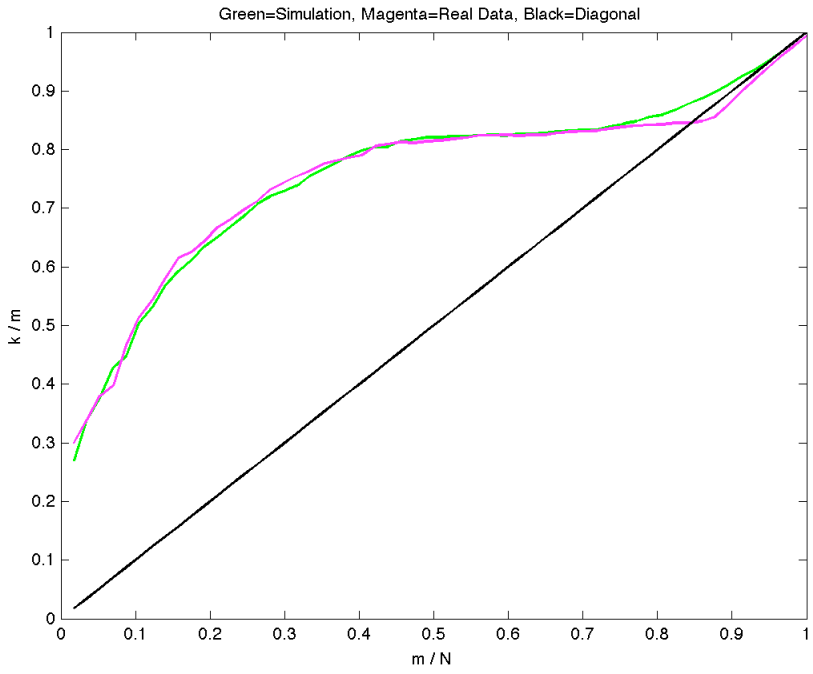
Supplement for CORaL: Comparison of Ranked Lists for Analysis of Gene Expression Data, by Antosh et al.

Table of Contents:

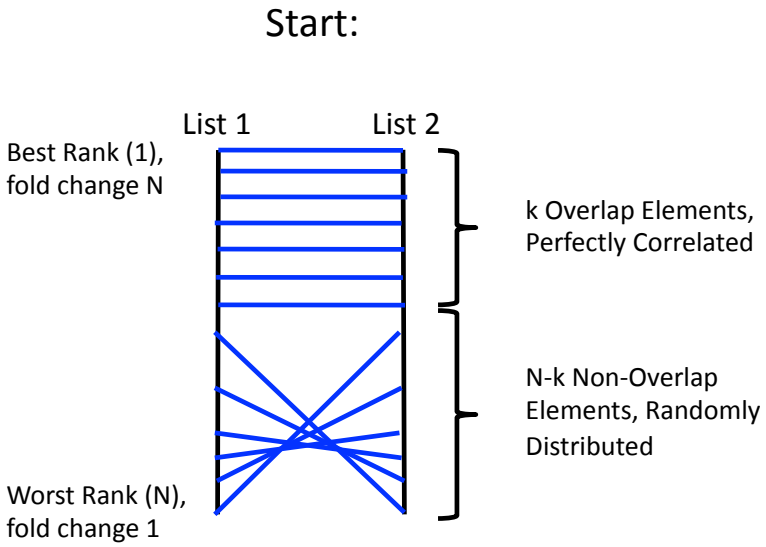
Diagnostic Plot Verification of Simulation Method



Diagnostic Plot is: for a given set size m, the overlap in the top m elements of each ranked list (k) divided by m versus m divided by the total list length N. We used a 1-

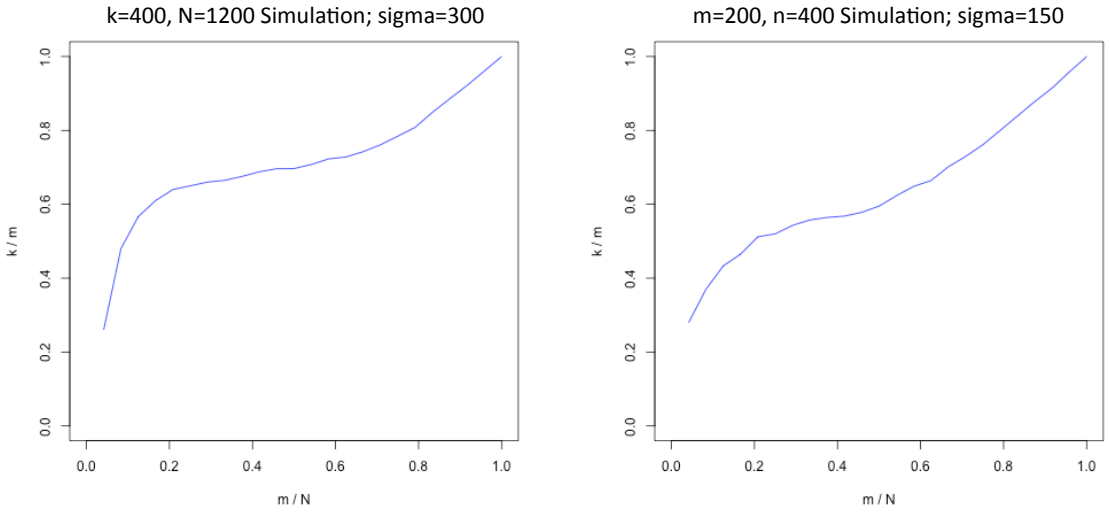
dimensional (m=n) approach as in Antosh et al. (2011) to represent the data because it is easier to visualize.

Illustration of Simulation Method



Then, subtract from each fold change the absolute value of a number drawn from N(0,sigma), where sigma is a parameter of the simulation

Example Diagnostic Plot for Realistic Noise Simulations



Compare these plots with the real data diagnostic figure on page 1. The desired output is to have no significant decreases in (k/m) with increasing (m/N).

Overlaps for Pearson Data in All 3 Algorithms:

Note: "resv" is resveratrol, "DR" is dietary restriction

Yang et al. (2006) Method OrderedList:

	fat DR	fat high resv	fat low resv	heart DR	heart high resv	heart low resv	liver DR	liver high resv	liver low resv	muscle DR	muscle high resv
fat high resv	0										
fat low resv	0	0									
heart DR	0	0	0								
heart high resv	0	0	0	0							
heart low resv	0	0	0	0	0						
liver DR	149	0	13	0	0	0					
liver high resv	103	0	0	0	0	0	1090				
liver low resv	35	0	0	0	0	0	0	0			
muscle DR	0	0	0	0	0	0	191	0	0		
muscle high resv	211	146	255	0	0	14	37	0	13	0	
muscle low resv	0	0	0	0	0	0	0	0	0	0	0

Plaisier et al. Method Rank-Rank Hypergeometric Overlap (RRHO):

	fat DR	fat high resv	fat low resv	heart DR	heart high resv	heart low resv	liver DR	liver high resv	liver low resv	muscle DR	muscle high resv
fat high resv	11205										
fat low resv	11115	11111									
heart DR	1970	1617	2306								
heart high resv	341	646	56	8573							
heart low resv	698	416	20	8451	10398						
liver DR	4906	5582	1209	2770	3151	919					
liver high resv	3804	1046	638	2837	3288	708	12729				
liver low resv	1079	1105	993	1467	1382	2286	9003	8363			
muscle DR	1274	0	21	1358	1760	1328	2387	2380	797		
muscle high resv	261	458	294	15	2257	118	38	917	363	9331	
muscle low resv	0	0	0	0	2032	2070	891	766	1514	9132	9985

Antosh et al. Method CORaL:

	fat DR	fat high resv	fat low resv	heart DR	heart high resv	heart low resv	liver DR	liver high resv	liver low resv	muscle DR	muscle high resv
fat high resv	5462										
fat low resv	6506	6976									
heart DR	21	0	7								
heart high resv	0	9	0	76							
heart low resv	14	0	14	42	2277						
liver DR	202	0	110	24	0	10					
liver high resv	111	8	18	7	24	0	5852				
liver low resv	23	0	13	0	0	10	781	2743			
muscle DR	27	0	19	28	17	8	32	13	8		
muscle high resv	193	36	134	10	24	31	35	19	11	2829	
muscle low resv	0	0	0	14	61	24	0	0	0	2110	4294

<u>Pearson Overlaps Correlation and Significance of Correlation</u>

Results of CORaL Analysis on Pearson Data

DR stands for dietary restriction, highRes and lowRes stand for high and low doses of resveratrol

The four tissues measured are fat, heart, liver, and muscle

Correlation is spearman correlation, statistical significance was tested using the cor.test function in R with option exact=FALSE

All overlaps greater than 35 genes in either direction have a correlation with statistical significance p value less than 0.05

Condition 1	Condition 2	Overlap Up	Correlation Up	Correlation p Value Up	Overlap Down	Correlation Down	Correlation p Value Down
fat DR	fat high resv	2562	0.63	7.7E-287	2900	0.48	5.0E-165
fat DR	fat low resv	2921	0.63	0.0E+00	3585	0.70	0.0E+00
fat DR	heart DR	0	0.00	1.0E+00	21	0.19	4.1E-01
fat DR	heart high resv	0	0.00	1.0E+00	0	0.00	1.0E+00
fat DR	heart low resv	0	0.00	1.0E+00	14	0.07	8.1E-01
fat DR	liver DR	14	0.26	3.7E-01	188	0.36	4.6E-07
fat DR	liver high resv	15	0.41	1.2E-01	96	0.41	3.3E-05
fat DR	liver low resv	11	-0.03	9.4E-01	12	-0.05	8.8E-01
fat DR	muscle DR	8	-0.07	8.7E-01	19	-0.02	9.2E-01
fat DR	muscle high resv	0	0.00	1.0E+00	193	0.31	1.1E-05
fat DR	muscle low resv	0	0.00	1.0E+00	0	0.00	1.0E+00
fat high resv	fat low resv	3192	0.74	0.0E+00	3784	0.63	0.0E+00
fat high resv	heart DR	0	0.00	1.0E+00	0	0.00	1.0E+00
fat high resv	heart high resv	0	0.00	1.0E+00	9	-0.12	7.7E-01
fat high resv	heart low resv	0	0.00	1.0E+00	0	0.00	1.0E+00
fat high resv	liver DR	0	0.00	1.0E+00	0	0.00	1.0E+00
fat high resv	liver high resv	0	0.00	1.0E+00	8	0.50	2.1E-01
fat high resv	liver low resv	0	0.00	1.0E+00	0	0.00	1.0E+00
fat high resv	muscle DR	0	0.00	1.0E+00	0	0.00	1.0E+00
fat high resv	muscle high resv	0	0.00	1.0E+00	36	0.48	2.8E-03
fat high resv	muscle low resv	0	0.00	1.0E+00	0	0.00	1.0E+00
fat low resv	heart DR	0	0.00	1.0E+00	7	0.32	4.8E-01
fat low resv	heart high resv	0	0.00	1.0E+00	0	0.00	1.0E+00
fat low resv	heart low resv	0	0.00	1.0E+00	14	0.07	8.2E-01
fat low resv	liver DR	0	0.00	1.0E+00	110	0.36	1.2E-04
fat low resv	liver high resv	0	0.00	1.0E+00	18	0.51	3.0E-02
fat low resv	liver low resv	0	0.00	1.0E+00	13	-0.20	5.1E-01
fat low resv	muscle DR	0	0.00	1.0E+00	19	0.51	2.6E-02
fat low resv	muscle high resv	0	0.00	1.0E+00	134	0.30	3.6E-04
fat low resv	muscle low resv	0	0.00	1.0E+00	0	0.00	1.0E+00
heart DR	heart high resv	36	0.62	5.0E-05	40	0.69	1.1E-06
heart DR	heart low resv	0	0.00	1.0E+00	42	0.67	1.0E-06
heart DR	liver DR	0	0.00	1.0E+00	24	0.14	5.2E-01
heart DR	liver high resv	0	0.00	1.0E+00	7	-0.64	1.2E-01
heart DR	liver low resv	0	0.00	1.0E+00	0	0.00	1.0E+00
heart DR	muscle DR	0	0.00	1.0E+00	28	0.41	3.0E-02
heart DR	muscle high resv	0	0.00	1.0E+00	10	0.38	2.8E-01
heart DR	muscle low resv	14	0.34	2.4E-01	0	0.00	1.0E+00
heart high resv	heart low resv	285	0.20	6.5E-04	1992	0.36	1.5E-60
heart high resv	liver DR	0	0.00	1.0E+00	0	0.00	1.0E+00
heart high resv	liver high resv	11	0.35	3.0E-01	13	0.19	5.4E-01
heart high resv	liver low resv	0	0.00	1.0E+00	0	0.00	1.0E+00
heart high resv	muscle DR	6	-0.26	6.2E-01	11	-0.22	5.2E-01
heart high resv	muscle high resv	0	0.00	1.0E+00	24	0.36	8.6E-02
heart high resv	muscle low resv	52	0.49	1.9E-04	9	0.48	1.9E-01
heart low resv	liver DR	0	0.00	1.0E+00	10	0.77	9.2E-03
heart low resv	liver high resv	0	0.00	1.0E+00	0	0.00	1.0E+00

heart low resv	liver low resv	0	0.00	1.0E+00	10	0.09	8.0E-01
heart low resv	muscle DR	0	0.00	1.0E+00	8	0.71	4.7E-02
heart low resv	muscle high resv	0	0.00	1.0E+00	31	0.33	6.6E-02
heart low resv	muscle low resv	11	-0.07	8.3E-01	13	0.38	2.0E-01
liver DR	liver high resv	2806	0.58	1.1E-253	3046	0.56	2.4E-248
liver DR	liver low resv	341	0.31	3.1E-09	440	0.37	1.1E-15
liver DR	muscle DR	9	0.17	6.7E-01	23	0.26	2.3E-01
liver DR	muscle high resv	0	0.00	1.0E+00	35	0.38	2.5E-02
liver DR	muscle low resv	0	0.00	1.0E+00	0	0.00	1.0E+00
liver high resv	liver low resv	611	0.43	2.9E-29	2132	0.51	1.4E-142
liver high resv	muscle DR	6	0.54	2.7E-01	7	0.07	8.8E-01
liver high resv	muscle high resv	0	0.00	1.0E+00	19	0.54	1.6E-02
liver high resv	muscle low resv	0	0.00	1.0E+00	0	0.00	1.0E+00
liver low resv	muscle DR	8	0.07	8.7E-01	0	0.00	1.0E+00
liver low resv	muscle high resv	0	0.00	1.0E+00	11	-0.47	1.4E-01
liver low resv	muscle low resv	0	0.00	1.0E+00	0	0.00	1.0E+00
muscle DR	muscle high resv	1188	0.37	1.4E-39	1641	0.41	1.2E-68
muscle DR	muscle low resv	650	0.38	2.5E-23	1460	0.36	1.9E-45
muscle high resv	muscle low resv	1757	0.41	4.0E-73	2537	0.46	5.0E-135

Psuedo Code For CORaL

CORaL Pseudo Code

- Read in data
- Rank data

- Define list_1_set_sizes as (step size) to (list length) by (step size)
- Define list_2_set_sizes as (step size) to (list length) by (step size)

- Calculate matrix of ranked list overlaps, with element (i,j) being the overlap between the top (list_1_set_sizes element i) in list 1 and the top (list_2_set_sizes element j) in list 2

- For every possible combination of values in list_1_set_sizes and list_2_set_sizes
  {
  Calculate step p values using equations 2, 3 and 4; calculate for for 3 directions: up one step in both lists, up one step in list 1, up one step in list 2
  }

- Correct all step p values with Benjamini-Yekutieli correction

- For every possible combination of values in list_1_set_sizes and list_2_set_sizes
  {
  Calculate T using equations 5 and 6. This can be done recursively, as T(list_1_set_sizes index i,list_2_set_sizes index j) depends on T(i-1,j-1), T(i-1,j) and T(i,j-1)
  Calculate S using equation 7
  Calculate L using equation 1
  }
- Define chosen set size list 1 and chosen set size list 2 as the values of list_1_set_sizes and list_2_set_sizes such that the corresponding value of L is maximized.

R Code for Plaisier Simulations

At the time of running simulations, the Plaisier group web site (systems.crump.ucla.edu/rankrank/) was not available. We used the following code to calculate the maximum (Benjamini-Yekutieli corrected) Fisher's Exact Test p value. The code is written in the language of the statistical computing language R.

```
Plaisier_Simulations<-function(input_file_1,input_file_2,delta,output_name,m){
  data1<-read.table(input_file_1,sep="\t",quote="",header=F,colClasses="numeric")
  data2<-read.table(input_file_2,sep="\t",quote="",header=F,colClasses="numeric")
  N<-dim(data1)[1]
  nsim<-dim(data1)[2]
  ij<-matrix(0,nsim,2)
  kRecovered<-rep(0,nsim) #measures all overlap found
  kRecoveredStartOnly<-rep(0,nsim) #measures the original overlap genes
  kStats<-matrix(0,nsim,4)

  for(simCount in 1:nsim){
```

```
    list1<-data1[,simCount]
    list2<-data2[,simCount]
    temp<-Plaisier_substitute_function(list1,list2,delta)
     ij[simCount,]<-temp[1:2]
     kRecovered[simCount]<-temp[3]
     if(temp[1]>0){
       kRecoveredStartOnly[simCount]<-
length(intersect(list1[1:(delta*temp[1])],intersect(list2[1:(delta*temp[2])],1:m)))
       }
       if(temp[1]==0){
              kRecoveredStartOnly[simCount]<-0
       }
  }

  summary<-matrix(0,N/delta,N/delta)
  for(tempCount in 1:nsim){
    summary[ij[tempCount,1],ij[tempCount,2]]<-summary[ij[tempCount,1],ij[tempCount,2]] +
1
  }

  png(paste("Overlap_Histogram_",output_name,".png",sep=""))
  hist(kRecovered)
  dev.off()


print(c(mean(kRecovered),sqrt(var(kRecovered)),mean(kRecoveredStartOnly),sqrt(var(kRecove
redStartOnly))))


write.table(summary,paste("Simulation_Results_Summary",output_name,".txt",sep=""),sep="\t
",quote=F,row.names=F,col.names=FALSE)
}


Plaisier_substitute_function<-function(list1,list2,delta){
  N<-length(list1)
  comparison_number<-seq(delta,N,delta)

  k<-matrix(0,length(comparison_number),length(comparison_number))
  pPlaisier<-matrix(0,length(comparison_number),length(comparison_number))

  for(i in 1:length(comparison_number)){
    for(j in 1:length(comparison_number)){
      k[i,j]<-
length(intersect(list1[1:comparison_number[i]],list2[1:comparison_number[j]]))
      pPlaisier[i,j]<-phyper(k[i,j]-1,comparison_number[i],N-
comparison_number[i],comparison_number[j],lower.tail=F)
    }
  }
  pPlaisier_FDR_BY<-pPlaisier
  dim(pPlaisier_FDR_BY)<-NULL
  pPlaisier_FDR_BY<-p.adjust(pPlaisier_FDR_BY,method="BY")
  dim(pPlaisier_FDR_BY)<-dim(pPlaisier)
  ij<-which(pPlaisier_FDR_BY==min(pPlaisier_FDR_BY),arr.ind=TRUE)   k_ij<-k[ij[1],ij[2]]
  if(pPlaisier_FDR_BY[ij[1],ij[2]] > 0.05){
  k_ij<-0
  ij<-c(0,0)
  }

  return(c(ij[1],ij[2],k_ij))
```

```
}
Plaisier_Simulations("m_400_n_400_list1_noise_0.txt","m_400_n_400_list2_noise_0.txt",50,"
Plaisier_Noise_0",400)
Plaisier_Simulations("m_400_n_400_list1_noise_300.txt","m_400_n_400_list2_noise_300.txt",
50,"Plaisier_Noise_300",400)
```