Evolution of the multi-tRNA synthetase complex and its role in cancer

Supplementary Information

Medical Subject Headings (MeSH) term-based collection of the research publications dealing with ARSs and related diseases.

To retrieve the articles describing the association of ARSs and disease, we queried Pubmed server with the MeSH terms representing ARS and disease. For “ARS”, we used a MeSH term of “amino acyl-trna synthetases”. For disease, we chose 6 MeSH terms representing major disease category including "Neoplasms", "Metabolic Diseases", "Autoimmune Diseases", "Central Nervous System Diseases", "Neurodegenerative Diseases", "Infection", and "Respiratory Tract Diseases”. To generate Pubmed query, we combined the MeSH terms ARS and each disease by “AND” operator and added Pubmed option “[MeSH Terms]”. This constraint is to find the research articles only if they carry both ARS and Disease MeSH terms. An example of keyword used in our research is “amino acyl-trna synthetases” [MeSH Terms] AND “Neoplasms” [MeSH Terms]. XML files from the Pubmed query results were downloaded and further analyzed to count cumulated number of the published papers by disease category and by year.

Functional enrichment analysis

For each species, we first collected interactors of MSC-ARSs/AIMPs and Free-ARSs. For each set of interactors, we then performed the enrichment analysis of gene ontology biological processes (GOBPs) to identify cellular processes represented by the interactors using DAVID (1). The GOBPs with $p < 0.05$ computed from DAVID were selected as the ones enriched by the set of the interactors.

Calculation of enrichment scores

To construct GOBP tree, we first collected directed acyclic graphs of GOBP terms describing parent-daughter relationships between GOBP terms from the Gene Oncology Consortium (release date: 2017-07-07) (2). For each of the representative cellular processes (e.g., DNA repair or RNA processing in Fig. 3C), we selected GOBP terms enriched by at least one set of MSC-ARS/AIMP interactors in five species (*S. cerevisiae, C. elegans, D. melanogaster, M. musculus* and *H. sapiens*) and then constructed a GOBP tree showing parent-daughter relationship among them using the selected GOBP terms. The GOBP tree is composed of a root GOBP and its daughter GOBPs. For the GOBP tree, the average enrichment score was evaluated as follows: 1) counter the number of species (N_{root}) in which the root GOBP was enriched by MSC-ARS/AIMP interactors; 2) select daughter GOBPs with which three or more MSC-ARS/AIMP interactors were annotated in more than half of N_{root}; and 3) compute the average of Z-scores of the root GOBP and the selected daughter GOBPs in the tree. The Z-score was computed as $-N(P)^{-1}$ where P is the enrichment P-value nd $N(.)^{-1}$ is the inverse normal distribution.

Construction of network models

To construct network models, we collected the list of MSC-ARSs/AIMPs interactors annotated with the GOBPs in *H. sapiens*, KEGG pathways related to two non-catalytic functions of ARSs/AIMPs (mTOR signaling and DNA repair) and interactors previously reported to have associations with these non-catalytic functions. For each non-catalytic function, we then constructed a network model describing the collected interactions among the non-ARS/AIMP

interactors of MSC-ARSs/AIMPs. The nodes in each network model were arranged based on their pathways in KEGG pathway database (3).

mRNA expression-survival correlation

To evaluate correlations of mRNA expression levels with patient survivals, we first collected read counts for 60,483 gene features for tumor samples of 26 types of human cancers (Table S1) from NCI Genomic Data Commons (GDC) Data Portal (4). For each type of human cancer, the read counts were normalized using the TMM normalization method (5) in the edgeR package (6). The normalized counts were converted to $\log_2$-read-counts after adding one to the normalized counts and the $\log_2$-read-counts for the samples were further normalized within the dataset for the type of cancer using the quantile normalization method (7). For each of ARSs/AIMPs and their interactors, we divided the samples in a cancer type into two groups (top and bottom 25% of patients with highest and lowest mRNA expression levels, respectively) based on the normalized $\log_2$-read-counts and evaluated differences in survival curves between the two groups using log-rank test with Kaplan-Meier estimation (8).

Supplementary figure legend

Figure S1. The illustration of noncatalytic domains of MSC components.

Human MSC component have several appended domains or motifs. The conserved catalytic domains and tRNA recognition domains are shown in dark or light gray boxes. Glutathion S-Transferase-like domain (GST) are shown in the EPRS, MRS AIMP2 and AIMP3, while WHEP domains are shown in ERPS, MRS. Leucine zipper motif is also observed in AIMP1, AIMP2, and RRS. AIMP1 has an EMAPII domain which is involved in several cellular response. While DRS and KRS have the lysine rich domains in the N-terminal region, LRS and IRS have the appended sequences. QRS has the appended sequences in the C-terminal regions.

Figure S2. The number of the published research articles on the disease association of ARSs from 1970 to 2018. The inset shows the list of 7 major diseases defined by MeSH term.

Figure S3. Relationship of MSC-ARSs/AIMPs to cancer patient survival

(**A**) The survival curves of liver hepatocellular carcinoma (LIHC) patients whose IRS expression levels were grouped to the top (red line) and bottom (black) 25% of the total patients. (**B**) The survival curves of breast invasive carcinoma (BRCA) patients whose MRS expression levels were grouped to the top (red line) and bottom (black) 25%. (**C**) The survival curves of head and neck squamous cell carcinoma (HNSC) patients whose AIMP1 expression levels were grouped to the top (red line) and bottom (black) 25%. P represents the significance of the mRNA expression-survival correlation.

Table S1. List of 26 types of human cancers used to analyze the mRNA expression-survival correlation

The mRNA-sequencing data generated form 26 types of human cancers by The Cancer Genome Atlas (TCGA) studies (number of primary tumor samples > 100) were collected from NCI Genomics Data Commons (GDC) Data Portal (5). Only the mRNA-sequencing data for primary tumor samples were used for evaluating the correlations

Type	Name	Primary site	Number of samples (primary tumor)
BRCA	Breast invasive carcinoma	Breast	1,102
UCEC	Uterine corpus endometrial carcinoma	Uterus	551
KIRC	Kidney renal clear cell carcinoma	Kidey	538
LUAD	Lung adenocarcinoma	Lung	533
LGG	Brain lower grade glioma	Brain	511
LUSC	Lung squamous cell carcinoma	Lung	502
THCA	Thyroid carcinoma	Thyroid	502
HNSC	Head and Neck squamous cell carcinoma	Head and Neck	500
PRAD	Prostate adenocarcinoma	Prostate	498
COAD	Colon adenocarcinoma	Colorectal	478
BLCA	Bladder urothelial carcinoma	Bladder	414
STAD	Stomach adenocarcinoma	Stomach	375
OV	Ovarian serous cystadenocarcinoma	Ovary	374
LIHC	Liver hepatocellular carcinoma	Liver	371

CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	Cervix	304
KIRP	Kidney renal papillary cell carcinoma	Kidney	288
SARC	Sarcoma	Soft Tissue	259
PCPG	Pheochromocytoma and paraganglioma	Adrenal Gland	178
PAAD	Pancreatic adenocarcinoma	Pancreas	177
READ	Rectum adenocarcinoma	Colorectal	166
ESCA	Esophagel carcinoma	Esophagus	161
GBM	Glioblastoma multiforme	Brain	156
LAMI	Acute myeloid leukemia	Bone Marrow	151
TGCT	Testicular germ cell tumors	Testis	150
THYM	Thymoma	Thymus	119
SKCM	Skin cutaneous melanoma	Skin	103

Table S2. PPI (protein-protein interaction) databases used for evolutionary analysis of ARS networks.

For each species, the PPI databases and the numbers of PPIs and proteins used for analysis are listed. Only the PPIs with experimental evidence were used. In the case of STRING, the PPIs with experimental score > 0.4 were used. The PPIs for dimerization were filtered out before the analysis.

Species	Number of interactions	Number of proteins	Databases
S. cerevisiae	214,967	5,862	BioGRID (9), CCSB interactome datasets (10-12), DIP (13), HitPredict (14), IntAct (15), MINT (16), STRING (17)
C. elegans	70,277	8,116	BioGRID (9), CCSB interactome datasets (18, 19), DIP (13), HitPredict (14), IntAct (15), MINT (16), STRING (17)
D. melanogaster	167,225	11,176	BioGRID (9), DIP (13), DroID (20), HitPredict (14), IntAct (15), MINT (23), STRING (24)
M. musculus	82,254	11,502	BioGRID (9), DIP (13), HitPredict (14), IntAct (15), MINT (16), STRING (17)
H. sapiens	337,756	17,445	BioGRID (9), CCSB interactome datasets (21-24), DIP (13), HitPredict (14), HPRD (25), HTRIdb (26), IntAct (15), MINT (16), STRING (17)

References

1. Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44-57 (2009).

2. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).

3. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

4. Grossman, R.L. et al. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* **375**, 1109-12 (2016).

5. Robinson, M.D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25 (2010).

6. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-40 (2010).

7. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-93 (2003).

8. Bewick, V., Cheek, L. & Ball, J. Statistics review 12: Survival analysis. *Critical Care* **8**, 389 (2004).

9. Stark, C. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-9 (2006).

10. Yu, H. et al. High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104-10 (2008).

11. Collins, S.R. et al. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* **6**, 439-50 (2007).

12. Reguly, T. et al. Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol* **5**, 11 (2006).

13. Salwinski, L. et al. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**, D449-51 (2004).

14. Patil, A., Nakai, K. & Nakamura, H. HitPredict: a database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Res* **39**, D744-9 (2011).

15. Orchard, S. et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358-63 (2014).

16. Licata, L. et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* **40**, D857-61 (2012).

17. Szklarczyk, D. et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**, D447-52 (2015).

18. Li, S. et al. A map of the interactome network of the metazoan C. elegans. *Science* **303**, 540-3 (2004).

19. Simonis, N. et al. Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nat Methods* **6**, 47-54 (2009).

20. Murali, T. et al. DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila. *Nucleic Acids Res* **39**, D736-43 (2011).

21. Rual, J.F. et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-8 (2005).

22. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212-26 (2014).

23. Venkatesan, K. et al. An empirical framework for binary interactome mapping. *Nat Methods* **6**, 83-90 (2009).

24. Yu, H. et al. Next-generation sequencing to generate interactome datasets. *Nat Methods* **8**, 478-80 (2011).

25. Keshava Prasad, T.S. et al. Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-72 (2009).

26. Bovolenta, L.A., Acencio, M.L. & Lemke, N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics* **13**, 405 (2012).