**Chapter 3**

**Inferring the Global Genetic Architecture of Gene Transcripts From Ultra-High Dimensional Molecular Data (ENAR Presentation Title)**

**3.1 Introduction**

Mapping of expression quantitative trait loci (eQTL) are used to target areas of the genome where the expression is controlled by specific gene or set of genes. Knowledge about how changes in gene expression are encoded by eQTLs is a key to construct the genotype-phenotype map for complex traits or diseases. Traditional eQTL mapping is to associate one transcript with a single marker at a time, thereby limiting our inference about a complete picture of the genetic architecture of gene expression. Since this is the case multiple markers should be considered. Among current approaches for gene detection and mapping, genome-wide association studies (GWAS) have emerged as the most popular one, achieving a considerable success since their first publication in 2005 (13) A pervasive challenge when considering multiple markers among the whole genome is the ultra-high dimensionality of such a dataset for both the response and the predictors. Along with estimating the markers effects, many would have little to no effect for a given gene expression. Therefore sure screening (2) and selecting markers in a computationally efficient way is highly desired to map out an entire genome.

Gene-gene interactions, or epistatic effects, and gene-environment interactions have been studied intensively in genome-wide association studies for the purposes of modern biological and medical research (Evans et al., 2006; Manolio & Collins, 2007; Kooperberg & LeBlanc, 2008; Cordell, 2009). Including such interactions would show improvement in approximation of the response, prediction accuracy and also provide insight to how predictors may affect each other for a given response. With these improvements, including such interactions may complicate the inference and also grows the dataset substantially.

*Interaction Screening for Ultra-High Dimensional Data*

A variety of variable selection methods have been developed and in common use, such as the LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001), Dantzig selector (Candes & Tao, 2007), elastic net (Zou & Hastie, 2005), minimax concave penalty (MCP) (Zhang, 2010), and others (Zou, 2006; Zou & Li, 2008). Many methods possess favorable theoretical properties such as model selection consistency (Zhao & Yu, 2006) and oracle properties (Fan & Lv, 2011). When p is much larger than n, sure screening is a more realistic goal to achieve than oracle properties or selection consistency (Fan & Lv, 2008; Wang, 2009). Sure screening assures that all important variables are identified with a probability tending to one, hence achieving effective dimension reduction without information loss and providing a reasonable starting point for low-dimensional methods to be applied.

Talk more about other methods of variable selection used in situation with high-dimensional data and the limitations of them. An application of variable selection is presented that approaches to systematically detect main effects and interaction effects among all possible loci. The application extends marker regression and equally uses all markers at one time, considering multiple markers as possible eQTLs. The eQTLs could be both cis-QTL, coming from the same physical location as the gene expression, and trans-QTL, coming from other areas of the genome. All types of interactions between possible QTLs selected in the model were also considered. The variable selection procedure for eQTLs mapping can be seen as one of deciding which subset of variables have effects on phenotypes, and identifying out all possible effects of those markers.(10)

(Look more at HanHao Proposal page 27 on the pdf)

A reanalysis of published genetic and genomic data collected in a mapping population of C. elegans, gaining new discoveries on the genetic origin of gene expression differentiation, which could not be detected by a traditional one-locus/one-transcript analysis approach

**3.2 Methods**  
 **3.2.1 Adaptation of iForm procedure or Procedure for General Mapping Population**

For the purpose of this model an inbred backcross population was considered. This type of population was first proposed by Wehrhahn and Allard (1965) to isolate and identify QTLs that contribute to a quantitative trait that may be inherited. By doing such a process you are restricting the majority of the inherited genome of the offspring to the recurrent parent collectively among the individuals in the population. There is a small proportion that would come from the donor parent and would contribute to the genetic noise in the sample. Statistically speaking this would translate to only having to estimate one parameter for an additive effect per marker included in the final model. This approach can be extended to include an backcross population with an additional parameter to estimate for the dominant effect for each marker included in the final model. This would result in twice as many parameters to estimate however.

Like in other Marker/QTL regression the model consists of a transcript of interest as the response with a linear combination of the Marker/SNP genotypes as possible eQTLs in the model plus some residual noise. Given a large genomic dataset of pairs which are independent and identically distributed copies of , where is a p-dimensional predictor vector and Y is the response.

(3.1)

The **’s** are the coefficients for the additive effects of each marker. Like most genome-wide datasets the number of markers in the data set grossly outnumber the number of observations in the dataset. This results with p>>n. Therefore selection procedures would need to implemented in order to fit a linear regression model such as (3.1). We are already at the point of high-dimensional data but if we would like to include epistasis effects between marker as predictors as well it would increase the amount of predictors by if considering every possible interaction. The resulting linear model would grow to b,

(3.2)

Where the are the coefficients for the epistatic effects for all the quadratic and two-way interactions between the predictors.

* Selection Procedure iForm (what about iFort?)
  + Marginality Principle
    - Selects both main effects and interaction effects simultaneously
  + Strong Heredity principle
    - Only considers interaction after both main effects have already been selected
      * (Hao and Zhang 2014 Note on …)the strong heredity condition is neither “strong” nor restrictive. Here is simple illustration. Consider model (3.4) with p = 2 without quadratic effects X21 and X22 . The entire parameter space for the coefficient vector (β1, β2, γ12)⊤ is R3. The strong heredity condition excludes a low dimensional subset {β1β2 = 0, γ2 12 > 0} from R3. Since the excluded set has Hausdorff measure zero in the Euclidean space, the strong heredity condition would cover the entire model space almost surely.
    - Think about running weak condition to compare
  + Cut off for number of predictors to use in final model
    - Inclusion criteria for marker to be selected

**3.2.2 Considerations made with data and model**

* Coded as 1 and 2 for dominant and recessive genotypes
  + Could also be coded as -1/2 and 1/2
* Redundant markers removed before running the procedure
  + Redundant markers were considered as possible cis-QTLs
  + One draw back that could arise with marker datasets when attempting to run multiple linear regression is the possibility of duplicate markers in the dataset. If two different markers would happen to have exactly the same genotypes for each subject it would show up as an exact linear combination of each other if both markers were to be placed in the linear model.

**3.2.3**

Testing Additive and Dominant effects after iForm procedure? (like in marker regression section of Statistical Genetics of Quantitative Traits) no dominant effect bc of inbred backcross?

**3.3 Application**

**3.3.1 Caenorhabditis elegans data (Rockman et al. 2010)**

To genetically dissect the causes of different variability among C. elegans traits, transcript abundances were measured by microarray in developmentally synchronized young adult hermaphrodites of 208 recombinant inbred advanced intercross lines from a cross between the laboratory strain, N2, and a wild isolate from Hawaii, CB4856. (Rockman et al. 2010) The microarray performed provides a genome-wide coverage of C. elegans with over 20,000 predicted genes from WormBase, a public C. elegans genome database. The microarray data was preprocessed through a normal–exponential convolution background correction and were normalized using quantile standardization. The strains used are considered relatively divergent for the C. elegans population, however they are closely related. The two strains differ roughly at approximately 1 base pair per 900. In the original dataset each line was genotyped at 1454 single-nucleotide polymorphism (SNP) markers. Each of these markers were further investigated as possible eQTLs

In the original paper, Interval mapping was conducted and the entire genome was scanned for possible eQTLs for each of the transcripts measured. The iForm procedure was conducted on the same set of transcripts to see if new insight could be found by testing all markers simultaneously and also including two-way interactions between markers.

* Results of running iForm on C. Elegans data
* Comparison to original paper
  + Focus on a few transcripts. Maybe one that resulted in Rsq 0.5, 0.6, 0.7, 0.8, 0.9’s when running the iForm procedure. Run single marker regression and interval mapping on them. Compare results with the iForm output.
  + Hot spots on chromosome X and two on Chromosome IV
  + Cis and trans qtls selected
* Comparison between iForm and forward seleciton
  + 1. **Simulation**
* Results of running iForm on simulated data
* Comparison of typical forward regression
* Run 1000 simulations and use interval mapping, forward selection and iForm
  + Compare R-square values
  + Compare number of QTL’s identified

**3.4 Discussion**