GSMA: an approach to identify robust global and test Gene Signatures using Meta-Analysis

Supplementary Material

Adib Shafi[1], Tin Nguyen[2], Azam Peyvandipour[1] and Sorin Draghici[1,3]*

[1]**Department of Computer Science, Wayne State University, Detroit, MI 48202**
[2]**Department of Computer Science and Engineering, University of Nevada, Reno, NV 89557**
[3]**Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI 48202**

1 Dataset Description

The R programming language was used to generate all the results included in the proposed manuscript. We download the raw probe level datasets from Gene Expression Omnibus. All the datasets used in this manuscript are described in Table S1.

Datasets from Affymetrix platform are normalized using RMA background adjustment, quantile normalization and median polish summarization. We use the *threestep* function from *affyPLM* package to achieve this goal [1]. For probe to gene mapping, standard genome wide annotation packages are used from bioconductor. Median values are taken whenever multiple probes mapped to the same gene. Datasets from Illumina platform are normalized using the *neqc* function from *limma* package [2]. Finally dataset from Agilent platform is normalized using *limma* package as well.

Table S1: Description of the 21 gene expression datasets used in the manuscript.
AD: Alzheimer's disease, Inf: Influenza, FC: Frontal cortex, TC: Temporal cortex, EC: Entorhinal cortex, HIP: Hippocampus, MTG: Medial Temporal Gyrus, PC: Posterior Cingulate, SFG: Superior Frontal Gyrus and VCX: Primary Visual Cortex

	Datasets	Disease	Discovery/ validation	Number of samples	Contrast	Tissue	Platform
1	GSE48350	AD	Discovery	253	173 Norm. vs 80 AD	EC, PCG	Affymetrix HG U133 Plus 2.0
2	GSE63061	AD	Discovery	273	134 Norm. vs 139 AD	Blood	Illumina HumanHT-12 4.0
3	GSE63060	AD	Discovery	249	104 Norm. vs 145 AD	Blood	Illumina HumanHT-12 3.0
4	GSE26927	AD	Discovery	118	18 Norm. vs 100 AD	EC	Illumina HumanRef-8 2.0
5	GSE1297	AD	Discovery	31	9 Norm. vs 22 AD	HIP	Affymetrix HG U133A
6	GSE15222	AD	Validation	363	187 Norm. vs 176 AD	TC	Illumina Sentrix HumanRef-8
7	GSE5281	AD	Validation	161	74 Norm. vs 87 AD	EC, MTG, PC, SFG, HIP, PVC	Affymetrix HG U133 Plus 2.0
8	GSE36980	AD	Validation	79	47 Norm. vs 32 AD	FC, TC, HIP	Affymetrix Human Gene 1.0 ST
9	GSE28146	AD	Validation	30	8 Norm. vs 22 AD	HIP	Affymetrix HG U133 Plus 2.0
10	GSE39420	AD	Validation	21	7 Norm. vs 14 AD	PC	Affymetrix Human Gene 1.1 ST
11	GSE12685	AD	Validation	14	8 Norm. vs 6 AD	FC	Affymetrix HG U133A
12	GSE17156	Inf	Discovery	34	17 Norm. vs 17 Inf	Peripheral blood	Affymetrix HG U133A 2.0
13	GSE42026	Inf	Discovery	52	33 Norm. vs 19 Inf	Whole blood	Illumina HumanHT-12 3.0
14	GSE21802	Inf	Discovery	23	4 Norm. vs 19 Inf	Whole blood	Illumina HumanWG-6 2.0
15	GSE40012	Inf	Discovery	75	36 Norm. vs 39 Inf	Whole blood	Illumina HumanHT-12 3.0
16	GSE29366	Inf	Validation	31	12 Norm. vs 19 Inf	Whole blood	Illumina HumanWG-6 3.0
17	GSE30550	Inf	Validation	33	16 Norm. vs 17 Inf	Peripheral blood	Affymetrix HG U133A 2.0
18	GSE20346	Inf	Validation	45	26 Bac. pneu. vs 19 Inf	Whole blood	Illumina HumanHT-12 3.0
19	GSE34205	Inf	Validation	50	22 Norm. vs 28 Inf	PBMC	Affymetrix HG U133 Plus 2.0
20	GSE82050	Inf	Validation	39	15 Norm. vs 24 Inf	Blood	Agilent SP G3 Human GE 3.0
21	GSE38900	Inf	Validation	46	30 Rhinovirus vs 16 Inf	Whole blood	Illumina HumanWG-6 3.0

Materials and Methods

The overall pipeline of the proposed framework is described in the main text. The algorithm used to perform *intra*- and *inter*-level analysis is described in the Figure S1. We utilize the *BLMA* package [3] from bioconductor to achieve this task.

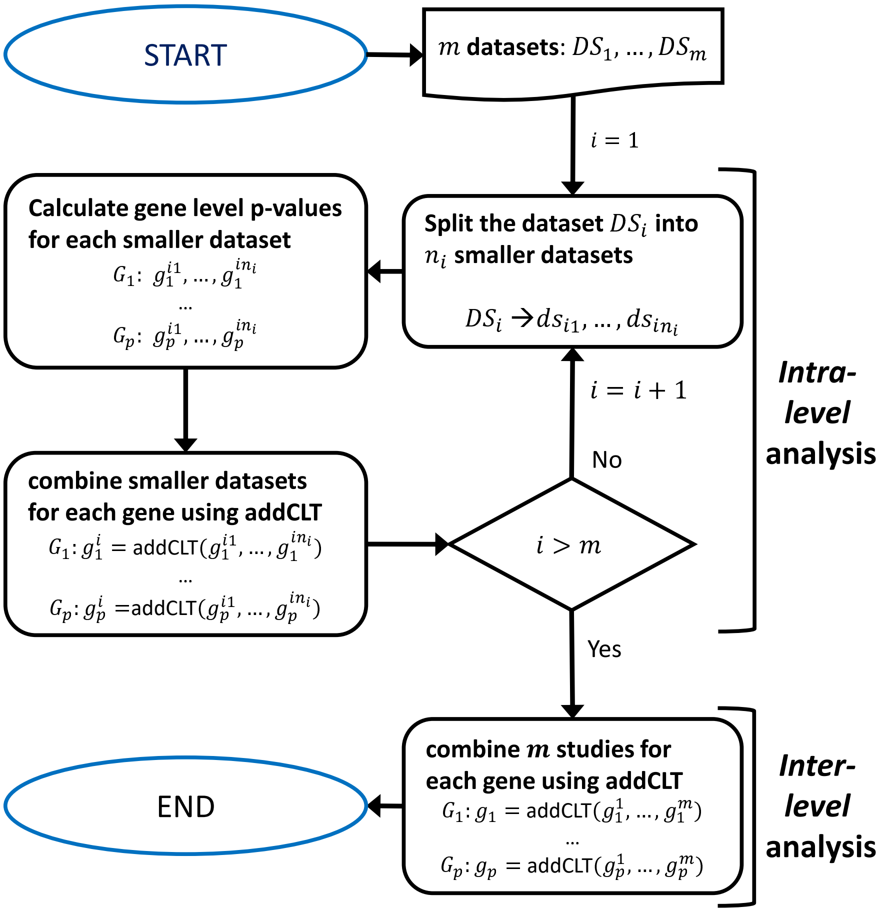
START
m datasets: $DS_1, \ldots, DS_m$
$i = 1$
Split the dataset DS_i into n_i smaller datasets
$DS_i \rightarrow ds_{i1}, \ldots, ds_{in_i}$
Calculate gene level p-values for each smaller dataset
G_1: $g_1^{i1}, \ldots, g_1^{in_i}$
...
G_p: $g_p^{i1}, \ldots, g_p^{in_i}$
$i = i + 1$
No
combine smaller datasets for each gene using addCLT
G_1: g_1^i = addCLT($g_1^{i1}, \ldots, g_1^{in_i}$)
...
G_p: g_p^i =addCLT($g_p^{i1}, \ldots, g_p^{in_i}$)
$i > m$
Intra-level analysis
Yes
combine m studies for each gene using addCLT
G_1: g_1 = addCLT($g_1^1, \ldots, g_1^m$)
...
G_p: g_p = addCLT($g_p^1, \ldots, g_p^m$)
END
Inter-level analysis

Figure S1: The gene level meta-analysis pipeline is performed in two stages: *intra-level* analysis and *inter-level* analysis. In the *intra-level* analysis, each dataset is divided into smaller datasets such that each smaller dataset consists of all the control samples and a subset of the disease samples. For each gene, p-values are calculated using moderated t-test and later combined using *addCLT*. In the *inter-level* analysis, *intra-level* p-values coming from individual datasets are combined using the same technique in order to compute meta p-value for each gene. Final output of the framework is a list of genes that are differentially expressed across the phenotypes of a given disease.

To compare with the p-value based approaches, we create four frameworks – each framework takes normalized gene expression datasets of a given disease as input, computes gene level p-values using moderated t-test for each dataset, and combines the individual studies for each gene using the chosen meta-analysis approach. Meta p-values are corrected using the FDR approach. *Leave-one-out* (LOO) analysis is performed for each of these four frameworks to select the genes that are not influenced by one single study. To select the significant genes, we use the same threshold that was used in the proposed framework. To compare with INMEX, we provide the gene level expression matrices as input in the web interface [4] and performed meta-analysis using both fixed effect model and random effect model, using the default settings. We rank the output genes with their absolute effect sizes (decreasing order). To compare with MetaIntegrator, we utilize the *MetaIntegrator* package from Bioconductor [5]. Similar to the above frameworks, we provide the gene level expression matrices and perform the LOO in the similar way as we did for the proposed framework. For the rank aggregation based method, we use the *RankAggreg* package from CRAN. This package requires the rank of the genes from multiple datasets as input and provides the combined ranks of the genes as output. At first, we compute gene level p-values for each dataset using moderated t-test and rank them based on their FDR corrected p-values. We choose the top significant genes from each dataset, using the same threshold used in the proposed framework. We use the default settings of the function provided in the package.

2 Results

We apply the proposed approach on 1108 samples from 9 independent training datasets related to two conditions: Alzheimer's disease (AD) and influenza. We evaluate the *global signature* using the target pathway enrichment approach and the *test signature* using an additional 912 samples from 12 independent validation datasets. Description of the findings are explained in the main text. The pathway enrichment results are computed using KEGG database [6] (version 84.0) that includes 204 signaling pathways. For both diseases, we compare the results of the proposed meta-analysis framework (GSMA) with the results of the eight other existing meta-analysis approaches. Among them, four approaches are p-value based (i.e., Fisher's method, Stouffer's method, minP, and maxP), three approaches are effect-size based (i.e., inmex fixed-effect model (inmex_FEM) [7], inmex random-effect model (inmex_REM) [7], and MetaIntergrator [8]), and one approach is rank aggregation based (i.e., RankAggreg [9]).

2.1 Alzheimer's Disease

We apply GSMA on 924 samples from 5 individual studies and identify 89 genes as the *global signature* and 7 genes as the *test signature*. We validate the *test signature* using an additional 668 samples from 6 individual validation studies. The two phenotypes for all the datasets are AD patients from different stages and healthy individuals.

Enriched pathways associated with the *global signature* identified by GSMA and the *global signature* identified by one given discovery dataset at a time are shown in the Table S2. The results show that GSMA is able to identify the target pathway at the very top. In contrast, the signatures obtained from any single analysis is not reproducible. In addition, they fail to identify the target pathway as significant in most of the cases. AUC plots of the 6 validation datasets based on the identified *test signatures* are illustrated in the Figure S2. The results indicate that the proposed approach outperforms any single analysis in 5 out of 6 cases by achieving higher AUC-ROC score.

The pathway enrichment results of the proposed framework and the existing approaches are shown in the Table S3. AUC scores of all 6 independent datasets are presented in the Table S4, whereas the AUC plots are presented in the Figure S3. Figure 3 in the main text explains the overall comparison between GSMA and the existing approaches.

2.2 Influenza

We apply the proposed framework on 184 samples from 4 individual studies and identify 153 genes as the *global signature* and 11 genes as the *test signature*. We validate the *test signature* using an additional 224 samples from 6 individual validation studies. In 8 out of 10 datasets, the two given phentypes are influenza patients of different stages and healthy patients. Among the remaining 2 datasets, GSE20346 compares 19 influenza patients and 26 bacterial pneumonia patients whereas GSE38900 compares 16 influenza patients and 30 Rhinovirus patients.

The pathway enrichment results of the proposed framework and the existing approaches are shown in the Table S5. AUC scores of all 6 independent datasets are presented in the Table S6, whereas the AUC plots are presented in the Figure S5. Figure 4 in the main text explains the overall comparison between GSMA and the existing approaches.

Table S2: The results of the enrichment analysis performed on the genes in the *global signatures* for **Alzheimer's disease** identified by the proposed meta-analysis framework (GSMA) and using one given discovery dataset at a time. The red line represents 0.5% threshold and the green highlighted cell represents the target pathway. GSMA is able to identify the target pathway - *Alzheimer's disease*, at the very top. On the other hand, using one single dataset at a time, the results are not reproducible and significantly influenced by the given individual dataset. In 3 out of 5 discovery datasets, the identified *global signatures* fail to identify the target pathway as significant.

	GSMA		Discovery_ds_1		Discovery_ds_2	
	Pathway	p.fdr	Pathway	p.fdr	Pathway	p.fdr
1	Alzheimer's disease	2.24E-07	Phagosome	0.0123	Pathogenic Escherichia coli infection	0.7973
2	Parkinson's disease	2.24E-07	Pathogenic Escherichia coli infection	0.0123	Cardiac muscle contraction	1
3	Non-alcoholic fatty liver disease (NAFLD)	3.63E-06	Gap junction	0.5592	Vasopressin-regulated water reabsorption	1
4	Huntington's disease	3.12E-05	Ferroptosis	0.7871	Cell cycle	1
5	Retrograde endocannabinoid signaling	0.0028	Platelet activation	0.8279	FoxO signaling pathway	1
6	Epithelial cell signaling in Helicobacter pylori infection	0.0435	Apelin signaling pathway	0.9143	Oxytocin signaling pathway	1
7	Cardiac muscle contraction	0.0621	cGMP-PKG signaling pathway	1	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	1
8	Chagas disease (American trypanosomiasis)	0.1430	Alcoholism	1	Protein processing in endoplasmic reticulum	1
9	Adipocytokine signaling pathway	0.2790	Focal adhesion	1	RNA degradation	1
10	Epstein-Barr virus infection	0.2790	mRNA surveillance pathway	1	Hypertrophic cardiomyopathy (HCM)	1

	Discovery_ds_3		Discovery_ds_4		Discovery_ds_5	
	Pathway	p.fdr	Pathway	p.fdr	Pathway	p.fdr
1	Adherens junction	0.0934	Parkinson's disease	8.41E-06	Parkinson's disease	4.35E-07
2	Proteoglycans in cancer	0.3461	Huntington's disease	8.41E-06	Alzheimer's disease	1.33E-06
3	Prolactin signaling pathway	0.3625	Alzheimer's disease	2.49E-05	Huntington's disease	2.85E-06
4	ErbB signaling pathway	0.4029	Non-alcoholic fatty liver disease (NAFLD)	9.63E-04	Non-alcoholic fatty liver disease (NAFLD)	6.20E-04
5	Th1 and Th2 cell differentiation	0.4029	Homologous recombination	0.0685	Retrograde endocannabinoid signaling	0.3385
6	Circadian entrainment	0.4029	Cardiac muscle contraction	0.3019	Homologous recombination	0.7155
7	AGE-RAGE signaling pathway in diabetic complications	0.4029	Retrograde endocannabinoid signaling	0.3019	Hedgehog signaling pathway	0.7775
8	Focal adhesion	0.4583	Epstein-Barr virus infection	0.7548	Vibrio cholerae infection	0.7775
9	Type II diabetes mellitus	0.4653	Vibrio cholerae infection	0.7760	Synaptic vesicle cycle	0.9895
10	Platelet activation	0.4653	Synaptic vesicle cycle	1	Phagosome	0.9895

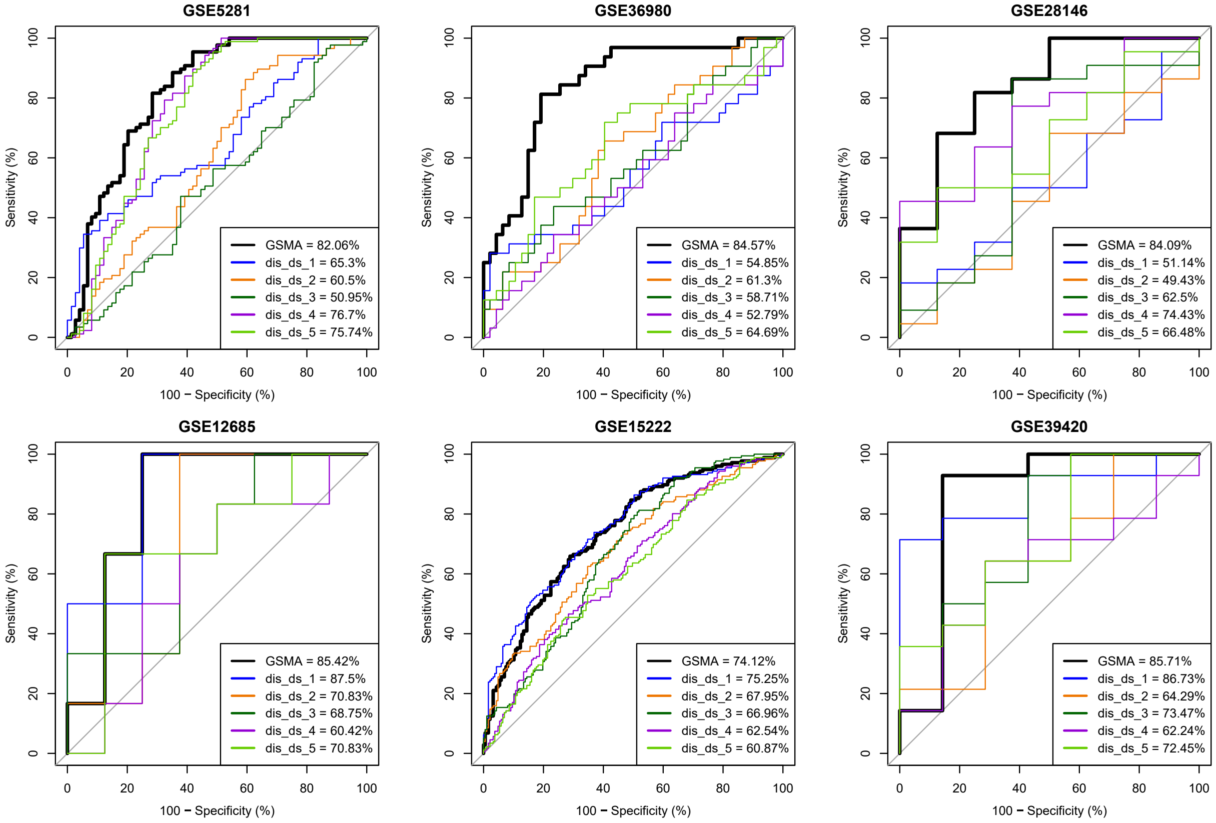

Figure S2: AUC plots across the 6 independent validation datasets related to **Alzheimer's disease**, based on the *test signature* identified by the proposed meta analysis framework - GSMA vs using one given discovery dataset at a time. The signature proposed by GSMA achieved higher AUC-ROC scores in 5 out of 6 independent datasets,

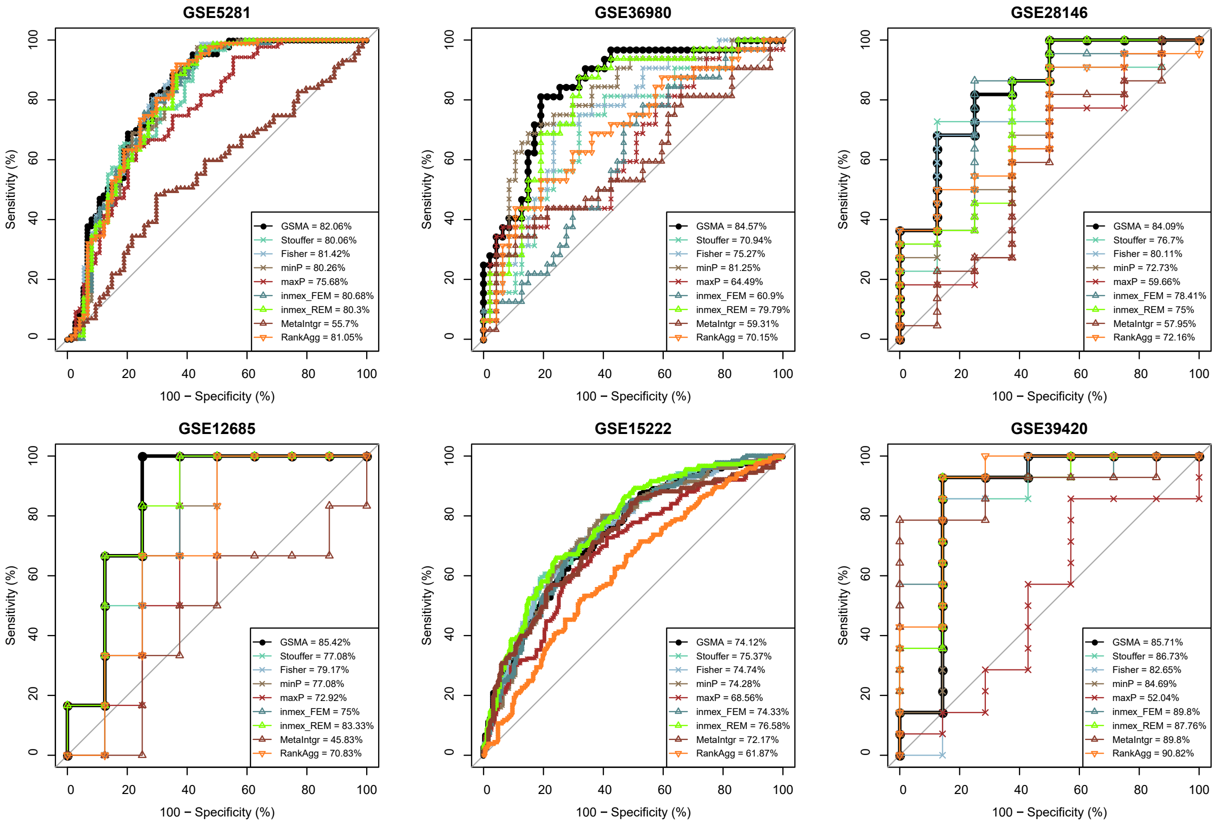

Figure S3: AUC plots across the 6 independent validation datasets related to **Alzheimer's disease** based on the *test signature*, identified by the proposed meta analysis framework - GSMA and the eight other existing meta-analysis approaches - Stouffer's method, Fisher's method, minP, maxP, inmex_FEM, inmex_REM, MetaIntegrator, and RankAggreg. The signature proposed by GSMA achieved highest AUC-ROC scores in 4 out of 6 independent datasets.

Table S3: The results of the enrichment analysis performed on the genes in the *global signatures* for **Alzheimer's disease** identified by the proposed meta-analysis framework (GSMA) and the eight other existing meta-analysis approaches (Stouffer's method, Fisher's method, minP, maxP, inmex_FEM, inmex_REM, MetaIntegrator, and RankAggreg). The red line represents 0.5% threshold and the green highlighted cell represents the target pathway. sing the *global signatures* identified by the GSMA, Stouffer's method, Fisher's method, and RankAggreg, the enrichment analysis finds all three neurological disorder pathways (*Alzheimer's disease, Parkinson's disease* and *Huntington's disease*) as significant. In contrast, the enrichment analysis performed with the *global signatures* identified by minP and maxP do not report any pathway as significant. The enrichment analysis performed on the signatures identified by MetaIntegrator reports two of the neurological disorder pathways as significant and rank them on top. The same analysis performed on the inmex_FEM and inmex_REM report some pathways as significantly enriched but none of them are the neurological disorder pathways. Interestingly, the other significant pathways reported by the enrichment analysis performed on the signatures identified by GSMA, *Non-alcoholic fatty liver disease (NAFLD)* and *Retrograde endocannabinoid signaling*, are also known to be involved in Alzheimer's disease [10, 11, 12, 13].

	GSMA		**Stouffer**		**Fisher**	
	Pathway	p.fdr	Pathway	p.fdr	Pathway	p.fdr
1	Alzheimer's disease	2.24E-07	Alzheimer's disease	3.78E-06	Alzheimer's disease	1.59E-04
2	Parkinson's disease	2.24E-07	Parkinson's disease	8.62E-05	Parkinson's disease	4.88E-04
3	Non-alcoholic fatty liver disease (NAFLD)	3.63E-06	Huntington's disease	5.80E-04	Huntington's disease	0.0019
4	Huntington's disease	3.12E-05	Non-alcoholic fatty liver disease (NAFLD)	0.0081	Non-alcoholic fatty liver disease (NAFLD)	0.0682
5	Retrograde endocannabinoid signaling	0.0028	Cardiac muscle contraction	0.0372	Homologous recombination	0.3336
6	Epithelial cell signaling in Helicobacter pylori infection	0.0435	Epithelial cell signaling in Helicobacter pylori infection	0.221	Vibrio cholerae infection	0.3687
7	Cardiac muscle contraction	0.0621	Retrograde endocannabinoid signaling	0.2708	Retrograde endocannabinoid signaling	0.3687
8	Chagas disease (American trypanosomiasis)	0.1430	Homologous recombination	0.5698	Synaptic vesicle cycle	0.4752
9	Adipocytokine signaling pathway	0.2790	Thyroid hormone signaling pathway	0.6203	Epithelial cell signaling in Helicobacter pylori infection	0.4877
10	Epstein-Barr virus infection	0.2790	Vibrio cholerae infection	0.655	Cardiac muscle contraction	0.5672

	minP		**maxP**		**inmex_FEM**	
	Pathway	p.fdr	Pathway	p.fdr	Pathway	p.fdr
1	Parkinson's disease	0.0147	Synaptic vesicle cycle	1	Epstein-Barr virus infection	5.76E-06
2	Alzheimer's disease	0.0152	Fluid shear stress & atherosclerosis	1	Pancreatic cancer	0.0014
3	Huntington's disease	0.0162	Herpes simplex infection	1	MAPK signaling pathway	0.0014
4	Vibrio cholerae infection	0.1690	Endocrine and other factor-regulated calcium reabsorption	1	Apoptosis	0.0021
5	Synaptic vesicle cycle	0.2060	Vibrio cholerae infection	1	Non-small cell lung cancer	0.0021
6	Epithelial cell signaling in Helicobacter pylori infection	0.2060	Epithelial cell signaling in Helicobacter pylori infection	1	Pathways in cancer	0.0024
7	Rheumatoid arthritis	0.3050	Adherens junction	1	Viral carcinogenesis	0.0024
8	Retrograde endocannabinoid signaling	0.5240	Bacterial invasion of epithelial cells	1	FoxO signaling pathway	0.0024
9	Non-alcoholic fatty liver disease (NAFLD)	0.5240	Gap junction	1	Osteoclast differentiation	0.0026
10	mTOR signaling pathway	0.5240	Rheumatoid arthritis	1	Bacterial invasion of epithelial cells	0.0037

	inmex_REM		**MetaIntegrator**		**RankAgg**	
	Pathway	p.fdr	Pathway	p.fdr	Pathway	p.fdr
1	Epstein-Barr virus infection	7.37E-06	Parkinson's disease	0.0024	Huntington's disease	1.03E-08
2	Leukocyte transendothelial migration	0.0055	Huntington's disease	0.0024	Parkinson's disease	4.79E-07
3	TNF signaling pathway	0.0082	Non-alcoholic fatty liver disease (NAFLD)	0.0139	Alzheimer's disease	5.38E-06
4	Kaposi's sarcoma-associated herpesvirus infection	0.0107	Alzheimer's disease	0.0242	Non-alcoholic fatty liver disease (NAFLD)	0.0032
5	HTLV-I infection	0.0112	Epithelial cell signaling in Helicobacter pylori infection	0.0853	Cardiac muscle contraction	0.0087
6	Hepatitis C	0.0128	NOD-like receptor signaling pathway	0.0853	NOD-like receptor signaling pathway	0.8569
7	Shigellosis	0.0128	Cardiac muscle contraction	0.1127	Prion diseases	1
8	Non-small cell lung cancer	0.0128	Rheumatoid arthritis	0.1642	Cell cycle	1
9	Fc gamma R-mediated phagocytosis	0.0169	Vibrio cholerae infection	0.1839	Central carbon metabolism in cancer	1
10	Pathways in cancer	0.0208	Necroptosis	0.2296	HIF-1 signaling pathway	1

Table S4: AUC-ROC scores on the 6 independent validation datasets related to **Alzheimer's disease**, based on the *test signatures* identified by different approaches. The results indicate that GSMA achieves the highest median AUC score among all other competitor approaches.

	Datasets	GSMA	Stouffer	Fisher	minP	maxP	inmex_FEM	inmex_REM	MetaIntgr	RankAgg
1	GSE5281	82.0596	80.0559	81.4228	80.2578	75.6757	80.6772	80.3044	55.7005	81.0500
2	GSE36980	84.5745	70.9441	75.2660	81.2500	64.4947	60.9043	79.7872	59.3085	70.1463
3	GSE28146	84.0909	76.7045	80.1136	72.7273	59.6591	78.4091	75.0000	57.9545	72.1591
4	GSE12685	85.4167	77.0833	79.1667	77.0833	72.9167	75.0000	83.3333	45.8333	70.8333
5	GSE15222	74.1249	75.3737	74.7387	74.2799	68.5586	74.3285	76.5800	72.1712	61.8711
6	GSE39420	85.7143	86.7347	82.6531	84.6939	52.0408	89.7959	87.7551	89.7959	90.8163
	Median	84.3327	76.8939	79.64015	78.67055	66.52665	76.70455	80.0458	58.6315	71.4962

Table S5: The results of the enrichment analysis performed on the genes in the *global signatures* for **influenza** identified by the proposed meta-analysis framework (GSMA) and the eight other existing meta-analysis approaches (Stouffer's method, Fisher's method, minP, maxP, inmex_FEM, inmex_REM, MetaIntegrator, and RankAggreg). The red line represents 0.5% threshold and the green highlighted cell represents the target pathway. The *global signatures* identified by the GSMA, Stouffer's method, inmex_FEM, and inmex_REM are significantly enriched in genes associated with the target pathway. The signatures produced by the other five existing methods are not enriched in genes associated with the target pathway to a significant level Interestingly, the other significant pathways reported by the enrichment analysis performed on the signatures identified by GSMA, such as *Herpes simplex infection*, *Staphylococcus aureus infection* and *Leishmaniasis*, are also known to have mechanisms similar to that of influenza[14, 15, 16, 17, 18].

	GSMA		**Stouffer**		**Fisher**	
	Pathway	p.fdr	Pathway	p.fdr	Pathway	p.fdr
1	Herpes simplex infection	5.01E-07	Staphylococcus aureus infection	5.92E-06	Staphylococcus aureus infection	4.28E-05
2	Influenza A	8.42E-06	Herpes simplex infection	7.19E-06	Herpes simplex infection	1.18E-04
3	Staphylococcus aureus infection	1.61E-05	Th1 and Th2 cell differentiation	1.56E-04	Leishmaniasis	0.0014
4	Leishmaniasis	0.0012	Leishmaniasis	2.09E-04	Systemic lupus erythematosus	0.0066
5	Systemic lupus erythematosus	0.0057	Th17 cell differentiation	3.31E-04	Influenza A	0.0066
6	Measles	0.0057	Asthma	0.0014	Tuberculosis	0.0071
7	Tuberculosis	0.0069	Systemic lupus erythematosus	0.0014	Toxoplasmosis	0.0091
8	Asthma	0.0115	Influenza A	0.0017	Phagosome	0.0091
9	Viral myocarditis	0.0142	Tuberculosis	0.0019	Asthma	0.0097
10	HTLV-I infection	0.0157	Toxoplasmosis	0.0019	Rheumatoid arthritis	0.0121

	minP		**maxP**		**inmex_FEM**	
	Pathway	p.fdr	Pathway	p.fdr	Pathway	p.fdr
1	Herpes simplex infection	0.0412	Cellular senescence	0.0901	Herpes simplex infection	2.85E-06
2	Systemic lupus erythematosus	0.0412	Complement and coagulation cascades	0.0901	Cell cycle	2.89E-06
3	Staphylococcus aureus infection	0.0437	Transcriptional misreg. in cancer	0.0978	Influenza A	1.02E-05
4	Viral myocarditis	0.0437	Staphylococcus aureus infection	0.0978	Measles	2.01E-05
5	Inflammatory bowel disease (IBD)	0.0437	Pertussis	0.1992	Viral carcinogenesis	5.09E-05
6	Antifolate resistance	0.0437	Influenza A	0.1992	Epstein-Barr virus infection	6.93E-05
7	Asthma	0.0437	Bladder cancer	0.1993	Cellular senescence	0.0002
8	Leishmaniasis	0.0537	Herpes simplex infection	0.1993	Th1 and Th2 cell differentiation	0.0008
9	Allograft rejection	0.0615	Measles	0.1993	Hepatitis B	0.0008
10	Graft-versus-host disease	0.0689	Estrogen signaling pathway	0.2862	p53 signaling pathway	0.0008

	inmex_REM		**MetaIntegrator**		**RankAgg**	
	Pathway	p.fdr	Pathway	p.fdr	Pathway	p.fdr
1	Herpes simplex infection	0.0001	Central carbon metabolism in cancer	0.5332	Systemic lupus erythematosus	0.0281
2	Influenza A	0.0028	Pertussis	0.5332	Asthma	0.0867
3	Viral carcinogenesis	0.0043	NOD-like receptor signaling pathway	0.5332	Non-alcoholic fatty liver disease (NAFLD)	0.0867
4	Measles	0.0043	Autophagy - animal	0.7784	Alzheimer's disease	0.0867
5	NOD-like receptor signaling pathway	0.0043	Apoptosis	0.7784	Influenza A	0.0867
6	Non-small cell lung cancer	0.0173	Legionellosis	0.7784	Parkinson's disease	0.0867
7	TNF signaling pathway	0.0200	Cytosolic DNA-sensing pathway	0.7784	Inflammatory bowel disease (IBD)	0.0867
8	Hepatitis B	0.0207	Renal cell carcinoma	0.7784	Graft-versus-host disease	0.1006
9	Kaposi's sarcoma-associated herpesvirus infection	0.0221	Prolactin signaling pathway	0.7784	Phagosome	0.1006
10	Epstein-Barr virus infection	0.0237	B cell receptor signaling pathway	0.7784	Staphylococcus aureus infection	0.1006

Table S6: AUC-ROC scores on the 6 independent validation datasets related to **influenza**, based on the *test signatures* identified by different approaches. The results indicate that GSMA achieves the highest median AUC score among all other competitor approaches.

	Datasets	GSMA	Stouffer	Fisher	minP	maxP	inmex_FEM	inmex_REM	MetaIntgr	RankAgg
1	GSE29366	97.3684	93.4211	89.0351	90.7895	93.4211	93.4211	93.4211	89.9123	92.9825
2	GSE34205	91.8831	81.1688	79.3831	86.0390	90.5844	86.8506	86.2013	81.1688	92.6948
3	GSE30550	91.1765	84.9265	80.1471	79.7794	77.2059	87.8676	88.6029	87.8676	71.6912
4	GSE38900	72.0833	54.7917	42.9167	55.2083	62.7083	57.7083	55.8333	72.5000	64.1667
5	GSE20346	83.8057	55.2632	57.8947	58.7045	56.2753	59.9190	60.1215	68.2186	96.9636
6	GSE82050	84.7222	81.9444	73.0556	85.5556	82.5000	95.5556	72.7778	88.8889	65.5556
	Median	87.9493	81.5566	76.2193	82.6675	79.8529	87.3591	79.4895	84.5182	82.1930

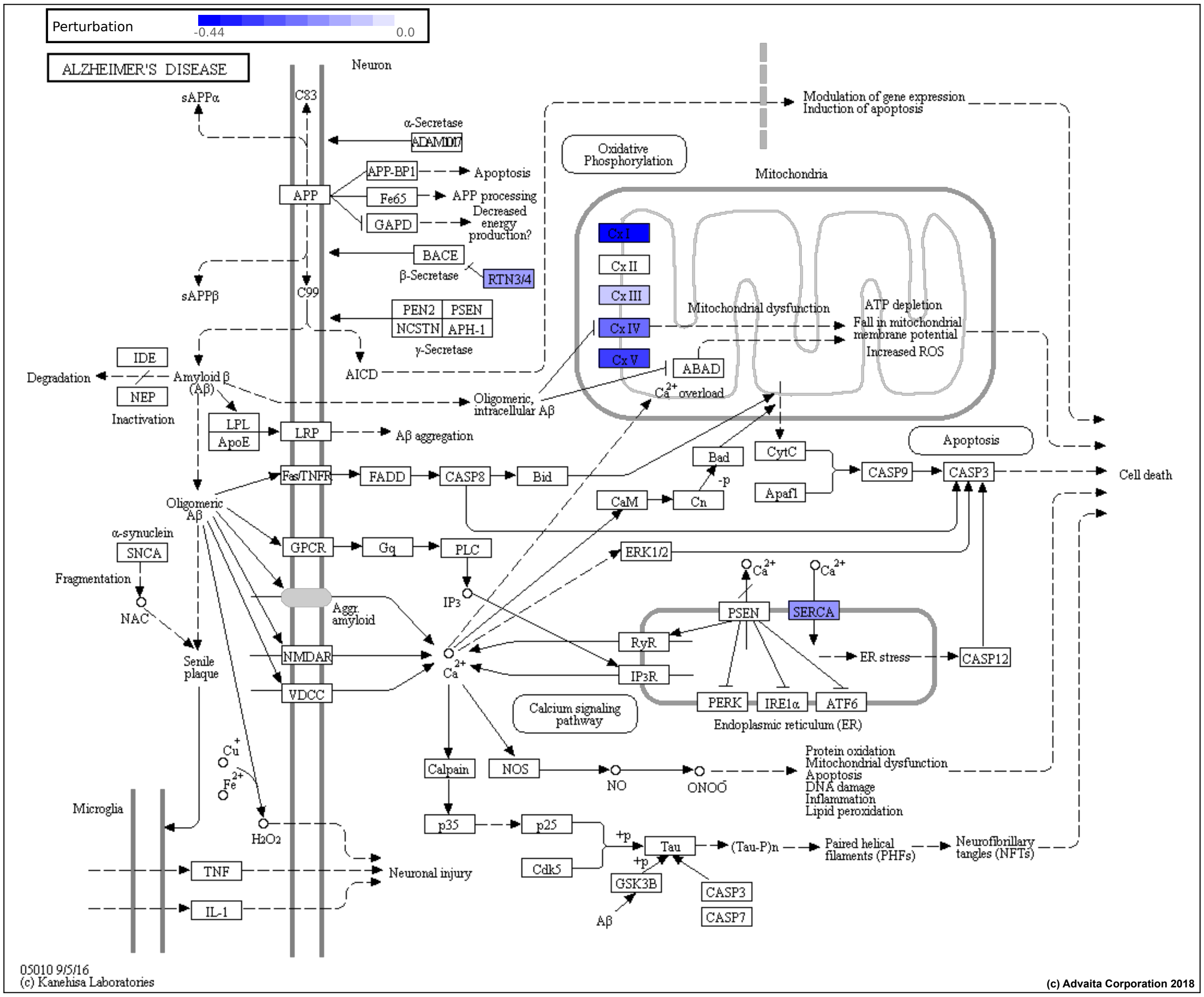

Figure S4: *Alzheimer's disease* pathway generated with iPathwayGuide [19, 20] using the *global signature* identified by the proposed framework - GSMA. Here the blue colors represent the negatively perturbed genes. The majority of the *global signature* genes present in the *Alzheimer's disease* pathway are part of the mitochondrial dysfunction process, which is one of the key factors for Alzheimer's disease progression [21, 22].

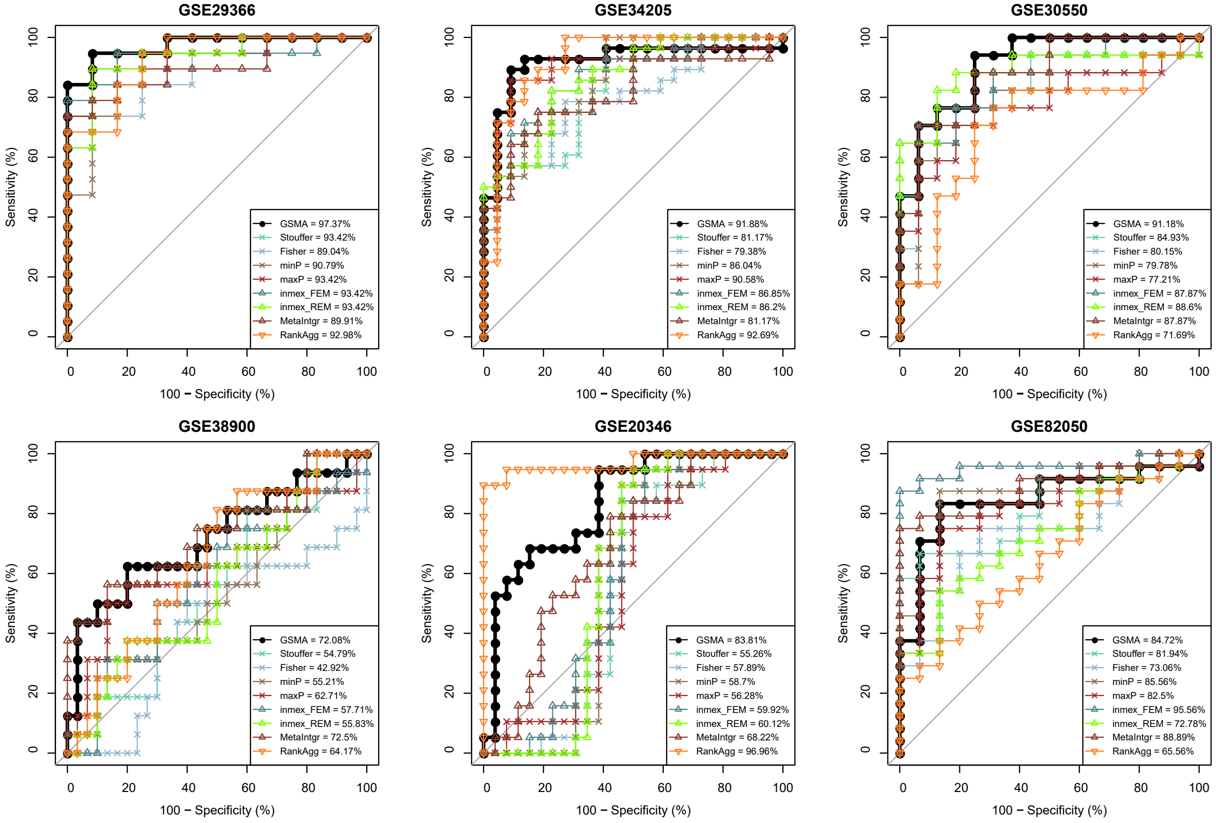

Figure S5: AUC plots across the 6 independent validation datasets related to **influenza** based on the *test signature*, identified by the proposed meta analysis framework - GSMA and eight other existing meta-analysis approaches - Stouffer's method, Fisher's method, minP, maxP, inmex_FEM, inmex_REM, MetaIntegrator, and RankAggreg. The signature proposed by GSMA and RankAggreg achieved highest AUC-ROC scores in 2 out of 6 independent datasets. In addition, Table S6 indicates that GSMA achieves the highest median AUC score among all other competitor approaches.

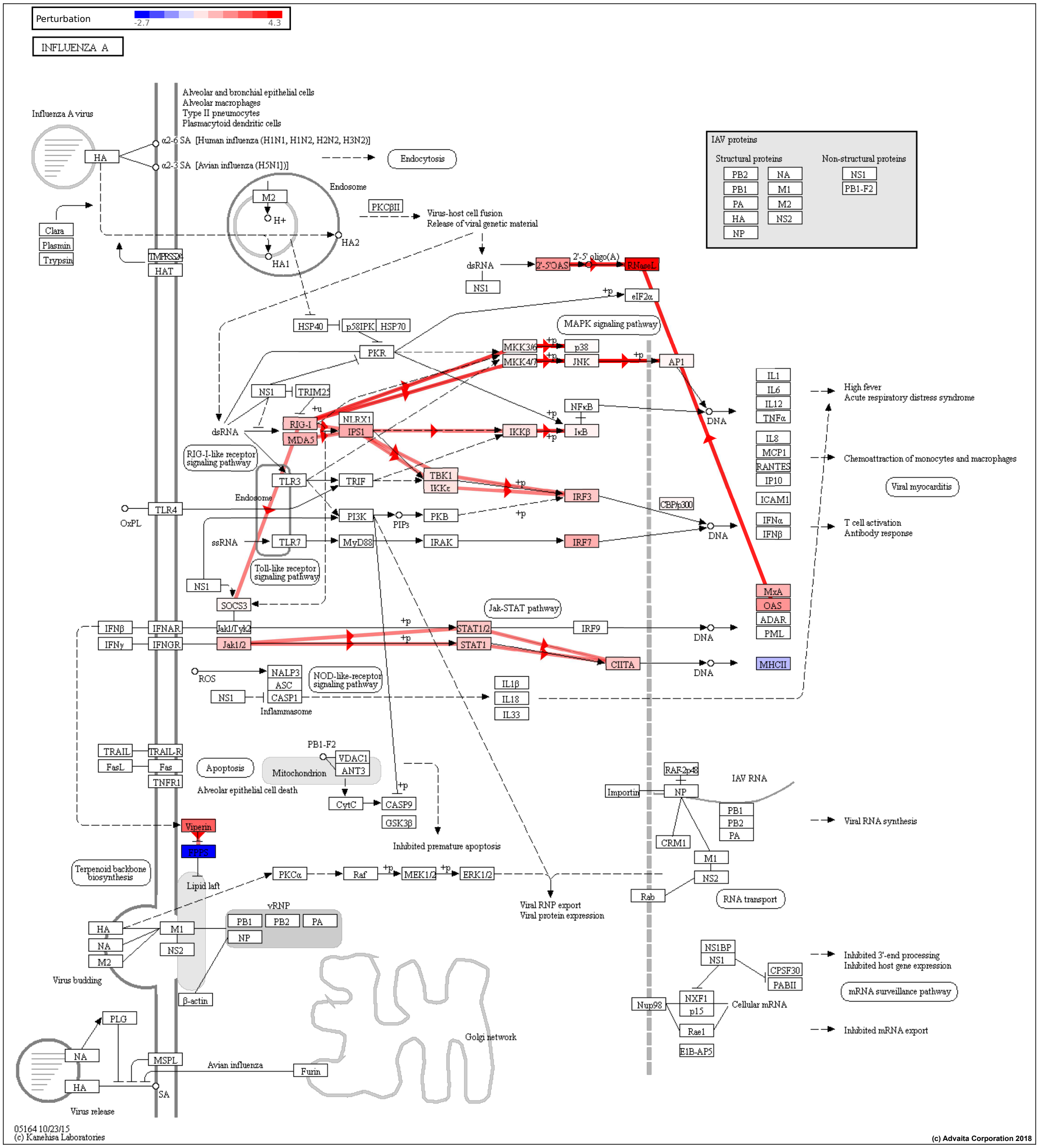

Figure S6: *Influenza A* pathway generated with iPathwayGuide [19, 20] using the *global signature* identified by the proposed framework - GSMA. Here the red colors represent the positively perturbed genes whereas the blue colors represent the negatively perturbed genes. The majority of the positively perturbed genes are part coherent chains of perturbation propagation which can be thought of as putative mechanisms.

Table S7: The top 20 pathways that are enriched with the *global signature* identified by GSMA, and the top 20 pathways that are identified by BLMA, using **Alzheimer's disease** datasets. The red line represents 0.5% threshold and the green highlighted cell represents the target pathway. The target pathway is significantly enriched and ranked at the top using the *global signature* identified by GSMA. On the other hand, BLMA is not able to identify the target pathway within the top 20 pathways. This shows that the *global signature* identified by GSMA is more powerful than BLMA, in terms of identifying the putative mechanism of a disease.

	GSMA		BLMA	
	Pathway	p.fdr	Pathway	p.fdr
1	Alzheimer's disease	2.24E-07	Retrograde endocannabinoid signaling	1.41E-06
2	Parkinson's disease	2.24E-07	Phagosome	9.18E-06
3	Non-alcoholic fatty liver disease (NAFLD)	3.63E-06	Synaptic vesicle cycle	0.0007
4	Huntington's disease	3.12E-05	Fc gamma R-mediated phagocytosis	0.0010
5	Retrograde endocannabinoid signaling	0.0028	HIF-1 signaling pathway	0.0010
6	Epithelial cell signaling in Helicobacter pylori infection	0.0435	Th1 and Th2 cell differentiation	0.0050
7	Cardiac muscle contraction	0.0621	Tuberculosis	0.1508
8	Chagas disease (American trypanosomiasis)	0.1433	Estrogen signaling pathway	0.1508
9	Adipocytokine signaling pathway	0.2786	Leukocyte transendothelial migration	0.1508
10	Epstein-Barr virus infection	0.2786	Leishmaniasis	0.1508
11	Rheumatoid arthritis	0.4724	Osteoclast differentiation	0.1508
12	Necroptosis	0.4744	Kaposi's sarcoma-associated herpesvirus infection	0.1508
13	NOD-like receptor signaling pathway	0.4744	Rheumatoid arthritis	0.1508
14	HIF-1 signaling pathway	0.4744	B cell receptor signaling pathway	0.1618
15	Th17 cell differentiation	0.4905	Vibrio cholerae infection	0.1855
16	TNF signaling pathway	0.4905	Amyotrophic lateral sclerosis (ALS)	0.1855
17	Herpes simplex infection	0.4905	Th17 cell differentiation	0.1910
18	Toxoplasmosis	0.4905	Cardiac muscle contraction	0.1930
19	Vibrio cholerae infection	0.4905	Salmonella infection	0.1930
20	cAMP signaling pathway	0.4905	Epithelial cell signaling in Helicobacter pylori infection	0.1989

Table S8: The top 20 pathways that are enriched with the *global signature* identified by GSMA, and the top 20 pathways that are identified by BLMA, using **influenza** datasets. The red line represents 0.5% threshold and the green highlighted cell represents the target pathway. The target pathway is ranked within the top two significant pathways in both cases. The list of significantly impacted pathways identified by BLMA include several false positive pathways. In contrast, the list of pathways that are significantly enriched with the *global signature* identified by GSMA is much more precise.

	GSMA		BLMA	
	Pathway	p.fdr	Pathway	p.fdr
1	Herpes simplex infection	5.01E-07	Staphylococcus aureus infection	2.60E-14
2	Influenza A	8.42E-06	Influenza A	4.65E-13
3	Staphylococcus aureus infection	1.61E-05	Intestinal immune network for IgA production	1.88E-11
4	Leishmaniasis	0.0012	Phagosome	9.81E-06
5	Systemic lupus erythematosus	0.0057	Systemic lupus erythematosus	1.27E-05
6	Measles	0.0057	Leishmaniasis	6.04E-05
7	Tuberculosis	0.0069	Transcriptional misregulation in cancer	0.0001
8	Asthma	0.0115	Acute myeloid leukemia	0.0001
9	Viral myocarditis	0.0142	Measles	0.0002
10	HTLV-I infection	0.0157	NOD-like receptor signaling pathway	0.0003
11	Allograft rejection	0.0185	Graft-versus-host disease	0.0004
12	Hepatitis B	0.0217	Herpes simplex infection	0.0010
13	Transcriptional misregulation in cancer	0.0217	Epstein-Barr virus infection	0.0012
14	Cell adhesion molecules (CAMs)	0.0217	Viral carcinogenesis	0.0012
15	Th17 cell differentiation	0.0217	Viral myocarditis	0.0013
16	Pertussis	0.0244	Allograft rejection	0.0014
17	Phagosome	0.0244	Asthma	0.0028
18	Complement and coagulation cascades	0.0265	Type I diabetes mellitus	0.0031
19	Intestinal immune network for IgA production	0.0278	Antigen processing and presentation	0.0058
20	Cellular senescence	0.0278	Autoimmune thyroid disease	0.0081

References

[1] Benjamin Milo Bolstad. *Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization.* PhD thesis, University of California, 2004.

[2] Gordon K Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

[3] Tin Nguyen and Sorin Draghici. *BLMA: A package for bi-level meta-analysis.* Bioconductor, 2017. R package.

[4] Jianguo Xia, Erin E Gill, and Robert EW Hancock. NetworkAnalyst - a web-based platform for gene expression profiling & biological network analysis. https://www.networkanalyst.ca/NetworkAnalyst/faces/home.xhtml, 2015.

[5] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

[6] Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[7] Jianguo Xia, Christopher D Fjell, Matthew L Mayer, Olga M Pena, David S Wishart, and Robert EW Hancock. INMEXa web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Research*, 41(W1):W63–W70, 2013.

[8] Winston A Haynes, Francesco Vallania, Charles Liu, Erika Bongen, Aurelie Tomczak, Marta Andres-Terrè, Shane Lofgren, Andrew Tam, Cole A Deisseroth, Matthew D Li, et al. Empowering multi-cohort gene expression analysis to increase reproducibility. In *Pacific Symposium on Biocomputing*, pages 144–153, New Jersey, 2017. World Scientific.

[9] Vasyl Pihur, Susmita Datta, and Somnath Datta. RankAggreg, an R package for weighted rank aggregation. *BMC bioinformatics*, 10(1):62, 2009.

[10] Do-Geun Kim, Antje Krenz, Leon E Toussaint, Kirk J Maurer, Sudie-Ann Robinson, Angela Yan, Luisa Torres, and Margaret S Bynoe. Non-alcoholic fatty liver disease induces signs of alzheimer's disease (ad) in wild-type mice and accelerates pathological signs of ad in an ad model. *Journal of Neuroinflammation*, 13(1):1, 2016.

[11] Gaurav Bedse, Adele Romano, Angelo M Lavecchia, Tommaso Cassano, and Silvana Gaetani. The role of endocannabinoid signaling in the molecular mechanisms of neurodegeneration in alzheimer's disease. *Journal of Alzheimer's Disease*, 43(4):1115–1136, 2015.

[12] Jan Mulder, Misha Zilberter, Susana J Pasquaré, Alán Alpár, Gunnar Schulte, Samira G Ferreira, Attila Köfalvi, Ana M Martín-Moreno, Erik Keimpema, Heikki Tanila, et al. Molecular reorganization of endocannabinoid signalling in Alzheimer's disease. *Brain*, 134(4):1041–1060, 2011.

[13] Sang Won Seo, Rebecca F Gottesman, Jeanne M Clark, Ruben Hernaez, Yoosoo Chang, Changsoo Kim, Kyoung Hwa Ha, Eliseo Guallar, and Mariana Lazo. Nonalcoholic fatty liver disease is associated with cognitive function in adults. *Neurology*, 86(12):1136–1142, 2016.

[14] Lynn M Hassman and David A DiLoreto. Immunologic factors may play a role in herpes simplex virus 1 reactivation in the brain and retina after influenza vaccination. *IDCases*, 6:47–51, 2016.

[15] Agnieszka Rynda-Apple, Keven M Robinson, and John F Alcorn. Influenza and bacterial superinfection: illuminating the immunologic mechanisms of disease. *Infection and Immunity*, 83(10):3764–3770, 2015.

[16] Mei-Ho Lee, Carlos Arrecubieta, Francis J Martin, Alice Prince, Alain C Borczuk, and Franklin D Lowy. A postinfluenza model of staphylococcus aureus pneumonia. *The Journal of Infectious Diseases*, 201(4):508–515, 2010.

[17] Keven M Robinson, Kevin J McHugh, Sivanarayana Mandalapu, Michelle E Clay, Benjamin Lee, Erich V Scheller, Richard I Enelow, Yvonne R Chan, Jay K Kolls, and John F Alcorn. Influenza a virus exacerbates staphylococcus aureus pneumonia in mice by attenuating antimicrobial peptide production. *The Journal of Infectious Diseases*, 209(6):865–875, 2013.

[18] Katherine Kedzierska, Joan M Curtis, Sophie A Valkenburg, Lauren A Hatton, Hiu Kiu, Peter C Doherty, and Lukasz Kedzierski. Induction of protective cd4+ t cell-mediated immunity by a leishmania peptide delivered in recombinant influenza viruses. *PLoS One*, 7(3):e33161, 2012.

[19] Advaita Corporation. Pathway Analysis with iPathwayGuide. http://www.advaitabio.com/ipathwayguide.html, 2014.

[20] S Ahsan and S Drăghici. Identifying significantly impacted pathways and putative mechanisms with ipathwayguide. *Current Protocols in Bioinformatics*, 57:7–15, 2017.

[21] Xinglong Wang, Wenzhang Wang, Li Li, George Perry, Hyoung-gon Lee, and Xiongwei Zhu. Oxidative stress and mitochondrial dysfunction in alzheimer's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(8):1240–1247, 2014.

[22] Michael H Yan, Xinglong Wang, and Xiongwei Zhu. Mitochondrial defects and oxidative stress in alzheimer's disease and parkinson disease. *Free Radical Biology and Medicine*, 62:90–101, 2013.