

SUBA Tutorial

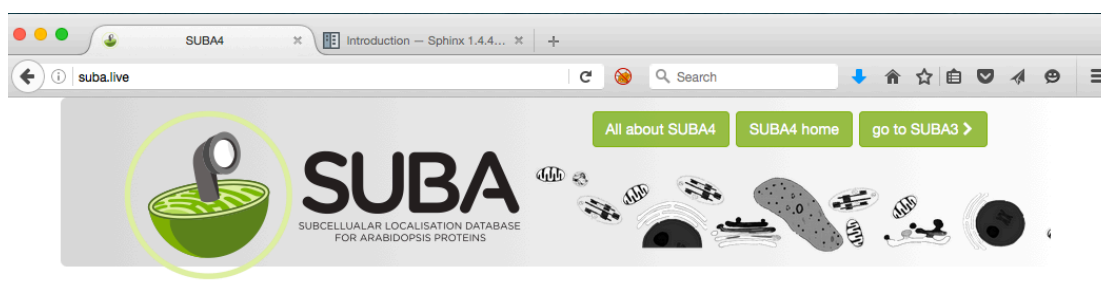
Version SUBA4 for using web portal access <http://SUBA.live>

SUBA4 HOME PAGE	2
TOP NAVIGATION BUTTONS	2
SUBA4 FUNCTION MENU	2
THE “WHAT IS SUBA?” INFORMATION BOX	3
THE SUBA4 NOTICE BOARD PANEL	3
SUBA4 STATISTICS	4
THE ‘QUICK SEARCH’ PANEL	4
THE BLAST PANEL	5
ABOUT SUBA4	6
SUBA4 SEARCH TAB	7
SUBCELLULAR LOCATION SEARCH TAB	8
PROTEIN PROPERTIES SEARCH PARAMETERS	10
PROTEIN-PROTEIN INTERACTION SEARCH PARAMETERS	11
AFFILIATIONS SEARCH PARAMETERS	12
BLAST SEARCH PARAMETERS	14
SUBA4 RESULTS TAB	15
SUBA4 TOOLBOX	16
THE MULTIPLE MARKER ABUNDANCE PROFILING (MMAAP) TOOL	16
THE COEXPRESSION ADJACENCY TOOL (CAT)	18
THE PPI ADJACENCY TOOL (PAT)	20
THE SUBA4 FACTSHEET	22

SUBA4 home page

Top navigation buttons

The very top of the SUBA4 homepage contains the welcome to SUBA4 banner and top hierarchy navigations buttons. The “All about SUBA” button leads to a broad overview of data and services contained in SUBA4. The “SUBA4 home” button gets you back to the home page from any other page in SUBA4. The “go to SUBA3” button leads to you to SUBA3 for finishing or checking work that was performed during the transition into SUBA4. About 12 months after introduction of SUBA4, SUBA3 will be decommissioned.



SUBA4 function menu

The SUBA4 functions are found underneath the banner and contain access points to the SUBA query builder (search) and web services (toolbox). Once a search query has been submitted the user will be guided to the “results” tab for viewing the query hit list. Other areas useful for users include the help and home.



Welcome to SUBA4

What is SUBA?

Proteins have specific functions and locations within the plant cell. They generate or are themselves products important for plant growth and response. In order to improve plants, protein function and location must be known. Protein subcellular location and the spatial relationship of proteins are important clues to function within the metabolic household. Subcellular location can be determined by fluorescent protein tagging or mass spectrometry detection in subcellular purifications as well as by prediction using protein sequence features. Location information helps represent plant cells as interacting protein networks that can be interrogated using SUBA.

For more information about data sets in SUBA, the subcellular locations reference standards, and SUBAtools go to [all about SUBA4](#)

Found SUBA useful? Please cite us. For an overview of publications on SUBA data and tools look at our [citation guide](#).

The “What is SUBA?” information box

The top information box on the home page gives a broad introduction to what SUBA4 is about and provides the link to the “all about SUBA4” site where users can find more information. You also find the link to the tutorial site, where there is a collection of tutorials about how to use different functions and services provided through the SUBA4 portal. The third link provided is leading the user to a citation guide explaining how to appropriately cite different data and services in downstream studies and applications.



Welcome to SUBA4

What is SUBA?

Proteins have specific functions and locations within the plant cell. They generate or are themselves products important for plant growth and response. In order to improve plants, protein function and location must be known. Protein subcellular location and the spatial relationship of proteins are important clues to function within the metabolic household. Subcellular location can be determined by fluorescent protein tagging or mass spectrometry detection in subcellular purifications as well as by prediction using protein sequence features. Location information helps represent plant cells as interacting protein networks that can be interrogated using SUBA.

For more information about data sets in SUBA, the subcellular locations reference standards, and SUBAtools go to [all about SUBA4](#)

Found SUBA useful? Please cite us. For an overview of publications on SUBA data and tools look at our [citation guide](#).

The SUBA4 notice board panel

The SUBA4 notice board contains important upgrades and news concerning SUBA. For releasing SUBA4 these news included the novel inclusions of tools and localisation data. It also indicates when experimental localisations were last updated and what data was used to train the SUBAcon version that is currently connected to the website.

SUBA4 Notice board

NEW data in SUBA4:
More localisations and experimental **suborganelle localisations** and protein-protein interaction localisations
The **SUBA4 toolbox** for estimating organellar proteins abundance (MMAAP tool) and protein-protein location relationships

Looking for crop protein localisations? New!!! Our subcellular localisation resource **cropPAL** » for wheat, barley, rice and maize.

Updates
Bibliographic references were last updated: **30th June 2016**.
SUBAcon was last retrained using data up to: **30th June 2016**.

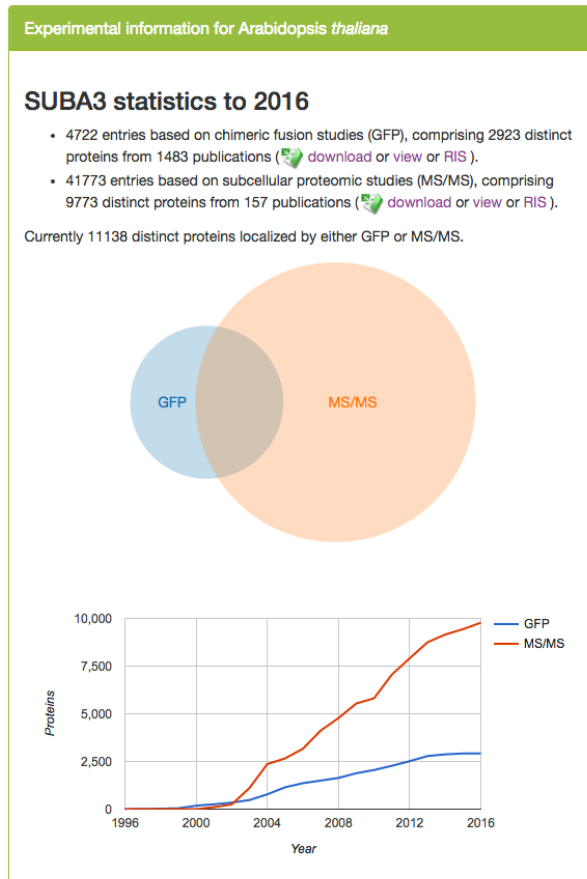
[Statistics](#)

Can't find what you are looking for? Email directly to Cornelia Hooper (cornelia.hooper [at] uwa.edu.au)

Try out an instant query for

SUBA4 statistics

You can gain immediate access to the localisation data using the “statistics” button. The display will show the data in SUBA accumulating over the year of publication. The images are downloadable using the download links in the pop up window.



Close

The ‘Quick Search’ panel

Using the quick search tab, the user can immediately start a search using an Arabidopsis Gene Identifier (AGI) or key words. Pressing the Query button will send the request to the database retrieving any hits for AGIs entered or for matches of the entered key words in protein description derived from TAIR10, titles or abstract of the SUBA4 literature.

Quick Search

Enter an AGI or text (keywords etc)

Retrieve hits

The BLAST panel

The panel labelled 'Find your closest AGI!' is a BLAST function, identical to the one found in the BLAST tab under the search function. The user can enter a sequence and retrieve data from Arabidopsis proteins with sequence similarity. The BLAST function will recognize A-Z plus '*'. The letters 'O' and 'U' will be converted to 'X' and 'N' will be recognized as 'unknown'. Gaps in the submitted sequence will be taken into account such as the best and 'longest' sequence matches are favourable.

The screenshot shows the 'Find your closest AGI!' interface. A text input field contains a protein fragment: MMRSRFLLFVFFSLFLFISSLIASDLGFCNEE...ISGEVESLARFAVDEHNKKNENLLEFARVWKAKEQVWAGTLHHLLTEILEAGQKKLYEAKVWVKPWLNFKELQEF...KQASDAPAITSSDLGCKQGEHESGWFEVPGI...DQRSNSLFPYELLEVHAKAEVTEAAKYNMLLKLKRGEEKEEFKVEVHKNHEGALHLNHAEQHHD. A dropdown menu is open, showing options: 'any', 'more than 200', 'more than 500', and 'more than 1000'. A red box labeled 'Choose BLAST score threshold' points to this dropdown. A red box labeled 'Enter your protein fragment' points to the text input field. A red box labeled 'Retrieve match' points to a 'Query' button. Below the input field, there is a small text box explaining the score: 'Score is $-\log_2(E)$ where $E = P_{\text{hit}} \times N$ is effective search space size. The larger the score the better. i.e. the p-value measures the statistical significance of the match but since we need to make a correction. (See here and here [PDF]).'

The BLAST hit AGIs are identified and the data for the AGI linked protein are retrieved from SUBA. The hits are displayed in the results tab and each hit shows the BLAST score and aligned protein sequence below the protein description.

The screenshot shows the results tab with a table of hits. The table has columns: AGI, SUBAcon, Predictions, Annotations, GFP, MS/MS, and PPI. The first hit is AT2G40880.1 with a score of 67.65. The second hit is AT3G12490.1 with a score of 64.18. The third hit is AT3G12490.2 with a score of 64.18. Red arrows point from the 'BLAST score' and 'BLASTed sequence' labels to the corresponding fields in the table.

AGI	SUBAcon	Predictions	Annotations	GFP	MS/MS	PPI
AT2G40880.1	extracellular	cytosol mitochondrion peroxisome plastid vacuole golgi endoplasmic reticulum extracellular	SwissProt: extracellular			
AT3G12490.1	cytosol	nucleus cytosol vacuole plastid mitochondrion	SwissProt: extracellular	cytosol	cytosol endoplasmic reticulum	AT3G56170.1: mitochondrion
AT3G12490.2	extracellular	cytosol mitochondrion plastid vacuole golgi endoplasmic reticulum extracellular	SwissProt: extracellular		cytosol	

About SUBA4

The “About SUBA4” contains a collection of descriptions, information and links to additional information about the data, functions and web services in SUBA4.

About SUBA4

What is SUBA?

SUBA provides a powerful tool to investigate subcellular localisation in Arabidopsis through the unification of disparate datasets and through the provision of web services through our accessible interface. Users can construct powerful queries or interrogate their protein sets resulting in a one-stop-shop for protein localisation and protein location relationships in the Arabidopsis model plant. **SUBA4** houses large scale proteomic, GFP localisation, Protein-Protein Interaction (PPI) data as well as PPI localisation data sets from cellular compartments of Arabidopsis. It also contains precompiled bioinformatic predictions for protein subcellular localisations and a consensus call taking predictive and experimental information into account. The **SUBA4 search** interface and **SUBA4 toolbox** provides flexible options of refining or interrogating protein data sets by location, expected abundance, interactions, co-expression, protein properties, bibliographic information.

Why SUBA? Subcellular localisation information can contribute towards our understanding of protein function, protein redundancy and of biological inter-relationships. While a variety of technologies are currently employed to determine the subcellular location of proteins, much of this information is not available in an integrated manner. In an attempt to get a clearer picture of our experimental data and to more generally understand subcellular partitioning we have brought together and expanded various data sources to build **SUBA**. The database has a web accessible interface for downstream applications.

More detailed description of SUBA

The resource in SUBA4

SUBA4 experimental data

Show me the SUBA4 experimental data statistics

SUBA4 tool box

The SUBA4 tool box is an interactive analysis tool (CAT). Linking the SUBAcon data to protein expression adjacency and purity of protein samples be estimated, predicted PPI data sets can be refined, spatial co-expression networks and more can be extracted by simply entering a set of AGIs and clicking a button. Try the **SUBA4 toolbox** now.

List of resources and services in SUBA4 with description and links to more details information

SUBA4 predictors

SUBA4 contains 22 predictors with use distinct training data sets, input variables and prediction methods. These have been reviewed and compared for their contribution to the consensus call in our recent study about **SUBAcon**. Predictors vary in their accuracy for each subcellular compartment. Using the table below you can use the most useful predictor for your target compartment. The classifier **SUBAcon** achieves the highest accuracy in any of the 10 subcellular categories.

SUBA consensus (SUBAcon) locations

Abundant experimental data from fluorescent protein (GFP) tagging or mass spectrometry (MS) are available for Arabidopsis, yet they only cover ~ 25% of the proteome. For the remaining 75% of proteins, many computational tools have been developed to predict proteome-wide subcellular location. None of the mentioned approaches are error-free and thus results are often contradictory. To help unify the multiple data sources contained in SUBA4, we have developed the SUBcellular Arabidopsis consensus (SUBAcon) algorithm, a naive Bayes classifier that integrates 22 computational prediction algorithms, experimental GFP and MS localizations, protein-protein interaction and co-expression data to derive a consensus call and probability. SUBAcon classifies protein location in Arabidopsis more accurately than single predictors. SUBAcon is a useful tool for recovering proteome-wide subcellular locations of Arabidopsis proteins. [More info about SUBAcon](#)


Arabidopsis Subcellular Reference (ASURE)

ASURE is the reference used for training SUBAcon and its built is described our recent study about **SUBAcon**. ASURE contains 5,393 proteins of which 2894 (53%) have been independently experimentally localized. Because experimental (GFP, MS) data were introduced in the SUBAcon classification algorithm, the assembly of ASURE sub-proteomes used additional inclusion criteria for curated ASURE proteins such as protein function and evidence from orthologues in other species. ASURE showed a discrepancy of less than 1% compared to the high-confidence Arabidopsis plastid proteome (Huang, et al., 2013) and to the peer-reviewed reference set used for training the classifiers MultiLoc2

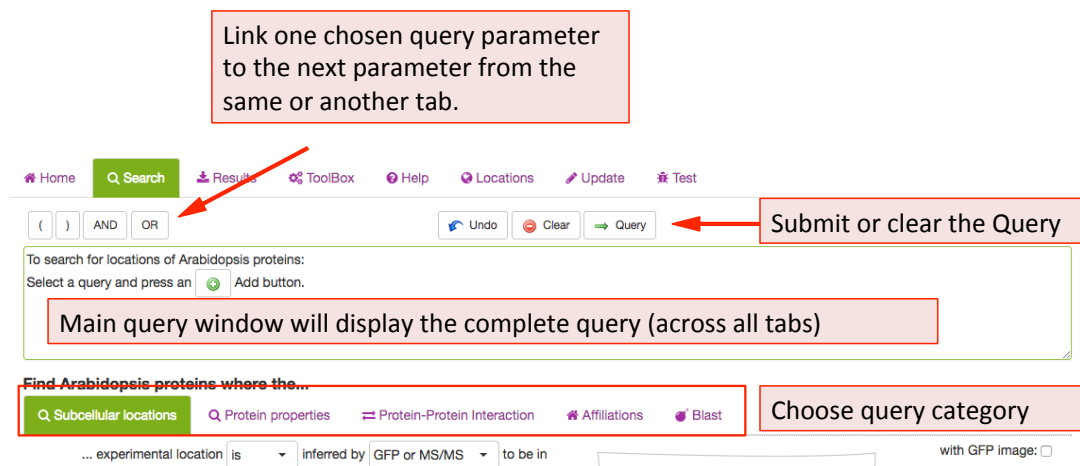
The “What is SUBA” provides a more detailed description of the purpose, collation and data provision intended by SUBA4. The ‘resource in SUBA4 panel provides a details description of each data set and data tool in SUBA including links to affiliated resources.

e.g. The user can find extra information about the predictive data sets. The performance for all predictors and SUBAcon for each compartment is given in the table. This may be useful for users in order to choose a best-suited predictor for their study purpose.

SUBA4 Search Tab

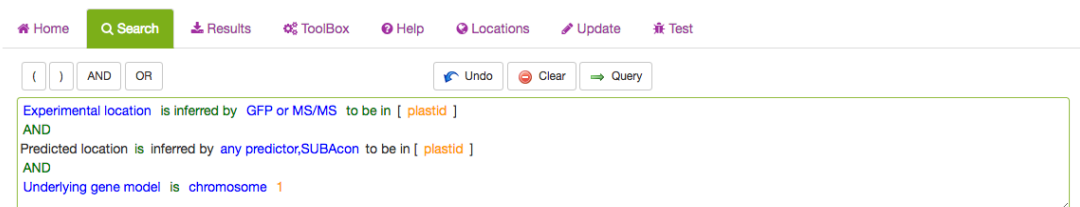
When clicking on the Search function the tab opens the query builder menu. In SUBA4 there are more options and categories of queries you can use to interrogate the SUBA data set. In order to enable the user to find the desired search parameters we have introduced search categories. Each category is stored under a tab. The user can choose a parameter from any tab and add it to the query. The query will appear in the query window at the bottom after clicking the  button. Different search categories can be combined using the AND/OR buttons in between parameters above the query window.

Link one chosen query parameter to the next parameter from the same or another tab.



The screenshot shows the SUBA4 search interface. At the top, there is a navigation bar with links for Home, Search, Results, ToolBox, Help, Locations, Update, and Test. Below this is a query builder menu with buttons for (),), AND, and OR. To the right of these buttons are buttons for Undo, Clear, and Query. A red arrow points from the 'Query' button to a text box that says 'Submit or clear the Query'. Below the menu is a main query window with a text input field and an 'Add button' icon. A red arrow points from the 'Add button' icon to a text box that says 'Main query window will display the complete query (across all tabs)'. Below the main query window is a 'Find Arabidopsis proteins where the...' section with a dropdown menu for 'Subcellular locations' and other categories like 'Protein properties', 'Protein-Protein Interaction', 'Affiliations', and 'Blast'. A red arrow points from the 'Subcellular locations' dropdown to a text box that says 'Choose query category'. Below this section are several dropdown menus for query parameters, including 'experimental location', 'inferred by', 'GFP or MS/MS', 'to be in', and 'with GFP image'.

The full query can be seen in the query window. For details about each category please see search category sections in the tutorial below.



The screenshot shows the SUBA4 search interface with the full query displayed in the query window. The query is: Experimental location is inferred by GFP or MS/MS to be in [plastid] AND Predicted location is inferred by any predictor,SUBAcon to be in [plastid] AND Underlying gene model is chromosome 1. The query is displayed in a text input field with a green border. The background shows the same navigation bar and query builder menu as in the previous screenshot.

Subcellular location Search Tab

This tab contains queries for limiting proteins based on their localisations. SUBA4 has 2 main categories of localisations. You can search for experimental localisations, which is the top query parameter.

The screenshot shows the SUBA4 search interface. At the top, there is a navigation bar with links for Home, Search, Results, ToolBox, Help, Locations, Update, and Test. Below this is a query editor with a search bar containing the query: "Experimental location is inferred by GFP or MS/MS to be in [plastid]". A red box with an arrow points to this query with the text "Check the full query".

Below the query editor is a section titled "Find Arabidopsis proteins where the...". It has several tabs: "Subcellular locations", "Protein properties", "Protein-Protein Interaction", "Affiliations", and "Blast". The "Subcellular locations" tab is active. There are two main sections for selecting parameters:

- Top section:** "experimental location is" (dropdown) "inferred by" (dropdown) "GFP or MS/MS" (dropdown) "to be in" (dropdown). To the right is a "with GFP image:" checkbox. Below this is a cell schematic with checkboxes for various organelles: nucleus, cytosol, mitochondrion, plastid, peroxisome, vacuole, Golgi, extracellular, endoplasmic reticulum, and plasma membrane. A red box with an arrow points to the "plastid" checkbox with the text "2. Choose the parameter". To the right of the schematic is a button with a plus sign and a minus sign, with a red box and arrow pointing to it.
- Bottom section:** "predicted location is" (dropdown) "inferred by" (dropdown) "any predicted" (dropdown). Below this is another cell schematic with similar checkboxes. A red box with an arrow points to the "plastid" checkbox with the text "3. Add parameter to query". To the right of the schematic is another button with a plus sign and a minus sign.


The parameters such as in/exclusion of particular compartments and methodology can be chosen from the drop down lists. For choosing a subcellular location, tick any of the box or structures in the cell schematic for the conventional SUBA location categories. For expanded suborganellar categories, click on the >> to expand the list.

This screenshot shows the same interface as the previous one, but with the "Subcellular locations" tab expanded. The "experimental location is" dropdown is set to "plastid". A red box with an arrow points to the ">>" button next to "plastid" with the text "Select >> to access suborganellar locations".

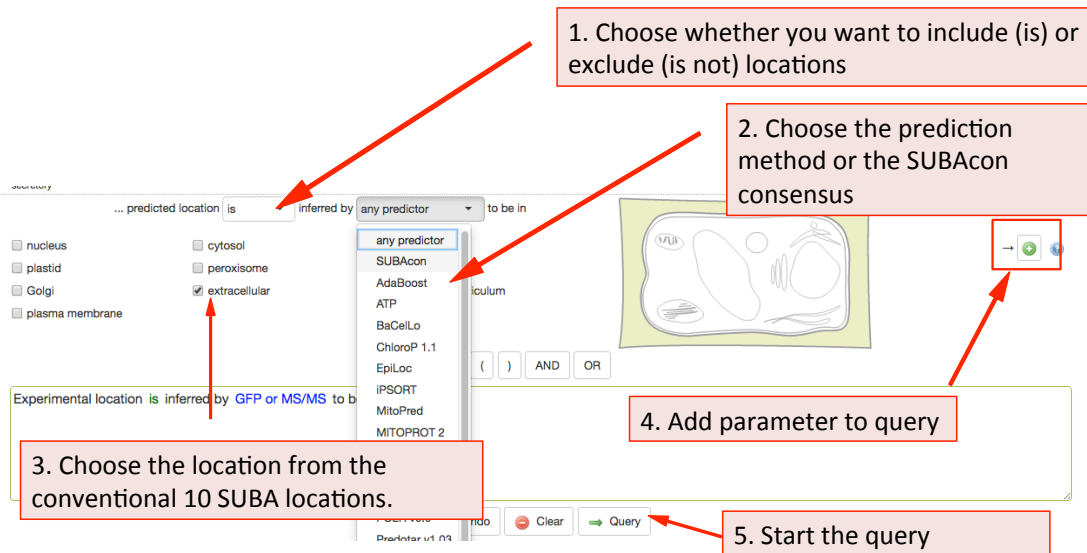
Below this, a dropdown menu is open, showing options: "GFP or MS/MS", "GFP assay", and "MS/MS assay". A red box with an arrow points to the "GFP or MS/MS" option with the text "Choose from the suborganellar locations".

At the bottom, a red box with an arrow points to the button with a plus sign and a minus sign with the text "Add the filter to the main query".

This will maximise the localisation view. Click on the desired location. For choosing more than one location, keep ticking more boxes. When choosing the whole compartments (extracellular), this will automatically include the

suborganellar locations (apoplast, cell wall). For only searching for apoplast, untick extracellular and only tick apoplast. Then add your parameter to the query by clicking the  button.

Similarly, to filter for prediction data choose the inclusion and exclusion and the type of predictor from the drop down list. Through this search option, you can also filter by our consensus call output when choosing SUBAcon.



The screenshot shows a search interface with the following elements and callouts:

- 1. Choose whether you want to include (is) or exclude (is not) locations:** Points to the dropdown menu showing 'is' selected.
- 2. Choose the prediction method or the SUBAcon consensus:** Points to the 'any predictor' dropdown menu with 'SUBAcon' selected.
- 3. Choose the location from the conventional 10 SUBA locations:** Points to the 'extracellular' checkbox, which is checked.
- 4. Add parameter to query:** Points to the plus icon button used to add parameters to the query.
- 5. Start the query:** Points to the 'Query' button at the bottom of the interface.

Once you have added all desired parameters to the query window you can check your query and submit it using the Query button. Your retrieved results will be automatically displayed in the Results tab when ready.

Protein Properties Search Parameters

The Protein property tab lets you filter SUBA data for protein annotations, physical properties and chromosomal locations. This tab also contains the option to enter a list of AGIs or text containing AGIs. A new query in SUBA4 lets you also filter for protein aliases, PFAM domains, EC numbers, pathway annotations, structural features and other annotations.

The screenshot shows the SUBA4 Protein Properties search interface. At the top, there is a navigation bar with links for Home, Search, Results, ToolBox, Help, Locations, Update, and Test. Below this is a search bar with buttons for AND, OR, Undo, Clear, and Query. The main query area contains the following text: "Experimental location is inferred by GFP or MS/MS to be in [plastid] AND Predicted location is inferred by any predictor,SUBAcon to be in [plastid] AND Underlying gene model is chromosome 1". A red box highlights this main query with the text "Main query will appear when adding parameters".

Below the main query is a section titled "Find Arabidopsis proteins where the...". It has tabs for Subcellular locations, Protein properties (selected), Protein-Protein Interaction, Affiliations, and Blast. The Protein properties tab is active, showing a list of search criteria. A dropdown menu is open for the "physical property of" field, showing options: "Number of supporting ESTs", "Number of supporting fl-cDNAs", "Number of amino acids", "Molecular weight", "Calculated IEP", and "GRAVY". A red box highlights this dropdown with the text "Enter a list of AGIs".

Another red box highlights the "Choose the query and enter thresholds, parameters or choose from the drop down menus then add to the main query" text, pointing to the search criteria fields and the "Add" (+) buttons.

At the bottom, there is a field for "EnsemblPlants identifier(s), alias or protein sequence feature is in the list of identifiers: (e.g. GO:0008270, IPR017986 etc.)". A red box highlights this field with the text "Enter PFAM domains, UniProt IDs and other features".

Protein-Protein Interaction Search Parameters


New in SUBA4: In addition to protein-protein interactions (PPI), there are now experimental localisations from observed protein-protein interactions (PPI) such as Bifocal completion Experimentation. The PPI search tab was included to provide a straightforward access to a number of PPI queries. Besides the conventional search for existing PPI partners by entering AGIs, SUBA4 users can now discover PPI proteins that have been experimentally shown to interact in a particular compartment. At the same time, the drop down menu also allows for the choice of PPI methodology. Other search options for PPI data include the isolation of PPI studies.

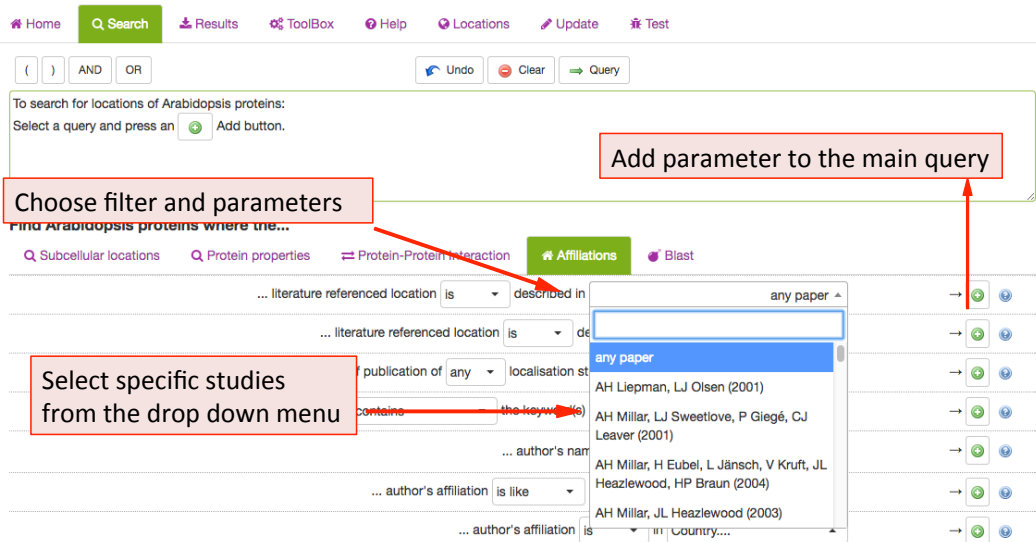
The screenshot shows the SUBA4 search interface with several annotations in red boxes and arrows:


- Annotation 1:** "Add the filter to the main query" points to the search query: "protein-protein interactions has been determined by methods [Bifocal Completion] AND location was experimentally observed in plastid".
- Annotation 2:** "Enter AGIs and find their interaction partners" points to the input field: "Enter AGI identifier(s) here (cut and paste/drag and drop)".
- Annotation 3:** "Search for interactions that have been observed in subcellular locations (experimentally by e.g. BiFC)" points to the subcellular location filter section, where "plastid" is selected.
- Annotation 4:** "Choose the methodology of interaction as a filter" points to the dropdown menu for "been determined by method of", which is open and shows "Bifocal Completion" selected.

The interface includes a navigation bar with Home, Search, Results, ToolBox, Help, Locations, Update, and Test. Below the search bar are buttons for AND, OR, Undo, Clear, and Query. The search results section includes a "Find Arabidopsis proteins where the..." section with tabs for Subcellular locations, Protein properties, Protein-Protein Interaction (selected), Affiliations, and Blast. There are also buttons for Clear, a dropdown for "protein does", and a button to "interact with protein(s) in list".

Affiliations Search Parameters


This tab allows SUBA4 users to find experimental localisation data from specific authors, institutions, countries or filter by year of publications. To limit the results a particular study, choose from the publication list in the drop-down menu. The results can also be filtered by any author (not just first author) as well as by year or range of years of publication. For adding any of the parameters to the main query press the  button.



The screenshot shows the SUBA4 search interface with the 'Affiliations' tab selected. The main query area contains the text: 'To search for locations of Arabidopsis proteins: Select a query and press an  Add button.' Below this, there are several search criteria with dropdown menus and 'Add' buttons. A red box highlights the 'Add parameter to the main query' button. Another red box highlights the 'Choose filter and parameters' dropdown menu. A third red box highlights the 'Select specific studies from the drop down menu' dropdown menu. The dropdown menu is open, showing a list of publications: 'any paper', 'AH Liepman, LJ Olsen (2001)', 'AH Millar, LJ Sweetlove, P Giegé, CJ Leaver (2001)', 'AH Millar, H Eubel, L Jänsch, V Kruff, JL Heazlewood, HP Braun (2004)', and 'AH Millar, JL Heazlewood (2003)'. A red arrow points from the 'Select specific studies from the drop down menu' box to the dropdown menu.

Home Search Results ToolBox Help Locations Update Test

() AND OR Undo Clear Query

To search for locations of Arabidopsis proteins:
Select a query and press an  Add button.

Add parameter to the main query

Choose filter and parameters

Find Arabidopsis proteins where the...

Subcellular locations Protein properties Protein-Protein interaction **Affiliations** Blast

... literature referenced location is described in any paper

... literature referenced location is de

publication of any localisation st

Select specific studies from the drop down menu

contains the keyword(s)

... author's nam

... author's affiliation is like

... author's affiliation is Country...

- any paper
- AH Liepman, LJ Olsen (2001)
- AH Millar, LJ Sweetlove, P Giegé, CJ Leaver (2001)
- AH Millar, H Eubel, L Jänsch, V Kruff, JL Heazlewood, HP Braun (2004)
- AH Millar, JL Heazlewood (2003)

SUBA4 allows the search for data by country of origin of the experimental study. Using the drop-down menu shows the countries and number of studies that have contributed to the SUBA4 data set. When using the map for choosing a country, the grey countries indicate a contribution to SUBA4 whereas white countries have not contributed data sets to SUBA4. Green indicates a chosen country.

Find Arabidopsis proteins where the...

Q Subcellular locations Q Protein properties Protein-Protein Interaction **Affiliations** Blast

... literature referenced location is described in any paper

... literature referenced location is described in pubmed:

... year of publication of any localisation studies is between and

... publication title or abstract of the localisation study contains the keyword(s)

... author's name is like

author's affiliation is like

... author's affiliation is in Australia

author's affiliation is in:

Choose a specific country from the drop-down menu or from the map to limit the results to data generated in a specific country

Country	Total Studies
Argentina	total 6
Australia	total 99
Austria	total 23
Belgium	total 59
Brazil	total 18
Canada	total 74
Chile	total 7
China	total 183
Czech Republic	total 21
Denmark	total 40
Finland	total 4
France	total 226
Germany	total 392
Greece	total 6
Hong Kong	total 25
Hungary	total 5
India	total 3
Ireland	total 3
Israel	total 35
Italy	total 38
Japan	total 391
Lithuania	total 1

plant energy biology ARC CENTRE OF EXCELLENCE

THE UNIVERSITY OF WESTERN AUSTRALIA

This work is licensed under a Creative Commons Attribution 4.0 International License. © 2005-2016 Harvev Millar (harvev.millar@uwa.edu.au)

BLAST Search Parameters

The BLAST tab contains the BLAST tool equal to the one in the BLAST panel labelled 'Find your closest AGI!' on the SUBA4 homepage. The user can enter a sequence and retrieve data from Arabidopsis proteins with sequence similarity. The results can be filtered using the BLAST score as a threshold. The score measures sequence similarity in respect to sequence length. The BLAST hit AGIs are retrieved and the data for the AGI linked protein is retrieved from SUBA.

Choose BLAST score threshold

Enter your protein fragment

Add to the query

Score is $-\log_2(E)$ where $E = p_{val} \times N_{eff}$ is the p-value times the effective search space size. The larger the score the more statistically significant the match but since we tried N_{eff} times to find a match we need to multiply the score by N_{eff} to get the true significance (PDF).

The hits are displayed in the results view and each hit shows the BLAST score and aligned protein sequence.

AGI	SUBAcon	Predictions	Annotations	GFP	MS/MS	PPI
AT2G40880.1	extracellular	cytosol mitochondrion peroxisome plastid vacuole golgi endoplasmic reticulum extracellular	SwissProt: extracellular			
cystatin A: Encodes a protein with cysteine proteinase inhibitor activity. Overexpression increases tolerance to abiotic stressors (i.e. salt, osmotic, cold stress).						
Blast Fragments score: 64.18 MADQQAGTIVGGVRDIDANANDLQVESLARFAVDEHNKNENLTLEYKRLGAKTQVQVAGTMHHLTVEVADGETNKVY						
AT3G12490.1	cytosol	nucleus cytosol vacuole plastid mitochondrion	SwissProt: extracellular	cytosol	cytosol endoplasmic reticulum	AT3G56170.1: mitochondrion
cystatin B: Encodes a protein with cysteine proteinase inhibitor activity. Overexpression increases tolerance to abiotic stressors (i.e. salt, osmotic, cold stress).						
Blast Fragments score: 67.65 MADQQAGTIVGGVRDIDANANDLQVESLARFAVDEHNKNENLTLEYKRLGAKTQVQVAGTMHHLTVEVADGETNKVY						
AT3G12490.2	extracellular	cytosol mitochondrion plastid vacuole golgi endoplasmic reticulum extracellular	SwissProt: extracellular		cytosol	
cystatin B: Encodes a protein with cysteine proteinase inhibitor activity. Overexpression increases tolerance to abiotic stressors (i.e. salt, osmotic, cold stress).						

BLAST score

BLASTed sequence

SUBA4 results tab

The results tab will automatically be activated when the query is submitted. SUBA4 users will be able to see the query by clicking on the “What’s this query” button in the top left. The results can be downloaded as a table format using the download button.

The results are presented in table format. The columns can be customized towards the preference of the user. The first column shows the AGI of the proteins fitting the submitted query and the description for the protein below. This is followed by the consensus call derived from SUBAcon. Each of the individual localisation data columns show the summary of the data for the category. For a more details view for each category the user can access the factsheet by clicking on the AGI.

The screenshot shows the SUBA4 results interface. At the top, there is a navigation bar with 'Home', 'Search', and 'Results' buttons. A search bar contains the text 'Your Main query' and 'The download of the results table'. Below the search bar, there are navigation controls including a page number '1', a search icon, and a 'page size: 20' dropdown. A 'Showing page 1 of 58 (1142 total hits)' indicator is present. On the right side, there are buttons for 'What's this query' and 'Download'. The main content area is a table with columns: AGI, SUBAcon, Predictions, Annotations, GFP, MS/MS, and PPI. The table lists several protein entries with their AGIs and associated localisation data. Annotations include 'golgi mitochondrion', 'endoplasmic reticulum', 'cytosol', 'plasma membrane (2x)', 'nucleus', 'mitochondrion', 'plastid', 'peroxisome', and 'extracellular endoplasmic reticulum'. Annotations are linked to their respective flat files. Red arrows point from text boxes to specific elements in the interface: 'What's this query' and 'Download' buttons, the 'MS/MS' column header, the AGI 'AT1G05190.1', the protein description for 'AT1G05320.1', the 'PPI' column header, and the AGI 'AT1G01910.1' in the PPI column.

Your Main query
The download of the results table

What's this query Download

page size: 20 Showing page 1 of 58 (1142 total hits)

AGI	SUBAcon	Predictions	Annotations	GFP	MS/MS	PPI
CLAVATA3/ESR-RELATED Clavata3 gene. Consists of						
AT1G05190.1	endoplasmic reticulum	golgi mitochondrion	golgi mitochondrion			
AT1G05320.1	endoplasmic reticulum	cytochrome P450, family 88, subfamily A, polypeptide 3; Encodes an ent-kaurenoic acid hydroxylase, a member of the	cytosol mitochondrion			
AT1G05520.1	endoplasmic reticulum golgi	endoplasmic reticulum plasma membrane nucleus cytosol mitochondrion				
Sec23/Sec24 protein transport family protein;						
AT1G05575.1	endoplasmic reticulum	plastid plasma membrane nucleus peroxisome extracellular endoplasmic reticulum		endoplasmic reticulum (2x)		AT1G01910.1: golgi

SUBA4 hits fitting your query. The AGI is linked and users can open the detailed flatfile by clicking on the AGI

Columns with the localisation summaries for each data category

Protein description as in TAIR10

PPI partners are linked to their flat file and their SUBAcon location is shown

SUBA4 Toolbox


The SUBA4 toolbox is a new feature and more detailed description of each tool is currently in preparation for submission. However, this section will provide the user with the information and instruction necessary for use.

The screenshot shows the SUBA4 Toolbox interface. At the top, there is a header with the SUBA logo and navigation links: "All about SUBA4", "SUBA4 home", and "go to SUBA3". Below the header is a navigation bar with "Home", "Search", "Results", "ToolBox", and "Help". The "ToolBox" tab is highlighted in green. Below the navigation bar, there are three tool tabs: "Abundance Tool (MMAP)", "COEX Adjacency Tool (CAT)", and "PPI Adjacency Tool (PAT)". The "MMAP" tab is selected. Below the tabs, there is a heading: "Estimate the compartmental protein abundance with the Multiple Marker Abundance Profiling (MMAP) tool". Below the heading, there is a text input field with the placeholder text "Enter a list of AGIs (or cut and paste or drag and drop a file):" and a "clear" button. Below the input field is a large dashed box labeled "Drag and drop window for AGI list". Below the input field and the drag-and-drop window, there are three buttons: "Calculate Abundances", "Images", and "Excel". Red arrows point from text boxes to these elements: "Choose one of the 3 tools" points to the tool tabs; "Tool short description" points to the heading; "Image download" points to the "Images" button; and "Tabular download" points to the "Excel" button.

There are 3 tools available in the SUBA4 toolbox. The initial toolbox view shows the front of the MMAP protein abundance tool (tab is highlighted in green). Below the tab index, the heading contains a short description of the main feature of the tool that is currently chosen. To choose another tool e.g. the Coexpression Adjacency tool (CAT), click on the CAT tab. All three tools have the AGI drag and drop window in common where AGIs can be dragged or pasted into. Besides tool-specific search functions, all tools have an image and tabular download function, which are situated underneath the AGI drag and drop window.

The Multiple Marker Abundance Profiling (MMAP) tool

The MMAP tool has been developed to provide a method to estimate compartmental protein abundance without additional experimental data. The relative protein abundance for each protein was estimated from global aggregated and normalized mass spectrometry data in MASC Gator (<https://omictools.com/masc-gator-tool>). A large number of marker proteins with high confidence was assembled for each compartment. When submitting a list of AGIs through the drag-and-drop window, the marker proteins are identified from the list and the relative abundance is summed for each compartment. This is compared to expected values in green and other non-predominantly photosynthetic tissues that have been estimated by MRM experimentation.

In short: For using this tool, all you need is a list of AGIs that is pasted into the window. Once you press the calculate abundances button () you receive a number of statistics back from your list .

A

choose the SUBA toolbox

then choose the Abundance tool

← Compartment legend

← Step 1. Input window the user AGIs

← Download options: download results graphs as images or as data table

← Step 2. Submit your AGI list and calculate abundances

B

Result 1. User AGI list statistics show the number of identified proteins in each compartment as estimated by SUBAcon (left) and the number of identified marker proteins in each compartment with assigned marker location (right)

Result 2. Relative compartmental abundance shows the estimated protein abundance of each compartment

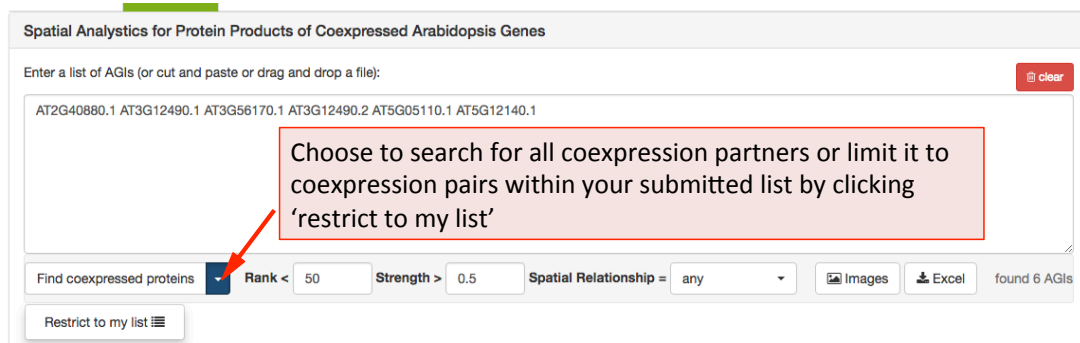
Result 3. Enrichment overview. The relative abundance of the user-submitted list is compared to the expected global protein abundance in green tissues or other tissues (left).

The ratio of the user-submitted relative abundance over the expected tissue abundance is calculated and shown in tabular format

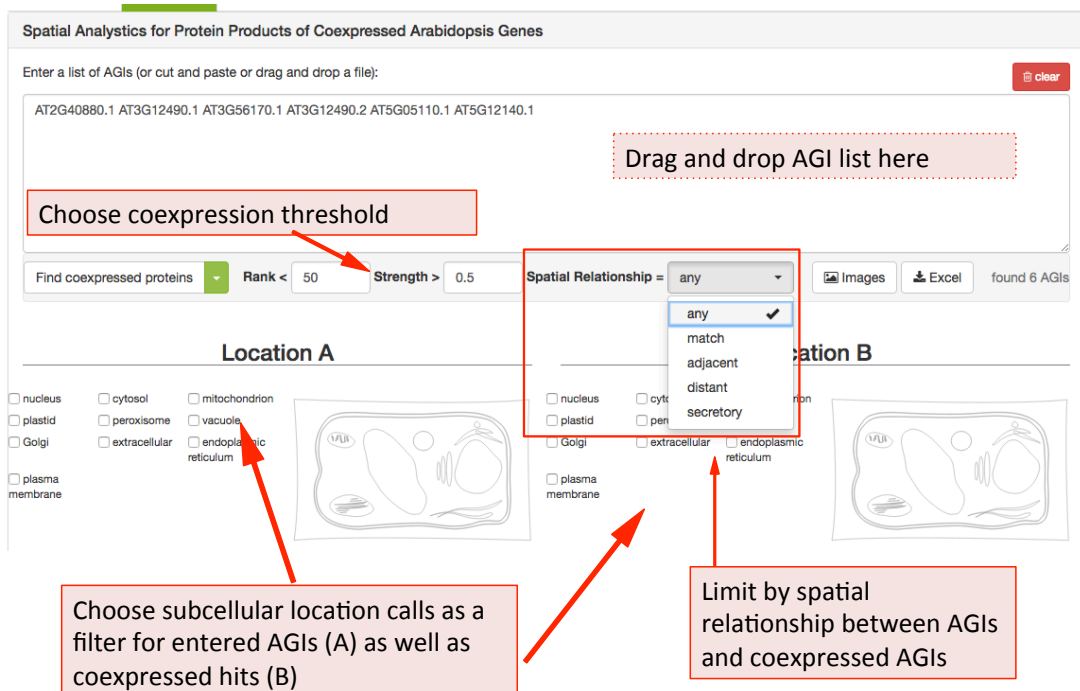
Location	User/Green	User/Other
cytosol	1.31	1.066
endoplasmic reticulum	0.09	0.06
extracellular	1.463	1.267
mitochondrion	0.815	0.654
nucleus	6.106	3.158
peroxisome	1.08	0.91
plasma membrane	1.053	0.93
plastid	0.433	0.784
vacuole	0.97	1.137

The Coexpression Adjacency Tool (CAT)

The coexpression adjacency tool provides spatial analysis statistics for coexpression partners using SUBAcon location calls. The user can drag a list of AGIs in the window on the top and look for coexpressed AGIs either within the list ('restrict to my list') or globally. SUBA4 contains the top 300 coexpressed partners and excludes self-coexpressed AGIs (rank = 1). Pressing the 'Find coexpressed proteins' or the 'restrict to my list' button submits the query and returns the statistics.

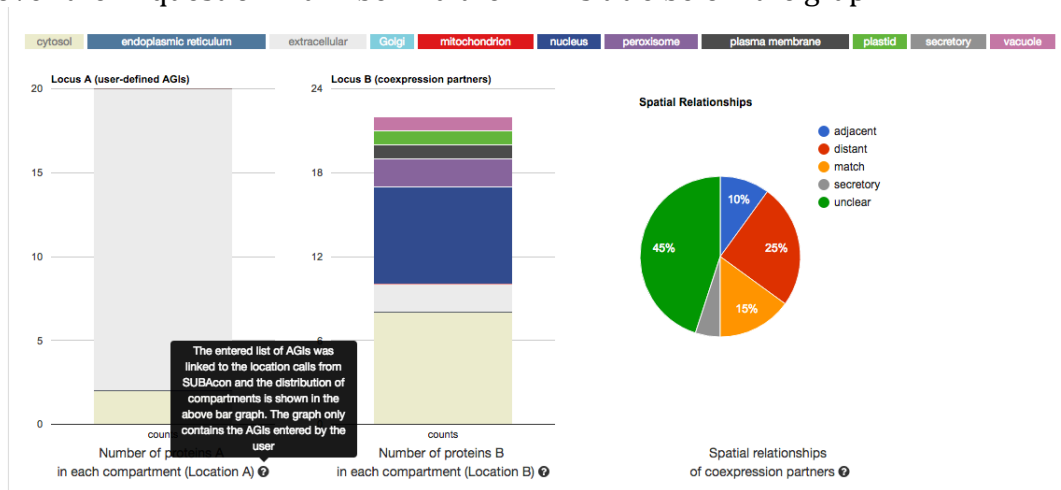


To limit the list of coexpressed partners, the coexpression rank (mutual rank 2-300) and strength (average coexpression coefficient 0-1) can be restricted. The list of coexpressed AGIs can also be filtered by the spatial relationship of the expressed proteins. The categories include matching, adjacent, distant location pairing as well as secretory for locations exclusively within ER, Golgi, vacuole, plasma membrane and extracellular. Mixed location calls with unclear biological implications are combined into the category unclear.



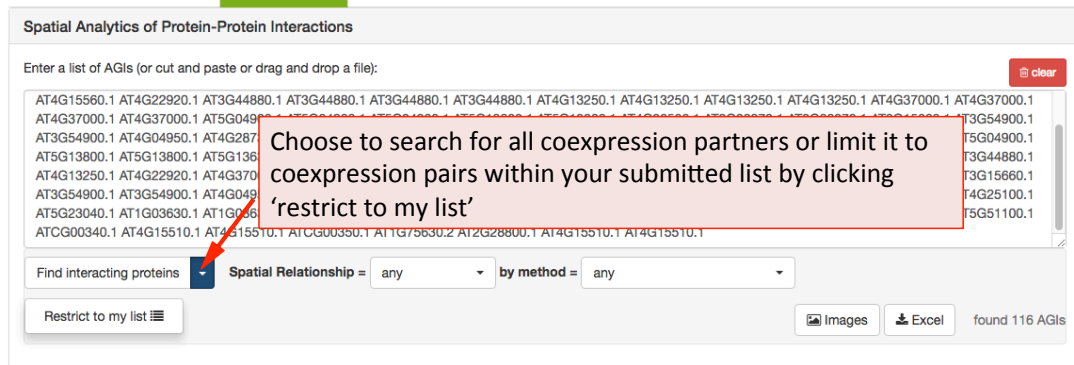
When searching for specific location pairings, the user-submitted list of AGIs can also be specifically limited to location calls for the proteins submitted (location A) and the location calls for the coexpressed proteins (location B)

Once submitted, the statistics for your user list will appear in format of 3 graphs. From left to right, it shows the distribution of the location calls for the user-submitted AGIs (Location A) and the distribution of the location calls for the retrieved coexpressed AGIs (Location B). If the 'restrict to my list' option was chosen, these distributions may look very similar. The pie chart on the right shows the distribution of the spatial relationships of the coexpression couples. This relative measure shows the % of matching, adjacent and distant compartment PPIs. The information for each graph can be obtained by gliding over the ? question mark behind the x-axis title below the graph.

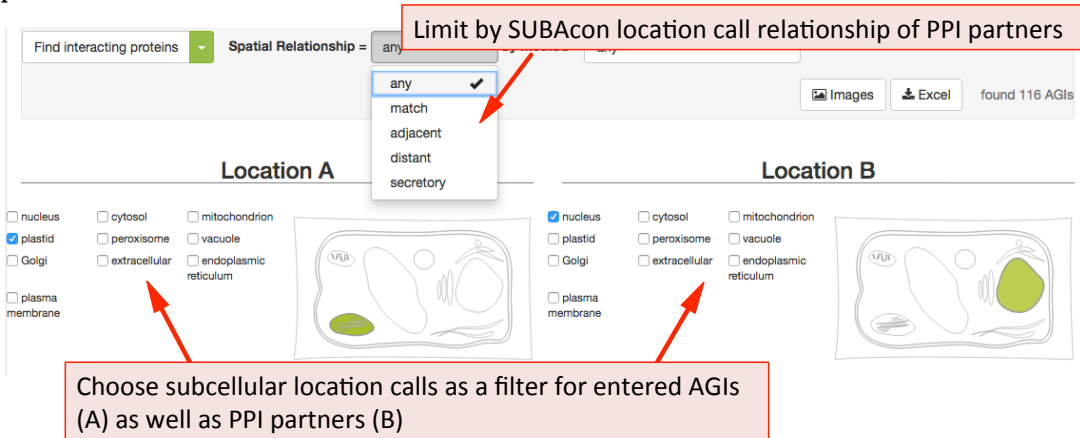


The PPI Adjacency Tool (PAT)

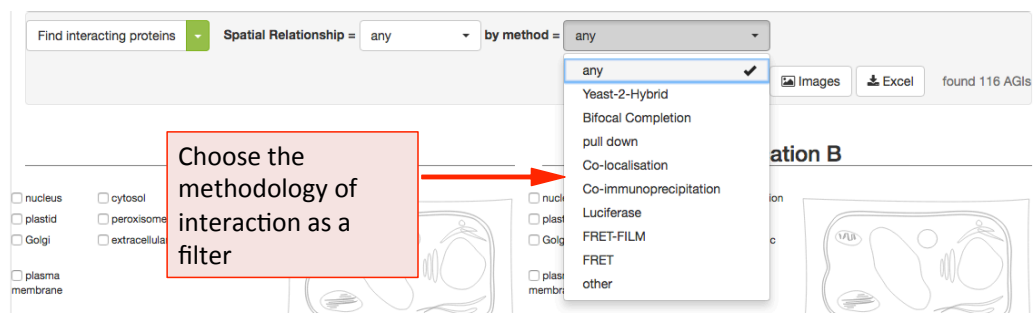
The Protein-Protein Interaction (PPI) adjacency tool provides spatial analysis statistics for PPI partners using SUBAcon location calls. The user can drag a list of AGIs in the window on the top and look for PPI partners either within the list ('restrict to my list') or globally. SUBA4 contains a set of experimentally verified PPI proteins that is searched through this tool. Pressing the 'Find interacting proteins' or the 'restrict to my list' button submits the query and returns the statistics.



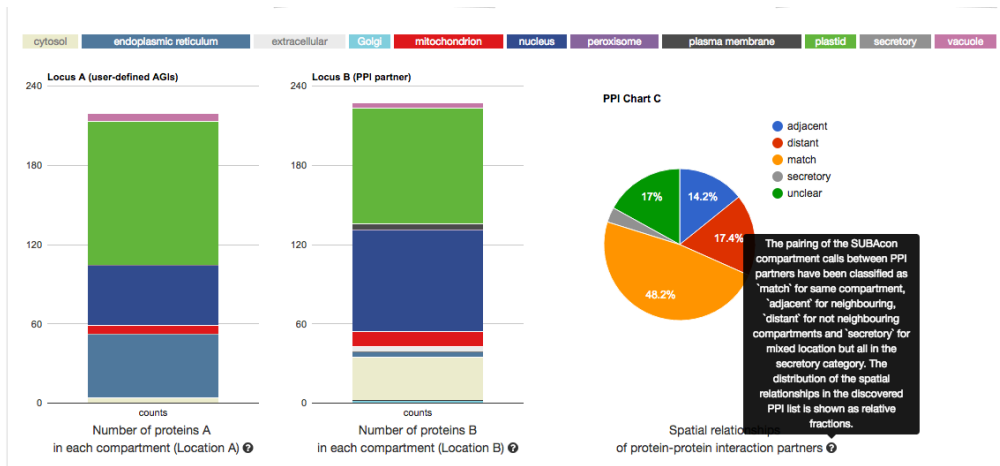
The PPI retrieval can be limited by the locations in two different ways. Firstly the user can choose to only view PPI proteins which have PPI partners in the same (match), neighbouring (adjacent) or non-neighbouring (distant) compartments. For PPI partners within the secretory but not matching (e.g. ER - Golgi), the category 'secretory' was introduced. This may help identify functionally related proteins within the PPI cohort.



The list of proteins can also be refined using the type of methodology that was used to identify the PPI. This includes Yeast-2-hybrid, immune co-precipitation, bifocal completion and other methodologies that can be chosen from the drop down menu.



Once submitted, the statistics for your user list will appear in format of 3 graphs. From left to right, it shows the distribution of the location calls for the user-submitted AGIs (Location A) and the distribution of the location calls for the retrieved PPI AGIs (Location B). If the 'restrict to my list' option was chosen, these distributions may look very similar. The pie chart on the right shows the distribution of the spatial relationships of the PPI couples. This relative measure shows the % of matching, adjacent and distant compartment PPIs. The information for each graph can be obtained by gliding over the [?] question mark behind the x-axis title below the graph.



The SUBA4 factsheet

Each protein in Arabidopsis has a SUBA4 factsheet that contains all the details about the protein properties, affiliated studies and localisations. The Factsheet can be accessed by clicking on the AGI in the results tab.

The screenshot shows the SUBA4 search results interface. At the top, there are navigation links: Home, Search, Results, ToolBox, Help, Locations, Update, and Test. Below these is a search bar with '1' entered and a 'page size' of 20. The main table has columns: AGI, SUBAcon, Predictions, Annotations, GFP, MS/MS, and PPI. The first row shows 'AT1G02470.1' under AGI, 'plastid' under SUBAcon, 'plastid' under Predictions, and 'goli' under Annotations. A red box highlights the AGI link 'AT1G02470.1' with an arrow pointing to it. A text box next to the arrow says 'The AGI is linked to the detailed factsheet. Access by clicking on the AGI'.

The top of the factsheet contains a widespread view of the localisation information. The left column contains the consensus that is formed by the Algorithm SUBAcon (red box). This takes into account the predicted as well as experimental localisation data. SUBAcon was trained using the subcellular reference standard ASURE. If the protein you are looking for was part of the reference standard, this will appear as well.

The screenshot shows the detailed SUBA4 factsheet for AT4G26500.1. At the top, there is a search bar with 'AT4G26500.1' and a 'Search AGI' button. Below the search bar is a 'Subcellular Consensus (Prediction and Experimental)' section with a diagram of a cell and a color scale. To the right are 'Predictors' and 'Annotations' sections. Below these are 'GFP', 'MS/MS', and 'PPI' sections.

Subcellular Consensus (Prediction and Experimental)		Predictors	Annotations	
<p>min: :max</p> <p>extracellular: 0 SUBAcon: plastid 1.000 Gold Standard: plastid What's this?</p> <p>Experimental Data</p>		<p>AdaBoost: plastid BaCellLoc: plastid iPSORT: plastid MultiLoc: plastid Predotar: plastid SLPFA: mitochondrion TargetP: plastid</p> <p>ATP: mitochondrion ChloroP: plastid MitoPred: mitochondrion PCLR: plastid PredSL: plastid SLP-Local: plastid WoLF PSORT: plastid</p> <p>ATP: plastid EpiLoc: nucleus Mitoprot 2: mitochondrion Plant-mPloc: plastid PProwler: plastid SubLoc: cytosol YLoc: plastid</p>	<p>AmiGO : mitochondrion 16437155 AmiGO : plastid 16437155 AmiGO : plastid 16455656 SwissProt : plastid 16381842 TAIR : plastid 18633119 TAIR : plastid 16437155 TAIR : plastid 18431481 TAIR : plastid 16455656 TAIR : mitochondrion 16437155</p>	
		GFP	MS/MS	PPI
		16455656 (2006): plastid	24872594 (2014): plastid 24361574 (2014): plastid » plastid matrix 21531424 (2011): plastid 20423899 (2010): plastid 18633119 (2008): plastid » plastid matrix 18431481 (2008): plastid	24203231 (2014): None 24203231 (2014): plastid 24203231 (2014): None 24203231 (2014): None 24203231 (2014): plastid 24203231 (2014): None 24203231 (2014): None

The top row displays the predictor outputs for any of the 22 predictors if there is one. If a predictor does not cover your protein, it will not appear. On the right you find the Annotations retrieved from other database resources.

The second row contains the experimental localisations including GFP localisations, MSMS localisations and PPI partners with localisations if available.

In the next row you find the AGI-AGI relationships. This is a selection of the top 10 coexpressed AGIs of your protein and the SUBAcon location call as well as the average coexpression coefficient (in red box). Below you find the protein name, description and any curator comments available through TAIR 10.

12938931 (2003): plastid		22366162 (2012): None
SUBAcon links AGI-AGI relationships	<p>AT3G02730.1 » plastid [0.45] AT5G65840.1 » plastid [0.42] AT1G63970.1 » plastid [0.43] AT1G77090.1 » plastid [0.43] AT5G37360.1 » plastid [0.41] AT4G36530.1 » cytosol [0.41] AT1G71480.1 » nucleus [0.41] AT4G27700.1 » plastid [0.42] AT4G22890.1 » plastid [0.41] AT1G32220.1 » plastid [0.4]</p>	
Description (TAIR10)	protein_coding : accelerated cell death 2 (ACD2)	
Curator Summary (TAIR10)	Mutants have spontaneous spreading cell death lesions and constitutive activation of defenses in the absence of pathogen infection. Its product was shown to display red chlorophyll catabolite reductase (RCCR), which catalyzes one step in the breakdown of the porphyrin component of chlorophyll. The enzyme was further assessed to be a Type-1 (pFCC-1-producing) RCCR. Upon P. syringae infection, ACD2 localization shifts from being largely in chloroplasts to partitioning to chloroplasts, mitochondria, and to a small extent, cytosol. Overexpression of ACD2 delayed cell death and the replication of P. syringae.	
Computational Description (TAIR10)	ACCELERATED CELL DEATH 2 (ACD2); FUNCTIONS IN: red chlorophyll catabolite reductase activity; INVOLVED IN: chlorophyll catabolic process, defense response, incompatible interaction, regulation of programmed cell death, regulation of plant-type hypersensitive response; LOCATED IN: in 6 components; EXPRESSED IN: 23 plant structures; EXPRESSED DURING: 13 growth stages; CONTAINS InterPro DOMAIN/s: Red chlorophyll catabolite reductase (InterPro:IPR009439); Has 181 Blast hits to 181 proteins in 30 species: Archae - 0; Bacteria - 8; Metazoa - 0; Fungi - 0; Plants - 170; Viruses - 0; Other Eukaryotes - 3 (source: NCBI BLINK).	

A new feature in SUBA4 is the Annotation box. This box contains a list of annotations from other data resources including functional domains, aliases, pathway annotations, enzyme annotations and others. Most of them are linked to their respective resource where more information about the annotation can be obtained.

Protein Annotations	<p>BioCyc:ARA:AT4G37000-MONOMER EMBL:AF326347 EnsemblPlants:AT4G37000 GO:GO:0005739 GO:GO:0010363 hmmpanther:PTRH34685 KEGG:ath:AT4G37000 PDBsum:3AGA PDBsum:3AGC PDBsum:3AGD ProteinModelPortal:Q8LDU4 UniGene:At.69554</p>	<p>BioCyc:MetaCyc:AT4G37000-MONOMER EMBL:AL161500 EnsemblPlants:AT4G37000.1 GO:GO:0005829 GO:GO:0019996 hmmpanther:PTRH34685:SF2 K0K13545 PDB:3AGB Pfam:PF06405 Proteomes:UPO00006548 UniPathway:UPA00674</p>	<p>BioGrid:15135 EMBL:AY045576 entrez:829854 GO:GO:0009507 GO:GO:0043067</p>	<p>BRENDA:1.3.1.80 EMBL:AY085797 EvolutionaryTrace:Q8LDU4 PDBsum:2Z0X PhylomeDB:Q8LDU4 SMR:Q8LDU4</p>	<p>EC:1.3.1.12 EMBL:AY093785 GeneID:829854 GO:GO:0009535 GO:GO:0051743 IntAct:Q8LDU4 PaxDb:Q8LDU4 PDBsum:2ZXL PIR:A85437 STRING:3702.AT4G37000.1</p>	<p>eggNOG:ENOG410JWI eggNOG:ENOG4111JQP EMBL:CP002687 GeneVisible:Q8LDU4 GO:GO:0005737 GO:GO:0009814 GO:GO:0009841 gramene_pathway:1.3.1.80 gramene_pathway:PWY-5098 InterPro:IPR009439 KEGG:00860+1.3.1.80 PDB:2Z0X PDBsum:3AGA PRIDE:Q8LDU4 PRO-PR:Q8LDU4 UniGene:At.4644</p>
Coordinates (TAIR10)	chr4:++17442627..17443762					
Molecular Weight (calculated)	36450.50 Da					
IEP (calculated)	5.90					
GRAVY (calculated)	-0.40					
Length	319 amino acids					
Sequence (TAIR10) (BLAST)	<p>001: MANIFONTLY SSSSPSYLSP LTKPSRFSK NLRPRAQF05 MEDHDDHLRR KFNEFPYVSP TRKQLMVOLM STVENRLQSQ LLPCNLPDVP RNFNPNFGSA 101: EASLHIRSGD KSSPIDFVIG SNIWCKIPTG VSLNITISIG FUNSTKAPN FVLELDQSS KSLVLLTLDLP HRKDLVLPND YLKEYYQDTA LDSHRQSLK 201: LPEVNPYVSP SLFVRSFVSP TASMCLKDAE EEEKLEELR DHVSPAKEV LEVLERCVK EEEKIVGVE EERMELELRD KSFRRKSIDL DLDFPFRMF 301: GEEVSRVWH AIKEAFGLV</p>					
<p>Hydropathy for AT4G37000.1</p>						
See Also	<p>AIP APP Aramemnon pep2pro DBGET Inparanoid MASC P Gator MIPS MPSS Plus PPDB PlantSpecDB ProMEX Proteins Wiki SALK (inserts) SALK (signal) TAIR SwissProt</p>					