Studying co-regulation and inter-regulation of genes via eQTL mapping

Tian Zheng

Department of Statistics
Columbia University

October 14, 2007
Calling from high-throughput biology

- The mRNA abundance (or gene expressions) can be simultaneously measured on thousands of genes in one microarray experiment—variation at the molecular level.
- Tens to hundreds of thousands of SNPs (single nucleotide polymorphisms) are available for genotyping to study the genetic variation of individuals—variation at the DNA sequence level.
- The challenge: identify and understand genetic variation contributing to complex phenotypes (e.g., diseases).
- Identifying gene-gene interactions and understanding how genes interact are of crucial importance.
- New annotations are being added to the biology literature on genetic pathways and regulatory networks.
Calling from high-throughput biology

- The mRNA abundance (or gene expressions) can be simultaneously measured on thousands of genes in one microarray experiment—variation at the molecular level.
- Tens to hundreds of thousands of SNPs (single nucleotide polymorphisms) are available for genotyping to study the genetic variation of individuals—variation at the DNA sequence level.
- The challenge: identify and understand genetic variation contributing to complex phenotypes (e.g., diseases).
- Identifying gene-gene interactions and understanding how genes interact are of crucial importance.
- New annotations are being added to the biology literature on genetic pathways and regulatory networks.
Calling from high-throughput biology

- The mRNA abundance (or gene expressions) can be simultaneously measured on thousands of genes in one microarray experiment—\textit{variation at the molecular level}.

- Tens to hundreds of thousands of SNPs (single nucleotide polymorphisms) are available for genotyping to study the genetic variation of individuals—\textit{variation at the DNA sequence level}.

- \textit{The challenge}: identify and understand genetic variation contributing to complex phenotypes (e.g., diseases).

- Identifying gene-gene interactions and understanding how genes interact are of crucial importance.

- New annotations are being added to the biology literature on genetic pathways and regulatory networks.
Calling from high-throughput biology

- The mRNA abundance (or gene expressions) can be simultaneously measured on thousands of genes in one microarray experiment—*variation at the molecular level*.

- Tens to hundreds of thousands of SNPs (single nucleotide polymorphisms) are available for genotyping to study the genetic variation of individuals—*variation at the DNA sequence level*.

- *The challenge*: identify and understand genetic variation contributing to complex phenotypes (e.g., diseases).

- Identifying gene-gene interactions and understanding how genes interact are of crucial importance.

- New annotations are being added to the biology literature on genetic pathways and regulatory networks.
The mRNA abundance (or gene expressions) can be simultaneously measured on thousands of genes in one microarray experiment—variation at the molecular level.

Tens to hundreds of thousands of SNPs (single nucleotide polymorphisms) are available for genotyping to study the genetic variation of individuals—variation at the DNA sequence level.

The challenge: identify and understand genetic variation contributing to complex phenotypes (e.g., diseases).

Identifying gene-gene interactions and understanding how genes interact are of crucial importance.

New annotations are being added to the biology literature on genetic pathways and regulatory networks.
Calling from high-throughput biology

- Improving methods and model performance versus understanding biology in bioinformatics research.
- In Genetic genomics, the analysis of data is heavily motivated by the biological hypotheses to be tested.
- Multiple sources of data can be combined in one analysis to identify interesting patterns for further biological investigation.
Calling from high-throughput biology

- Improving methods and model performance versus understanding biology in bioinformatics research.
- In Genetic genomics, the analysis of data is heavily motivated by the biological hypotheses to be tested.
- Multiple sources of data can be combined in one analysis to identify interesting patterns for further biological investigation.
Calling from high-throughput biology

- Improving methods and model performance versus understanding biology in bioinformatics research.
- In Genetic genomics, the analysis of data is heavily motivated by the biological hypotheses to be tested.
- Multiple sources of data can be combined in one analysis to identify interesting patterns for further biological investigation.
What is eQTL mapping?

- **Quantitative traits**: as opposed to dichotomous traits, have continuous trait values (e.g., plasma cholesterol).
- **Quantitative trait locus**: a chromosomal region that affects the levels of a heritable quantitative trait.
- QTL mapping methods have been developed to identify potential QTL for a specific phenotype trait via either linkage information or association information, or both.
- Gene expression—mRNA abundance of a given gene—is also heritable.
- eQTL (expression Quantitative Trait Locus) is then the regulatory region for an expression trait.
- eQTL mapping is then QTL mapping with an expression trait used as the phenotype.
What is eQTL mapping?

- **Quantitative traits**: as opposed to dichotomous traits, have continuous trait values (e.g., plasma cholesterol).

- **Quantitative trait locus**: a chromosomal region that affects the levels of a heritable quantitative trait.

- QTL mapping methods have been developed to identify potential QTL for a specific phenotype trait via either linkage information or association information, or both.

- Gene expression—mRNA abundance of a given gene—is also heritable.

- eQTL (expression Quantitative Trait Locus) is then the regulatory region for an expression trait.

- eQTL mapping is then QTL mapping with an expression trait used as the phenotype.
What is eQTL mapping?

- **Quantitative traits**: as opposed to dichotomous traits, have continuous trait values (e.g., plasma cholesterol).
- **Quantitative trait locus**: a chromosomal region that affects the levels of a heritable quantitative trait.
- QTL mapping methods have been developed to identify potential QTL for a specific phenotype trait via either linkage information or association information, or both.
- Gene expression—mRNA abundance of a given gene—is also heritable.
- **eQTL** (expression **Quantitative Trait Locus**) is then the regulatory region for an expression trait.
- eQTL mapping is then QTL mapping with an expression trait used as the phenotype.
What is eQTL mapping?

- **Quantitative traits**: as opposed to dichotomous traits, have continuous trait values (e.g., plasma cholesterol).
- **Quantitative trait locus**: a chromosomal region that affects the levels of a heritable quantitative trait.
- QTL mapping methods have been developed to identify potential QTL for a specific phenotype trait via either linkage information or association information, or both.
- Gene expression—mRNA abundance of a given gene—is also heritable.
- eQTL (expression Quantitative Trait Locus) is then the regulatory region for an expression trait.
- eQTL mapping is then QTL mapping with an expression trait used as the phenotype.
What is eQTL mapping?

- **Quantitative traits**: as opposed to dichotomous traits, have continuous trait values (e.g., plasma cholesterol).
- **Quantitative trait locus**: a chromosomal region that affects the levels of a heritable quantitative trait.
- QTL mapping methods have been developed to identify potential QTL for a specific phenotype trait via either linkage information or association information, or both.
- Gene expression—mRNA abundance of a given gene—is also heritable.
- eQTL (**expression Quantitative Trait Locus**) is then the regulatory region for an expression trait.
- eQTL mapping is then QTL mapping with an expression trait used as the phenotype.
What is eQTL mapping?

- **Quantitative traits**: as opposed to dichotomous traits, have continuous trait values (e.g., plasma cholesterol).
- **Quantitative trait locus**: a chromosomal region that affects the levels of a heritable quantitative trait.
- QTL mapping methods have been developed to identify potential QTL for a specific phenotype trait via either linkage information or association information, or both.
- Gene expression—mRNA abundance of a given gene—is also heritable.
- eQTL (expression **Quantitative Trait Locus**) is then the regulatory region for an expression trait.
- eQTL mapping is then QTL mapping with an expression trait used as the phenotype.
eQTL mapping

- Referred as “genetical genomics”
  - also as “expression genetics”
eQTL mapping

- Referred as “genetical genomics”
- also as “expression genetics”
eQTL mapping

- Referred as “genetical genomics”
- also as “expression genetics”
eQTL mapping

- Referred as “genetical genomics”
- also as “expression genetics”
QTL mapping

Searching for QTL’s in a genome scan:

- Markers
- Observed Trait
- Association due to linkage
- QTL
- Genomic Location
- Linkage (short biological distance)
Studying gene regulations using eQTL mapping

A hypothetical example of multiple genetic factors (located in physically distinct regions) may affect the mRNA abundance of the “black” gene.

Mutations within the promoter, upstream, intronic region, “blue” TF, “red” enhancer, or the “purple” RNA-binding protein can all affect the mRNA abundance of “black.”

Studying gene regulations using eQTL mapping

▶ A hypothetical example of multiple genetic factors (located in physically distinct regions) may affect the mRNA abundance of the “black” gene.

![Diagram showing gene regulation](image)


▶ Mutations within the promoter, upstream, intronic region, “blue” TF, “red” enhancer, or the “purple” RNA-binding protein can all affect the mRNA abundance of “black.”
Gene co-expression and co-regulation

- In traditional gene expression analysis, genes that demonstrate strongly positive correlated expression levels are clustered together, which sometimes are termed as *co-expressed genes*.

- The clustering based on co-expression patterns are then compared to known gene-gene regulation.

- Co-expression can be result of two genes being regulated by the same regulator.

- OR, co-expression can ALSO be result of that gene A regulates gene B.
Gene co-expression and co-regulation

- In traditional gene expression analysis, genes that demonstrate strongly positive correlated expression levels are clustered together, which sometimes are termed as *co-expressed genes*.

- The clustering based on co-expression patterns are then compared to known gene-gene regulation.

- Co-expression can be result of two genes being regulated by the same regulator.

- OR, co-expression can ALSO be result of that gene A regulates gene B.
In traditional gene expression analysis, genes that demonstrate strongly positive correlated expression levels are clustered together, which sometimes are termed as *co-expressed genes*.

The clustering based on co-expression patterns are then compared to known gene-gene regulation.

Co-expression can be result of two genes being regulated by the same regulator.

OR, co-expression can ALSO be result of that gene A regulates gene B.
Gene co-expression and co-regulation

- In traditional gene expression analysis, genes that demonstrate strongly positive correlated expression levels are clustered together, which sometimes are termed as *co-expressed genes*.

- The clustering based on co-expression patterns are then compared to known gene-gene regulation.

- Co-expression can be result of two genes being regulated by the same regulator.

- **OR**, co-expression can ALSO be result of that gene A regulates gene B.
Regulatory activities for a set of genes

- For a set of genes (in a pathway or related to a disorder), we can study their genome regulators.
  - When studied as a set, their common activities and "connections" are of more interests.
  - **Co-regulation**: two genes are regulated with the same regulator.
  - **Inter-regulation**: one gene regulates the other gene (within a gene set of interests).
Regulatory activities for a set of genes

- For a set of genes (in a pathway or related to a disorder), we can study their genome regulators.

- When studied as a set, their common activities and "connections" are of more interests.

- **Co-regulation**: two genes are regulated with the same regulator.

- **Inter-regulation**: one gene regulates the other gene (within a gene set of interests).
Regulatory activities for a set of genes

- For a set of genes (in a pathway or related to a disorder), we can study their genome regulators.
- When studied as a set, their common activities and “connections” are of more interests.
- **Co-regulation:** two genes are regulated with the same regulator.
- **Inter-regulation:** one gene regulates the other gene (within a gene set of interests).
Regulatory activities for a set of genes

- For a set of genes (in a pathway or related to a disorder), we can study their genome regulators.
- When studied as a set, their common activities and "connections" are of more interests.
- **Co-regulation**: two genes are regulated with the same regulator.
- **Inter-regulation**: one gene regulates the other gene (within a gene set of interests).
Part I: Study gene co-regulation.
Evidence of gene co-regulation in eQTL mapping

- Identified eQTL do not distribute completely random on the genome.
- Some locus seems to have more regulatory “duties” than others—(transcription) hotspots.
- Among possible explanations, a regulator may have pleiotropic consequences on multiple downstream targets.
- Morley et al. (2004): “the shared expression control regions suggest co-regulation.”
Evidence of gene co-regulation in eQTL mapping

- Identified eQTL do not distribute completely random on the genome.

- Some locus seems to have more regulatory “duties” than others—(transcription) hotspots.

- Among possible explanations, a regulator may have pleiotropic consequences on multiple downstream targets.

- Morley et al. (2004): “the shared expression control regions suggest co-regulation.”
Evidence of gene co-regulation in eQTL mapping

- Identified eQTL do not distribute completely random on the genome.
- Some locus seems to have more regulatory “duties” than others—(transcription) *hotspots*.
- Among possible explanations, a regulator may have pleiotropic consequences on multiple downstream targets.
- Morley et al. (2004): “the shared expression control regions suggest co-regulation.”
Evidence of gene co-regulation in eQTL mapping

- Identified eQTL do not distribute completely random on the genome.
- Some locus seems to have more regulatory “duties” than others—(transcription) hotspots.
- Among possible explanations, a regulator may have pleiotropic consequences on multiple downstream targets.
- Morley et al. (2004): “the shared expression control regions suggest co-regulation.”
Transcription hotspots


- Transcription hotspots are defined as “bins” on the genomes with more eQTL mapped than random.
  - 491 bins of 5 Mb each.
  - 142 phenotypes with 318 eQTL hits.
  - The average number of hits per bin is just 0.65.
Transcription hotspots


- Transcription hotspots are defined as “bins” on the genomes with more eQTL mapped than random.
- 491 bins of 5 Mb each.
- 142 phenotypes with 318 eQTL hits.
- The average number of hits per bin is just 0.65.
Transcription hotspots


- Transcription hotspots are defined as “bins” on the genomes with more eQTL mapped than random.
- 491 bins of 5 Mb each.
- 142 phenotypes with 318 eQTL hits.
- The average number of hits per bin is just 0.65.
Transcription hotspots


- Transcription hotspots are defined as “bins” on the genomes with more eQTL mapped than random.
- 491 bins of 5 Mb each.
- 142 phenotypes with 318 eQTL hits.
- The average number of hits per bin is just 0.65.
Transcription hotspots

- eQTL hotspots can be an artifact of gene co-expression?

- Simulation studies have been used to study this speculation and found part of the hotspot phenomenon can be caused by correlated gene expression along but not all.

Transcription hotspots

- eQTL hotspots can be an artifact of gene co-expression?
- Simulation studies have been used to study this speculation and found part of the hotspot phenomenon can be caused by correlated gene expression along but not all.

Another idea of studying “joint” regulation activities is to construct “supergenes” that combines the expression levels of *multiple traits*.

Gene expression transcripts can be grouped according to pathway information, disease relevance, genome locations, etc.

The “supergene” phenotype can be a quantitative characteristic of this group of transcripts, such as sum, principal components.

Regulators of such “supergenes” can also be viewed to have pleiotropic regulatory function.
Regulator of joint patterns of a gene set

- Another idea of studying “joint” regulation activities is to construct “supergenes” that combines the expression levels of multiple traits.

- Gene expression transcripts can be grouped according to pathway information, disease relevance, genome locations, etc.

- The “supergene” phenotype can be a quantitative characteristic of this group of transcripts, such as sum, principal components.

- Regulators of such “supergenes” can also be viewed to have pleiotropic regulatory function.
Regulator of joint patterns of a gene set

- Another idea of studying “joint” regulation activities is to construct “supergenes” that combines the expression levels of *multiple traits*.
- Gene expression transcripts can be grouped according to pathway information, disease relevance, genome locations, etc.
- The “supergene” phenotype can be a quantitative characteristic of this group of transcripts, such as sum, principal components.
- Regulators of such “supergenes” can also be viewed to have pleiotropic regulatory function.
Another idea of studying “joint” regulation activities is to construct “supergenes” that combines the expression levels of *multiple traits*.

Gene expression transcripts can be grouped according to pathway information, disease relevance, genome locations, etc.

The “supergene” phenotype can be a quantitative characteristic of this group of transcripts, such as sum, principal components.

Regulators of such “supergenes” can also be viewed to have pleiotropic regulatory function.
Part I case study: study breast cancer related genes without breast cancer patients
Study a disorder without any patients

- Gene expression have been found to have predictive power of etiological properties of breast cancer.
- Genes that are predictive of breast cancer may play an important role in the disease molecular process.
- Studying the regulation of these genes may provide important insights on the disease, even without using any data from patients.
Study a disorder without any patients

- Gene expression have been found to have predictive power of etiological properties of breast cancer.
- Genes that are predictive of breast cancer may play an important role in the disease molecular process.
- Studying the regulation of these genes may provide important insights on the disease, even without using any data from patients.
Study a disorder without any patients

- Gene expression have been found to have predictive power of etiological properties of breast cancer.
- Genes that are predictive of breast cancer may play an important role in the disease molecular process.
- Studying the regulation of these genes may provide important insights on the disease, even without using any data from patients.
Data

- **Affymetrix Human Focus Arrays**, with 8500 transcripts were measured on 194 individuals in 14 CEPH families (Morley et al., 2004).
- Genotypes of these CEPH individuals on 2819 autosomal SNPs across the genome were obtained from The SNP Consortium (http://snp.cshl.org/linkage_maps/).
- We examined 18 transcripts that are related to several candidate genes of breast cancer, discussed in OMIM.
- We ran both association scan and linkage scan.
- Linkage scan was done on all 194 members of 14 CEPH families using MERLIN.

Data

► Affymetrix Human Focus Arrays, with 8500 transcripts were measured on 194 individuals in 14 CEPH families (Morley et al., 2004).

► Genotypes of these CEPH individuals on 2819 autosomal SNPs across the genome were obtained from The SNP Consortium (http://snp.cshl.org/linkage_maps/).

► We examined 18 transcripts that are related to several candidate genes of breast cancer, discussed in OMIM.

► We ran both association scan and linkage scan.

► Linkage scan was done on all 194 members of 14 CEPH families using MERLIN.

Data

- Affymetrix Human Focus Arrays, with 8500 transcripts were measured on 194 individuals in 14 CEPH families (Morley et al., 2004).
- Genotypes of these CEPH individuals on 2819 autosomal SNPs across the genome were obtained from The SNP Consortium (http://snp.cshl.org/linkage_maps/).
- We examined 18 transcripts that are related to several candidate genes of breast cancer, discussed in OMIM.
- We ran both association scan and linkage scan.
- Linkage scan was done on all 194 members of 14 CEPH families using MERLIN.

Data

- Affymetrix Human Focus Arrays, with 8500 transcripts were measured on 194 individuals in 14 CEPH families (Morley et al., 2004).
- Genotypes of these CEPH individuals on 2819 autosomal SNPs across the genome were obtained from The SNP Consortium (http://snp.cshl.org/linkage_maps/).
- We examined 18 transcripts that are related to several candidate genes of breast cancer, discussed in OMIM.
- We ran both association scan and linkage scan.
  - Linkage scan was done on all 194 members of 14 CEPH families using MERLIN.

Data

- Affymetrix Human Focus Arrays, with 8500 transcripts were measured on 194 individuals in 14 CEPH families (Morley et al., 2004).
- Genotypes of these CEPH individuals on 2819 autosomal SNPs across the genome were obtained from The SNP Consortium (http://snp.cshl.org/linkage_maps/).
- We examined 18 transcripts that are related to several candidate genes of breast cancer, discussed in OMIM.
- We ran both association scan and linkage scan.
- Linkage scan was done on all 194 members of 14 CEPH families using MERLIN.

Multilocus association scan

▶ Multilocus association scores have been proposed for gene mapping to study the association information content of a set of markers, with respect to a dichotomous phenotype. (Lo and Zheng, 2002; Lo and Zheng 2004; Zheng et al. 2006)

▶ We proposed a new statistic, qGTD (quantitative genotype-trait distortion) for the association scan.
Multilocus association scores have been proposed for gene mapping to study the association information content of a set of markers, with respect to a dichotomous phenotype. (Lo and Zheng, 2002; Lo and Zheng 2004; Zheng et al. 2006)

We proposed a new statistic, qGTD (quantitative genotype-trait distortion) for the association scan.
Multilocus association scan

- qGTD is defined on the ranks of observed quantitative phenotype values of \( n \) individuals, i.e., \( \{R_1, \ldots, R_n\} \).
- Given a set of \( k \) markers, there are \( 3^k \) possible multilocus genotypes, denoted by \( \{G_1, \ldots, G_{3^k}\} \).
- For individual \( i \), let \( g_i \) be his/her unphased multilocus genotype on these \( k \) loci.

\[
qGTD = \frac{12}{n^3 - n} \sum_{i=1}^{3^k} \left( S_i - n \frac{n + 1}{2} \right)^2,
\]

- where \( S_i = \sum_{g_j = G_i} R_j \) is the rank sum of individuals with genotype \( G_i \).
Multilocus association scan

- qGTD is defined on the ranks of observed quantitative phenotype values of \( n \) individuals, i.e., \( \{ R_1, \ldots, R_n \} \).
- Given a set of \( k \) markers, there are \( 3^k \) possible multilocus genotypes, denoted by \( \{ G_1, \ldots, G_{3^k} \} \).
- For individual \( i \), let \( g_i \) be his/her unphased multilocus genotype on these \( k \) loci.

\[
qGTD = \frac{12}{n^3 - n} \sum_{i=1}^{3^k} \left( S_i - n_i \frac{n + 1}{2} \right)^2,
\]

where \( S_i = \sum_{g_j = G_i} R_j \) is the rank sum of individuals with genotype \( G_i \).
Multilocus association scan

- qGTD is defined on the ranks of observed quantitative phenotype values of \( n \) individuals, i.e., \( \{R_1, \ldots, R_n\} \).
- Given a set of \( k \) markers, there are \( 3^k \) possible multilocus genotypes, denoted by \( \{G_1, \ldots, G_{3^k}\} \).
- For individual \( i \), let \( g_i \) be his/her unphased multilocus genotype on these \( k \) loci.

\[
q\text{GTD} = \frac{12}{n^3 - n} \sum_{i=1}^{3^k} \left( S_i - n_i \frac{n + 1}{2} \right)^2,
\]

where \( S_i = \sum_{g_j = G_i} R_j \) is the rank sum of individuals with genotype \( G_i \).
Multilocus association scan

- qGTD is defined on the ranks of observed quantitative phenotype values of \( n \) individuals, i.e., \( \{R_1, \ldots, R_n\} \).
- Given a set of \( k \) markers, there are \( 3^k \) possible multilocus genotypes, denoted by \( \{G_1, \ldots, G_{3^k}\} \).
- For individual \( i \), let \( g_i \) be his/her unphased multilocus genotype on these \( k \) loci.

\[
qGTD = \frac{12}{n^3 - n} \sum_{i=1}^{3^k} \left( S_i - n_i \frac{n + 1}{2} \right)^2,
\]

where \( S_i = \sum_{g_j = G_i} R_j \) is the rank sum of individuals with genotype \( G_i \).
Multilocus association scan

- qGTD is defined on the ranks of observed quantitative phenotype values of $n$ individuals, i.e., $\{R_1, \ldots, R_n\}$.
- Given a set of $k$ markers, there are $3^k$ possible multilocus genotypes, denoted by $\{G_1, \ldots, G_{3^k}\}$.
- For individual $i$, let $g_i$ be his/her unphased multilocus genotype on these $k$ loci.

$$q\text{GTD} = \frac{12}{n^3 - n} \sum_{i=1}^{3^k} \left( S_i - n_i \frac{n + 1}{2} \right)^2,$$

where $S_i = \sum_{g_j = G_i} R_j$ is the rank sum of individuals with genotype $G_i$. 
Multilocus association scan

- \( q \text{GTD} \) has expectation 1 under the null hypothesis of no association;
- and has expectation greater than 1 when there is association between the set of markers and the phenotype.
- A greedy screening algorithm guided by \( q \text{GTD} \) is then used to screen out markers that do not contribute to increase the value of \( q \text{GTD} \)
- and retain a cluster of markers that contribute important information to the score.
- This greedy algorithm is repeated a large number of times (5,000,000 times) on random subsets of SNPs. The returned clusters and their scores are recorded.
- For the association scan, we used the 56 unrelated grandparents in the 14 CEPH families.
### Multilocus association scan

- \( q_{\text{GTD}} \) has expectation 1 under the null hypothesis of no association;

- and has expectation greater than 1 when there is association between the set of markers and the phenotype.

- A greedy screening algorithm guided by \( q_{\text{GTD}} \) is then used to screen out markers that do not contribute to increase the value of \( q_{\text{GTD}} \)

- and retain a *cluster of* markers that contribute important information to the score.

- This greedy algorithm is repeated a large number of times (5,000,000 times) on random subsets of SNPs. The *returned* clusters and their scores are recorded.

- For the association scan, we used the 56 unrelated grandparents in the 14 CEPH families.
Multilocus association scan

- $q$GTD has expectation 1 under the null hypothesis of no association;
- and has expectation greater than 1 when there is association between the set of markers and the phenotype.
- A greedy screening algorithm guided by $q$GTD is then used to screen out markers that do not contribute to increase the value of $q$GTD
- and retain a cluster of markers that contribute important information to the score.
- This greedy algorithm is repeated a large number of times (5,000,000 times) on random subsets of SNPs. The returned clusters and their scores are recorded.
- For the association scan, we used the 56 unrelated grandparents in the 14 CEPH families.
Multilocus association scan

- $q$GTD has expectation 1 under the null hypothesis of no association;
- and has expectation greater than 1 when there is association between the set of markers and the phenotype.
- A greedy screening algorithm guided by $q$GTD is then used to screen out markers that do not contribute to increase the value of $q$GTD
- and retain a cluster of markers that contribute important information to the score.
- This greedy algorithm is repeated a large number of times (5,000,000 times) on random subsets of SNPs. The returned clusters and their scores are recorded.
- For the association scan, we used the 56 unrelated grandparents in the 14 CEPH families.
Multilocus association scan

- $qGTD$ has expectation 1 under the null hypothesis of no association;
- and has expectation greater than 1 when there is association between the set of markers and the phenotype.
- A greedy screening algorithm guided by $qGTD$ is then used to screen out markers that do not contribute to increase the value of $qGTD$
- and retain a cluster of markers that contribute important information to the score.
- This greedy algorithm is repeated a large number of times (5,000,000 times) on random subsets of SNPs. The returned clusters and their scores are recorded.
- For the association scan, we used the 56 unrelated grandparents in the 14 CEPH families.
Multilocus association scan

- \( q \text{GTD} \) has expectation 1 under the null hypothesis of no association;
- and has expectation greater than 1 when there is association between the set of markers and the phenotype.
- A greedy screening algorithm guided by \( q \text{GTD} \) is then used to screen out markers that do not contribute to increase the value of \( q \text{GTD} \)
- and retain a cluster of markers that contribute important information to the score.
- This greedy algorithm is repeated a large number of times (5,000,000 times) on random subsets of SNPs. The returned clusters and their scores are recorded.
- For the association scan, we used the 56 unrelated grandparents in the 14 CEPH families.
Overall association and interaction association

- SNPs are first ranked by the numbers of times (return frequencies) that they are retained by the screening algorithm, which measure the overall importance of individual SNPs.
- Secondly, we filtered the retained SNP clusters by their qGTD scores and only selected the top 1000 distinctive clusters with the highest qGTD values.
- Using these 1000 clusters, we computed the qGTD return frequencies for each SNP.
- SNPs that present more frequently in clusters with higher qGTD values play a more critical role in gene-gene interactions that decide the variation of the phenotype.
- We selected the top 30 overall important SNPs and the top 30 important interaction SNPs with the highest qGTD return frequencies.
Overall association and interaction association

▶ SNPs are first ranked by the numbers of times (return frequencies) that they are retained by the screening algorithm, which measure the overall importance of individual SNPs.

▶ Secondly, we filtered the retained SNP clusters by their qGTD scores and only selected the top 1000 distinctive clusters with the highest qGTD values.

▶ Using these 1000 clusters, we computed the qGTD return frequencies for each SNP.

▶ SNPs that present more frequently in clusters with higher qGTD values play a more critical role in gene-gene interactions that decide the variation of the phenotype.

▶ We selected the top 30 overall important SNPs and the top 30 important interaction SNPs with the highest qGTD return frequencies.
Overall association and interaction association

- SNPs are first ranked by the numbers of times (return frequencies) that they are retained by the screening algorithm, which measure the overall importance of individual SNPs.

- Secondly, we filtered the retained SNP clusters by their qGTD scores and only selected the top 1000 distinctive clusters with the highest qGTD values.

- Using these 1000 clusters, we computed the qGTD return frequencies for each SNP.

- SNPs that present more frequently in clusters with higher qGTD values play a more critical role in gene-gene interactions that decide the variation of the phenotype.

- We selected the top 30 overall important SNPs and the top 30 important interaction SNPs with the highest qGTD return frequencies.
Overall association and interaction association

- SNPs are first ranked by the numbers of times (return frequencies) that they are retained by the screening algorithm, which measure the overall importance of individual SNPs.
- Secondly, we filtered the retained SNP clusters by their qGTD scores and only selected the top 1000 distinctive clusters with the highest qGTD values.
- Using these 1000 clusters, we computed the qGTD return frequencies for each SNP.
- SNPs that present more frequently in clusters with higher qGTD values play a more critical role in gene-gene interactions that decide the variation of the phenotype.
- We selected the top 30 overall important SNPs and the top 30 important interaction SNPs with the highest qGTD return frequencies.
Overall association and interaction association

- SNPs are first ranked by the numbers of times (return frequencies) that they are retained by the screening algorithm, which measure the overall importance of individual SNPs.
- Secondly, we filtered the retained SNP clusters by their qGTD scores and only selected the top 1000 distinctive clusters with the highest qGTD values.
- Using these 1000 clusters, we computed the qGTD return frequencies for each SNP.
- SNPs that present more frequently in clusters with higher qGTD values play a more critical role in gene-gene interactions that decide the variation of the phenotype.
- We selected the top 30 overall important SNPs and the top 30 important interaction SNPs with the highest qGTD return frequencies.
Overview of results

STK11 41657_s_at
PTEN 219370_s_at
PPM1D 203120_at
NCOA3 203421_at
RASSF1 217244_s_at
RAD51L1 203116_at
RAD51C 207046_at
BRCA1 206066_at
204146_at
TP53 207066_at
TP53BP2 206966_at
TP53I3 207066_at
PPM1D 207046_at
TP53 207066_at
TP53BP2 206966_at
TP53I3 207066_at
PPM1D 207046_at
BRCA1 206066_at
TP53 207066_at
TP53BP2 206966_at
TP53I3 207066_at
PPM1D 207046_at
BRCA1 206066_at
TP53 207066_at
TP53BP2 206966_at
TP53I3 207066_at
PPM1D 207046_at
BRCA1 206066_at
Hotspots?

- Overlap with genomic location of breast cancer candidate genes:
  - linkage at 2q, 11q, and 17q;
  - overall association at 1q, 2q, and 17q;
  - interaction association at 8, 17p and 20q.

- Linkage and association regulatory loci in common:
  - BARD1 [MIM 601593] on 2q34-35 and BRCA1 [MIM 113705] on 17q21.
Hotspots?

- Overlap with genomic location of breast cancer candidate genes:
  - linkage at 2q, 11q, and 17q;
  - overall association at 1q, 2q, and 17q;
  - interaction association at 8, 17p and 20q.

Hotspots?

- Overlap with genomic location of breast cancer candidate genes:
  - linkage at 2q, 11q, and 17q;
  - overall association at 1q, 2q, and 17q;
  - interaction association at 8, 17p and 20q.

- Linkage and association regulatory loci in common:
  BARD1 [MIM 601593] on 2q34-35 and BRCA1 [MIM 113705] on 17q21.
Hotspots?

- Overlap with genomic location of breast cancer candidate genes:
  - linkage at 2q, 11q, and 17q;
  - overall association at 1q, 2q, and 17q;
  - interaction association at 8, 17p and 20q.

- Linkage and association regulatory loci in common:
  BARD1 [MIM 601593] on 2q34-35 and BRCA1 [MIM 113705] on 17q21.
Hotspots?

- Overlap with genomic location of breast cancer candidate genes:
  - linkage at 2q, 11q, and 17q;
  - overall association at 1q, 2q, and 17q;
  - interaction association at 8, 17p and 20q.

Hotspots?

- Clustering using four sets of information:
  - the phenotype,
  - the number of shared interacting regulatory pairs,
  - the qGTD return frequencies (interaction),
  - the return frequencies (overall).

- BRCA1 and RAD51AP1 are found to share much more interacting regulatory loci than other transcript pairs.

- The grouping based on interacting regulatory activities is different from that based on overall regulatory activities.
Hotspots?

- Clustering using four sets of information:
  - the phenotype,
  - the number of shared interacting regulatory pairs,
  - the qGTD return frequencies (interaction),
  - the return frequencies (overall).

- BRCA1 and RAD51AP1 are found to share much more interacting regulatory loci than other transcript pairs.

- The grouping based on interacting regulatory activities is different from that based on overall regulatory activities.
Hotspots?

- Clustering using four sets of information:
  - the phenotype,
  - the number of shared interacting regulatory pairs,
  - the qGTD return frequencies (interaction),
  - the return frequencies (overall).

- BRCA1 and RAD51AP1 are found to share much more interacting regulatory loci than other transcript pairs.

- The grouping based on interacting regulatory activities is different from that based on overall regulatory activities.
Hotspots?

- Clustering using four sets of information:
  - the phenotype,
  - the number of shared interacting regulatory pairs,
  - the qGTD return frequencies (interaction),
  - the return frequencies (overall).

- BRCA1 and RAD51AP1 are found to share much more interacting regulatory loci than other transcript pairs.

- The grouping based on interacting regulatory activities is different from that based on overall regulatory activities.
Hotspots?

- Clustering using four sets of information:
  - the phenotype,
  - the number of shared interacting regulatory pairs,
  - the qGTD return frequencies (interaction),
  - the return frequencies (overall).

- BRCA1 and RAD51AP1 are found to share much more interacting regulatory loci than other transcript pairs.

- The grouping based on interacting regulatory activities is different from that based on overall regulatory activities.
Hotspots?

- Clustering using four sets of information:
  - the phenotype,
  - the number of shared interacting regulatory pairs,
  - the qGTD return frequencies (interaction),
  - the return frequencies (overall).

- BRCA1 and RAD51AP1 are found to share much more interacting regulatory loci than other transcript pairs.

- The grouping based on interacting regulatory activities is different from that based on overall regulatory activities.
Hotspots?

- Clustering using four sets of information:
  - the phenotype,
  - the number of shared interacting regulatory pairs,
  - the qGTD return frequencies (interaction),
  - the return frequencies (overall).

- BRCA1 and RAD51AP1 are found to share much more interacting regulatory loci than other transcript pairs.

- The grouping based on interacting regulatory activities is different from that based on overall regulatory activities.
Overview of results
Evaluating overlapping eQTL

Compare linkage signals for two gene expression traits (NBR1 and RAD51AP). (Overlapped linkage signals indicate evidence for co-regulation of these two transcripts.)
Association between two rankings on the same objects

- $X_i, i = 1, \ldots, n$ and $Y_i, i = 1, \ldots, n$ be two sets of independent rankings of $n$ objects.
- (Throughout, we discuss rankings in decreasing order.)
- Denote $\alpha_i$ as the importance of object $i$.
- $X_i$’s and $Y_i$’s are random representations of the true ranking, $\text{Rank}(\alpha_i)$.
- We assume that
  \[
  X_i = \text{Rank}(\alpha_i + \varepsilon_i), \quad Y_i = \text{Rank}(\alpha_i + \delta_i),
  \]
  where $\varepsilon_i \overset{iid}{\sim} F$ and $\delta_i \overset{iid}{\sim} G$.
- $\alpha$’s, $F$ and $G$ are introduced for the convenience of discussion.
Association between two rankings on the same objects

- \( X_i, i = 1, \ldots, n \) and \( Y_i, i = 1, \ldots, n \) be two sets of independent rankings of \( n \) objects.

- (Throughout, we discuss rankings in decreasing order.)

- Denote \( \alpha_i \) as the importance of object \( i \).

- \( X_i \)'s and \( Y_i \)'s are random representations of the true ranking, \( \text{Rank}(\alpha_i) \).

- We assume that

\[
\begin{align*}
X_i &= \text{Rank}(\alpha_i + \varepsilon_i), \\
Y_i &= \text{Rank}(\alpha_i + \delta_i),
\end{align*}
\]

where \( \varepsilon_i \overset{iid}{\sim} F \) and \( \delta_i \overset{iid}{\sim} G \).

- \( \alpha \)'s, \( F \) and \( G \) are introduced for the convenience of discussion.
Association between two rankings on the same objects

- $X_i, \ i = 1, \ldots, n$ and $Y_i, \ i = 1, \ldots, n$ be two sets of independent rankings of $n$ objects.
- (Throughout, we discuss rankings in decreasing order.)
- Denote $\alpha_i$ as the importance of object $i$.
- $X_i$’s and $Y_i$’s are random representations of the true ranking, $\text{Rank}(\alpha_i)$.
- We assume that

$$X_i = \text{Rank}(\alpha_i + \varepsilon_i),$$
$$Y_i = \text{Rank}(\alpha_i + \delta_i),$$

where $\varepsilon_i \overset{iid}{\sim} F$ and $\delta_i \overset{iid}{\sim} G$.
- $\alpha$’s, $F$ and $G$ are introduced for the convenience of discussion.
Association between two rankings on the same objects

- $X_i, i = 1, \ldots, n$ and $Y_i, i = 1, \ldots, n$ be two sets of independent rankings of $n$ objects.
- (Throughout, we discuss rankings in decreasing order.)
- Denote $\alpha_i$ as the importance of object $i$.
- $X_i$'s and $Y_i$'s are random representations of the true ranking, $\text{Rank}(\alpha_i)$.
- We assume that

\[
X_i = \text{Rank}(\alpha_i + \varepsilon_i), \\
Y_i = \text{Rank}(\alpha_i + \delta_i),
\]

where $\varepsilon_i \sim \text{iid } F$ and $\delta_i \sim \text{iid } G$.
- $\alpha$'s, $F$ and $G$ are introduced for the convenience of discussion.
Association between two rankings on the same objects

- $X_i, \ i = 1, \ldots, n$ and $Y_i, \ i = 1, \ldots, n$ be two sets of independent rankings of $n$ objects.
- (Throughout, we discuss rankings in decreasing order.)
- Denote $\alpha_i$ as the importance of object $i$.
- $X_i$’s and $Y_i$’s are random representations of the true ranking, $\text{Rank}(\alpha_i)$.
- We assume that

\[
X_i = \text{Rank}(\alpha_i + \varepsilon_i), \\
Y_i = \text{Rank}(\alpha_i + \delta_i),
\]

where $\varepsilon_i \overset{iid}{\sim} F$ and $\delta_i \overset{iid}{\sim} G$.

- $\alpha$’s, $F$ and $G$ are introduced for the convenience of discussion.
Association between two rankings on the same objects

- $X_i, i = 1, \ldots, n$ and $Y_i, i = 1, \ldots, n$ be two sets of independent rankings of $n$ objects.
- (Throughout, we discuss rankings in decreasing order.)
- Denote $\alpha_i$ as the importance of object $i$.
- $X_i$’s and $Y_i$’s are random representations of the true ranking, $\text{Rank}(\alpha_i)$.
- We assume that

  $$
  X_i = \text{Rank}(\alpha_i + \varepsilon_i),
  $$

  $$
  Y_i = \text{Rank}(\alpha_i + \delta_i),
  $$

  where $\varepsilon_i \overset{iid}{\sim} F$ and $\delta_i \overset{iid}{\sim} G$.

- $\alpha$’s, $F$ and $G$ are introduced for the convenience of discussion.
Association between two rankings on the same objects

Without loss of generality, assume that the objects are arranged in the order of their importance, that is

\[ \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n. \]

If \( \alpha_1 = \alpha_2 = \cdots = \alpha_n \), \( X \) is reduced to \( \text{Rank}(\varepsilon) \), would be independent of ranking \( Y=\text{Rank}(\delta) \).

If \( \alpha_1 > \alpha_2 > \cdots > \alpha_n \), \( X \) and \( Y \) will be positively correlated and the degree of correlation depends on the random variation of \( \varepsilon \)'s and \( \delta \)'s.

The correlation between rankings \( X \) and \( Y \) can be used to measure the variation among the objects’ importance (signal), relative to the amount of noises.
Association between two rankings on the same objects

- Without loss of generality, assume that the objects are arranged in the order of their importance, that is
  \[ \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n. \]

- If \( \alpha_1 = \alpha_2 = \cdots = \alpha_n \), \( X \) is reduced to \( \text{Rank}(\varepsilon) \), would be independent of ranking \( Y = \text{Rank}(\delta) \).

- If \( \alpha_1 > \alpha_2 > \cdots > \alpha_n \), \( X \) and \( Y \) will be positively correlated and the degree of correlation depends on the random variation of \( \varepsilon \)'s and \( \delta \)'s.

- The correlation between rankings \( X \) and \( Y \) can be used to measure the variation among the objects’ importance (signal), relative to the amount of noises.
Association between two rankings on the same objects

- Without loss of generality, assume that the objects are arranged in the order of their importance, that is

\[ \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n. \]

- If \( \alpha_1 = \alpha_2 = \cdots = \alpha_n \), \( X \) is reduced to \( \text{Rank}(\varepsilon) \), would be independent of ranking \( Y = \text{Rank}(\delta) \).

- If \( \alpha_1 > \alpha_2 > \cdots > \alpha_n \), \( X \) and \( Y \) will be positively correlated and the degree of correlation depends on the random variation of \( \varepsilon \)'s and \( \delta \)'s.

- The correlation between rankings \( X \) and \( Y \) can be used to measure the variation among the objects’ importance (signal), relative to the amount of noises.
Association between two rankings on the same objects

- Without loss of generality, assume that the objects are arranged in the order of their importance, that is
  \[ \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n. \]

- If \( \alpha_1 = \alpha_2 = \cdots = \alpha_n \), \( X \) is reduced to \( \text{Rank}(\varepsilon) \), would be independent of ranking \( Y=\text{Rank}(\delta) \).

- If \( \alpha_1 > \alpha_2 > \cdots > \alpha_n \), \( X \) and \( Y \) will be positively correlated and the degree of correlation depends on the random variation of \( \varepsilon \)'s and \( \delta \)'s.

- The correlation between rankings \( X \) and \( Y \) can be used to measure the variation among the objects' importance (signal), relative to the amount of noises.
Kendall rank-order correlation coefficient

The Kendall rank-order correlation coefficient (Kendall, 1955) is formulated as

\[ T = \frac{\# \text{ agreements} - \# \text{ disagreements}}{\text{total number of pairs}} \]

- Consider all possible pairs of \((X_i, X_j)\) in which \(X_i\) is lower than \(X_j\), if
  - if \(X_i\) is lower than \(X_j\), it is then an agreement;
  - if \(X_i\) is higher than \(X_j\), it is then an disagreement

\[
\# \text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} 1(X_i < X_j) 1(Y_i < Y_j);
\]

\[
\# \text{ disagreements} = \sum_{i=1}^{n} \sum_{i \neq j} 1(X_i < X_j) 1(Y_i > Y_j);
\]
Kendall rank-order correlation coefficient

The Kendall rank-order correlation coefficient (Kendall, 1955) is formulated as

\[ T = \frac{\# \text{ agreements} - \# \text{ disagreements}}{\text{total number of pairs}} \]

Consider all possible pairs of \((X_i, X_j)\) in which \(X_i\) is lower than \(X_j\), if

- if \(Y_i\) is lower than \(Y_j\), it is then an agreement;
- if \(Y_i\) is higher than \(Y_j\), it is then an disagreement.

\[ \# \text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} 1(X_i < X_j) 1(Y_i < Y_j), \]

\[ \# \text{ disagreements} = \sum_{i=1}^{n} \sum_{i \neq j} 1(X_i < X_j) 1(Y_i > Y_j). \]
Kendall rank-order correlation coefficient

The Kendall rank-order correlation coefficient (Kendall, 1955) is formulated as

$$T = \frac{\# \text{ agreements} - \# \text{ disagreements}}{\text{total number of pairs}}$$

► Consider all possible pairs of \((X_i, X_j)\) in which \(X_i\) is lower than \(X_j\), if
  ► if \(Y_i\) is lower than \(Y_j\), it is then an agreement;
  ► if \(Y_i\) is higher than \(Y_j\), it is then an disagreement.

\[
\# \text{ agreements} = \sum_{i=1}^{n} \sum_{i\neq j} 1(X_i<X_j)1(Y_i<Y_j),
\]

\[
\# \text{ disagreements} = \sum_{i=1}^{n} \sum_{i\neq j} 1(X_i<X_j)1(Y_i>Y_j).
\]

Kendall rank-order correlation coefficient

The Kendall rank-order correlation coefficient (Kendall, 1955) is formulated as

$$T = \frac{\# \text{ agreements} - \# \text{ disagreements}}{\text{total number of pairs}}$$

► Consider all possible pairs of \((X_i, X_j)\) in which \(X_i\) is lower than \(X_j\), if
  ► if \(Y_i\) is lower than \(Y_j\), it is then an agreement;
  ► if \(Y_i\) is higher than \(Y_j\), it is then an disagreement.

\[
\# \text{ agreements} = \sum_{i=1}^{n} \sum_{i\neq j} 1(X_i<X_j)1(Y_i<Y_j),
\]

\[
\# \text{ disagreements} = \sum_{i=1}^{n} \sum_{i\neq j} 1(X_i<X_j)1(Y_i>Y_j).
\]
Kendall rank-order correlation coefficient

The Kendall rank-order correlation coefficient (Kendall, 1955) is formulated as

\[
T = \frac{\# \text{ agreements} - \# \text{ disagreements}}{\text{total number of pairs}}
\]

- Consider all possible pairs of \((X_i, X_j)\) in which \(X_i\) is lower than \(X_j\), if
  - if \(Y_i\) is lower than \(Y_j\), it is then an agreement;
  - if \(Y_i\) is higher than \(Y_j\), it is then an disagreement.

\[
\# \text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} 1(X_i < X_j)1(Y_i < Y_j),
\]

\[
\# \text{ disagreements} = \sum_{i=1}^{n} \sum_{i \neq j} 1(X_i < X_j)1(Y_i > Y_j).
\]
Kendall rank-order correlation coefficient

The Kendall rank-order correlation coefficient (Kendall, 1955) is formulated as

$$T = \frac{\# \text{ agreements} - \# \text{ disagreements}}{\text{total number of pairs}}$$

- Consider all possible pairs of \((X_i, X_j)\) in which \(X_i\) is lower than \(X_j\), if
  - if \(Y_i\) is lower than \(Y_j\), it is then an agreement;
  - if \(Y_i\) is higher than \(Y_j\), it is then an disagreement.

\[
\# \text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} 1(X_i < X_j) 1(Y_i < Y_j),
\]

\[
\# \text{ disagreements} = \sum_{i=1}^{n} \sum_{i \neq j} 1(X_i < X_j) 1(Y_i > Y_j).
\]
Kendall rank-order correlation coefficient

- If there are no ties,
  \[ \text{# agreements} + \text{# disagreements} = \frac{n(n - 1)}{2}. \]
- Under the null hypothesis,
  \[
  \begin{align*}
  E(\text{# agreements}) &= E(\text{# disagreements}) = \frac{1}{4} n(n - 1), \\
  \text{var(\#agreements)} &= \frac{1}{16} \left( \frac{4n}{9} + \frac{10}{9} \right) n(n - 1). \\
  E(T) &= 0 \quad \text{and} \quad \text{var}(T) = \frac{2(2n + 5)}{9n(n - 1)}. \n  \end{align*}
  \]
Kendall rank-order correlation coefficient

- If there are no ties,
  \[ \text{# agreements} + \text{# disagreements} = \frac{n(n-1)}{2}. \]
- Under the null hypothesis,

  \[
  \begin{align*}
  \mathbb{E}(\text{# agreements}) &= \mathbb{E}(\text{# disagreements}) = \frac{1}{4}n(n-1), \\
  \text{var(\#agreements)} &= \frac{1}{16}(\frac{4n}{9} + \frac{10}{9})n(n-1). \\
  \mathbb{E}(T) &= 0 \text{ and } \text{var}(T) = \frac{2(2n+5)}{9n(n-1)}. 
  \end{align*}
  \]
When the number of the objects is large

- In most current studies, the number of objects is large, while the number of objects with higher importance is small.
  - \( H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_n = \alpha \)
  - versus a local alternative
    - \( H_a : \exists 1 \leq k_0 \ll n \), s.t.,
    - \( \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{k_0} > \alpha_{k_0+1} = \cdots = \alpha_n \).

- The strength of association measured by the original statistic is weakened by the large number of objects with undifferentiated importance.
When the number of the objects is large

- In most current studies, the number of objects is large, while the number of objects with higher importance is small.
  - \( H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_n \equiv \alpha \)
  - versus a local alternative
    \[ H_a : \exists 1 \leq k_0 \ll n, \text{ s.t.}, \]
    \[ \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{k_0} > \alpha_{k_0+1} = \cdots = \alpha_n. \]
- The strength of association measured by the original statistic is weakened by the large number of objects with undifferentiated importance.
When the number of the objects is large

In most current studies, the number of objects is large, while the number of objects with higher importance is small.

- \( H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_n \equiv \alpha \)
- versus a local alternative
  \( H_a : \exists 1 \leq k_0 \ll n, \text{ s.t.}, \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{k_0} > \alpha_{k_0+1} = \cdots = \alpha_n. \)

The strength of association measured by the original statistic is weakened by the large number of objects with undifferentiated importance.
When the number of the objects is large

- In most current studies, the number of objects is large, while the number of objects with higher importance is small.
  
  ▶ $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_n \equiv \alpha$
  
  ▶ versus a *local* alternative
    
    $H_a : \exists 1 \leq k_0 \ll n$, s.t.,
    
    $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{k_0} > \alpha_{k_0+1} = \cdots = \alpha_n$.

- The strength of association measured by the original statistic is weakened by the large number of objects with undifferentiated importance.
Modified Kendall association test

- Consider the truncated rankings $X_i^c = \min(X_i, k)$.
- The number of agreements can then be computed and tested on the truncated $X$ and $Y$.
- Using the truncated rankings, the noises from the objects with no signals are reduced.

\[
\text{# agreements} = \sum_{i=1}^{n} \sum_{i\neq j} 1(X_i^c < X_j^c) 1(Y_i^c < Y_j^c)
\]

and under the null hypothesis

\[
E(\text{# agreements}) = \frac{1}{4} n(n-1) \left( 1 - \frac{(n-k+1)}{\binom{n}{2}} \right)^2
\]
Modified Kendall association test

- Consider the truncated rankings \( X_i^c = \min(X_i, k) \).
- The number of agreements can then be computed and tested on the truncated \( X \) and \( Y \).
- Using the truncated rankings, the noises from the objects with no signals are reduced.

\[
\# \text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} 1(X_i^c < X_j^c) 1(Y_i^c < Y_j^c)
\]

and under the null hypothesis

\[
E(\# \text{ agreements}) = \frac{1}{4} n(n-1) \left( 1 - \left( \frac{n-k+1}{\binom{n}{2}} \right)^2 \right)
\]
Modified Kendall association test

- Consider the truncated rankings \( X^c_i = \min(X_i, k) \).
- The number of agreements can then be computed and tested on the truncated \( X \) and \( Y \).
- Using the truncated rankings, the noises from the objects with no signals are reduced.

\[
\# \text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbb{1}(X^c_i < X^c_j) \mathbb{1}(Y^c_i < Y^c_j)
\]

and under the null hypothesis

\[
E(\# \text{ agreements}) = \frac{1}{4} n(n - 1) \left( 1 - \frac{(n-k+1)}{\binom{n}{2}} \right)^2
\]
Modified Kendall association test

\[
\text{var} \left( \sum_{i=1}^{n} \sum_{j \neq i} 1(x_i^c < x_j^c) 1(y_i^c < y_j^c) \right) \\
= n(n-1) \left\{ \frac{1}{4} \frac{1}{n(n-1)} \left( 1 - \frac{(n - k + 1)(n - k)}{n(n-1)} \right)^2 \\
+ (n-2)(n-3) \left( \frac{1}{4} \frac{1}{n} \frac{(k-1)}{4} + \frac{1}{4} \frac{k-1}{3} \frac{n-k+1}{1} + \frac{1}{6} \frac{k-1}{2} \frac{n-k+1}{2} \right)^2 \\
+ (n-2) \frac{1}{6} \left( \frac{k}{3} + \frac{k-1}{2} \frac{n-k}{1} \right)^2 \\
+ (n-2) \frac{1}{9} \left( 1 - \frac{n-k+1}{3} \right)^2 \\
- \frac{1}{16} (n^2 - n) \left( 1 - \frac{(n - k + 1)(n - k)}{n(n-1)} \right)^4 \right\}.
\]
The modified Kendall rank-order test statistic is defined as

\[ T^c = \frac{\# \text{ agreements} - E(\# \text{ agreements})}{\sqrt{\text{Var}(\# \text{ agreements})}}. \]
Evaluating overlapping eQTL

Compare linkage signals for two gene expression traits (NBR1 and RAD51AP). (Overlapped linkage signals indicate evidence for co-regulation of these two transcripts.)
Evaluating overlapping eQTL

From OMIM (Online Mendelian Inheritance in Man):
“Dong et al. (2003) isolated a holoenzyme complex containing BRCA1 (113705), BRCA2, BARD1 (610593), and RAD51, which they called the BRCA1- and BRCA2-containing complex (BRCC). . . . . . . concluded that the BRCC is a ubiquitin E3 ligase that enhances cellular survival following DNA damage.”
Discussion and conclusion

- Use interaction information in eQTL mapping may recover more overlapped regulatory activities.
- Studying gene co-regulation can be more biological relevant when constrained to gene sets.
- Transcription hotspots summarize extent of overall overlap.
- Detailed tests on pairwise genetic overlap may be of interests to study interactions among genes in a gene set.
Discussion and conclusion

- Use interaction information in eQTL mapping may recover more overlapped regulatory activities.
- Studying gene co-regulation can be more biological relevant when constrained to gene sets.
- Transcription hotspots summarize extent of overall overlap.
- Detailed tests on pairwise genetic overlap may be of interests to study interactions among genes in a gene set.
Discussion and conclusion

- Use interaction information in eQTL mapping may recover more overlapped regulatory activities.
- Studying gene co-regulation can be more biological relevant when constrained to gene sets.
- Transcription hotspots summarize extent of overall overlap.
- Detailed tests on pairwise genetic overlap may be of interests to study interactions among genes in a gene set.
Discussion and conclusion

- Use interaction information in eQTL mapping may recover more overlapped regulatory activities.
- Studying gene co-regulation can be more biological relevant when constrained to gene sets.
- Transcription hotspots summarize extent of overall overlap.
- Detailed tests on pairwise genetic overlap may be of interests to study interactions among genes in a gene set.
Part II: Study gene inter-regulation.

Motivation


- Mutations within the promoter, or “blue” TF genomic location may affect the co-expression between the “red” enhancer, the “purple” RNA-binding protein and the “black” gene.

- The extent of co-expression among genes may be regulated as well.
Motivation


- Mutations within the promoter, or “blue” TF genomic location may affect the co-expression between the “red” enhancer, the “purple” RNA-binding protein and the “black” gene.

- The extent of co-expression among genes may be regulated as well.
The new *liquid association* method can be used to identify the regulators of co-expression for each gene pair. (Dr. Ker-Chau Li’s lab, UCLA and Statistics Sinica)
Regulator of a gene’s expression and regulator of inter-regulation

Mutation occurs in the regulatory region of a gene that is in the upper cascade of a pathway may affect the co-expression patterns among the genes in this pathway.
Differential allelic co-expression (DACE) test

- We are exploring a test that examines overall changes of co-expression due to a mutation at a genomic locus.
- For $n$ subjects and an SNP, subjects are divided into its $G$ genotypes.
- To study a set of $p$ expression phenotypes, first compute, within each genotype group, the Pearson correlation coefficients of the expression levels between all pairs of transcripts in a set.
- Denote $r_{ijg}$ as the correlation between genes $i$ and $j$ within genotype group $g$.
- We perform the "Fisher’s $z'$ transformation" on the original correlation values:

$$z_{ijg} = \frac{1}{2} \ln \left( \frac{1 + r_{ijg}}{1 - r_{ijg}} \right)$$
Differential allelic co-expression (DACE) test

- We are exploring a test that examines overall changes of co-expression due to a mutation at a genomic locus.
- For \( n \) subjects and an SNP, subjects are divided into its \( G \) genotypes.
- To study a set of \( p \) expression phenotypes, first compute, within each genotype group, the Pearson correlation coefficients of the expression levels between all pairs of transcripts in a set.
- Denote \( r_{ijg} \) as the correlation between genes \( i \) and \( j \) within genotype group \( g \).
- We perform the "Fisher's z' transformation" on the original correlation values:

\[
    z_{ijg} = \frac{1}{2} \ln \left( \frac{1 + r_{ijg}}{1 - r_{ijg}} \right)
\]
Differential allelic co-expression (DACE) test

- We are exploring a test that examines overall changes of co-expression due to a mutation at a genomic locus.
- For $n$ subjects and an SNP, subjects are divided into its $G$ genotypes.
- To study a set of $p$ expression phenotypes, first compute, within each genotype group, the Pearson correlation coefficients of the expression levels between all pairs of transcripts in a set.
- Denote $r_{ijg}$ as the correlation between genes $i$ and $j$ within genotype group $g$.
- We perform the "Fisher’s z' transformation" on the original correlation values:

$$z_{ijg} = \frac{1}{2} \ln \left( \frac{1 + r_{ijg}}{1 - r_{ijg}} \right)$$
Differential allelic co-expression (DACE) test

- We are exploring a test that examines overall changes of co-expression due to a mutation at a genomic locus.
- For $n$ subjects and an SNP, subjects are divided into its $G$ genotypes.
- To study a set of $p$ expression phenotypes, first compute, within each genotype group, the Pearson correlation coefficients of the expression levels between all pairs of transcripts in a set.
- Denote $r_{ijg}$ as the correlation between genes $i$ and $j$ within genotype group $g$.
- We perform the "Fisher’s $z'$ transformation" on the original correlation values:

$$z_{ijg} = \frac{1}{2} \ln \left( \frac{1 + r_{ijg}}{1 - r_{ijg}} \right)$$
Differential allelic co-expression (DACE) test

- We are exploring a test that examines overall changes of co-expression due to a mutation at a genomic locus.
- For $n$ subjects and an SNP, subjects are divided into its $G$ genotypes.
- To study a set of $p$ expression phenotypes, first compute, within each genotype group, the Pearson correlation coefficients of the expression levels between all pairs of transcripts in a set.
- Denote $r_{ijg}$ as the correlation between genes $i$ and $j$ within genotype group $g$.
- We perform the "Fisher's z' transformation" on the original correlation values:

$$z_{ijg} = \frac{1}{2} \ln \left( \frac{1 + r_{ijg}}{1 - r_{ijg}} \right)$$
Differential allelic co-expression (DACE) test

Consider an (over-simplified) model

\[ z_{ijg} = \beta_0 + \beta_1 X_g + \varepsilon_{ijg} \]

where \( X_g \) is 0, 1, or 2 (the number of the associated allele).

DACE test is then to test \( H_0 : \beta_1 = 0 \).

This test projects a number of open problems for statistical research.

FDR was used to corrected for multiple comparison.
Differential allelic co-expression (DACE) test

- Consider an (over-simplified) model

\[ z_{ijg} = \beta_0 + \beta_1 X_g + \varepsilon_{ijg} \]

where \( X_g \) is 0, 1, or 2 (the number of the associated allele).

- DACE test is then to test \( H_0 : \beta_1 = 0 \).

- This test projects a number of open problems for statistical research.

- FDR was used to corrected for multiple comparison.
Differential allelic co-expression (DACE) test

- Consider an (over-simplified) model

\[ z_{ijg} = \beta_0 + \beta_1 X_g + \varepsilon_{ijg} \]

where \( X_g \) is 0, 1, or 2 (the number of the associated allele).

- DACE test is then to test \( H_0 : \beta_1 = 0 \).

- This test projects a number of open problems for statistical research.

- FDR was used to corrected for multiple comparison.
Differential allelic co-expression (DACE) test

- Consider an (over-simplified) model

\[ z_{ijg} = \beta_0 + \beta_1 X_g + \epsilon_{ijg} \]

where \( X_g \) is 0, 1, or 2 (the number of the associated allele).

- DACE test is then to test \( H_0 : \beta_1 = 0 \).

- This test projects a number of open problems for statistical research.

- FDR was used to corrected for multiple comparison.
Data


- 22 recombinant inbred strains were generated by repeated inbreeding of F2 mice derived from 2 parental inbred strains, C57BL/6 (B6) and DBA/2 in the case of BXD RI strains.

- Gene expression: Affymetrix U74Av2 arrays, 12,488 transcripts.

- 3,033 SNPs were used.

- Studied 233 pathways with 30 members or less.
Data


- 22 recombinant inbred strains were generated by repeated inbreeding of F2 mice derived from 2 parental inbred strains, C57BL/6 (B6) and DBA/2 in the case of BXD RI strains.

- Gene expression: Affymetrix U74Av2 arrays, 12,488 transcripts.

- 3,033 SNPs were used.

- Studied 233 pathways with 30 members or less.
Data


▶ 22 recombinant inbred strains were generated by repeated inbreeding of F2 mice derived from 2 parental inbred strains, C57BL/6 (B6) and DBA/2 in the case of BXD RI strains.

▶ Gene expression: Affymetrix U74Av2 arrays, 12,488 transcripts.

▶ 3,033 SNPs were used.

▶ Studied 233 pathways with 30 members or less.
Data

- 22 recombinant inbred strains were generated by repeated inbreeding of F2 mice derived from 2 parental inbred strains, C57BL/6 (B6) and DBA/2 in the case of BXD RI strains.
- Gene expression: Affymetrix U74Av2 arrays, 12,488 transcripts.
- 3,033 SNPs were used.
- Studied 233 pathways with 30 members or less.
Data

- 22 recombinant inbred strains were generated by repeated inbreeding of F2 mice derived from 2 parental inbred strains, C57BL/6 (B6) and DBA/2 in the case of BXD RI strains.
- Gene expression: Affymetrix U74Av2 arrays, 12,488 transcripts.
- 3,033 SNPs were used.
- Studied 233 pathways with 30 members or less.
15 biological pathways were found to have significant regulators of co-expression, after adjusting for multiple comparison.
Current efforts

- Improve and device tests that take into consideration the dependence among correlation coefficients and different sample sizes for the genotype groups.
- Device a more general test of differential co-expression patterns.
- Bootstrap-based evaluation of significance.
Current efforts

- Improve and device tests that take into consideration the dependence among correlation coefficients and different sample sizes for the genotype groups.
- Device a more general test of differential co-expression patterns.
- Bootstrap-based evaluation of significance.
Current efforts

- Improve and device tests that take into consideration the dependence among correlation coefficients and different sample sizes for the genotype groups.
- Device a more general test of differential co-expression patterns.
- Bootstrap-based evaluation of significance.
Acknowledgement

- Collaborators: Shaw-Hwa Lo, Yuejing Ding and Lei Cong (Statistics, Columbia University); Shuang Wang (Biostatistics, Columbia University); Joon Hung Woo and Ju Han Kim (Seoul National University)
- National institutes of health for research support.
- The conference committee for organizing such a great intellectually stimulating and idea-exchanging conference.
Acknowledgement

- Collaborators: Shaw-Hwa Lo, Yuejing Ding and Lei Cong (Statistics, Columbia University); Shuang Wang (Biostatistics, Columbia University); Joon Hung Woo and Ju Han Kim (Seoul National University)

- National institutes of health for research support.

- The conference committee for organizing such a great intellectually stimulating and idea-exchanging conference.
Acknowledgement

- Collaborators: Shaw-Hwa Lo, Yuejing Ding and Lei Cong (Statistics, Columbia University); Shuang Wang (Biostatistics, Columbia University); Joon Hung Woo and Ju Han Kim (Seoul National University)
- National institutes of health for research support.
- The conference committee for organizing such a great intellectually stimulating and idea-exchanging conference.