New Addhealth Gene Expression Dataset

Brandt Levitt Ph.D.

Carolina Population Center

University of North Carolina at Chapel Hill

Motivations

- Why do we care – linking social exposures to biological outcomes, personalized interventions, policy implications, deeper understanding of human experience
- What datasets will be discussed –three batches of gene expression
- How can you use these – search for linked gene expression variation, biomarkers, determinant of disease, linking social exposures to health outcomes, immune cell distributions, biological aging clocks, hormone production, prognostic indicator of disease

Outline of Presentation

- Gene expression background
- Data – description, QC, format
- Composition of sample
- Paired samples
- Example of use
- Access
- Limitations

Birth of Social Genomics – Gene Expression

Research

Open Access

Social regulation of gene expression in human leukocytes

Steve W Cole*†‡, Louise C Hawkley§, Jesusa M Arevalo*, Caroline Y Sung†, Robert M Rose¶ and John T Cacioppo§

Addresses: *Department of Medicine, Division of Hematology-Oncology, UCLA School of Medicine, Los Angeles CA 90095-1678, USA. †UCLA AIDS Institute, UCLA Molecular Biology Institute, Jonsson Comprehensive Cancer Center. ‡Norman Cousins Center. §Department of Psychology, and Center for Cognitive and Social Neuroscience, University of Chicago. ¶Institute for Medical Humanities, University of Texas Medical Branch at Galveston, and the John D and Catherine T MacArthur Foundation.

Correspondence: Steve W Cole. Email: coles@ucla.edu

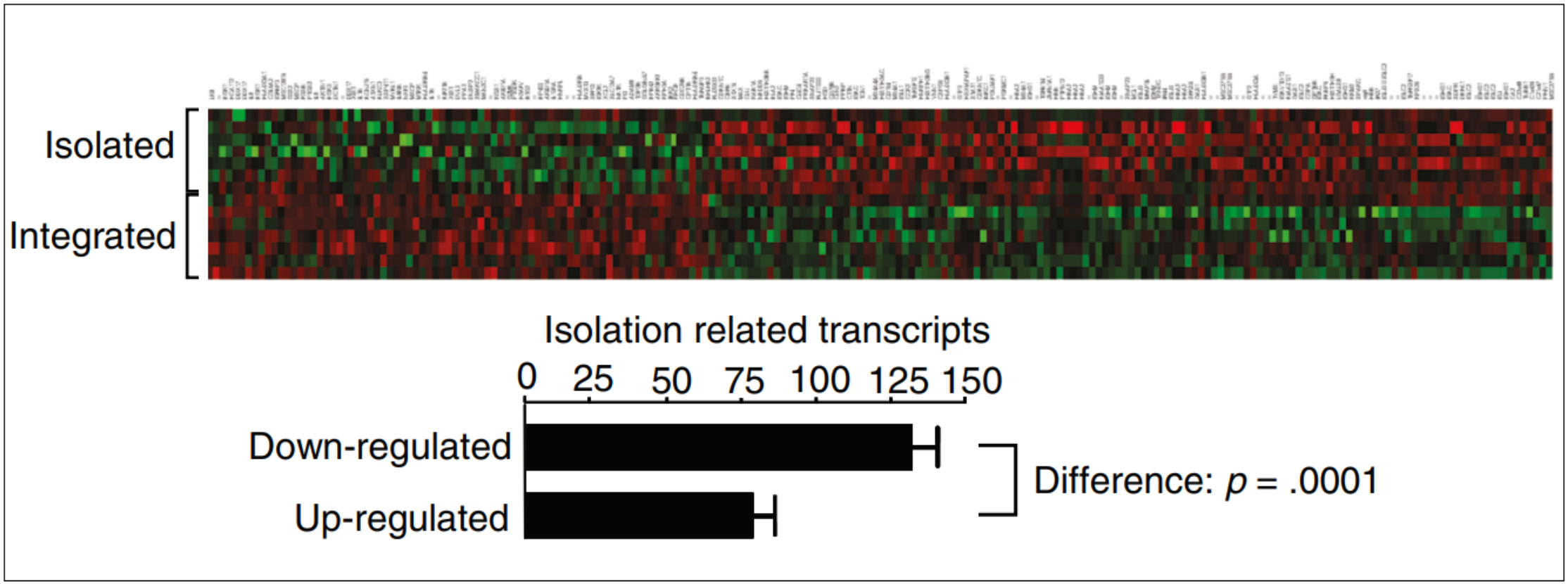
Published: 13 September 2007

Genome ***Biology*** 2007, **8:**R189 (doi:10.1186/gb-2007-8-9-r189)

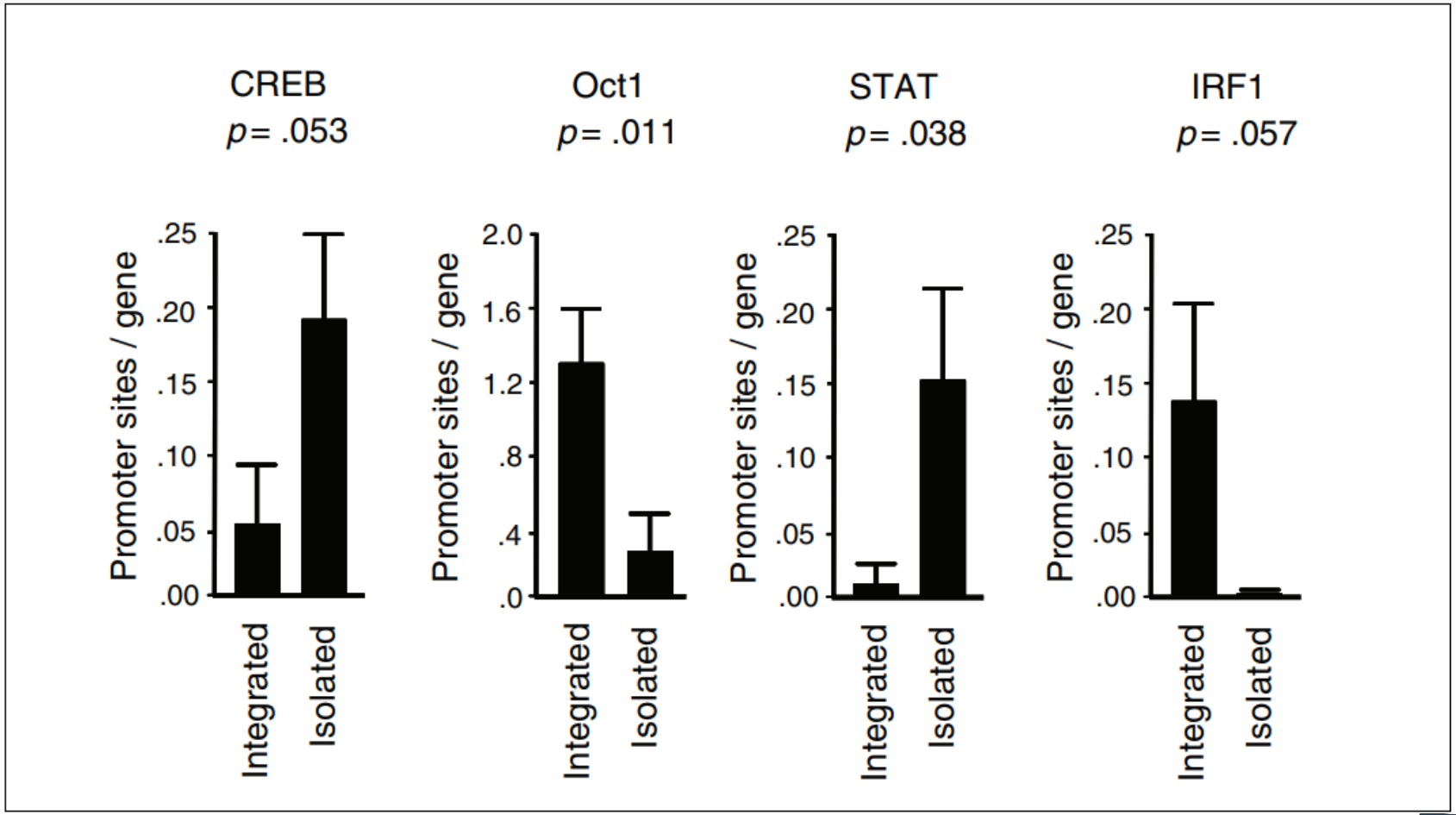
The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/9/R189

Received: 2 March 2007
Revised: 30 July 2007
Accepted: 13 September 2007

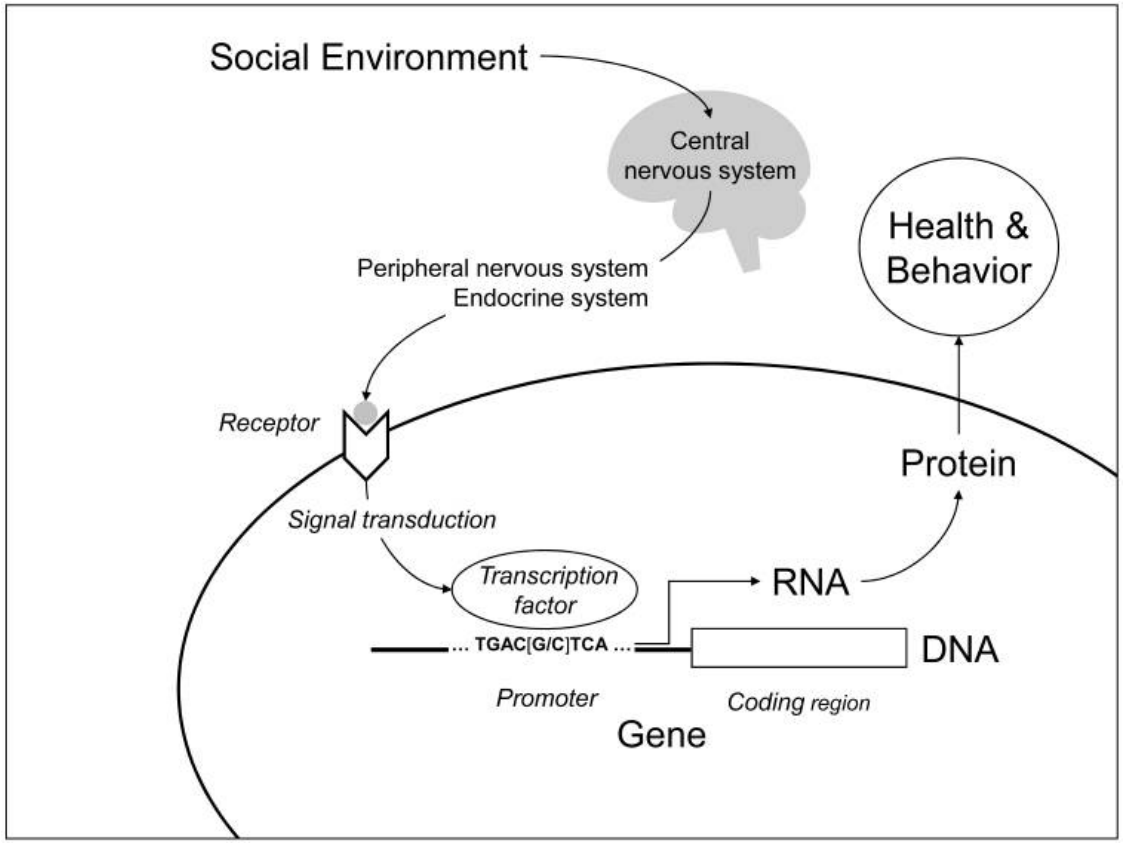
Birth of Social Genomics – Gene Expression



Birth of Social Genomics – Gene Expression



Background



Cole SW. Social regulation of human gene expression. Curr Dir Psychol Sci. 2009 Jun 1;18(3):132-137.

Sample Collection

- 2.5 ml of whole blood from venipuncture
- Deposited in PAXgene RNA tubes
- Lyses cells and stabilizes RNA
- Samples sent within 24 hours to UCLA social genomics core
- 1-8 micrograms of RNA extracted
- Converted to cDNA for data collection

Sample Collection Questionnaire

H5Q013A	Q013a Gum disease/tooth loss in last 4 weeks
H5Q013B	Q013b Active infection in last 4 weeks
H5Q013C	Q013c Injury in last 4 weeks
H5Q013D	Q013d Acute illness in last 4 weeks
H5Q013E	Q013e Surgery in last 4 weeks
H5Q013F	Q013f Active seasonal allergies in last 4 weeks
H5INFECT	Q013 Count of infectious/inflammatory diseases
H5Q014A	Q014a Cold or Flu-like symptoms in last 2 weeks
H5Q014B	Q014b Fever in last 2 weeks
H5Q014C	Q014c Night sweats in last 2 weeks
H5Q014D	Q014d Nausea or vomiting or diarrhea in last 2 weeks
H5Q014E	Q014e Blood in stool or feces or urine in last 2 weeks
H5Q014F	Q014f Frequent urination in last 2 weeks
H5Q014G	Q014g Skin rash or abscess in last 2 weeks
H5SUBCLN	Q014 Count of subclinical symptoms

Sample Composition

		year1	year2	year3	total
age					
	<=34	98	14	47	159
	35	150	105	153	408
	36	182	175	274	631
	37	216	184	321	721
	38	209	291	372	872
	39	189	230	396	815
	40	70	190	336	596
	41	10	89	160	259
	>=42	2	19	48	69
sex	female	680	796	1250	2726
	male	452	505	860	1817
race	white	762	860	1304	2926
	black	198	253	417	868
	other	7	11	29	47
	asian	51	50	136	237
	hispanic	114	127	224	465
twin	no	1097	1258	2059	4414
	yes	35	43	51	129
					4543

Batches

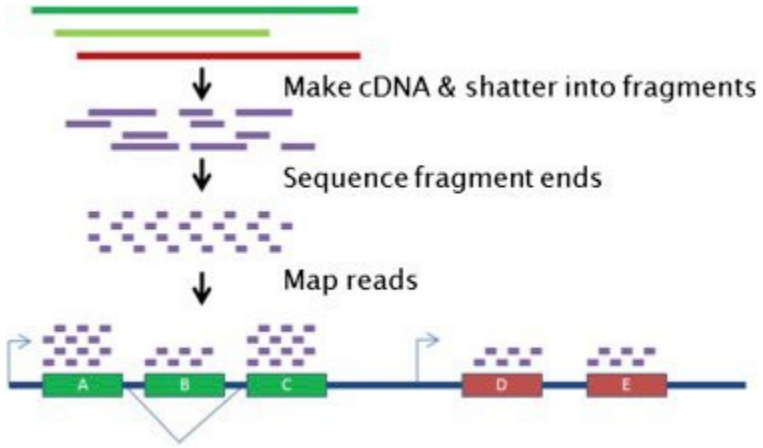
3 Batches

- Year 1 - n=1126
- Year 2 – n=1468
- Year 3 – n=2110

Includes intraindividual variation (IIV) samples n=83

N=4543 unique samples

Data Collection



DEFINITIONS

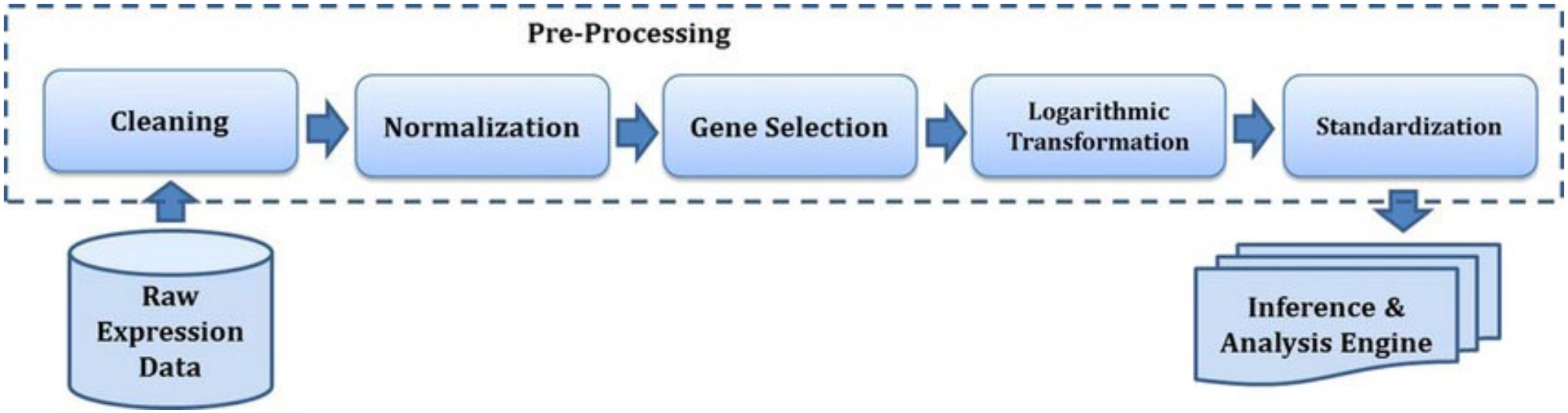
(n x m) Data Matrix

m Samples

n Samples

Gene	Sample 1	Sample 1	.....	Sample m
a				
b				
c				
...				
n				

Data Preprocessing and Normalization

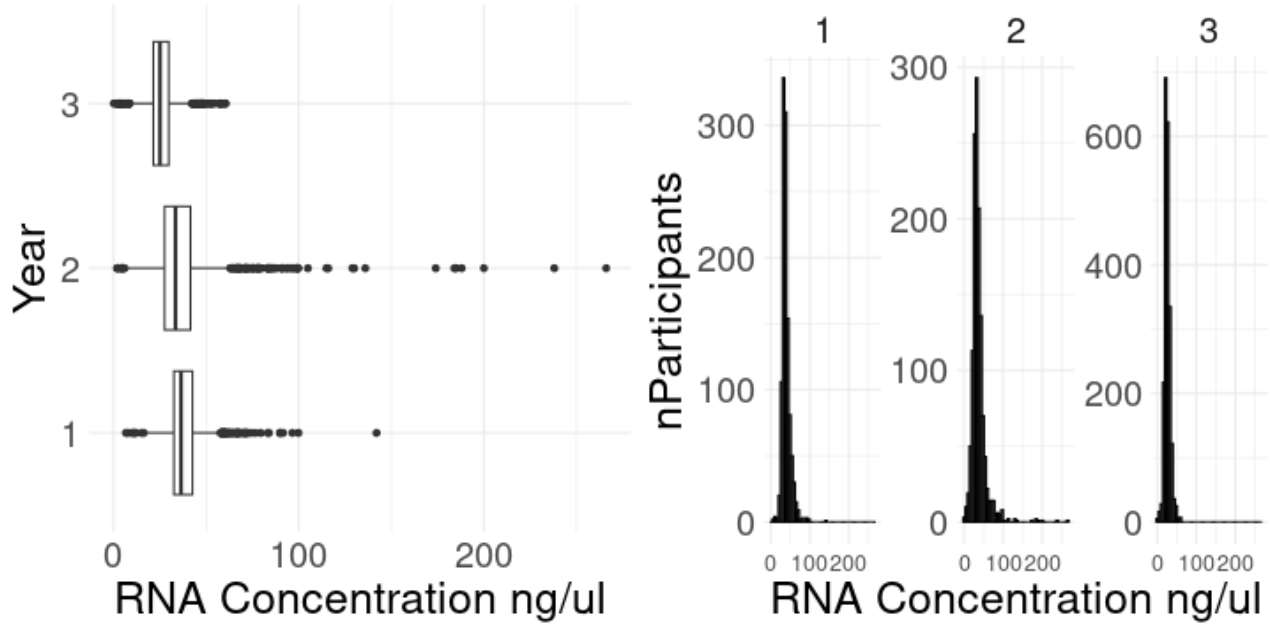


Roy, Swarup & Sharma, Pooja & Nath, Keshab & Bhattacharyya, Dhruba K & Kalita, Jugal. (2018). Pre-Processing: A Data Preparation Step. 10.1016/B978-0-12-809633-8.20457-3.

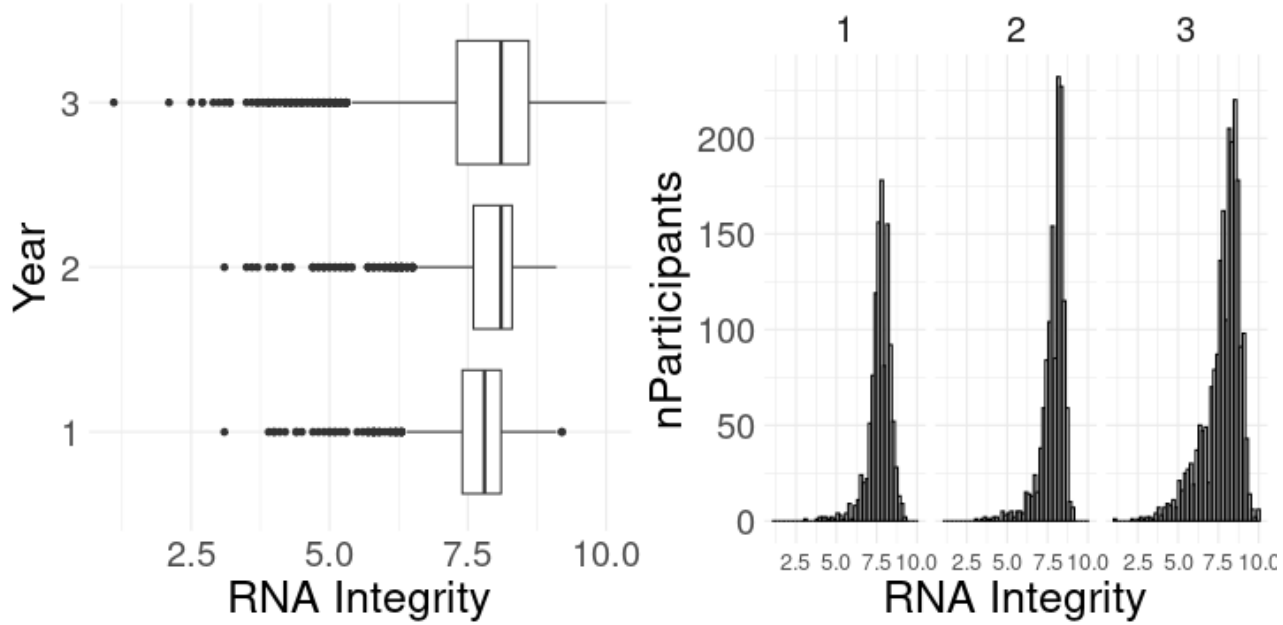
Sample Flags

- Low input sample quality (1.1-10 RIN, median 8 RIN, >3 filter, 7 flagged)
- Low input sample quantity (0.2 -166 ng/ul, median 30 ng/ul, no filter)
- Low reads per sample (135,233-80,059,632 reads, median 12,907,214 reads, no filter)
- Low mapping of reads to genome (8.9%-99.4% mapped, median 97.8%, <85% mapped, 69 flagged)
- High sample dissimiliarity (0.08-0.97 AvgCorrelogram, median 0.90, >0.85, 184 flagged)

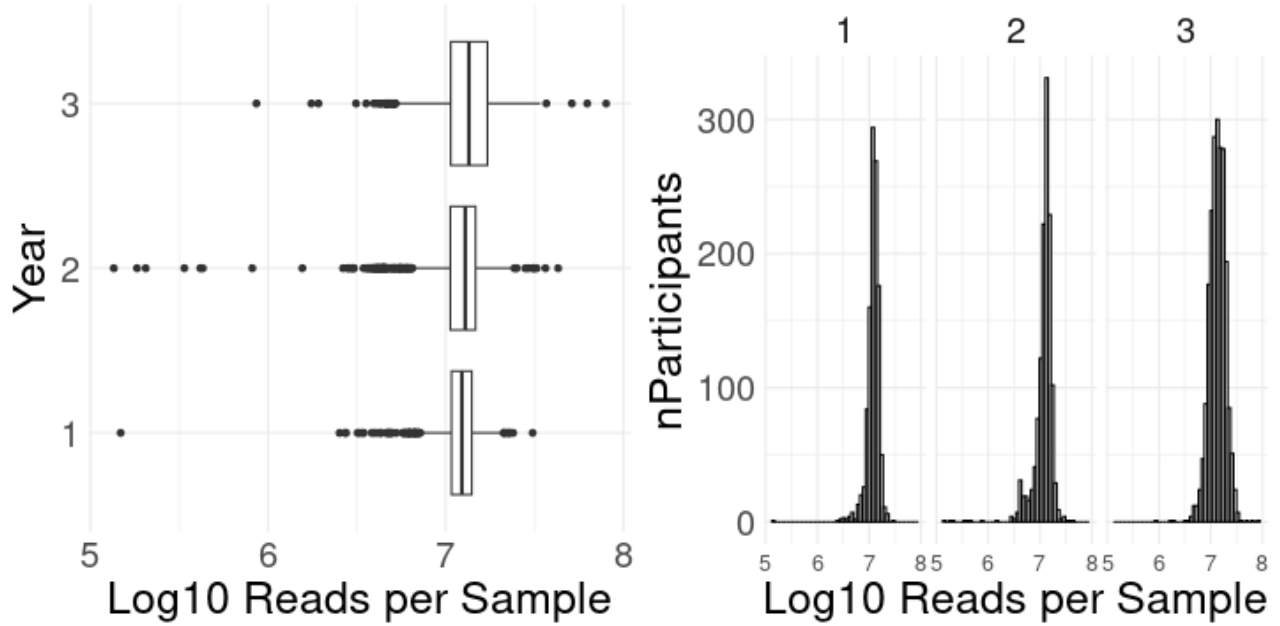
Input Sample Quantity



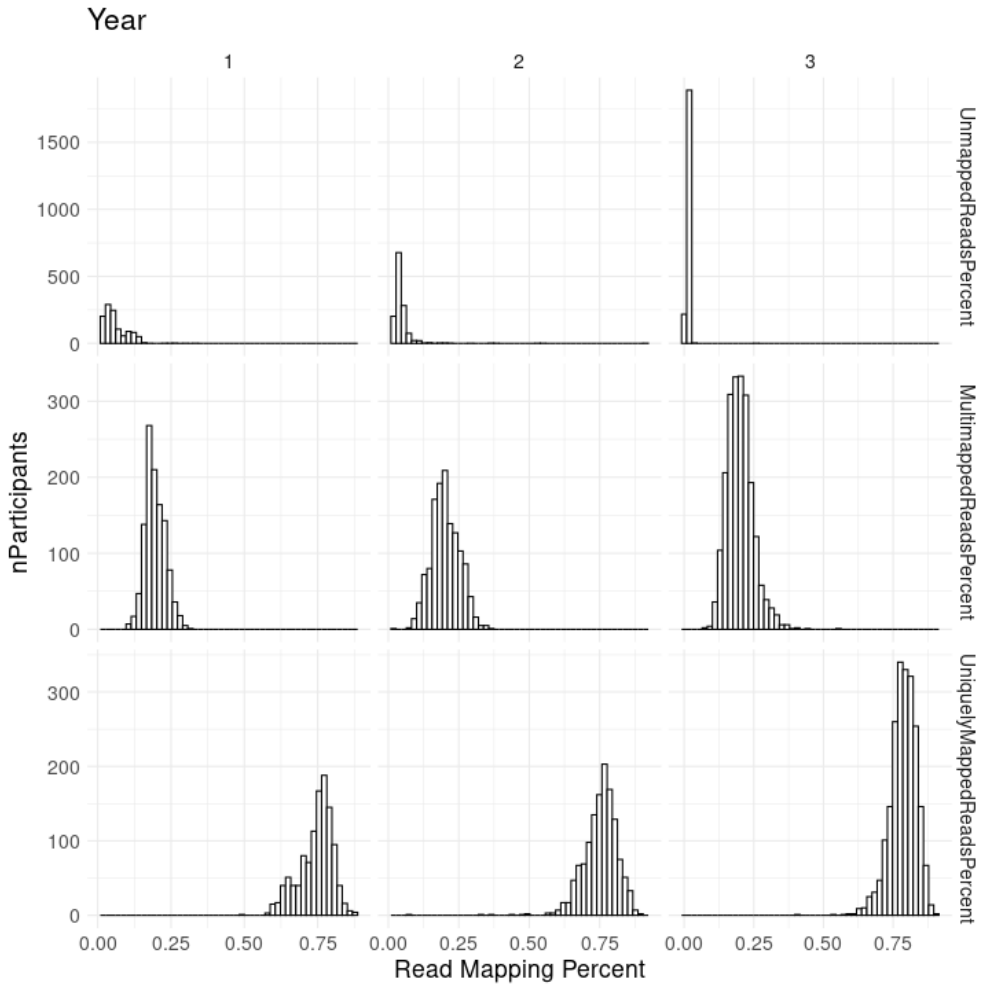
Input Sample Quality



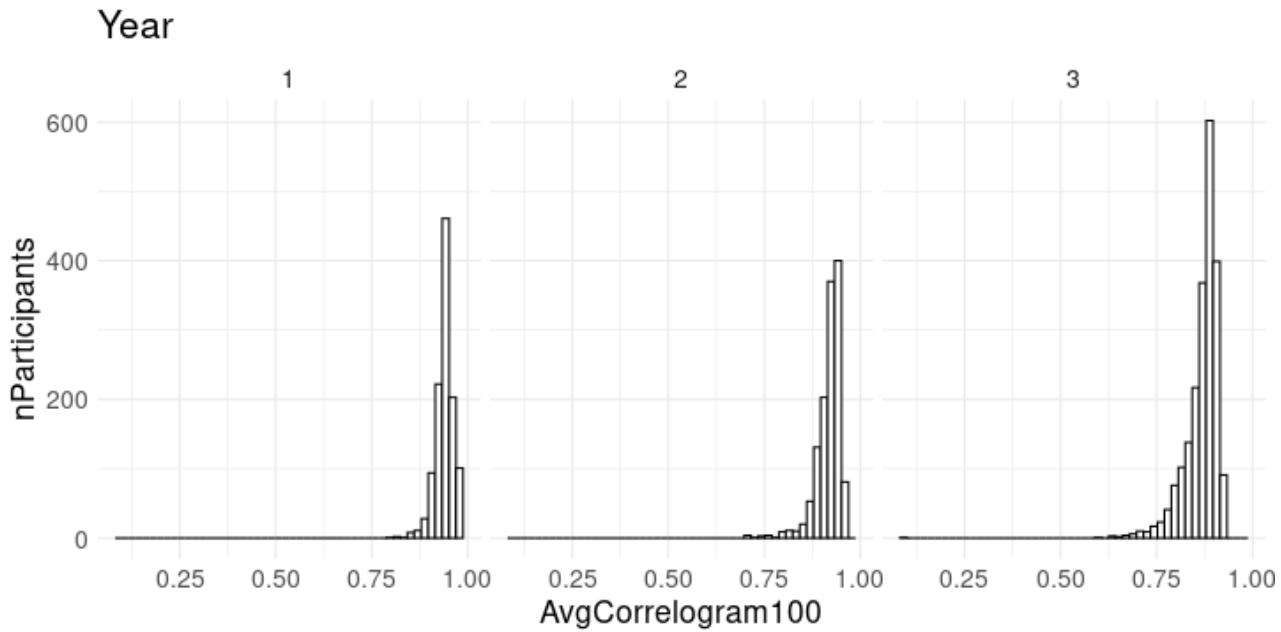
Reads per Sample



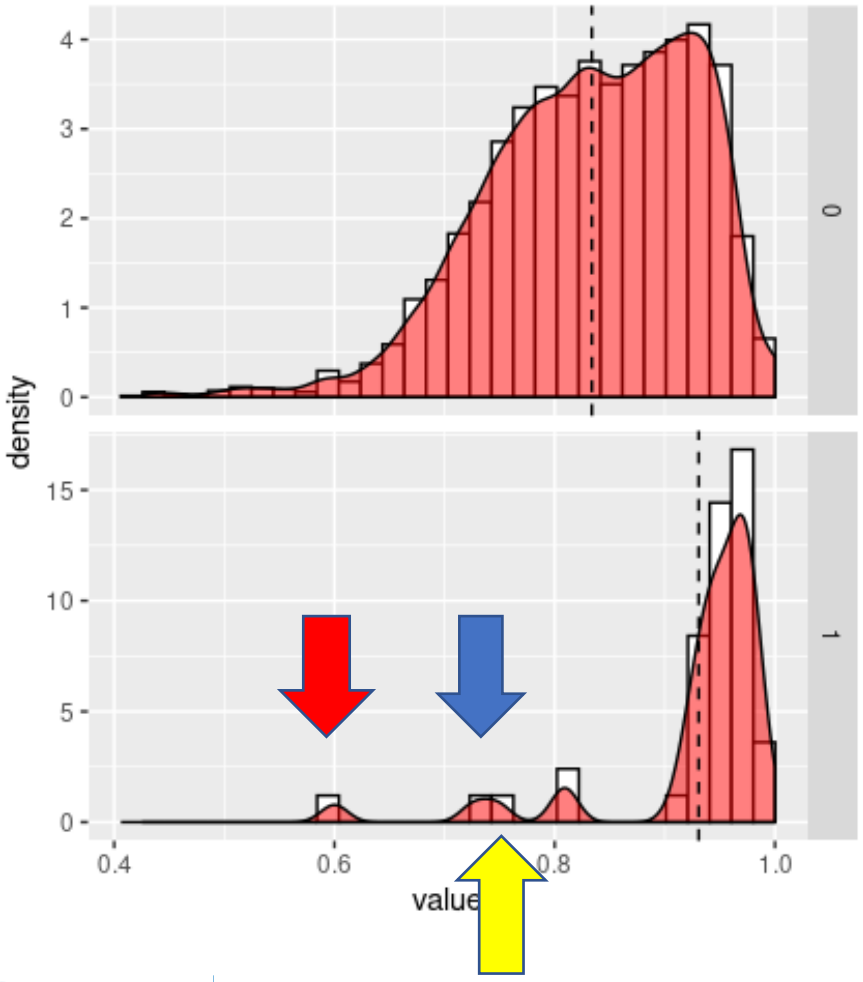
Read Mapping



Sample Correlation

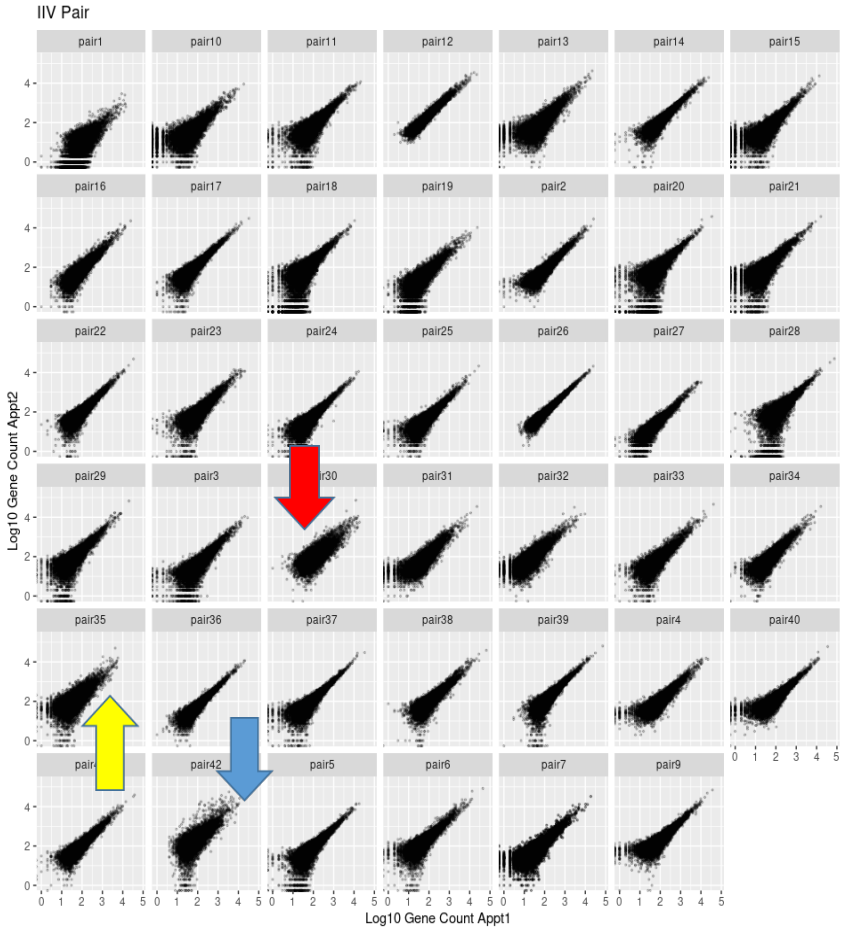


Genome-wide Replication Correlation

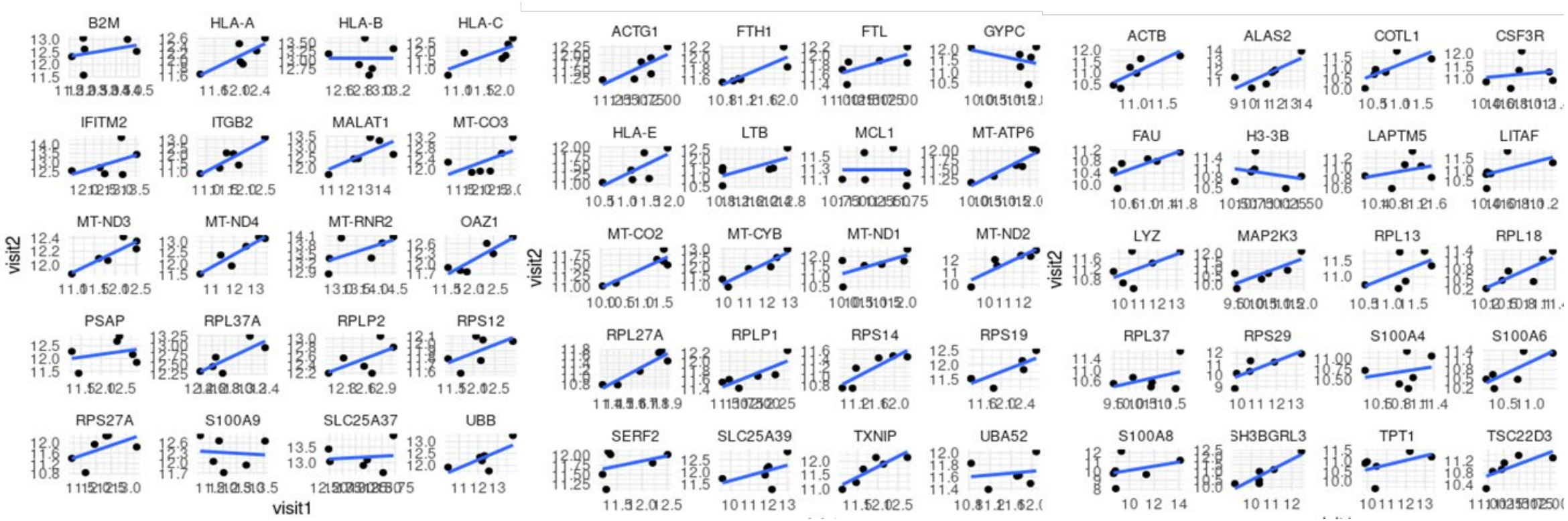


Not matched IIV

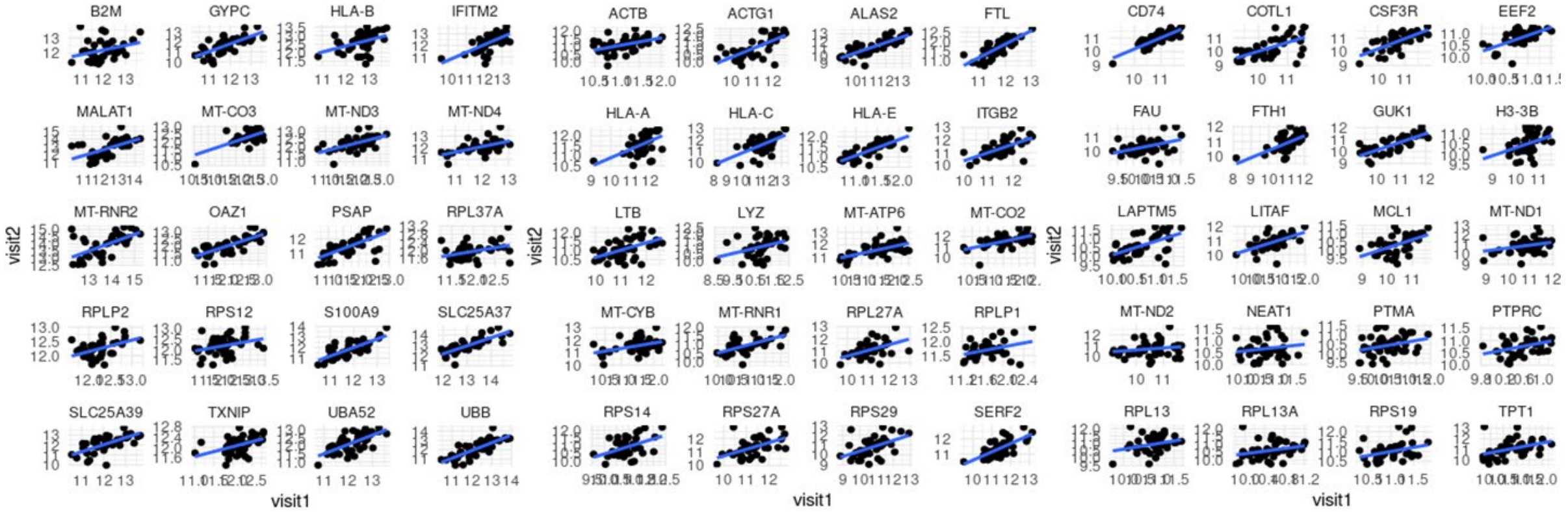
Matched IIV



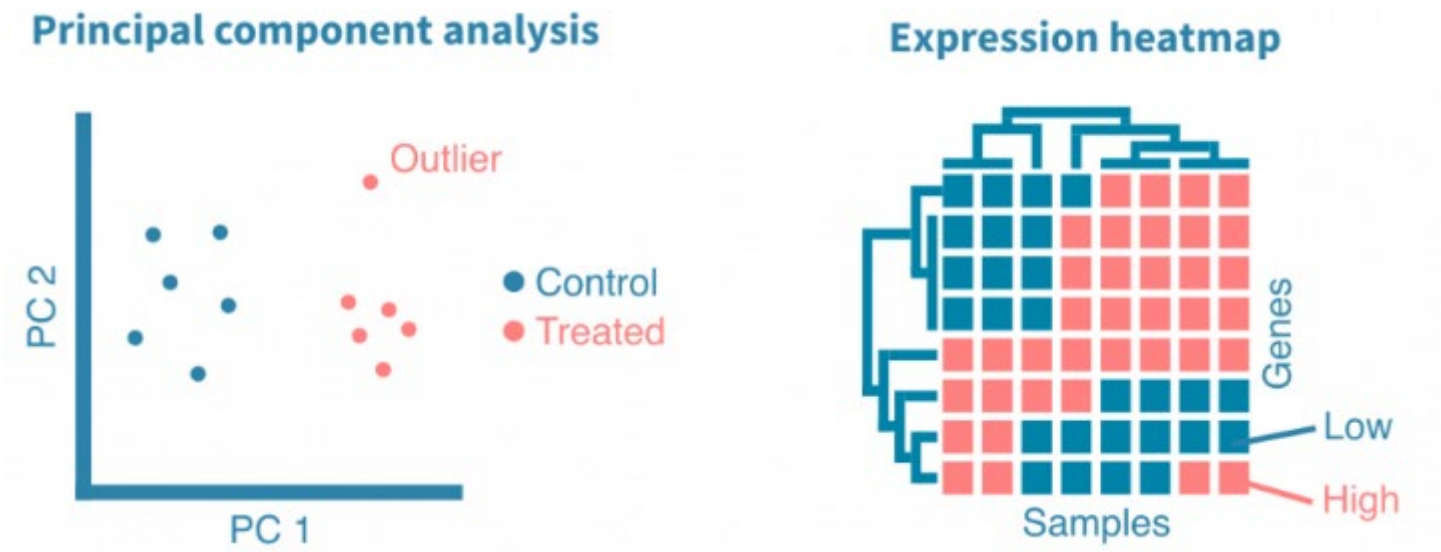
Gene-wise Replicate Correlation



Gene-wise Replicate Correlation

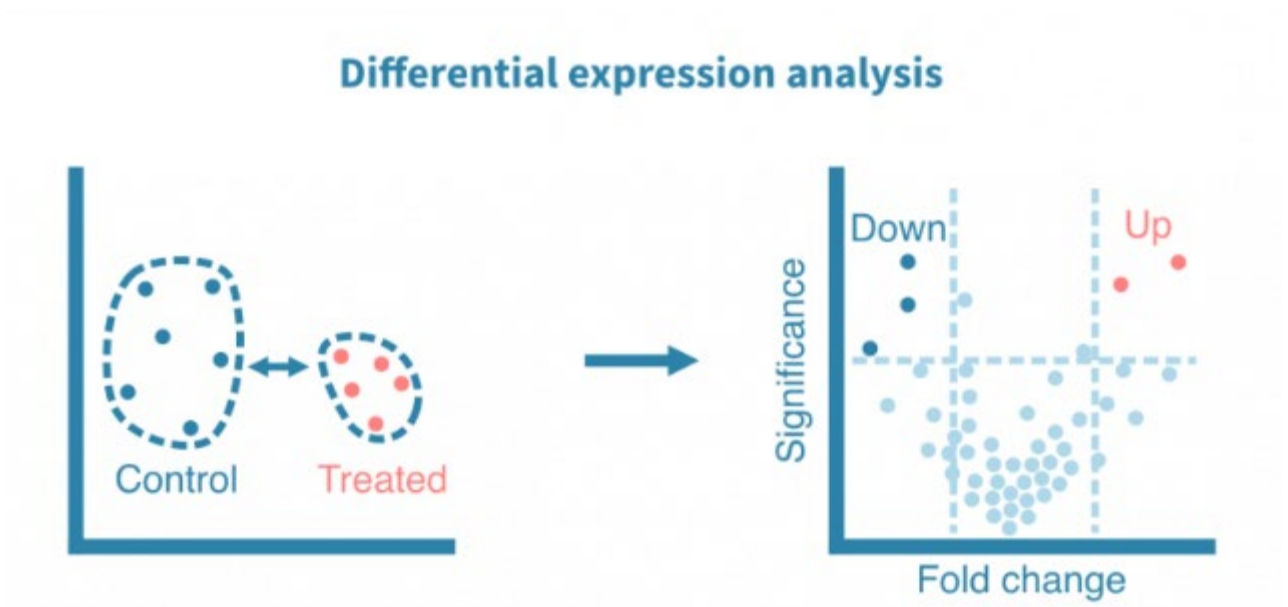


Gene Expression Visualization



GeneVia Technologies Infographics. Retrieved 06/2024

Differential Gene Expression



GeneVia Technologies Infographics. Retrieved 06/2024

Gene Sets and Biology

- Genes work together in gene sets
- KEGG gene onotology/pathways
- Reactome and disease processes
- Bespoke gene sets (CTRA)

Gene ontology

Live display

https://www.gsea-msigdb.org/gsea/msigdb/human/genesets.jsp?collection=HPO

Gene Pathways

Live display

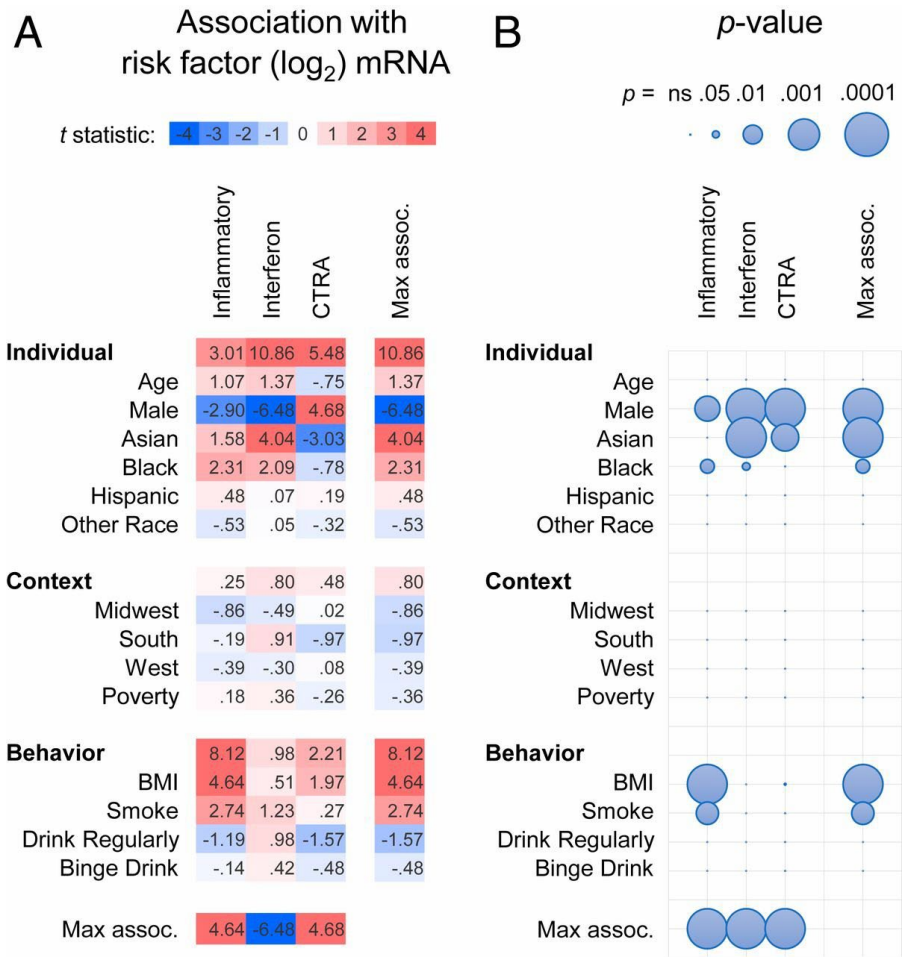
https://www.genome.jp/kegg/pathway.html

Reactome Browser

Live display

https://reactome.org/PathwayBrowser/

Bespoke Gene Sets



Cole SW, Shanahan MJ, Gaydosh L, Harris KM. Population-based RNA profiling in Add Health finds social disparities in inflammatory and antiviral gene regulation to emerge by young adulthood. Proc Natl Acad Sci U S A. 2020 Mar 3;117(9):4601-4608.

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Cold-Flu Like Symptoms

H5Q014A - Q014a COLD/FLU-LIKE SYMPTOMS (2 WKS) -W5

Type	Code
Measurement Unit	Numeric
H5Q014A	Q014. In the last TWO weeks, have you had any of the following? a. Cold or Flu-like symptoms such as sore throat, runny nose, or cough

			Frequency	% of total	% of valid
Valid	0	no	4,504	83.7%	83.9%
	1	yes	862	16.0%	16.1%
		Total	5,366	99.7%	100%
Missing	98	don't know	15	0.3%	
		Total	15	0.3%	

Example Code

```
#load R libraries
library(edgeR)
library(sva)

#convert csv or data frame to DGEList object
x=DGEList(counts=dataset)

#filter lowly expressed genes
keep = filterByExpr(x)
x = x[keep,  keep.lib.sizes = F]

#normalize samples based upon library size
x = calcNormFactors(x, method = "TMM")
```

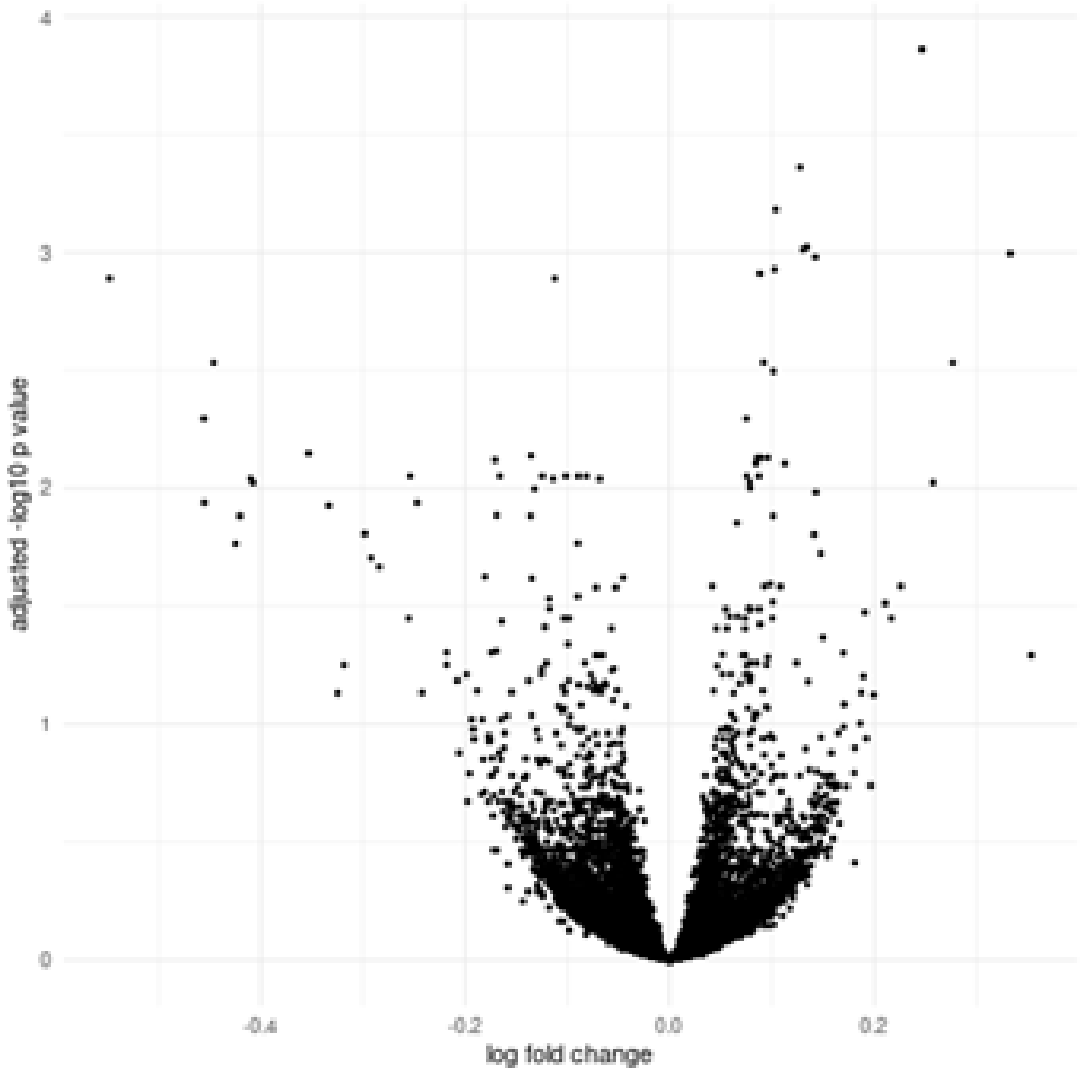
Example Code

```
#construct design matrix
design = model.matrix(~0+exposure+covariate1+covariate2...covariateN, dat=dataset)

#remove heteroscedasticity from counts
v=voom(x, design, plot = F)

#batch correct
v$E=ComBat(v$E, dataset$batch)

#construct linear model for each gene with exposure/covariates
fit = lmFit(v$E, design)
fit = eBayes(fit,trend = T)

#determine differentially expressed genes
hits= topTable(fit,number = Inf, coef = 'exposure')
```

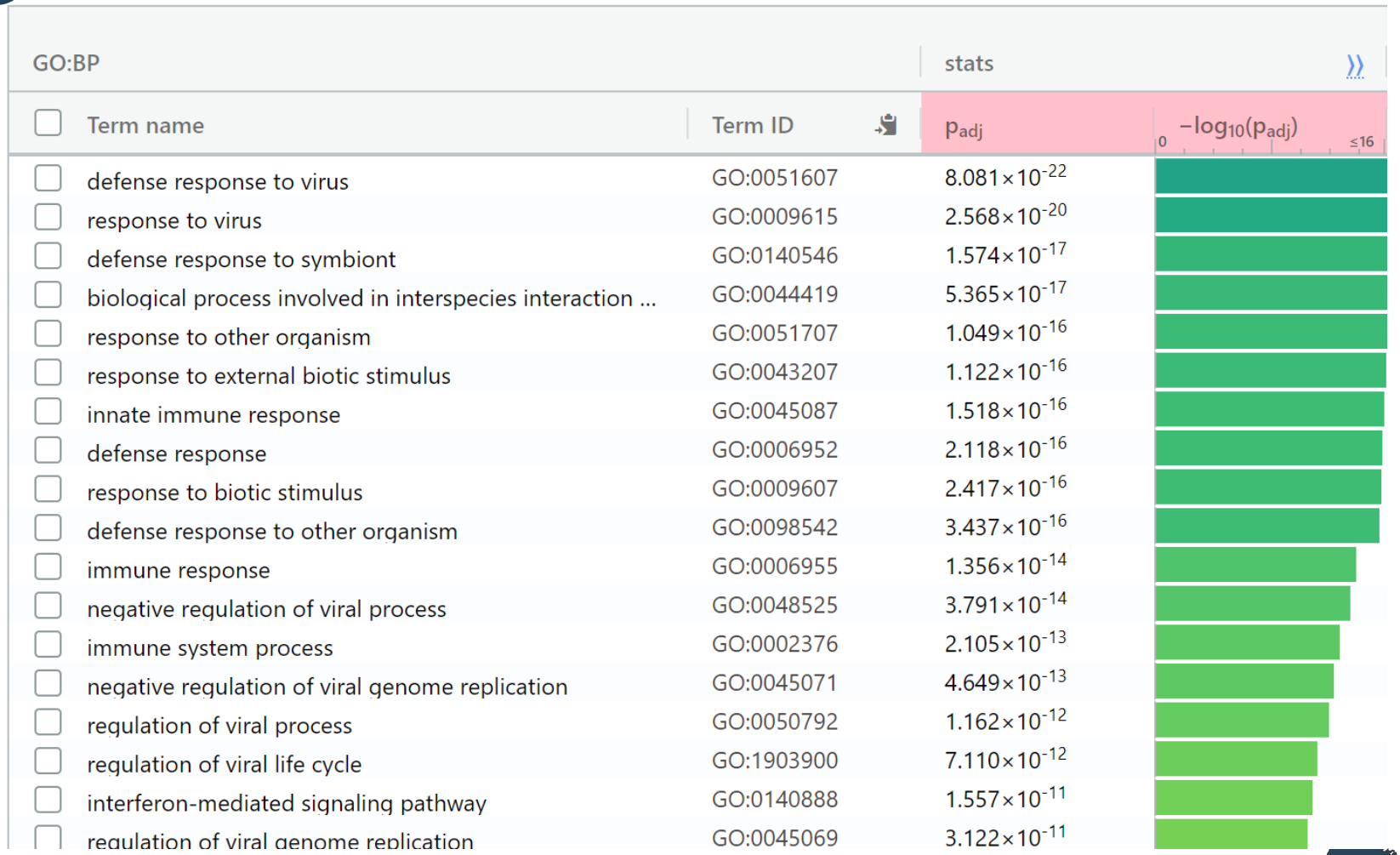
Genes Associated with Cold/Flu-like Symptoms

gene	logFC	AveExpr	t	P.Value	adj.P.Val	B	hgnc_symbol	entrezgene_id
ENSG00000133106	0.65	5.27	6.91	7.89E-12	1.06E-07	16.3	EPSTI1	94240
ENSG00000165949	0.99	3.64	6.49	1.24E-10	8.32E-07	13.33	IFI27	3429
ENSG00000068079	0.48	4.04	6.3	4.17E-10	1.40E-06	12.27	IFI35	3430
ENSG00000119917	0.52	6.97	6.14	1.08E-09	2.47E-06	11.7	IFIT3	3437
ENSG00000137965	0.85	2.71	6.3	4.13E-10	1.40E-06	11.66	IFI44	10561
ENSG00000111331	0.57	4.79	6.14	1.10E-09	2.47E-06	11.59	OAS3	4940
ENSG00000123610	0.55	4.99	6.07	1.68E-09	3.23E-06	11.22	TNFAIP6	7130
ENSG00000160932	0.48	6.18	5.69	1.62E-08	2.73E-05	9.14	LY6E	4061
ENSG00000137959	0.62	5.39	5.55	3.52E-08	4.73E-05	8.4	IFI44L	10964
ENSG00000134321	0.77	4.05	5.56	3.33E-08	4.73E-05	8.32	RSAD2	91543
ENSG00000185745	0.63	4.21	5.29	1.44E-07	0.000164	7.02	IFIT1	3434
ENSG00000102524	0.25	7.74	5.26	1.66E-07	0.000172	6.9	TNFSF13B	10673
ENSG00000187608	0.48	6.45	5.17	2.75E-07	0.000246	6.46	ISG15	9636
ENSG00000134326	0.78	2.4	5.29	1.46E-07	0.000164	6.45	CMPK2	129607
ENSG00000136514	0.61	2.84	5.19	2.41E-07	0.000232	6.17	RTP4	64108
ENSG00000186407	0.28	5.52	4.9	1.11E-06	0.000932	5.19	CD300E	342510
ENSG00000089127	0.38	5.17	4.82	1.63E-06	0.001289	4.84	OAS1	4938

Genes Associated with Cold/Flu-like Symptoms



Biological Processes

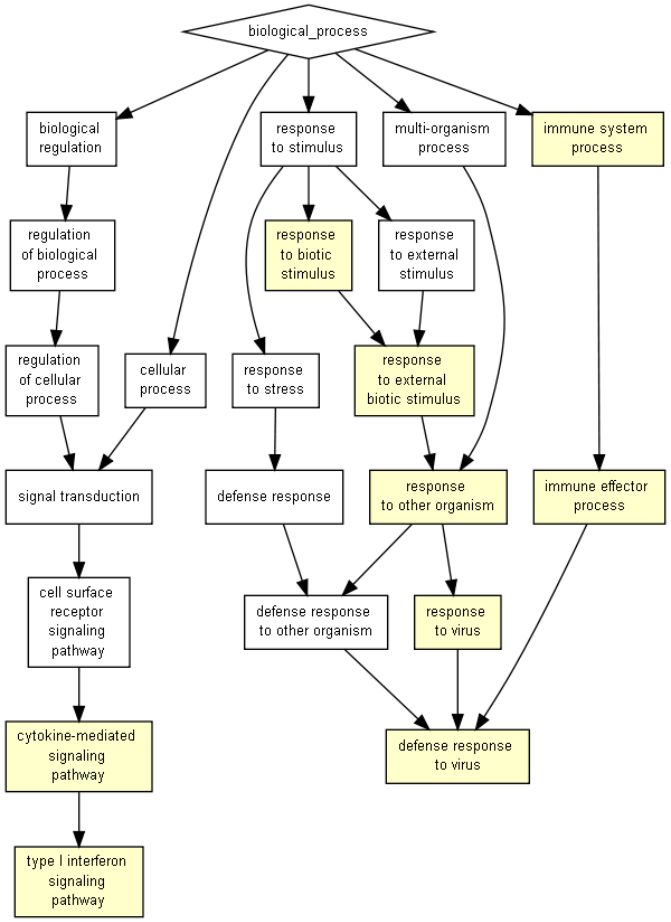


GO:BP

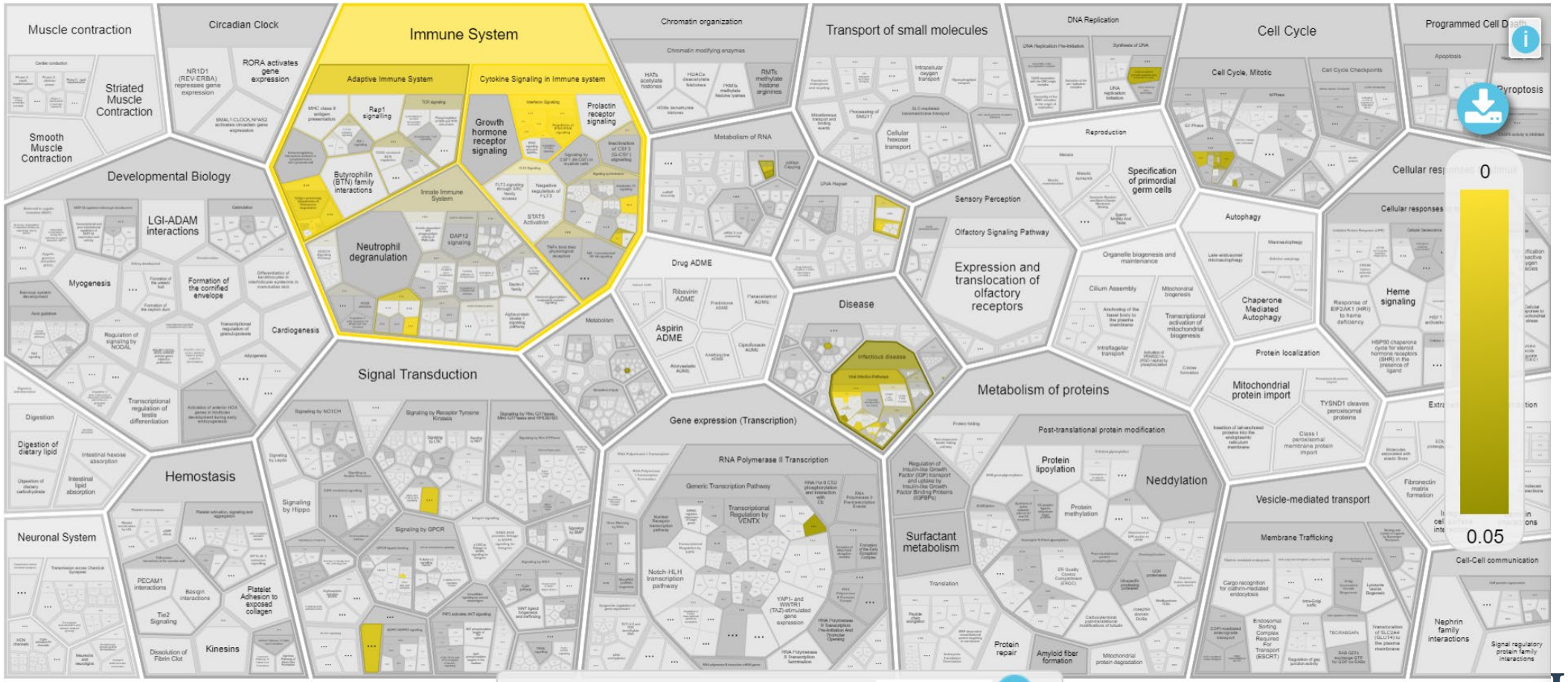
stats

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ (0 – ≤16)
defense response to virus	GO:0051607	8.081×10^{-22}	
response to virus	GO:0009615	2.568×10^{-20}	
defense response to symbiont	GO:0140546	1.574×10^{-17}	
biological process involved in interspecies interaction ...	GO:0044419	5.365×10^{-17}	
response to other organism	GO:0051707	1.049×10^{-16}	
response to external biotic stimulus	GO:0043207	1.122×10^{-16}	
innate immune response	GO:0045087	1.518×10^{-16}	
defense response	GO:0006952	2.118×10^{-16}	
response to biotic stimulus	GO:0009607	2.417×10^{-16}	
defense response to other organism	GO:0098542	3.437×10^{-16}	
immune response	GO:0006955	1.356×10^{-14}	
negative regulation of viral process	GO:0048525	3.791×10^{-14}	
immune system process	GO:0002376	2.105×10^{-13}	
negative regulation of viral genome replication	GO:0045071	4.649×10^{-13}	
regulation of viral process	GO:0050792	1.162×10^{-12}	
regulation of viral life cycle	GO:1903900	7.110×10^{-12}	
interferon-mediated signaling pathway	GO:0140888	1.557×10^{-11}	
regulation of viral genome replication	GO:0045069	3.122×10^{-11}	

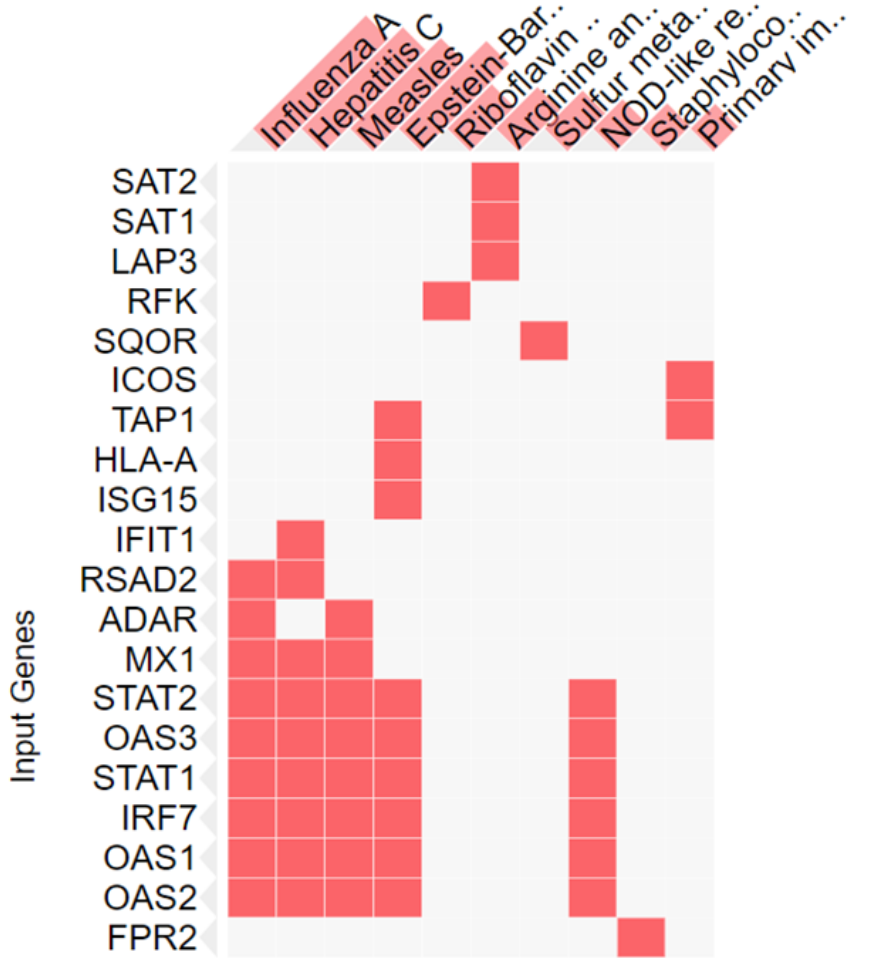
Gene Ontology Terms

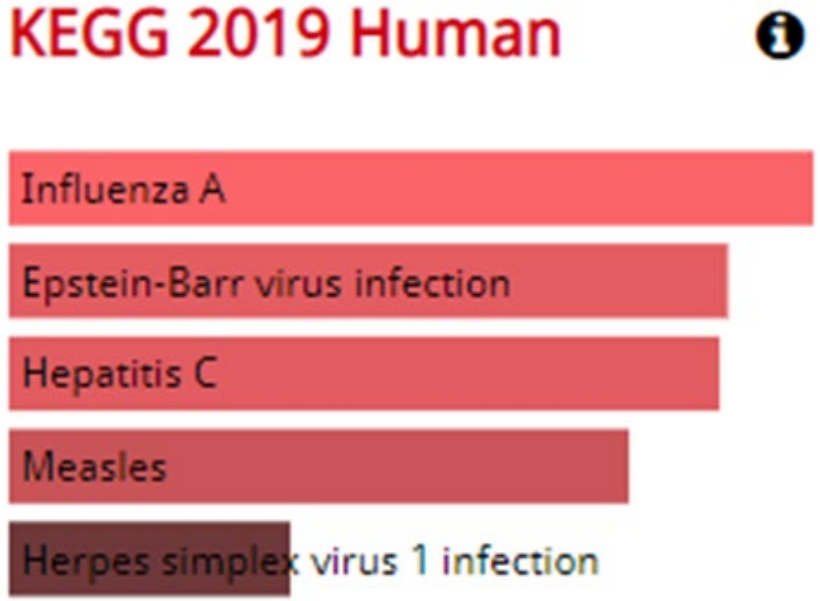


Reactome

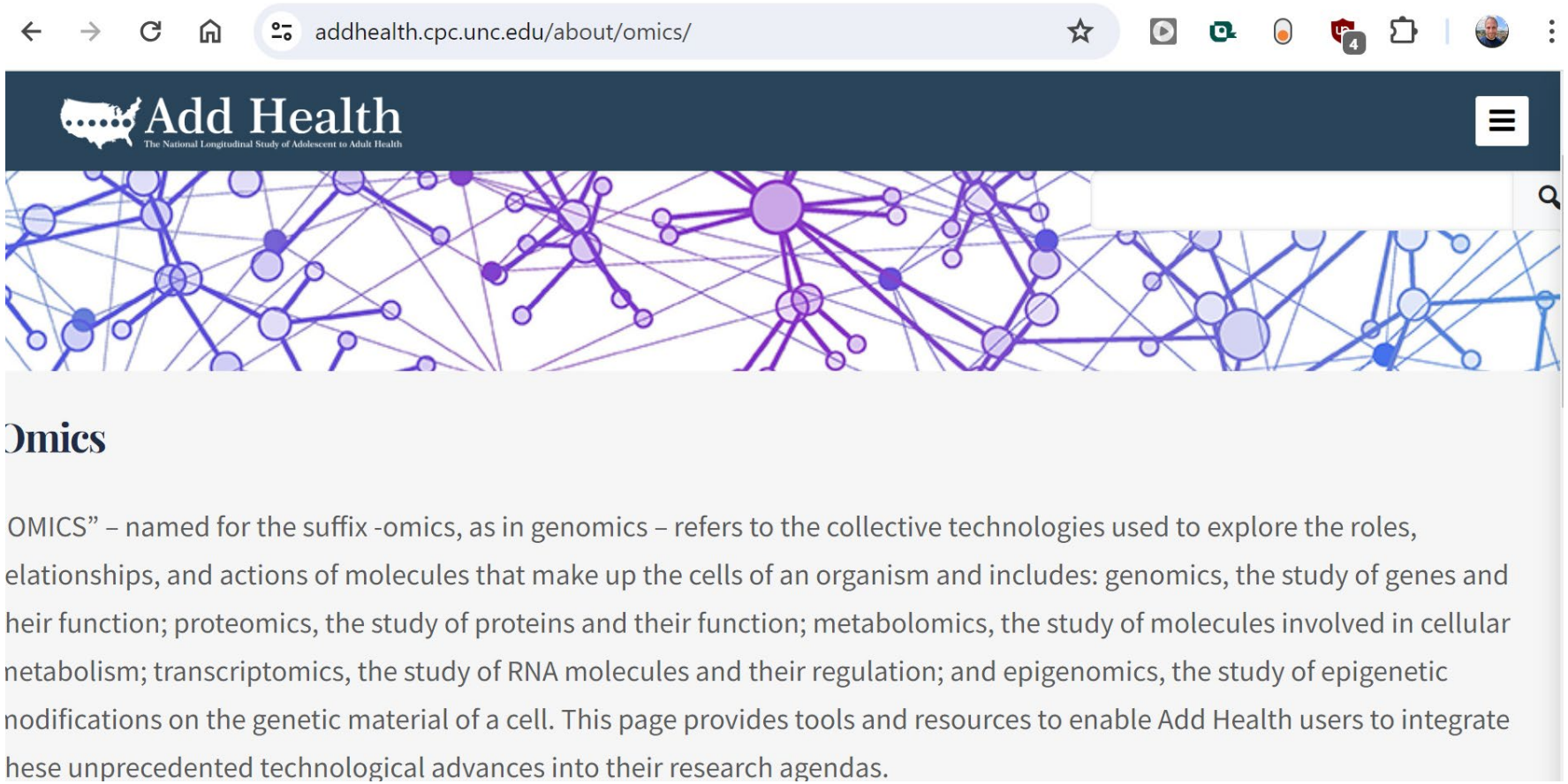

Muscle contraction
Striated Muscle Contraction
Smooth Muscle Contraction
Circadian Clock
NR1D1 (REV-ERBA) represses gene expression
RORA activates gene expression
Developmental Biology
LGI-ADAM interactions
Myogenesis
Formation of the cornified envelope
Cardiogenesis
Transcriptional regulation of testis differentiation
Digestion
Digestion of dietary lipid
Intestinal hexose absorption
Digestion of dietary carbohydrate
Intestinal lipid absorption
Neuronal System
Hemostasis
PECAM1 interactions
Basigin interactions
Platelet Adhesion to exposed collagen
Tie2 Signaling
Dissolution of Fibrin Clot
Kinesins
Immune System
Adaptive Immune System
Cytokine Signaling in Immune system
Rap1 signaling
Butyrophilin (BTN) family interactions
Growth hormone receptor signaling
Prolactin receptor signaling
Innate Immune System
Neutrophil degranulation
DAP12 signaling
STAT5 Activation
Signal Transduction
Signaling by Hippo
Signaling by GPCR
Chromatin organization
Transport of small molecules
DNA Replication
Cell Cycle
Cell Cycle, Mitotic
Programmed Cell Death
Pyroptosis
Metabolism of RNA
Reproduction
Specification of primordial germ cells
Sensory Perception
Olfactory Signaling Pathway
Expression and translocation of olfactory receptors
Autophagy
Chaperone Mediated Autophagy
Cellular responses
Heme signaling
Drug ADME
Aspirin ADME
Disease
Infectious disease
Metabolism
Gene expression (Transcription)
RNA Polymerase II Transcription
Generic Transcription Pathway
Transcriptional Regulation by VENTX
Notch-HLH transcription pathway
YAP1- and WWTR1 (TAZ)-stimulated gene expression
Metabolism of proteins
Post-translational protein modification
Protein lipoylation
Neddylation
Protein methylation
Surfactant metabolism
Protein repair
Amyloid fiber formation
Protein localization
Mitochondrial protein import
TYSND1 cleaves peroxisomal proteins
Vesicle-mediated transport
Membrane Trafficking
Cell-Cell communication
Nephrin family interactions
0
0.05
OVERREPRESENTATION
Showing
pValue

Disease Processes





Access


addhealth.cpc.unc.edu/about/omics/
Add Health
The National Longitudinal Study of Adolescent to Adult Health
Omics
OMICS" – named for the suffix -omics, as in genomics – refers to the collective technologies used to explore the roles,
elationships, and actions of molecules that make up the cells of an organism and includes: genomics, the study of genes and
heir function; proteomics, the study of proteins and their function; metabolomics, the study of molecules involved in cellular
netabolism; transcriptomics, the study of RNA molecules and their regulation; and epigenomics, the study of epigenetic
nodifications on the genetic material of a cell. This page provides tools and resources to enable Add Health users to integrate
hese unprecedented technological advances into their research agendas.

Access


addhealth.cpc.unc.edu/about/omics/
Add Health
The National Longitudinal Study of Adolescent to Adult Health
How To Access Data
How to access data
Need help accessing the Add Health GWAS data?
How to Obtain Add Health OMICS data from dbGaP
Interested in accessing a file to link genomic data to phenotype data?
How to Obtain Add Health Phenotype Data and/or dbGaP Linkage File

Limitations

- Only from peripheral blood, not other tissue types
- Only from one point in time
- Noisy data type
- Ancestry differences in gene function
- Gene redundancies
- Technical complexity

Acknowledgements

Add Health is directed by Robert A. Hummer and funded by the National Institute on Aging cooperative agreements U01 AG071448 (Hummer) and U01AG071450 (Aiello and Hummer) at the University of North Carolina at Chapel Hill. Waves I-V data are from the Add Health Program Project, grant P01 HD31921 (Harris) from *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), with cooperative funding from 23 other federal agencies and foundations. Add Health was designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill.

Thank you

- Kathleen Mullan Harris
- Lauren Gaydosh
- Michael Shanahan
- Steve Cole
- Justin Chumbley
- Ravi Sudharshan
- Cecilia Potente