Published on May 30, 2024 by Arcadia Science **DOI**: 10.57844/arcadia-74ad-345f

How can we biochemically validate protein function predictions with the Ras GTPase family?

We're using the well-studied superfamily of small monomeric GTPases, the Ras GTPases, to evaluate our structure-based clustering tool, ProteinCartography. We're seeking feedback on working with this protein family and determining which individual proteins to study.

Contributors (A-Z)

Prachee Avasthi, Audrey Bell, Brae M. Bigge, Megan L. Hochstrasser, Atanas Radkov , Dennis A. Sun, Harper Wood, Ryan York

Version 1 · Apr 01, 2025

Purpose

ProteinCartography is a tool for computational comparison of protein structures across species **[1]**. It uses the sequence and structure of an input protein to identify similar proteins. It then produces clusters of structurally similar proteins, displayed in an interactive map. We've outlined a rough plan to biochemically validate the two foundational hypotheses underlying the pipeline **[2]**.

The first step of this plan was to select protein families for analysis. We selected the Ras GTPase superfamily because it's previously been biochemically analyzed and because it presented many opportunities to test our foundational hypotheses **[2]**. Here, we present our ProteinCartography results for the Ras GTPases.

We'd like feedback on how we should select individual clusters and proteins and how we might test the function of this protein across species *in vitro*. We'd particularly love to hear from those who've studied Ras GTPases.

- This pub is part of the **platform effort**, "Functional annotation: mapping the functional landscape of proteins across biology." Visit the platform narrative for more background and context.
- This pub is part of our **validation strategy** series of pubs that starts with "A strategy to validate protein function predictions *in vitro*." We're also considering **deoxycytidine kinases** as an orthogonal protein family for validation. To learn more about them, visit the accompanying pub **[3]**.
- The **ProteinCartography pipeline** used to run these analyses is available in this GitHub repo. To create the custom overlays, we used this notebook and added our custom color dictionaries, which can be found in the associated Zenodo repositories.
- The **data** associated with this pub, including the full ProteinCartography analysis for the Ras GTPase family, can be found in this Zenodo repository.

Background

Why use RasGTPases?

For our first round of validation, we want to focus on protein families that will help us test our foundational hypotheses in a straightforward way. We started our search for candidate protein families by looking at the 200 most-studied human proteins in the Protein Data Bank, as these have likely been purified and biochemically studied

previously **[4]**. We first narrowed down this list by looking for proteins under 1,280 amino acids, as this is the cutoff that AlphaFold uses (as listed in the FAQ at the time of writing), and ProteinCartography uses structures from the AlphaFold database **[5][6]**. Each AlphaFold structure has per-residue confidence scores in the form of pLDDT scores, which approximate the amount of disorder in a protein's structure **[7]**. We chose to focus only on proteins with a mean pLDDT score over 80, which implies that the proteins are generally modeled well. Given that the ProteinCartography pipeline relies on AlphaFold for structural comparison, these cutoffs increased the chances that our structural predictions would be high-confidence. We next narrowed down the list by looking for proteins with commercially available assay kits.

We found that the Ras GTPases, namely HRas and KRas, not only fit these criteria **[2]** but also result in a ProteinCartography map that revealed clearly defined clusters that should let us test our hypotheses (Figure 1).

What do RasGTPases do and why are they important?

Ras GTPases are a well-studied superfamily of small monomeric GTPases that are key participants in myriad signal transduction pathways, including membrane trafficking, apoptosis, and cell differentiation **[8]**. In these processes, they function as binary molecular switches controlled by the action of GAPs (GTPase-activating proteins), which facilitate cleavage of the phosphate in GTP molecules, and GEFs (guanine exchange factors), which allow for rapid dissociation of the bound GDP **[9]**. The Ras superfamily includes the Ras, Rab, Ran, Rho, and Arf subfamilies **[8]**. Our analysis includes members from each of these subfamilies, but we're primarily focused on the Ras subfamily. The name Ras comes from the cancer-causing **Ra**t **s**arcoma viruses from which these genes were first sequenced **[10]**. Three human Ras genes encode Ras subfamily members: HRas, KRas, and NRas **[11]**. HRas and KRas are ranked 28th and 29th (respectively) in a list of the most-studied human proteins, so we've chosen to focus on them here **[4]**.

Mutations in Ras genes are implicated in up to 30% of cancers, as constitutively active Ras results in uncontrollable cell proliferation **[12]**. As such, many studies have aimed to reverse the constitutive activity of oncogenic Ras mutants. Despite a long-term reputation as "undruggable," recent focus on allele-specific inhibition of Ras has led to

multiple promising cancer therapeutics **[13]**. In 2021, the FDA approved the first KRas inhibitor, sotorasib, which binds mutated KRas as a therapy for KRas-related non-small cell lung cancer **[14]**. Alternative work has focused on inhibiting Ras–effector interactions and preventing activation of the signaling cascade **[13]**. Looking at Ras proteins across species could give us more information about the function of this master regulator, and a deeper structural and functional understanding of Ras proteins might inform further therapeutic avenues.

Diving into the ProteinCartography results for the Ras GTPase family

Running ProteinCartography on Ras GTPases

To identify similar proteins to our inputs and explore the structural variation in this protein family, we ran ProteinCartography analysis in "search mode" using human HRas and KRas as inputs (UniProt IDs: P01112 and P01116). ProteinCartography fetches similar proteins based on structure and sequence. It compares every structure to every other structure and generates TM-scores, or structural similarity scores, between each pair of structures **[15]**. It uses these to create interactive UMAP and t-SNE projections with overlaid Leiden clusters and metadata for exploration **[16][17][18]**. To learn more about how ProteinCartography works, visit our ProteinCartography pub **[1]**.

For this analysis, we requested 3,000 Foldseek hits, 7,000 BLAST hits, and 10,000 total structures for both inputs combined. This run generated 5,421 unique structure hits that the pipeline grouped into 12 clusters (Figure 1 and Figure 2, A). Both HRas and KRas are in LC00 (Figure 1 and Figure 2, A). Since HRas and KRas are very similar, we focus on just HRas in our downstream discussion. When we refer to the structural similarity of clusters to an input protein, we perform those calculations by comparing them to HRas alone (Figure 2, E).

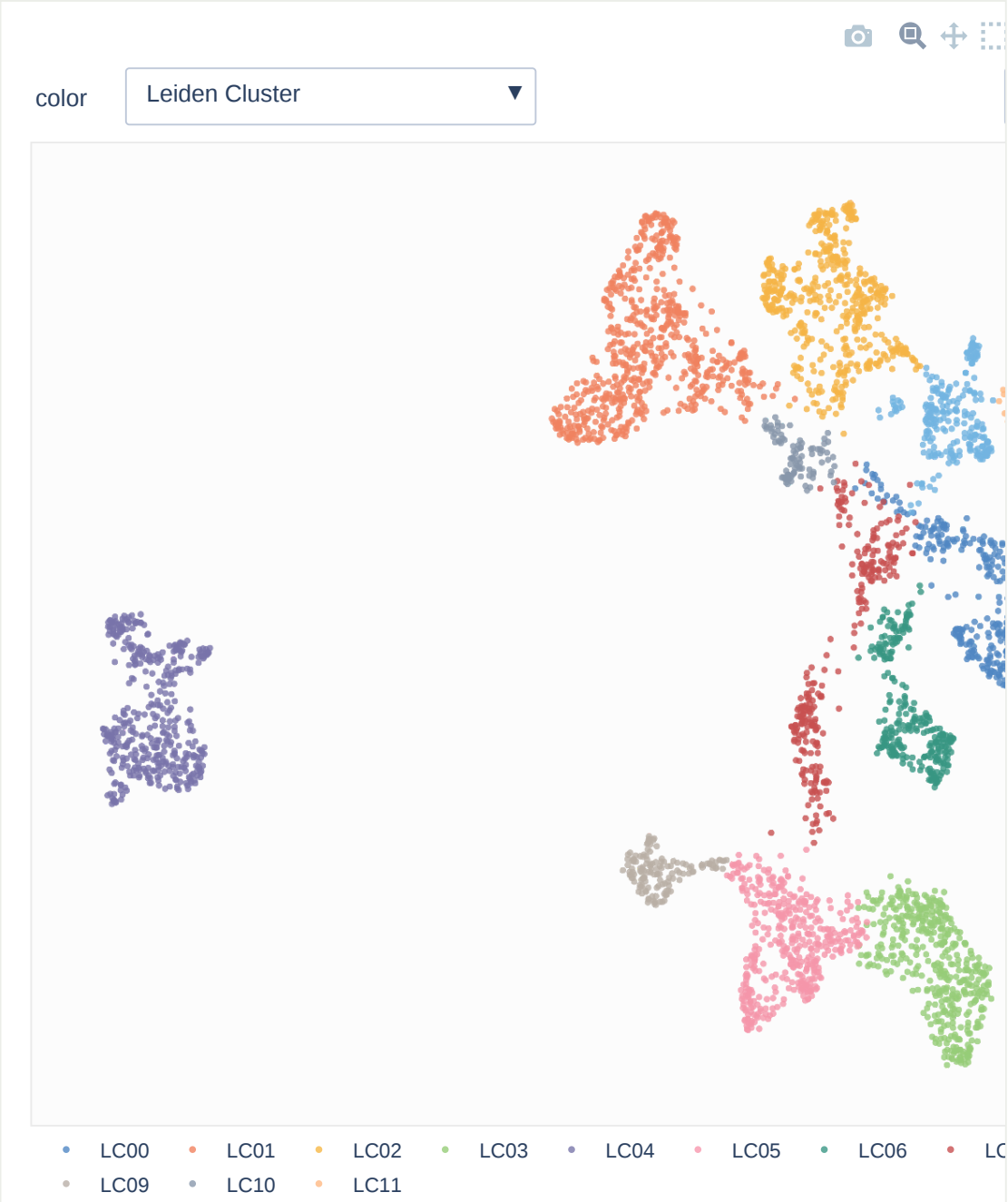


Figure 1

Interactive protein space with metadata overlays.

UMAP projection generated with ProteinCartography for human HRas. Our input protein, human HRas, is in LC00 and indicated by a four-pointed star. The overlays can be changed via the drop-down “color” menu.

A full list of all the proteins in this analysis, plus all the aggregated information from the pipeline is available in the aggregated features file linked below:

tsv GTPase_HRas_KRas_aggregated_features_pca_umap.tsv Download

Assessing compactness and overall quality

Our first step was to assess the cluster similarity matrix (Figure 2, B) for inter- and intra-cluster similarity. This can help us understand how well the clustering approach separated the proteins. These values are determined by calculating the mean TM-score of each protein in each cluster compared to every other protein in every cluster. The TM-score tells us how similar two protein structures are, with a score of 1 indicating the structures are identical **[15]**. The diagonal of the matrix represents how similar the structures of a cluster's constituent proteins are to each other, and the average of the diagonal is the "cluster compactness" score for the run. For the Ras GTPases, that value is 0.68. This indicates that most clusters are quite compact — in fact, all clusters except LC02, LC07, and LC10 have compactness scores over 0.6 (Figure 2, B). Additionally, some clusters show cross-cluster similarity (i.e., they have a high between-cluster mean TM-score), but many clusters appear distinct.

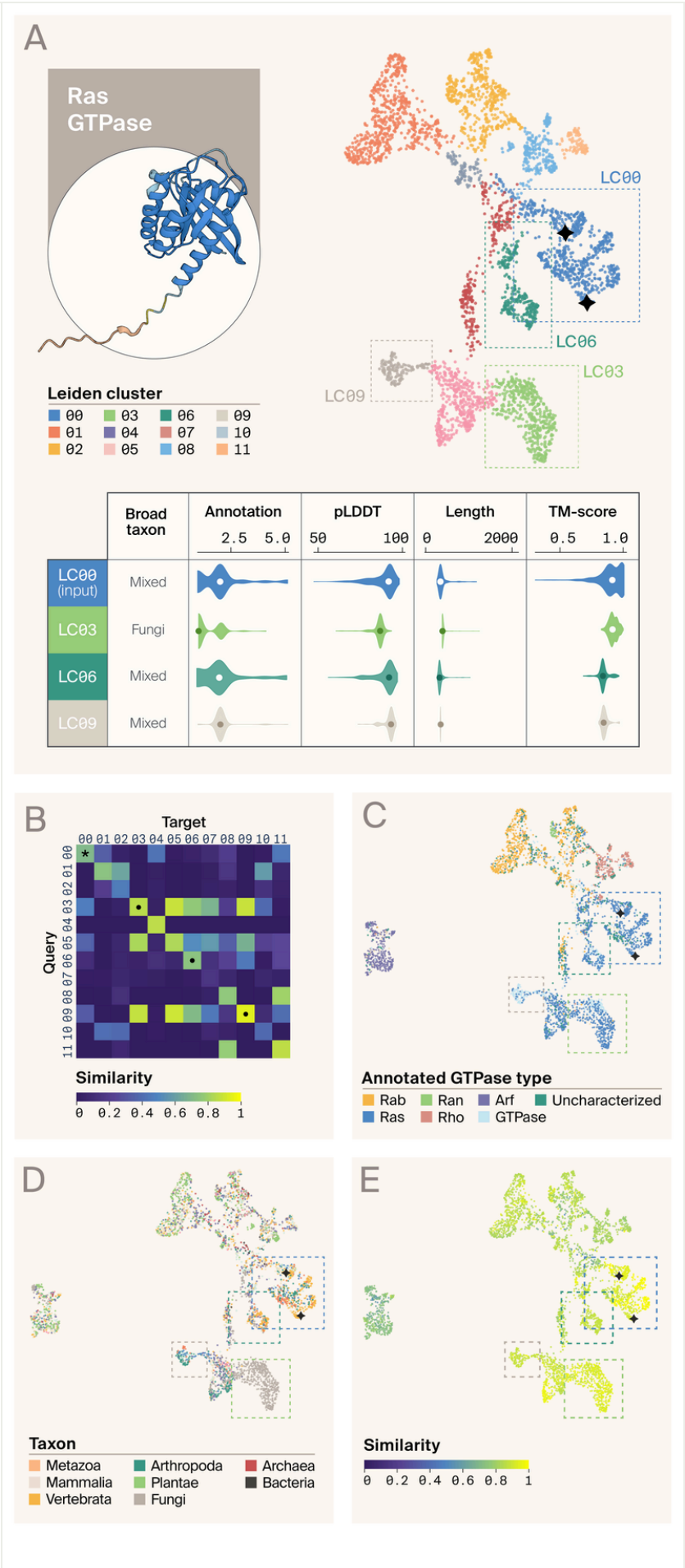
Next, we did a few quality checks on these outputs (Figure 1 and Figure 2). We first used the structural confidence, or mean pLDDT, overlay to assess the structure quality and the level of disorder of our output protein structures. In this case, the majority of the structures have mean pLDDT scores around 80 (Figure 1). This value gave us reasonably high confidence in the predicted structures and tells us that they likely don't contain large regions of disorder.

We next explored the TM-score overlay, which tells us the similarity of the fold of each output protein to the fold of the input protein (here, human HRas). This can also serve as a confidence metric. If our 5,421 hits were all very structurally similar to the input (only high TM-scores), we might lack enough variation to find functional differences between clusters. Conversely, if our hits were all extremely dissimilar (only low TM-scores), it might suggest that we haven't captured closely related proteins. We found a range of TM-scores, but overall this protein family had high TM-scores across the board. In this case, the lowest TM-scores were around 0.5 (found in LC04, the Arfs), which suggests even these structures adopt the same fold as our input (Figure 1 and Figure 2, E). LC00 has, on average, quite high TM-scores (around 0.92) — an

encouraging sign, as this cluster contains the input protein itself (Figure 1 and Figure 2, E). Once we confirmed that the outputs could yield informative results, we moved on to assessing the distribution of taxonomic origins, lengths, and annotation scores across clusters.

Exploring the data

In the following subsections, we walk through the most interesting clusters from our ProteinCartography analysis. We use the metadata overlays and semantic analysis to learn more about these clusters and to find proteins we can use to test our two foundational hypotheses about ProteinCartography



(that proteins within a cluster function similarly and those in different clusters function differently).

SHOW ME THE DATA: Our full ProteinCartography analysis for the Ras GTPase family is in this Zenodo repository (DOI: 10.5281/zenodo.11288430).

LC00: How does our input protein cluster?

We began by exploring LC00, which contains our input proteins, to assess if the outputs of ProteinCartography seem reliable and match what we'd expect. Taxonomically, LC00 mostly comprises

Figure 2

ProteinCartography outputs reveal interesting clusters for proteins with structural similarity to human HRas.

(A) The structure of human HRas, where orange indicates regions of higher disorder, alongside the UMAP projection with Leiden cluster overlay. Black diamonds indicate the locations of the input proteins (top, human HRas; bottom, human KRas). Note that LC04 is cropped out. Below the projection are violin plots showing the distribution of key values for each of our clusters of interest where the circle indicates the median value. White dots mean the median is below the threshold for significance, while filled-in dots denote significance in a Mann–Whitney *U* test. "Broad taxon" indicates taxonomic groups that are represented in each cluster. "Annotation" is the UniProt annotation score, or the relative confidence in each functional annotation (scale: 1–5). "pLDDT" is the confidence in the AlphaFold structural prediction for each structure (scale: 0–100). "Length" is the number of amino acids in each protein. "TM-score" is the similarity of each structure to that of human HRas (scale: 0–1).

(B) Cross-cluster similarity matrix. Each box represents the average TM-score (structural similarity) when comparing all structures in one cluster to all structures in another, where a higher score means the structures are more similar. The input cluster is marked with an asterisk (*) and our clusters of interest are marked with dots (•).

(C) UMAP projection with custom overlay showing existing annotation. Annotations were manually sorted

into the known subfamilies of the Ras GTPase superfamily.

(D) UMAP projection with taxonomic origin overlaid.

(E) UMAP projection with TM-scores (compared to the input protein) overlaid. TM-scores indicate higher structural similarity to human HRas.

(A, C–E) Dashed boxes mark our clusters of interest.

metazoa, vertebrates, and arthropods (Figure 2, D). The average TM-score, or structural similarity, of proteins in this cluster compared to human HRas is 0.92 (Figure 2, A), which suggests these proteins have extremely similar structures. Though the length of human HRas is only 189 amino acids, the average length for proteins in this cluster is 236 amino acids (Figure 2, A). This means that at least some proteins in this cluster are longer than the human protein. We could investigate whether these length differences within a cluster have meaningful effects on biochemical function. Although LC00 contains both of our well-annotated input proteins, the average annotation score for this cluster is 1.96 (Figure 2, A), which is still quite low and indicates plenty of room for discovery even within the input-protein-containing cluster. If we find that representative proteins from this cluster indeed share a function, it would support annotating all the proteins in the cluster as Ras GTPases.

LC03: Fungal homologs close in structure to human HRas

Our next focus cluster was LC03. While most of the clusters contain some combination of taxonomic origins, LC03 comprises entirely fungal proteins (Figure 2, A and Figure 2, D). The average TM-score (structural similarity to the input) for proteins in this cluster is quite high, at 0.93, implying that most of them adopt a highly similar fold to human HRas (Figure 2, A and Figure 2, E). The majority of these proteins are annotated as “Ras-like proteins” or “small monomeric GTPases,” though the average annotation score for the cluster is low — only 1.4 (Figure 2, A and Figure 3). The average length of proteins in this cluster is 226 (Figure 2, A). This is closer to the length of human HRas (189 amino acids) than the average length of the proteins that co-clustered with both HRas and KRas. The mean pLDDT, or structural confidence, for proteins in this cluster is 81.4, suggesting that these proteins have some regions of disorder (Figure 2, A). While this is within what we consider an acceptable range, it’s

lower than our other clusters of interest and it could point to these proteins having disordered regions and it may result in lower-confidence functional predictions.

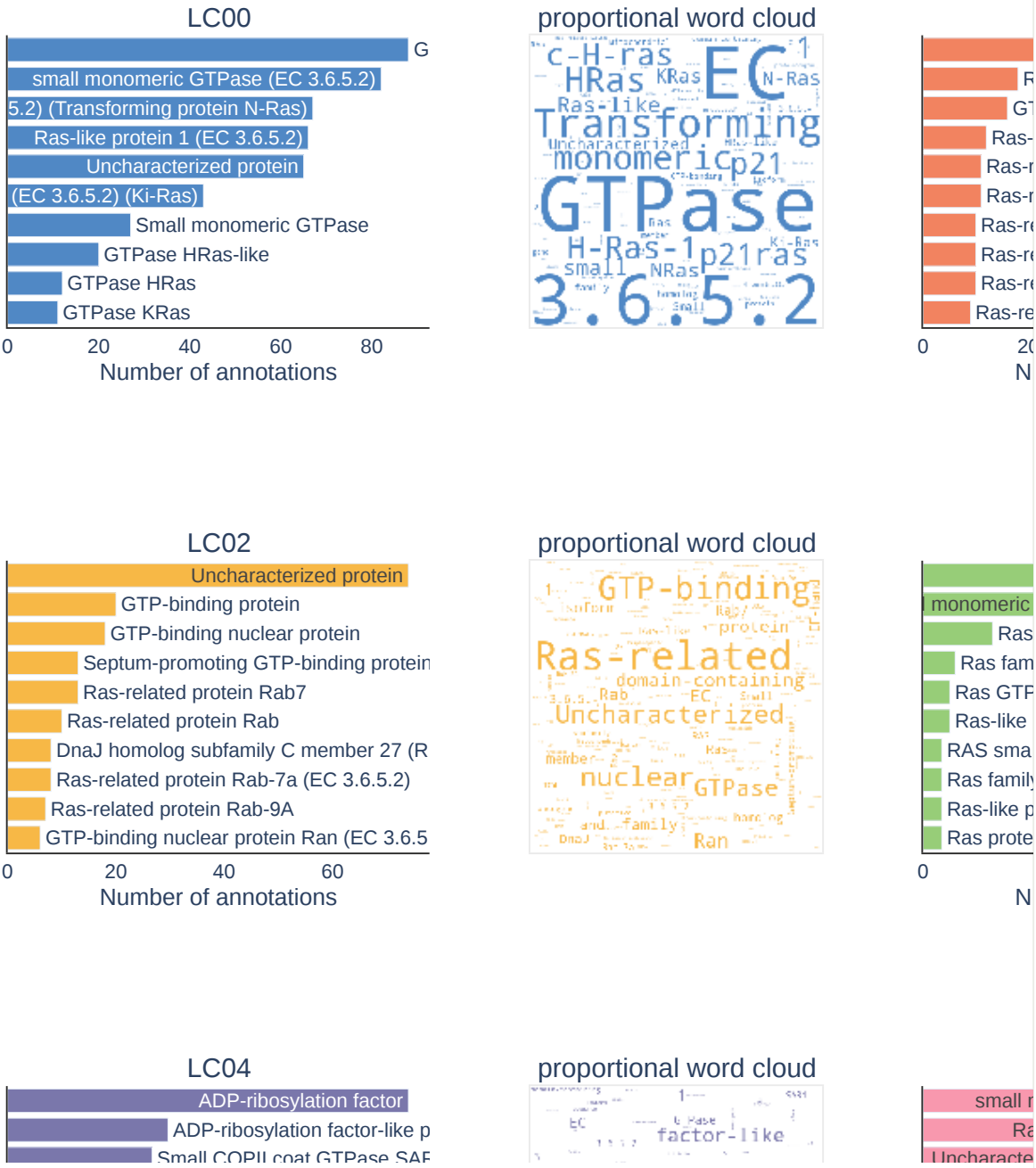


Figure 3

Interactive semantic analysis plot from ProteinCartography analysis of human HRas, showing the relationship between clustering and protein annotations.

The colors in the plot correspond to the colors of the Leiden clusters. For each cluster, the plot contains a ranked bar chart and a word cloud. The chart includes the top ten full annotation strings, while the word cloud shows the top individual words that appear in the annotations. Hovering over a bar in the chart displays the full annotation string and the number of proteins with that annotation.

LC06 and LC09: Taxonomically diverse homologs with generic annotations

After LC03, we explored LC06. This cluster contains proteins of mixed taxonomic origins, including fungi, vertebrates, and even a few archaea (Figure 2, A and Figure 2, D). Despite this apparent diversity, there are no plants or bacteria represented. The average TM-score for this cluster is 0.88; though lower than that of LC03, this still indicates that the proteins adopt the same general fold (Figure 2, A and Figure 2, E). The average length is 208 amino acids, slightly closer to the length of human HRas (189 amino acids) than either LC00 or LC03 (Figure 2, A). Interestingly, this cluster has an average annotation score of 2.0 (Figure 2, A), which is higher than we expected. This is because there are quite a few well-annotated proteins mixed in with many that are vaguely characterized or even entirely uncharacterized. The top annotation for this cluster is simply “small monomeric GTPase,” a descriptor shared by all members of the Ras superfamily (Figure 2, A and Figure 3).

Our final cluster of interest is LC09. The average length of proteins in this cluster is 188 amino acids, similar to the 189 amino acids length of human HRas (Figure 2, A). In many ways, LC09 is similar to LC06. This cluster, too, comprises proteins from mixed taxonomic origins, with especially high representation from arthropods and other ecdysozoans (Figure 2, A and Figure 2, D). There are two fungal proteins and quite a few proteins from rotarians, but no representation of plants or bacteria. Similar to LC06, the average TM-score of these proteins is 0.88 and their average annotation score is 2.1, suggesting the proteins share a fold with the input protein and that many proteins in this cluster have confident annotations (Figure 2, A). However, the top annotation for this cluster is the general annotation, “small monomeric GTPase” (Figure 2, A and Figure 3). Interestingly, the cross-cluster compactness matrix indicates that proteins in LC03 (all fungal proteins) and LC09 have highly similar folds to each other (Figure 2, B).

Overlaying annotation data

We produced custom metadata overlays to visualize trends between clusters. As mentioned, the Ras family is part of the Ras superfamily, alongside the Ras, Rab, Ran, Rho, and Arf families **[8]**. Did our clusters separate proteins into these well-known groups simply based on structural comparisons? We first assessed the semantic

analysis, an output of the ProteinCartography pipeline that provides the top annotations by cluster along with their counts. We saw that clusters tend to be composed primarily of a single subfamily (Figure 3). We then went through and manually categorized each protein into its subfamily (for example, we'd categorize a protein annotated as "mitochondrial Rho GTPase (EC 3.6.5.-)" as simply "Rho"). The file with manual annotation groups can be found here. Overlaying these general annotations on top of our Leiden clusters, we recognized some patterns that support this clustering strategy. First, each of the subfamilies cluster together quite well (Figure 2, C). For example, the Arf GTPases form a distinct cluster, LC04 (Figure 2, C). This is expected, as Arfs are generally less related to the other Ras GTPase family members **[8]**. Inspecting these more closely reveals that the Ran family clusters with the Rab family; this is also expected because Rans are generally considered part of the Rab family **[8]** (Figure 2, C). We also noticed "uncharacterized proteins" and vague annotations like "GTP-binding protein" throughout the map.

Summary

We'll be testing the hypotheses that proteins clustered together function similarly and proteins in different clusters have different functions. We can do so by comparing proteins within LC00, which contains our input protein HRas, and by comparing proteins from various additional clusters to those in LC00. Three candidate clusters jumped out at us for this analysis due to their high TM-scores and low annotation scores — LC03, LC06, and LC09. The high TM-scores suggest these clusters have captured proteins with strong structural similarity to human HRas, while their low annotation scores indicate that they are under-studied (particularly experimentally). If we can confirm their function in the lab, these are strong candidates for additional functional annotation. You'll have the opportunity to vote on a favorite research direction or comment with any further ideas below.

What do you think?

Do proteins within clusters function similarly?

Here are our ideas about how we might answer this question.

1. We could characterize uncharacterized proteins from the cluster containing our input protein to see if they have similar functions (in LC00). To start, we'll be testing their GTPase activity compared to human HRas.
2. We could also refine the current annotations of proteins that are annotated too broadly. Many proteins throughout the analysis are annotated as "GTP-binding protein" or "small monomeric GTPase."

Do these seem like reasonable approaches to test this hypothesis?

Do proteins in different clusters have different functions?

Here are the clusters we're considering to answer this question. We plan to compare proteins from these clusters to our input protein, which is in LC00. This cluster primarily contains metazoa, vertebrates, and arthropods.

1. LC03 contains all fungal proteins with a highly similar fold to human HRas. The top annotation for this cluster is "Ras-like proteins" or "small monomeric GTPases," but these annotations rank poorly in terms of quality and experimental support. By studying this cluster, we might learn why these proteins cluster separately from the input protein even though their fold is so similar.
2. LC06 has mixed taxonomic origins, but lacks plants and bacteria. The structures of proteins in this cluster are also highly similar to human HRas, although slightly less than those in LC03. Though the annotations in this cluster have slightly higher confidence than LC03, there are still many proteins that are uncharacterized or

vaguely annotated. Like LC03, we'd be interested in understanding why these structurally similar proteins cluster separately from the input protein.

3. LC09 has mixed taxonomic origins but includes many arthropods. The structures of proteins in this cluster are about as similar to human HRas as those in LC06. Additionally, these proteins are generally shorter than the other two clusters, similar in length to human HRas. The proteins in this cluster are primarily annotated as "GTP-binding protein" or something similarly generic. In addition to learning why these proteins cluster separately from the input cluster, we could look into why LC03 and LC09 cluster separately from each other even though they seem to share a fold.

How should we approach working with Ras GTPase proteins *in vitro*?

Once we select individual clusters and proteins, we'll purify each protein and test its GTPase activity using an *in vitro* assay.

Are there tips/tricks/challenges to biochemical analysis of Ras GTPase?

Do you have ideas for functions or mechanisms of Ras GTPases that we might want to test other than or in addition to intrinsic GTPase function?

Additional methods

We used ChatGPT to suggest wording ideas and then chose which small phrases or sentence structure ideas to use.

Next steps

We're seeking feedback on selecting individual clusters and protein families for further analysis *in vitro*. We aim to characterize the biochemical activity of a handful of these proteins to test our overall hypotheses about how ProteinCartography clusters proteins. However, there are additional analyses we can tackle in the meantime that might tell us more about this protein family.

Align functional data in the literature with ProteinCartography clustering

Because this protein family has been studied extensively, we wondered if we might find information in the literature about the biochemical function of the proteins in our analysis. Could we use the available data to help validate ProteinCartography and to help narrow down which proteins we bring into the lab?

There are several annotated, biochemically characterized Ras superfamily proteins that fit into the families we found in our analysis. We plan to curate available experimental data on Ras GTPase homologs and see how well this info aligns with our clustering.

Learn more about clusters and individual proteins by studying specific, conserved structural features

While ProteinCartography compares global protein structures, there's much we could learn by comparing specific aspects of the structures in this analysis. For example, we could look at surface vs. buried residues, electrostatics, topology, hydrophobicity, secondary structural elements, and more.

We know that the function of these Ras GTPases depends on binding GTP, GAPs, GEFs, and effectors. Because we know the regions responsible for each of these functions, we can look for conservation of these structural features across the family. By doing so, can we predict which GAPs and GEFs a given Ras GTPase interacts with? Can we predict if proteins from certain organisms are more or less susceptible to mutations that cause cancer in humans?

Summary

While we prepare for *in vitro* validation of ProteinCartography with Ras GTPases, we hope to use additional information from the literature and from the structures themselves to help us better understand the relationship between clustering and function.

References

1 Avasthi P, Bigge BM, Celebi FM, Cheveralls K, Gehring J, McGeever E, Mishne G, Radkov A, Sun DA. (2024). ProteinCartography: Comparing proteins with structure-based maps for interactive exploration. https://doi.org/10.57844/ARCADIA-A5A6-1068

2 Avasthi P, Bigge BM, Radkov A, Wood H, York R. (2024). A strategy to validate protein function predictions in vitro. https://doi.org/10.57844/ARCADIA-CAE9-96C4

3 Avasthi P, Bigge BM, Radkov A, Wood H, York R. (2024). How can we biochemically validate protein function predictions with the deoxycytidine kinase family? https://doi.org/10.57844/ARCADIA-1E5D-E272

4 Li Z, Buck M. (2021). Beyond history and “on a roll”: The list of the most well-studied human protein structures and overall trends in the protein data bank. https://doi.org/10.1002/pro.4038

5 Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. (2021). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. https://doi.org/10.1093/nar/gkab1061

6 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. (2021). Highly accurate protein structure prediction with AlphaFold. https://doi.org/10.1038/s41586-021-03819-2

7 Mariani V, Biasini M, Barbato A, Schwede T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. https://doi.org/10.1093/bioinformatics/btt473

8 Wennerberg K, Rossman KL, Der CJ. (2005). The Ras superfamily at a glance. https://doi.org/10.1242/jcs.01660

9 Vigil D, Cherfils J, Rossman KL, Der CJ. (2010). Ras superfamily GEFs and GAPs: validated and tractable targets for cancer therapy? https://doi.org/10.1038/nrc2960

10 Malumbres M, Barbacid M. (2003). RAS oncogenes: the first 30 years. https://doi.org/10.1038/nrc1097

11 Valencia A, Chardin P, Wittinghofer A, Sander C. (1991). The ras protein family: evolutionary tree and role of conserved amino acids. https://doi.org/10.1021/bi00233a001

12 Saxena N, Lahiri SS, Hambarde S, Tripathi RP. (2008). RAS: Target for Cancer Therapy. https://doi.org/10.1080/07357900802087275

13 Moore AR, Rosenberg SC, McCormick F, Malek S. (2020). RAS-targeted therapies: is the undruggable drugged? https://doi.org/10.1038/s41573-020-0068-6

14 Nakajima EC, Drezner N, Li X, Mishra-Kalyani PS, Liu Y, Zhao H, Bi Y, Liu J, Rahman A, Wearne E, Ojofeitimi I, Hotaki LT, Spillman D, Pazdur R, Beaver JA, Singh H. (2021). FDA Approval Summary: Sotorasib for *KRAS G12C*-Mutated Metastatic NSCLC. https://doi.org/10.1158/1078-0432.ccr-21-3074

15 Zhang Y, Skolnick J. (2004). Scoring function for automated assessment of protein structure template quality. https://doi.org/10.1002/prot.20264

16 Traag VA, Waltman L, van Eck NJ. (2019). From Louvain to Leiden: guaranteeing well-connected communities. https://doi.org/10.1038/s41598-019-41695-z

17 Belkina AC, Ciccolella CO, Anno R, Halpert R, Spidlen J, Snyder-Cappione JE. (2019). Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. https://doi.org/10.1038/s41467-019-13055-y

18 McInnes L, Healy J, Saul N, Großberger L. (2018). UMAP: Uniform Manifold Approximation and Projection. https://doi.org/10.21105/joss.00861