X02 Application

# Interdisciplinary Consortium on the Genetics and Co-Morbidity of Stress

*Submitted:* April 18, 2006

*PI:* David W. Threadgill
Department of Genetics, CB#7264
University of North Carolina
Chapel Hill, NC 27599
TEL: (919) 843-6472
FAX: (919) 966-3292
E-MAIL: dwt@med.unc.edu

*Co-I:* Elissa J. Chesler, Oak Ridge National Laboratory, Oak Ridge, TN
Gary A. Churchill, The Jackson Laboratory, Bar Harbor, ME
Michael A. Langston, University of Tennessee, Knoxville, TN
Leonard McMillan, University of North Carolina, Chapel Hill, NC
Darla Miller, Oak Ridge National Laboratory, Oak Ridge, TN
Kenneth F. Manly, University of Tennessee Health Science Center, Memphis, TN
Fernando Pardo Manuel, University of North Carolina, Chapel Hill, NC
Cynthia B. Peterson, University of Tennessee, Knoxville, TN
Susan S. Smyth, University of North Carolina, Chapel Hill, NC
Brynn H. Voy, Oak Ridge National Laboratory, Oak Ridge, TN
Edward K. Wakeland, University of Texas Southwestern Medical Center, Dallas, TX
Wei, Wang, University of North Carolina, Chapel Hill, NC
Roberts W. Williams, University of Tennessee Health Science Center, Memphis, TN

## Project Summary

Modern American society is faced with a number of common diseases whose etiologies are highly heterogeneous, yet are connected through complex networks of genetic, molecular, cellular, and tissue interactions. A common thread across many of the most debilitating diseases, particularly those that contribute to high mortality like neurological and cardiovascular, is an increased level of psychological and environmental stress. While we now know much about fundamental mechanisms, particularly those close to the hypothalamic-pituitary-adrenal axis (HPA), we do not yet understand the web of genetics and genomic factors that account for highly variable susceptibility to stress among humans. Stress is in many ways an ideal health problem to exploit as a "case study" for developing a new field of integrated biomedical research called *systems genetics*. Systems genetics is a non-reductionist field that was simply not practical even a few years ago: it relies on diverse multiscale and multiorgan phenotype data sets obtained from large segregating populations, sophisticated statistical and graph-theoretical methods, and high-performance computing. Systems genetics has the goal and potential to dissect and reassemble complex molecular and phenotypic networks in the context of natural genetic variation within populations. It represents a new synthetic phase of genetics. It also represents a genuine paradigm shift in the way of thinking about biological problems and solutions. Individualized and predictive medicine, or more accurately probabilistic medicine, will not be possible over the next two decades without a strong foundation in systems genetics.

To catalyze the development of this emerging field we have assembled a unique multidisciplinary group of scientists called The Interdisciplinary Consortium on the Genetics and Co-Morbidity of Stress (ICOGS). Our group will use complementary approaches to dissect stress responses and consequences in a genetically diverse reference population of mice appropriately called the Collaborative Cross (CC). To achieve this goal, we will 1) Institute a governance plan, 2) Develop key resources and a conceptual framework for systems genetics, 3) Assemble a research consortium to investigate the systems genetics of stress, and 4) Design programs to foster the growth and broader application of systems genetics.

*"Stress is not a vague concept, somehow related to the decline in the influence of traditional codes of behavior, dissatisfaction with the world, or the rising cost of living, but rather that it is clearly a definable biological and medical phenomenon whose mechanisms can be objectively identified and with which we can cope much better once we know how to handle it."*

H. Selye, 1956

## Specific Aims

Modern American society is faced with a number of common diseases whose etiologies are highly heterogeneous, yet are connected through complex networks of genetic, molecular, cellular, and tissue interactions. Despite major advances in the treatment of many diseases, adverse environmental changes have often offset gains. An obvious example that applies with force in the United States is the sharp reduction in exercise, concomitant gains in body weight, and the epidemic of type II diabetes. A common thread across many of the most debilitating diseases, particularly those that contribute to high mortality, is an increased level of psychological and environmental stress. Going back to catalytic work by Hans Selye (1936, 1956), stress has been recognized as a key modulator of diseases that range from behavioral disorders, metabolic syndrome, cardiovascular and infectious diseases, and cancer. While we now know much about fundamental mechanisms, particularly those close to the hypothalamic-pituitary-adrenal axis (HPA), we do not yet understand the web of genetics and genomic factors that account for highly variable susceptibility to stress among humans. Stress is in many ways an ideal health problem to use as a "case study" for deployment of a new field of integrated biomedical research called *systems genetics*. Systems genetics is a non-reductionist field that was simply not practical even a few years ago: it relies on diverse multiscale and multiorgan phenotype data sets obtained from large segregating populations, sophisticated statistical and graph-theoretical methods, and high-performance computing. Systems genetics has the goal and potential to dissect and reassemble complex molecular and phenotypic networks in the context of natural genetic variation within populations. It represents a new synthetic phase of genetics. It also represents a genuine paradigm shift in the way of thinking about biological problems and solutions. Individualized and predictive medicine, or more accurately probabilistic medicine, will not be possible over the next two decades without a strong foundation in systems genetics.

To catalyze the development of this emerging field we have assembled a unique multidisciplinary group of scientists called The Interdisciplinary Consortium on the Genetics and Co-Morbidity of Stress (ICOGS). Our group will use complementary approaches to dissect stress responses and consequences in a genetically diverse reference population of mice appropriately called the Collaborative Cross (CC). The CC is a novel mouse model system that supports integration of multiple data types on fixed genomes that can also be used to generate reproducible testable models for predictive medicine. To achieve this goal, we will:

**Specific Aim 1: Institute a governance plan**. To successfully implement the ICOGS requires a system that melds each discipline. This requires steady communication, collegial interactions and joint publications. We will develop an administrative structure that supports these activities.

**Specific Aim 2: Develop key resources and a conceptual framework for systems genetics.** Complex biological questions require a robust and innovative model system. A research core will provide a central resource to study stress in the context of highly diverse cell, tissue and organism responses across a wide variety of genotypes with sufficient diversity to model human populations. A bioinformatic infrastructure will also be built to support data collection, curation, error-checking, integration, and distribution with a powerful portal to ICOGS resources.

**Specific Aim 3: Assemble a research consortium to use systems genetics to investigate the biological consequences of stress**. The ICOGS team will collect and integrate high-dimensional data sets using state-of-the-art tools and protocols. We will also develop new methods to analyze data streams in the context of causal linkage between stress and susceptibility-to-disease phenotypes.

**Specific Aim 4: Design programs to foster the growth and broader application of systems genetics.** Our goal is not only to develop models describing how biological information is altered by external stress leading to altered physiology and disease susceptibility, but also to cultivate interdisciplinary systems genetics research through novel interactions and learning environments, and to educate current and future generations of researchers. TO achieve this goal, we will develop outreach and fellowship programs focusing on the applications of systems genetics.

## Background and Significance

**Stress: a common etiological factor for diseases with rising incidences.** Stress is any physical, circumstantial or emotional change that requires a biological adjustment. Estimates suggest that 80% of all illnesses and health problems are directly or indirectly related to stress. Additionally, stress, either psychological or environmental, is thought to contribute to eight of the top ten causes of death, accounting for 71% of all deaths, in the US according to statistics from the Centers for Disease Control (CDC). This includes the top four: heart disease, malignant neoplasia, cerebrovascular disease and chronic lower respiratory disease. Of even greater concern is that stress is increasingly a characteristic for many Americans. Although the range of stressors is highly heterogeneous, significant similarities exist between different types of stressors and the spectrum of biological changes they induce.

*Obesity and metabolic syndrome.* Health statistics show that more than 50% of Americans are overweight and that 25% are obese, numbers that have increased dramatically in the last few decades (CDC). Being overweight increases risk for many health problems. Chronic stress is strongly associated with obesity and its co-morbidities through a number of mechanisms. A well-examined effect of stress is on glucocorticoids (GC), which inhibit activity in the hypothalamus-pituitary-adrenal axis (Dallman *et al*., 2003). The chronic actions of GCs on brain are excitatory. GCs increase expression of corticotropin-releasing factor (CRF), which induces a chronic stress response network. This increases measurable compulsive activities, including food intake that increases abdominal fat depots. This leads to further disinhibition of CRF activity and in humans, further increases in food intake and weight gain in some or decreased intake and body weight loss in others. This chronic stress-response network and its associated anxiety alters the metabolic state, effecting susceptibilities to other disease as well.

Other types of stress, like psychological stress, can cause food to be a comfort and distraction, causing many people to gain weight. This creates a downward spiral as stress increases the propensity to become obese while additional obesity further increases stress levels, both emotionally and metabolically. Since obesity is a risk factor for many co-morbidities like hypertension, stroke and heart disease, elevated metabolic stress caused by obesity has far reaching consequences. Fat depots secrete hormone-like substances that decrease insulin and increase development of metabolic syndrome. Abdominal fat is correlated with increased levels of C-reactive protein (CRP), a marker of inflammation, suggesting obesity leads to a chronic state of inflammatory stress. An additional stress response is cortisol secretion. Consequently, stress-induced metabolic changes and cortisol release leads to increases in abdominal fat depots, further exacerbating the stress response.

In additional to the other effects, obesity can lead to increased oxidative stress. Production of reactive oxygen species (ROS) is increased selectively in adipose tissue in obese mice, creating systemic oxidative stress (Furukawa *et al.,* 2004).

*Cardiovascular disease*. Environmental, emotional or psychological stress is commonly thought to contribute to the development of cardiovascular disease (CVD), the leading cause of death in the United States. Probably the most dramatic example of the effects of environmental stress on the cardiovascular system is Takotsubo syndrome, an acute reversible form of profound heart failure typically observed in middle-aged women and frequently precipitated by the experience of emotional or physiologic stress. Relatively little is understood about the relationship of chronic stressors, such as job strain, social isolation, and personality traits, on the development of CVD. Epidemiologic studies have provided evidence of connections between stress and CVD, although it still not clear if stress represents an independent risk factor for the development of CVD or acts by influencing known CVD risk factors such as blood pressure, lipid levels, diet, and physical inactivity. Additionally, chronically elevated levels of hormones like cortisol or adrenaline and sex differences in physiological response to stress vary and most likely contribute to CVD prevalence. Research in the field suffers from a perceived lack of pathophysiologically reliable mechanisms to account for stress responses, which is, in large part, a consequence of a poor understanding of the complex interactions of stress with genetically-mediated susceptibility (reviewed in DeRijk and de Kloet, 2005).

*Inflammation, cancer and infection susceptibilities.* Although not clearly demonstrated in humans, chronic stress like that achieved by social disruption can promote cancer development in mouse models (Liu and Wang, 2005). This effect is hypothetically mediated by an altered immune response or changes to the endocrine system. Natural killer (NK) cells are a possible link. Stress decreases NK-cell activity in as little as an hour after exposure and this inhibition can last for as long as two days after removal of the stressor (Ben-Eliyahu *et al.,* 1999). NK-cell levels have also been shown to decrease during job stress (Morikawa *et al*., 2005). An alternative link with stress is through altered B-cells function. In fact, B-cells appear to be the immune cells most influenced by stress. Interestingly, changes in B-cell activity may be associated with

decreased NK-cell activity. B-cells produce Il12, which activates NK-cells. Stress can also impede cells ability to repair DNA damage (Glaser *et al*., 1985b) and T-cell activity (Glaser *et al.,* 1985a). Breast cancer patients who feel higher levels of stress upon diagnosis show evidence of a weakened immune system when compared to those with less stress upon diagnosis. Coincidently, the women with the highest levels of post-diagnosis stress also had lower levels of NK-cells (Anderson, 2003).

*Behavioral and neurological disease.* The central nervous system (CNS) appears to be a key mediator of the stress response. An initial step to most stressors involves interpreting diverse sensory input and assigning them affective and emotional value. Emotional valence can be assigned strictly by innate mechanisms (the freezing response evoked in naïve mice by the first smell of cat or rat) or may be learned (the freezing response evoked by a conditioned stimulus previously paired with mild foot shock). Evaluation involves complex cortical and subcortical networks, especially reciprocal connections among prefrontal and limbic areas, the basolateral and central amygdala, the striatum, nucleus accumbens, and the bed nuclei of the stria terminalis. Affective states, such as anxiety, fear, pleasure, anticipation, and aggressive tone are partially generated in these telecephalic derivatives. As a group they communicate downstream to the medial preoptic area, hypothalamus, and numerous thalamic and brainstem nuclei. This system is rich with extensive parallel feedback connections and it is clear that notions of causality and feedback cannot be tagged to either single CNS nuclei or directed connections (nodes or vertices, to use the vocabulary of our computer science colleagues in the ICOGS). Links have also been made between stress-induced HPA changes and inflammation (Chrousos and Kino, 2005; Fig. 1) and cardiovascular disease (Benarroch, 2005), implying a tangled web of connection and co-morbidities.

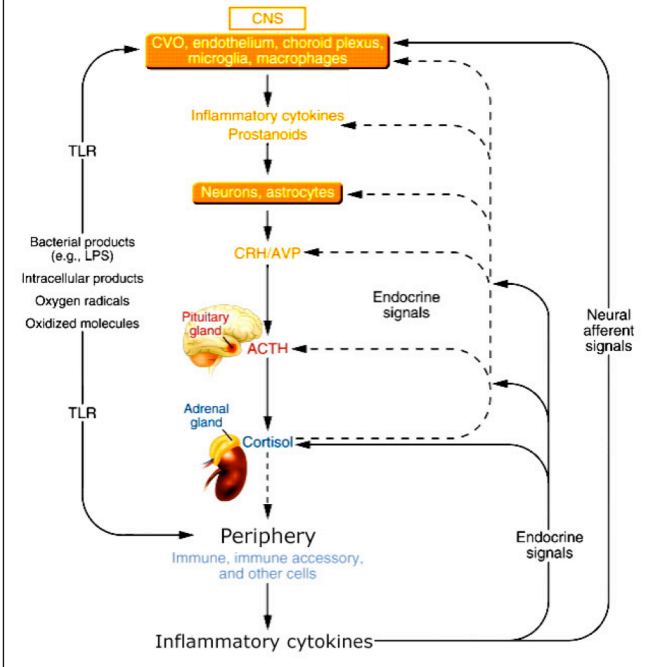



**Figure 1.** Relationship between HPA and immune system. Modified from Chrousos and Kino, 2005.

Affective states are truly emergent properties of this complex system, which has many different states in different individuals. Some of these internal network states result in extreme pathological conditions. The dominant paradigm of reductionist assignment of function X to biological element Y is a general approach that belies the true complexity of biological systems, which act in contexts of genetic and environmental perturbation. This should not be equated with a defeatist position—we can gain real understanding of the complex biological systems that underlie much of the initial stress response, but success will fundamentally require a new synthetic approach that models and tests the whole complex structure simultaneously rather than each bit in isolation. For example at the CNS level, stress is a complex emergent property and no one molecule —even cortisol, CFR, or ACTH—will be an accurate surrogate or endophenotype. Response will depend on context and on an individual's unique genomic constitution. As in humans, what is stressful to one strain may be pleasurable to another and neutral to a third. By examining the complex interactions among stress responses across hundreds of genetically defined individuals, we will have an objective measure (covariance of traits and correlation between traits and alleles) to decide which parts of the genetic and biological circuitry covary most closely with stress, however we define it—corticotropin levels, thymus introgression, or even white blood cell counts.

Despite all we know about the epidemiological links and biological consequences of stress, we know little about the early molecular and physiological changes induced by stress that set the body on a course for elevated susceptibility to many types of diseases or how these early events are inter-related among highly diverse disease processes. The elucidation of these early biological changes, along with the identification of stress-related biomarkers, is a major aim of the ICOGS. This will be achieved through an integrated molecular and computational dissection of a generalized stress response.

**Systems Genetics: a new scientific discipline at the nexus of classical and quantitative genetics, systems biology, and the computational sciences.** In order to properly examine the effects of stress on biological systems and disease susceptibility and to elucidate the inter-relationships between biological effects,

a variety of disciplines that have largely developed independently will need to be assimilated to form a new interdisciplinary field (Table 1). The team that has been assembled to achieve this goal comprises many of those considered to be leaders in developing the ideas behind a new field called systems genetics. This field will evolve through the anchoring of systems biology approaches and technologies onto fixed maps of genetic polymorphisms, which will support the identification of the critical variable biological network nodes responsible for differential susceptibility to stress all the while supporting the identification of co-modulated networks of genes and phenotypes. Systems genetics will be enabled by the concurrent introduction of new statistical tools for multivariate analysis and introduction of methods from other data-intensive content fields. The goal will be to develop empirically supported functional networks on the foundation of novel data architectures and data mining tools that are robust, quick and extendable. The results can only be fully appreciated through novel visualization and presentation portals that incorporate the latest Web-based interaction tools. In this application, we lay out our vision for achieving this integration through the combination of biological and computational sciences represented in the ICOGS.

**Table 1.** Disciplines represented by PIs of ICOGS projects.

| *Discipline* | *ICOGS Investigators* |
|---|---|
| Mammalian genetics/genomics | Threadgill, Williams, Wakeland, Chesler, Manly, Pardo Manuel, Wakeland, Churchill |
| Metabolism | Voy, Threadgill |
| Neuroscience | Chesler, Williams |
| Behavior | Chesler, Williams |
| Cardiovascular biology | Smyth |
| Immunology | Wakeland |
| Environmental exposure | Threadgill, Voy |
| Statistics | Churchill, Manly, Chesler |
| Graph theory | Langston, Wang |
| High performance computing | Wang, Langston, McMillan |
| Complex data visualization | McMillan, Williams, Manly |
| Evolution | Pardo Manuel |

**Predictive medicine**. Predictive genetics has been practiced for over three decades at an elementary level. Precedence was set by universal screens of neonates for mutations in PAH that lead to phenylketonuria (PKU, incidence of 1: 15000) and severe cognitive impairment. The utility of a simple dietary adjustment—avoidance of phenylalanine—is now an icon of the power of gene-by-environment interactions and the utility of knowledge of biochemical pathways to modify disease progression even of strong Mendelian traits with high heritability.

We are now in a second phase of genetically tailored medicine that directly affects much larger populations. Effective drug dosage for diseases ranging from childhood leukemias, to HIV and congestive heart failure (CHF) depend on information about variants of the cytochrome P450 genes and the *ABCB1* transporter. A single SNP in exon 26 of *ABCB1* is associated with a 2-fold difference in protein expression and membrane trafficking of xenobiotics such as protease inhibitors (HIV), digoxin (CHF), vincristin and numerous other anticancer drugs (Schwab *et al*., 2003). Armed with this information, the “standard” trail and error “titration” approach to drug dosage is slowly giving way to direct prediction of best treatment regime based on genomic and somatic genotype, age, and history.

The third phase of predictive medicine will rely on whole genome analyses (Collins *et al*., 2004). A survey of 1 million SNPs is close to the prognostic and therapeutic holy grail of the $1000 personal genome (Church, 2006). Comprehensive sequence data will be coupled with massive transcriptome, proteome, and metabolome data sets for lymphocytes and biopsy samples. The trick will be converting these complex data sets into knowledge that can be used to improve health by modifying risk exposure, and when that fails, to tailor treatment for best outcome. However, many of these approaches that are still on the horizon are being pursued in isolated biological systems of limited complexity or in systems where experiments cannot be replicated or easily validated. Most importantly, they are being pursued in circumstances that do not permit the development of integrated models, models that would describe how information flow from nuclei of diverse cells and tissues to disease state is dependent upon other biological changes. The platform for interdisciplinary research we describe in this application is ideally suited not only to investigate the biological consequences of stress, but also to support the development of information-rich predictive biology.

*"We may, by deliberately varying in each case some of the conditions of the experiment, achieve a wider inductive basis for our conclusions, without in any degree impairing their precision."*

R. A. Fisher, 1935

**Previous efforts to investigate traits with complex etiologies**. To understand a complex system with many potentially interacting parts, we must perturb the system and observe and quantify the consequences. Under appropriate conditions, we can conclude that perturbations are causal. However, causal relationships are more difficult in situations in which we need to know about how environmental factors—diet, infection, and toxins—interact with gene variants. Gene-by-environment interaction (GXE) is a crucial problem in genetics, and a

major contributor to most common phenotypic and medical traits as described above, that has been difficult to study in any mammalian population. GXE analysis requires the use of isogenic strains that can be studied in large numbers in different environments, usually over a period of years. Mouse experimental geneticists usually have adequate control over many environmental factors, but have not had sufficiently large isogenic mapping panels.

In the first century of genetics, the basis for heritable phenotypes was established largely through an approach that involved perturbing biological systems one component at a time, whether using experimental model systems to isolate independent activities of polymorphic genes or exploiting spontaneous or induced "simple" Mendelian mutations. The success of this approach has laid the groundwork for much more efficient and effective approaches that recognize the subtlety and complexity of genetic regulation (Fig. 2). Perturbing a system one component at a time can completely miss effects that are manifest only through the interactions of multiple parts of the system (Hartwell, 2004). A one-at-a-time approach also does not scale well when faced with a genome consisting of 20,000 – 30,000 genes. Factorial experimentation, in which many components of a system are perturbed simultaneously, is both more efficient and more comprehensive (Fisher, 1935). The Collaborative Cross (described in the Preliminary Data section) achieves this goal by randomizing existing natural genetic variation sampled from eight diverse founder inbred strains, allowing us to observe gene effects in a multitude of combinations. A fixed set of genomes, interrogated by a community of researchers, will enable an enormous and systematic accumulation of data on the complex interplay of genes and environments that will support a previously untenable unifying theory of mammalian biology. This approach is ideally suited to investigate the influence of external stress on biological systems and the consequential changes in disease susceptibility.

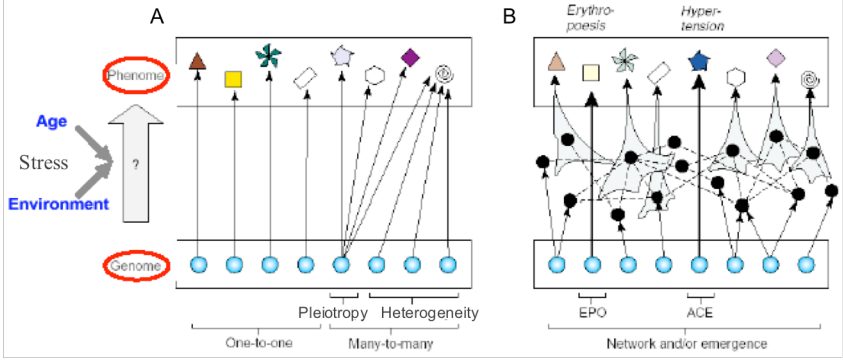



**Figure 2.** Gene-phenotype relationships. (A) Model used in studies over most of the previous century and (B) more accurate representation based on data from previous decade. Adapted from Huang, 2002.

## Preliminary Studies

In this section we present work that has led up to the ICOGS application and briefly describe resources, regents, unique expertise, and complementary projects that will be leveraged to investigate pleiotropic effects of stress at different scales and in different organ systems in a complex genetic milieu that more accurately models human populations.

**Complex Trait Consortium**. During the previous century, isogenic lines of mice became one of the major resources responsible for advancing biomedical and genetic research. Yet, with the development and refinement of powerful genetic tools including transgenesis (Gordon *et al*., 1980), gene targeting (Doetschman *et al*., 1987; Thomas and Capecchi, 1987), and ENU mutagenesis (Russell *et al*., 1979; Bode, 1984), the impressive power of inbred strains in functional genomics was overshadowed, but not forgotten. During a meeting at the 15th International Mouse Genome Conference in 2001, a group of geneticists interested in complex trait analysis, lead by Drs. David Threadgill and Robert Williams (ICOGS director and associate director, respectively), gathered to discuss potential community resources that could bring the concept of complex trait dissection much closer to reality (Threadgill et al., 2002). The Complex Trait Consortium (CTC; www.complextrait.org) emerged from this meeting with the goal to galvanize research into more complex biological questions and to develop an infrastructure to facilitate rapid advances.

The vast majority of phenotypes associated with human health, like stress-induced changes to disease susceptibility, are not caused by Mendelian variants but have complex multigenic and environmental etiologies. Mechanisms that underlie disease are, with few exceptions, influenced by numerous genetic, developmental and environmental factors. Since current resources are not sufficient to solve most complex biological systems where the mouse could contribute substantially, members of the CTC decided that a well-planned and specifically designed genetic resource was needed for analysis of complex and quantitative traits in mice (Threadgill *et al*., 2002; Churchill *et al*., 2004). Because the experimental reproducibility afforded by inbred strains permits measurements of even small variation in quantitative traits and enables virtual collaboration across time and space, the CTC also decided that the inbred strain, and derivatives thereof, would provide the foundation for future complex and quantitative trait dissection, fulfilling the promise originally espoused by luminaries like Clarence Little almost a century ago.

We believe that computational, statistical, and genomic resources are now sufficiently mature to address complex biological systems using models that more accurately reflect the genetic structure of human populations. Furthermore, global analysis of complex biological systems can be implemented most efficiently using experimental designs that employ multifactorial perturbations (Fisher, 1935; Jansen, 2003). A well designed, model organism resource for a new synthetic phase of genetic studies will have a direct, positive, and long lasting impact on many traits of medical relevance.

**Collaborative Cross genetic reference population.** With the realization that a new model population was needed to understand human diseases with complex etiologies, the Collaborative Cross (CC) was designed (Threadgill *et al.,* 2002; Churchill *et al*., 2004). The CC provides a translational tool to integrate single gene functional studies into genetic networks, which will be essential to understand the intricacies of biological processes as complex as altered disease susceptibility caused by differential response to stress.

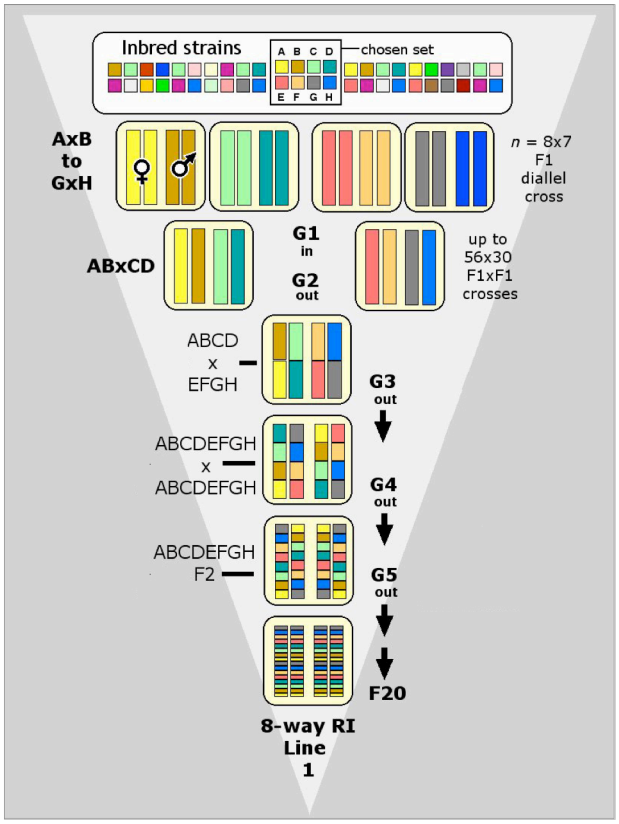
The CC is a large panel of recombinant inbred (RI) lines derived from a genetically diverse set of eight founder strains and designed specifically for complex trait analysis. Furthermore, since the CC is a genetic reference population that can be reproduced *ad infinitum*, it supports a model of community collaboration that will generate a comprehensive body of molecular and physiological data anchored on the common and reproducible CC reference population (Fig. 3). This is in contrast to current efforts exploiting the mouse as a model organism based on isolated and transient crosses. By providing a large, common set of genetically defined mice, the CC will become a focal point for cumulative and integrated data collection within the ICOGS, giving rise to a synergy of new

**Figure 3.** The CC allows direct and tight integration of multiscalar and interdisciplinary data sets. This is a resource that in combination with new analytic and synthetic quantitative approaches will catalyze an important change conceptualizing gene function in a more natural and far more realistic population context than single gene studies.

information and a new view of the mammalian organism as a whole and interconnected system.

The CC is an ideal genetic reference population. CC lines can produce unlimited numbers of genetically identical animals that can be exposed to a variety of experimental perturbations and interventions. Similarly, the genetic structure of the CC will enable the ICOGS to establish networks of functionally important relationships within and among diverse sets of biological and physiological phenotypes that can be altered by external stress. The CC lines have the potential to support studies by the larger scientific community incorporating multiple genetic, environmental, and developmental variables into comprehensive statistical models describing disease susceptibility and progression. Equally important, the CC can be used as a test bed for predictive, or more accurately, probabilistic medicine with the development of genetically defined hybrids between CC lines as described in Specific Aim 2.

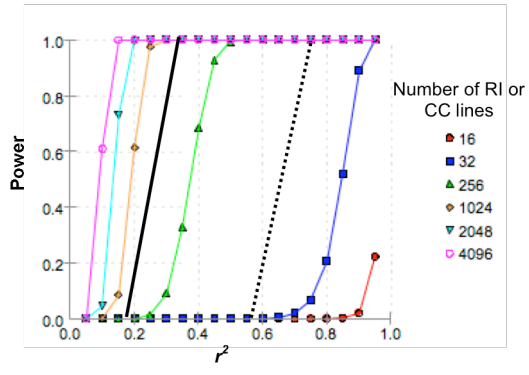
**Figure 4.** Simple structure of the derivitization of each CC line.




**Establishing the Collaborative Cross**. With support from The Jackson Laboratory, The Ellison Medical Foundation (EMF), and the Department of Energy, production of nearly 500 CC lines was begun in 2004 using a breeding scheme that ensures random allele segregation and balanced contributions of sex chromosomes and mitochondria. The CC is an unusual cross that combines the genomes of eight inbred strains of mice : A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO/HiLtJ, CAST/Ei, PWK/PhJ, and WSB/EiJ. These strains have been bred to each other using a combinatorial design to yield a large number of genetically independent hybrids (Fig. 4). Reciprocal matings were initially made between all viable pairs of strains at The Jackson Laboratory by Dr. Gary Churchill (PI, Project 9). F1 hybrids were distributed to ORNL, where pairs of the different F1 progeny were intercrossed, and a balanced subset of non-overlapping 4-way and 8-way progeny were subsequently mated. The final 8-way stocks were sib mated to produce unique incipient CC lines. Details on recombination, inbreeding rates, and statistical power of this complex cross have been described (Broman, 2005; Valdar *et al.*, 2006). The ultimate goal of the CTC has always been to initiated all 1680 (8x7x6x5) independent lines, and retain as many of these lines as prove viable. The primary reason for needing to produce such a large set of lines is the power achieved for trait and phenotype correlation when there are many random perturbations of allele combinations (Fig. 5). However, for the purposes of the ICOGS we will rely on the first wave of 500 CC lines now in progress at ORNL, which will provide sufficient power to detect statistically significant correlations greater than $r^2 = 0.25$.

A major new DOE resource—the Russell Vivarium at DOE's Oak Ridge National Laboratory (ORNL) will be leveraged by the ICOGS. This 36,000 ft$^2$ specific-pathogen-free facility is fully AAALAC-accredited and has a capacity over 20,000 cages, with the major use currently being the production of the CC. ORNL has been performing large and complex mouse-genetics experiments for six decades. With this new DOE User Facility, ORNL can offer vivarium, laboratory guest space, and expert technical support. The per diem charges are far lower than those of most institutions, a fact that reflects substantial cost sharing with DOE (~$0.29/cage/day, overhead included). The EMF grant to Drs. Dabney Johnson (Participant, Project 2) and David Threadgill (ICOGS director) now provide reliable funding for much of the per diem costs associated with the CC breeding, including all work that has lead up to this application (over 4 years of planning and two years of breeding). Both the DOE and

**Figure 5.** Relative power of varying sizes of RI or CC lines for detecting correlations. The largest RI set currenlty available is about 75 lines, which can be used to detect $r^2 = 0.75$ at 0.8 power (dotted black line). The 500 CC lines that will be used in the ICOGS project can detect $r^2 = 0.25$ at 0.8 power (solid black line).



EMF support are examples that qualify as non-NIH "parent" awards (to use the term of RFA-RM-06-008). Both awards have annual budgets through 2010 in excess of $250,000, and both are leveraged to achieve the goals of the ICOGS without overlap. Both of these non-NIH awards support the initial stages of the CC and but neither supports specific scientific applications as will be proposed by the ICOGS.

**Tracking progress and use of the Collaborative Cross.** Driving the implementation of the CC is a software system called the Collaborative Cross Database (CCDB) developed by Dr. Kenneth Manly (PI, Project 11) for use in the ORNL colony by the husbandry team. CCDB consists of a MySQL database and a set of Python scripts that provide a web-based interface (Fig. 6). The scripts create a user interface that is integrated with the workflow of weaning one generation and setting up matings to create the next. CCDB was designed with the following goals: a) to ensure the randomization of progeny selection and mate choice at each generation, b) to allow data entry as mice are weaned and mated, c) to minimize data entry time and data entry errors, d) to monitor and maintain data integrity, and e) to allow data entry, monitoring, and reporting from multiple locations. Manual data entry into CCDB is minimized by batch loading of G1 mouse information and mating assignments. These mating assignments are generated by a custom Python script that rapidly tests thousands of assignments in an attempt to maximize the number of lines that can be created from a given set of available G1 mice. This script generates tab-separated files that can be imported directly into CCDB, assuring accurate records of mouse information and mating assignments.

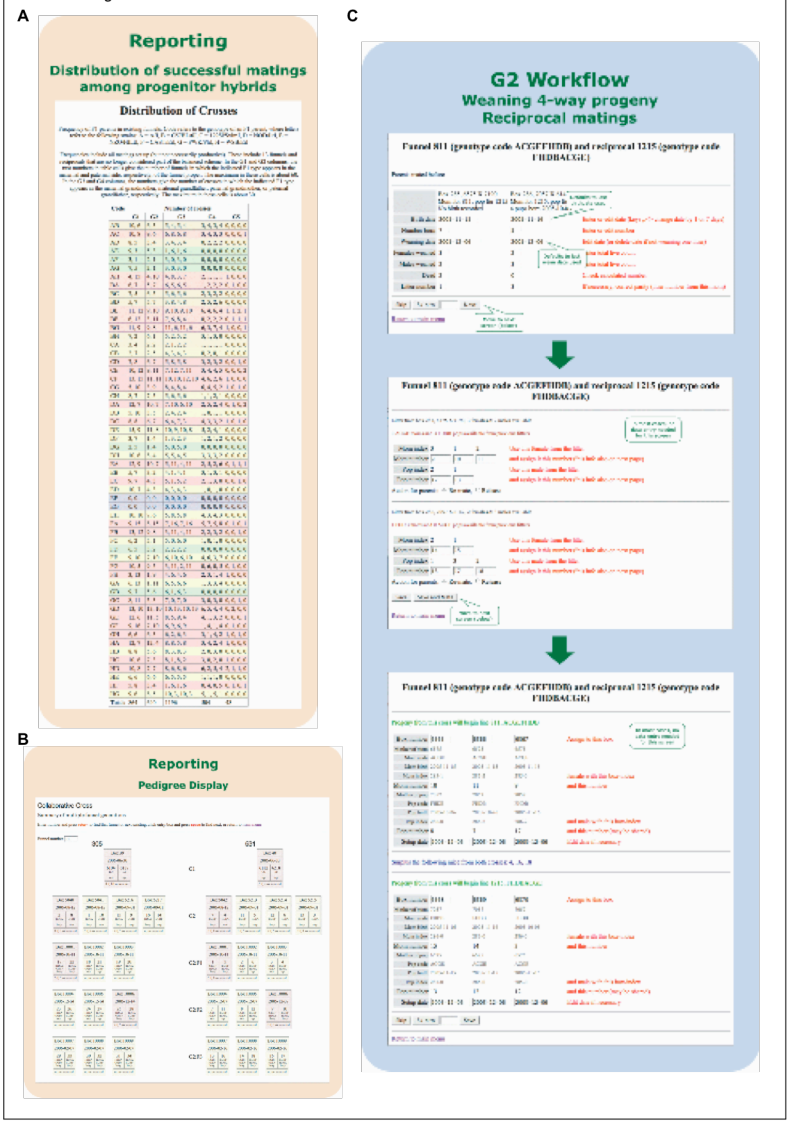



**Figure 6.** Screen shots from the CCDB showing reports for (A) status of matings from parental F1 hybrids and (B) individual funnels for each CC line. (C) Representative workflow summary and mating instructions at the G2 stage.

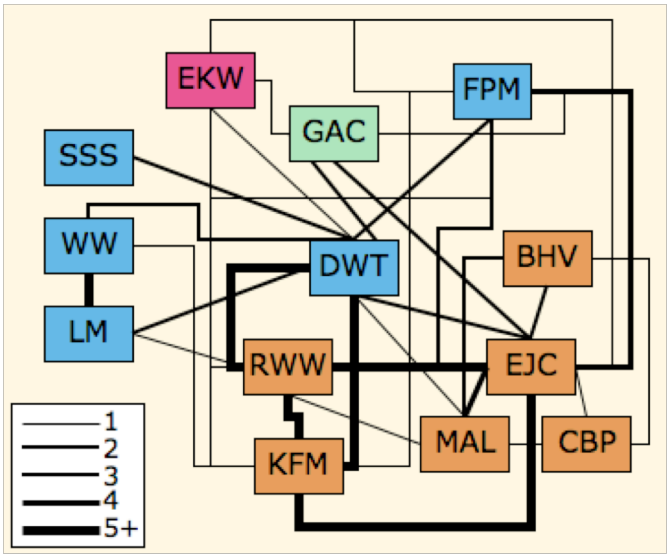
Team members use laptop computers directly at the station where they are setting up matings. Each generation of the cross has a specific set of routines to ensure optimal use of breeding capacity. The actual individual mice chosen for mating from within each litter are randomly selected using the software, preventing unintentional selection, like ease of handling. This mating assignment is substantially different than all other RI lines and mouse strains developed to date; most previous lines had significant human selection during their derivation. The system allows entry of litter size, sex ratio and other parameters at weaning. The result is a population of lines bred to retain genetic diversity and balance, with sufficient constraint to allow tractable genetic analysis of multi-factorial disease.

**Alternative approaches.** Inbred strains of mice harbor a tremendous amount of natural variation and it has been suggested that the variation among extant strains can be used for *in siico* (haplotype association) mapping, obviating the need for genetic crosses or RI lines like the CC (Grupe *et al*., 2001; Pletcher *et al*., 2004). However, several facts limit our ability to exploit existing inbred strains as a reference panel. Primarily, the number of readily available strains is too small to provide reliable statistical support for most genetic effects and phenotypic correlations. Moreover, the genetics of these mice is confounded by a complex and uncertain history and inter-relatedness that prevents inference of causation due to strong patterns of linkage disequilibrium (Petkov *et al*., 2005; Mhyre *et al*., 2005). Controlled randomization of genetic factors, essential to causal inference, is achieved in the construction of the CC. The common occurrence of transgressive

segregation and novel traits, seen only in the progeny of crosses, indicates that a vast potential for phenotypic diversity is hidden away in the common inbred strains and that this diversity can be expressed when these genomes are mixed in many different combinations and minimizing the breeding selection that frequently accompanies derivation of new inbred lines.

**Assembling the ICOGS team.** The group of investigators in the ICOGS was selected for their individual expertise, willingness to work in a new interdisciplinary field, and interest in complex biological systems of medical relevance. In fact, there already exists a significant track record of collaboration among the project PIs, demonstrating their interest and more importantly, eagerness to collaborate within the ICOGS (Fig. 7). Many of the ICOGS members are active participants in the development of major resources or experimental programs that will be leverage by the ICOGS project PIs, including: a) the new Department of Energy-Oak Ridge National Laboratory vivarium and user facility, b) the Complex Trait Consortium (CTC), c) the Carolina Center for Genome Science, d) the Renaissance Computing Center at UNC and the UT-ORNL Supercomputer facilities, e) links to major bioinformatics projects including the Biomedical Informatics Research Network (NCRR), Mouse Models for Human Cancer Consortium (NCI), Integrative Neuroscience Initiative on Alcoholism (NIAAA), Human Brain Project/GeneNetwork (NIMH, NIDA, NIAAA), and a new Center of Excellence for Chromosome Dynamics (NIGMS).

**Figure 7.** Web of establlished interactions between ICOGS project PIs. Each co-authored paper or active/pending grant shared by investigators is counted as an interaction. EJC, Elissa J. Chesler; GAC, Gary A. Churchill; MAL, Michael A. Langston; LM, Leonard McMillan; KFM, Kenneth F. Manly; FPM, Fernando Pardo Manuel; CBP, Cynthia B. Peterson; SSS, Susan S. Smyth; DWT, David W. Threadgill; BHV, Brynn H. Voy; EKW, Edward K. Wakeland; WW, Wei Wang; RWW, Robert W. Williams. Color-coding is by institution.



## Research Design and Methods

The ICOGS has been assembled to ensure effective interdisciplinary synergy. Many of our ICOGS members have already worked together – biologists, computer scientists and statiticians (see Fig. 7). Our collaboration has lead to a new integrative scientific discipline that we call systems genetics. We will drill down on a specific area of research – namely, the broad spectrum of disease and morbidity associated with high stress levels. We have laid out our vision of systems genetics in the Background and Significance section and in this section, we provide a specific collaborative plan. The twelve components that comprise the ICOGS are listed in Table 2 (color coded by institution).

**Table 2.** Projects, predicted funding levels and PIs for the ICOGS application. Projects are color-coded by institution. The qualifications of each PI are not elaborated upon in the text. Please see the associated biosketches, which documents their expertise through the published literature.

| Project | Type | Support | Title | PI | Institution |
|---|---|---|---|---|---|
| 1 | U54 | $150,000 ($300,000 Y02+) | Interdisciplinary Consortium on the Genetics and Co-Morbidity of Stress | Threadgill | University of North Carolina |
| 2 | P30 | $800,000 | Collaborative Cross Reference Population and Bioinformatics Core | Williams & Chesler | University of Tennessee Oak Ridge National Lab |
| 3 | R01 | $200,000 | Systems Genetics of the HPA | Chesler | Oak Ridge National Lab |
| ~~4~~ | ~~R01~~ | ~~$200,000~~ | ~~Metabolic Consequences of Stress~~ | Threadgill & Voy | University of North Carolina Oak Ridge National Lab |
| 5 | R01 | $200,000 | Cardiovascular Disease and Stress | Smyth | University of North Carolina |
| 6 | R01 | $200,000 | Immunological Changes Induced by Stress | Wakeland | University of Texas Southwestern |
| 7 | R01 | $150,000 | Epistasis and Drift in the Collaborative Cross | Pardo Manuel | University of North Carolina |
| 8 | R01 | $150,000 | High Performance Computing for Data Mining and Modeling | Wang | University of North Carolina |
| 9 | R01 | $150,000 | Advanced Statistical Modeling of Complex Genetic Networks | Churchill | The Jackson Laboratory |
| 10 | R01 | $150,000 | Graph Theory and Clique Analysis of Multiscale Networks | Langston | University of Tennessee |
| 11 | R01 | $200,000 | Visual Data Portal Interfaces | McMillan | University of North Carolina |
| 12 | R25 | $100,000 | Systems Genetics Workshop and Outreach | Manly | University of Tennessee |
| 13 | T90/ R90 | $150,000 | Interdisciplinary Training Program in Systems Genetics | Peterson | University of Tennessee |

**Specific Aim 1: Institute a governance plan**. To successfully implement the ICOGS requires a system that incorporates and melds each discipline. This requires steady communication, collegial interactions and co-authored publications. We can point to a number of recent examples of joint publications and collaborations among our members, docmenting our interactions. Our NIH biosketches demonstrate substantial interactions among card-carrying computer scientists, statisticians and biomedical researchers with wide ranging interests. Prior experiences with other successful large-scale projects has shown us the importance of a) an active, experienced team of senior leaders who set directions and policy, b) an engaged day-to-day consortium director and coordinator, and c) well-defined research, integration and dissemination plans. We have been surprisingly successful at these objectives even without the benefit of NIH funding by using the Complex Trait Consortium as an informal catalyst for collaboration and as a virtual U54 pilot. But ICOGS will allow us to achieve substantial and specific gains using a formal administrative structure, data sharing protocols, meetings, and pilot projects not acheivable by any other mechanism. The foundation supporting these activities will be the parent U54 consortium project.

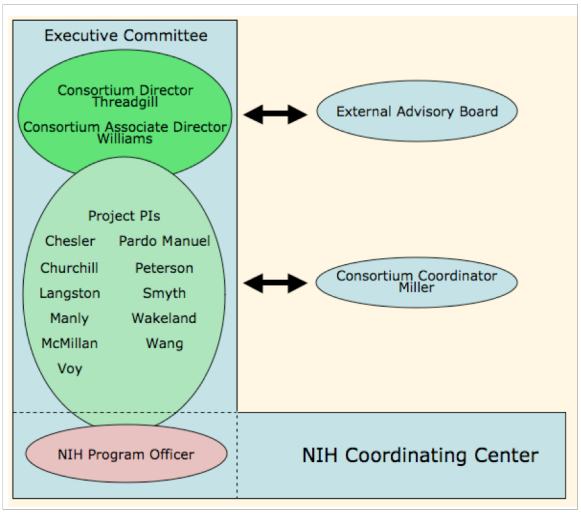
**PROJECT 1: (U54) Interdisciplinary consortium on the genetics and co-morbidity of stress**

*PI, Consortium Director:* Threadgill (UNC)
*Co-I, Consortium Associate Director:* Williams (UT)
*Consortium Coordinator:* Miller (ORNL)
*Executive Committee (EC):* Component PIs and NIH Program Official
*External Advisory Board (EAB):* To be appointed.

This project will administer and oversee the ICOGS to ensure the overall goals are met. The ICOGS administrative team will consist of experienced leaders in the different disciplines contributing to the development of systems genetics, including classical and quantitative genetics, evolutionary theory, physiology, molecular biology, high-performance computing, graph theory, and data visualization among others (Table 1). The EC will be complemented by a consortium coordinator and an EAB to provide requisite strategic advice and guidance, serving as *eminences grises*, with counsel on long-term policies and approaches, inter-organizational interaction and technical approaches. The specific aims of the U54 will be to:

*Aim 1: Develop an administrative structure to support interdisciplinary investigations into the effects of stress.* The ICOGS will be organized around an EC, consisting of the PI from each component part and an NIH Program Official (Fig. 8), that will provide advice on the overall direction and strategy. There will be weekly conference calls of the EC to evaluate progress and to identify, as early as possible, any changes in experimental plan to ensure the continued success of the ICOGS. The EC will be chaired Dr. Threadgill, who will commit a minimum of 35% effort leading the ICOGS. He will also serve as the point of contact with NIH for all consortium matters, independent of those that are project specific, and have final authority and responsbility for all management decisions. Dr. Threadgill has experience assembling groups of investigators to collaborate across disciplines. As director of the Transgenomics Research Core within the NIEHS-sponsored Center for Environmental Health and Susceptibility, he has been responsible for generating interactions and collaborations between scientists focusing on different –omic technologies and bioinformatics. This research core provided critical support for two successful applications, one focusing on the integrative genetics of cancer susceptibility, on which Dr. Threadgill serves as PI as part of the NCI-sponsored Mouse Models of Human Cancer Consortium (MMHCC) and an NIAAA-sponsored grant proposing to combine metabolomic and genomic data to investigate the genetic mechanisms of ethanol toxicity. The associate director of the ICOGS will be Dr. Williams, who is a co-investigator on the MMHCC grant and also has extensive leadership experience, being director of the Informatics Center for Mouse Neurogenetics (a Human Brain Project), and Coordinator of the UT Center of Genomics and Bioinformatics. In the event that the director needs to be replaced, the EC would meet and with the input of the EAB, review the credentials for a new director, which will be recommended to NIH. At this time, the leadership has agreed that the associate director, Dr. Williams will serve as the new director pending approval from NIH.

**Figure 8.** Organizational structure of the ICOGS.




An EAB will be established to provide unbiased, external advice to the consortium director and EC on matters dealing with the organization and inter-project interactions within the ICOGS. The EAB will include experienced and respected leaders, with technical expertise in biology, medicine, educational programs, mathematics, computing and information exchange. The EAB's experience base will span academia and government, providing a cross-section of the management and community interactions expected of the ICOGS.

Previous experience with other large-scale projects has shown that success depends on both a strategic research vision but also on the vision's operational implementation. Consquently, Ms. Miller will serve as the consortium coordinator to ensure the coordinated implementation of the ICOGS strategic plan. The consortium coordinator will report directly to the consortium director and will have responsibility for day-to-day activities of the ICOGS. Additionally, the consortium coordinator will provide weekly reports to the EC. Ms. Miller brings a strong record of organizational skills having served as the Business Manager for the International Mammalian Genome Society (IMGS). Through these activities, she has also organized the annual IMGS meeting that regularly draws 250-350 attendees. Ms. Miller will also draw on her experience as Central Coordinator of the Tennessee Mouse Genome Consortium, which recently completed a large, multi-institutional mouse mutagenesis program that required significant reagent and data transfer among institutions.

The consortium will also be supported by a full-time administrative assistant who will be cost-shared between NIH and UNC and whose role will be to assist the consortium director and coordinator and ensure timely communication among component projects. This person will organize conference calls, arrange

meetings of the EC and EAB, assist in developing ICOGS reports, coordinate travel and provide necessary administrative support.

*Aim 2: Organize an annual retreat for all ICOGS participants.* An annual retreat, organized with Project 12 (Specific Aim 4), will be held for all ICOGS participants and pilot grant awardees. Attendees will also include the EAB and appropriate NIH program staff. This two-day meeting will be held at ORNL and will focus on project integration and progress towards the scientific goal of the ICOGS. There will be talks from each project highlighting the interfaces between projects and new knowledge that has been developed providing insight into stress responses. There will also be an ICOGS round table discussion to provide feedback to the EC on organizational structures that may be impacting, either negatively or positively, the interaction among projects and disciplines. The meeting will conclude with an executive session attended by the consortium director and associate director and the EAB.

*Aim 3: Establish a pilot project program.* A pilot project program will be developed with the first awards made in year 2 of the ICOGS. This program will be announced yearly and will predominantly focus on providing support for non-ICOGS investigators to exploit the CC and the data and tools that will be developed by ICOGS projects. Pilot projects focusing on biological areas impacted by stress or computational and analysis methods not represented by ICOGS projects will have highest priority. The primary goal of the pilot projects will be to provide sufficient preliminary data using CC parental strains or their hybrids (as described in Project 2) to support a competitive grant application that uses the CC or the development of new computational, modeling, statistical analysis or data structures that would further enhance the use of the CC to investigate the biological and physiological consequences of stress. The pilot project program will be advertised through various sources like the Mouse Genome Informatics list-serve, CTC, and relevant medical, statistical and computer science list serves. Applications will be reviewed by the EC and the EAB to identify those with the highest probability of providing scientific breadth to the ICOGS. The pilot project program will also support travel to ORNL or other ICOGS sites for collection or analysis of data.

**Specific Aim 2: Develop key resources and a conceptual framework for systems genetics.** Complex biological questions require a robust and innovative model system. A research core will be provide a central resource to study stress in the context of highly diverse cell, tissue and organismal responses across a wide variety of genotypes with sufficient diversity to model human populations. A bioinformatic infrastructure will also be built to support data collection, curation, error-checking, integration, and distribution. In an era where biomedical science is advancing rapidly, our ability to perform integrative research *in vivo* is still hampered by inappropriate research tools and resources. As described in the Background and Significance section, the CC has the potential to fill this void.

**PROJECT 2: (P30) Collaborative cross reference population and bioinformatics core**

*Co-PIs:* Williams (UT) and Chesler (ORNL)

*Additional participants:* Threadgill (UNC), Manly (UT), Churchill (JAX), Dabney Johnson (ORNL), Miller (ORNL)

The CC will be the focal point of this research core project. It will be responsible for breeding, genotyping, initial characterization, and distribution of tissue and cell samples to all research projects. This Core will also apply a uniform model stressor that combines psychological (social isolation) and environmental (dietary modification) stress, which is of greatest relevance to current trends in Western life styles. A bioinformatics component will also be developed to track breeding and samples, to acquire and host all ICOGS biological data from Projects 3-7 and computational programs from Projects 8-10, and educational tutorials developed in Project 12. Finally, this component will build and maintain the ICOGS web-portal in collaboration with Project 11 to support integrated consortium activities and information distribution to the broader research and lay communities. The proposed specific aims will be to:

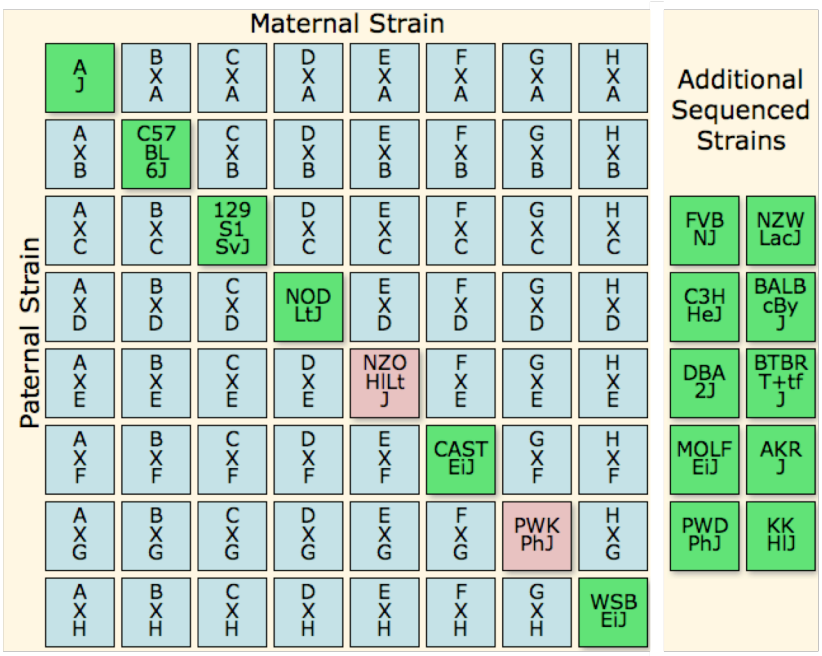
*Aim 1. Establish the Collaborative Cross as a genetic reference population.* The new 20,000+ cage Russell Vivarium at ORNL will be used to produce, maintain and use the CC for the goals of the ICOGS. As described in the Preliminary Data section, funding from DOE and EMF for initial breeding of the CC is being leveraged to develop the CC resource. Although the ultimate goal of the CTC has always been to initiated all 1680 independent CC lines, and retain as many of these strains as prove viable, for the purposes of the ICOGS we will rely on the first wave of 500 inbred CC lines being developed now at ORNL.

As of April 2006, approximately 450 CC lines have been initiated and several of these have already progressed to the fifth generation (G3F5) of inbreeding. An addition 100 lines will be started in mid 2006. By the start of the ICOGS award, late 2007 of early 2008, a significant number of the CC lines will have progressed to 12 generations of inbreeding (~G3F11–13), and will be available for high density genotyping.

When the CC lines reach F12 or greater of inbreeding (>90% inbred), marker-assisted selection will be used to accelerate the final stages of inbreeding over the first two years of the ICOGS (Wakeland *et al*., 1997). SNP marker selection and genotyping will be performed in collaboration with Project 7. We will be able to provide all research teams with essentially isogenic lines by late in 2009 or early in 2010, corresponding with the completion of parental line characterization as described below.

*Aim 2. Develop a sample pipeline to support experimental investigations into the biological consequences of stress.* Genotypes and samples that support biological analyses described in Specific Aim 3, will be from three sources. During the first two years of the ICOGS, a comparatively large panel of highly diverse inbred strains will be phenotyped for stress response (Fig. 9). These will consist of: a) the 8 CC parental strains and the 15 strains being resequenced by NIEHS (18 total strains since there is overlap in the two sets) and b) all reciprocal F1 progeny from the 8 CC parental strains in a complete diallel cross. There are several excellent reasons to being our study of stress using these resources: a) these lines are all completely isogenic, b) they represent strains that have been very well characterized for many traits as part of the Mouse Phenome Project and other phenotyping efforts, and c) they include the first two generations of the CC—the parental and entire F1 generations, providing a framework for phenotypic diversity.

**Figure 9.** Representation of lines and hybrids that will be analyzed in year 1 and 2. Green (NIH and NIEHS sequenced lines) and pick (non-sequenced) parental strains and blue reciprocal F1 hybrids from CC parental lines.



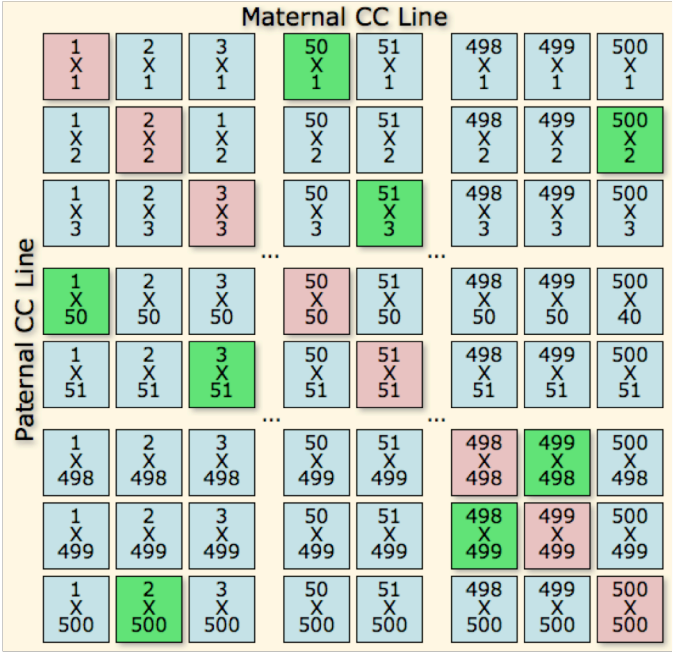

As described in the Preliminary Data section, the ideal genetic reference population would consist of a large number of isogenic but non-inbred lines. We can achieve this goal easily by making F1 hybrids, also known as RI intercross (RIX) progeny between CC lines (Fig. 10). With a set of 500 CC strains, almost 250,000 non-reciprocal, unique isogenic RIX can be generated. The major animal resource that will be used and shared by all components of the ICOGS is a subset of 250 reciprocal RIX progeny that will be produced beginning late in year 2 by intercrossing the 500 CC lines such that each CC line only contributes to a single RIX, ensuring all are genetically equidistant. The particular selection of CC lines chosen for the production of these 250 RIX will be made in Project 9. The remaining potential RIX will then be used to validate predictive models describing molecular and physiological responses to stress that will be developed within Projects 3-6. The RIX method has been extensively described and tested, both theoretically and empirically, by members of the ICOGS (Zou *et al*., 2005; Tsaih *et al*., 2005).

Progeny from each hybrid or strain will be weaned at approximately 21-days-of-age and group housed by sex by strain. At 60-days-of-age, all mice will be housed in isolation and their consumption of food monitored for one week. Half of the mice will then be exposed to combination stressor of social isolation (individually housed) and a high-fat diet (HFD), and half will be group housed using a conventional diet to which mice were weaned. This combination stressor is not meant to test a specific type of stress, but rather to represent a generalized state of stress, experienced by most Americans. Stress exposed and control mice will be subject in series to three broad protocols: a) behavioral assays of anxiety and response to conventional behavioral stressors, b) a cardiovascular function assay (echocardiogram), and c) terminal analysis to acquire tissues and cells, including hippocampus, hypothalamus, kidney, adrenal gland, thymus, spleen, heart, peritoneal fat, jejunum, lymphocytes, and serum. Most of the assays that will populate the ICOGS database will be high-throughput mRNA, metabolite, and cell ssays of littermate control and stressed mice. These data will be complemented by strain genotype at one extreme and behavior data at the other extreme. As quantitative proteomic methods mature in the next few years we will also test whether we can capture data on protein level

and modifications. For this purpose, aliquots of cells and tissues will be set aside for prospective analysis by other investigators and pilot projects.

*Aim 3. Design, implement and maintain an ICOGS web portal.* An essential component of the ICOGS will be data tracking, integration and distribution. Several members of the ICOGS have significant experience from a previous large-scale ENU mutagenesis project in which mice mutagenized and bred at ORNL were distributed for phenotyping throughout Tennessee. Mouse husbandry, shipping, tracking, data storage, and analysis were all achieved using the MuTrack system, which features an Oracle database with SAS statistical and analytic tools. We will modify some aspects of the database schema to improve its integration with the CCDB (described in the Preliminary Data section). Individual animal phenotype data and breeding history will be integrated and displayed using a large open source set of tools for downloading and statistical comparison. While the implementation details are obviously not yet set in stone, we are enthused about the use of the Resource Description Framework (RDF) extension of XML (part of the Semantic Web initiative) as a method and convention to make ICOGS data much more useful for distributed data processing and querying.

**Figure 10.** Generation of CC samples for biological analysis. Pink represents parental CC lines, green are the CC RIX hybrids used for characterization, and blue are the remaining CCRIX hybrids that can be used for predictive phenotype testing.




The ICOGS web portal will also provide a repository for posters and talks presented by ICOGS member laboratories and a forum designed to manage thread-oriented data presentation and discussion. This forum will also be the site through which ICOGS members present background material in preparation for the monthly interdisciplinary clinics (described in Project 12), and it will provide a center for discussion of topics presented there. The forum will include an issue tracker to record and publish suggestions or concerns, from any ICOGS member. Whereas the forum will take the form of articles followed by comments from readers, the ICOGS web portal will also provide for less structured communication. Blogs or wikis for each of the project areas will provide within project communication, but also allow members of other projects to stay informed. Like the forum and interdisciplinary clinics, these discussion tools will be restricted to scientists associated with the ICOGS.

The forum will be deployed using an existing open-source content management system. Several such systems are available. Ideally, the software we use for the forum would have the following characteristics: a) thread-oriented display of information; that is, information and comments related to a user-defined topic displayed in chronological order, with separate display and indexing of topics available, b) ability to display graphics and video and simple methods for including these in a posting, c) access control that would allow visibility and posting to be restricted to consortium laboratories but provide access automatically within that group, d) functions to allow users to comment on any topic or to start a new topic. Unlike a typical wiki site, this forum would not give users the ability to edit information posted by others, e) automatic email notification of new postings and comments on existing postings, with email lists controlled by both forum posters and potential recipients, f) functions to tag postings with keywords chosen by the poster or by other users, and the ability to index postings according to those keywords, and g) available source code and a design that facilitates customization and distribution to other interdisciplinary teams.

Finally, the ICOGS web portal will incorporate novel tools and data visualizations developed by projects described in Specific Aim 3 for use by the broader scientific research community.

**Specific Aim 3: Assemble a research consortium to use systems genetics to investigate the biological consequences of stress**. The ICOGS team will collect and integrate high-dimensional data sets using state-of-the-art tools and protocols. We will also develop new methods to analyze data streams in the context of causal linkage between stress and susceptibility-to-disease phenotypes.

**Integration across R01s.** We will have four R01 applications that focus on particular biological domains and their relations to stress (nervous system, cardiovascular system, metabolomsm, and immune function), but we recognize that the whole point of the ICOGS is to begin to merge results from these domains to assemble a more holistic model of the genetics of stress. This can be done in part by: a) appropriate selection of a shared model of stress that will be used by all R01s—in our case the isolation-diet model of generalized stress, b) the use of a common set of highly diverse isogenic lines of mice, including inbred parents of the CC and a fixed panel of isogenic CC RIX hybrids, and c) close collaboration of the lead scientists of the four biological projects with the five R01 from computational groups who will help to synthesize more complex and (we expect) more realistic models of multisystem and polygenic responses to stress.

**PROJECT 3: (R01) Systems Genetics of the Hypothalamic-Pituitary-Adrenal Axis**

*PI:* Chesler (ORNL)

*Additional participants:* Williams (UT), Langston (UT), Lu Lu (UT)

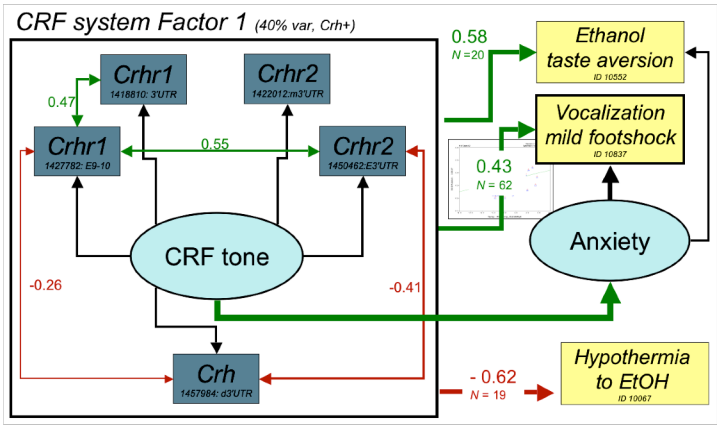
Homeostasis is a dynamic process, which is optimized through allostatic processes in response to internal and external conditions. Perturbations like drug addiction, or, as in the case of the ICOGS, stress, require high levels of compensation and place a long duration allostatic load on the CNS via the hypothalamic-pituitary-adrenal (HPA) axis (McEwen, 2004), which may not be easily reset. We suspect that post-traumatic stress syndrome, fibromyalgia, chronic fatigue syndrome, irritable bowel syndrome and other complex disorders represent maladaptive allostatic set points to generalized or specific stressors. Genetic variability determines the relative consequences of an allostatic load for different individuals. This genetic variability can be harnessed to construct biological networks underlying the stress response and concomitant co-morbid diseases.

In this component of the ICOGS, we will focus on the first limb of a crucial circuit involved in responses to stress—the HPA axis. The design of the experiment is simple—we will acquire multiple data sets, from mRNAs to behavior followed by an analysis of the inter-relationships between measurements and stress load, as determined by alterations in physiological, metabolic and immunological states. The aims will be to:

*Aim 1. Acquire HPA data sets across diverse genotypes in the context of a generalized stress.* We will acquire data on: a) anxiety levels (elevated circle maze, open field activity), b) corticosterone and other neuroactive steroid levels in blood and brain (hypothalamus, hippocampus, prefrontal cortex), adrenal weight, thymus weight, and c) measures of steady-state mRNA abundance in hypothalamus, pituitary, and adrenals. These large data sets will be combined with those collected in Projects 4-6 and evaluated using tools developed in Project 9 to identify common and organ-specific responses to stress.

*Aim 2. Combine multi-organ data sets to generate genetic-based models of HPA responsivity to stress.* To anneal data into a cohesive structure we will take advantage of genetic covariance using methods we have pioneered (Chesler *et al*., 2003; Chesler *et al.,* 2005). We will exploit multivariate statistical techniques including structural equation modeling (Fig. 11), Bayesian networks and large-scale combinatorial graph analytic approaches being developed in Project 10. Interesting relationships will also be evaluated with respect to allele co-adaptation, which will be investigated in Project 7 as a mechanism for organismic preservation of optimality in the system. The results should identify novel intersections between the regulation of stress-induced changes in the HPA with other physiological changes identified in Projects 4-6 that may alter disease susceptibility.

**Figure 11.** Example of using genetic covariance to study the role of CRF tone in modulating anxiety. Expression of *Crh, Crhr1*, and *Crhr2* were measured in 86 strains of mice. Colored lines are correlations between traits (green positive, red negative). Ethanol taste aversion and hypothermia are correlated to the first principal component of CRF tone.




*Aim 3: Validate and test predictions using CC RIX progeny.* The data resulting from an analysis of the parental and F1 hybrid lines as well as the subset of CC RIX described in Project 2 will serve as the "training" data set for model development. To test and cross-validate models, we will use a different subset of CC RIX progeny

representing unique genome combinations. Thus we will deliver sophisticated network models of the impact of genetic variation on the HPA-centered response to stress.

**PROJECT 4: (R01) Metabolic consequences of stress**

*Co-PI:* Threadgill (UNC) and Voy (ORNL)

*Additional participants:* Wang (UNC), Christoph Borchers (UNC)

The interplay between neurologic, metabolic and immunologic systems in response to stress is crucial in disease outcome. In particular, it is commonly acknowledged that the Western life style is driving the present epidemic of obesity and type 2 diabetes. We know gene-environment interactions are a major contributing factor since people from ethnic groups not previously exposed to the Western life style develop a higher rate of obesity once exposed.

Of particular interest are the adipokines, which mediate whole body consequences of stress, placing adipose tissue at the nexus of a stress response. Multiple types of stressors, including those in our generalized stressor, converge on a set of common signaling effectors in adipocytes and regulate adipokine release into the circulation. The net outcome to the individual depends largely on undefined genetic factors that control the complex interplay between stress, adipokines and end-organ targets. Predicting the response is further complicated by the fact that adipokine functions continue to emerge and expand beyond their roles in endpoints classically related to metabolism, such as insulin sensitivity. Leptin represents the best-described (but by no means the only) example; it was originally identified as an adipose-derived satiety factor that signaled energy status to the brain but is now recognized as a multi-functional hormone linking adipose tissue to a spectrum of systems including immunity, reproduction and neuroendocrine function.

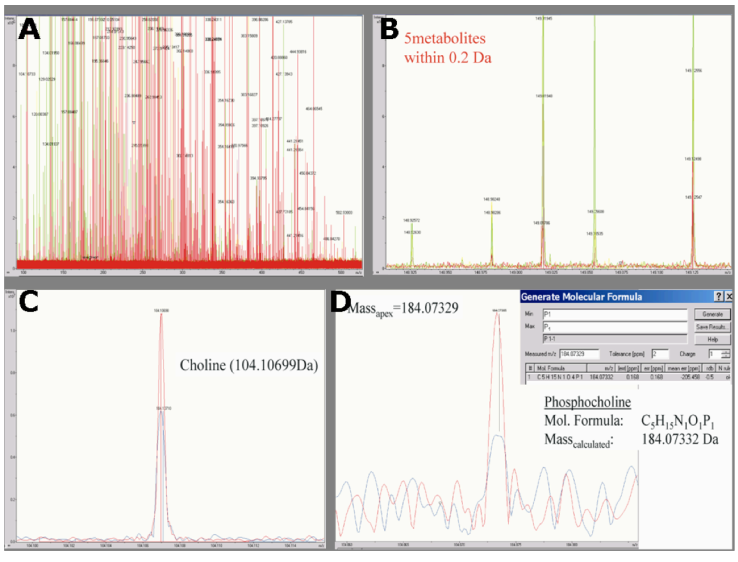
This project will investigate early changes in metabolic homeostasis elicited by a generalized stress. The aims will be to:

*Aim 1. Measure metabolic responses to stress.* Stress is known to elicit a wide range of metabolic changes. Using the combined generalized stress descrirbed in Project 2, the metabolic state of CC lines will be characterized before and after exposure. Basic metabolic parameters (fasting glucose, hemoglobin A1C, serum creatine, lipid and adipokine profiles) will be quantified as well as gene expression levels in liver, abdominal fat and jejunum. We will aslo collect serum metabolomic profiles, using a highly sensitive FTICR-MS (Fig. 12), and relative levels of gastrointestinal bacteria, by qPCR (Alexander *et al.,* submitted), which may be highly sensitive biomarkers of stress response. The data collected in this aim will be analyzed for genetic co-regulation similar to that described in Project 3.

*Aim 2. Identify links between metabolic dysregulation and changes in behavior, immune status and cardiovascular function.* Metabolic data will be analyzed with in the context of the data collected in Projects 3, 5 and 6 for co-morbid effects of altered metabolism on behavioral or cardiovascular phenotypes or coincidnce with altered immune system function. The metabolic changes will also be analyzed for relationship to co-adapted alleles identified in Project 7. All analyses will be supported by interactions with Projects 9 and 10.

*Aim 3. Develop predictive biomarkers of metabolic response to stress.* Using the large data sets collected in Aim 1, biomarkers and genetic variants will be identified that can accurately predict an individuals metabolic response to a generalized stress. These models will be validated and refined using independent cohorts of CC RIX mice.

**Figure 12.** FTICR-MS analysis of the metabolome (80-500 Da) in plasma. (A) Full-range scan and (B) a cluster of individual metabolites resolved within 0.2 Da mass range. When spectra from two samples (choline-deficient subject is shown in blue) are analyzed together, it is possible to quantitate (C) and identify (D) individual metabolites directly from mass analysis.




**PROJECT 5: (R01) Cardiovascular disease and stress**

*PI:* Smyth (UNC)

*Additional participants:* Threadgill (UNC), McMillan (UNC), Mauricio Rojas (UNC)

Environmental, emotional or psychological stress is commonly thought to contribute to the development of cardiovascular disease

(CVD), the leading cause of death in the United States. However, research in the field suffers from a perceived lack of pathophysiologically reliable mechanisms to account for stress responses, which is, in large part, a consequence of a poor understanding of the complex interactions of stress with genetically-mediated susceptibility. Our long-term goal is to determine the molecular mechanisms by which genetic susceptibility contributes to the effects of stress on the development of CVD and apply that knowledge to drive the prevention of CVD in humans.

We plan to accomplish the overall objective of this application through a two-tiered approach to analyze comprehensively the genetic susceptibility of mice to stress-induced changes in molecular and physiologic parameters and pathologic findings in heart, vessel, and blood. The specific aims will be to:

*Aim 1. Measure CVD parameters in response to stress.* The combinatorial effects of stress and genetic predisposition on cardiac function/integrity, measurements of cardiac structure/function (heart rate, chamber size, wall thickness, and left ventricular contractility), and histologic evidence of cardiac pathology (fibrosis and cardiomyocyte hypertrophy) will be determined. Similarly, stress and genetic effects on vascular function (*e.g.* sVCAM, E-selectin), physiologic measurements of vascular function (blood pressure and aortic compliance), and histologic evidence of vascular pathology (aortic atherosclerosis, fibrosis, calcification, and inflammation) will be determined. Measurements of blood thrombogenicity (*e.g.* sCD40L, sP-selectin, vWF, fibronectin, fibrinogen, PAI-1, homocysteine), platelet function (shear-induced thrombus formation), and histologic evidence of thrombosis (coronary artery thrombosis and thrombosis associated with atheroma) will be performed for comparison with physiological and morphometic measurements of cardiac and vessel function. Finally, the left ventricle will be collected for gene expression profiling to identify early molecular responses to stress.

*Aim 2. Determine molecular and physiological links to differenial sex response to stress.* Sex differences in physiological response to stress vary and most likely contribute to CVD prevalence. However, the lack of pathophysiologically reliable mechanisms to account for stress responses has hindered advances in understanding sex-specific responses. To identify sex differences in stress response that might be linked to altered CVD susceptibility, computational and statistical methods from Projects 9 and 10 will be applied to determine how genetic variants impact sex differences in pathophysiological responses in heart, vessel and blood to stress.

*Aim 3. Identify novel biomarkers and behavioral traits associated with heart, vessel and blood stress responses.* Tools and data mining approachces developed in Projects 8-10 will be applied in combination with data from Aim 1 and Projects 3, 4, 6 and 7 to determine how genetic variants interact with established CVD risk factors, including inflammatory biomarkers (CRP, MPO) and metabolic parameters (fasting glucose, hemoglobin A1C, serum creatinine, lipid profiles), to culminate in heart, vessel, and blood stress responses. In addition, the interaction of genetic variants with molecular profiles, inflammatory markers, metabolic profiles, and behavioral traits will be associated with heart, vessel and blood stress responses to identify common links to generalized stress.

**PROJECT 6: (R01) Immunological changes induced by stress**

*PI:* Wakeland (UTSW)

*Additional participants:* Churchill (JAX)

The view of the immune system has evolved substantially over the last decade, with the realization that it plays critical roles in many diverse diseases. As such, a detailed analysis of immunological status and function is critical to understanding how the body copes with stress and how a single stimulus can elicit such wide-ranging effects. To capture the effect of generalized stress on the immune system and inflammatory processes, like what occurs in obese individuals, a global analysis will be performed on the major components of both the innate and acquired immune system. The aim of this project will be:

*Aim 1. Quantify changes in immune status caused by exposure to stress.* General inflammatory biomarkers (CRP, MPO) will be quantified along with detailed cytokine profiling using a Luminex BeadSorter. Lymphocyte subpopulations will be measured by FACS. Since NK-cells appear to have a central role in stress response that may modulate effects in other tissues, we will perform an analysis of NK-cell number and function. High dimensional gene expression profiling will be performed in lymphocytes, spleen and thymus to capture the molecular state of the immune system. The resulting data sets will be analyzed using approaches that will be developed in Projects 9 and 10 to identify co-regulated components of the immune system and to anchor the regulators to the genetic map of the CC lines.

*Aim 2. Identify links between immune system changes and behavioral, metabolic, and cardiovascular homeostasis.* Using out put from Aim 1, we will work with Projects 3-5 to investigate causal links between

changes in immune response after stress and phenotypic changes indicative of early disease processes. One goal of this aim will be to determine whether inflammatory biomarkers can predict phenotypic changes in behavior, metabolism or cardiovascular homeostasis.

*Aim 3. Develop predictive models for immunological response to stress.* A major goal of the ICOGS is to use the data generated from a generalized stress response to develop predictors based upon either biomarkers or genetic variants. This will be an iterative process using the original CC RIX training set and then an independent CC RIX test set. The goal will be to develop accurately predictors of the inflammatory response to a generalized stressor.

**PROJECT 7: (R01) Epistasis and drift in the collaborative cross**

*PI:* Pardo Manuel (UNC)

*Additional participants:* Churchill (JAX), Williams (UT), Manly (UT)

Over 70 years ago Sewall Wright proposed his "Shifting Balance" theory of evolutionary selection in which populations acquire favorable combinations of alleles at multiple genes and that it is these combinations, rather than individual loci, on which selection acts (summarized in Wright, 1980). Recent experimental evidence of linkage disequilibrium (LD) among inbred stains of mice supports the idea that coordinated evolution has occurred and that a substantial fraction of the genome is organized into large domains of co-adapted genes and that these domains determine phenotype by interacting across the genome in complex functional networks (Petkov *et al.,* 2005).

Beginning nearly a hundred years ago, laboratory scientists began breeding inbred strains that are artificial inter-subspecific hybrids (Fig. 13A), scrambling the genomes of different subspecies and then driving the animals to complete homozygosity, a process that involves intense selection for survival and reproduction, in addition to human selection for mate choice. The result is a unique, albeit involuntary, evolutionary experiment in which preliminary data has lead to the proposal that co-adapted alleles at different loci were retained in specific combinations (Petkov *et al*., 2005). The CC provides an exceptional opportunity to gain insight on the extent of epistasis between alleles in functionally related genes in the mouse and on how this process influences phenotypic variation like response to stress.

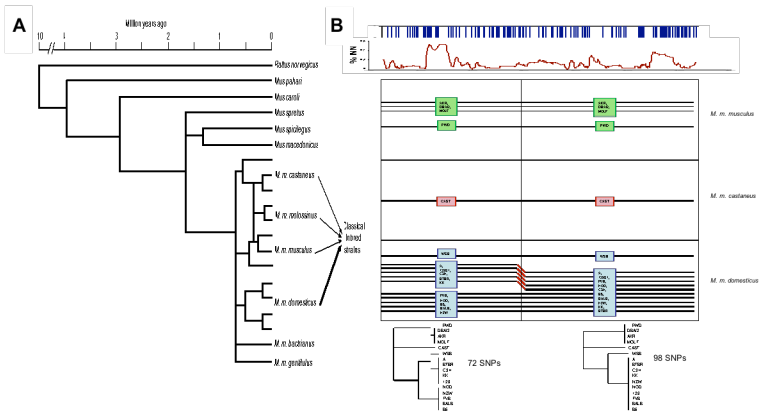



**Figure 13.** Phylogenetic relationships of mouse strains. (A) Classical representtation of the genus *Mus* showing divergence times for each lineage (Guenet and Bonhomme, 2003). (B) Haplotype identity and phylogenetic origin in a 50kb region of mouse Chr 11 from a web browswer under development by the Center of Excellence for Chrmoosome Dynamics (NIGMS, PI: Churchill), which will be used in Project 7.

Here we propose to identify and define LD domains (syntenic) and networks of co-inherited alleles (*i. e*., nonsyntenic "linkage" disequilibrium) in both the final set of 500 CC lines and at intermediate steps during the derivation process prior to and after the arising of inbreeding depression. This is possible since the starting population used to generate the CC is already inbred, meaning that inbreeding depression and the assembly of LD domains and networks should be mostly due to epistatic interactions between co-evolved genes, not selection against recessive lethal alleles. The specific aims will be to:

*Aim 1. Genotype early generations of CC lines during and after the inbreeding process.* When genomes are mixed, as in the derivation of RI lines, unfavoralble allele combinations lead to inbreeding depression. To investigate this phenomenon of allele co-inheritance, approximately 10,000 animals from F1 to F10 will be genotyped at 1,500 maximally informative SNPs evenly distributed across the genome. This data will be used to identify: a) nascent LD domains and networks, b) regions of residual heterozygosity, and c) correlation of these genetic features with the reproductive performance of sib-pairs (litter size, time to litter, sex ratio). After inbreeding is complete, the 500 CC lines will then be genotyped at 100,000 SNP markers to identify fixed LD domains and networks.

*Aim 2. Compare LD domains and networks between CC lines and extant inbred lines.* Since the vast majority of extant inbred strains and RI lines underwent significant human selection during inbreeding, the CC lines, which will have far less human selection due to the randomly assigned breeding scheme (see Project 2), will provide a way to identify which LD domains and networks are due to natural mating selection versus those

caused by human intervention. We will also be able to determine whether alternative alleles found in the LD domains and networks have co-evolved (*i. e*., have originated in the same phylogenetic lineage, Figure 13B).

*Aim 3. Use information from Aims 1 and 2 to identify dysfunctional combinations of alleles impacting physiological response to stress.* The LD domains and networks of co-inherited alleles will be compared to phenotypes and markers of stress response studied in Projects 3-6 to identify those lines that have adverse responses to stress due to retention of unfavorable allele combinations during inbreeding.

**PROJECT 8: (R01) High performance computing for data mining and modeling**

*PI:* Wang (UNC)

*Additional participants:* Manly (UT), McMillan (UNC), Langston (UT), Edward Uberbacher (ORNL)

The data generated by the ICOGS will eventually contain high-density SNPs, or even whole genome sequences, for 500 CC lines and millions of phenotypic measurements (molecular and physiological) and other derived variables. Several key characteristics of this large data matrix complicate its analysis: a) the dimensionality is high since the data matrix contains massive amounts of information on (relatively) few subjects and there exist both complex correlations and causal relationships between variables, b) the data matrix is comprised of disparate measurements including both continuous and discrete variables, which may not be directly comparable to each other, c) the data matrix is not static, but growing, both in terms of adding new CC lines and measurements through the pilot project program or other users, d) individual items may be contaminated, noisy or simply missing, which makes detectable relationships hard to "see", and thus hard to interpret, and e) the number of unknowns far exceeds the number knowns since relatively little is known about associations between polymorphisms to gene expression pathways to phenotypic observations. Consequently, the number of potential hypotheses is extremely large, making it intractable to generate and test every possibility.

As described above, a significant computing challenge will arise from the large quantities of data generated in Projects 3-7 and the computational intense applications developed in Projects 9 and 10. This project aims to address this challenge through the development of novel, scalable data structures and access methods that are applicable to many types of data sets, conditions and queries. The specific aims will be to:

*Aim 1. Investigate novel structures to support ICOGS data.* Approaches to organize and reduce the complexity without loosing information will be developed. This should enable highly efficient data mining with realtime feedback of results thus allowing repetitive data queries involving extensive permutation analysis to evaluate significance. The data structures will also be designed to be compatible with the visualization tools that will be built in Project 11. To achieve this goal, we will design scalable mining algorithms and compact data representations that enable efficient discovery and summarization of relevant patterns. The criteria for the new computational models and algorithms developed in this project are to be efficient, scalable, incrementally updatable, robust to noise, and easily parallelizable.

*Aim 2. Develop efficient representations for high-density data.* One type of data requiring a novel structure for efficient analysis is SNPs. We are developing a SNP pattern data structure and search engine that is optimized for fast queries and correlation calculations. The key idea employed in our approach is the use of a pattern-based, rather than position-based, SNP graph representation, with edges between highly associated patterns. This allows highly correlated SNP pattern queries to proceed as depth-first graph traversals. Our approach leverages the biological observations that the number of SNP association patterns observed is far smaller than the space of possibilities. It also takes advantage of their local coherence. The net result is a SNP query engine that allows for thousands of permutation evaluations and association tests and is insensitive to the size of the SNP strings. This general approach will also be extended to other data types.

*Aim 3. Address computational challenges in system genetics.* The data sets that will be generated in the ICOGS, as well as the intermediate data structures they imply, are too large for in-core processing. Yet, we intend to provide access, summaries, and visualizations in Projects 2 and 11 for these data sets in their entirety. We intend to address these problems using a family of compression methods called dimensionality reduction. Current dimensionality reduction methods are suitable for a few hundreds-of-strains and, then, only off-line. We propose to scale this for the potential to deal with thousands-of-strains in real time. We have already employed dimensionality reduction methods, specifically Multidimensional Scaling (MDS), to visualize the impact of parameter choices in Microarray clustering analysis. MDS is also an integral part of many modern non-linear dimensionality reduction methods such as isomaps (Tanenbaum et al., 2000). The problem with MDS is that it does not easily scale to large data sets. We have developed a fast MDS approximation algorithm that is targeted at interactive rates and large data sets (Yang *et al*., 2006). This new approach is orders-of-

magnitude faster than previous published methods and can generalize to higher dimensions better than other fast methods.

**PROJECT 9: (R01) Advanced statistical modeling of comlex genetic networks**

*PI:* Churchill (JAX)

*Additional participants:* Manly (UT), Karl Broman (Johns Hopkins University), Fei Zou (UNC)

The ability to assay phenotypes associated with stress response presents a unique opportunity to investigate the mechanisms and biomarkers underlying altered disease susceptibility, but also a unique challenge to statistically evaluate the resulting associations. To accurately represent the associations and to build predictive models, we will need to know which phenotypes share common genetic determinants, which genetic factors are specific to different phenotypes or networks of interactions, and the nature of non-genetic interactions among phenotypes that can be modulated by stress. The high-dimensionality of the data sets developed in Projects 3-8 pose a particular challenge to develop appropriate statistical tools to evaluate results using computational analyses from Project 10. The specific aims will be to:

*Aim 1. Explore multivariate analysis methods for quantitative variables.* Appropriate tools are needed to investigate the structure of a genetic system that includes allelic variation at multiple loci, intermediate phenotypes and disease states. We will build upon our experience in complex and quantitative genetics to provide methods to statistically evaluate results obtained using algorithms developed in this project as well as those in Project 10. This will include methods for building multivariant and random effects models to deal with the large numbers of estimated parameters that will result from the data sets collected in Projects 3-7. Additionally, we will explore methods for multivariant modeling using Bootstrapped data, Bayesian averaging, and Monte-Carlo Markov Chains. An inherent objective of these analysis is to efficiently detect epistatic interactions.

*Aim 2. Develop statistically sound approaches to extract biological knowledge.* Mapping studies that investigate clusters of related phenotypes often reveal a network of genetic effects, in which each phenotype is influenced by multiple loci (heterogeneity) and different phenotypes share one or more loci in common (pleiotropy) (Stoll *et al*., 2001; Nadeau *et al.,* 2003). The complexity of observed networks will vary depending on the traits and it is also likely that physiological interactions, independent of genetic factors, may result in correlated phenotypic responses. One method that we will extend to extract information is through structural equation models (SEM) (McArdle and Goldsmith, 1990), a class of graphical model in which measured variables are represented as nodes and causal relationships are directed edges. Multivariate probability distributions are defined by the conditional dependencies represented in the graph. They impose structure on the correlation through a system of linear equations that define the causal relationships among measured variables in a system. Covariates – factors like sex, batch, or litter that are external to direct genetic causality but introduce variations in phenotypes of interest – can be incorporated into SEM analysis. We have applied SEM methods to small numbers phenotypes (Fig. 14). The challenge is to bring these methods to bear now on more extensive phenotype collections like those that will be collected on the CC; this will present additional analytic and computational challenges.

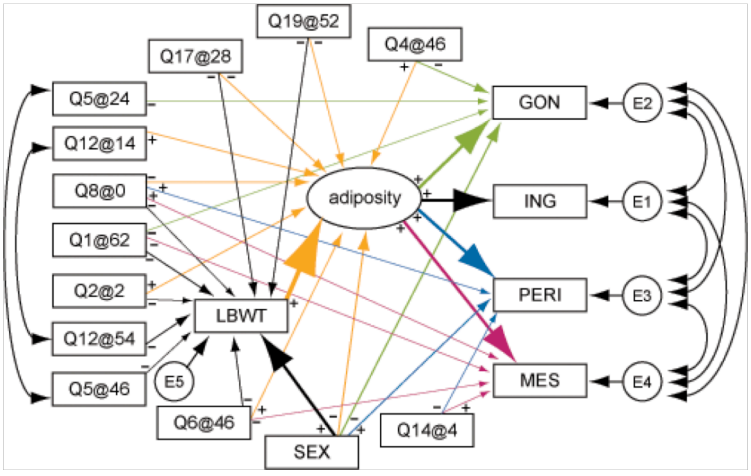



**Figure 14.** Graphical representation of the structural equation model for adiposity and lean body weight. The four fat pad traits are: ING = inguinal; GON = gonadal; PERI = peritoneal; MES = mesenteric. Single headed arrows indicate causal paths and the thickness of each arrow is proportional to the effect sizes. Doubled-headed arrows denote unresolved covariance. The boxes indicate measured traits or QTL and the oval denotes a latent variable. E1, E2, E3, E4 and E5 denote unobserved residual error.

**PROJECT 10: (R01) Graph theory and clique analysis of multiscale networks**

*PI:* Langston (UT)

*Additional participants:* Chesler (ORNL), Williams (UT), Wang (UNC), Churchill (JAX), Bruce Aronow (Cincinnati Children's Hospital)

The large data sets being collected as part of Projects 3-7 can be exploited to study the genetic regulatory mechanisms that control cellular and physiological responses to generalized stress. These responses are probably highly complex, and involve many genes and biomolecules. To increase our understanding of stress response, we seek a) to develop novel algorithms to generate highly distilled gene sets, b) to produce scalable implementations for cutting-edge high performance computing platforms, c) to use these implementations to extract gene sets suggestive of co-regulation, and d) to perform genomic data mining to highlight the most promising gene sets for detailed scrutiny and sequence comparisons. Our primary target is the elucidation of genetic regulatory mechanisms that control cellular responses to stress.

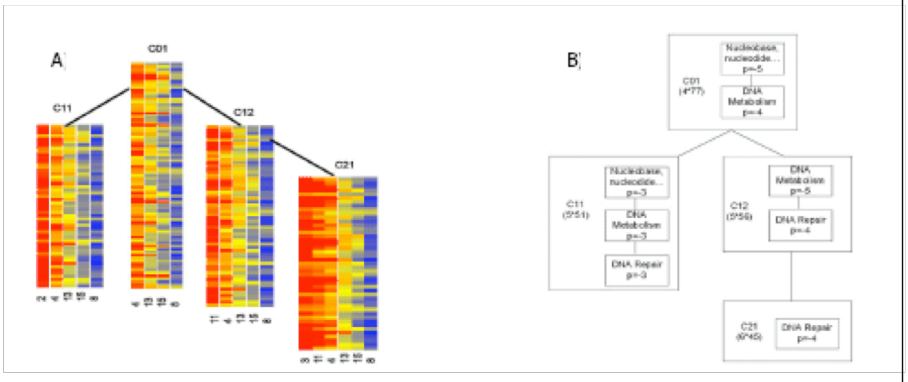
The traditional common clustering approaches are either hierarchical, in which all genes begin in their own clusters and are eventually merged into one, or centroid, in which genes are organized into a predefined number of clusters by iterative adjustments based on similarity (Slonim, 2002). The resulting clusters are typically disjoint, which places an artificial limitation on the biology under study in that many genes have pleiotropic effects with important roles in multiple but distinct pathways (Butte *et al.,* 2000). More recent clustering techniques, such as factor analysis (Alter *et al*., 2000) and relevance networks (Butte *et al.,* 2000; Patti *et al.,* 2003; Allocco *et al.,* 2004), that do not require exclusive cluster membership for single genes. Unfortunately, these tend to produce biologically uninterpretable factors without the incorporation of prior biological information or lack algorithmic means to extract the aggregate relationships between multiple genes. It is therefore necessary to develop more powerful tools to extract subsets of coordinately regulated genes and biomolecules from large aggregates of data. The specific aims of this project will be to:

*Aim 1. Use graph theory to elucidate biological networks.* Graph theory offers unique advantages to the types of analyses proposed in the ICOGS. Graph algorithms are based on decades of basic research, and constitute a class of tools that can help elucidate relationships in highly complex data structures, in our case as matrices of correlations across thousands of genes. In this respect, graph algorithms offer a means to extract lightly-connected aggregates of genes from within the relevance network framework, resulting in unweighted correlation graphs whose properties may convey many forms of biological meaning.

*Aim 2. Apply clique-centric approaches to identify putative co-regulated networks.* The challenge, once a network graph is created, is not to study the graph in its entirety, but rather to extract its embedded subgraphs, or small, tightly connected regions of the graph that represent subsets of genes or other measurements with strong correlations between every pair of its members and thus are likely to represent biologically significant interactions. In the most extreme case, in which a subgraph contains all possible edges between its vertices, this structure is called a *clique*, which will be used to identify co-regulation of genes and other measurements. Each and every pair of vertices is joined by an edge, from which we can infer some form of co-regulation among the corresponding genes. Clique is widely known for its application in a variety of combinatorial settings, a great number of which are relevant to computational molecular biology (Setubal and Meidanis, 1997). It is important to note that cliques need not be disjoint -- a vertex can reside in more than one clique, just as a gene can be pleiotropic and operate in more than one regulatory network. In terms of gene expression, clique represents the most trusted potential for identifying a set of interacting genes. The degree of co-regulation of genes within cliques will be analyzed using common sequence motifs, ontology and transcription factor databases to identify and validate the molecular mechanisms of co-regulation.

*Aim 3. Develop subspace clustering algorithms to identify meaningful data subsets.* Subspace Clustering (SSC) is an alternative approach that will also be developed to investigate the effects of stress on biological systems. SCC is an unsupervised learning method for identifying patterns of association among trait measurements and variables. Similar to clique analysis, it does not rely on external categorization or labeling. Patterns will typically be identified by: a) dividing the CC lines into groups with similar responses, b) dividing the variables into groups with similar values across lines, and c) identifying distinguished line-variable associations (Fig. 15). With appropriate similarity and association criteria, all these patterns can be modeled via SSC.

**Figure 15.** Subspace clusters. (A) A list of subspace clusters, each of which has consistent patterns of their gene expressions. (B) Annotation of each cluster using gene ontology.



It allows cluster overlap that leads to improved interpretability and supports incorporation of prior biological knowledge. Given a data matrix along with some partitioning criteria, we have developed SSC algorithms that can efficiently enumerate every submatrix satisfying a given criterion (Liu *et al.,* 2004; Liu *et al*., 2004; Liu *et al.,* 2006).

**PROJECT 11: (R01) Visual data portal interfaces**

*PI:* McMillan (UNC)

*Additional participants:* Manly (UT), Threadgill (UNC), Williams (UT), Chesler (ORNL), Wang (UNC)

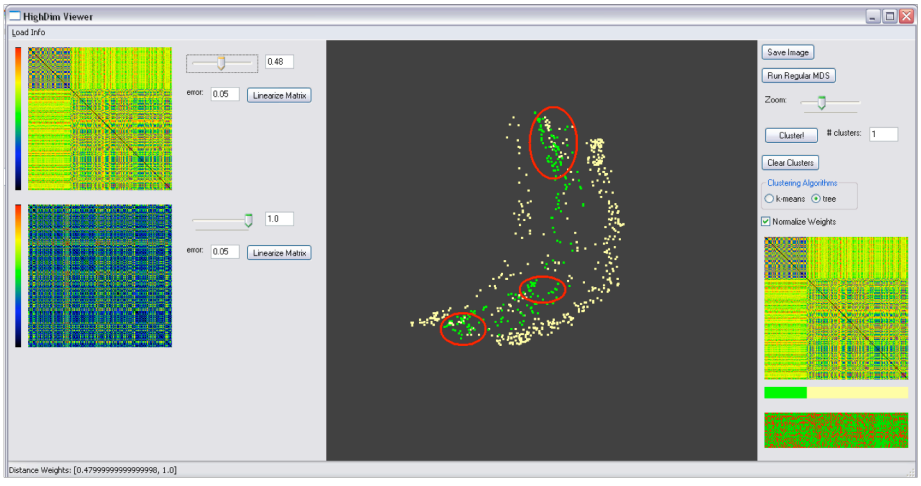
Because of the complex nature of the data that will be generated by the ICOGS, innovative data visualization and access approaches are needed to support the analyses proposed in Projects 3-7. A primary goal is to provide customized navigation tools that provide easy access to complex data sets and to aid in its comprehension. In addition to supporting the ICOGS, we will provide interfaces suitable for use by novices that will be integrated into the ICOGS web portal as part of Project 2. This effort includes the development of two-way portals for the CC results, tightly integrated with the data structures generated in Project 8 to permit realtime data exploration, even using large-scale permutations, and tools from Projects 9-10. Training in the use of the ICOGS web interfaces will be through Project 11. The specific aims will be to:

*Aim 1. Develop visual data exploration interfaces via smart clients.* The proposed projects in the ICOGS will generate immense volumes of data, only a small fraction of which might be relevant to particular query. We will address the problem of how to interactively narrow the scope of a query, and to provide intelligent client interfaces to the underlying databases. Google Earth for geographical data is an example of the sort of smart client that we envision applied to systems genetics data. Providing this level of access requires the development of new and innovative user-interaction paradigms using smart database clients, which can provide for efficient filtering, searching, indexing, and depiction of large, high-dimensional datasets.

*Aim 2. Explore approaches for depicting high-dimensional data.* A major challenge to support data interpretation will be to develop techniques for efficiently exploring large data sets as a whole or in parts. We will build on our current efforts to develop new visualization techniques that aid the process of exploring large high-dimensional data spaces (Fig. 16). Our approach is to create lower dimensional (2D and 3D) analogs which maintain user-specified properties of the raw data. By doing so, we allow users to gauge both local and global effects of the parameter choices, and to see how these choices impact data clustering and statistical correlations. These tools should facilitate modeling, understanding, and ultimately, discovery of new relationships.

*Aim 3. Develop information visualization tools.* We will create a mix of new visualization elements for viewing, analyzing and exploring data sets to aid in the comprehension of genetic and physiological links to stress-induced changes in health status. One of our primary views will be the dissimilarity matrix, which has the property of reducing a wide-range of properties (both discrete and continuous) to a matrix of scalar values. This permits the visual exploration of seemingly incompatible data types. Since dissimilarity matrices are insufficient for viewing the formation of clusters in response to parameter changes, we also plan to employ a point-cloud analogy of the data set that treats the dissimilarity matrices as the distance between points. This visualization provides a spatial representation of the clusters in a low-dimensional setting suitable for direct viewing in 3D. In this view, it is possible to see clusters of similar data points as well as to track the evolution of clusters resulting from parameter changes.

**Figure 16.** Screen captures of a prototype viewer applied to SNP-phenotype correlations. The top left matrix is a permuted view of two dissimilarity matrices for a SNP data set. The center panel depicts a three-dimensional analogy of the SNP-phenotype correlation. The right square matrix illustrates a user-defined combination of the dissimilarity matrices from the left column.




*Aim 4. Provide universal data access points.* A goal of the ICOGS is to make our data and tools accessible to a wider audience. Visualization and portal access is a key to this dissemination. This includes

access for other researchers and educators at all levels. We believe that ease-of-use and flexibility are essential to making this a success. We plan to build with Projects 2 and 11, a single user interface that will span the various needs of domain experts, with specific scientific inquiries, to high-school-aged children, who are being introduced to the concepts of complex genetics.

**Specific Aim 4: Design programs to foster the growth and broader application of systems genetics.** Our goal is not only to develop models describing how biological information is altered by external stress leading to altered physiology and disease susceptibility, but also to cultivate interdisciplinary systems genetics research through novel interactions and learning environments, and to educate current and future generations of researchers to exploit the power of systems genetics, particularly focusing on the use of the CC, which takes advantage of natural allelic variation present among mice as a model for human populations. This will be achieved through two complementary programs.

**PROJECT 12: (R25) Systems genetics workshop and outreach**

*PI:* Manly (UT)

*Additional participants:* Chesler (ORNL), Threadgill (UNC), Langston (UT), Miller (ORNL)

This project will develop methods and procedures that encourage communication and information exchange through a program that educates ICOGS members by focusing on mutual awareness and dialog and educating the larger scientific and lay communities. The proposed specific aims will be to:

*Aim 1. Organize and manage a monthly Interdisciplinary Clinic.* This clinic, presented by scheduled Webcast or video conference and organized with Project 1, will describe a current interdisciplinary or transdisciplinary problem encountered by groups of investigators involved with Projects 2-11, what the desired outcome is, and the proposed solution. Open discussion and feedback will be elicited from the participants. To allow free discussion of unpublished data, this clinic will be restricted to authorized ICOGS laboratories and staff. A secondary goal of this aim is to establish the methods and infrastructure to encourage regular interdisciplinary meetings, including how such meetings are organized, moderated and documented using contemporary web-based tools. The results will provide a template for wider adoption by other interdisciplinary groups.

*Aim 2. Organize a workshop associated with the annual meeting of the CTC.* This workshop will explain systems genetics applications of the CC using the stress-disease focus of the ICOGS. The content of the workshop will consist of ICOGS members presenting specific results obtained from the CC that were the result of interdisciplinary approaches, training on the use of ICOGS-developed software and resources, and tutorials on systems genetics to support predictive medicine. Each day-long workshop will also have one or two invited non-ICOGS speakers who are performing related systems genetics-type research, who are using the CC, who are developing computational or analytical tools that can be applied to the CC, or who are applying results from CC studies to humans. Electronic versions of the workshops and workshop materials, in the form of videos and slide shows, will be published on a public section of the ICOGS web portal.

*Aim 3. Develop web tutorials and classroom-based educational programs on systems genetics and its application to human health.* A major goal of this project will be the development of educational materials that bring systems genetics to life within the high school classroom. Both web-based tutorials classroom projects describing how biological and computational sciences are being integrated to address the major health issue of our times will be developed. Several ICOGS participating institutions have a long track record of supporting public high schools through cooperative programs that bring select students into the laboratory for hands on research experiences. This will be expanded to allow more students to oportunity to experience the excitement of interdisciplinary science.

**PROJECT 13: (T90/R90) Interdisciplinary training program in systems genetics**

*PI:* Cynthia Peterson (UT)

*Additional participants:* All members of the EC

With the explosive growth in the scale and complexity of biological data, the computationally intensive data processing and the formation of multidisciplinary teams, the training of future scientists must be more interdisciplinary. The scales at which data are now gathered and analyzed require scientists who are able to integrate hypothesis-driven experimental design with facility at informatics and computational science to elucidate complex biological relationships. They also must be capable of traversing smoothly between fields. This project will develop a novel post-doctoral training program with three funded positions to train the next generation of scientist, ensuring that they become fluent in the multiple disciplines and enabling them to

actively contribute and develop research projects at the disciplinary interfaces of systems genetics with which to launch their own independent research careers. The proposed specific aims will be to:

*Aim 1. Implement a competitive application process for training positions.* A major goal of the training program is to support trainees that have the drive, dedication and potential to become leaders in systems genetics. Consequently, there will be a formal application and review process for the available positions. Given the requisite skills needed to train this unique cohort of scientists, the positions will also be open for non-NRSA eligible applicants. However, of the three available slots, one will be reserved for a US citizen or permanent resident. Applicants can apply directly to the program or be nominated by ICOGS-associated laboratories. A requisite of the program is that applicants must identify two mentors from different disciplines and propose a research project encompassing both disciplines, spending at least one-fourth of their time in each laboratory. Fellows will be supported for two years, with the option for a third year of training upon competitive re-application. All applicants will be trained in ethics, grant writing and be expected to submit an independent fellowship application.

*Aim 2. Develop a novel web-based course in systems genetics.* The applicants to the training program will be expected to have previous training in one discipline and a working knowledge of at least one other. To extend the prior PhD training of the fellows, a web-based distance learning course will be developed, allowing fellows at all ICOGS member locations to participate, for non-fellows in ICOGS laboratories to participate and instructors to be drawn from all ICOGS-associated faculty. The focus of the course is to emphasize the relation between experimental design, analytic methods and the construction of biological knowledge, and will be organized around case conferences. An example case might include the presentation of a complex biological problem for which large datasets are available and that requires a combination solution drawing on computational, statistical, and/or analytic solutions. Background information required for the case conferences will be posted on the ICOGS web site. Basic programming skills using PERL, R or other script languages will be required.

*Aim 3. Organize a seminar and journal club focusing on discipline interfaces and stress responses.* In addition to formal training, scientists-in-training need to develop skills to present and converse across disciplines. This becomes particularly relevant as team science and interdisciplinary research becomes essential to enable investigation of ever more complex biological questions. Similarly, investigators at different locations will need to interact as efficiently as scientists that are physically close. To gain experience in remote collaborations, a weekly meeting using multimedia capabilities will be organized. This distance learning activity will combine a research colloquium and journal club focusing on systems genetics. ICOGS faculty working in the area of the weekly topic will also participate to stimulate thought questions and alternative perspectives.

## ICOGS Summary

*"Solving the puzzle of complex diseases, from obesity to cancer, will require a holistic understanding of the interplay between factors such as genetics, diet, infectious agents, environment, behavior, and social structures."*

E. Zerhouni, 2003

The ICOGS will be administered by the U54 (Project 1) component, which will oversee and maintain direction by providing organizational structure to the interdisciplinary consortium (Fig. 17). The output of the consortium will be presented to the consortium and relayed to the broader scientific community through the R25 (Project 12), R01 (Project 11) and P30 (Project 2) components, which will develop workshops and 'Interdisciplinary Clinics" to both educate and elicit comments on pertinent problems encountered by the individual projects and develop a web portal for all consortium activities. To train the next generation in the new science of systems genetics, a T90/R90 (Project 13) post-doc training program will be established.

The consortium is organized around the CC and the bioinformatics portal in the P30 (Project 2) research core, leveraging a substantial investment by the EMF and DOE. The CC provides a unique platform for developing the new field of systems genetics, which is the study and interpretation of interactions between genetic variation and environmental perturbations, as modeled by stress in this proposal. The outcome of these interactions is alterations in intermediate biomolecules, physiological states and subsequent differential susceptibility to many types of common diseases afflicting humans, which will be measured and quantified in a set of R01 (Projects 3-7) components. To enable the integration and interpretation of the large data sets collected collected by the ICOGS, a set of R01 (Projects 8-10) components will develop new methods of organizing data and approaches for analysis. A major overarching goal of the ICOGS is to use the resulting

data to develop approaches to predict phenotypic outcome from defined genotypes or generalized external stressors. The platform that will be developed should dramatically propel the field of predictive medicine much close to reality.

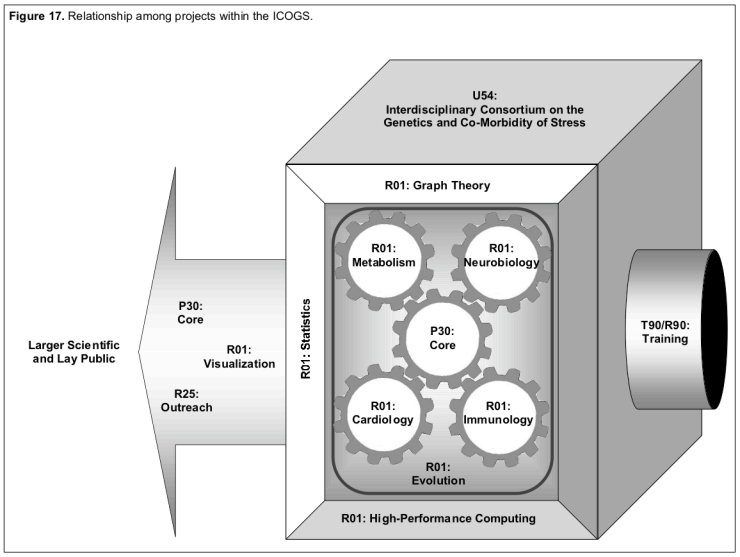


**Figure 17.** Relationship among projects within the ICOGS.

At this early stage in the development of these new research tools, we can not fully appreciate either the scope or limitations of the powerful new approach called systems genetics, particularly as applied to the complex relations between stress and biological homeostasis leading to altered disease susceptibility. Nonetheless, the plan we have presented has the potential to radically alter our understanding of complex biological systems by pursuing a holistic approach to biological systems and integrating across disciplines with powerful new computational and mouse reagent tools. Although highly anticipated, predictive medicine is currently a dream as we know too little about how biological systems are interconnected. Ultimately, we would like to know, based solely on genetic information, what the likely future health of an individual will be. Yet, how can we have confidence in genetically-based predictions in light of the known complexities and the multitude of factors that determine individual phenotypes? Our proposal, using the novel and powerful attributes of the CC to investigate one major factor, generalized stress, is a step in the right direction and will be the prototype for application to human health by laying the ground work for developing methods and tools to identify the probability that an individual will develop specific disease characteristics. The application of predictive genetics to humans will obviously involve greater uncertainties but if there is any hope to achieve a time when we can predict and alter disease processes, the principles must be developed and refined in the context of a mammalian model system with sufficient genetic diversity to accurately represent humans.

## References Cited


Allocco DJ, Kohane S, Butte AJ. (2004) Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics 5:18.

Alter A, Brown PO, Botstein D. (2000) Singular value decomposition for genome-wide expression data processing and modeling, Proc Natl Acad Sci USA 97:10101–10106.

Benarroch EE. (2005) Paraventricular nucleus, stress response, and cardiovascular disease. Clin Auton Res 15:254-263.

Ben-Eliyahu S, Page GG, Yirmiya R, Shakhar G. (1999) Evidence that stress and surgical interventions promote tumor development by suppressing natural killer cell activity. Int J Cancer 80:880-888.

Bode VC. (1984) Ethylnitrosourea mutagenesis and the isolation of mutant alleles for specific genes located in the T region of mouse chromosome 17. Genetics 108:457-470.

Broman KW. (2005) The genomes of recombinant inbred lines. Genetics 169:1133-1146.

Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci U S A 97:12182-12186.

Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, Williams RW. (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet. 37:233-242.

Chesler EJ, Wang J, Lu L, Qu Y, Manly KF, Williams RW. (2003) Genetic correlates of gene expression in recombinant inbred strains: a relational model system to explore neurobehavioral phenotypes. Neuroinformatics 1:343-357.

Chrousos GP, Kino T. (2005) Ikaros transcription factors: flying between stress and inflammation. J Clin Invest. 115:844-848.

Church GM. (2006) Genomes for all. Sci Am. 294:46-54.

Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, Beavis WD, Belknap JK, Bennett B, Berrettini W, Bleich A, Bogue M, Broman KW, Buck KJ, Buckler E, Burmeister M, Chesler EJ, Cheverud JM, Clapcote S, Cook MN, Cox RD, Crabbe JC, Crusio WE, Darvasi A, Deschepper CF, Doerge RW, Farber CR, Forejt J, Gaile D, Garlow SJ, Geiger H, Gershenfeld H, Gordon T, Gu J, Gu W, de Haan G, Hayes NL, Heller C, Himmelbauer H, Hitzemann R, Hunter K, Hsu HC, Iraqi FA, Ivandic B, Jacob HJ, Jansen RC, Jepsen KJ, Johnson DK, Johnson TE, Kempermann G, Kendziorski C, Kotb M, Kooy RF, Llamas B, Lammert F, Lassalle JM, Lowenstein PR, Lu L, Lusis A, Manly KF, Marcucio R, Matthews D, Medrano JF, Miller DR, Mittleman G, Mock BA, Mogil JS, Montagutelli X, Morahan G, Morris DG, Mott R, Nadeau JH, Nagase H, Nowakowski RS, O'Hara BF, Osadchuk AV, Page GP, Paigen B, Paigen K, Palmer AA, Pan HJ, Peltonen-Palotie L, Peirce J, Pomp D, Pravenec M, Prows DR, Qi Z, Reeves RH, Roder J, Rosen GD, Schadt EE, Schalkwyk LC, Seltzer Z, Shimomura K, Shou S, Sillanpaa MJ, Siracusa LD, Snoeck HW, Spearow JL, Svenson K, Tarantino LM, Threadgill D, Toth LA, Valdar W, de Villena FP, Warden C, Whatley S, Williams RW, Wiltshire T, Yi N, Zhang D, Zhang M, Zou F; Complex Trait Consortium. (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet. 36:1133-1137.

Collins A, Lau W, De La Vega FM. (2004) Mapping genes for common diseases: the case for genetic (LD) maps. Hum Hered. 58:2-9.

Dallman MF, Pecoraro N, Akana SF, La Fleur SE, Gomez F, Houshyar H, Bell ME, Bhatnagar S, Laugero KD,

Manalo S. (2003) Chronic stress and obesity: a new view of "comfort food". Proc Natl Acad Sci U S A. 100:11696-11701.

DeRijk R, de Kloet ER. (2005) Corticosteroid receptor genetic polymorphisms and stress responsivity. Endocrine 28:263-270.

Doetschman T, Gregg RG, Maeda N, Hooper ML, Melton DW, Thompson S, Smithies O. (1987) Targetted correction of a mutant HPRT gene in mouse embryonic stem cells. Nature 330:576-578.

Fisher, RA (1951) The Design of Experiments, 6th ed., Oliver and Boyd, London.

Furukawa S, Fujita T, Shimabukuro M, Iwaki M, Yamada Y, Nakajima Y, Nakayama O, Makishima M, Matsuda M, Shimomura I. (2004) Increased oxidative stress in obesity and its impact on metabolic syndrome. J Clin Invest. 114:1752-1761.

Glaser R, Kiecolt-Glaser JK, Stout JC, Tarr KL, Speicher CE, Holliday JE. (1985) Stress-related impairments in cellular immunity. Psychiatry Res. 16:233-239.

Glaser R, Thorn BE, Tarr KL, Kiecolt-Glaser JK, D'Ambrosio SM. Effects of stress on methyltransferase synthesis: an important DNA repair enzyme. Health Psychol. 4:403-412.

Gordon JW, Scangos GA, Plotkin DJ, Barbosa JA, Ruddle FH. (1980) Genetic transformation of mouse embryos by microinjection of purified DNA. Proc Natl Acad Sci U S A 330:7380-7384.

Grupe A, Germer S, Usuka J, Aud D, Belknap JK, Klein RF, Ahluwalia MK, Higuchi R, Peltz G. (2001) In silico mapping of complex disease-related traits in mice. Science 292:1915-1918.

Hartwell L (2004) Genetics. Robust interactions. Science 303:774-5.

Huang S. (2002) Rational drug discovery: what can we learn from regulatory networks? Drug Discov Today 7:S163-169.

Jansen RC. (2003) Studying complex biological systems using multifactorial perturbation. Nat Rev Genet. 4:145-151.

Kelmenson PM, Petkov P, Wang X, Higgins DC, Paigen BJ, Paigen K. (2005) Genetics 169:833-841.

Liu H, Wang Z (2005) Effects of social isolation stress on immune response and survival time of mouse with liver cancer. World J Gastroenterol. 11:5902-5904.

Jinze L, Wei W, Yang J. (2004) A framework for ontology-driven subspace clustering, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 623-628.

Jinze L, Strohmaier K, Wang W. (2004) Revealing true subspace clusters in high dimensions, Proceedings of the 4th IEEE International Conference on Data Mining (ICDM), pp. 463-466.

Liu, J, Paulsen S, Xu X, Wang W, Nobel A, Prins J. (2006) Mining Approximate frequent itemset in the presence of noise: algorithm and analysis, Proceedings of the 6th SIAM Conference on Data Mining (SDM).

McArdle JJ, Goldsmith HH (1990) Alternative common factor models for multivariate biometric analyses. Behav Genet 20:569-608.

McEwen BS (2004) Protection and damage from acute and chronic stress: allostasis and allostatic overload and relevance to the pathophysiology of psychiatric disorders. Ann N Y Acad Sci. 1032:1-7.

Mhyre TR, Chesler EJ, Thiruchelvam M, Lungu C, Cory-Slechta DA, Fry JD, Richfield EK. (2005) Heritability,

correlations and in silico mapping of locomotor behavior and neurochemistry in inbred strains of mice. Genes Brain Behav. 4:209-228.

Morikawa Y, Kitaoka-Higashiguchi K, Tanimoto C, Hayashi M, Oketani R, Miura K, Nishijo M, Nakagawa H. (2005) A cross-sectional study on the relationship of job stress with natural killer cell activity and natural killer cell subsets among healthy nurses. J Occup Health 47:378-383.

Nadeau JH, Burrage LC, Restivo J, Pao YH, Churchill G, Hoit BD. (2003) Pleiotropy, homeostasis, and functional networks based on assays of cardiovascular traits in genetically randomized populations. Genome Res. 13:2082-2091.

Patti ME, Butte AJ, Crunkhorn S, Cusi K, Berria R, Kashyap S, Miyazaki Y, Kohane I, Costello M, Saccone R, Landaker EJ, Goldfine AB, Mun E, DeFronzo R, Finlayson J, Kahn CR, Mandarino LJ. (2003) Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1. Proc Natl Acad Sci U S A. 100:8466-8471.

Petkov PM, Graber JH, Churchill GA, DiPetrillo K, King BL, Paigen K. (2005) Evidence of a large-scale functional organization of mammalian chromosomes. PLoS Genet. 1:e33.

Pletcher MT, McClurg P, Batalov S, Su AI, Barnes SW, Lagler E, Korstanje R, Wang X, Nusskern D, Bogue MA, Mural RJ, Paigen B, Wiltshire T. (2004) Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. PLoS Biol. 2:e393.

Russell WL, Kelly EM, Hunsicker PR, Bangham JW, Maddux SC, Phipps EL. (1979) Specific-locus test shows ethylnitrosourea to be the most potent mutagen in the mouse. Proc Natl Acad Sci U S A 76:5818-5819.

Schwab M, Eichelbaum M, Fromm MF. (2003) Genetic polymorphisms of the human MDR1 drug transporter. Annu Rev Pharmacol Toxicol. 43:285-307.

Selye H (1936) A syndrome produced by diverse nocuous agents. Nature 138:32.

Selye H (1956) The Stress of Life. New York: McGrawHill.

Setubal JC, Meidanis J. (1997) Introduction to Computational Molecular Biology. Boston: PWS Publishing Company.

Slonim DK (2002), From patterns to pathways: gene expression data analysis comes of age. Nature Genetics Supplement 32:502-508.

Stoll M, Cowley AW Jr, Tonellato PJ, Greene AS, Kaldunski ML, Roman RJ, Dumas P, Schork NJ, Wang Z, Jacob HJ. (2001) A genomic-systems biology map for cardiovascular function.
Science 294:1723-1726.

Tenenbaum JB, Silva V, Langford JC. (2000) A Global Geometric Framework For Nonlinear Dimensionality Reduction. Science 290:2319-2323.

Thomas KR, Capecchi MR. (1987) Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. Cell 51:503-512.

Threadgill DW, Hunter KW, Williams RW. (2002) Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. Mamm Genome 13:175-178.

Tsaih SW, Lu L, Airey DC, Williams RW, Churchill GA. (2005) Quantitative trait mapping in a diallel cross of recombinant inbred lines. Mamm Genome 16:344-355.

Valdar W, Flint J, Mott R. (2006) Simulating the collaborative cross: power of quantitative trait Loci detection

and mapping resolution in large sets of recombinant inbred strains of mice. Genetics 172:1783-1797.

Wakeland E, Morel L, Achey K, Yui M, Longmate J. (1997) Speed congenics: a classic technique in the fast lane (relatively speaking). Immunol Today 18:472-477.

Yang T, Liu J, Wang W, McMillan L. (2006) A fast approximation to multidimensional scaling, ECCV Workshop on Computation Intensive Methods for Computer Vision. Graz, Austria.

Zou F, Gelfond JA, Airey DC, Lu L, Manly KF, Williams RW, Threadgill DW. (2005) Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. Genetics 170:1299-1311.