Supporting Information for

AyurPhenoClusters define common molecular roots for rare diseases and resolves complex ciliopathies

Aditi Joshi, Deepika Jangir,Ashish Sharmaa, Tannay Anand, Hemendra Verma, Sandeep Kumar, Shipra Girdhar, Manvi, Pallavi Joshi, Nupur Rangani, Ravi Pratap Singh, Rakesh Sharma, Abhimanyu Kumar, Lipika Dey, Mitali Mukerji

Mitali Mukerji & Lipika Dey

E-mail: mitali@iitj.ac.in, lipika.dey@ashoka.edu.in

This file includes:

1. **SI Methods**
2. **SI Figures S1-S4**

1. SI Methods

Identification of clusters of diseases based on doshic proportions using Expectation Maximization (EM) algorithm

Examination of the dataset revealed that the disease to HPO association is probabilistic in nature. The dataset shows that for a single disease, the associated features may or may not have been recorded for all the patients. This is quite understandable as not all patients who suffer from the same disease report all the symptoms, nor do doctors look for an exhaustive set of symptoms for a particular disease.

Expectation-Maximization (EM) clustering is a powerful algorithm used for probabilistic clustering, especially in situations where the data may comprise multiple underlying distributions. This algorithm iteratively finds local maximum likelihood parameters of the underlying distributions. The distributions may involve latent variables and unknown parameters which have to be inferred from given data observations. The method can support missing values in the data, which was highly suitable for our purpose. EM algorithm assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. The number of clusters may be specified a priori or determined by the algorithm through cross validation.

The key steps of the algorithm are explained below:

- **Initialization:**
    - The parameters of the Gaussian distributions i.e. the means, covariables and mixture weights are initialised with random values. The input value of the cluster can be given as an input or as in our case, the algorithm can find it iteratively through exploration and using the convergence properties.
- **Expectation Step (E-step):**

- For each data point xi, the probability ric , typically referred to as “responsibility” that it belongs to a Gaussian distribution c is computed using the multivariate Gaussian probability density function as follows

$$r_{ic} = \frac{\pi_c N(x_i \mid \mu_c, \Sigma_c)}{\sum_{k=1}^{K} \pi_k N(x_i \mid \mu_k, \Sigma_k)}$$

where K is the total number of clusters, πc is the mixing coefficient of the weight for the Gaussian distribution c, which was initialised in the earlier step, and N(x|μ, **Σ**) describes the probability density function (PDF) of a Gaussian distribution with mean μ and covariance **Σ**, with respect to datapoint x. N(x|μ, **Σ**) is computed as given below:

$$N(x_i, \mu_c, \Sigma_c) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_c|^{\frac{1}{2}}} exp(-\frac{1}{2}(x_i - \mu_c)^T \Sigma_c^{-1}(x_i - \mu_c))$$

The responsibility measures how much the c-th Gaussian distribution is responsible for generating the i-th data point.

- **Maximization Step (M-step):**

  In the M-step, the algorithm uses the responsibilities of the Gaussian distributions computed earlier to update the estimates of the model’s parameters.

- The weights πc, the means μc and the covariance **Σ**c are updated using the following equations -

$$\pi_c = \frac{\sum_{i=1}^{m} r_{ic}}{m}$$

$$\Sigma_c = \frac{\sum_{i=1}^{m} r_{ic}(x_i - \mu_c)^2}{\sum_{i=1}^{m} r_{ic}}$$

- The updated estimate is used in the next E-step to compute new responsibilities for the data points.

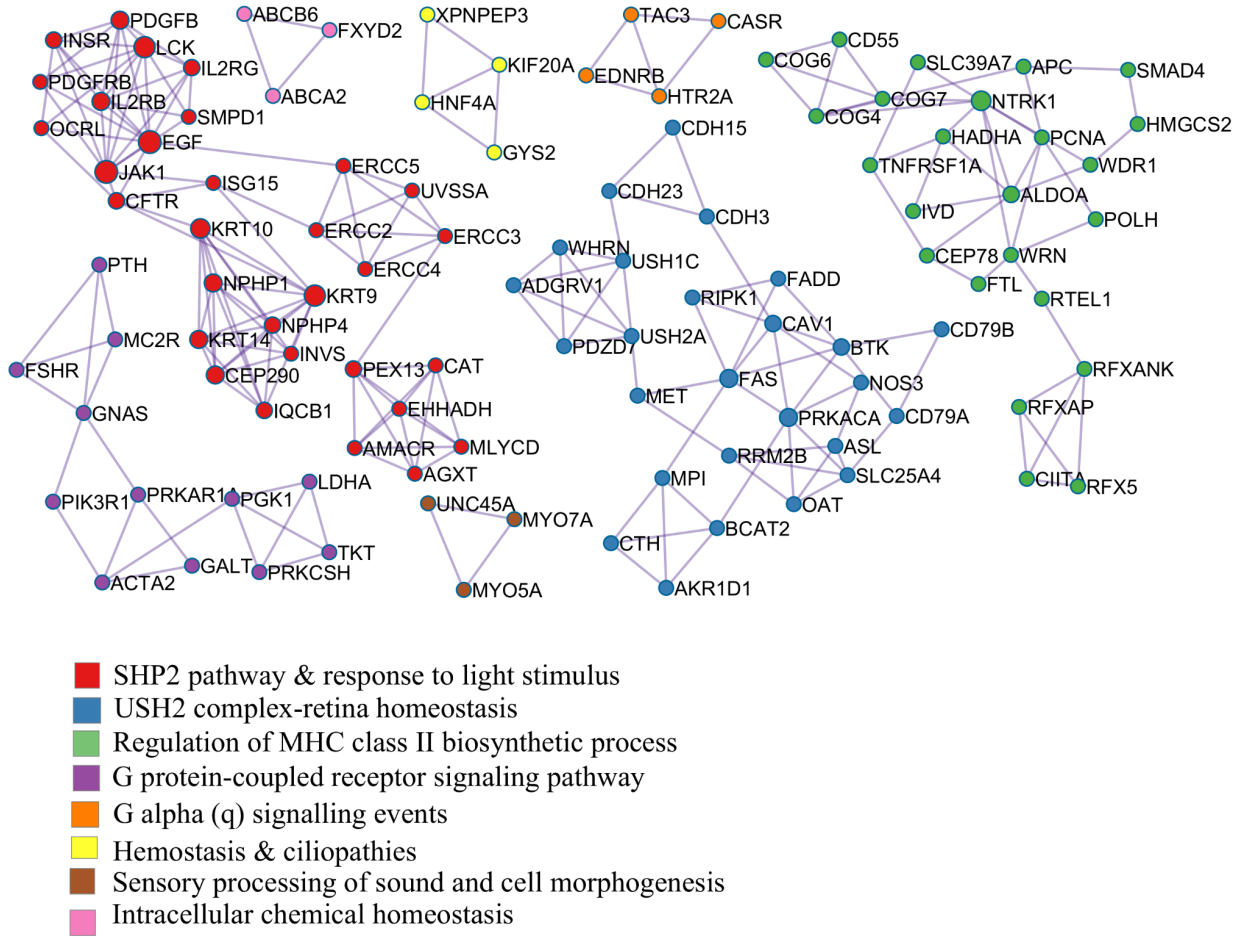
- The E-step and M-step are iteratively repeated till either a convergence or maximum number of iterations is reached.
- Checking for Convergence
  - Convergence is checked by evaluating the change in log-likelihood of the data, using the equation given below

- $L(\Theta) = \sum_i \sum_k r_{ik} \log(\pi_c * N(x_i \mid \mu_k, \Sigma_k))$

If the change falls below a specified threshold then it is said to converge. Once convergence is reached, the final parameters of the Gaussian mixture model are accepted as estimates of the underlying distributions. We have used the Weka package for implementing the EM algorithm. The exercise was repeated thrice with different initializations.

2. SI Figures

Fig S1(a): The modules represented belong to the cluster C1. This cluster is a mix type with sharing characteristics from Vata and Pitta. The processes shown here includes immune response, homeostasis and transport of small molecules



SHP2 pathway & response to light stimulus
USH2 complex-retina homeostasis
Regulation of MHC class II biosynthetic process
G protein-coupled receptor signaling pathway
G alpha (q) signalling events
Hemostasis & ciliopathies
Sensory processing of sound and cell morphogenesis
Intracellular chemical homeostasis

Fig S1(b): The modules represented belong to the cluster C2. This cluster is dominated with characteristic Pitta features. The processes observed here majorly include immune and inflammation along with cell activation and signaling.

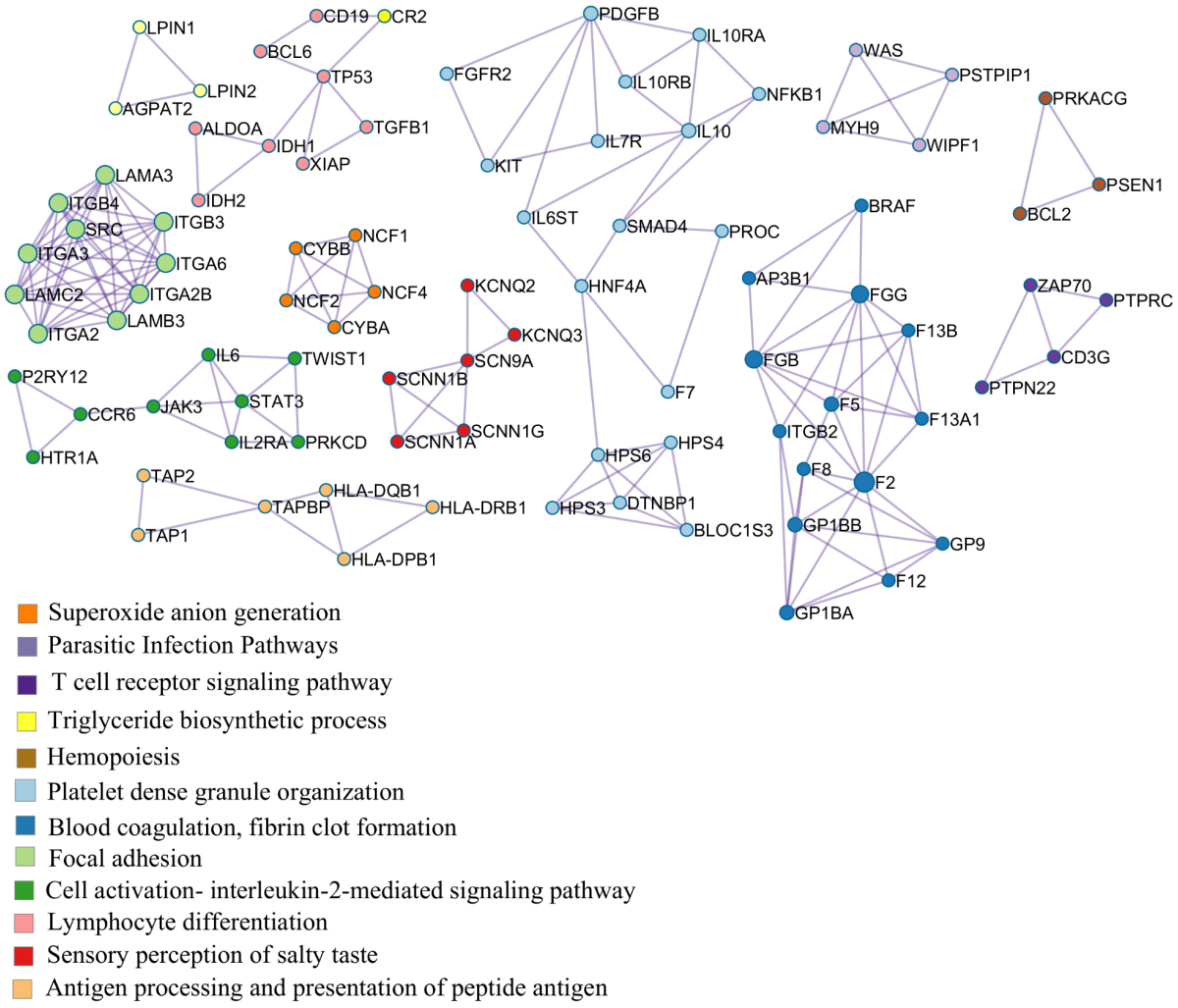


Fig S1(c): The modules represent the cluster C3. This cluster is dominantly Pitta with a small share of features from Vata. The processes here include inflammation and telomere activity.

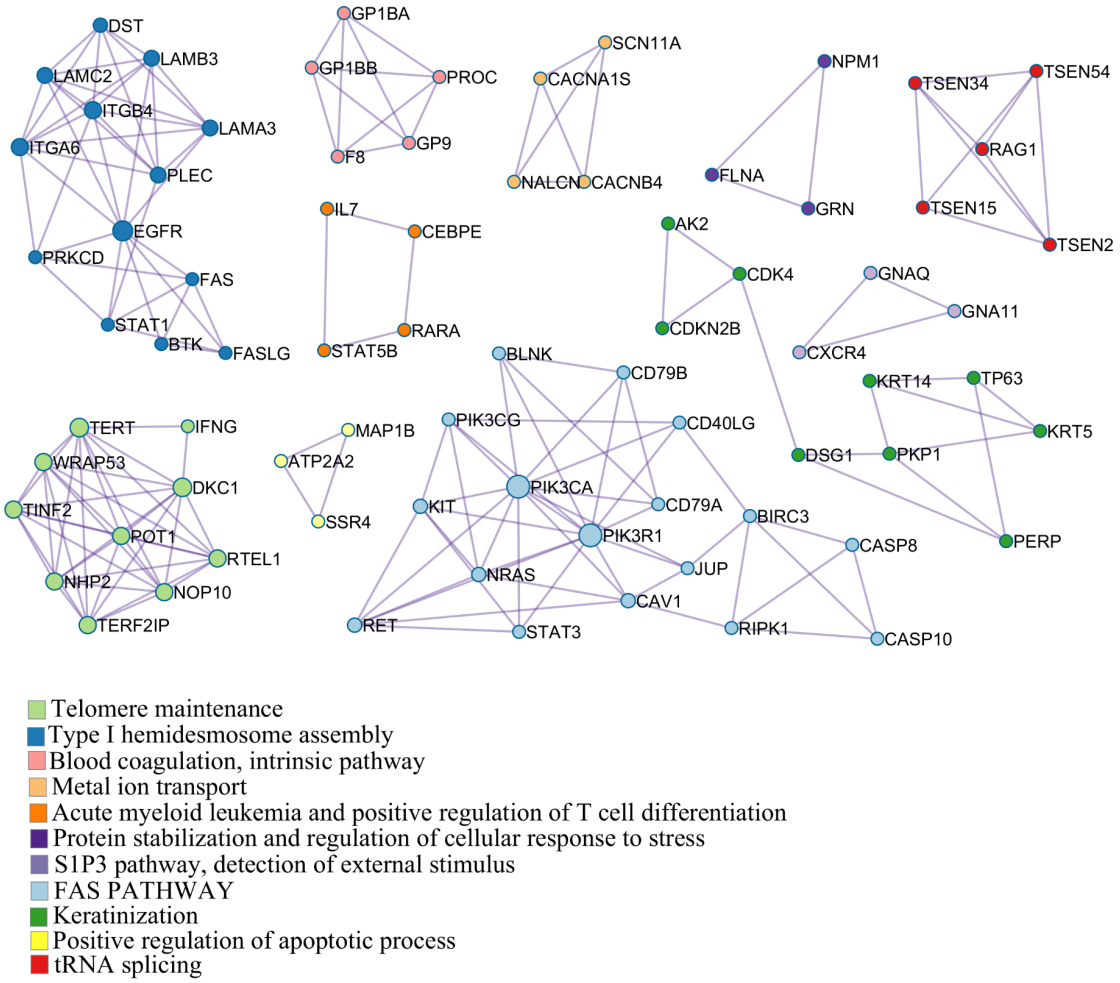


Fig S1(d): The modules represent the cluster C4. The cluster is dominantly kapha. Hence the modules are associated with ciliopathies, visual perception and metabolism.

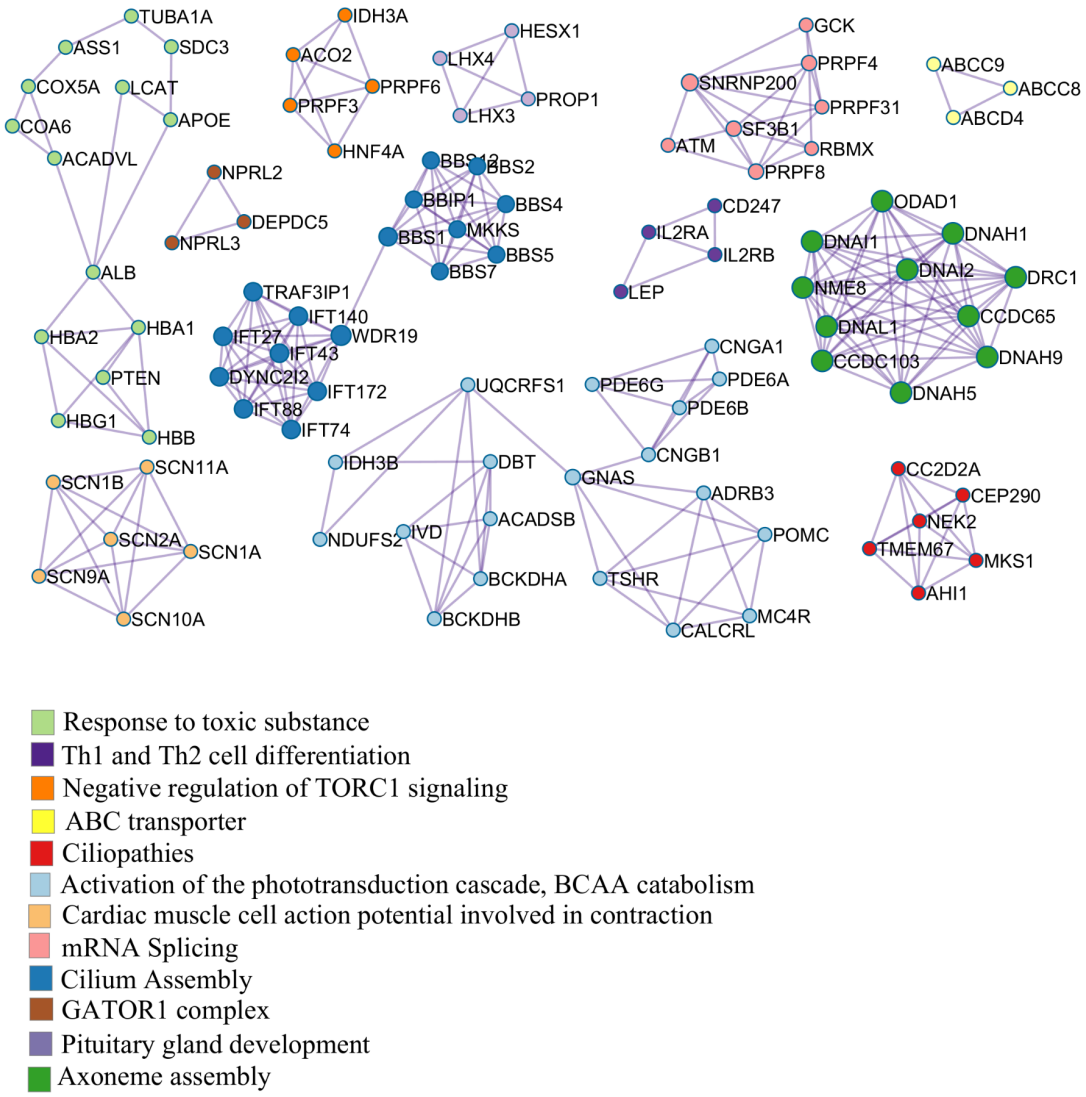


Fig S1(e): The modules represented are from cluster C5. This cluster is highly rich with characteristics of Vata. The modules for this cluster are distinct from other clusters sharing characteristics of Vata. The distinct processes associated with this cluster are DNA damage, DNA damage response, chromatin organization and regulation of cell cycle.

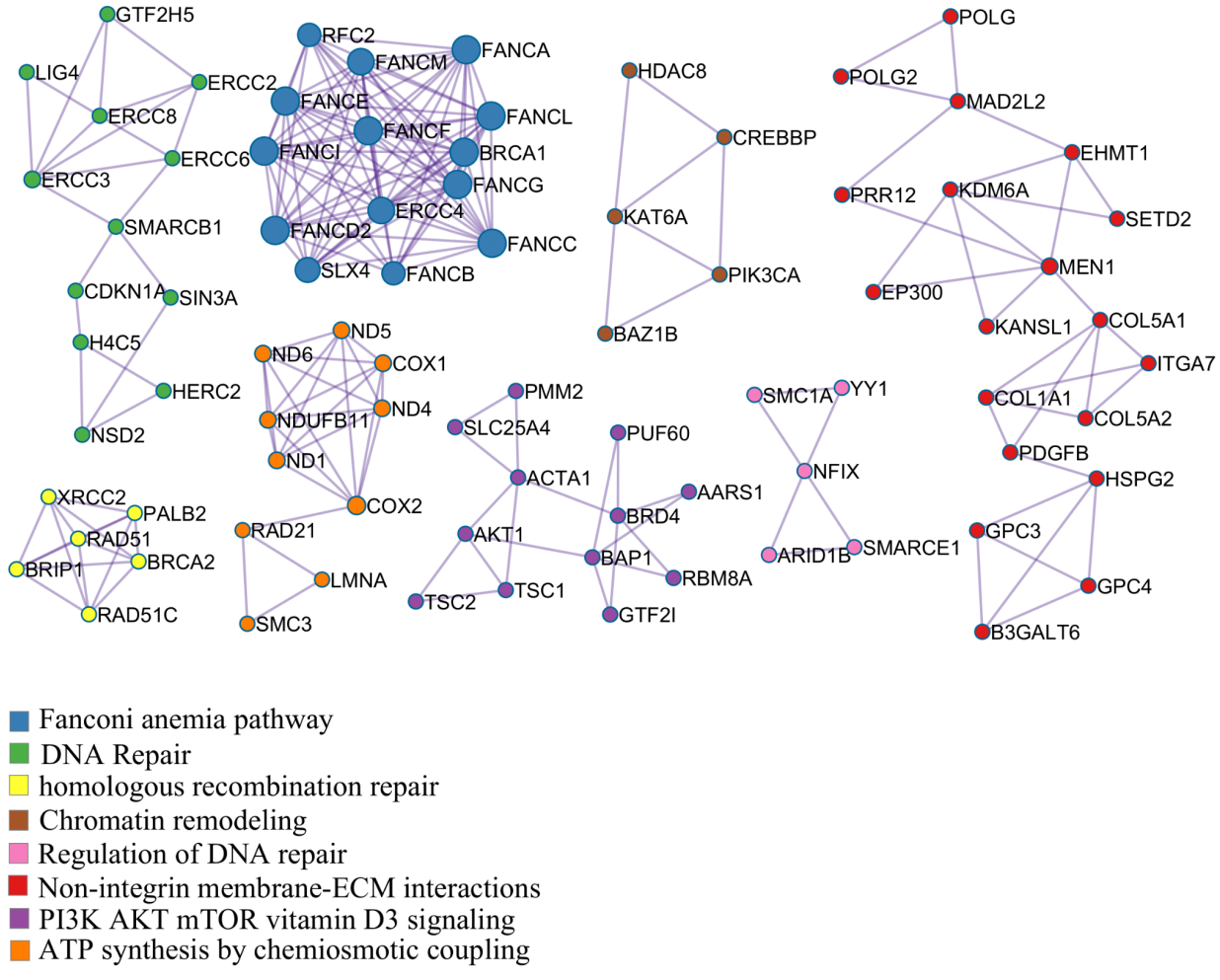


Fig S2: This figure shows the network for all ciliary genes from all the clusters, this contains 5 major hubs of pathways most of them are dominated by Cluster C4 & C0.

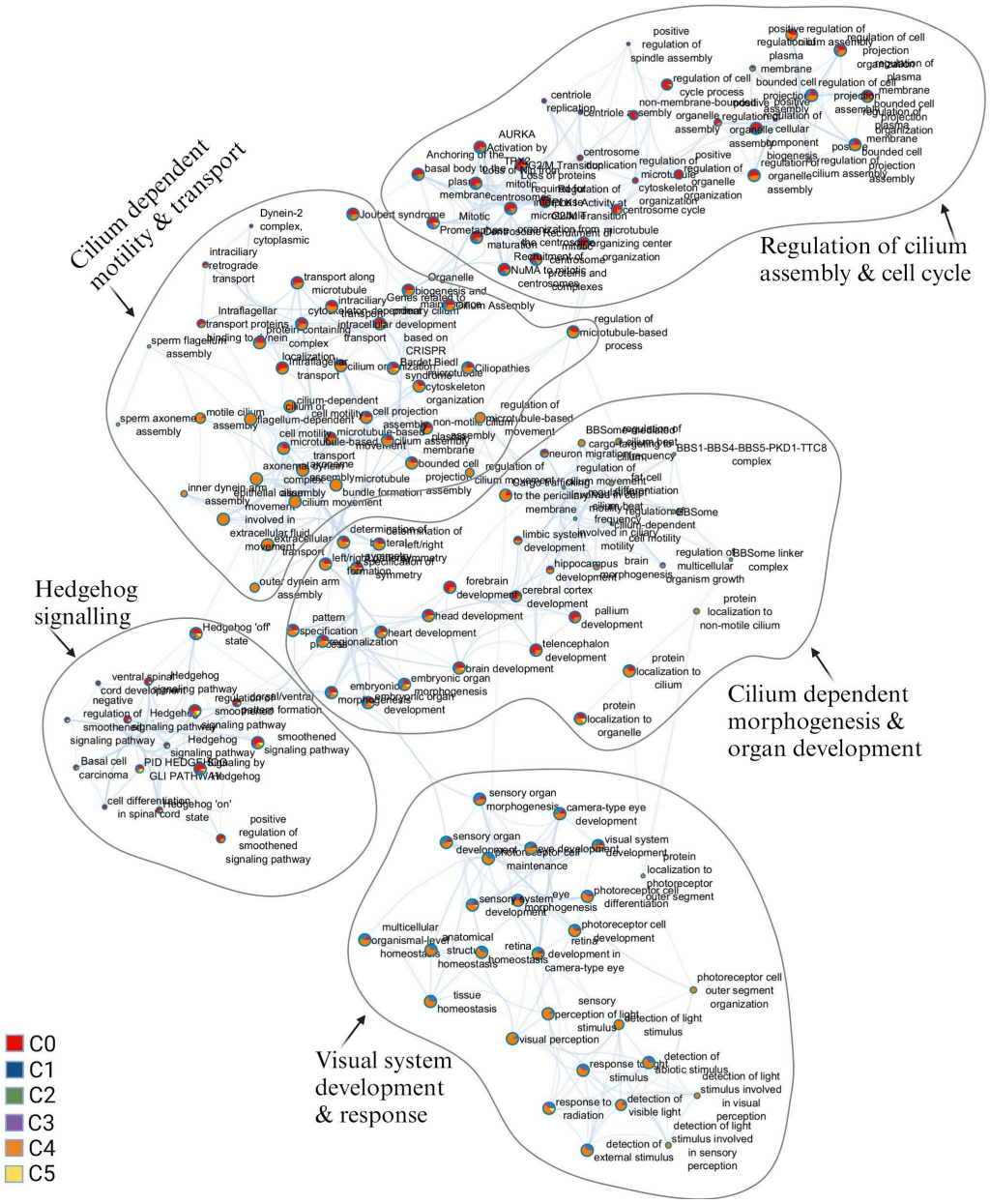


Fig S3(a): This figure shows the contribution of ciliary and remaining genes from cluster C4 in cellular pathways. The network has two inter-connected hubs, one cilium assembly & functions, which is dominated by overlapping ciliary genes, and other cellular responses to biomolecules dominated by non- overlapping genes.

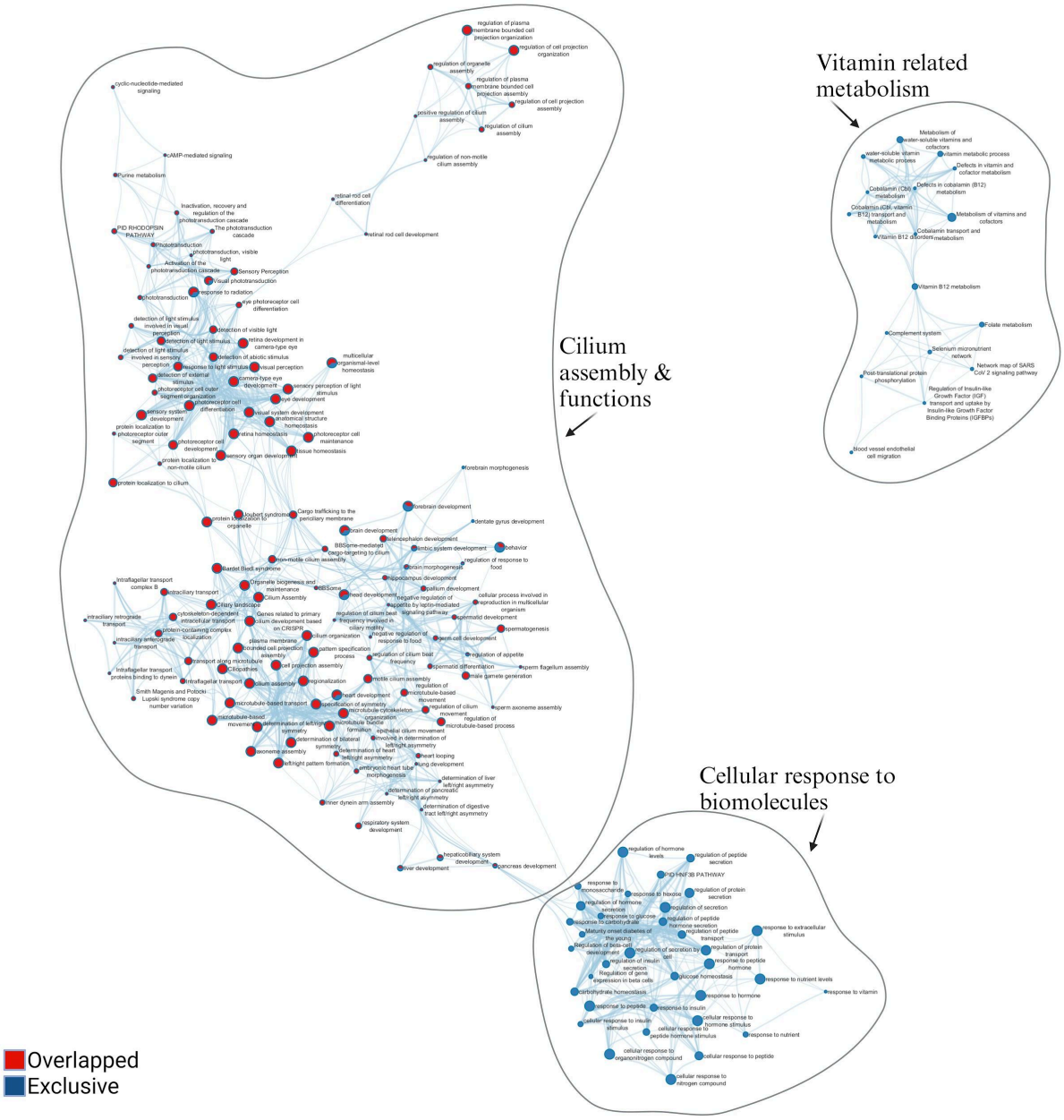

Vitamin related metabolism
Cilium assembly & functions
Cellular response to biomolecules
Overlapped
Exclusive

Fig S3(b): A network analysis of cluster C0 overlapping ciliary known genes & non-overlapped both, showing the interconnectedness between cilium-related signalling and morphogenesis & organ development, dominated by ciliary genes and non-overlapped ones, respectively.

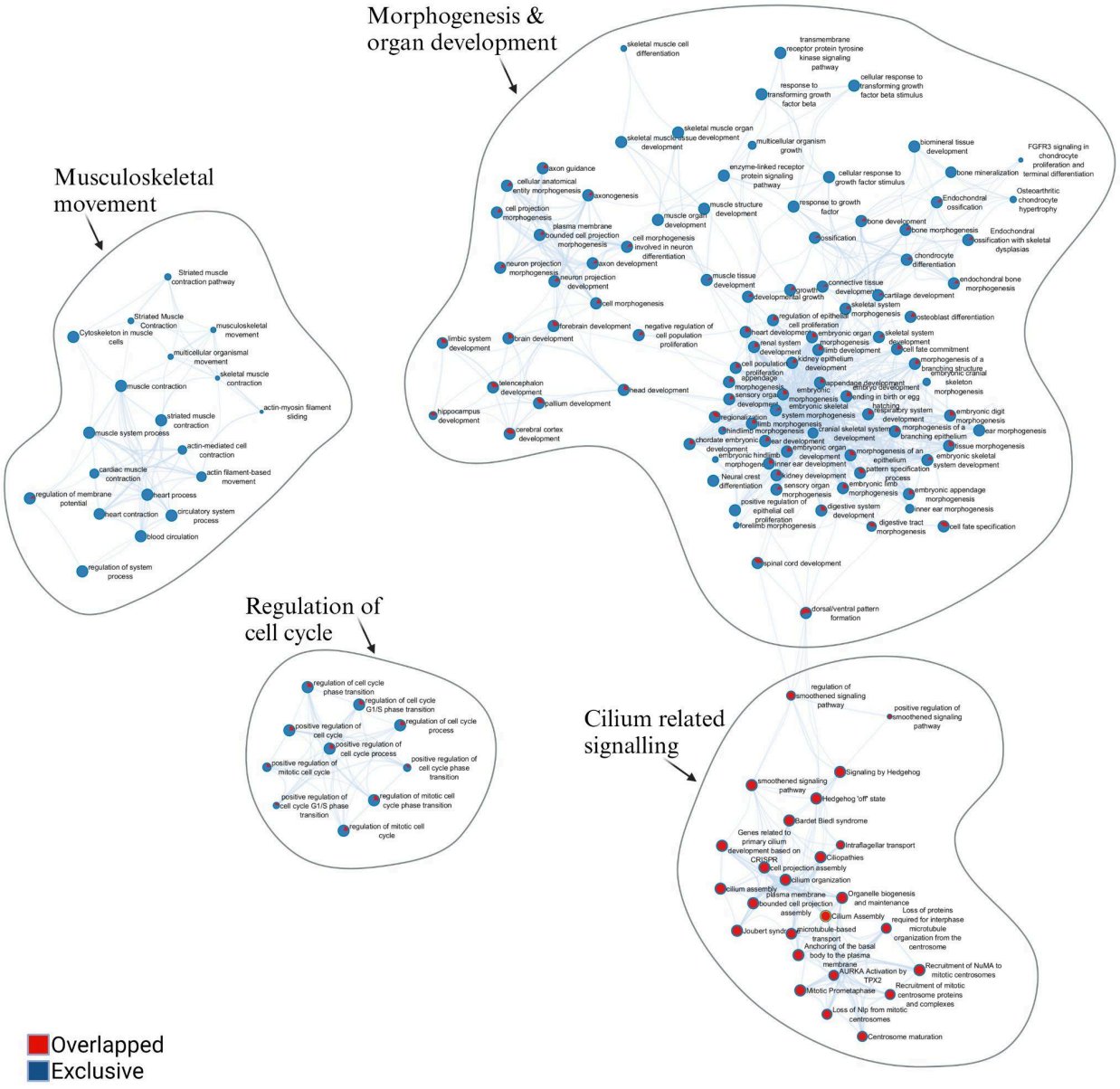


Fig S4: Network analysis of cluster C5. Three clusters i.e. C0, C1and C3 have been observed with characteristic Vata features in share with Pitta features. Cluster C5 is dominant in Vata features involved in distinct processes like DNA damage, DNA damage response, chromatin organization and regulation of cell cycle, which have been found absent in other three clusters.