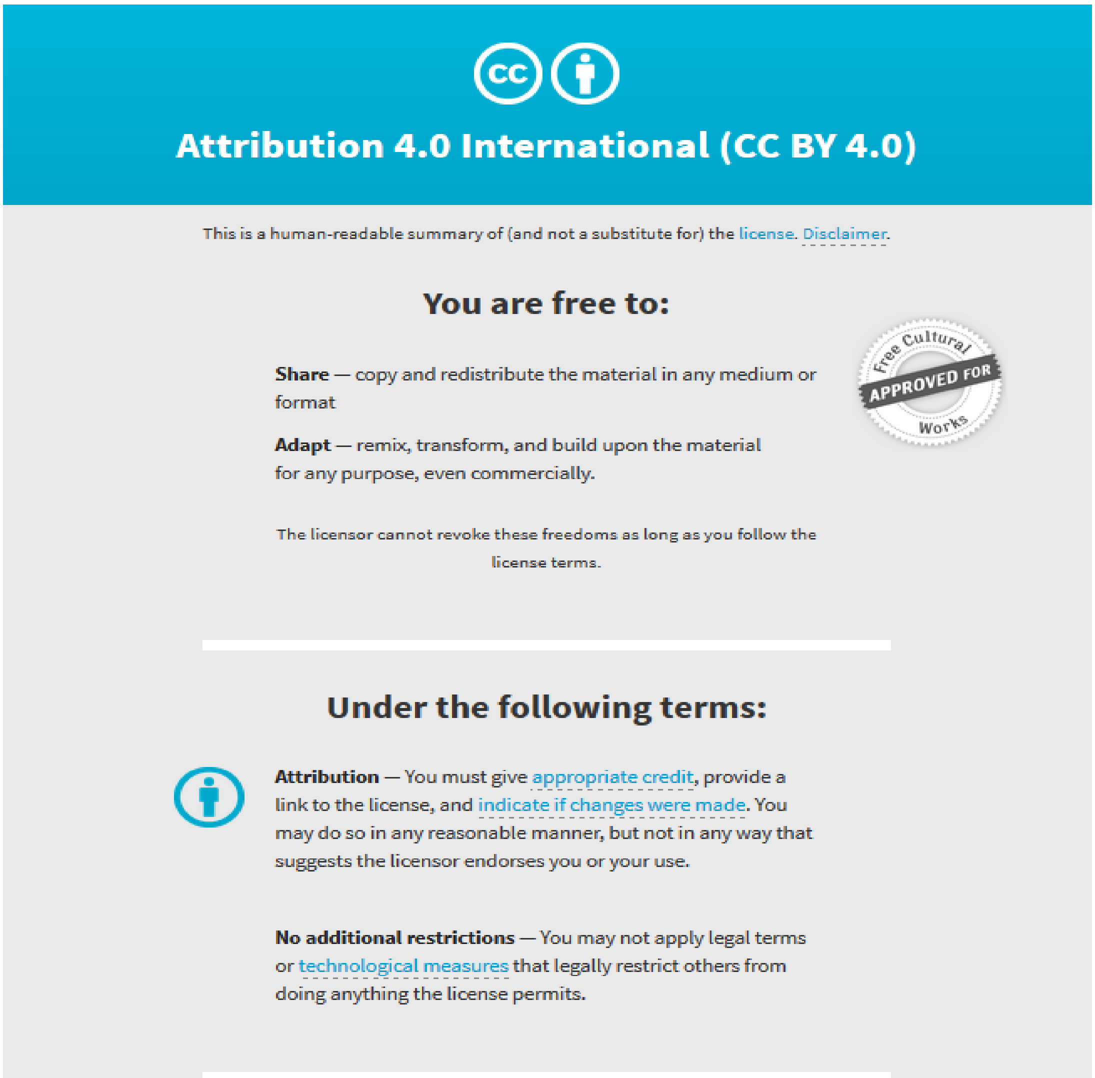


Free Cultural
APPROVED FOR
Works

independent association studies of different ethnicities, platforms or even species, while avoiding the technical difficulties when performing meta-analysis directly on the marker-level association data [11]. To run meta-MSEA, users simply need to navigate to the Meta-MSEA tab, and upload multiple datasets following the same workflow as previously described for MSEA to generate results for individual datasets as well as the pathway/network-level meta-analysis results. The result files produced by Meta-MSEA follow the same layout as MSEA.

Weighted key driver analysis (wKDA)

wKDA aims to pinpoint key regulator genes or key drivers (KDs) of the disease related gene sets from MSEA or meta-MSEA using gene network topology and edge weight information. Specifically, wKDA first screens the network for candidate hub genes. Then the disease gene sets are overlaid onto the subnetworks of the candidate hubs to identify KDs whose neighbors are enriched with disease genes.

Data preparation

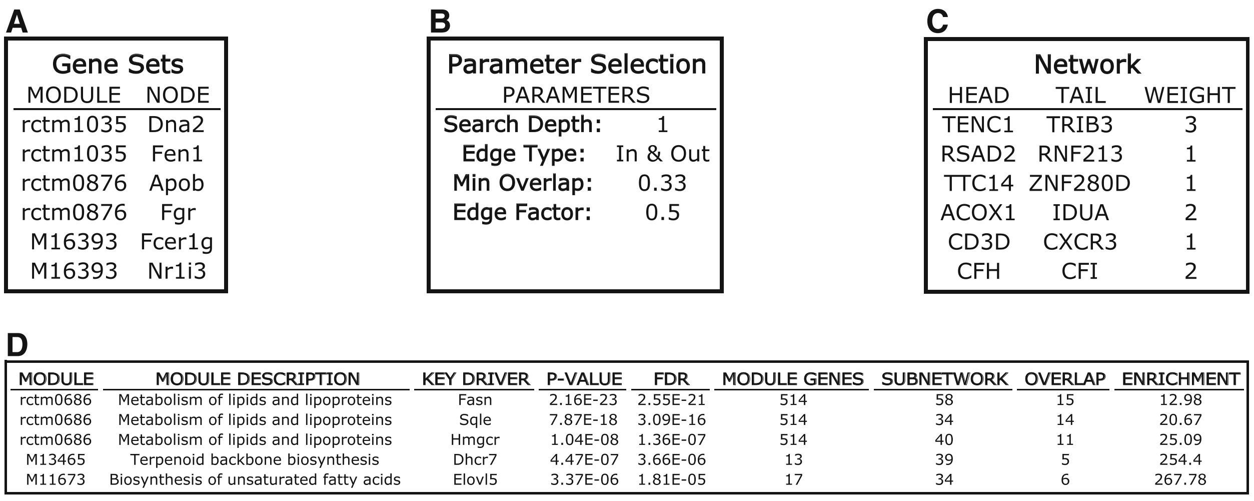
Two types of files are required for wKDA: 1) disease-associated gene sets (Fig. 4a) and 2) molecular networks (Fig. 4c). wKDA can be run as either the continuing step of MSEA or meta-MSEA or as an independent step (Fig. 2). If the user elects to continue wKDA from MSEA or meta-MSEA, then the enriched gene sets from these analyses will be used as the disease-associated gene sets. If the user elects to run wKDA as a separate module, they must upload their own gene sets to the web server or they can use the pre-loaded sample gene set for testing. With regards to molecular networks, wKDA supports a wide range of directed and undirected regulatory networks. wKDA is designed to utilize edge weight information in gene networks, which could be connection strength or reliability measures. If no edge weight information is available, wKDA can also operate by considering equal weights to all edges. The web server provides a collection of tissue-specific Bayesian networks previously constructed in human and mouse studies (Additional file 1: Table S1). There are also a large number of publicly available network resources, such as protein-protein interactions (PPI) [12], BioGRID [13], GeneMANIA [14] and GIANT [15], which could be used in wKDA.

Parameter setting

Core wKDA parameters include 1) search depth, which specifies the number of layers to expand in the network when determining the neighboring genes of candidate KDs for enrichment assessment, and 2) edge directionality, which specifies whether to neglect edge directionality (incoming and outgoing) or require the candidate hubs to be upstream of neighbor genes (only outgoing) for networks that carry directionality information (Fig. 4b). Additional parameters are described in detail in the online tutorial.

Result interpretation

Summary results of wKDA will be displayed on the webpage (Fig. 4d), which reports top 5 KDs for each disease gene set along with the statistics. Users could also download four detailed results files: 1) "wKDA_kd_*p*values.txt", a summary table of all KDs ranked by *p* -values and FDR; 2) "wKDA_kd_full_results.txt", providing detailed statistics on all KDs identified; 3) "wKDA_kd_tophits.txt", *p*-value summary table for only the top KDs for each disease gene set. 4) "wKDA_hub_structure.txt", specifying hub-cohub



Gene Sets

MODULE	NODE
rctm1035	Dna2
rctm1035	Fen1
rctm0876	Apob
rctm0876	Fgr
M16393	Fcer1g
M16393	Nr1i3

Parameter Selection

PARAMETERS	
Search Depth:	1
Edge Type:	In & Out
Min Overlap:	0.33
Edge Factor:	0.5

Network

HEAD	TAIL	WEIGHT
TENC1	TRIB3	3
RSAD2	RNF213	1
TTC14	ZNF280D	1
ACOX1	IDUA	2
CD3D	CXCR3	1
CFH	CFI	2

MODULE	MODULE DESCRIPTION	KEY DRIVER	P-VALUE	FDR	MODULE GENES	SUBNETWORK	OVERLAP	ENRICHMENT
rctm0686	Metabolism of lipids and lipoproteins	Fasn	2.16E-23	2.55E-21	514	58	15	12.98
rctm0686	Metabolism of lipids and lipoproteins	Sqle	7.87E-18	3.09E-16	514	34	14	20.67
rctm0686	Metabolism of lipids and lipoproteins	Hmgcr	1.04E-08	1.36E-07	514	40	11	25.09
M13465	Terpenoid backbone biosynthesis	Dhcr7	4.47E-07	3.66E-06	13	39	5	254.4
M11673	Biosynthesis of unsaturated fatty acids	Elovl5	3.37E-06	1.81E-05	17	34	6	267.78

Fig. 4 Screenshots showing a step-by-step overview of the wKDA module using sample data on the web server. **a** Gene set format. **b** Parameters. **c** Network format. **d** Example result summary table containing the IDs of the disease associated gene sets, their KDs, *p* - values and FDRs of the KDs, gene counts of KD subnetworks, and fold enrichment of the KD subnetworks for disease genes

relationship. The co-hub structure is useful to group KDs with highly overlapping network topology, and to retrieve list of independent KDs for more efficient prioritization. Additionally, wKDA provides Cytoscape-ready files that can be used in Cytoscape [16] for a more customized visualization than the included web-based network visualization module.

Network visualization

Our web server provides a convenient module to allow users visualize top KDs and subnetworks using Cytoscape Web v0.8 [17]. The top 5 KDs for each disease gene set from wKDA will be automatically visualized, as exemplified in Fig. 5a. The visualization is interactive so that users can make real-time changes such as zooming in on a node of interest by only considering that particular subnetwork (Fig. 5b) or by filtering a subnetwork based on the edge weight information (Fig. 5c), as detailed in the tutorial page.

Application example

As an illustration of our web server's workflow and analysis results, we applied the Mergeomics web server to a publically available low-density lipoprotein (LDL) GWAS dataset from the GLGC consortium [18]. All files mentioned within this section are provided as example files on the web server. To correct for LD between

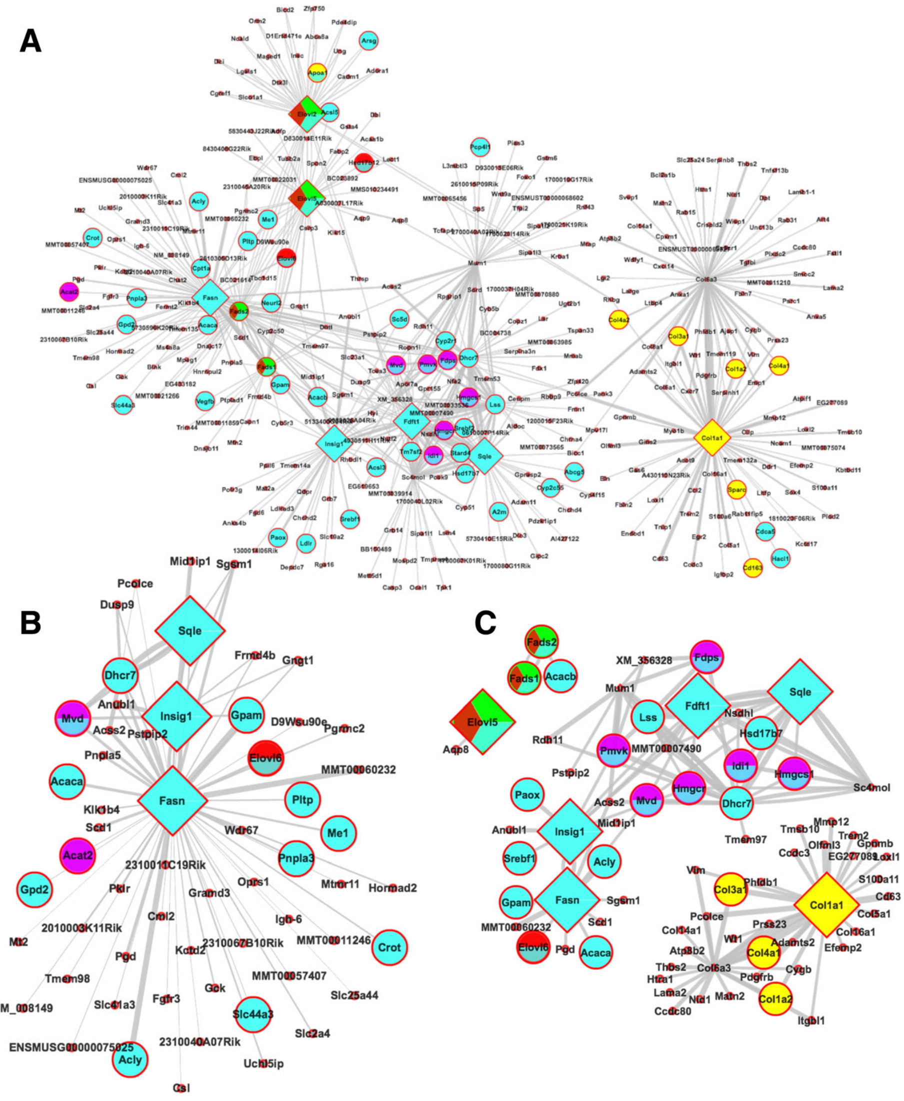


Fig. 5 Screenshots of the network visualization module output generated by the default sample files in the wKDA module. **a** Overview of a network comprised of top KDs for LDL-associated gene sets, with 5 KDs for each gene set displayed. KDs are depicted as diamond nodes and node colors indicating gene set membership. **b** Focused subnetwork view of key driver of interest *Fasn* when filtering the network in (**a**) by submitting a "Fasn" query or right clicking the node. **c** The network can be filtered to display only nodes with edge weights higher than a certain threshold: 4.0 in this case

SNPs, we used the MDF analysis module and the following input files: the GLGC LDL GWAS summary statistics (SNPs, –log10 *p* - values), SNP-gene mapping based on 50 kb chromosomal distance, and the Hapmap CEU LD file containing SNPs with $r^2 > 0.7$ as the Marker Dependency file. We also filtered the GWAS loci by only consider the top 50 % SNPs ranked by *p* - values to reduce random noise from the weaker association spectrum.

After correction for marker-dependency using MDF, the resulting association and mapping files were used directly as input for MSEA (Fig. 3a-b) using gene permutation and default setting for the other parameters (Fig. 3c), to test for enrichment for canonical pathways collected from KEGG, Biocarta and Reactome databases (Fig. 3d-e). Upon completion of MSEA, a summary table was produced which details the top pathways ranked by FDR along with descriptions of the pathways and top associated genes and SNPs in each pathway. As exemplified in Fig. 3f, "Metabolism of lipids and lipoproteins", "biosynthesis of unsaturated fatty acids", and "terepenoid backbone biosynthesis" were three of the top pathways identified among others. *APOC2*, *APOE*, and *LDLR* were listed as the top associated genes in the "metabolism of lipids and lipoproteins" pathway, and their corresponding SNPs were also provided in the summary table. Links to detailed result tables were also displayed for file download. Furthermore, these top pathways were checked for overlaps and merged if significant overlaps in gene membership between pathways were identified. A summary table of the merged pathways was also displayed (not shown).

To identify potential KDs and subnetworks for the LDL-associated pathways, the merged pathways were used directly as input for wKDA (Fig. 4a). wKDA was run using default parameters (Fig. 4b) and a liver Bayesian network (Fig. 4c). Upon completion, a summary table is produced (Fig. 4d) which lists the top 5 KDs for each merged module and information about their local subnetwork structure. For example, *Fasn* was a KD for the Metabolism of Lipids and Lipoproteins pathway. Links to detailed result tables were also displayed for file download.

The wKDA results can be viewed directly using the interactive visualization feature, which by default illustrates the top 5 KDs for each gene set and their local subnetworks with disease genes highlighted (Fig. 5a). The networks can be filtered by selecting a particular KD of interest to focus on (Fig. 5b) or by removing edges below an edge weight cutoff to focus on high confidence network connections (Fig. 5c). To facilitate further customization of network views, Cytoscape-ready files can be downloaded for external visualization.

Conclusions

We have implemented the Mergeomics pipeline as a user-friendly, publicly available web server that can facilitate multidimensional omics data integration to expedite novel discoveries. The web server also pre-populates a wide range of publically available data sources. Users can apply the pipeline to their own data in conjunction with any preloaded data to identify disease-associated pathways, gene networks, and key regulators. The web server includes step-by-step tutorials, examples and visualization tools in a web-based platform. The flexibility of the web server to accommodate various omics data types and to conduct pathway and network-level meta-analysis of multiple studies of different design will boost our ability to integrate big data.

Additional file

Additional file 1: Table S1. List of public datasets available for access in the Mergeomics web server. (DOCX 128 kb)

Acknowledgements

This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group. We thank Brian Pape and Rafaella D'Auria from UCLA Hoffman2 IDRE for technical support, and Qingying Meng, Yuqi Zhao, Zeyben Kurt, Katherine Wei, Brandon Tsai, Rio Barrere-Cain, Xingyi Shi, Christine Sun for valuable feedbacks.

Funding

DA is supported by the NIH-NCI National Cancer Institute T32CA201160. XY is supported by the American Heart Association Scientist Development Grant 13SDG17290032, Leducq Foundation, and NIDDK R01DK104363. VPM is supported by the American Heart Association Postdoctoral Fellowship 13POST17240095. LS is supported by China Scholarship Council Scholarship and UCLA Eureka Scholarship.

Availability of supporting data and materials

The datasets supporting the conclusions of this article are available at http://mergeomics.research.idre.ucla.edu/Download/Sample_Files/

- **Project name:** Mergeomics
- **Project home page:** http://mergeomics.research.idre.ucla.edu/
- **Operating system(s):** Platform independent
- **Programming language:** R, C++ (server side scripts); no language required client-side
- **Other requirements**: Internet browser; Flash Player required for network visualization
- **License:** Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/)
- **Any restrictions to use by non-academics:** no restriction

Authors' contributions

DA, AB, and XY designed the web server. DA and AB implemented the web server. DA, LS, and XY wrote the manuscript, with inputs from AB and VPM. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

All data sources mentioned in the study are publically available summary level information that requires no ethical approval or consent.

Author details

[1]Department of Integrative Biology and Physiology, University of California, Los Angeles, CA 90095, USA. [2]South Australian Health and Medical Research Institute, Adelaide, Australia. [3]School of Biological Sciences, University of Adelaide, Adelaide, Australia. [4]Institute of Health Sciences, University of Oulu, Oulu, Finland.

Received: 21 March 2016 Accepted: 30 August 2016
Published online: 09 September 2016

References

1. Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. Nat Rev Genet. 2014;15(1):34–48.
2. Joyce AR, Palsson BØ. The model organism as a system: integrating'omics' data sets. Nat Rev Mol Cell Biol. 2006;7(3):198–210.
3. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet. 2005;37(7):710–7.
4. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467(7317):832–8.
5. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
6. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580–5.
7. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, Bell JT, Yang TP, Meduri E, Barrett A, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012;44(10):1084–9.
8. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014;42(Database issue):D472–7.
9. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
10. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
11. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet. 2013;14(6):379–89.
12. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell. 2005;122(6):957–68.
13. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, et al. The BioGRID interaction database: 2013 update. Nucleic Acids Res. 2013;41(Database issue):D816–23.
14. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38(Web Server issue):W214–20.
15. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet. 2015;47(6):569–76.
16. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
17. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. Cytoscape Web: an interactive web-based network browser. Bioinformatics. 2010;26(18):2347–8.
18. Global Lipids Genetics C, Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45(11):1274–83.