Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets

Q. Xiong[1], N. Ancona[5], Elizabeth R. Hauser[2], Sayan Mukherjee[1,3,4], Terrence S. Furey[1,3]

[1]Institute for Genome Sciences & Policy, [2]Section of Medical Genetics, Department of Medicine, Center for Human Genetics, [3]Department of Computer Science, [4]Departments of Statistical Science and Mathematics, Duke University
[5]Institute of Intelligent Systems for Automation National Research Council Bari, IT.

September 3, 2010

Objective

Dissect genetic and molecular mechanism underlying complex (disease) traits.

Objective

Dissect genetic and molecular mechanism underlying complex (disease) traits.
Standard approaches:

(1) Genome wide association studies (GWAS): Correlations between genetic variants and trait variation.

Objective

Dissect genetic and molecular mechanism underlying complex (disease) traits.
Standard approaches:

(1) Genome wide association studies (GWAS): Correlations between genetic variants and trait variation.

(2) Gene expression studies: correlations between gene expression and trait variation.

Objective

Dissect genetic and molecular mechanism underlying complex (disease) traits.
Standard approaches:

(1) Genome wide association studies (GWAS): Correlations between genetic variants and trait variation.

(2) Gene expression studies: correlations between gene expression and trait variation.

Integration of both approaches for complementary evidence.

Genome wide association studies

(1) Find single variants, independently contributing to disease.

Genome wide association studies

(1) Find single variants, independently contributing to disease.

(2) Issues with population structure, control for LD, etc...

Genome wide association studies

(1) Find single variants, independently contributing to disease.

(2) Issues with population structure, control for LD, etc...

(3) Genetic variations have been identified for a wide variety of common complex diseases (GWAS catalog).

Genome wide association studies

(1) Find single variants, independently contributing to disease.

(2) Issues with population structure, control for LD, etc...

(3) Genetic variations have been identified for a wide variety of common complex diseases (GWAS catalog).

(4) Missing heritability: genetic variation explains 5% of hight variation.

(5) Very weak predictive power.

Expression based studies

(1) Signatures or gene lists predictive of disease.

Expression based studies

(1) Signatures or gene lists predictive of disease.

(2) Sensitive to many environmental factors.

Expression based studies

(1) Signatures or gene lists predictive of disease.

(2) Sensitive to many environmental factors.

(3) Is a complex trait itself.

Expression based studies

(1) Signatures or gene lists predictive of disease.
(2) Sensitive to many environmental factors.
(3) Is a complex trait itself.
(4) Causal versus reactive.

Expression based studies

(1) Signatures or gene lists predictive of disease.
(2) Sensitive to many environmental factors.
(3) Is a complex trait itself.
(4) Causal versus reactive.
(5) Can we find evidence that expression variation predictive of trait variation is genetic.

Expression quantitative trait loci eQTL

Given expression data and genetic variation data on a set of individuals:

Expression quantitative trait loci eQTL

Given expression data and genetic variation data on a set of individuals: eQTLs or eQTNs are SNPs or loci that association with gene expression.

(1) SNPs associated with complex traits are enriched in eQTLs.

Expression quantitative trait loci eQTL

Given expression data and genetic variation data on a set of individuals: eQTLs or eQTNs are SNPs or loci that association with gene expression.

(1) SNPs associated with complex traits are enriched in eQTLs.

(2) This association is robust across eQTL thresholds.

Expression quantitative trait loci eQTL

Given expression data and genetic variation data on a set of individuals: eQTLs or eQTNs are SNPs or loci that association with gene expression.

(1) SNPs associated with complex traits are enriched in eQTLs.

(2) This association is robust across eQTL thresholds.

(3) Can help with causal versus reactive.

Expression quantitative trait loci eQTL

Given expression data and genetic variation data on a set of individuals: eQTLs or eQTNs are SNPs or loci that association with gene expression.

(1) SNPs associated with complex traits are enriched in eQTLs.

(2) This association is robust across eQTL thresholds.

(3) Can help with causal versus reactive.

(4) Need expression data and SNP data from same individuals.

Expression quantitative trait loci eQTL

Given expression data and genetic variation data on a set of individuals: eQTLs or eQTNs are SNPs or loci that association with gene expression.

(1) SNPs associated with complex traits are enriched in eQTLs.
(2) This association is robust across eQTL thresholds.
(3) Can help with causal versus reactive.
(4) Need expression data and SNP data from same individuals.
(5) Missing heritability still a problem.

Pathway based analysis

For common diseases that are polygenic the "missing heritability" may come from the sum of many alleles with small effect.

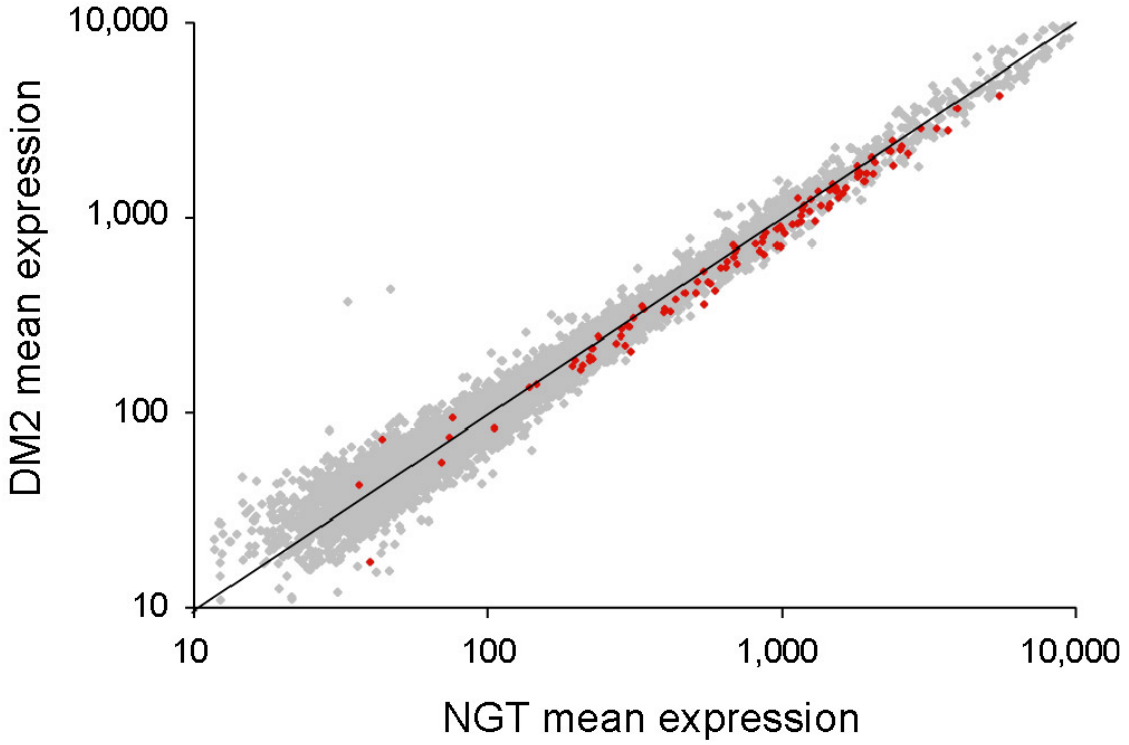
Pathway based analysis

For common diseases that are polygenic the "missing heritability" may come from the sum of many alleles with small effect.

A possible model is of pathway disruption causing complex disease.

Pathway based analysis

For common diseases that are polygenic the "missing heritability" may come from the sum of many alleles with small effect.
A possible model is of pathway disruption causing complex disease.



Mootha et al 2005

Advantages of pathway based methods

(1) Cellular processes involving the interaction of multiple genetic components are better modeled using pathway-based approaches.

Advantages of pathway based methods

(1) Cellular processes involving the interaction of multiple genetic components are better modeled using pathway-based approaches.

(2) Capture joint effect of multiple loci and better capture small changes across many loci.

Advantages of pathway based methods

(1) Cellular processes involving the interaction of multiple genetic components are better modeled using pathway-based approaches.

(2) Capture joint effect of multiple loci and better capture small changes across many loci.

(3) False positives can be reduced.

Advantages of pathway based methods

(1) Cellular processes involving the interaction of multiple genetic components are better modeled using pathway-based approaches.

(2) Capture joint effect of multiple loci and better capture small changes across many loci.

(3) False positives can be reduced.

(4) Facilitates interpretation of the results from association studies.

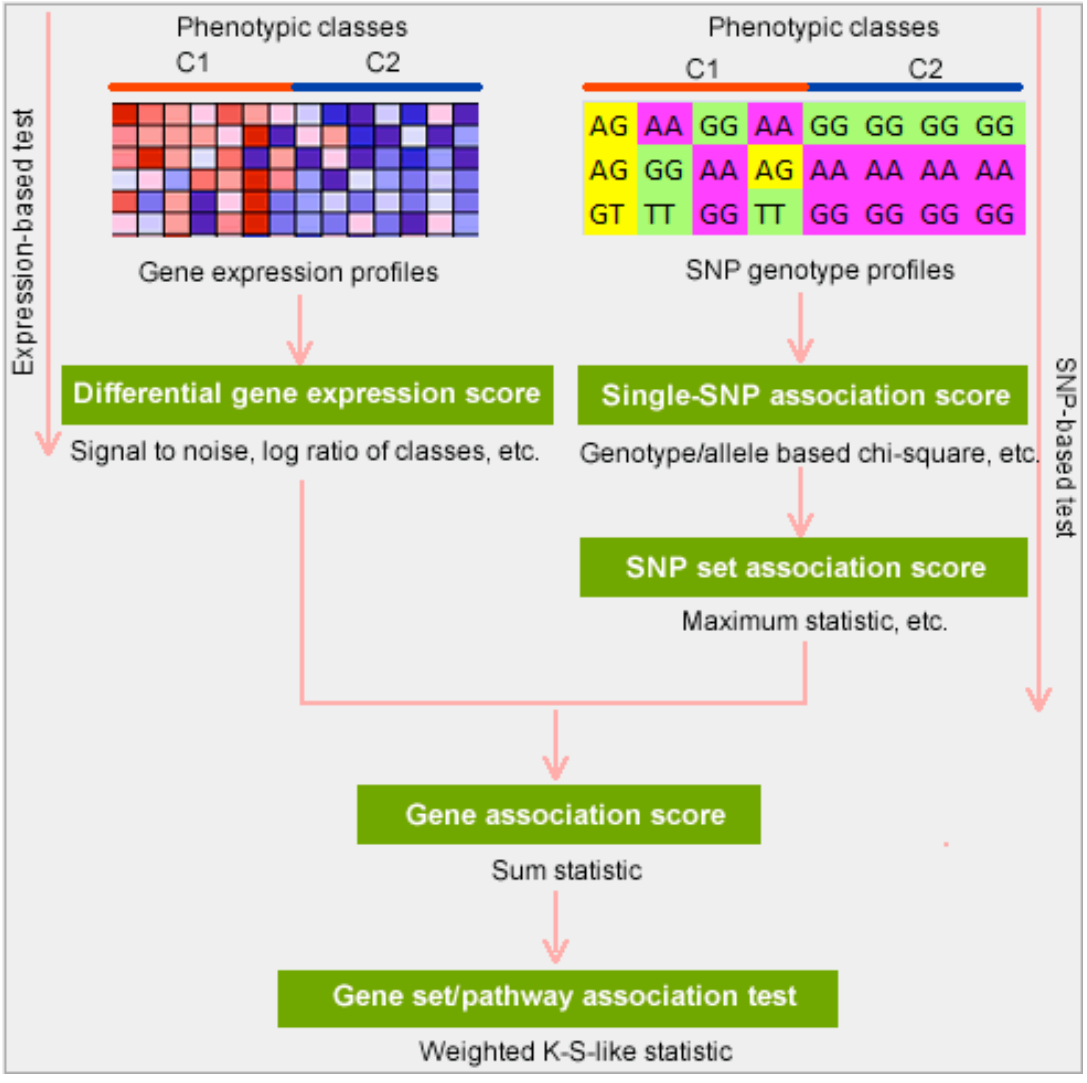
Gene Set Association Analysis

The joint analysis of gene expression and SNP genotype data using a pathway-based strategy for more robust and comprehensive inference of associations.

Gene Set Association Analysis

The joint analysis of gene expression and SNP genotype data using a pathway-based strategy for more robust and comprehensive inference of associations.

Gene Set Association Analysis



Inputs

Gene expression data $X_1, ..., X_n$ and corresponding (Categorical) phenotypic labels: $Y_1, ..., Y_n$, with $X_i \in \mathbb{R}^N$.

Inputs

Gene expression data $X_1, ..., X_n$ and corresponding (Categorical) phenotypic labels: $Y_1, ..., Y_n$, with $X_i \in \mathbb{R}^N$.
SNP data $S_1, ..., S_m$ and corresponding (Categorical) phenotypic labels: $Y_1, ..., Y_m$, with S_i a V-dimension categorical vector.

Inputs

Gene expression data $X_1, ..., X_n$ and corresponding (Categorical) phenotypic labels: $Y_1, ..., Y_n$, with $X_i \in \mathbb{R}^N$.
SNP data $S_1, ..., S_m$ and corresponding (Categorical) phenotypic labels: $Y_1, ..., Y_m$, with S_i a V-dimension categorical vector.
Collections of a priori defined gene sets.

Evidence of differential expression

For each gene, $1, \ldots, N$, compute a differential expression score

$$r_i = \frac{\hat{\mu}_0 - \hat{\mu}_1}{\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2}}.$$

Evidence of differential expression

For each gene, $1, \ldots, N$, compute a differential expression score

$$r_i = \frac{\hat{\mu}_0 - \hat{\mu}_1}{\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2}}.$$

Use any correlation statistic: t-statistic, shrinkage models, effect size estimates....

Evidence of differential expression

For each gene, $1, ..., N$, compute a differential expression score

$$r_i = \frac{\hat{\mu}_0 - \hat{\mu}_1}{\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2}}.$$

Use any correlation statistic: t-statistic, shrinkage models, effect size estimates....
Result: $\{r_1, ..., r_N\}$.

What to do if only given p-values

For each gene, $1, ..., N$, you are given $\{p_1, ..., p_N\}$. How do you calibrate p-value to provide evidence ?

What to do if only given p-values

For each gene, $1, ..., N$, you are given $\{p_1, ..., p_N\}$. How do you calibrate p-value to provide evidence ?
Odds ratio

$$e_j = \frac{f_j(s_j \mid M_1)}{f_j(s_j \mid M_2)}.$$

What to do if only given p-values

For each gene, $1, ..., N$, you are given $\{p_1, ..., p_N\}$. How do you calibrate p-value to provide evidence ?
Odds ratio

$$e_j = \frac{f_j(s_j \mid M_1)}{f_j(s_j \mid M_2)}.$$

P-value fallacy p-value of .001 $\not\Rightarrow$ $1/.0001 = 1,000$ more evidence.

What to do if only given p-values

For each gene, $1, ..., N$, you are given $\{p_1, ..., p_N\}$. How do you calibrate p-value to provide evidence ?
Odds ratio

$$e_j = \frac{f_j(s_j \mid M_1)}{f_j(s_j \mid M_2)}.$$

P-value fallacy p-value of .001 $\not\Rightarrow$ $1/.0001 = 1,000$ more evidence.
P-value calibration

$$B(p_j) = \begin{cases} = \frac{1}{-ep_j \log(p_j)} & \text{if } p_j \in (0, \frac{1}{e}] \\ 1 & \text{otherwise.} \end{cases}$$

Single SNP association score

Use your favorite single-SNP association score

(1) genotype-based chi-square statistic

Single SNP association score

Use your favorite single-SNP association score

(1) genotype-based chi-square statistic

(2) allele-based chi-square statistic

Single SNP association score

Use your favorite single-SNP association score

(1) genotype-based chi-square statistic

(2) allele-based chi-square statistic

(3) frequency differences in major/minor alleles.

Single SNP association score

Use your favorite single-SNP association score

(1) genotype-based chi-square statistic

(2) allele-based chi-square statistic

(3) frequency differences in major/minor alleles.

Simulations suggest genotype-based chi-square test (more power).

SNP Set association score

SNPs need to be assigned to a gene and then a summary score needs to be computed for the SNPs assigned to the gene.

SNP Set association score

SNPs need to be assigned to a gene and then a summary score needs to be computed for the SNPs assigned to the gene.
Assignment: All SNPs within 1kB upstream and down stream of TSS.

SNP Set association score

SNPs need to be assigned to a gene and then a summary score needs to be computed for the SNPs assigned to the gene.
Assignment: All SNPs within 1kB upstream and down stream of TSS.
Summary statistic: Default is maximum. Weighted average. Bayes factors.
Result: $\{s_1, ..., s_N\}$.

SNP Set association score

(1) Max score – the region harbors only one risk variant; more effectively eliminate the negative effects of correlation structure between SNPs. one causal variant but markers with strong LD.

SNP Set association score

(1) Max score – the region harbors only one risk variant; more effectively eliminate the negative effects of correlation structure between SNPs. one causal variant but markers with strong LD.

(2) Weighted mean score – multiple independent risk variants.

SNP Set association score

(1) Max score – the region harbors only one risk variant; more effectively eliminate the negative effects of correlation structure between SNPs. one causal variant but markers with strong LD.

(2) Weighted mean score – multiple independent risk variants.

Gene association score

Given: $\{r_1, ..., r_N\}$ and $\{s_1, ..., s_N\}$.

(1) impose directionality on SNP evidence: $s_i \equiv s_i \times \text{sign}(r_i)$.

Gene association score

Given: $\{r_1, ..., r_N\}$ and $\{s_1, ..., s_N\}$.

(1) impose directionality on SNP evidence: $s_i \equiv s_i \times \text{sign}(r_i)$.

(2) normalize

$$\tilde{r}_i = \frac{r_i}{\sum_{j=1}^{N} |r_j| \times \mathbf{I}[\text{sign}(r_i) = \text{sign}(r_j)]}$$

$$\tilde{s}_i = \frac{s_i}{\sum_{j=1}^{N} |r_j| \times \mathbf{I}[\text{sign}(s_i) = \text{sign}(s_j)]}.$$

Gene association score

Given: $\{r_1, ..., r_N\}$ and $\{s_1, ..., s_N\}$.

(1) impose directionality on SNP evidence: $s_i \equiv s_i \times \text{sign}(r_i)$.

(2) normalize

$$\tilde{r}_i = \frac{r_i}{\sum_{j=1}^{N} |r_j| \times \mathbf{I}[\text{sign}(r_i) = \text{sign}(r_j)]}$$

$$\tilde{s}_i = \frac{s_i}{\sum_{j=1}^{N} |r_j| \times \mathbf{I}[\text{sign}(s_i) = \text{sign}(s_j)]}.$$

(3) combine evidence $c_i = \tilde{s}_i + \tilde{r}_i$.

eQTL setting (in progresss)

The previous gene association score can be thought of as

$$c_j = e(Y \mid S_j) + e(Y \mid X_j).$$

eQTL setting (in progresss)

The previous gene association score can be thought of as

$$c_j = e(Y \mid S_j) + e(Y \mid X_j).$$

If we have expression and genetic data on same individuals we can adjust evidence

$$c_j = e(Y \mid S_j) + \mathrm{P}(Y \mid X_j) \times e(X_j \mid S_j).$$

eQTL setting (in progresss)

The previous gene association score can be thought of as

$$c_j = e(Y \mid S_j) + e(Y \mid X_j).$$

If we have expression and genetic data on same individuals we can adjust evidence

$$c_j = e(Y \mid S_j) + \mathrm{P}(Y \mid X_j) \times e(X_j \mid S_j).$$

This idea can be used for other genomic features.

Gene set association score

For a gene set S with h genes a the rank ordered association scores $\{c_{(1)}, ..., c_{(N)}\}$ compute running association score

$$\mathrm{RAS}_S(i) = \frac{1}{N_s} \sum_{j=1}^{i} |c_j| \times \mathbf{I}(j \in S) - \frac{1}{N-h} \mathbf{I}(j \notin S),$$

with $N_S = \sum_{j=1}^{N} |c_j| \times \mathbf{I}(j \in S)$.

Gene set association score

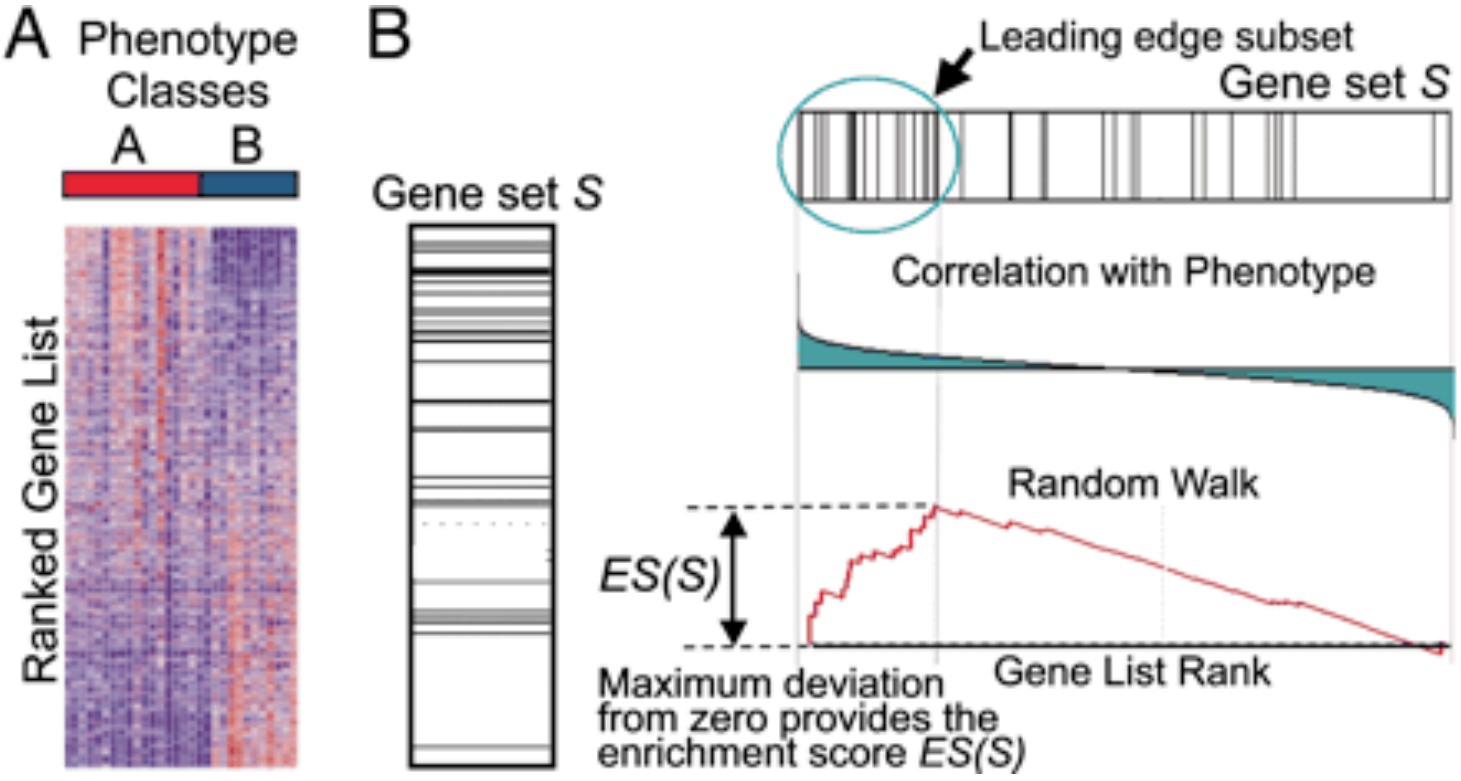
For a gene set S with h genes a the rank ordered association scores $\{c_{(1)}, ..., c_{(N)}\}$ compute running association score

$$\mathrm{RAS}_S(i) = \frac{1}{N_s} \sum_{j=1}^{i} |c_j| \times \mathbf{I}(j \in S) - \frac{1}{N-h} \mathbf{I}(j \notin S),$$

with $N_S = \sum_{j=1}^{N} |c_j| \times \mathbf{I}(j \in S)$.

The association score $\mathrm{AS}(S)$ is the maximum deviation of $\{\mathrm{RAS}_S(i)\}$ from zero.

Gene set association score



Statistical significance and adjustment for multiple hypothesis testing

Use permutation procedure and FDR corrections. This automatically corrects for linkage structure and population.

Simulation studies

We compared four methods

(1) GSAA – SNP and expression data

Simulation studies

We compared four methods

(1) GSAA – SNP and expression data

(2) GSEA – only expression data

Simulation studies

We compared four methods

(1) GSAA – SNP and expression data

(2) GSEA – only expression data

(3) GSEA-SNP – only SNP data

Simulation studies

We compared four methods

(1) GSAA – SNP and expression data

(2) GSEA – only expression data

(3) GSEA-SNP – only SNP data

(4) two step regression model – two step regression model, filter genes with un/weakly associated SNPs, regression on remaining SNPs.

Simulated data

We generated simulated expression data and SNP data

(1) SNP data was generated using SIMLA. The parameters correspond to marker and disease placement, locus heterogeneity, disequilibrium between markers and between markers and disease loci.

Simulated data

We generated simulated expression data and SNP data

(1) SNP data was generated using SIMLA. The parameters correspond to marker and disease placement, locus heterogeneity, disequilibrium between markers and between markers and disease loci.

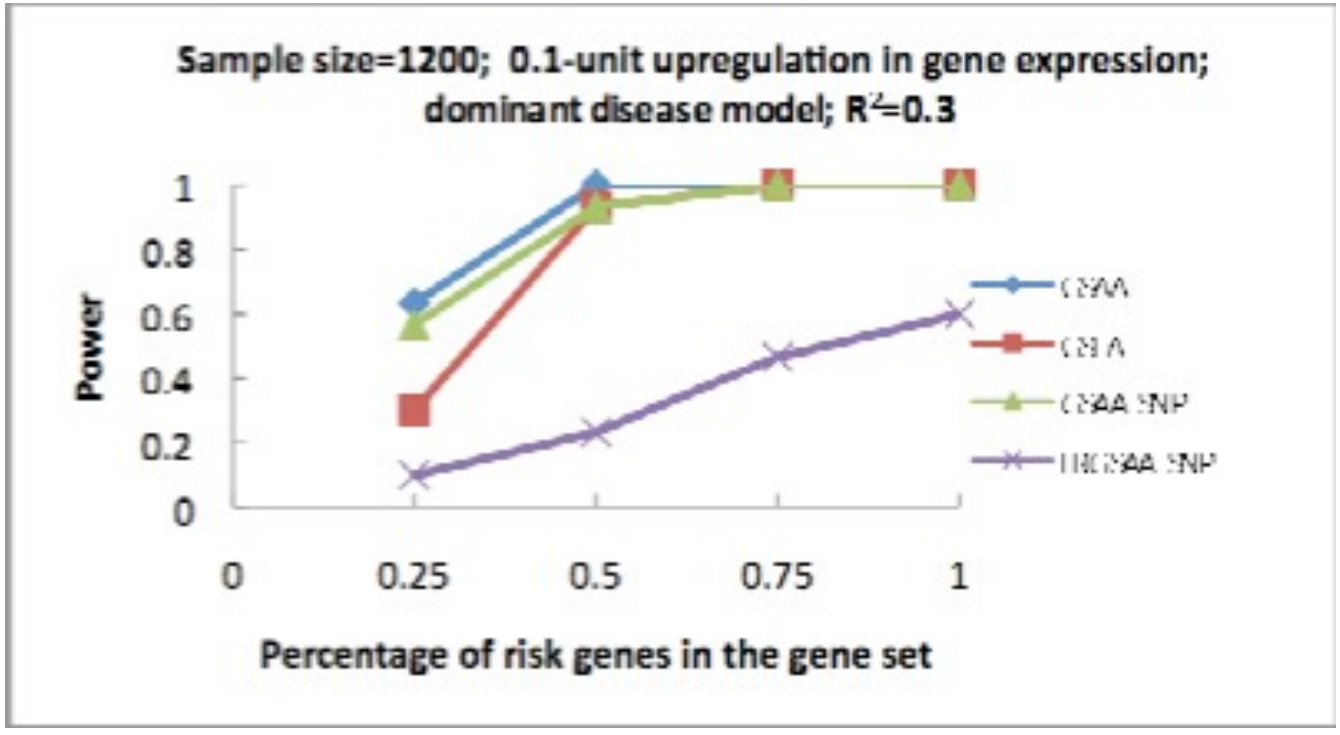
(2) Expression data was simulated using normals.

Simulated data

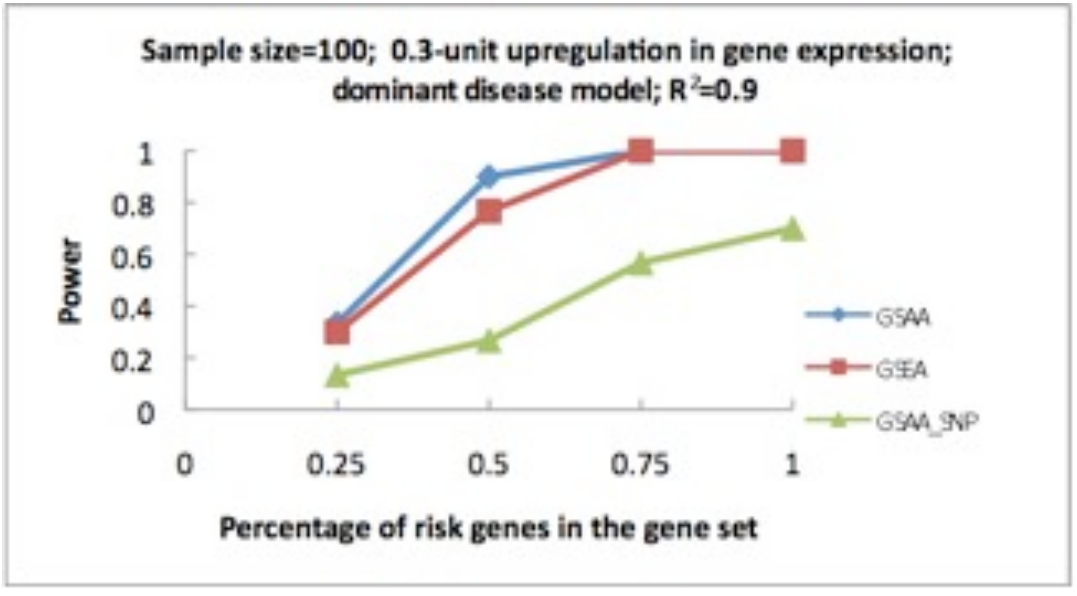
(1) 1000 genes, first 20 causal, 3 SNPs in each causal gene with the second marker is in LD with the disease variant with varying R^2.

(2) 100 gene sets, the first set is causal including 5-20% of causal genes.

Power calculations



Power calculations



The Cancer Genome Atlas

An excellent source for integrated genomic data for various tumors, currently
glioblastoma multiforme, ovarian (serous cystadenocarcinoma) and lung (squamous carcinoma).

The Cancer Genome Atlas

An excellent source for integrated genomic data for various tumors, currently
glioblastoma multiforme, ovarian (serous cystadenocarcinoma) and lung (squamous carcinoma).

Collection of clinical, expression, SNP, copy number, and high-throughput sequencing data.

Glioblastoma data

Expression data: 258 tumor samples and 11 normal samples
SNP data: 205 tumor samples and 89 normal samples

Glioblastoma data

Expression data: 258 tumor samples and 11 normal samples
SNP data: 205 tumor samples and 89 normal samples
357 "canonical pathways" from MSigDB and 658 GO gene sets.

Pathways associated

Table 2. Most significant canonical pathways associated with tumor samples with FDR≤0.25

Gene Set Name	P-value	FDR
[a,b] P53PATHWAY	0.011	0.0955
[a,b] RELAPATHWAY	0.0146	0.1482
[a] ATRBRCAPATHWAY	0.0846	0.1574
[a] HSA03030_DNA_POLYMERASE	0.1319	0.1764
[a] G1PATHWAY	0.045	0.1853
[b] CASPASEPATHWAY	0.0184	0.1949
[a,b] HSA04115_P53_SIGNALING_PATHWAY	0.0034	0.1961
[a] CELL_CYCLE_KEGG	0.0669	0.2034
[a] DNA_REPLICATION_REACTOME	0.2094	0.2035
[a] G2PATHWAY	0.0432	0.2151
INTRINSICPATHWAY	0.0476	0.2311
[a,b] ATMPATHWAY	0.1239	0.2371
STATIN_PATHWAY_PHARMGKB	0.0256	0.2403

[a] Canonical pathways related to the cell cycle, proliferation, cell cycle transitions, or checkpoints.
[b] Canonical pathways related to or contain genes involved in the induction of apoptosis.
For full results, see Table S15.

Expression and association signatures

For P53PATHWAY:

(1) 5 genes TP53, RB1, E2F1, ATM, and MDM2 show evidence in our single-SNP analysis.

Expression and association signatures

For P53PATHWAY:

(1) 5 genes TP53, RB1, E2F1, ATM, and MDM2 show evidence in our single-SNP analysis.

(2) 6 genes TP53, RB1, CDK2, CDK4, PCNA, p21 show evidence in our expression analysis.

Expression and association signatures

For P53PATHWAY:

(1) 5 genes TP53, RB1, E2F1, ATM, and MDM2 show evidence in our single-SNP analysis.

(2) 6 genes TP53, RB1, CDK2, CDK4, PCNA, p21 show evidence in our expression analysis.

Top ranked genes with respect to combined expression and SNP association:

(1) ADAM12 – evidence of transcriptional regulation.

Expression and association signatures

For P53PATHWAY:

(1) 5 genes TP53, RB1, E2F1, ATM, and MDM2 show evidence in our single-SNP analysis.

(2) 6 genes TP53, RB1, CDK2, CDK4, PCNA, p21 show evidence in our expression analysis.

Top ranked genes with respect to combined expression and SNP association:

(1) ADAM12 – evidence of transcriptional regulation.

(2) CDKN2A – locus associated in recent GWA study.

WTCC data

Expression data: 7 cases 16 controls
SNP data: 1748 cases samples and 2938 controls

WTCC data

Expression data: 7 cases 16 controls
SNP data: 1748 cases samples and 2938 controls
357 "canonical pathways" from MSigDB and 658 GO gene sets.

Pathways associated

Table 3. Most significant canonical pathways and GO gene sets associated with case samples with FDR≤0.25

Gene Set Name	P-value	FDR
Canonical Pathways		
PROTEASOME	0.0121	0.0539
[b] SA_MMP_CYTOKINE_CONNECTION	0.0059	0.0652
CHOLESTEROL_BIOSYNTHESIS	0.1532	0.0823
AMINOACYL_TRNA_BIOSYNTHESIS	0.0794	0.0882
PROTEASOMEPATHWAY	0.0133	0.0925
[b] LAIRPATHWAY	0.017	0.1035
[b] STEMPATHWAY	0.0113	0.111
HSA00970_AMINOACYL_TRNA_BIOSYNTHESIS	0.1063	0.1176
[b] HYPERTROPHY_MODEL	0.0171	0.1193
[b] IL6PATHWAY	0.0122	0.1196
[a] ATMPATHWAY	0.0006	0.126
[a,b] TNFR2PATHWAY	0.0075	0.139
[b] ERYTHPATHWAY	0.0551	0.1395
HSA03050_PROTEASOME	0.0102	0.1464
[a,b] NTHIPATHWAY	0.0147	0.2121
[b] NO2IL12PATHWAY	0.0541	0.216
[a,b] TIDPATHWAY	0.0177	0.2235
HSA00530_AMINOSUGARS_METABOLISM	0.0181	0.2269
[a,b] RELAPATHWAY	0.0251	0.2429
GO Gene Sets		
[b] CHEMOKINE_RECEPTOR_BINDING	0.0022	0.0138
[b] G_PROTEIN_COUPLED_RECEPTOR_BINDING	0.003	0.0261
[b] CHEMOKINE_ACTIVITY	0.0022	0.0276
PROTEIN_DOMAIN_SPECIFIC_BINDING	$<10^{-15}$	0.0543
INDUCTION_OF_APOPTOSIS_BY_INTRACELLULAR_SIGNALS	$<10^{-15}$	0.1114
[b] VIRAL_GENOME_REPLICATION	0.002	0.1253

[a] Canonical pathways including NF-κB signaling.
[b] Canonical pathways and GO gene sets related to cytokines, chemokines, and immune response.
For full results, see Table S28, S30.

Extensions

(1) Sequencing data

Extensions

(1) Sequencing data

(2) Other genomic sources: methylation, protemics ?,...

Extensions

(1) Sequencing data
(2) Other genomic sources: methylation, protemics ?,...
(3) Better associations of SNPs to genes

Extensions

(1) Sequencing data

(2) Other genomic sources: methylation, protemics ?,...

(3) Better associations of SNPs to genes

(4) Variation for cases where expression and SNP appears on same individuals

Extensions

(1) Sequencing data

(2) Other genomic sources: methylation, protemics ?,...

(3) Better associations of SNPs to genes

(4) Variation for cases where expression and SNP appears on same individuals

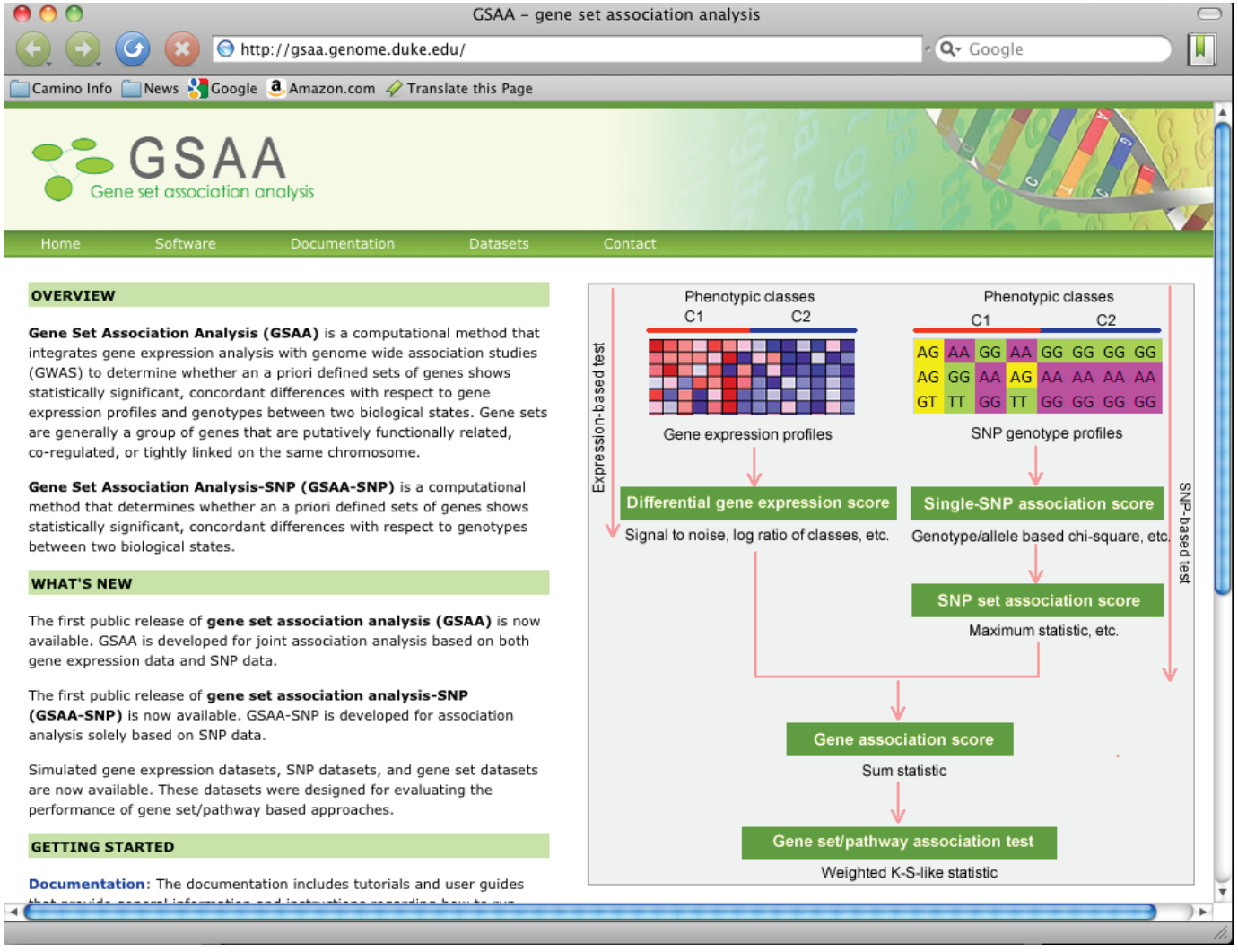
(5) A proper model.

Software

GSAA software.

Software

GSAA software.



Acknowledgements

Funding:

- Center for Systems Biology at Duke
- NSF
- NIH