}$$

As mentioned earlier, we use dynamic programming to calculate this equation in linear time complexity. Lastly, $P(X_g|\widehat{\lambda}_g)$ is the likelihood of gene g being in its own singleton ECM, and this probability can be calculated by using the marginalization property of the Dirichlet process as:

$$P\left(X_g | \widehat{\lambda}_g\right) = \int P\left(X_g | \widehat{\lambda}_g, \theta_g\right) dD_0\left(\theta_g\right). \quad \text{(Equation S16)}$$

Using this MCMC sampling scheme, genes with similar evolutionary history will be clustered together to form ECMs, and genes without any close neighbor will stay in their own singleton ECMs. This process helps CLIME to automatically estimate the number of ECMs in the input gene set. CLIME calculates the marginal likelihood $P(X|I)$ at the end of each MCMC iteration, and the ECM assignments with highest marginal likelihood among all iterations, denoted as $\widehat{I}$, will be reported as the final ECM partitioning of the input gene set. The marginal likelihood $P(X|I)$ for our model does not have closed form. Therefore we used Chib's method to approximate it with MCMC samples (Chib, 1995). We used a simulated annealing scheme to increase the efficiency for finding the highest marginal likelihood partitioning $\widehat{I}$. The initial temperature of the simulated annealing algorithm was set to be 10 and gradually decreased to 0.05 during the MCMC sampling.

In summary, the MCMC sampling of the Bayesian mixture model with Dirichlet process prior enables CLIME to automatically estimate the number of ECMs, the ECM partitioning of input gene set G, and the loss probability parameters for each ECM.

Collapsed Gibbs Sampler by Integrating θ' s out of the Model

In the implementation of the Gibbs sampler, we integrated the θ' s out of the model and ran the collapsed Gibbs sampler (Liu, 1994). By integrating out the θ' s from the model, we are able to skip step (b) in the MCMC iterations, which dramatically increases the rate of Markov chain convergence, thus improving the efficiency of the Gibbs sampler. The collapsed Gibbs sampler will converge to exactly same target distribution (Liu, 1994), therefore the efficiency of the algorithm will improve without any influence on the results. After M (default 1,000) iterations of collapsed Gibbs sampling, we run another 1,000 MCMC iterations to sample from the posterior distribution of θ s conditional on the optimal partitioning $\widehat{I}$. The loss probability of branch s for each ECM k, $\widehat{\theta}_{k,s}$, can be estimated by the posterior mean of $\theta_{k,s}$ conditional on $\widehat{I}$,

$$\widehat{\theta}_{k,s} = E\left[\theta_{k,s} | X, \widehat{I}\right], \quad \text{(Equation S17)}$$

which is approximated by the average of 1,000 MCMC samples.

Definition of ECM Strength

After partitioning the input gene set G into ECMs, it is of great interest to determine which of the ECMs share more informative and coherent evolutionary histories than others, since the ranking of ECMs leads to different priorities for further low throughput experimental investigations. In our Bayesian model-based framework, the strength of ECM k, ϕ_k, is defined as the logarithm of the Bayes Factor for two models normalized by the number of genes in that ECM: one model is under the assumption that all genes in this ECM have coevolved, and the other is under the alternative assumption that each gene has evolved independently in its own singleton ECM under the background null model. Specifically, the strength for ECM k is formulated as

$$\phi_k = \frac{1}{N_k} \log \left[\frac{\int \left[\prod_{l_g = k} P\left(X_g | \widehat{\lambda}_g, \theta\right) \right] P(\theta) d\theta}{\prod_{l_g = k} P\left(X_g | \widehat{\lambda}_g, \theta_0\right)} \right], \quad \text{(Equation S18)}$$

where N_k is the number of genes in ECM k and $p(\theta)$ denotes the prior distribution of loss rates.

Expansion Step

In the "Expansion" step, we use the inferred HMM of each ECM to identify other genes in the reference genome X that also share the same evolutionary history. In particular, for each gene and each ECM we calculate the log-likelihood ratio (LLR) between two models: one is that the gene evolved under the HMM of ECM and the other is that the gene evolved under the background null HMM. The inferred HMM of each ECM k consists of two parts: the inferred gain branch of this ECM, $\widehat{\lambda}_k$, which is defined as the last common ancestor of inferred gain branches of genes in this ECM, and inferred loss probabilities of this ECM over all branches, $\widehat{\theta}_k = (\widehat{\theta}_{k,1}, \ldots, \widehat{\theta}_{k,2S-1})$. The LLR score for gene g coevolved with ECM k is formulated as

$$LLR_{g,k} = 2\left(L_{g,k} - L_{g,0}\right), \quad \text{(Equation S19)}$$

where $L_{g,k}$ is the log-likelihood of X_g generated by the HMM of ECM k,

$$L_{g,k} = \log P\left(X_g \middle| \widehat{\lambda}_k, \widehat{\theta}_k\right), \quad \text{(Equation S20)}$$

and $L_{g,0}$ is the log-likelihood of X_g generated by the background null HMM,

$$L_{g,0} = \log P\left(X_g \middle| \widehat{\lambda}_g, \widehat{\theta}_0\right). \quad \text{(Equation S21)}$$

High value of $LLR_{g,k}$ indicates that the HMM of ECM k explains the phylogenetic profile X_g much better than the background null model, which suggests that gene g is more probable to share the same evolutionary history with the genes in ECM k, than the null model.

For each ECM, CLIME scores all genes in the reference genome X (green matrix in Figure 1), ranks the genes by their LLR scores, and sets a cutoff at a certain threshold (e.g., 0) to get the expanded gene list (ECM+).

CLIME Model of Gene Evolution

CLIME is different from other tree-based phylogenetic profiling methods in the way by which it models gene evolution. CLIME assumes a single gain branch and branch-specific loss probabilities for each gene module. Only tree topology (not branch length) is utilized. Our probabilistic approach is distinct from methods using a fixed model of gene evolution (e.g., dollo parsimony) and from the BayesTraits probabilistic algorithm – which models gene births and losses as Poisson processes that depend on branch length (Barker et al., 2007; Barker and Pagel, 2005). Our model is branch-length independent since there are many biological examples in which closely related species (with short branch lengths) have widely different number of genes (thus have undergone rapid and extensive gene gain/loss). The BayesTraits assumptions enable it to have only 8 parameters to estimate – rendering it more stable, but also vulnerable to inaccuracies in branch length estimation. Although it can be challenging to estimate loss probabilities for all branches, CLIME's assumption of a single gain branch helps to stabilize the estimates of loss probabilities. Modeling branch-specific loss probabilities enables CLIME to detect surprising loss events not expected based on the branch length. CLIME is not expected to perform well in bacteria, where horizontal gene transfer may be rampant and violates CLIME's single gain assumption.

Choice of Parameter Default Values

Based on our 138-species tree and our phylogenetic matrices, we estimated default values for two CLIME parameters:

- **Error rate of observed phylogenetic profiles, ε (default 0.01).** Initially, we designed CLIME to estimate ε and calculate its posterior probability distribution from the data using the MCMC sampling processes. Under this scenario, ε was estimated to be 0.01, with small uncertainty. However, we observed that CLIME's algorithm was robust to the choice of ε: setting ε to be 0.005, 0.01 or 0.02 had little effect on the CLIME results. Therefore we made ε a user-defined parameter with default 0.01 in order to decrease the complexity of CLIME's inference and increase the computation speed. Based on our experience, other ε settings will have little effect on the resulting partitioning or expansion.
- **Branch-independent loss rate (default 0.03).** During the preprocessing step only, CLIME uses a branch-independent loss rate in order to estimate each gene's gain branch, and to estimate the null model (in which each branch has its own loss probability). The branch-independent loss rate is set to 0.03 – which is the average loss rate for all human genes across all branches of our 138 eukaryotic species tree. Since this parameter is merely the starting point, CLIME should be fairly robust to the value of this parameter.

Software Inputs

CLIME program inputs (i) a binary phylogenetic tree topology, T, in Newick format, (ii) a binary phylogenetic matrix, X, in tab-delimited format, with one row per gene and one column per species, along with two initial columns containing unique gene identifiers and unique gene symbols, (iii) an optional paralogy matrix consisting of two tab-delimited columns with pairs of gene identifiers from the reference species that show sequence similarity, (iv) a gene set G, formatted as a list of gene identifiers.

Software Outputs

CLIME program outputs (i) a PDF file containing the partition of G into disjoint ECMs, a visual representation of each ECM's evolutionary model (with gain branches shown in blue and loss branches show in different shades of red corresponding to probability of loss), the phylogenetic profiles and gene symbols of the genes in each ECM, and the phylogenetic profiles and symbols of the genes in the ECM+. Paralogous genes (based on the input paralogy matrix) are assigned to "paralogy groups," and the paralogy group is indicated in the PDF file along with gray font; (ii) a text file containing the ECMs and ECM+ genes and phylogenetic profiles.

Running Time/Resources

CLIME is implemented in an algorithm of complexity $O(MSn^2)$ and memory usage $O(NS + MSn)$, where N is the number of genes in the reference genome, n is the number of genes in the input set, S is the number of species, and M is the number of MCMC iterations. Using the default setting on 1,000 MCMC iterations and input of 20,000 human genes across 138 species, on a standard single computer processor CLIME's running times for input gene sets are roughly: 2-3 min (10 genes), 20 min (100 genes), 1-2 hr (200 genes), or 12 hr (1000 genes). Clustering large gene sets (5000 genes) takes less than two days (using method described below).

Clustering Large Input Gene Sets

For clustering large gene sets (3 model organisms) we applied two solutions to avoid local trapping within the MCMC partitioning step. First, we used the SAME_GL flag of CLIME software that ensures that only genes with same pre-estimated gain branch can be partitioned into same ECM. This helps to dramatically increase the computational efficiency of the algorithm, and this restriction does not have much impact on the results. Second, to avoid local trapping we started the MCMC from different initializations. Specifically, we launched in parallel 10,000 identical CLIME jobs (each using a single processor). Each CLIME job uses a different starting initialization and a different random number generator seed for its 1,000 MCMC iterations and then retains the partitioning with the highest marginal likelihood. We then selected the highest marginal likelihood from the 10,000 independent CLIME runs to return the optimal partitioning. CLIME analysis of *S. cerevisiae* (5822 genes) took less than two days on a single processor, and the entire parallel procedure took less than two days on a compute farm. We report in the software package the random number generator seeds for the three large data sets (3 model organisms) corresponding to the highest marginal likelihood of 10,000 runs.

CLIME Analysis of Human Mitochondrial Proteins

Mitochondrial genes were compiled from MitoCarta (Pagliarini et al., 2008), excluding entries recently discarded from the NCBI database. Gain branches and loss events for each mitochondrial gene were estimated in Preprocessing step of CLIME. We calculated the cumulative gain proportions of mitochondrial genes versus all human on the 27 potential gain branches between human and the eukaryotic least common ancestor (LCA) (Figure 6B). The average loss probabilities for all mitochondrial genes were calculated for each tree branch, and plotted against the average loss probabilities for all human genes (Figure 6C). CLIME partitioned the 1,007 human MitoCarta genes into 120 nonsingleton ECMs containing 606 genes, 61 of which were significantly enriched for known biological processes or cellular components (hypergeometric p value < 1e-4) (Figure S6). Evolutionary patterns alone were able to cluster together key functional pathways such as fatty acid biosynthesis, folate biosynthesis, lipoic acid metabolism, the mevalonate pathway, and the TIMM/TOMM protein import machinery. Expansions of these modules can reveal interacting proteins, for example in heme biosynthesis (Figure S6). Heme biosynthesis is accomplished by eight enzymatic reactions that originate in mitochondria and continue in the cytosol before returning to mitochondria (Nilsson et al., 2009). Three of the 4 known mitochondrial enzymes (*PPOX*, *CPOX*, *FECH)* were automatically grouped together into ECM 36 ($\phi = 4.5$), and its expansion ECM+ contained 4 genes with LLR > 15, including 3 of the 4 cytosolic heme enzymes (*UROD*, *HMBS*, *ALAD*). The fourth ECM+ gene, *TMEM63A*, does not have any known function annotation. Surprisingly, 4 genes involved in folate metabolism (*MTHFD2L*, *MTFMT*, *MTHFD2*, *MTHFD1*) were also a part of the ECM+. This module showed specific losses in *Cryptosporidium*, *Piroplasmida* and *Nematoda*. Although folate deficiency is known to give rise to anemia, shared evolutionary history between heme biosynthesis and folate metabolism is unexpected.

Impact of Incomplete or Inaccurate Genome Annotation

Inclusion of incomplete genome annotations (due either to draft assemblies or systematic biases in the genome annotation pipeline) can cause spurious evolutionary signals. Incomplete genome annotations will manifest as inaccurate "absence" calls in the phylogenetic matrix for many genes in a given species. While CLIME models profile inaccuracies (ε profile-error parameter), this works best when inaccuracies are independent (e.g., BLASTP will introduce different errors for gene A versus gene B). In two cases, we observed that systematic genome annotations inaccuracies led to false evolutionary signals: (i) a spurious OXPHOS cluster was generated by incorrect "absence" calls of mtDNA proteins, since 27 of 138 eukaryotic species annotations did not include mtDNA annotations, and (ii) a spurious ribosome cluster was generated by incorrect absence calls of 15 short ribosomal proteins – which were systematically not annotated in 8 fungal species (although present in the underlying genome sequences based on TBLASTN analysis). Therefore, care should be taken to apply CLIME only to species with largely complete annotations – as assessed using criteria such as CEGMA (Parra et al., 2007).

SUPPLEMENTAL REFERENCES

Aldous, D.J. (1985). Exchangeability and related topics. In École d'Été De Probabilités De Saint-Flour XIII—1983, P.L. Hennequin, ed. (Berlin Heidelberg: Springer-Verlag), pp. 1–198.

Chib, S. (1995). Marginal likelihood from the Gibbs output. J. Am. Stat. Assoc. *90*, 1313–1321.

Durbin, R. (1998). Biological Sequence Analysis (Cambridge: Cambridge University Press).

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. *17*, 368–376.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. Ann. Stat. *1*, 209–230.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. *85*, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013). Bayesian Data Analysis, Third Edition (Boca Raton: CRC Press).

Liu, J.S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. J. Am. Stat. Assoc. *89*, 958–966.

Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. *9*, 249–265.

Nilsson, R., Schultz, I.J., Pierce, E.L., Soltis, K.A., Naranuntarat, A., Ward, D.M., Baughman, J.M., Paradkar, P.N., Kingsley, P.D., Culotta, V.C., et al. (2009). Discovery of genes essential for heme biosynthesis through large-scale gene expression analysis. Cell Metab. *10*, 119–130.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061–1067.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. Lecture Notes—Monograph Series *30*, 245–267.

Scott, S.L. (2002). Bayesian methods for hidden Markov models: recursive computing in the 21st century. J. Am. Stat. Assoc. *97*, 337–351.

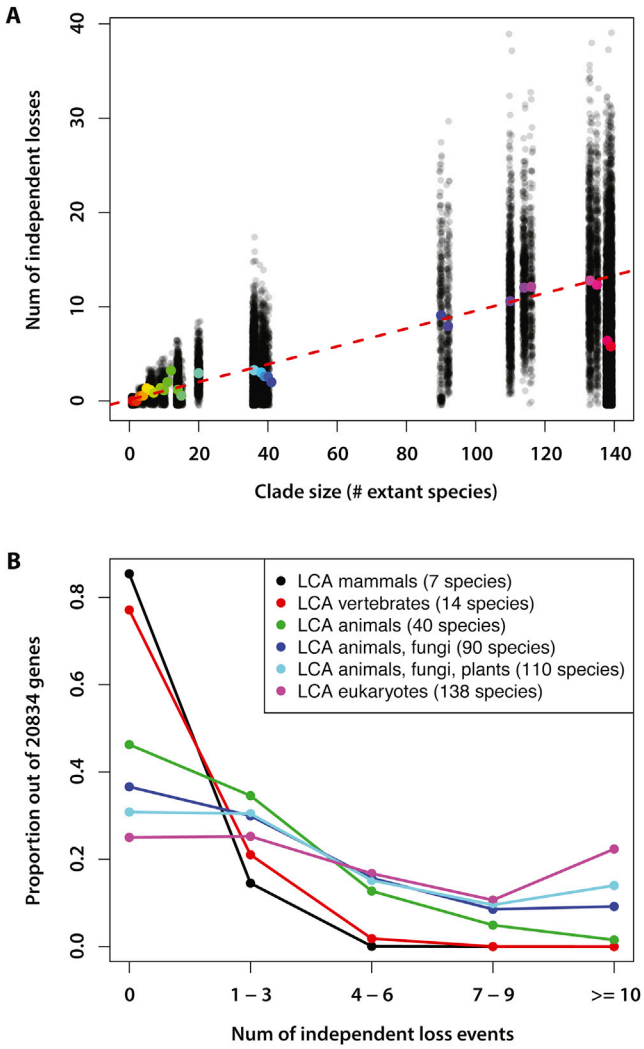

Figure S1. Relationship between Phylogeny Size and Number of Independent Losses, Related to Figure 1

(A) Clade size (number of extant species) versus number of independent losses is plotted for each of the 27 potential gain branches between human and the eukaryotic least common ancestor (LCA). Each black dot shows one human gene, where the x axis shows the clade size representing where the gene was gained and the y axis shows the number of independent losses derived by CLIME. Dots are jittered for visualization. Colored dots show the average number losses for each of the 27 potential gain sites. The red dashed line shows the least square regression line for all colored dots excluding the two rightmost (deepest eukaryotic branches).

(B) Number of independent losses versus percent of human genes is plotted for different phylogenies (colored lines). Each phylogeny represents a subset of the 138-species eukaryotic tree. Unlike (A), the genes were not first analyzed by their gain branch but just the number of losses present in given phylogenetic subset.

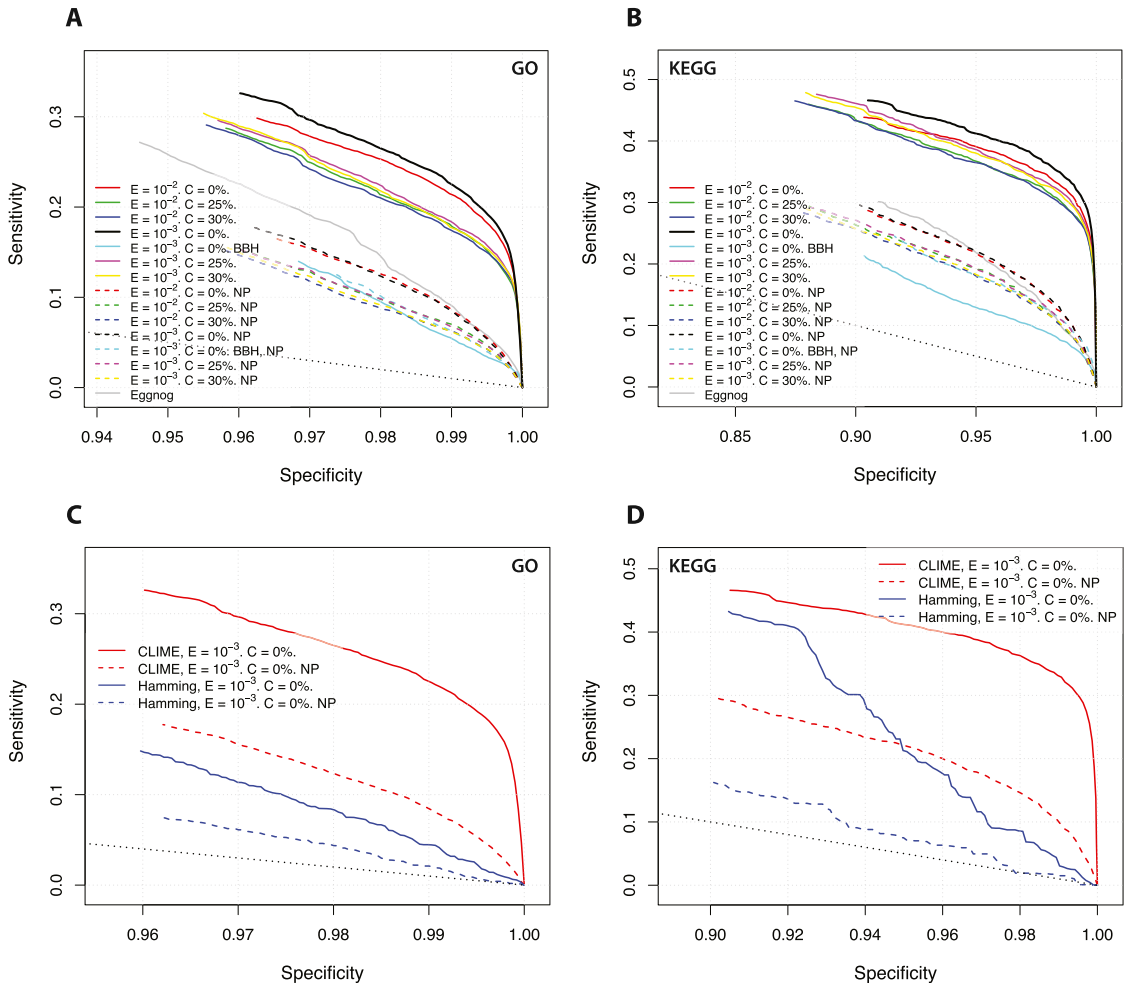

Figure S2. Comparison of Phylogenetic Matrices and Distance Metrics, Related to Figure 1

(A–D) CLIME performance is assessed using receiver operator curves (ROC) based on leave-one-out analysis of curated pathways from GO cellular location (A,C) and KEGG metabolic and signaling pathway gene sets (B,D). Each curve plots the relationship between sensitivity and specificity as LLR is increased from 0 (leftmost point on each curve) to its maximum value (right). Panels A and B both show results on different input matrices. Panels C and D both show results on different distance metrics: CLIME versus naive phylogenetic profiling (Hamming distance method, Pellegrini et al., 1999) using the same input matrix.

Solid lines show different input matrices (A, B) or different distance metrics (C, D). Dashed lines show input matrices from which paralogous genes were removed. Dotted lines show the line of no discrimination (results of random chance).

Abbreviations: E: expect; C: query coverage; BBH: best bidirectional hit; NP: nonparalogous.

Figure S3. Comparison of Phylogenetic Profiling Methods Based on Simulation Studies, Related to Figure 1

Results from simulation studies for tree-based model (A) and tree-independent model (B) of evolution are shown for each of three phylogenetic profiling methods: CLIME (black bars), hierarchical clustering with Hamming distance (gray bars) (Pellegrini et al., 1999), and hierarchical clustering with anticorrelation (Glazko and Mushegian, 2004) distance (white bars). Clustering accuracy was measured by Adjusted Rand Index (ARI, y axis) (Hubert and Arabie, 1985). Each bar shows the average ARI from 100 simulated data sets, with each data set containing a mixture of 50 ECMs (with 10 genes per ECM). Simulation parameters include: N_L: number of loss branches for each ECM; P_L: probability of loss for each loss branch; N_S: number of singleton genes in each simulated data set.

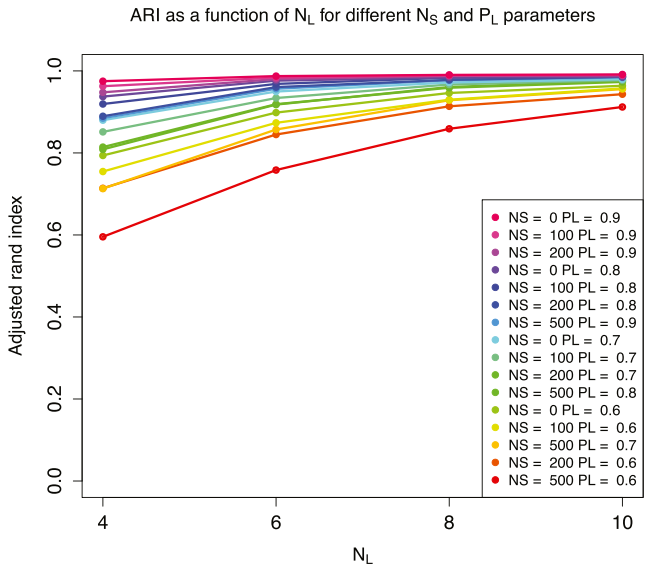

Figure S4. Power of CLIME as a Function of Simulated Loss Events, Related to Figure 1

CLIME performance (adjusted Rand index) based on number of independent loss events (N_L), under the tree-based simulation study with different parameters for number of singletons (N_S) and the probability of gene loss per branch (P_L). The legends are in the same vertical order of the curves.

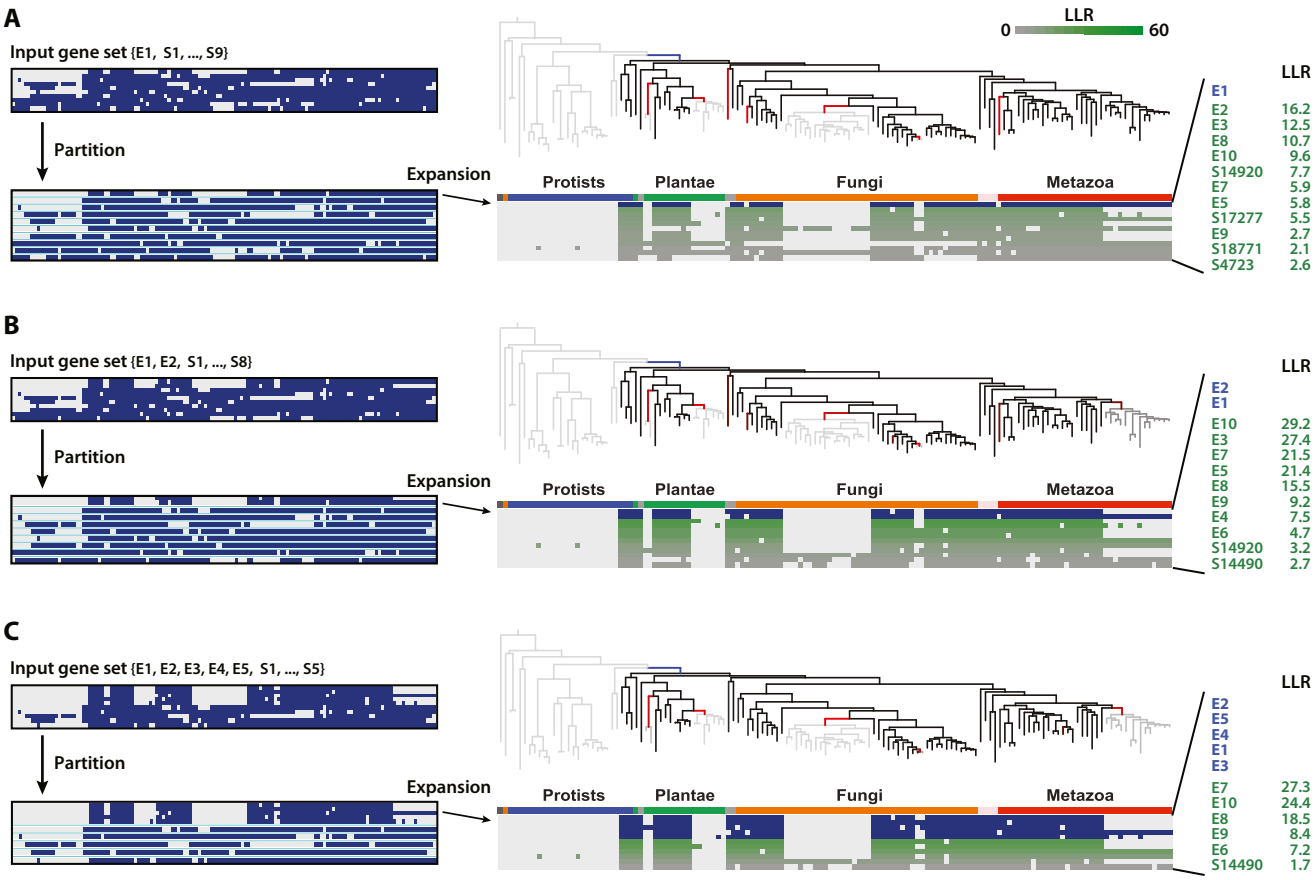

Figure S5. CLIME Inference on a Simulated Data Set, Related to Figure 1

Results from simulation study on a data set with 10 genes from simulated ECM, E* (E1, E2, …, E10) and 19,990 singleton genes (S1, S2, …, S19990). Shown are CLIME partitions of three input gene sets:

(A) 1 gene from E* (E1) and 9 singleton genes,

(B) 2 genes from E* (E1, E2) and 8 singleton genes,

(C) 5 genes from E* (E1, E2, E3, E4, E5) and 5 singleton genes.

Insets show the ECMs that contain genes from E*. The inferred evolutionary model is shown on the phylogenetic tree, with gain branches shown in blue and loss branches show in different shades of red corresponding to probability of loss. The ECM+ expansion genes are shown in green/white matrices with different shades of green indicating their prediction LLR scores (listed on right side).

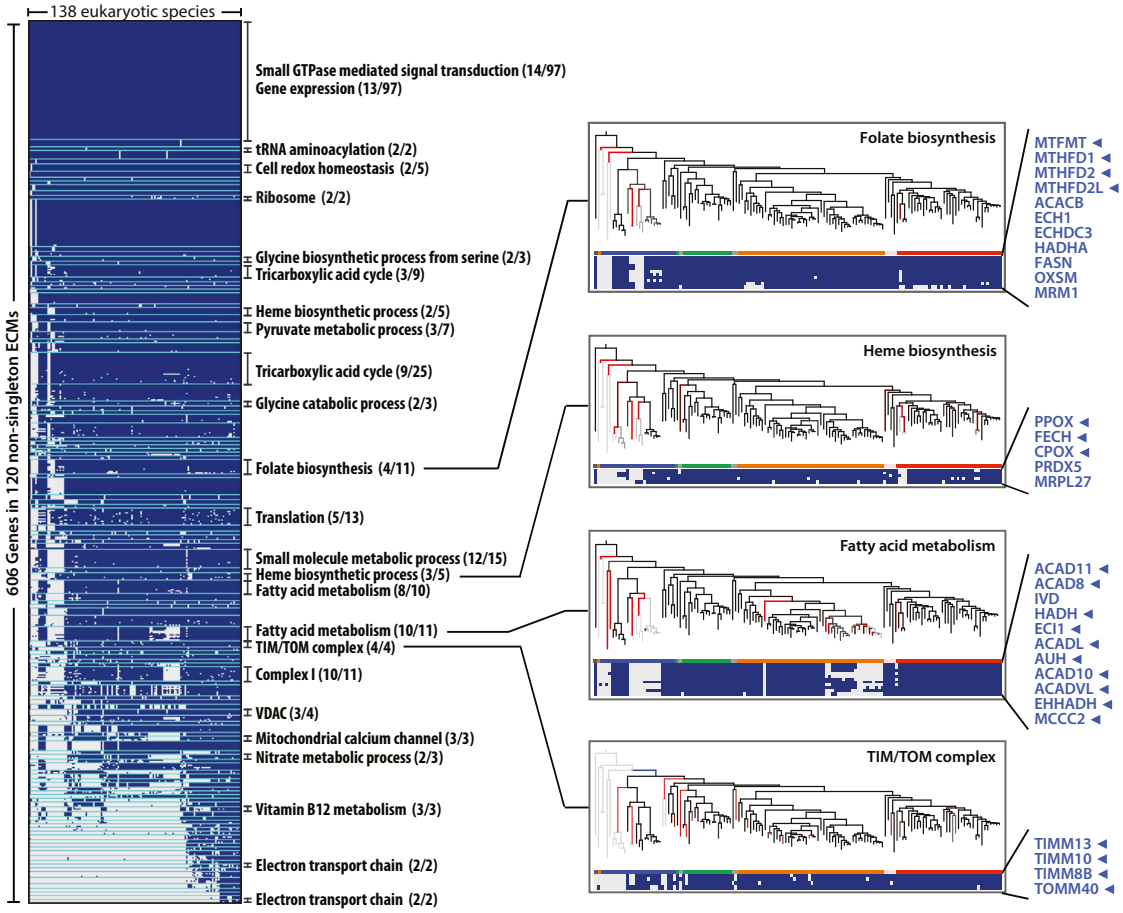

Figure S6. CLIME Partitioning of the Mitochondrial Proteome, Related to Figure 6

CLIME partitioning of 1,007 human MitoCarta genes into 120 nonsingleton ECMs (separated by aqua lines; 401 singletons not shown). Selected ECMs are labeled with GO/KEGG pathways having significant enrichment (parenthesis show fraction of all ECM genes in given pathway). ECMs are ordered by mean number of homologs present across taxa. Insets show four ECMs with their inferred independent losses (red branches) and ECM members. Blue arrows indicate genes in the enriched GO/KEGG pathway.

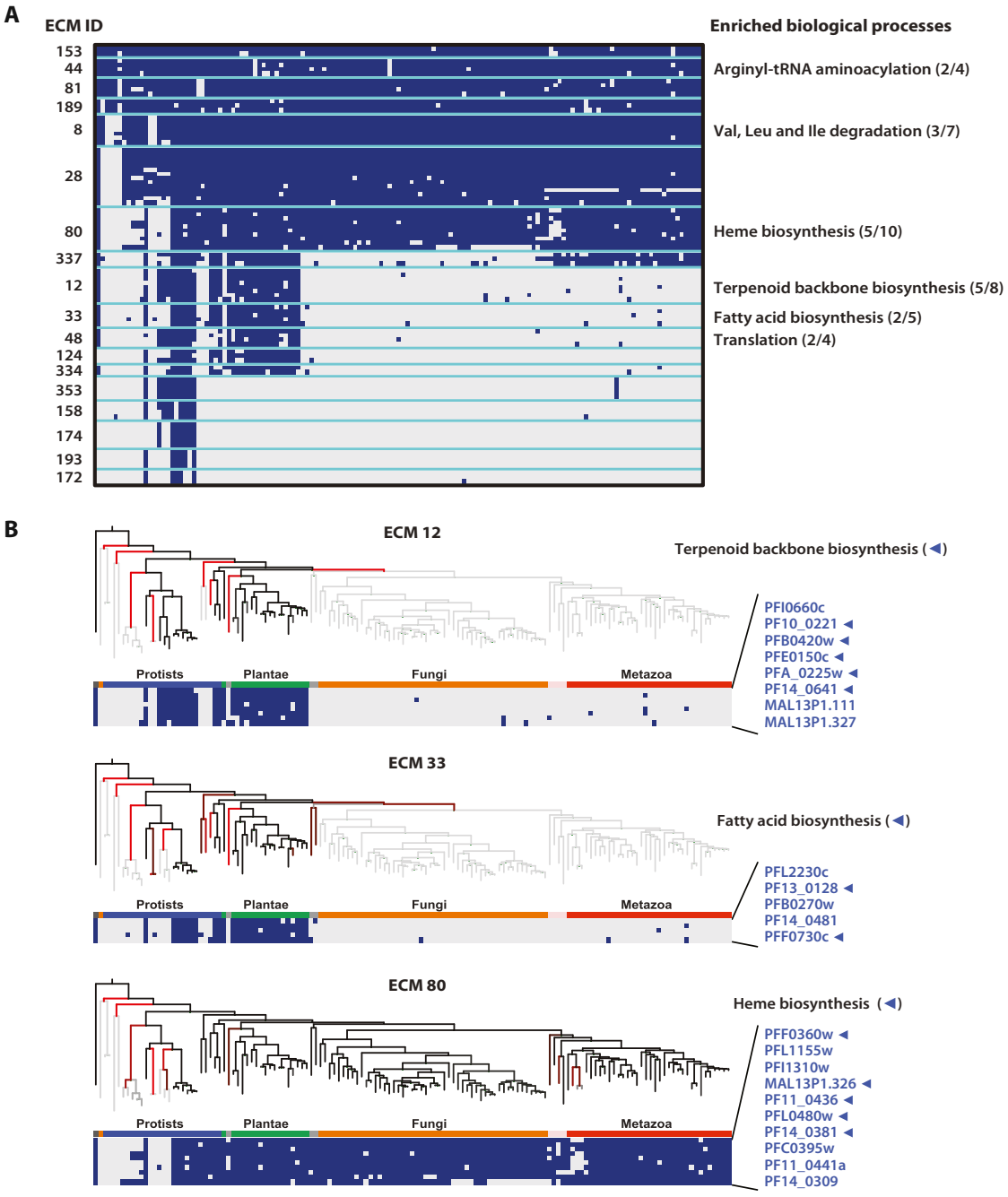

Figure S7. 18 Apicoplast-Enriched ECMs from *P. falciparum* ECMs, Related to Figure 7

(A) 18 ECMs enriched for GO cellular component "Apicoplast" (cumulative hypergeometric $p < 0.01$). The ECM IDs of the ECMs are listed on the left, and the other biological processes enriched for each ECM are listed on the right. The gene function annotations are from GO biological processes and KEGG metabolic and signaling pathways (Ashburner et al., 2000; Kanehisa et al., 2006).

(B) Apicoplast-related ECMs enriched for Terpenoid backbone biosynthesis (ECM12), fatty acid biosynthesis (ECM33) and heme biosynthesis (ECM80), including the independent loss events (red branches), the phylogenetic profile for the ECM genes (blue/white matrix and blue text).