The Collaborative Cross: Rationale, Implementation, and Costs

Robert W. Williams and Gary A. Churchill

"Most genetic traits of interest in populations of humans and other organisms are determined by many factors, including genetic and environmental components, which interact in often unpredictable ways. For such complex traits, the whole is not only greater than the sum of its parts, it may be different from the sum of its parts. Thus, complex traits have a genetic architecture that consists of all of the genetic and environmental factors that contribute to the trait, as well as their magnitude and their interactions."

National Institute of General Medical Sciences, 1998, Complex Trait Workshop Report: www.nigms.nih.gov/news/reports/genetic_arch.html

---

Abstract

The goal of the Complex Trait Consortium (CTC) is to promote the development of resources that can be used to understand, treat, and ultimately prevent pervasive human diseases. Essentially all human diseases are complex in the sense that incidence, severity, and outcome are determined by interactions among many gene variants and environmental factors. Cancer, diabetes, heart and lung disease, Alzheimer's, and infectious diseases fall into this category. Members of the CTC have spent the past year devising a detailed plan to generate a collaborative genetic resource that will greatly accelerate the study of mouse models of human disease. The resource is called the Collaborative Cross. This Report explains our objectives and provides an outline of the methods and costs associated with making the Collaborative Cross.

Introduction

Mechanisms that underlie disease susceptibility and progression in humans and mice are with few exceptions influenced by intricate molecular networks that interact with numerous developmental and environmental factors. While reductionist tactics continue to provide an astounding wealth of mechanistic insight, we believe that computational, statistical, and genomic resources are now sufficiently mature to address these questions in the context of an experimental model system that more accurately reflects the genetic structure of human populations (SM Williams et al., 2004). Global analysis of complex biological systems can be implemented most efficiently by using experimental designs that employ multifactorial perturbations (Jansen, 2003). The Collaborative Cross is intended to provide such a resource for the new synthetic phase of genetic investigations that we see opening up this decade. We anticipate that the analysis of the Collaborative Cross will have a direct, positive, and long lasting impact on the diagnosis and treatment of common and chronic human disease, including cancer, diabetes, cardiovascular disease, obesity, behavioral abnormalities and neurodegenerative diseases.

From their initial use in cancer genetics in the early 1900s, mice have been highly effective models for a broad spectrum of human diseases. Mice breed rapidly (3–4 generations/year) and can be housed or cryopreserved economically and in large numbers. Inbred strains and other genetically defined lines are now exploited in an astonishing variety of experiments that are having an intense impact on our understanding of normal variation and the genetic basis of disease. The availability of sequence data for five common strains (C57BL/6J,

DBA/2J, A/J, 129S1/SVEVJ, 129X1/SvJ) has provided direct access to the genetic variants that underlie complex phenotypes. In many instances it is becoming clear that the same genes that modulate phenotypes in experimental crosses can also contribute to disease related phenotypes in human populations. Loci implicated in studies of mice are providing human geneticists with a stream of candidates suitable for association and linkage studies.

A wide range of resources exists for the dissection of monogenic and multigenic traits in mice. However, these resources share one dominant characteristic, namely, the intentional reduction of genetic complexity compared to human and wild populations. An explicit design goal of essentially all mouse resources has been to convert polygenic traits into more tractable sets of near-Mendelian traits; congenic and consomic strains being powerful examples. While current resources are optimized to study actions of isolated genetic loci on a fixed background, they are less effective for studying intact polygenic networks and interactions among genomes, environments, pathogens, and other exposures. The Collaborative Cross represents a radical departure and novel extension of current resources that has been specifically designed for collaborative and integrative analysis of complex polygenic gene-phenotype networks. Using this cross it will be possible to make effective headway with difficult challenges in genetics including epistatic interactions, gene pleiotropy, and norms of reaction.

The Collaborative Cross will provide a common reference panel for investigation of mammalian biology at a systems level. Its will be used to develop and test hypotheses of gene function across different scales, organ systems, ages, and environments. This in turn will lead to a dramatic shift in the way we approach human health and disease. The use of the Collaborative Cross as reference panel for complex trait analysis contrasts most current efforts relying on crosses that generate transient populations. Results from isolated intercross and backcross mapping populations do not lend themselves to data integration. In contrast, the Collaborative Cross will provide the research community with a focal point for research. Cumulative and integrated data will play a central role in new systems level approaches to biological questions.

A large panel of recombinant inbred (RI) mouse strains represents a powerful resource for collaborative genetic studies (Gibson and Mackay, 2002). Similar resources have been developed or are currently being developed for studies in maize (~ 4000 RI lines by Pioneer) and Arabidopsis (~1000 RI lines). RI lines are particularly efficient because they need to be genotyped only once, removing a technical (and statistical) hurdle and opening up genetic methods to a much larger community of researchers. More significantly, isogenic RI lines can be distributed essentially in perpetuity. The same genetic individual can be studied repeatedly, an essential feature for studies of gene-by-environment interaction and when traits have low heritability.

To broaden the utility and diversity of the Collaborative Cross, members of the CTC have decided after long discussion, argument, and simulation to construct RI strains using a starting material from eight highly divergent inbred strains of mice. With genetic contributions from eight parental strains, including three wild strains, the Collaborative Cross will capture an abundance of genetic diversity in a single set of RI lines. By generating a sufficiently large number of lines, we can guarantee high resolution in mapping applications and large sample sizes that will be essential for applications such as the mapping of extended networks of epistatic interactions. Moreover, it will be possible to generate a tremendous combinatorial diversity of genotypes in the form of F1 progeny of the RI lines (RIX). These RIX animals will have reproducible genotypes with low formal inbreeding coefficients, more representative of the genetic state of human populations. A set of one thousand RI lines can generate as many as one million distinct but genetically

well defined and reproducible hybrid lines that will represent a vast resource for the discovery of new animal models of human diseases.

Now is an appropriate time to begin the development of the resources that will be needed to push genetic studies of the mouse forward. With the best current breeding practices it takes approximately seven years to produce 8-way RI strains. Once developed, each strain represents a long-lived resource that can be used again and again to generate cumulative information. Each new data set can be immediately integrated into and interrogated with respect to all previous data sets collected on the collaborative cross. Given the rapid pace of development of new technologies and methods for interrogating mammalian phenotypes, it is hard to predict where we will be seven years from now. One thing is certain; many new phenotypic assays will be available and many of the phenotypes that we currently obtain only with great effort and expense will have evolved to a point where high throughput applications will be economically and logistically feasible. It will be wise to generate the Collaborative Cross in anticipation of rapid developments in assays of transcriptomes, proteomes, and metabolomes. Systematic and sophisticated analyses of molecular-phenotypic networks will be greatly enriched by the combination of high throughput phenotyping of the Collaborative Cross.

The remainder of this report is divided into three parts.

Part 1 (Aims and Rationale) summarizes the aims of The Collaborative Cross and answers the question: What will this resource accomplish in terms of precision and power of mapping QTLs? How will researchers exploit this cross over the next several decades to study networks of genes that contribute to complex phenotypes? What impact do we expect the cross will have on biomedical research and the diagnosis and treatment of human disease?

Part 2 (Implementation) summarizes the design and implementation of the Collaborative Cross. This section incorporates initial results of simulations and calculations concerning the CTC consensus recommendation for how to implement a Collaborative Cross.

Part 3 (Costs, Housing, Distribution, Archiving, and Maintenance) is an outline of costs and how the effort leading to the Collaborative Cross can be structured and distributed across a diverse and international community of researchers using mice in a wide variety of research projects. We provide provisional estimates of what it will cost to make and maintain the cross, how generating the cross will be organized and controlled.

PART I. AIMS AND RATIONALE

In this section we briefly review eight aims that were considered in designing a single integrated cross for complex trait analysis.

1. **Broad utility for the investigations of many complex traits**
2. **Freedom from genotyping and lower barriers to initiating genetic studies**
3. **Unrestricted access to strains, tissues and data**
4. **Improved power and precision for QTL mapping**
5. **Powerful new approaches to the genetic analysis of complex systems**
6. **Analysis of gene-by-environment interactions**
7. **A resource for systems biology**
8. **New and more realistic animal models of human disease**

1. Broad utility. It is not possible to design a single resource that can be used to address all of the questions posed by the biology of complex diseases. However we expect the Collaborative Cross to be directly or indirectly useful to a majority of researchers using mouse models and to many researchers and clinicians studying complex human diseases. The Collaborative Cross is designed to incorporate a broad spectrum of natural allelic variants that have been sequestered within inbred strains over the past several centuries. These genetic variants provide a collection of natural perturbations which, when present in many distinct combinations, will result in tremendous phenotypic diversity. Thus, for essentially any common human disease, relevant phenotypic and genetic variation will be present in the Collaborative Cross population.

A recent analysis of haplotypes of several major strains and subspecies of mice (Wade et al., 2002) emphasizes that some of the most common inbred strains are hybrids of a small number of ancestral haplotypes. By combining eight strains into a single cross we gain a tremendous resource in the form of both the haplotype structure that have been archived in each inbred strain, as well as the novel alleles fixed within the past few hundred generations of inbreeding in each of eight lines. After extensive and animated discussion members of the CTC have recommended that the Collaborative Cross include the use of three wild-derived inbred stains. These wild strains will significantly enrich the allelic diversity of the cross.

Why an 8-way Cross? We have settled on a cross that combines eight strains (an 8-way cross) as a good compromise between embracing genetic diversity while coping with computational and statistical challenges of a multiallele cross. Estimates by Dr. James Cheverud, based on multidimensional scaling analysis of a survey of 48 strains, has demonstrated that a well chosen set of eight strains (for example, C57BL/6J, A/J, CAST/Ei, NOD/LtJ, NZO, WSB/Ei, PWK/Ph, 129S1/SvImJ) can sample more than 20% of the total genetic diversity based on SSLPs and up to 50% of the SNP diversity. With the inclusion of three wild derived we expect that the cross will segregate one polymorphism every 200 to 250 bases of DNA. Approximately one half of these variants will be SNPs. Ongoing simulation studies by Dr. Daniel Gaile and colleagues are providing us with more precise estimates of power and optimal analysis strategies for exploiting a large 8-way RI set.

Trade-offs. There are inevitably trade-offs between numbers of progenitor strains and the power to dissect trait variance with a given sample size. Mouse geneticists currently have a wealth of mouse resources in which genetic complexity has been intentionally reduced to improve power. Conventional inbred strains, chromosome substitution strains, congenic sets, recombinant congenic sets, and 2-way recombinant inbred strains are some examples of crosses in which good power can be obtained with relatively modest sample size. In contrast, we currently lack community resources that incorporate the kind of allelic diversity often encountered in studies of humans. The Collaborative Cross provides this complementary and more complex resource. In order to achieve good power in the face of relatively high genetic diversity, the cross must be large by current standards.

2. Freedom from genotyping. Members of the CTC will be responsible for genotyping the Collaborative Cross. A panel of 1000 8-way RI lines will carry roughly 100,000 unique recombination breakpoints at an average spacing of ~25 kilobases. Initial genotyping efforts would employ a marker density sufficient to place all recombinations into 1 megabase (Mb) bins. Eventually all recombination breakpoints would be located as precisely as possible given the availability of markers and limitations imposed by haplotype diversity among the parental strains. Ultimately, we anticipate that each of the eight parental strains will be sequenced either selectively or completely. The effort required to fully genotype the Collaborative Cross needs to be undertaken only once and the data produced by this effort will be freely available to all investigators who utilize these mice.

The availability of mice from a fully genotyped panel will greatly reduce the barrier to entry for new studies, particularly for non-geneticists. In addition, the common reference panel will have been extensively characterized for many phenotypes that have already been acquired. New work will start with a strong genetic and phenotypic foundation. Data sets, results, and understanding will accumulate synergistically. We expect more widespread usage of methods that have largely been the domain of a small group of statistical geneticists.

3. Unrestricted access to strains, tissues and data. Use of strains of mice, tissues, genotypes, and community-acquired phenotypes must be available without restriction, but with adequate cost recovery and subject to availability. In the case of resource limits and conflicts, a committee will rank requests and devise ways to share material and mice more effectively. Likewise, the resource should be useable by labs that do not have access to large mouse facilities. This will be accomplished through resource populations maintained at distribution research centers that will also provide visiting research teams access and equipment for phenotyping resource populations. It is crucial that all CTC RI strains be part of a single publicly available resource without legal restriction. The SNP Consortium's solution to the issue of ownership and rights will be a useful model to study.

4. Improved power and precision for QTL mapping. The Collaborative Cross was designed to achieve 0.1 cM precision—equivalent to approximately 200,000 bp—when mapping the types of QTLs with which most investigators now routinely work (additive effects of > 0.25 standard deviation units). This level of precision will exist in a single well genotyped panel. It will therefore generally not be necessary to generate custom secondary mapping resources. Achieving this level of precision with high statistical power in a single mapping panel requires archiving approximately 100,000 independent recombinations in the set of strains. The critical issue of statistical power also makes it necessary to generate large numbers of strains. A set of 1000

How many mice will need to be phenotyped? The number of individuals that need to be phenotyped is a function of the heritability, the number of QTLs influencing a trait, and of course, the number of strains available from stock at a reasonable cost. A Mendelian trait is most efficiently mapped using only a single animal from each of a large number of strains. In contrast, a trait involving an incidence, dose-response function, or a variability score may require a sample size of 10 to 40 individuals per line. The number of animals and strains that are typed will often be influenced by factors other than efficiency of gene mapping. In general, we note that it is better to study fewer individuals and larger numbers of strains than more individuals per line and smaller numbers of strains.

Most QTL studies begin with an initial step of QTL discovery and low precision QTL mapping, followed by one or two additional stages of mapping designed to achieve higher and higher precision. As a practical matter, a QTL mapping study using the Collaborative Cross may also need to be conducted using a two or three stage step-down approach. In the first stage, a set of 100 to 200 RI strains would be phenotyped to locate the approximate positions a small number of major QTLs. The 2-LOD support intervals of these QTLs will often be in the range of 2 to 20 cM. The more refined analysis of each of these QTLs will involve a second stage of analysis that exploits a different set of 100 to 200 RI strains (but also from the Collaborative Cross) all of which have recombinations in targeted chromosomal intervals. This will provide confirmation of the QTL and simultaneously refines the QTL position. A third stage of mapping, using all strains with relevant recombinations, could be applied as needed to achieve the maximum possible resolution. If phenotyping is relatively efficient, it should be possible to confine a QTL with effect size >5% of the total variance to an intervals of 0.5cM (~1 MB) using fewer than 1000 animals.

strains containing 100,000 breakpoints is a far more powerful research tool than 100 strains containing the same number of breakpoints. Precisely mapped QTLs and interactions will provide direct leads to the most relevant human candidate genes without the need for the extensive and costly follow up that is currently required to "clone" QTLs (Glazier et al., 2002; Flaherty et al., 2003).

Epistatic interactions are often difficult to detect and characterize due to the small sample size of most experimental crosses. There is a pressing need for a resource with enough power to systematically map sets of interacting QTLs. This consideration has been a major factor influencing the CTC recommendation to generate a large number of strains. Mounting evidence suggests that gene-gene (as well as gene-environment) interactions play a critical role in complex disease etiology (Reifsnyder et al., 2000; SM Williams et al. 2004). Epistatic interactions also provide strong biological constraints that can simplify candidate gene analysis.

With an 8-way cross, a private allele (an allele present in only one of the eight strains) will typically appear in 32 of 256 RI strains. Four RI strains will typically be homozygotes at any two unlinked loci in the desired combination. This small sample size is not sufficient to test epistatic interactions among private alleles. In contrast, a set of 1024 RI strains (a 1K set) leads to much higher *Ns* and excellent prospects for analysis of epistasis. The 1K set will typically include at least 16 two-locus RI strains with any given pairwise combination of private alleles. For biallelic loci (four *A*-type and 4 *B*-type alleles in an 8-way cross), three- and even four-way interactions can be explored effectively. For example, in a best-case scenario, 4-way epistatic interactions can be tested in 16 RI strains. The obvious point to make here is that epistatic effects are major factors motivating the need for a large number of strains in Collaborative Cross.

5. New approaches to the genetic analysis of complex systems. The focus of human genetics is shifting from Mendelian traits to complex diseases. Understanding why humans differ so greatly in their genetic vulnerability is a major challenge, and finding answers quickly and at a reasonable cost is vital if functional genomics is to live up to its promise. Perhaps even more significant than the yield of closely mapped candidate regions will be the greatly enriched analysis of gene pleiotropy and molecular networks that will become possible with access to a large reference panel of recombinant inbred strains. Biologists are already beginning to exploit recombinant inbred strains to define the covariance structure among traits at many different levels of organization. These networks of phenotypes are extremely valuable tools in functional genomics (Nadeau et al., 2004).

> "Due in part to research advances, the burden of disease is now shifting from more acute and lethal forms of disease to chronic illness. Our success in conditions like myocardial infarction and infectious diseases is leading to better survival rates. As the results of such prolonged survival and aging of the population, the incidence of chronic and long-term diseases, such as congestive hear failure, cancer, Alzheimer's disease, Parkinson's disease, diabetes, and obesity, among others, in increasing… "
>
> "It is imperative that we develop more comprehensive strategies to address such emerging challenges. In all likelihood these strategies will require a better understanding of 1) the series of molecular events that lead to disease in the hope of affecting its course before the disease develops, so-called Molecular Prevention; 2) the interactions between genes, the environment, and lifestyle as they relate to the etiology and progression of disease; ways of delaying the onset of the disease and/or ways to reduce the severity of its course and its impact on quality of life."

Dr. Elias Zerhouni. *Opening Statement to Congress on the FY 2004 Budget Request*

6. Analysis of Gene-by-Environment Interactions. As highlighted by Dr. Zerhouni's statement to Congress, there are many situations in which we need to know far more about how environmental factors—diet, stress, infection, and toxins—interact with gene variants. Ethical considerations protect humans from many highly informative studies that can be readily carried out using mice. Factors can be added and subtracted singly or in combination in mice to test causal relations and to expose additional gene variants that might otherwise remain hidden in the background.

Gene-by-environment interaction (GXE) is a crucial problem in genetics that has been difficult to study in any mammalian population. In the last two years gene-by-pathogen interactions have also unfortunately become far more important. GXE analysis requires the use of isogenic lines that can be studied in large numbers in different environments and 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. The Collaborative Cross solves this problem by providing a large collection of isogenic lines that can be assayed under different sets of environmental conditions.

7. A systems biology resource. The CTC will collect, curate, and distribute large amounts of phenotype data. A major enticement to motivate the adoption of the Collaborative Cross will be to acquire high quality data on the transcriptome, proteome, and metabolome of 10 to 25 key organ systems and tissue types in adults of both sexes across a substantial faction of the panel (ideally all extant strains). The number of RI and RIX lines that can be systematically phenotyped as part of this effort will depend on scientific motivation and funds. High throughput array and proteome methods will be much more practical by the time the RI set is ready for use and distribution. Our intent is to ensure that all investigations of complex traits begin with extensive data on transcripts and proteins in several major organ systems.

8. New Animal Models of Human Diseases (Isogenic but non-inbred lines). We want to retain the advantages of mouse experimental genetics, including large sample size, known genotypes, and replicated genomes. But we would also like to be able to model the complex and non-inbred genetic structure of human populations more faithfully. The Collaborative Cross makes this possible for the first time. Although RI strains are fully inbred, this limitation can be finessed by systematically producing isogenic F1 intercrosses from the RI parents. These F1 hybrids are called RIX lines (Threadgill et al., 2002). RIX hybrids made using 8-way RI stock have formal inbreeding coefficients of 0.125 and thus realistic levels of heterozygosity compared to human populations. Inbred lines have been criticized as models of disease in this respect (Hartwell, 2004) as they often lack the natural buffering that occurs in animals with reasonable amounts of heterozygosity. Thus, an RIX mapping panel has some similarity to typical human populations in consisting of a complex and non-inbred admixture of ancestral genomes. Unexpected phenotypes can emerge by mixing genomes. With one million distinct genotypes at our disposal, the cross will provide a wealth of new models for human disease.

Furthermore, using the Collaborative Cross strains, one can generate cohorts of isogenic but non-inbred RIX animals that have specific multi-allele genotypes—what one might call *designer mice*. For example, if we hypothesize that three genetic loci are interacting in their effects on a trait, one will often be able to construct a set of RIX animals with specific

combinations of alleles at all three loci in a single generation. This provides a novel and efficient approach to prove complex genetic models by synthesis.

PART II. IMPLEMENTATION

Synopsis: The breeding strategy for the Collaborative Cross employs the most direct method to combine genomes from eight parental strains that ensures that all recombinations in the final set represent independent meiotic events. Each of these 8-way RI strains will incorporate an average of approximately 100 recombinations; twice that of a typical 2-way RI strain and representing a 7x genetic map expansion compared to intercross populations. We will initiate production of a balanced set of RI strains (a total of 8x7x6x5 or 1680 lines) and we expect to push ~1000 RI strains through complete inbreeding (~F25). Selected tissues and DNAs will be collected at intermediate generations leading to each RI strain. We will routinely genotype many of these intermediate generations. Error checking and quality control will be continuous and stringent. Finally, we will cryopreserve several hundred embryos of each strain as soon as inbreeding is complete.

Starting material for each incipient RI strain consists of eight parental strains, or more specifically, of sets of four F1 intercrosses. These four F1s are intercrossed over two generations to merge and recombine parental haplotypes. Each "haplotype funnel" is an independent and easily replicated process defined by a unique pattern of mating. We are striving to minimize cost and cage counts throughout this cross while generating a total of approximately 1000 surviving strains. We have opted not to struggle to maintain those strains with low viability. In order to compensate for inevitable attrition, we plan to initiate significantly more strains than we anticipate being able to retain. 1680 funnels (8x7x6x5) will be initiated. We expect ~1000 will survive through inbreeding. Importantly, the breeding design is modular, simple, and robust. This will allow for efficient distributed implementation of the cross with no need to ship animals between sites.

The simple starting structure of the cross makes it feasible for a single site to initiate the cross. CTC coordination will involve subdivision of breeding effort at steps G2, G3, etc. Colony informatics, phenotyping, and genotyping will need to be tightly coordinated. Friendly competition and collaborative help should improve morale and performance. This distribution of breeding sites will also provide insurance against colony infection or other catastrophes.

In general, our workgroup supports and encourages efforts to ensure international participation in the CTC resource development. One caveat is that all sites (US and elsewhere) should have the necessary cage capacity to place stock behind a pathogen barrier once the inbreeding begins (see below). We discussed distributing efforts across 4–6 major sites. This number of centers would provide adequate dispersion of animals, expertise, and interest, and would also protect the effort.

Strain selection. Members of the CTC and outside advisors met at the Whitehead Institute in September 2003 to discuss the selection of strains for inclusion in the Collaborative Cross. There was surprising unanimity in the selection process and 10 strains were selected as prime candidates. Five factors weighed heavily in the choice:

1. Nominations made by members of the CTC over a 6-month solicitation period, as well as the likely use of the Collaborative Cross by those who are in the midst of mutagenesis and studies of knock-out lines. This cross must serve a large research

community that is interested in mapping modifier loci of Mendelian traits; not only those interested in complex and quantitative traits.

2. Availability of complementary resources. In our discussion on strain selection we emphasized the availability of sequence data from both the Mouse Genome Sequencing Consortium and Celera Genomics. One intriguing possibility was raised; namely a complementary sequencing initiative for all eight strains that are ultimately selected. This would ideally involve the acquisition of at least 4X coverage of each of seven strains (other than C57BL/6J which is now is sequenced with roughly 8X coverage). The current Celera data is approximately 1.3X for DBA/2J, 1.5X for A/J, and 1.5X collectively for the two 129 substrains. Our final consensus was to include C57BL/6J and two of the four other partially sequenced strains.

Other complementary resources. We also considered complementary mapping panels and strain resources, including existing RI sets (primarily AXB and BXD), consomic and congenic strain sets (B6.A, B6.129, B6.PWD, B6.D2, B6.CAST), BAC libraries, and other resources that could complement the Collaborative Cross.

3. Genetic diversity. Dr. James Cheverud presented an analysis of genetic diversity among strains using complementary SNP data from GNF (Pletcher and Wiltshire) and microsatellite data from the CTC (Morahan, Williams, Lu, and Gu). Since SNPs were selected from Celera sequences they predictably showed a bias that arises due to selection (A/J, DBA/2J, 129, and C57BL/6J are apparent outliers). The SSLP data generated by Morahan and Williams for the CTC were also noted to have some biases. Nonetheless, the microsatellites (about 750 across 47 strains) clearly delineated the known major grouping of mice and suggested some strains as being particularly 'far from center.' Lastly, a set of highly detailed (but very localized) sequence data were presented by Dr. Fernando Pardo-Manuel de Villena. His data provided insight into the level of genetic diversity among strains (only half of the nucleotide differences are SNPs). He stressed the substantial gain in allelic diversity that we can achieve by including wild strains.

4. Phenotypic diversity weighed in less heavily. Drs. Ken Paigen and Molly Bogue demonstrated that any choice of strains will show roughly the same level of phenotypic diversity. Moreover, we expect that the genetic diversity of the cross will deliver much new phenotypic diversity.

5. Breeding performance. Some strains don't breed well; others suffer in combination. For example PWD x B6 males are thought to be infertile. Other strains carry deleterious alleles that could potential cause high mortality among RI lines during the breeding process.

The leading candidates for the five laboratory strains are:

1. **C57BL/6J**: regarded by most mouse geneticists as obligatory
2. **129S1/SvImJ**: also regarded by a vocal community as obligatory
3. **A/J** OR **DBA/2J**: Both strains have marvelous resources (animals and sequence), but in the interests of genetic diversity we prefer not to include both of these Castle strains. We are leaning toward A/J.
4. **NOD/LtJ**: our Swiss Webster representative and a favorite of immunogenetics
5. **NZO/HILtJ** or **BALB/cByJ** or **KK/HIJ**-- KK adds greater genetic diversity, but NZO is a favorite for immunogenetics. We are leaning toward NZO.

The leading candidates for the three wild-derived inbred strains are:

6. **CAST/Ei** (*M. musculus castaneus*; a *M. m. m.* lineage members)

7. **CZECHII/Ei** (*M. musculus musculus*), **PWD/Ph** or **PWK/Ph** (*M. musculus musculus*). We are leaning toward PWD/Ph due to Dr. Jiri Forjet's BAC library and other resources. We are currently generating F1 progeny from these strains to test fertility.
8. **WSB/Ei** (*M. musculus domesticus*), a wild derivative of the common *M. m. d.* subspecies.

Stages of Making the Collaborative Cross

G0. Parental Strains. We will cryopreserve embryos of the eight parental strains within the first year of project initiation (2004–2005). We must preserve the original ancestral founder stock. This is necessary because these eight ancestral strains will themselves drift slowly away from their Year 2005 "vintage" due to the accumulation and fixation of novel mutations. Without frozen founder stock we will have difficulty adding new strains without introducing systematic subset differences. For example, the original 26 BXD RI strains introduced in the late 1970s differ as a group from the new set of 10 BXD strains introduced in 1999 for reasons that are known to involve the fixation of new alleles in C57BL/6J and DBA/2J over a 20-year interval.

G1. Production of F1. We will produce a full diallel cross of the eight strains. Imagine an 8 x 8 chess board in which each of the squares is populated by an F1 intercross and the diagonal consists of parental strains. A total of 56 F1s (28 reciprocal pairs) will be produced at a single location (JAX). The F1 animals will be shipped as breeding trios (1 AXB male with 2 CxD females). Two complementary trios are required to initiate an 8-way RI line. This first step could begin as early as Fall 2004. The CTC and the Jackson Laboratory (Dr. Gary Churchill) will coordinate distribution of F1s. Dr. Churchill is currently performing a diallel cross among eight inbred strains to produce different types of F1 hybrids. These matings can generate >100 F1 progeny per week using only two racks of cages.

> ***Use of F1 animals.*** The isogenic F1 progeny are a valuable resource in their own right and they can be used for in-depth phenotyping and mapping efforts. These animals do not incorporate any recombinations and have formal inbreeding coefficients of 0. The primary function of an F1 analysis will be to study genetic architecture as well as maternal and cytoplasmic effects. However, with a dense haplotype map of the eight parental strains, the diallel progeny may be useful for some types of mapping.

G2. Production of a G2 (a 2-way by 2-way cross). The 56 types of F1 progeny produced in the first generation (G1) will be intercrossed. AB will be crossed to CD and to DC; BA will be crossed to DC and CD, etc. We will generate a set of all 1680 non-intersecting pairs (pairs that do not share parental strains). This complete coverage ensures that we will be starting with a balanced representation of the sex chromosomes and of all haplotypes in all combinations. Single chromosomes of G2 progeny will have AxB recombinations or CxD recombinations (not both). Each pair of genomes (A and B) will recombine during meiosis 60 times—30 AB and 30 BA pairing across the whole set of 1680 crosses. These G2 progeny are themselves a remarkably valuable mapping resource. We will definitely want to obtain DNA samples from all individuals that are used in subsequent breeding and we would also hope to cryopreserve selected tissues for eventual transcriptome and proteome studies.

> ***Design note.*** We considered the possibility of intercrossing ABF1 animals to produce F2, F3, or even F4 populations. F4 advanced intercross animals would incorporate 2x the recombination load of a conventional ABF2 and they could be generated without any shared breakpoints. In principle, such F3 or F4 animals could be used to initiate the 1680 G2 strains above. However, simulations

indicate that the gain is negligible and we have therefore decided against this method in favor of an immediate cross of ABF1 x CDF1. We can generate sufficiently high numbers of recombination events while ensuring balanced representation of parental alleles in a shorter period of time.

G3. Production of G3 (4-way by 4-way). ABCD will be crossed to EFGH to generate G3 offspring. The number of ways to breed non-intersecting 4-way G2 parents is 8-factorial (40,320), but we will limit the G3 to a balanced sub set of 1680 strains (e.g., ABCD to EFGH; ABEF to CDGH). As with the G2 production, cage space will be limited at the breeding sites and it is planned to only fund the production of a single G3 litter. Again we will encourage CTC investigators and other potential collaborators to make use of this unusual G3 resource that should be available starting one year into the project

G4. Production of G3F1 (8-way by 8-way cross). G3 8-way animals produced in the previous generation will be intercrossed. These can be sib matings without any loss of productive recombinations. The G3F1 progeny of sib-mated G3 parents will often have single chromosomes with [(AxB)x(CxD)] x [(ExF)x(GxH)] recombination blocks. These animals will have an inbreeding coefficient of 0.5 and are in many ways like an F2.

G5. Production of G3F2. Second sib mating of 8-way parents.

G6–G13. G3F3 through G3F10. This is the initial critical phase of inbreeding. At F10 the strains should be about 75–85% inbred We expect to incur losses during this phase of breeding that will reduce the numbers of surviving to lines to approximately 1000. Loss of lines will free additional cage space for more aggressive maintenance of the surviving lines. DNA will be retained from breeding pairs of all lines, including those that go extinct.

Genotyping note. At F10 it might be worthwhile to genotype the pools of DNA from 4 to 10 animals at 500 to 1000 markers to gauge the ultimate balance and performance of the whole RI set. Tissues from these animals could then be used for molecular phenotyping assays.

G14–G20. The final stages of inbreeding. We might consider the application of speed inbreeding at this stage with selection against heterozygotes and fixation of alleles and haplotypes that will improve balance of whole set. The efficacy of speed inbreeding will depend primarily on genotyping cost. The gain in recombinations load and genotype balance could be substantial. With speed inbreeding the F15 may be considered fully inbred.

Selection of Major Breeding Centers

Prior to the initiation of the cross, the major breeding sites must be chosen and committed to six years of developing the RI strains. For example, four sites might each be individually responsible for the initiation of 420 strains. Members of the CTC have already identified several excellent candidate breeding centers including the Oak Ridge National Laboratory (Dabney Johnson and colleagues), the Jackson Laboratory (Gary Churchill and colleagues), Pennsylvania State University (Byron Jones), and the University of Tennessee (Robert Williams and colleagues), and UNC Chapel Hill (David Threadgill and colleagues). We also have colleagues in Australia (Grant Morahan) and in the United Kingdon (Jonathan Flint and colleagues) who are willing to initiate breeding programs as part of the Collaborative Cross.

Scientific Use and Analysis

The design of the Collaborative Cross enables CTC members to engage in cutting-edge complex trait analysis throughout the project; not just after the lines have been fully

inbred. However, the amount of science done during the production phase is entirely dependent on a budget for genotyping and phenotyping intermediate generations.

Year 01. In year 01 we focus on the phenotypes associated with the eight inbred strains and the associated 56 reciprocal F1 intercrosses. This is essentially a phenome project of vigorous non-inbred but isogenic strains of mice for which we expect to have superb haplotype maps. The intent here is not to duplicate the efforts of the mouse inbred strain phenome project but rather to supplement this effort, especially for phenotypes not currently included in the database. One such area would be the development of a transcriptome database for various tissues. Once the inbred strain database is developed, attention should turn to the F1 intercrosses, which can be easily recreated at numerous sites. Issues of dominance, maternal effects and so on are easily measured. The quantitative genetics core of the consortium will be responsible for helping investigators analyze these data and for developing new analytical tools. These services will be made available to all investigators studying the eight strains and associated crosses used to form the RI panel.

Year 02-05. Collaborative QTL mapping using G2 through G6 individuals. The data obtained from the analysis of the eight inbred strains and the reciprocal F1 crosses will generate numerous hypotheses as to phenotypic distribution among the 4-way crosses. As noted previously, the number of 4-way animals that would be available to a single investigator may be limited; however, sufficient numbers should be available to obtain strong preliminary data to justify recreating crosses of interest. A similar argument applies to moving from the 4-way to the 8-way crosses. This point is important since it will provide a foundation for understanding the genetic architecture of the final RI strains. This effort will segue perfectly into the eventual analysis of RI and RIX strains.

Year 05 to 07. Analysis of near isogenic RI strains with inbreeding coefficients above 0.95. The near isogenic strains will be particularly interesting, especially in view of the fact that the precise locations of many QTLs will already be known. Any QTL located in a region that is still segregating can be converted into sibling strain pairs in which the QTL is fixed for alternate alleles. This method can be used to maximize recombinations and to explore QTL effects.

Year 07 onward. Start of systematic phenotyping of fully inbred RI and the associated RIX crosses. Of the expected 1K strains that will be viable at year 07, those strains showing the highest level of recombination will be chosen for the core set of 256.

PART III. COSTS, HOUSING, DISTRIBUTION, ARCHIVING, and MAINTENANCE

Cage costs

The costs for producing F1 animals are comparative modest: Assuming that we plan to ship two trios of each F1xF1 cross to each of four sites, shipping costs will be about $20,000. We will make all 56 F1 crosses from the eight parental strains. This will require about 250 duplex cages. Cage costs to generate and maintain the parents and the F1 offspring (until 6-weeks-of-age) are ~$25,000.

In total across four breeding centers, we will initiate all 1680 types of G2 progeny from the F1 by F1 crosses. We will set up two independent cages of each type of G2. This requires 3360 duplex cages. Each cage will be used to produce about 10 animals. At the G3 generation we will double our cage usage to roughly 2000 cages per site for the production

of RI lines. We are assuming a cost of no more than $0.50 per diem. Per diem charges at each of four sites would be approximately $350,000/year. Experienced colony managers will be hired to oversee production at each site at an additional cost of $60,000 to $70,000 per year. We anticipate that these colony costs would stay relatively constant over the period of RI development.

Estimated costs: $1.5M/yr.

Genotyping

The 6720 G2 animals used in breeding will be genotyped at 100 SSLPs, and at a cost of $40 x 6720 = $270,000. These animals will be used to set up 3360 G2 x G2 breeding cages in the next step.

Like the G2 animals, the G3 animals are also all non-inbred segregating heterozygotes. However, they have twice the load of recombinations. Since these animals are more recombinant we would ideally double the marker density to 200 per animal for a cost of $500,000. Note, that the genotypes of the G2 parents will provide additional information about the genotypes of the G3s.

DNA will be collected from all breeding pairs in subsequent generations for genotyping on an as needed basis for troubleshooting and genetic quality control.

Estimated costs: $0.5M/yr

Tissue Collection and Additional Phenotyping

Tissue will be saved from all animals, but DNA will be extracted and genotypes for only 20% of these 33,600 animals (one male, one female per cage). We expect any major phenotyping efforts will fall outside the scope of the breeding project, a minimal level should be maintained for quality control and scientific value. Excess mice from production will be made available for phenotyping by other investigators on a cost recovery basis.

Estimated costs: $0.5M/yr

Bioinformatics + Mouse Tracking. Colony management of a cross this large will require a robust and powerful mouse tracking system. We have developed and implemented a powerful tool, the Mouse Tracking System (MTS), at The Jackson Laboratory. The system has been extensively tested in multiple labs. One group has generated more than 120,000 mouse records. The system is designed assist mouse room staff in the day-to-day routine of mouse colony management and is capable of tracking phenotype and genotype data. We anticipate costs associated with maintenance, upgrades and modifications of MTS.

In addition we will continue the development of database and analysis tools required for management and interpretation of data generated by the Collaborative Cross at intermediate and final breeding stages. By concurrently developing analysis tools we will ensure that tools are available, tested, and reliable on the same day that the first finished mice are shipped to a new investigator.

Estimated cost: $0.5M/yr

Cryopreservation. As noted, above it will be essential to cryopreserve founder stocks of the Collaborative Cross. Although the number of lines to freeze down is small (eight), some of these lines may be difficult and we will want to ensure that a large number of embryos are available for multiple future recoveries. Current costs for freezing an inbred strain are about $2000. With a $25,000 budget we can be sure of cryopreserving an number of specimens of each founder line.

Cryopreservation of 1000 strains will be required at the end phase of the project. At this time, it is likely lines will somewhat out of synch in their approach to full inbreeding. After consultation with staff at The Jackson Laboratory and at The Oakridge National Laboratory, we have determined that it will be feasible to freeze down ~500 strains per year at a cost of $1000 to 2000 per strain. Much of the expense associated with preservation is due the need to expand the stocks to obtain ~20 females for ovary harvest. However, it is usually possible to generate a sufficient number of embryos for 10 recoveries using only two trio matings (Dabney Johnson, pers. comm.). We will most likely not make heroic efforts the freeze the most difficult lines.

Estimated cost: $2M, one time.

Which Strains to Keep. A core set of 256 RI strains will be chosen as the consensus mapping resource set and maintained at high availability. This set and derivative RIX strains will be optimized for recombination load and distribution in gene-rich regions and for balanced representation of the eight haplotypes and alleles. The core set of RI strains will be housed and bred at several international sites. The full 1K set will be retained at maintenance level (~6 cages/line). The 1K set will be used for systematic high-throughput molecular phenotyping at major centers and for final mapping projects that require the ultimate resolution. Strains with reproductive problems will be cryopreserved. Prior to cryopreservation we will collect tissue for an archive of normal adult RI and RIX tissues.

Distribution of RIX strains. We would like to encourage scientists to exploit the non-inbred RIX derivatives of the RI strains. RIX strains provide users with twice the haplotype diversity and twice the load of recombinations of an equal number of RI strains, with lower within-line variances, with access to dominance signal, with reduced sensitivity to collateral damage from recessive alleles (e.g., blindness in behavioral studies), and with the ability to assess parental effects. Viability and robustness of RIX progeny is also better, and we will be able to provide and distribute them more inexpensively to researchers. The RIX progeny of an 8-way cross have an inbreeding coefficient that is essentially zero. They are good mouse mimics of admixed human populations. We recognize that for some applications, even 256 RI strains will be statistically limiting. In some situations, the RIX crosses can overcome this limitation.

Organization of the Consortium (preliminary)

Production. As noted previously, the production of the RI strains will be confined to four to six large centers each of which will be capable of devoting a minimum of 2000 cages to this project for a period of no less than five years. All databases will be maintained centrally, with local mirrors. Once the first phase is complete (the RI strains are formed), demand is expected to justify the need for four distribution centers, at least one of which should be overseas. A single center will be needed to handle the cryopreservation of the strains – this need not be one of the production centers.

Informatics Core. There are several significant informatics components. These components include coordination of breeding, data repositories, the development of data mining tools and outreach to the larger genetics community. All efforts will be implemented (when possible)

using open source software using a portable programming style. The key features of the informatics core include:

A. Breeding schema testing and colony operations database operations (now in progress).

B. High-throughput genotyping cores and informatics resources to ensure optimal breeding.

C. Final genotyping informatics of RI output.

D. High-level analysis of genome-type structure of the entire RI set to select optimal set of strains for mapping and other applications.

E. Phenotype database development and web interface. Transcriptome-QTL database development. Proteomics and metabolomic informatics.

F. Order, tracking, and distribution database system.

G. Automatic and guided on-line QTL analysis using hybrids of WebQTL and R/qtl. Focus of this system is on end-user tools and interface for multiple-trait mapping, permutation analysis, and query and display of a wide variety of phenotypes and their QTLs.

Quantitative Genetics Analysis Core. This core will have the responsibility for statistical genomic tool development. It will focus on developing new computational tools to reach deeper into multidimensional data sets (sequence, genotype, haplotypes, phenotypes, and environments/pathogens).

Administrative Core. The administrative core will be responsible for the day-to-day operation of the consortium and will be responsible for coordinating regular meetings of participating groups, including internal and external scientific advisory panels.

Genotyping Core(s). Although these cores could be separate, it seems likely that their integration will facilitate the development of the required databases. The initial goal of the Genotyping Core is to locate recombination breakpoints in all RI strains with a precision of approximately 1 Mb. This will require 10,000 or more markers, many of which will have to be custom generated to distinguish alleles and haplotypes over short intervals. With full genome sequence data, it is now possible to rapidly and selectively expand the current set of ~6200 MIT dinucleotide markers. SNP databases are now expanding rapidly and we expect to have more than adequate resources to define breakpoints into 1 Mb bins in the final stages of inbreeding. Since each line will incorporate ~100 breakpoints—typically 3 to 10 per chromosome—it should be possible to achieve high resolution maps using only a small fraction of the available markers. It is likely that SNP genotyping protocols will be sufficiently advanced and inexpensive to achieve 100 kb resolution for the majority of breakpoints.

Phenotyping Cores. As the RI strains are developed, it is proposed that a number of specialized phenotyping cores should come online to facilitate the analysis of the large RI panel. The ability to analyze 256 RI strains and perhaps an equal number of RIX crosses, even if only a few animals/group are needed, will require resources (in terms of space and equipment) on a much larger scale then is usually needed. These should include a metabolic core, a behavioral core, a cancer core and a histology core. In addition, there will be a need for at least one core with sufficient space to allow visiting investigators engage in collaborative phenotyping. These cores will not be part of the founding application but it is assumed that the funding institutions will be committed to their development.

References

Belknap J (1998) Effect of within-strain sample size on QTL detection and mapping using recombinant inbred mouse strains. Behav Genet 1998 28:29-38.

Broman KW, Rowe LB, Churchill GA, Paigen K (2002) Crossover interference in the mouse. Genetics 160:1123–1131. Online at www.complextrait.org

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: An extensible QTL mapping environment. Bioinformatics 19: 1-2.

Darvasi A (1998) Experimental strategies for the genetic dissection of complex traits in animal models. Nat Gen 18:19–24.

Gibson G, Mackay TFC (2002) Enabling population and quantitative genomics. Genet Res Camb 80:1–6.

Hartwell L (2004) Robust interactions. Science 303:774-775.

Hitzemann RW, Malmanger B, Cooper S, Coulombe S, Reed C, Demarest K, Koyner J, Cipp L, Flint J, Talbot C, Rademacher B, Buck K, McCaughran Jr. J (2002) Multiple cross mapping (MCM) markedly improves the localization of a QTL for ethanol-induced activation. Genes, Brain and Behav 1: 214–222.

Hunter KW, Williams RW (2002) Complexities of cancer research: mouse genetic models. ILAR J 43:80–88. Online at www.complextrait.org

Jansen RC (2003) Studying complex biological systems using multifactorial perturbation. Nature Reviews Genetics 4: 145-151.

Felsenstein J (1989) PHYLIP—Phylogeny inference package (v. 3.2). Cladistics 5:164–166.

Nadeau JH, Burrage LC, Revisto J, Pao Y-H, Churchill GA, Hoit BD (2003) Pleiotropy, homeostasis, and functional networks based on assays of cardiovascular traits in genetically randomized populations. Genome Reseach 13:2082-2091.

Reifsnyder PC, Churchill GA, Leiter EH (2000) Maternal environment and genotype interact to establish diabesity in mice. Genome Research 10:1568-1578.

Schalkwyk LC, Jung M, Daser M, Weiher M, Walter J, Himmelbauer H and Lehrach H (1999) Panel of microsatellite markers for whole-genome scans and radiation hybrid mapping and a mouse family tree. Genome Res 9:878-887.

Strachan T, Reed AP (1996) Human molecular genetic. Wiley-Liss, New York.

Threadgill DW, Hunter KW, Williams RW (2002) Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort, Mamm Gen 13:175–178. Online at www.complextrait.org

Williams RW, Gu J, Qi S, Lu L (2001) The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. Genome Biol. 2: RESEARCH0046.

Williams SM, Haines JL, Moore JH (2004) The use of animal models in the study of complex disease BioEssays 26:170-179.