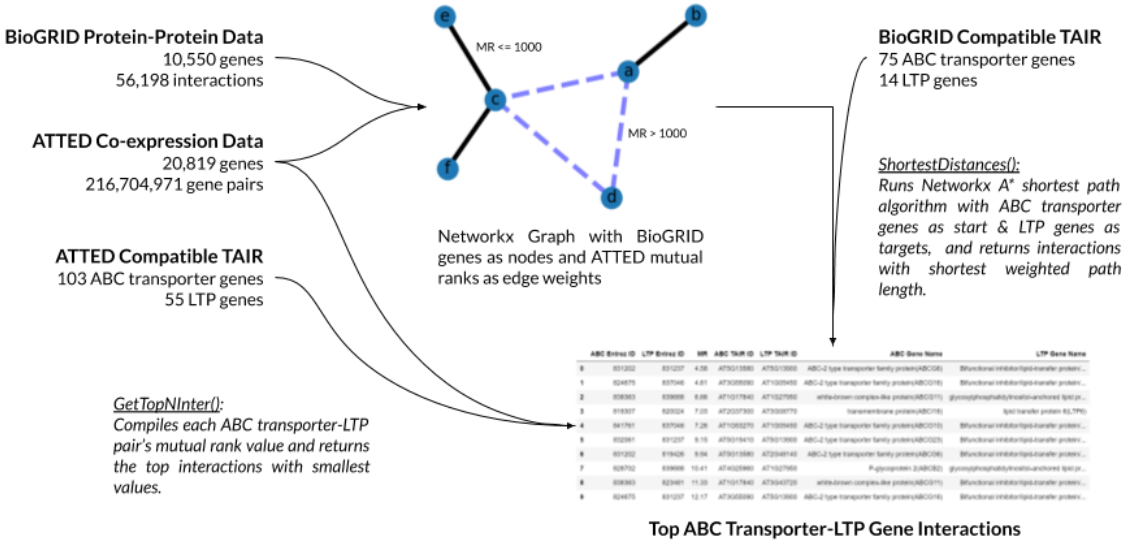
Linking of ATP-Binding Cassette (ABC) Transporters and Lipid Transfer Proteins (LTPs) in Arabidopsis

Yashwanth Manne
Michigan State University
CMSE 410: Bioinformatics and Computational Biology
Dr. Arjun Krishnan
April 24, 2020

Abstract

Gene networks have become quite popular for genome-wide representation of the complex functional organization of biological systems. In particular, gene co-expression networks prove quite useful in the functional annotation of unknown genes.[13] This project combined the mutual ranks of co-expression data with protein-protein networks in an exploratory analysis of pairs of ATP-binding cassette (ABC) transporter and Lipid Transfer Protein (LTP) genes to try to find functional linkage between the two protein types. ABC transporters are known to function in cross-membrane transport while LTPs are known to taxi small hydrophobic molecules. The question lies in whether or not these proteins directly interact in similar pathways. The analysis resulted in a list of 50 different ABC transporter-LTP gene pairs with 7 pairs having mutual ranks below 10, indicating that they are very likely co-expressed and may have a functional linkage. Of the seven most promising pairs identified, one interaction has already been experimentally characterized.[6] As such, these results prove to be promising and further exploration of the gene pairs may shed more light on the lipid transport mechanisms of vital cellular pathways.



Top ABC Transporter-LTP Gene Interactions

Introduction

Arabidopsis thaliana is the principal genetic model and standard reference organism in plant and crop science. This is due to its short generation time, small size and space requirements, prolific seed production, and small, genetically tractable genome.[1] Due to its role as a reference organism, it is imperative that it be fully functionally characterized in order to aid in future research of other plant organisms. Ongoing research of Arabidopsis is that of ATP-binding cassette transporters and lipid transfer proteins and their role in various metabolic pathways.

ATP-binding cassette (ABC) transporters are one of the largest protein families in all living organisms. They function as a method of transmembrane transport driven by ATP hydrolysis. In prokaryotes, they not only serve as importers (prokaryotes only) and exporters but are also thought to function in DNA repair and translation.[2] In Arabidopsis, 22 out of 130 ABC transporter proteins have been functionally analyzed.[3] All of these have been localized to membranes and are involved in roles including but not limited to detoxification, organ growth, plant nutrition, plant development, abiotic stress response, pathogen resistance, plant-environment interaction, pathogen response, surface lipid deposition, phytate accumulation in seeds, and transport of the phytohormones auxin and abscisic acid. Generally, each ABC transporter contains two

transmembrane domains (TMDs) and two cytosolic/nucleotide-binding domains (NBDs). In plants, the encoding of each subunit varies by subgroups.[3] All may be coded by individual genes, two genes may each encode an NBD and TMD pair and form heterodimers, a single gene may encode an NBD and TMD pair and form homodimers, or a single gene can encode all four domains. Note that there are 8 different ABC subgroups (A, B, C, D, E, F, G, I) and the I subgroup has bacterial origins.

Lipid Transfer Proteins (LTPs) are small, compact proteins folded around a hydrophobic cavity, enabling it to transport other hydrophobic molecules like lipids. In fact, it has been shown that the majority of lipid traffic is done by LTPs rather than vesicles.[4] It is thought to be key in the colonization of land as it is encoded by gene families in all land plants but not algae or other organisms.[5] There is evidence that LTPs are involved in the transfer and deposition of monomers required for the assembly of water-proof lipid barriers and signaling during pathogen attacks. Additionally, it has been theorized that LTPs facilitate the transfer of barrier materials and adhesion between barriers and extracellular materials.

It has been shown that ABC transporters often work in parallel with LTPs that 'pick up' the compound from ABC transporters.[6] This project aims to identify possible gene pairs of ABC transporters and LTPs to provide a branching point for future research of cellular transport mechanisms in plants. This is vital as transport mechanisms are vital in most plant pathways and its understanding can aid not just future research but also the production of high-value plant products or crops.

Data and Methods

Datasets

ATTED

ATTED-II is a plant co-expression database created from numerous DNA microarray analysis studies.[7,8] The ATTED-II data seems to be a rma.mrgeo.d file. Specifically, the folder that was downloaded, "Ath-mB.v17-08.G20819-S16033.rma_combat.mrgeo.d," contains multiple files, with each being a co-expressed gene list for that specific query gene titled by its Entrez ID. Each file has the gene ID, a mutual rank of the gene (MR), and the Pearson's correlation coefficient (PCC) of each gene to a given gene. The creators of the data recommended that MR be used rather than the PCC for data analysis.

BioGRID

Biological General Repository for Interaction Datasets (BioGRID), which is a curated database of genetic, protein, and chemical interactions as well as post-translational modifications.[9] The "BIOGRID-ORGANISM-3.5.181.tab2.zip" file that was downloaded contains the protein-protein interactions for seventy organisms but only the file for *Arabidopsis thaliana*: "BIOGRID-ORGANISM-Arabidopsis_thaliana_Columbia-3.5.181.tab2" was used. This file is formatted as a Tab 2.0 Delimited Text file and contains all interaction and associated annotation data. There are some missing annotations in the dataset, but they are not relevant to this project and can be ignored. The relevant columns are 'Entrez Gene Interactor A,' 'Entrez Gene Interactor B,' Organism Interactor A,' and 'Organism Interactor B.'

TAIR

The Arabidopsis Information Resource (TAIR) is a collective database of genetic and molecular biology data for the model plant Arabidopsis thaliana.[10] The site was used to gather a list of ABC transporter and LTP genes. The ABC transporter gene data was already compiled as an *.xls* file.[11] The LTP data was collected manually into an *.xlsx* document using the search engine with the keyword 'LTP'. Preprocessing of the data

was conducted in Excel to convert AGI codes to TAIR IDs. DAVID's Gene ID Conversion tool was used to convert the genes to their Entrez IDs and outputted as *.txt*.[12]

Approach

The project aimed to take co-expression data of Arabidopsis genes and rule out spurious relations by cross-referencing with protein-protein interactions. Co-expression data was obtained from the ATTED data and the protein-protein interactions were sourced from BioGRID. These two datasets were then combined into a weighted, undirected Networkx graph that is the framework for our analysis. The co-expression mutual ranks between genes in the ATTED dataset made up the edge-weights for the interactions between genes described in the BioGRID dataset. Once the graph was generated, separate lists of ABC transporter and LTP genes were fed into a "ShortestDistances" function that calculated the shortest-path for each ABC transporter-LTP pair using Networkx's A* algorithm for very large graphs as a matrix. The A* algorithm is a heuristic that drastically reduces computational time and memory space. Using this generated matrix, the top fifty unique ABC transporter-LTP gene pairs with the lowest path lengths were selected. Additionally, the lists of ABC transporter and LTP genes were passed into the "GetTopNInter" function that compiles each ABC transporter-LTP pair's mutual rank value using the ATTED dataset and returns the top N ABC transporter-LTP pairs.

Generation of Networkx Graph

In order to create the Networkx graph, the 'Entrez Gene Interactor A' and 'Entrez Gene Interactor B' columns of the BioGRID dataset were read in and isolated as a Pandas DataFrame and a list of unique genes was created. Then, OS was used to check for the existence of an ATTED file titled by each gene in the list. If the ATTED file existed, it was read in as a column of a Pandas DataFrame with the columns and row indexes as Entrez IDs and the values being MRs. Next, the simplified BioGRID DataFrame was iterated through and a third column labeled 'MR' was added to the BioGRID DataFrame with the values being pulled from the ATTED data for the corresponding genes. If the BioGRID genes were not present in the ATTED data, np.nan was added instead. Finally, Networkx was used to create edges from the BioGRID DataFrame with the 'MR' column making up the edge weights.

Shortest Path Data Generation: 'ShortestDistances', 'SimplifyGraph'

First, the lists of ABC transporter genes and LTP genes were processed to remove the genes not in the nodes of the Networkx graph generated. Next, 'ShortestDistances' was used to iterate through the lists of ABC transporter & LTP genes and calculate the distance between each ABC transporter gene to each LTP gene using Networkx's 'astar_path_length()' function and output a Pandas DataFrame with ABC transporter genes as the columns and LTP genes as the rows with the shortest distances between the two as the values. Unfortunately, this approach could not run on the computer. As such, the previous Networkx graph was simplified to only nodes within one degree of separation from the target ABC transporter and LTP genes. The general idea was to start off with a very small graph and slowly expand the degree of separation from the target genes until the computer memory was reached. This was done with the 'SimplifyGraph' function, which took in a list of ABC transporter genes, list of LTP genes, the original graph to simplify, and an integer, K, representing the degree of separation. The function iterates through the target genes (originally the concatenated lists of ABC transporter & LTP genes) and gathers all neighboring nodes into a new list using Networkx's 'Graph.neighbors()' function. The new list becomes the target genes and the process continues in a while loop and K is subtracted until it is less than or equal to 0. The now greater target nodes list is passed into Networkx's 'Graph.subgraph()' function to create the new, simplified graph. Unfortunately, the simplified graph with one degree of separation in the 'ShortestDistances' function still crashed the computer. As such, a Python script called 'ShortestPath' was created and fed into the High Performance

Computing Center at the Institute (HPCC) for Cyber-Enabled Research (ICER). Alas, there were some technical errors in running the script and the use of the HPCC proved to be unfeasible for this project within the timeframe. It is the author's hope that this will be rectified in future editions of this research.

Top N ABC transporter-LTP Interactions: 'GetTopNInter', 'GetTopGenes'

First, all ATTED data files were read in as a Pandas DataFrame. Using this data, a separate function called 'GetTopGenes' iterating through two lists of ABC transporter and LTP genes, checks if the ABC transporter-LTP gene-gene MR value is below an inputtable cutoff,C, and the interaction appends to a list as a tuple. The 'GetTopInter' function takes in 'N', the cutoff for the number of interactions, as input and continually runs the 'GetTopGenes' function inside the while loop and checks to see whether or not the length of the interaction list returned by that function is less than the N and quickly increases the cutoff, C, for the 'GetTopGenes' by 5. Once the first while loop concludes, a second while loop starts by slowly decreasing C by increments of 0.1 until the length of the interaction list is less than or equal to N. At the end, the function outputs a list of tuples for each ABC/LTP interaction, a dictionary with ABC transporter genes as keys and LTP genes as values, and a list of unique genes. These were used to make a DataFrame with Entrez ID, TAIR ID, and common name for each ABC transporter and LTP gene along with MR values for each interaction.

Software

The majority of data analysis was conducted using Python 3.7.7 and the Pandas (1.0.3), Networkx (2.4), Matplotlib (3.1.3), Numpy (1.18.1), Seaborn (0.10.0), Pickle (4.0), OS, and Collections libraries. The complete Python code archive can be found here: https://github.com/yashmanne/Arab_ABC-LTP_Linkage

In addition to Python, the National Center for Biotechnology Information (NCBI) Taxonomy Name/ID Status Report was used to find the organisms of the BioGRID interactions. Finally, the Database for Annotation, Visualization and Integrated Discovery (DAVID) Gene ID Conversion tool was used to convert the TAIR IDs of the ABC transporter and LTP genes to their Entrez IDs.

Results and Discussion

Preliminary Data Analysis

ATTED

There are 20,819 total genes in the ATTED dataset out of which 8,782 are also present in the BioGRID set. The dataset contains no missing values.

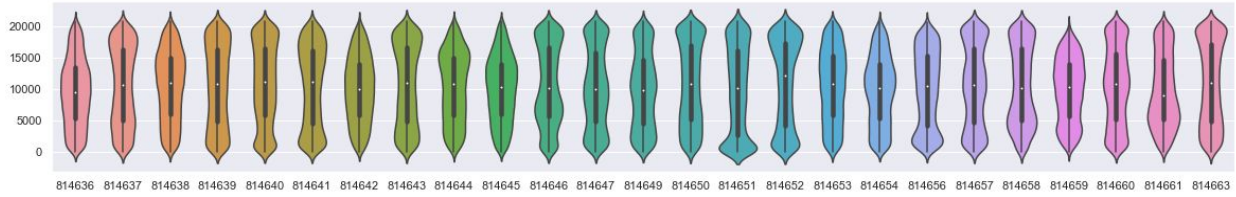


Figure 1. Violin Plot of First Few ATTED Genes

The violin plot shows that the data is normalized and the distribution of the mutual rank values of the ATTED genes are relatively similar across multiple genes. As such, the MR can be used readily for analysis.

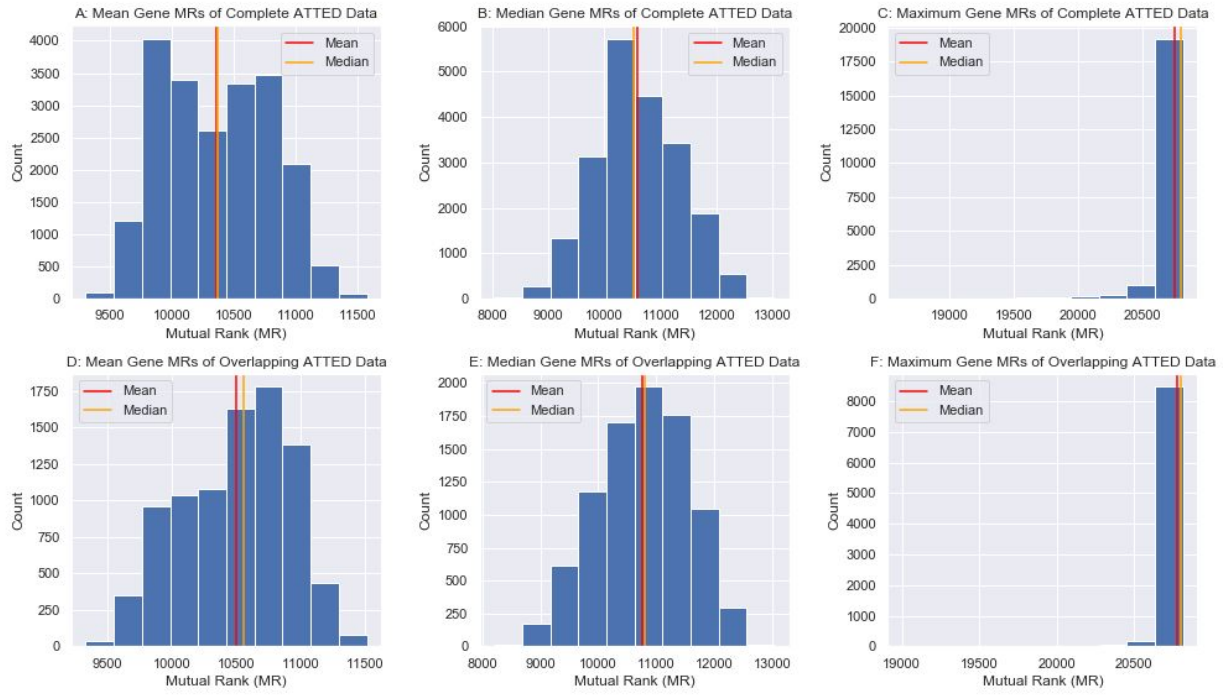


Figure 2. Means, Medians, and Maximums of Complete & Overlapping ATTED Data

The first row of graphs are for the complete ATTED dataset. (A) The overall distribution of mean MR values for each gene is fairly normal. The minimum is 9,303.55, the maximum is 11,570.47, the mean is 10,360.99, and the median is 10,363.80. (B) The overall distribution of median MR values for each gene is normal. The minimum value is 8,025.39, the maximum is 13,042.22, the mean is 10,583.59, and the median is 10,527.91. (C) The overall distribution of maximum MR values for each gene is heavily skewed left. The minimum value is 18,632.12, the maximum is 20,818, the mean is 20,749.13, and the median is 20,799.90. The next row of graphs are for just the overlapping genes of the ATTED data. (D) The overall distribution of mean MR values for each gene is fairly normal. The minimum is 9,333.04, the maximum is 11,517.99, the mean is 10,497.96, and the median is 10,556.60. (E) The overall distribution of median MR values for each gene is normal. The minimum value is 8,203.10, the maximum is 13,042.22, the mean is 10,759.11, and the median is 10,804.69. (F) The overall distribution of maximum MR values for each gene is heavily skewed left. The minimum value is 19,002.36, the maximum is 20,818, the mean is 20,782.26, and the median is 20,806.20.

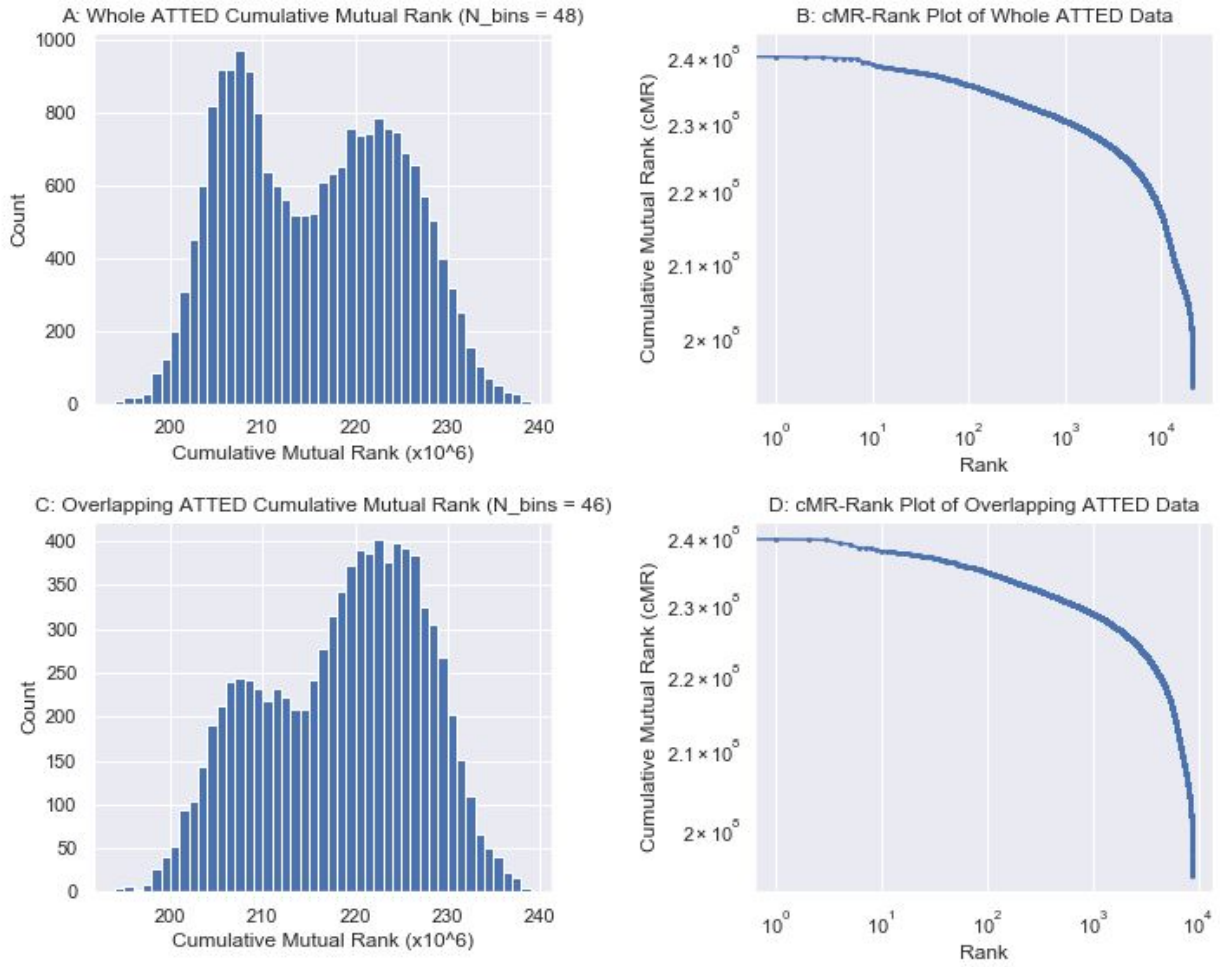


Figure 3. Cumulative Mutual Rank of Complete & Overlapping ATTED Data
(A, B) The cumulative mutual rank of the entire ATTED dataset follows a bimodal distribution with peaks at 207 million and 222 million with about 1,000 and 800 genes having cumulative MRs in that range. (C, D) The cumulative rank of just the overlapping genes from the ATTED dataset follows a similar distribution with peaks at 207 million and 222 million but with counts of 250 and 400 genes respectively.

BioGRID

The BioGRID data shows that there are 15 other organisms interacting with Arabidopsis. Of the 56,198 total interactions, 55,814 are Arabidopsis:Arabidopsis interactions, 384 are Arabidopsis: Other interactions, and 0 that don't involve Arabidopsis. There are 10,550 unique genes in the dataset but only 10,367 genes involved in only Arabidopsis: Arabidopsis interactions. The original Networkx graph made from this data has 10,550 nodes and 48,981 edges. A simplified graph of nodes just 1 node away from ABC transporter & LTP genes has 399 nodes and 2,907 edges.

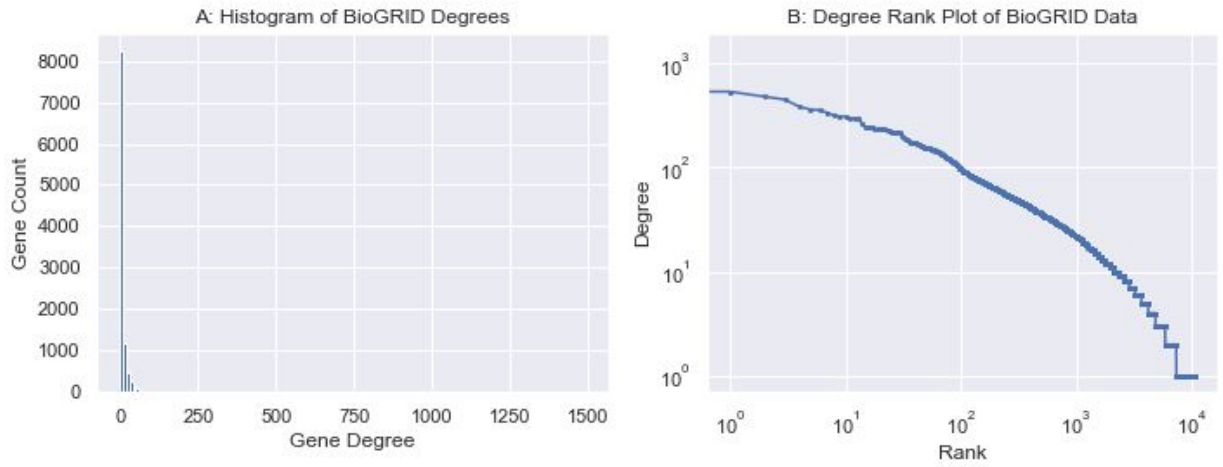


Fig. 4 Degrees of BioGRID genes
(A) 8,238 genes have between zero and ten connections but there is one gene with 1,341 connections. (B) The degree-rank further illustrates that most genes have very few connections but there are still 296 genes with over fifty connections and 97 genes that have well over a hundred connections.

TAIR
Out of the 131 ABC transporter genes gathered from TAIR, 129 were convertible to an Entrez ID, 103 were compatible with the ATTED data, and 75 were compatible with the BioGRID data. All 85 LTP genes gathered were convertible to an Entrez ID, 55 were compatible with the ATTED data, and 14 were compatible with the BioGRID data.

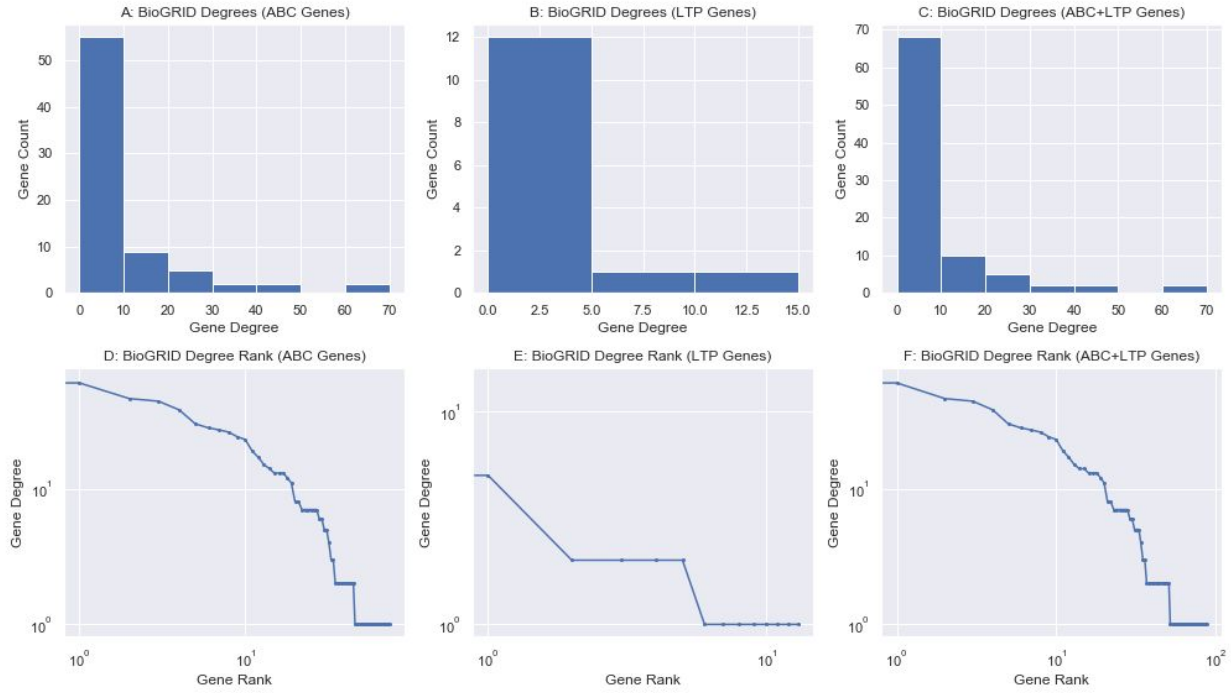


Fig. 5 Degrees of the ABC/LTP subset of BioGRID dataset.
(A) 55 of the 75 BioGRID compatible ABC transporter genes have between zero and ten connections with

one gene having the greatest degree of 63. (D) The degree-rank further illustrates that most genes have very few connections and there are only 2 genes with over fifty connections and 0 genes with over a hundred connections. (B, E) LTP genes seem to have much fewer connections than ABC transporter genes, with 13 out of 14 genes having less than or equal to five connections and a single gene with 14 genes. (C, F) Overall, the LTP genes don't seem to contribute much to the overall degrees of the BioGRID subset.

Gene Discovery

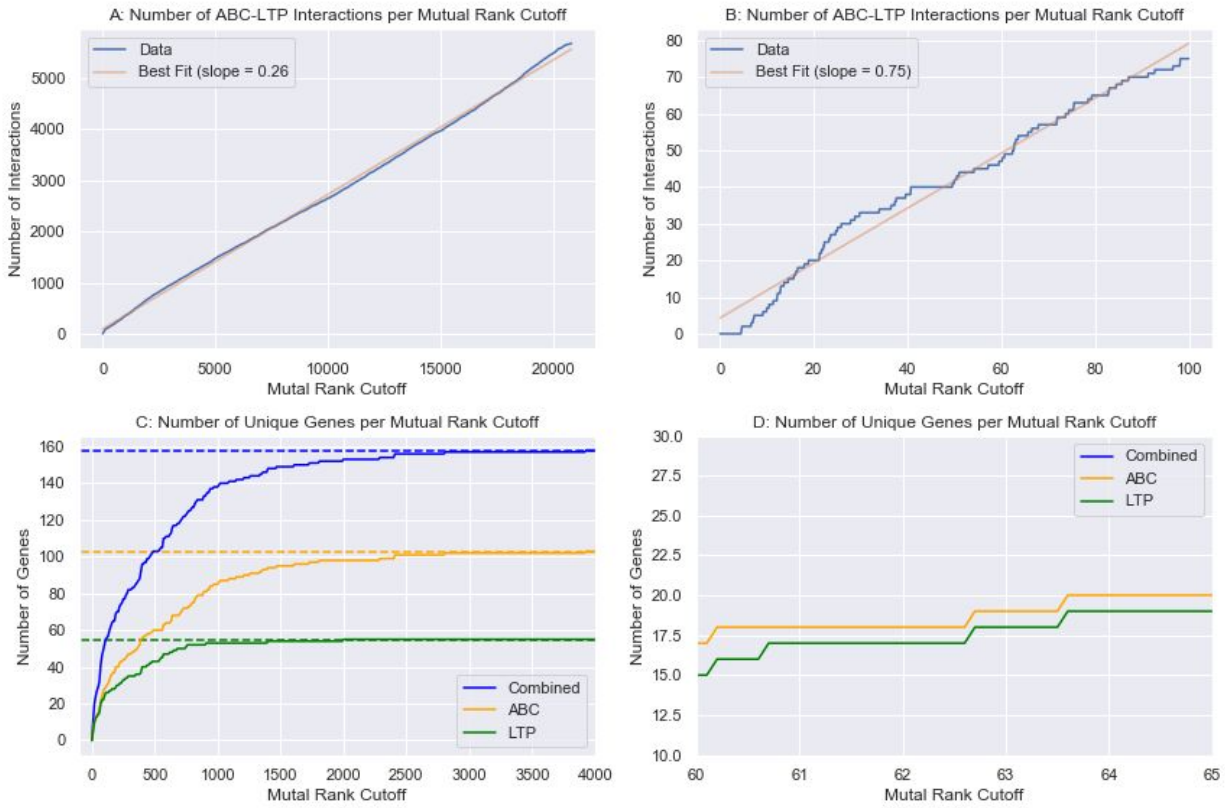


Fig. 6 Number of ABC Transporter-LTP Interactions and Number of Unique Genes for Increasing Mutual Rank Cutoffs

By running the 'GetTopGenes' function multiple times, the number of ABC transporter-LTP interactions were plotted against increasing MR cutoff values. (A) In the full-scale graph, the maximum number of interactions, 5,665, is reached at a cutoff of greater than 20811.77, which is the maximum MR value of an ABC transporter-LTP interaction. It is important to note that the maximum possible MR value in the entire ATTED dataset is 20,818 since there are 20,819 total genes. Overall, there was a clear linear trend with the line of best fit having a slope of 0.26 when including all ABC transporter-LTP interactions with MR cutoffs up to 20812.77. The linear trends tell us that the ABC transporter-LTP interactions are not clustered around specific regions of MR values but are instead evenly spread out. (B) When narrowing down the scope with MR values only up to a 100, the slope of the line of best fit increases up to 0.75. It was interesting to note that while the number of interactions increases steadily with the MR values, there was a brief stall between MR values of 40 and 50. (C, D) As before, the 'GetTopGenes' function was run multiple times to plot the number of unique genes against increasing MR cutoff values. Unlike before, there is not a linear trend. Instead, the data appears to follow a logarithmic trend. (C) It appears that it takes a lower MR cutoff to reach all unique LTP genes, taking an MR cutoff of 2,000. All unique ABC transporter genes are reached

with a cutoff of 3,940 and 1,195 interactions are present. While more interactions are charted in MRs past 3,940 in A, no new genes are involved. As such, any interactions past this point is most-likely spurious.

	ABC Entrez ID	LTP Entrez ID	MR	ABC TAIR ID	LTP TAIR ID	ABC Gene Name	LTP Gene Name
0	831202	831237	4.58	AT5G13580	AT5G13900	ABC-2 type transporter family protein(ABCG6)	Bifunctional inhibitor/lipid-transfer protein/...
1	824675	837046	4.61	AT3G55090	AT1G05450	ABC-2 type transporter family protein(ABCG16)	Bifunctional inhibitor/lipid-transfer protein/...
2	838363	839688	6.66	AT1G17840	AT1G27950	white-brown complex-like protein(ABCG11)	glycosylphosphatidylinositol-anchored lipid pr...
3	818307	820024	7.03	AT2G37300	AT3G08770	transmembrane protein(ABCI16)	lipid transfer protein 6(LTP6)
4	841761	837046	7.26	AT1G53270	AT1G05450	ABC-2 type transporter family protein(ABCG10)	Bifunctional inhibitor/lipid-transfer protein/...
5	832061	831237	9.15	AT5G19410	AT5G13900	ABC-2 type transporter family protein(ABCG23)	Bifunctional inhibitor/lipid-transfer protein/...
6	831202	819426	9.94	AT5G13580	AT2G48140	ABC-2 type transporter family protein(ABCG6)	Bifunctional inhibitor/lipid-transfer protein/...
7	828702	839688	10.41	AT4G25960	AT1G27950	P-glycoprotein 2(ABCB2)	glycosylphosphatidylinositol-anchored lipid pr...
8	838363	823481	11.33	AT1G17840	AT3G43720	white-brown complex-like protein(ABCG11)	Bifunctional inhibitor/lipid-transfer protein/...
9	824675	831237	12.17	AT3G55090	AT5G13900	ABC-2 type transporter family protein(ABCG16)	Bifunctional inhibitor/lipid-transfer protein/...

Fig. 7 First 10 values of Top Fifty ABC Transporter-LTP Interactions
The table details the Entrez ID, TAIR ID, and common name for each ABC transporter and LTP gene along with MR values for each interaction. There are seven promising interactions with MRs less than 10. These seven interactions are made up of 11 unique genes (6 ABCs and 5 LTPs) while the total 50 interactions are made of 35 unique genes (18 ABCs and 17 LTPs).

The 18 unique ABC transporter genes are ABCA7, ABCB2, ABCC7, ABCG1, ABCG2, ABCG6, ABCG10, ABCG11, ABCG12, ABCG13, ABCG16, ABCG18, ABCG20, ABCG23, ABCG28, ABCG29, ABCG32, and ABCI16. The 17 unique LTP genes include 11 members of the bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily, a member of the glycolipid transfer protein (GLTP), glycosylphosphatidylinositol-anchored LTP (LTPG1), and four other LTPs (LTP, LTP1, LTP2, LTP6). The functions of all these LTPs are not yet clear.

Of the 18 identified ABC transporter genes, ABCG11, ABCG12, ABCG13, and ABCI16 have been functionally characterized in 2011.[3] ABCG11 plays a role in cutin formation (flower, seed), cuticle formation (leaf, shoot), and suberin formation (root). ABCG12 is involved in cuticle formation (leaf, shoot). ABCG13 takes part in cutin formation (flower), and ABCI16 is linked with aluminum tolerance of root tissue. According to DAVID, ABCB2 is involved in auxin transport. ABCG1 is involved in the cellular response to nematode and pollen wall assembly. ABCG2 is involved in suberin formation. ABCG6 is involved in nematode response and suberin formation. ABCG11 has additional roles in wound response, salt stress response, abscisic acid response, stem and cotyledon vascular tissue pattern formation. ABCG12 has additional roles in salt stress response, abscisic acid response, and wax biosynthesis. ABCG13 has additional roles in petal epidermis patterning, and salt stress response. ABCG16 is involved in pollen wall assembly. ABCG20 is involved in suberin formation. ABCG29 is involved in lignin formation. ABCG32 is involved in cuticle formation. Finally, ABCC7, ABCG29, and ABCG32 all function in various drug resistance.

Since many of the above ABC transporters have already been characterized, literature reviews should be conducted to look for experimental confirmations of LTP interactions involving these ABC transporters. In fact, Allan DeBono, Ph.D, from the University of British Columbia experimentally characterized a linkage between ABCG11 and ABCG12 to a glycosylphosphatidylinositol-anchored lipid transfer protein (LTPG) that is highly expressed in the epidermis during cuticle biosynthesis in Arabidopsis thaliana inflorescence stems.[6] This linkage is the third entry in the table! Thus, it seems the analysis of the ATTED co-expression dataset seems to be effective in finding potential interactions between ABC transporters and LTPs.

Limitations and Future Directions

The main limitation in this project was the inability to run my 'ShortestDistances' function to compile shortest path data for each ABC transporter-LTP pair because the Networkx A* shortest path algorithm required more memory than available and a python script was unable to run properly on the High Performance Computing Center at the Institute for Cyber-Enabled Research (ICER). Even upon simplifying the Networkx graph by only selecting the nodes within one node of the target genes, the algorithm would not run. Without this step, it is hard to rule out spurious correlations from the ATTED dataset. Another concern was that only 75 of the 131 ABC transporter genes and 14 of the 85 LTP genes were present in the BioGRID dataset. This issue was resolved by looking at the raw MR values from the ATTED dataset but even this dataset had 28 missing ABC transporter genes and 30 LTP genes. In the future, it would be a good idea to integrate more co-expression data to accomodate for the missing genes. While my project revealed probable ABC transporter-LTP interaction, they require experimental testing with knockouts and pulldown assays of the list of genes to confirm any co-functionality and rule out spurious relations. Additionally, a literature review can be conducted on the identified genes to see whether any other experimental links between the given proteins encoded by the genes have been found.

Citations

1. Krämer, Ute. "Planting Molecular Functions in an Ecological Context with Arabidopsis Thaliana." ELife, vol. 4, 25 Mar. 2015, doi:10.7554/elife.06100.
2. Davidson AL, Dassa E, Orelle C, Chen J (Jun 2008). "Structure, function, and evolution of bacterial ATP-binding cassette systems". Microbiology and Molecular Biology Reviews. 72 (2): 317–64, table of contents. doi:10.1128/MMBR.00031-07.
3. Kang, Joohyun, et al. "Plant ABC Transporters." The Arabidopsis Book, vol. 9, 6 Dec. 2011, doi:10.1199/tab.0153.
4. Salminen, Tiina A., et al. "Lipid Transfer Proteins: Classification, Nomenclature, Structure, and Function." Planta, vol. 244, no. 5, 25 Aug. 2016, pp. 971–997., doi:10.1007/s00425-016-2585-4.
5. Edqvist, Johan, et al. "Plant Lipid Transfer Proteins: Are We Finally Closing in on the Roles of These Enigmatic Proteins?" Journal of Lipid Research, vol. 59, no. 8, 19 Mar. 2018, pp. 1374–1382., doi:10.1194/jlr.r083139.
6. DeBono, Allan, et al. "Arabidopsis LTPG Is a Glycosylphosphatidylinositol-Anchored Lipid Transfer Protein Required for Export of Lipids to the Plant Surface." The Plant Cell, vol. 21, no. 4, Apr. 2009, pp. 1230–1238., doi:10.1105/tpc.108.064451.
7. Takeshi Obayashi, Yuichi Aoki, Shu Tadaka, Yuki Kagaya, Kengo Kinoshita, ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index, Plant and Cell Physiology, Volume 59, Issue 1, January 2018, Page e3, doi:10.1093/pcp/pcx191
8. Obayashi T., Hayashi S., Saeki M., Ohta H. & Kinoshita K. (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. Nucleic Acids Research 37,D987–D991.
9. Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, Mike Tyers, BioGRID: a general repository for interaction datasets, Nucleic Acids Research, Volume 34, Issue suppl_1, 01 January 2006, Pages D535–D539, doi:10.1093/nar/gkj109
10. Berardini, Tanya Z., et al. "The Arabidopsis Information Resource: Making and Mining the 'Gold Standard' Annotated Reference Plant Genome." Genesis, vol. 53, no. 8, 22 July 2015, pp. 474–485., doi:10.1002/dvg.22877.
11. Verrier, Paul J., et al. "Plant ABC Proteins – a Unified Nomenclature and Updated Inventory." Trends in Plant Science, vol. 13, no. 4, 1 Apr. 2008, pp. 151–159., doi:10.1016/j.tplants.2008.02.001.
12. Huang, Da Wei, et al. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." Nature Protocols, vol. 4, no. 1, 2008, pp. 44–57., doi:10.1038/nprot.2008.211.
13. Serin, Elise A. R., et al. "Learning from Co-Expression Networks: Possibilities and Challenges." *Frontiers in Plant Science*, vol. 7, 8 Apr. 2016, doi:10.3389/fpls.2016.00444.

Glossary

- **Pearson correlation coefficient (PCC):** A statistic that measures linear correlation between two variables, in this case Gene A and Gene B, and has a range of -1 to 1.
- **Co-expression Correlation Rank:** $Rank_{A \rightarrow B}$ is the rank of the Pearson correlation coefficient (PCC) for Genes A and B relative to all the PCCs of A. Many genes have low-PCC gene pairs but are still

relevant, so the PCCs are converted to ranks to allow for weak but still significant gene co-expression by mapping the PCC to between 0 and the number of other genes in the co-expression set.

- **Mutual Rank (MR):** The geometric average of the two correlation ranks for a given gene-gene interaction to correct for the asymmetry of the correlation rank. In other words, $Rank_{A\to B} \neq Rank_{B\to A}$. $MR_{AB} = \sqrt{Rank_{A\to B} * Rank_{B\to A}}$
- **TAIR ID:** A plain text systematic name used by the TAIR database.
- **Entrez ID:** The gene identifier from the Entrez gene database
- **Nematodes:** Common parasitic insects also known as roundworms
- **Cuticle:** A protective film made up of hydrophobic polymers that covers the epidermis of leaves, young shoots and other above-ground, non-bark covering surface.
- **Cutin:** One of the two main hydrophobic polymers that make up the cuticle.
- **Suberin:** Waxy substance in cell walls of corky tissue.
- **Lignin:** Polymer that provides rigid structure for cell walls of plant tissue, especially bark.
- **Abscisic acid (ABA:** Common developmental plant hormone.
- **Inflorescence:** Modified part of the shoot of seed plants where flowers are formed.

Project Experience

Going into this, I was without a doubt terrified of having to conduct an open-ended project within the span of a semester. Having no major previous experience in bioinformatics or code generation, I struggled to find an interesting topic for which I had sufficient previous knowledge to progress. While daunting at first, I feel that the open-endedness pushed me to understand the work of a real researcher: only a few major deadlines are provided but the rest has to be self-paced to avoid wasting time and money. The project helped me not only develop skills in time & project management but also taught me file management and data processing. I definitely have a newfound appreciation of the intellectual and technical challenges of a bioinformatician. I remember it took me almost a week to just find a way to read in my data correctly and structure it in an efficient manner. Additionally, I found the proposal-writing aspect of the project enlightening since I wasn't aware that there was so much to do before even beginning a project. While conducting the project itself, I learned many new tricks with the file manipulation and Networkx algorithms. While at times, the project proved to be frustrating like when the shortest path algorithm crashed my computer, it was fun applying all the things I learned in CMSE 201 and 202 to a completely new topic. On the whole, the project definitely better prepared me for future computational biology research and general code development.

Supplementary Information

Calculation of Mutual Rank by ATTED From ChipGenes

First, the microarrays of 16033 ChipGenesare normalized using Robust Multi-array Average (RMA) and combined into a single gene expression table. Next, the Pearson correlation coefficients (PCC) were calculated for each sample pair using the following formula where $RE_{g,s}$ is the relative expression of gene G in sample S, $\overline{RE_{S1}}$ is the average relative expression of all genes in sample S1, and $\overline{RE_{S2}}$ is the average relative expression of all genes in sample S2:

$$R_{s1,s2} = \frac{\sum_g (RE_{g,s1} - \overline{RE_{s1}})(RE_{g,s2} - \overline{RE_{s2}})}{\sqrt{\sum_g (RE_{g,s1} - \overline{RE_{s1}})^2 \sum_g (RE_{g,s2} - \overline{RE_{s2}})^2}}$$

Next, $J_{S1,S2} = \frac{max(0, R_{S1,S2} - C)}{1 - C}$ represents the pairwise sample redundancy and C is the cutoff threshold optimized to be 0.4. J_S is the sum of all pairwise sample redundancies between sample S and all samples. This is then used to calculate the weight of the specific sample S. W_S is the inverse square root of J_S. The weights of samples are used to calculated a weighted PCC for a specific gene pair using the following formula where

$RE_{g,s}$ is the relative expression of gene G in sample S, $\overline{RE}_{G1}$ is the weighted average relative expression of gene G1, and $\overline{RE}_{G2}$ is the average relative expression of gene G2:

$$COR_{g1,g2} = \frac{\sum_s W_s (RE_{g1,s} - \overline{RE}_{g1})(RE_{g2,s} - \overline{RE}_{g2})}{\sqrt{\sum_s W_s (RE_{g1,s} - \overline{RE}_{g1})^2 \sum_s W_s (RE_{g2,s} - \overline{RE}_{g2})^2}}$$

This PCC is converted to MR using $MR_{G1,G2} = \sqrt{Rank_{G1 \to G2} * Rank_{G1 \to G2}}$.

	ABC Entrez ID	LTP Entrez ID	MR	ABC TAIR ID	LTP TAIR ID	ABC Gene Name	LTP Gene Name
0	831202	831237	4.58	AT5G13580	AT5G13900	ABC-2 type transporter family protein(ABCG6)	Bifunctional inhibitor/lipid-transfer protein/...
1	824675	837046	4.61	AT3G55090	AT1G05450	ABC-2 type transporter family protein(ABCG16)	Bifunctional inhibitor/lipid-transfer protein/...
2	838363	839688	6.66	AT1G17840	AT1G27950	white-brown complex-like protein(ABCG11)	glycosylphosphatidylinositol-anchored lipid pr...
3	818307	820024	7.03	AT2G37300	AT3G08770	transmembrane protein(ABCI16)	lipid transfer protein 6(LTP6)
4	841761	837046	7.26	AT1G53270	AT1G05450	ABC-2 type transporter family protein(ABCG10)	Bifunctional inhibitor/lipid-transfer protein/...
5	832061	831237	9.15	AT5G19410	AT5G13900	ABC-2 type transporter family protein(ABCG23)	Bifunctional inhibitor/lipid-transfer protein/...
6	831202	819426	9.94	AT5G13580	AT2G48140	ABC-2 type transporter family protein(ABCG6)	Bifunctional inhibitor/lipid-transfer protein/...
7	828702	839688	10.41	AT4G25960	AT1G27950	P-glycoprotein 2(ABCB2)	glycosylphosphatidylinositol-anchored lipid pr...
8	838363	823481	11.33	AT1G17840	AT3G43720	white-brown complex-like protein(ABCG11)	Bifunctional inhibitor/lipid-transfer protein/...
9	824675	831237	12.17	AT3G55090	AT5G13900	ABC-2 type transporter family protein(ABCG16)	Bifunctional inhibitor/lipid-transfer protein/...
10	831202	837046	12.29	AT5G13580	AT1G05450	ABC-2 type transporter family protein(ABCG6)	Bifunctional inhibitor/lipid-transfer protein/...
11	831202	819425	12.86	AT5G13580	AT2G48130	ABC-2 type transporter family protein(ABCG6)	Bifunctional inhibitor/lipid-transfer protein/...
12	841761	831237	12.95	AT1G53270	AT5G13900	ABC-2 type transporter family protein(ABCG10)	Bifunctional inhibitor/lipid-transfer protein/...
13	838363	841970	13.64	AT1G17840	AT1G55260	white-brown complex-like protein(ABCG11)	Bifunctional inhibitor/lipid-transfer protein/...
14	841575	815994	14.53	AT1G51500	AT2G15050	ABC-2 type transporter family protein(ABCG12)	lipid transfer protein(LTP)
15	828702	823481	15.73	AT4G25960	AT3G43720	P-glycoprotein 2(ABCB2)	Bifunctional inhibitor/lipid-transfer protein/...
16	838363	818436	16.11	AT1G17840	AT2G38540	white-brown complex-like protein(ABCG11)	lipid transfer protein 1(LP1)
17	824519	819426	16.57	AT3G53510	AT2G48140	ABC-2 type transporter family protein(ABCG20)	Bifunctional inhibitor/lipid-transfer protein/...
18	832061	837046	17.98	AT5G19410	AT1G05450	ABC-2 type transporter family protein(ABCG23)	Bifunctional inhibitor/lipid-transfer protein/...
19	824677	820024	18.80	AT3G55110	AT3G08770	ABC-2 type transporter family protein(ABCG18)	lipid transfer protein 6(LTP6)
20	832061	819426	21.19	AT5G19410	AT2G48140	ABC-2 type transporter family protein(ABCG23)	Bifunctional inhibitor/lipid-transfer protein/...
21	823932	830121	21.28	AT3G47780	AT4G39670	ABC2 homolog 6(ABCA7)	Glycolipid transfer protein (GLTP) family prot...
22	841575	823481	21.70	AT1G51500	AT3G43720	ABC-2 type transporter family protein(ABCG12)	Bifunctional inhibitor/lipid-transfer protein/...
23	831202	821833	22.01	AT5G13580	AT3G22620	ABC-2 type transporter family protein(ABCG6)	Bifunctional inhibitor/lipid-transfer protein/...
24	828702	818436	22.26	AT4G25960	AT2G38540	P-glycoprotein 2(ABCB2)	lipid transfer protein 1(LP1)
25	841575	841970	23.24	AT1G51500	AT1G55260	ABC-2 type transporter family protein(ABCG12)	Bifunctional inhibitor/lipid-transfer protein/...
26	841571	820024	23.57	AT1G51460	AT3G08770	ABC-2 type transporter family protein(ABCG13)	lipid transfer protein 6(LTP6)
27	824519	831237	24.66	AT3G53510	AT5G13900	ABC-2 type transporter family protein(ABCG20)	Bifunctional inhibitor/lipid-transfer protein/...
28	824519	837046	25.04	AT3G53510	AT1G05450	ABC-2 type transporter family protein(ABCG20)	Bifunctional inhibitor/lipid-transfer protein/...
29	828702	841970	25.80	AT4G25960	AT1G55260	P-glycoprotein 2(ABCB2)	Bifunctional inhibitor/lipid-transfer protein/...
30	832061	819425	27.88	AT5G19410	AT2G48130	ABC-2 type transporter family protein(ABCG23)	Bifunctional inhibitor/lipid-transfer protein/...
31	824519	819425	28.74	AT3G53510	AT2G48130	ABC-2 type transporter family protein(ABCG20)	Bifunctional inhibitor/lipid-transfer protein/...
32	841575	839688	29.71	AT1G51500	AT1G27950	ABC-2 type transporter family protein(ABCG12)	glycosylphosphatidylinositol-anchored lipid pr...
33	841761	819426	33.93	AT1G53270	AT2G48140	ABC-2 type transporter family protein(ABCG10)	Bifunctional inhibitor/lipid-transfer protein/...
34	841575	818436	36.49	AT1G51500	AT2G38540	ABC-2 type transporter family protein(ABCG12)	lipid transfer protein 1(LP1)
35	817232	823481	37.20	AT2G26910	AT3G43720	pleiotropic drug resistance 4(ABCG32)	Bifunctional inhibitor/lipid-transfer protein/...
36	841761	819425	37.57	AT1G53270	AT2G48130	ABC-2 type transporter family protein(ABCG10)	Bifunctional inhibitor/lipid-transfer protein/...
37	818520	818435	39.56	AT2G39350	AT2G38530	ABC-2 type transporter family protein(ABCG1)	lipid transfer protein 2(LTP2)
38	818312	819426	40.64	AT2G37360	AT2G48140	ABC-2 type transporter family protein(ABCG2)	Bifunctional inhibitor/lipid-transfer protein/...
39	818312	837046	40.65	AT2G37360	AT1G05450	ABC-2 type transporter family protein(ABCG2)	Bifunctional inhibitor/lipid-transfer protein/...
40	818312	816352	49.57	AT2G37360	AT2G18370	ABC-2 type transporter family protein(ABCG2)	Bifunctional inhibitor/lipid-transfer protein/...
41	817232	841970	49.99	AT2G26910	AT1G55260	pleiotropic drug resistance 4(ABCG32)	Bifunctional inhibitor/lipid-transfer protein/...
42	818312	831237	50.43	AT2G37360	AT5G13900	ABC-2 type transporter family protein(ABCG2)	Bifunctional inhibitor/lipid-transfer protein/...
43	824675	821833	50.97	AT3G55090	AT3G22620	ABC-2 type transporter family protein(ABCG16)	Bifunctional inhibitor/lipid-transfer protein/...
44	836195	840520	54.19	AT5G60740	AT1G36150	ABC transporter family protein(ABCG28)	Bifunctional inhibitor/lipid-transfer protein/...
45	831202	816352	57.13	AT5G13580	AT2G18370	ABC-2 type transporter family protein(ABCG6)	Bifunctional inhibitor/lipid-transfer protein/...
46	820498	830121	59.42	AT3G13100	AT4G39670	multidrug resistance-associated protein 7(ABCC7)	Glycolipid transfer protein (GLTP) family prot...
47	820881	819038	60.14	AT3G16340	AT2G44300	pleiotropic drug resistance 1(ABCG29)	Bifunctional inhibitor/lipid-transfer protein/...
48	831202	825024	60.67	AT5G13580	AT3G58550	ABC-2 type transporter family protein(ABCG6)	Bifunctional inhibitor/lipid-transfer protein/...
49	818312	825024	62.36	AT2G37360	AT3G58550	ABC-2 type transporter family protein(ABCG2)	Bifunctional inhibitor/lipid-transfer protein/...

Fig. 2S: Full length version of Fig. 7.
This is the full table of the top fifty ABC transporter-LTP gene interactions.