Mergeomics: Integrative Network Analysis of Omics Data

May 3, 2016

Ville-Petteri Makinen, Le Shu, Yuqi Zhao, Zeyneb Kurt, Bin Zhang, Xia Yang
Department of Integrative Biology and Physiology, University of California, Los Angeles
South Australian Health and Medical Research Institute
Department of Genetics and Genomics Sciences, Mount Sinai School of Medicine

zhaoyuqi616@ucla.edu, zeyneb@ucla.edu, icestrike@ucla.edu, xyang123@ucla.edu

If you use mergeomics in published research, please cite:
Shu L, Zhao Y, Kurt Z, Byars S, Tukiainen T, Kettunen J, Ripatti S, Zhang B, Inouye M, Makinen VP, Yang X. Mergeomics: integration of diverse genomics resources to identify pathogenic perturbations to biological systems. bioRxiv doi: http://dx.doi.org/10.1101/036012

Contents

1 Introduction

The recent revolution in genomic technologies has enabled the generation of massive amount of molecular data encompassing genetic, transcriptomic, epigenomic, metabolomics, and proteomics, which have become easily accessible in public domains and private sectors. It is increasingly recognized that analysis of individual types of data separately only reveals a fraction of the complex biology and often misses the key players driving diseases, making

multi-dimensional big data integration an urgent need. Mergeomics package is developed as a capable and flexible pipeline to integrate various types of disease association data such as genetic association (e.g, GWAS or exome sequencing), transcriptome-wide association (e.g., TWAS from microarray or RNA sequencing studies), and epigenetic association (e.g., EWAS from methylome association studies), functional genomics (such as eQTLs and ENCODE annotations), biological pathways, and gene networks to identify disease-associated gene sub-networks and key regulatory genes.

In this tutorial, we will assume that you have downloaded the standalone Mergeomics package from

http://mergeomics.research.idre.ucla.edu/Download/Package/

Mergeomics_0.99.1.tar.gz and install it using the following commands:

```
> # install.packages("Mergeomics_0.99.1.tar.gz", repos = NULL,
> # type="source")
```

However, if you want, you can also try out these examples with our web-server at http://mergeomics.research.idre.ucla.edu. If you have any questions regarding this document or the described software, please contact us: Xia Yang (xyang123@ucla.edu).

2 Tutorials

The whole Mergeomics pipeline can be divided into two major components: 1) Marker Set Enrichment Analysis (MSEA), 2) Weighted Key Driver Analysis (wKDA). In the following step, we will illustrate how to implement Mergeomics using one sample dataset, the descriptions of which are as follows.

Marker-disease association summary result: Genome-wide association studies (GWAS) of high-density lipoprotein cholesterol (HDLC) from Global Lipids Genetics Consortium [1];

Gene-marker mapping file: The GWAS loci were projected to genes within 40 Kb distance from gene region;

Functionally related gene-sets: The coexpression networks were reconstructed in mouse/human liver tissues, containing 2674 modules, were adopted [2-6];

Gene regulatory networks: The Bayesian networks generated from mouse/human liver tissues were included [2-6].

2.1 Data Preprocessing

Before we take any further steps, the datasets should be preprocessed and filtered for dependency structure among markers (e.g. Linkage disequilibrium), as marker dependency may cause artifacts and biases in the downstream analysis. We have provided an external C++ program named MDPrune

(http://mergeomics.research.idre.ucla.edu/Download/MDPrune/), which can remove marker dependency while preferentially keeping those with a strong statistical association with traits.

The commands are listed as follows:

mdprune MARFile MAPFile MDSFile OutPath [Cutoff]

All files should contain tab delimited plain text where;

MARFile is the association study summary file where the VALUE column contains the -log10 association P values.

MAPFile is the Marker-Gene mapping file;

MDSFile is the marker dependency structure file (MARKERa MARKERb WEIGHT).

Here, we chose a moderate cutoff (R^2>0.7) for the dataset.

2.2 One-step Mergeomics

Here, we first provide an one-step way to run the whole pipeline. Then, the step-by-step processing will be described in detail in separate sections.

```
> ############################################################
> #######    One-step analysis for Mergeomics       #########
> ############################################################
> ## Import library scripts.
> # library(Mergeomics)
> ################ MSEA (Marker set enrichment analysis)  ###
> # job.msea <- list()
> # job.msea$label <- "hdlc"
> # job.msea$folder <- "Results"
> # job.msea$genfile <- system.file("extdata",
> # "genes.hdlc_040kb_ld70.human_eliminated.txt", package="Mergeomics")
> # job.msea$marfile <- system.file("extdata",
> # "marker.hdlc_040kb_ld70.human_eliminated.txt", package="Mergeomics")
> # job.msea$modfile <- system.file("extdata",
```

```
> # "modules.mousecoexpr.liver.human.txt", package="Mergeomics")
> # job.msea$inffile <- system.file("extdata", "coexpr.info.txt",
> # package="Mergeomics")
> # job.msea <- ssea.start(job.msea)
> # job.msea <- ssea.prepare(job.msea)
> # job.msea <- ssea.control(job.msea)
> # job.msea <- ssea.analyze(job.msea)
> # job.msea <- ssea.finish(job.msea)
> #########  Create intermediary datasets for KDA ############
> # syms <- tool.read(system.file("extdata", "symbols.txt",
> # package="Mergeomics"))
> # syms <- syms[,c("HUMAN", "MOUSE")]
> # names(syms) <- c("FROM", "TO")
> # job.kda <- ssea2kda(job.msea, symbols=syms)
> #######   wKDA (Weighted key driver analysis)    ###########
> # job.kda$netfile <- system.file("extdata",
> # "network.mouseliver.mouse.txt", package="Mergeomics")
> # job.kda <- kda.configure(job.kda)
> # job.kda <- kda.start(job.kda)
> # job.kda <- kda.prepare(job.kda)
> # job.kda <- kda.analyze(job.kda)
> # job.kda <- kda.finish(job.kda)
> ######  Prepare network files for visualization   #########
> ## Creates the input files for Cytoscape (http://www.cytoscape.org/)
> # job.kda <- kda2cytoscape(job.kda)
```

2.3 Marker set enrichment analysis (MSEA)

The purpose of the MSEA is to leverage genome/epigenome wide association data, function genomics, canonical pathways and/or data-driven gene modules for identifying causal subnetworks for disease/traits.

```
> ############################################################
> ## Import Mergeomics library.
> # library("Mergeomics")
> ## create an empty list for setting parameters
> # job.msea <- list()
> ## Next, label your project
> # job.msea$label <- "HDLC"
> ## The pathway size varies from 1 to a few thousands and will
```

```
> ## introduce bias to the analysis. We set criteria for the
> ## min. (mingenes) and max. (maxgenes) gene size for the pathways.
> # job.msea$maxgenes <- 500
> # job.msea$mingenes <- 10
> ## set the output folder
> # job.msea$folder <- "./Result"
> ## The parameter genfile defines the Marker-to-Gene mapping file
> ## It contains two columns, GENE and MARKER, delimited by tab
> # job.msea$genfile <- system.file("extdata",
> # "genes.hdlc_040kb_ld70.human_eliminated.txt", package="Mergeomics")
> ## The parameter marfile defines the Disease association data file
> ## It contains two columns, MARKER and VALUE, delimited by tab
> ## Here, the marfile comes from the GWAS file after marker
> ## dependency pruning, so the VALUE is the minus log10 transformed
> # job.msea$marfile <- system.file("extdata",
> # "marker.hdlc_040kb_ld70.human_eliminated.txt", package="Mergeomics")
> ## The modfile defines the pathway information, which could come
> ## from knowledge-based databases (such as KEGG, and Reactome)
> ## or data-driven data sets (such as co-expression modules).
> ## It contains two columns, MOUDLE and GENE, delimited by tab
> # job.msea$modfile <- system.file("extdata",
> # "modules.mousecoexpr.liver.human.txt", package="Mergeomics")
> ## The inffile provides the basic descriptions for the pathways
> ## It contains three columns, MODULE, SOURCE, and DESCR, which
> ## provide information for pathway IDs corresponding to the
> ## pathway names in modfile, the sources of the pathways, and
> ## pathway annotations
> # job.msea$inffile <- system.file("extdata", "coexpr.info.txt",
> # package="Mergeomics")
> ## Then, MSEA will run for ~30 minutes to ~2 hours
> # job.msea <- ssea.start(job.msea)
> # job.msea <- ssea.prepare(job.msea)
> # job.msea <- ssea.control(job.msea)
> # job.msea <- ssea.analyze(job.msea)
> # job.msea <- ssea.finish(job.msea)
> ############################################################
```

Part of the original result is listed in Table 1. The pathways satisfying certain cutoffs (e.g. FDR<30%) will be used for the next step and can be further annotated using diverse tools, such as DAVID (http://david.abcc.ncifcrf.gov/).

Table 1: Summary result of MSEA

Pathways	Pvalues	FDRs	Descriptions
5099	3.48E-04	6.05%	5099:liver
4588	1.05E-03	13.64%	4588:liver
4737	2.01E-03	21.02%	4737:liver
4128	4.44E-03	30.24%	4128:liver
5117	5.22E-03	32.03%	5117:liver
6645	5.24E-03	32.05%	6645:liver
5104	5.50E-03	32.29%	5104:liver
4094	7.65E-03	33.93%	4094:liver
4932	7.95E-03	34.12%	4932:liver

2.3.1 ModuleMerge

Usually, pathways/modules collected from different sources will have certain degree of redundancies, or refer to the same processes. In this case, users could merge the disease-related pathways into relatively independent gene sets after running MSEA.

```
> ############################################################
> # job <-list()
> # job$folder <- c("module_merge")
> ## The moddata and modinfo come from the significant pathways
> ## in MSEA
> # moddata <- tool.read("PATHtoDATAFILES/Significant_pathways.txt",
> # c("MODULE","GENE"))
> # modinfo <- tool.read("PATHtoDATAFILES/Significant_pathways.info.txt",
> # c("MODULE","SOURCE","DESCR"))
> ## Merge and trim overlapping modules.
> # rmax <- 0.2
> # moddata$OVERLAP <- moddata$MODULE
> # moddata <- tool.coalesce(items=moddata$GENE, groups=moddata$MODULE,
> # rcutoff=rmax)
> # moddata$MODULE <- moddata$CLUSTER
> # moddata$GENE <- moddata$ITEM
> # moddata$OVERLAP <- moddata$GROUPS
> # moddata <- moddata[,c("MODULE", "GENE", "OVERLAP")]
> # moddata <- unique(moddata)
> ## Mark modules with overlaps.
```

```
> # for(i in which(moddata$MODULE != moddata$OVERLAP))
> #     moddata[i,"MODULE"] <- paste(moddata[i,"MODULE"], "..", sep=",")
> ## Save module info for KDA.
> # modfile <- "merged_modules.txt"
> # tool.save(frame=unique(moddata[,c("MODULE", "GENE", "OVERLAP")]),
> # file=modfile, directory=job$folder)
> ############################################################
```

Users are also recommended to analyze the merged genesets with MSEA again to confirm that they are still significantly associated with disease.

2.3.2 Meta-MSEA

When multiple disease association datasets are available, meta-MSEA can be used to conduct meta-analysis at pathway or network level. This function allows user to achieve maximal power by combining results from independent association studies of different ethnicity, platform or even species, while evading the technical difficulties when performing meta-analysis directly on the marker-level association data.

```
> ############################################################
> ## Assume there are three MSEA objects passed down by
> ## ssea.finish()
> # job.metamsea = list()
> # job.metamsea$job1 = job.msea1
> # job.metamsea$job2 = job.msea2
> # job.metamsea$job3 = job.msea3
> # job.metamsea = ssea.meta(job.metamsea,"meta_label",
> # "meta_folder")
> ############################################################
```

2.4 Weighted key driver analysis (wKDA)

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

```
> ############################################################
> # job.kda <- list()
```

```
> # job.kda$label<-"HDLC"
> ## parent folder for results
> # job.kda$folder<-"./Results"
> ## Input a network
> ## columns: TAIL HEAD WEIGHT
> # system.file("extdata", "network.mouseliver.mouse.txt",
> # package="Mergeomics")
> ## Gene sets derived from ModuleMerge, containing two columns,
> ## MODULE, NODE, delimited by tab
> # job.kda$modfile<-"HDLC_Combined.txt"
> ## Annotation file for the gene sets
> # job.kda$inffile<-"HDLC_Combined.anno.txt"
> ## "0" means we do not consider edge weights while 1 is
> ## opposite.
> # job.kda$edgefactor<-0.0
> ## The searching depth for the KDA
> # job.kda$depth<-1
> ## "0" means we do not consider the directions of the
> ## regulatory interactions
> ## while 1 is opposite.
> # job.kda$direction<-0
> ## Let us run KDA!
> # job.kda <- kda.start(job.kda)
> # job.kda <- kda.prepare(job.kda)
> # job.kda <- kda.analyze(job.kda)
> # job.kda <- kda.finish(job.kda)
> ############################################################
```

2.5 Network visualization

For the current version, we generate Cytoscape-ready (http://www.cytoscape.org/) input files to visualize the network of key driver (KD) genes. Cytoscape-ready files are generated for illustrating either the neighborhoods of top five KDs of each module or the neighborhoods of a particular gene (node) list. By default, log files for each module's top five KDs' neighborhoods are prepared. Number of the top KDs can be specified by the user. Node and edge list files are created separately for Cytoscape. Additionally, top key driver name list for each module and color map of the modules are logged in the relevant files. kda2cytoscape function provides users:

Table 2: Summary result of wKDA

Key Drivers	MODULES	Pvalues	FDRs	Descriptions
Itih4	4932,..	2.90E-14	0.00%	4932:liver
Clstn3	136	6.92E-14	0.00%	136:liver
Pmvk	4407	1.40E-13	0.00%	4407:liver
Acss2	4407	1.54E-13	0.00%	4407:liver
Gpam	4407	3.82E-13	0.00%	4407:liver
Aqp8	4407	8.26E-13	0.00%	4407:liver
Itih4	5099	1.56E-12	0.00%	5099:liver
Lpin1	4084	2.21E-12	0.00%	4084:liver

- To illustrate top five KD neighborhoods for the modules, whose pathway or module name is listed within "modules" parameter (all modules are selected by default),
- To specify a particular number of KDs for each module with "ndrivers" (default value is five),
- To search and illustrate the neighborhoods of a given node list with "node.list" parameter (default value is NULL; if this is not specified by the user, top 5 KDs of each module and their neighborhoods will be illustrated; if "node.list" parameter is specified, it will be prioritized rather than top 5 KDs illustration),
- To examine the KD neighborhoods for a given depth (default depth is 1; hence, only the directly connected neighbor nodes of key drivers are taken into account).

```
> ############################################################
> # job.kda <- kda2cytoscape (job.kda, node.list=NULL,
> # modules=NULL, ndrivers=5, depth=1)
> ############################################################
```

Generated file names and their descriptions are given below:

- **kda2cytoscape.top.kds.txt:** Top KDs of the modules are listed in this file. Number of the key drivers can be set by user with "ndrivers" parameter.
- **kda2cytoscape.edges.txt:** Edge lists of the generated graph.

- **kda2cytoscape.nodes.txt:** Node lists of the generated graph. This file includes the optional formatting information for the nodes of the graph. Field/attribute list provided for each node is as follows:
  - NODE: gene symbol
  - LABEL: node label (same as above)
  - COLOR: background color of the node, it also shows the module membership of the node. If a node is member of multiple modules/pathways, only one color among them will be assigned as background color, but in the URL field this multiple module membership will be represented by a pie chart (node is divided into several sectors to show the multi-membership of the node). If a node is not a member for any of the pathways, the color will be "grey"
  - SIZE: size of the node. If the node is a KD it will be twice larger than ordinary nodes.
  - SHAPE: shape of the node. If the node is a KD, it will be in diamond-shape, otherwise it is illustrated with circle.
  - URL: pie chart created by Google chart tools. If the node is member of multiple modules, node circle will be divided as a pie chart and colored by the module colors it belongs to
  - LABELSIZE: text size of the node. If the node is a KD, its font size will be larger than ordinary nodes.
- **module.color.mapping.txt:** Color mapping for the modules, i.e. one color is assigned to each module.

An example graph for top 5 KDs'neighborhoods of a given pathway is illustrated in Figure 1. Different colors represent modules/pathways, while the larger nodes are the key driver genes.

Note that: SHAPE feature is not used in the Cytoscape in Fig 1. All the listed attributes for nodes are optional.

3 Citation

If you use Mergeomics in published work, please cite: Shu L, Zhao Y, Kurt Z, Byars S, Tukiainen T, Kettunen J, Ripatti S, Zhang B, Inouye M, Makinen VP, Yang X. Mergeomics: integration of diverse genomics resources to identify pathogenic perturbations to biological systems.

bioRxiv doi: http://dx.doi.org/10.1101/036012.

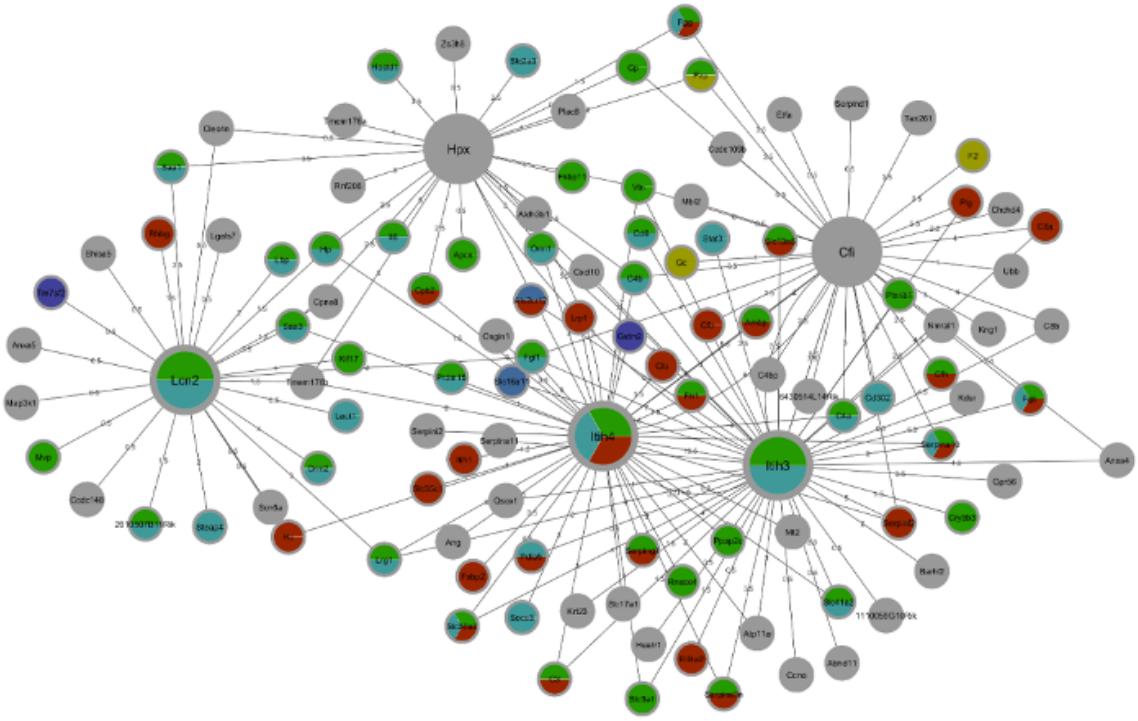

Figure 1: The key driver gene subnetwork

4 References

[1] Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, et al. (2013) Discovery and refinement of loci associated with lipid levels. Nature Genetics 45.

[2] Derry JMJ, Zhong H, Molony C, MacNeil D, GuhaThakurta D, et al. (2010) Identification of genes and networks driving cardiovascular and metabolic phenotypes in a mouse F2 intercross. PLoS ONE 5: e14319. doi:10.1371/journal.pone.0014319.

[3] Wang SS, Schadt EE, Wang H, Wang X, Ingram-Drake L, et al. (2007) Identification of pathways for atherosclerosis in mice: integration of quantitative trait locus analysis and global gene expression data. Circ Res 101: e11-e30. doi:10.1161/CIRCRESAHA.107.152975.

[4] Yang X, Schadt EE, Wang S, Wang H, Arnold AP, et al. (2006) Tissue-specific expression and regulation of sexually dimorphic genes in mice. Genome Res 16: 995-1004. doi:10.1101/gr.5217506.

[5] Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6: e107. doi:10.1371/journal.pbio.0060107.

[6] Tu ZZ, Keller MPM, Zhang CC, Rabaglia MEM, Greenawalt DMD, et al. (2012) Integrative analysis of a cross-loci regulation network identifies App as a gene regulating insulin secretion from pancreatic islets. PLoS Genet 8: e1003107-e1003107. doi:10.1371/journal.pgen.1003107.