Quantitative Analysis of Regulatory Variation
Quantitative Genetics applies to any trait showing variation, and molecular traits are no exception.

Quantitative genetics is now being applied to many areas of genomics, such as variation in gene and metabolic regulation.

**Genome**: the entire collection of genes (DNA) in the cell and their dynamics

**Transcriptome**: the entire collection of (RNA) transcripts and their dynamics in the cell

**Proteome**: the entire collection of proteins and their dynamics within the cell

**Metabolome**: the elements and dynamics of metabolism in a cell
QTLs Involved in Protein Regulation

One of the first studies of QTLs involved in regulation was Damerval et al. (1994), who analyzed the spot volumes of 72 anonymous proteins (from maize seed tissue) on high-resolution 2-D protein gels.

Genes controlling protein volume are, by definition, regulatory genes.

60 F$_2$ individuals scored, and a total of 70 QTLs were detected, affecting 46 of the 72 proteins.

25 of these proteins influenced by 2 (or more) QTLs.

33/70 QTLs showed strict additivity, 37/70 showed at least some dominance.
The amount of variation accounted for by the QTLs ranged from 16% (lower detection limit in this experiment) to 67%.

Cumulative amount of variation accounted for by all QTLs influencing a protein ranged from 37 - 90%.

4 QTLs detected only through their epistatic effects (no significant main effects).

15% of the proteins showed detectable epistasis.
More recent explorations into regulatory variation

Microarrays are currently very popular.

Their analysis yields a high-dimension gene expression phenotype, looking at perhaps thousands to tens of thousands of mRNA levels at once.
More recent explorations into regulatory variation

Variation in mRNA levels detected. What fraction of such variation is heritable?

What is the nature of eQTLs (expression QTLs) underlying mRNA (and other regulatory) variation?
The basic idea behind gene expression arrays

• With a complete (or partial) genome sequence in hand, one can array sequences from genes of interest on small chip, glass slide, or a membrane

• mRNA is extracted from cells of interest and hybridized to the array

• Genes showing different levels of mRNA can be detected
Types of microarrays

- **Synthetic oligonucleotide arrays**
  - Chemically synthesize oligonucleotide sequences directly on slide/chip/membrane (e.g., using photolithography)
  - Affymetrix, Agilent

- **Spotted cDNA arrays**
  - PCR products from clones of genes of interest are spotted on a glass slide using a robot
  - Extracted cellular mRNA is reverse-transcribed into cDNAs for hybridization
Cell type 1

Extract mRNA

Label mRNA with red fluorescent dye (Cy5)

Cell type 2

Extract mRNA

Label mRNA with Green fluorescent dye (Cy3)
Hybridize mRNA to array

Each spot (or feature) corresponds to a different gene.

The color of the spot corresponds to the relative concentrations of mRNAs for that gene in the two cell types.
mRNA mainly from Cell Type 2

mRNA equal mix from cell Types 1 and 2

mRNA mainly from Cell Type 1

Each spot (or feature) corresponds to a different gene.
mRNAs for these genes more abundant in cell type 2
mRNAs from these genes more abundant in cell type 1
mRNAs from these genes of roughly equal abundance in both cell types
Analysis of microarray data

• Image processing and normalization
• Detecting significant changes in expression
• Clustering and classification
  - Clustering: detecting groups of co-expressed genes
  - Classification: finding those genes at which changes in mRNA expression level predict phenotype
Significance testing-- GLM

\[ Y_{klij} = u + A_k + R_{kl} + T_i + G_j + TG_{ij} + e_{lki} \]

- Array k
- Replicate l in array k
- Gene j
- Replicate block within array
- Interaction between gene i and treatment j
- Treatment i
Problem of very many tests (genes) vs. few actual data vectors

- **Expectation**: A large number of the GxT interactions will be significant
  - Controlling experiment-wide p value is very overly conservative (further, tests may be strongly correlated)

- **Generating a reduced set of genes for future consideration** (data mining)
  - FDR (false discovery rate)
  - Empirical Bayes approaches
Example: Storey & Tibshirani’s (2003) analysis of a BRCA-1 vs. BRCA-2 cancer cell lines. 3,226 genes compared.

Setting a Bonferroni p value of 0.001 detects 51 significant genes. The probability that at least one of these is a false-positive is \( 1 - (1-0.001)^{3226} = 0.96 \). The expected number of false-positives is \( 3226*0.001 = 3.2 \), or 6% of declared significant.

Setting an FDR rate of 0.05 detects 160 genes, of which 5% are expected to be FP.

FDR detects more genes and have a lower FP rate!

Consider a specific gene: MSH. p value = \( 5.5 \times 10^{-5} \). q value of 0.013.
Which loci control array-detected changes in mRNA expression?

- **Cis-acting factors**
  - Control regions immediately adjacent to the gene
- **Trans-acting factors**
  - Diffusible factors unlinked (or loosely linked) to the gene of interest
- **Global (Master) regulators**
  - Trans-acting factors that influence a large number of genes
David Treadgill's (UNC) mouse experiment

- Recombinant Inbred lines from a cross of DBA/2J and C57BL
- The level of mRNA expression (measured by array analysis) is treated as a quantitative trait and QTL analysis performed for each gene in the array
- Plot of gene producing mRNA vs. eQTL locations, a waterfall plot.
Distribution of >12,000 gene interactions

Genomic location of mRNA level modifiers

CIS-modifiers

Genomic location of genes on array

Distribution of >12,000 gene interactions

Genomic location of mRNA level modifiers
Genomic location of mRNA level modifiers

Distribution of >12,000 gene interactions

Genomic location of genes on array

Genomic location of mRNA level modifiers
Candidate loci: Differences in Gene Expression between lines

- Correlate differences in levels of expression with trait levels
- Map factors underlying changes in expression
  - These are (very) often trans-acting factors
- Difference between structural alleles and regulatory alleles
General Patterns of Transcriptional Variation

Although it is still early, there appear to be some generalization coming out from evolutionary studies of the whole-genome transcript phenotype.

First, recall that when considering the conservation (or lack thereof) of gene expression, three features can change:

- **Level** (amount) of transcript
- **Breadth** (tissue distribution) of transcript
- **Temporal** (the timing) of transcript

Most studies (to date) has focused on level and perhaps breadth.
**Gene Expression Levels are Highly Heritable**

Between-individual variation in expression appears not only heritable, but in some cases highly heritable.

Brem et al (2002) looked at mRNA expression at 6200 yeast genes in a natural population x lab cross. Over 1500 were differentially expressed, and for many the expression difference segregated in haploid spores.

A follow-up study (Brem & Kruglyak 2005) of 5700 of these genes detected at least one eQTL in over half (2,984). Variation was highly heritable, with over half (3,546) of the genes having broad-sense heritabilities exceeding 70%.

Looking over maize, humans & mice, 40% of expressed genes had at least one detectable eQTL (Schadt et al 2003).
Cheunh et al (2003) looked at five highly expression-variable Human genes in three groups: unrelated individuals, full sibs, and monozygotic twins. Genes showed less within-group variation in expression the more similar the relatives.

Morely et al. (2004) followed 3,554 human genes, and for 1000 of these significant evidence of an eQTL was found.
Are there correlations between the rates of Sequence and expression divergence?

Genes vary in the amount of sequence conservation across species. Is conservation (or lack thereof) correlated between the sequence divergence and expression of a gene? A few general patterns appear to be emerging, although these may change as additional data comes on line.

- **Codon usage bias tracks level of expression in many, but not all, organisms.**

In yeast, bacteria and Drosophila, strong codon bias is seen.

In mammals, codon bias is very small, often not significant.

In Humans, Cheuch et al. Found a (small) positive correlation between expression level and codon bias.
• More highly expressed genes tend to evolve slower. Pal et al. (2001) looked at yeast (no tissues), finding that more highly expressed genes evolve slower.

• Genes expressed later in development evolve more rapidly than those expressed earlier.

• There are mixed signals on correlations between expression profile divergence (tissue variation) and sequence divergence.

Some studies have found a correction between expression profile divergence (tissue distribution) and sequence divergence, others have not.

Hence, while the amount of expression correlated with sequence conservation, distribution is not
Correlations Between Regulatory Divergence and Expression Level/Pattern

Now turn from correlations between expression patterns and sequence conservation to correlation between expression patterns and expression conservation.

• **Within- and between-species variation in expression level is generally correlated.**

  Genes with the large within-population variation tend to show the greatest between-population variance. However, there are genes with high within-species variance that does not translate into between-species variance.

• **The conservation of expression level of a gene is directly proportional to its range of tissue expression.**

  Runs counter to the logic that tissue-specific genes are usually highly specific, and hence expression likely to be conserved.
• **Different tissues can show different rates of expression level conservation.**

  Khaitovich et al. (2005) looked at expression in five tissues in humans and chimps, finding significant differences between tissues in conservation of expression level.

  Brain showed the least divergence, the liver the most.

  *Genes expressed in only a single tissue showed both the highest divergence in expression level and the highest divergence in sequence conservation.*

  *Genes expressed in all five tissues showed the least divergence*
Constraints on expression levels

Looking in yeast, Fay et al. 2004 found 5448 genes with expression changes, a ratio of 0.887 expression changes/synonymous substitution.

Non-synonymous /syn. ratio = 0.175 (5 times higher)
Inter-genic /syn. ratio = 0.291 (3 times higher)

Hence, only slightly less constrained that synonymous sites!
Does Divergence in Expression follow a neutral model?


Under a strict neutral model (no selection constraints) mean expression level changes via Brownian motion.

Khaitovich et al., looking at 12,000 genes in 6 humans, 3 chimps, an orangutan and a rhesus macaque observed that the variance in the amount of expression increased linearly with time, as predicted from Brownian motion.

More strikingly, they observed that rates of expression divergence between species does not vary between intact and pseudogenes.
Mutation variance in expression levels and divergence

Three studies involving mutation variance suggest the strictly neutral view may be incorrect.

Rifkin et al. (2005) looked at the mutation variance for gene expression levels in 12 D. melanogaster lines that have been accumulating mutations for 200 gens.

Estimates of the mutation heritability had medium values of $2.5 \times 10^{-5}$.

Genes with higher $\sigma^2_m$ tend to have a larger between-species difference.

The observed between-species divergence far less than predicted from these $\sigma^2_m$ values.
Denver et al. (2005) compared levels of gene expression in mutation-accumulation (MA) lines of C. elegans vs. natural isolates (NI)

9% of MA lines had significant differential expression vs. 2% for NI.

The between-line variance in MA >> between-NI variance, although the later have been separated for much longer

Denver et al. took this excessive divergence in MA to be consistent with stabilizing selection on MA lines. (MA lines have small $N_e$ and hence weak selection)

Assuming NI lines in mutation-drift equilibrium, ratio of standing variation to mutation variance should be $4N_e$ (elegans self, hence $4N_e$ vs. $2N_e$). For all genes, this ratio fall below expected value of $N_e = 20,000$.

Most of the MA differences appeared due primarily to a few trans-acting factors with multiple down-stream effects.
In the final study, Lemos et al. (2005) looked at two lab strains of mice, human vs. chimp, two mice species, and two Drosophila species.

Using standard tests for neutral drift on a trait, 60-90% of the gene comparisons more consistent with stabilizing selection than genetic drift.

Lemos et al. Also suggest the striking observation of primate pseudogenes and active genes showing similar levels of divergence may result because the “pseudogenes” are really under selection, being conserved for several million years.
Analysis of Pathways

Microarrays offer a static snap-shot of the global pattern of mRNA expression.

Other (evolving) tools are allowing us to look at other features in the cell as well.

2-D protein gels are the (rough) equivalent of microarrays, offering a picture of protein expression. Protein chips based on specific antibodies to specific proteins will provide greater resolution and allow global protein analysis to approach microarrays in scope.

Two-hybrid screens allow protein-protein interactions (based on physical contact) to be screened in a cell
By (quite laboriously) looking at all pairwise-interactions of all of the proteins in the cell, a protein-protein interaction map can be constructed of which proteins physically contact which other proteins.
Drosophila protein-protein interaction map
Yeast protein-Protein map
C. elegans protein-protein map
Flux and Pathways

How best to increase the flux through this pathway?

Perhaps we increase the concentration of $e_1$.

However, it may be more efficient to increase the concentration of $e_4$.

The flux control coefficient, introduced by Kascer and Burns, provides a quantitative solution to this problem.
Kascer-Burns Sensitivity Analysis (aka. Metabolic Control Analysis)

“No theory should fit all the facts because some of the facts are wrong” (N. Bohr)

“All models are wrong, although some models are Useful” (Box)
Flux Control Coefficients, $C_{ji}^j$

The control coefficient for the flux at step $i$ in a pathway associated with enzyme $j$,

$$C_{ji}^j = \frac{\partial F_i}{\partial E_j} \frac{E_j}{F_i} = \frac{\partial \ln F_i}{\partial \ln E_j}$$

Roughly speaking, the control coefficient is the percentage change in flux divided by percentage change in enzyme activity.
Why many mutations are recessive: a 50% reduction in activity (the heterozygote) results in only a very small change in the flux.

When the activity of E is large, C is close to zero.
Why many mutations are recessive: a 50% reduction in activity (the heterozygote) results in only a very small change in the flux.

When the activity of E is near zero, C is close to 1.
Kacser-Burns Flux summation theorem:
\[ \sum_i C^j_i = 1 \]

- While most values of $C$ for proteins are positive, negative regulators (repressors) give negative values, allowing for $C$ values $> 1$.
- If a control coefficient is greatly increased in value, this decreases the values of other control coefficients.
- Coefficients are not intrinsic properties of an enzyme, but rather a (local) system property.
- Truly rate-limiting steps are rare.
“Rate-limiting” steps in pathways

Small-Kacser theorem: the factor $f$ by which flux is increased by an $r$-fold increase in activity of $E$ is

$$f = \frac{1}{1 - \frac{r - 1}{r} C_E^j}$$

Hence, the limiting increase in $f$ is

$$f = \frac{1}{1 - C_E^j}$$
Regulatory networks and graph theory

One mathematical presentation of a regulatory or metabolic network is given by a graph.

A graph is collection of nodes and edges (or lines) that connect those nodes that interact.

For example, suppose proteins A and B interact. We denote each protein with a node and connect them with an edge.

A node with multiple edges means that element interacts with a number of other nodes.
A graph represents the basic topology of a network. The actual dynamics of that network, its rates and types of interactions as well as its output and behavior, depend on both the topology and the dynamical equations typically given by a set of differential equations) for how these elements interact.

An intermediate set between the simple topology and the complete dynamical equations is a directed graph, where nodes are connected by arrows representing the directions of interaction.
Finally, we can represent the basic topology of a graph by a matrix $M$ containing zeros and ones.

In a simple graph, $M_{ij}$ is one if nodes $i$ and $j$ are connected by an edge. Hence, such a matrix is a square, symmetric matrix.

In a directed graph, a one means that node $i$ influences node $j$. In this case, $M_{ij} = M_{ji}$ only if both nodes influence each other, so that the matrix is typically not symmetric.
Erdos-Renyi Random Graphs and Random Boolean Networks

In order to understand a regulatory graph, we first need to consider the features of a random graph.

Random graphs, where edges are randomly collected to nodes, are often called Erdos-Renyi graphs after the two mathematicians who considered their basic properties.

Formally, we start with n nodes, with p the probability that any two randomly-chosen nodes are connected.
Descriptive properties of a Graph

The degree distribution $P(k)$ is a central feature of any graph, and simply is the probability that a randomly-chosen node is linked to exactly $k$ other nodes.

Under ER graphs, $P(k)$ follows a Poisson distribution, with success parameter $z = n*p$.

The path length distribution $d$ is a measure of how well connected all of the nodes in the graph are to each other. This is simply the distribution of smallest number of steps that it takes to go from one node in the graph to any other.
For an ER graph, the average path length between any two nodes is roughly given by

\[ d_{ER} \sim \frac{\ln(n)}{\ln(z)}, \quad z = np \]

A final important property of a graph is its cluster coefficient, \( C \), the probability that two neighbors of a given node are also neighbors of each other.

For an ER graph, \( C \sim z/n \)

ER graphs display a very striking phase transition at \( z = 1 \) (on average, one connection between two randomly-chosen nodes)
A giant component forms in the graph where a large fraction of the nodes in the graph are interconnected.
The biological implication of a giant component is that most of the elements in the graph influence (or at least interact with) most other elements.

Thus, the regulatory network moves from a series of discrete, non-overlapping compartments into a single integrated structure.

In a classic paper, Kaufmann (1967) modeled gene circuits as random Boolean networks. Recall Boolean simple mean binary (on/off etc).

In Kaufmann’s model, genes were randomly connected (formed an ER graph) and simple on/off rules applied (such as if your neighbor is on, you are one).
Kaufmann found that such random networks showed giant regulatory components, features that appear to show complex order and give all the appearance of being highly evolved, yet they were entirely random.

Key: random graphs can show considerable structure and one must be care in inferring evolution simply because a graph shows a complex, and highly integrated, structure.

So what do real graphs, such as the protein-protein interaction graph (from two-hybrid data) show?
Small worlds, Scale-free Graphs and Power Laws

Regulatory and metabolic graphs examined to date share two critical features

First, they are small-world graphs, which means that the mean path distance between any two nodes is short. The members live in a small world (Bacon, Erdos numbers)

Formally, a small-world graph means that the average path distance is close to that predicted for an ER graph, $d_{ER} \sim \ln(n) / \ln(z)$.

One can transform a highly regular graph into a small world graph by simply randomly rewiring a small fraction of the nodes.
The critical feature of small-world graphs is that they propagate information very efficiently.

The second feature that studied regulatory/metabolic networks showed is that the degree distribution (probability distribution that a node is connected to k other others) follows a power law

\[ P(k) \sim k^{-\gamma}, \quad P(k) \sim z^k e^{-k} /k! \]

By contrast, recall that for ER graphs, the degree distribution follows a Poisson distribution, and hence following a maximal value, the change of more links falls off exponentially,
Under a Poisson, a modal value for number of links, after which the chance falls off exponentially.

$P(k)$ represents the probability distribution of the number of links $k$. The graph shows two distributions: a Power law and a Poisson distribution. The modal value for links occurs at a specific $k$ value, after which the chance decreases exponentially.
Under a power law, no modal value, but the probability of many links falls off as a power, not exponentially, resulting in a few nodes with a large number of links.
Graphs with a power distribution of links are called 
**scale-free graphs**.

In a scale-free graph, a few of the nodes will have **very many connections**. Such nodes are often called **hubs**.

We have seen the presence of such hubs in the protein-protein graph, and you can also see them in the waterfall plot for eQTLs, where a few locations influence a large number of transcripts.

A plot of $P(k)$ (the degree distribution) vs $k$ will be **linear** if a power law holds so that one can check this for any given graph.
Hubs in the mRNA graph
The yeast protein-protein interaction data show such a linear relationship (with $\gamma \sim 2$, e.g. $P(k) \sim k^{-2}$),

Jeong et al (2000) examined the metabolic networks of 43 organisms (spanning the three major kingdoms of life), and found at all also followed a power law, again with $\gamma \sim 2$.

Scale-free graphs show they very important feature that they are fairly robust to perturbations. Most randomly-chosen nodes can be removed with little effect on the system.

While removing a hub has a critical effect, the chance that a randomly-chosen node is a hub is small.
Gene knock-out experiments in yeast, where every single gene was deleted one at a time, showed they only a very small fraction had any effect on phenotype. This is entirely consistent with the developmental pathways leading to phenotypes resulting from scale-free graphs.

Since a scale-free structure gives a regulatory network inherent stability, biological homeostasis may just be a simply consequence of this structure, rather than a highly evolved feature.

How might such scale-free graphs evolve? The answer turns out to be rather simple: when we add new nodes, they have a slight preference to attach to already established nodes.
This feature is exactly what is thought to happen as gene duplication adds more nodes (elements) to a network.

Jeong et al (2000) offer some support for this model, by looking at the hubs in their survey of the metabolic networks of 43 species.

Here the nodes are substrates for metabolic reactions.

They found that the rankings of the most highly-connected substrates (the hubs of their metabolic graph) was essentially the same across all 43 species.

This is especially striking given that only 4% of the substrates were present in all 43 species.
Conversely, species-specific substrates in general are much less connected.

Consistent with a model of evolution where ancient features serve as the hubs, with new features being preferentially connected to the ancient hubs.