Chapter 28

The Collaborative Cross Resource for Systems Genetics Research of Infectious Diseases

Paul L. Maurizio and Martin T. Ferris

Abstract

An increasing body of evidence highlights the role of host genetic variation in driving susceptibility to severe disease following pathogen infection. In order to fully appreciate the importance of host genetics on infection susceptibility and resulting disease, genetically variable experimental model systems should be employed. These systems allow for the identification, characterization, and mechanistic dissection of genetic variants that cause differential disease responses. Herein we discuss application of the Collaborative Cross (CC) panel of recombinant inbred strains to study viral pathogenesis, focusing on practical considerations for experimental design, assessment and analysis of disease responses within the CC, as well as some of the resources developed for the CC. Although the focus of this chapter is on viral pathogenesis, many of the methods presented within are applicable to studies of other pathogens, as well as to case–control designs in genetically diverse populations.

Key words Use case, Infectious diseases, Collaborative Cross, Influenza

1 Introduction

A confluence of factors interacts to result in adverse infectious disease outcomes, including demographic, environmental, and genetic contributions from the host and pathogen. Given the many challenges of studying viral infections during primary human outbreaks, small animal model systems have been and continue to be essential for the assessment of host genes that drive differences in infection susceptibility and outcomes [1–5]. Differential immune regulation before and after infection is often modulated by complex genetic effects, such as gene-by-gene/gene-by-environment interactions, and allelic variation at individual genes (e.g., hypomorphs, deletions), as an increasing number of studies have begun to illustrate [6–9]. These complex effects are best uncovered and studied in the context of genetically diverse and multi-allelic systems. Therefore, in order to dissect the role of genetic variation on

Klaus Schughart and Robert W. Williams (eds.), *Systems Genetics: Methods and Protocols*, Methods in Molecular Biology, vol. 1488,
DOI 10.1007/978-1-4939-6427-7_28,

host interactions with viruses and other pathogens, it is critical that novel frameworks are developed for the analysis of these complex traits within these genetically diverse systems.

Genetic reference populations (GRPs) have long proven to be powerful for studying complex traits and their underlying causal genetic variants. Many of the classical GRPs (e.g., the BxH [10–12], AxB [13] and BxD panels [14]), as well as classical backcrosses and intercrosses, have been critical in identifying polymorphic host genome regions that influence disease susceptibility and pathology. Building upon the utility of the classical systems, the Collaborative Cross (CC) GRP and the Diversity Outbred (DO) heterogeneous stock were created. These populations advanced the progress of complex trait studies in mice, while also modeling the genetic and allelic complexity present in naturally occurring populations [15–17]. Briefly, both the CC and the DO were derived from a common set of eight founder strains, which are comprised of the three major *Mus musculus* subspecies: *musculus*, *domesticus*, and *castaneus*. As a result of their breeding designs, both populations have high levels of genetic diversity (~45 million SNPs, and ~4 million indels) spread roughly evenly across the genome. Furthermore, in these GRPs, up to eight unique alleles may exist at any gene/locus, and novel epistatic (gene-by-gene) interactions have been introduced that are not present in any of the classical inbred laboratory mouse strains.

Concurrently with advances in the development and genetic characterization of GRPs, a variety of statistical and computational advances have been made. These have improved our ability to identify and characterize unique genetic variants driving differential traits. Improved power and precision for detecting QTL and causative underlying haplotypes, as evinced in refs. [18], have resulted specifically from: our enhanced ability to identify founder strain haplotypes [19]; the publication of annotated whole-genome sequences for the eight founder strains [20, 21]; and the development of powerful software packages for genetic mapping [22–24]. These advances have also enabled the narrowing of QTL regions down directly to candidate causative polymorphisms. Concurrently, RNA-seq and a variety of computational pipelines [25–27] allow for precise and accurate quantification of transcripts, allele-specific expression, and isoform expression within genetically heterogeneous populations. Together, these methods provide powerful new tools in the systems genetics arsenal.

To understand the contribution of host genetic effects on differential infectious disease responses, the CC recombinant inbred (CC-RI) lines and a variety of related populations, including the eight CC founder strains, the partially inbred incipient CC (preCC), the DO, and CC-F1 (recombinant inbred intercrosses, or CC-RIX), have been used in a number of recent studies. In the following sections, we summarize results from across these studies,

and use them to provide a framework for researchers interested in using multiparent populations (MPPs) to study host responses to infection. Although we largely focus on viral pathogens, this guidance is equally useful for other pathogen systems, as well as for systems genetics studies using a case–control design.

Resources describing other uses of the CC and related populations are available in the accompanying chapters of this book, and are also referenced in the following reviews: ref. [28], which covers informatics resources for the Collaborative Cross; ref. [29], which discusses behavioral studies in complex genetic populations; ref. [30], which specifically deals with systems genetics of coronaviruses; and ref. [31], which reviews systems genetics and the utility of network modeling for inference. In addition, there have been several studies examining baseline immune status, autoimmunity, allergy, and inflammation in the CC [32–35]. While highlighting and expanding on the approaches described below, expansion to include specific autoimmune and allergic responses are beyond the scope of this chapter.

2 Methods

In this section, we provide some suggestions for experimental design of infectious disease studies in the CC. A general and useful basic guideline, as adapted from a chapter by [36] is as follows: (1) formulate statistical and biological hypotheses; (2) determine treatment variables, phenotypes of interest, and nuisance variables; (3) determine the population, selecting and/or excluding mouse lines, and simulate and estimate the number of mice that will be required; (4) decide on a randomization protocol; and (5) decide on tools for computational and statistical analysis; and we expand on these points below. We note that there are a large number of different approaches and goals for studying pathogens within the CC. These include, but are not limited to: identifying novel models of pathogenesis [37]; determining the effects, across genetic backgrounds, of variants at previously characterized genes of major effect (e.g., *Mx1* [38] or *Oas1b* [39]); and mapping genetic variants driving differential disease responses [38, 40]. We focus the methodology within in this section as if a researcher were interested in genetic mapping. The general principles and basic protocol are enumerated below:

1. Determine the range of phenotypes to be collected within the study. While many phenotypes are classically considered to be linked during viral infection in traditional inbred lines, it is likely that: (a) these phenotypes will become unlinked due to segregating variants within the CC; and (b) phenotypes causing severe pathology may be differentiated from those simply correlated with disease. Thus, collecting a variety of related phenotypes will

allow for better inferences about the pathways involved in disease pathogenesis. Additionally, in order to avoid confounding effects, potentially important baseline measures, prior to infection, should be considered and recorded. Genetically diverse mice also have phenotypically diverse baseline measures, such as body mass, coat color, litter size, susceptibility to spontaneous disease during aging, etc. Some of these measures may be important for causal inference of the effect of infection, or for clarifying misallocated sample identities, when the data is analyzed.

2. Consider the impact of genes of major effect in order to determine experimental design and/or select a subset of lines. For many pathogens, host genes or loci, e.g., MHC [41, 42], that exert major effects on control of viral disease have already been identified, e.g., *Cmv1* for cytomegalovirus, *Oas1b* for flaviviruses, *Mx1* for influenza, and *CCR5* for HIV. Furthermore, for *Oas1b* [39] and *Mx1* [38], there are both functional and nonfunctional variants segregating within the CC. Using the genetic sequence information available for CC-RIs, experimenters may wish to exclude specific lines from their experimental population. For example, a researcher interested in identifying genetic variants that enhance lung damage during influenza A virus infection might wish to exclude lines with a functional *Mx1* from their study.

   An analogous approach deals with those cases where reagents required to properly assess disease responses are genotype-sensitive or genotype-specific. One example includes a specific viral peptide or tetramer with a major histocompatibility complex (MHC) haplotype restriction. In this case, although CC lines with (e.g.) a C57BL/6J MHC haplotype should generate robust disease response data, CC lines with other founder haplotypes at the MHC locus might not be compatible with the reagent, and therefore accurate assessments of the antiviral states of these lines will not be possible. Thus, exclusion of specific lines, stratified analysis of all lines, or alternative experimental designs may be needed to address these issues. In both of these cases, the best approach to identify specific lines is to examine the founder-strain haplotypes at the genes/loci of interest. The CC status website (http://csbio.unc.edu/CCstatus/index.py) contains a variety of tools, reviewed in ref. [28], with which researchers can identify and visualize the haplotypes present in all available CC lines at given loci. In this way, specific lines can be identified, and subsequently included or excluded, based on the investigator's desired and required haplotypes. We note that while the DO might provide a greater number of genetically unique individuals for a study, the outbred nature of the DO and not being able to preselect animals with given haplotypes from the DO before purchase might strongly affect the ability to assess phenotypic variation if these haplotype-specific reagents and/or genes of major effect are present.

3. Assess a range of phenotypes in a preliminary subset of lines. Host responses to viral infections can differ in a variety of ways, including disease magnitude, kinetics, duration and infection dose responses. Depending on the question of interest, any number of study designs may be optimal for analysis. However, in all cases it is useful to understand the potential range of phenotypic variation being driven by host genetic variants in the CC. This can be achieved by assessing a preliminary subset of mouse lines. A common and useful approach is to screen the eight founder strains of the CC and DO, using a standard, well-characterized dose of virus and relatively long experimental timecourse. In this way, estimates of the range of variation in kinetics, magnitude, onset, and duration of disease can be obtained. Importantly, it is likely, due to transgressive segregation and allele shuffling, that some CC lines will express more extreme viral resistance or susceptibility phenotypes than the eight founder strains. Within these eight strains, one can collect data on the full range of viral pathogenesis phenotypes of interest (e.g., clinical disease, viral replication/dissemination, and tissue damage), following step (1), and determine the phenotypes which vary the most due to host genetic differences.

   In some cases, assessment of the founder strains will be insufficient for estimating phenotypic ranges within the CC. As mentioned above, prior knowledge may dictate that a specific founder strain haplotype should be included or excluded to accommodate experimental needs. In these cases, assuming that the founder haplotype distributions within the CC allow it, an initial screen may be performed using a subset of CC lines rather than the eight founder lines. To illustrate, if only one of the eight founder haplotypes is informative (e.g., seven of the founders are highly resistant to infection due to their allele at a major effect locus), screening several CC strains that contain the one haplotype may be preferable. In contrast, if seven or eight founder haplotypes are informative, screening the seven or eight founders of interest may be preferable to using a subset of CC lines.

4. Determine the batching/blocking and covariates to be used in the study. After an initial screen using the subset of lines in **step 3**, it will be useful to revisit and modify, as necessary, the experimental design for the larger CC study, including experimental block designs and specific covariate data collection, and design of the linear model to be used in the analysis. Again, such decisions are likely to be driven by the investigator's questions of interest, the infectious disease system, and experimental approaches that will be used. However, a few general rules may be helpful.

   One type of idealized experimental design might include an assessment of every treatment group, timepoint, and sex across multiple replicate animals in a single infection batch, with several full batches studied to confirm and generalize these results.

However, we note that even for those examining a single timepoint post-infection, a screen of replicate animals from the entire library of available CC lines might be logistically difficult. In such cases, some form of well-reasoned batching (or "blocking") is required to improve experimental feasibly, while still maintaining an ability to assess statistical significance. The investigator may also want to ensure that the characteristics of the various blocks are well-balanced with respect to the sample size and factors of interest.

There are a large number of ways to design blocking, and we suggest a few simple guidelines. First of all, attempt to randomize, where possible, such that if there is a choice to be made, mice of a given line and sex should be randomly selected from among those available. In order to simplify the screen, it may be preferable to assess and perform QTL mapping in a single sex, with follow-up studies of single lines or timepoints expanded into both sexes to broaden conclusions and to examine sex-specific differences. The inclusion of specific timepoints or subsets of lines will likely depend on the resources available and the phenotypes of interest, such as discovery of new models of previously restricted pathogens, genetic mapping of host variants affecting specific pathologic outcomes, or analysis of differentially expressed transcriptional pathways. For example, if genetic mapping at a single specific timepoint is critical, then ensuring that some lines are repeated across multiple batches, and that each batch contains lines that are repeated in other batches, can be useful for normalizing data across batches. In contrast, if examining the kinetics of differential transcriptional networks is the goal, batches should include all animals of each line in the experiment, with a subset of the total lines to be used. Most importantly, when mock samples are to be paired with samples from a specific timepoint post-infection (DPI, e.g., to study transcriptional differences at 2 DPI or to contrast immune cell infiltration into specific tissues), the mock animals and infected animals from each line should be assayed on the same day to explicitly control for any batch effects. To generalize, for a given contrast or factor of interest (sex, treatment/condition, dose, etc.), including all the levels of interest within each given batch (or even each cage), is preferable when feasible, so that the effect of confounding variables is reduced.

5. Collect phenotype data. Once an appropriate experimental blocking strategy is determined, the study should proceed following the investigator's appropriate infection protocols and design. We note that it is critical to carefully observe and record potentially important, yet previously undescribed disease responses. Such phenotypes might be useful for characterizing novel disease phenotypes in follow-up studies and/or for

improving disease classifications for transcriptional analysis. Be aware of and carefully annotate aberrant or unexpected phenotypes that might be useful as covariates in further analyses (e.g., tumors within tissues of importance that could impact immune phenotypes in those tissues).

6a. Examine the distribution of and correlation between phenotypes. Following data collection, quantify and visualize the within-strain means and variances, as well as the aggregate mean and variance for each phenotype. The use of a Box–Cox transformation on the raw pathogenesis phenotype data will ensure that the residuals follow a more normal phenotypic distribution, enabling a more robust array of statistical analyses. Once data are appropriately transformed, one may determine the genetic contribution to the variance in the data, otherwise known as the broad-sense heritability (H^2), and related measures.

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$$

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}$$

$$h^2 = \frac{\sigma_A^2}{\sigma_G^2 + \sigma_E^2}$$

$$\mathcal{g}^2 = \frac{MS_B - MS_W}{MS_B + (2n-1)MS_W}$$

where σ_P^2 is the total phenotypic variance, σ_G^2 is the total variance attributable to genetics, and σ_E^2 is the remaining variance, attributable to environment and residual noise. σ_G^2 can be partitioned into additive (σ_A^2), dominance (σ_D^2), and epistatic (σ_I^2) components. Broad-sense heritability (H^2) is calculated as the ratio of the genetic variance (σ_G^2), to the total phenotypic variance. Narrow-sense heritability (h^2), which is a subset of H^2, is calculated as the ratio of the additive (σ_A^2) to the total phenotypic variance. The "coefficient of genetic determination" ($\mathcal{g}^2$), which is used for estimating broad-sense heritability in inbred lines [35, 43], is a function of the number of animals tested per strain (n), and the between- and within-strain mean-squared errors (MS_B, MS_W).

Furthermore, a reexamination of the correlation structure of the disease phenotypes (both stratified by strain, and in aggregate) can help to clarify relationships between different

aspects of viral pathogenesis, and can strengthen decision making regarding the phenotypes to be used for mapping causal loci. Simple packages such as corrplot (https://cran.r-project.org/web/packages/corrplot/index.html) and corrgram (https://cran.r-project.org/web/packages/corrgram/index.html) in R can be used to visualize the correlation and covariance structure of a phenotype matrix.

6b. Identify/select samples for transcriptional analysis. In some cases, researchers may wish to add whole-genome transcriptional analysis to further clarify the genes and pathways that are differentially expressed in concordance with specific phenotypes. In many cases, it will be cost-prohibitive to run transcriptional analyses on all samples. Transcriptional analyses that are focused on extreme phenotypic outcomes (e.g., contrast individuals with high vs. undetectable titers), such as is used in bulk segregant analysis, may provide increased power to identify transcripts associated with differential disease. This approach has been illustrated in ref. [44], where a combination of titer and weight loss extremes was used to identify reactive transcriptional networks differentiating the extreme phenotypic groups. Consider, additionally, whether banking a variety of specific immune-related tissues (bone marrow, lymph nodes, CNS, spleen), as well as "unrelated" control tissues may be helpful in follow-up studies, following transcriptional analysis or mapping. Also consider exploring other CC-related in vitro resources (cell-culture, such as mouse embryonic fibroblasts from CC-related mice) that would be useful for your study, especially in the follow-up stages.

7. Conduct genetic mapping. Once phenotypes with high heritability and sufficiently large variation have been identified, genetic mapping can be carried out. A number of software packages exist for multiparent mapping, including Bagpipe (http://valdar-lab.unc.edu/software.html), HAPPY (http://www.well.ox.ac.uk/happy/) [19], and DOQTL, a package for the R statistical computing environment [24], which also works for mapping in the CC. We currently recommend using the DOQTL package, as it is stably supported on Bioconductor (https://www.bioconductor.org/packages/release/bioc/html/DOQTL.html), and also has features to integrate SNP and gene variant features based on the Sanger Institute's resequencing of the eight founder strains of the CC, as described in refs. [17, 24]. For mapping in the CC or the DO, one uses a file(s) to describe the founder haplotype probabilities in the mapping population. This is used both for fully inbred lines, where probabilities should theoretically be 1.0 or 0.0 for any given founder haplotype at each locus, as well as for heterozygous populations, with fractional probabilities and haplotype uncertainties due to

recombination. A separate file is used, containing the transformed phenotypes and any covariates one might consider important (e.g., batch, sex, starting weight). The software uses a linear regression approach, asking whether there is a significant association between phenotypes and haplotype probabilities at each haplotype block along the genome. Significance thresholds are determined using permutation or false discovery rate approaches, both of which take into account the distribution of genotypes and phenotypes within the test population.

Covariates that can be included in the model may take a variety of forms, including demographic (age, sex), experimental (batches), and genetic (e.g., genes of major effect, previously discovered QTL). Proper accounting for these covariates often increases one's power to detect causative genetic polymorphisms underlying virus-induced disease. However, care must be taken in the analysis procedure to ensure that inclusion of covariates does not mask true genetic effects. For example, consider the scenario where two batches of CC lines are tested. In batch 1, all of the lines with a polymorphism enhancing viral titer are tested. In batch 2, all of the lines with a polymorphism repressing viral titer are tested. In this case, including a batch covariate will cause much of the polymorphism's effects to be attributed, incorrectly, to differences between experimental batches. We suggest that QTL scans are run both with and without inclusion of such covariates, with the investigator carefully examining the mapping results (maximum significance scores, significance thresholds, etc.) from all situations. Furthermore, if there are QTL that appear when scans are run without a covariate, but are absent when a covariate is included, we suggest follow-up studies that replicate some of the lines and timepoints that were previously split by covariates. Where possible, such studies will confirm and validate such QTL without the confounding of covariates. Additional, more rigorous model selection protocols for determining the inclusion/exclusion of covariates may be considered, but they are beyond the scope of this chapter.

8. Identify causative polymorphisms. Once QTL are identified, a number of approaches may be utilized to determine specific causative genetic variants. Toward this end, most multiparent mapping tools will generate allele effect plots. These plots display the estimated scaled effects of each founder allele on a trait of interest within a given genomic locus. In this way, one can distinguish groups of haplotypes that enhance, suppress, or have no effect on disease. By identifying SNPs and other genetic variants within the QTL that follow the allele effects patterns, one can narrowly focus on subsets of candidate genes or features that are likely to be causative for the phenotype of interest. For example,

at a given SNP, if both "high" and "low" phenotype groups share a founder allele, it is unlikely to be causative, whereas SNPs that contain alleles that segregate between the high- and low-responder strains are much more likely to be causative. Further integration of already available whole-genome expression data, or post-hoc gene expression analysis (e.g., qPCR) can help to narrow and refine candidates. For example, genes underneath a QTL locus which are differentially expressed between high and low groups, in a relevant tissue or compartment and at a relevant timepoint, will help lead to potential candidate genes and/or pathways impacting disease outcomes.

9. Consider alternate studies and experimental approaches. In the preceding text, we highlighted a case where follow-up studies might be useful in identifying genetic variants, i.e., where initial QTL scans suggested a locus, but the effect of that locus was confounded by a covariate such as experimental batch. Following the collection of initial data, there are a variety of other experiments which can help clarify and enhance the initial studies. One possibility is that only one or two CC lines show a desired or extreme disease response [37, 45]. Such outcomes may indicate either complex gene–gene interactions (epistasis) or de novo mutations arising in lines. In both cases, one should consider either a tailored follow-up genetic cross, such as an F2, as in refs. [45], or follow-up intensive expression analysis, as in refs. [37], to focus on likely causative loci or networks driving these unique disease responses. Another possible outcome is that a gene of major effect has been discovered. This would be a case where a QTL explains a large fraction (e.g., 50%) of trait variation for one or more of the pathology traits of interest. In these cases, it may be useful to either (as recommended above) redo analysis with the QTL of major effect as a covariate in the main QTL scans OR to subset your set of lines into those with the high versus low haplotypes at the QTL. These subpopulation style analyses can help in identifying further genetic variants that affect disease only in the context of a gene of major effect. For example, if a variant impacts viral dissemination from a primary tissue, its effect can be masked if there is an additional polymorphism that abrogates the viral receptor within the CC. Only by mapping with the receptor positive population will it be possible to identify the dissemination variant.

3 Expected Results

Given that there are a variety of possible genetic architectures underlying host responses to multiple aspects of viral infection, it is difficult to precisely predict outcomes for any given study type. However, based on the breadth of work conducted so far within

the CC, the DO, and related populations, one can expect that there are genetic variants segregating within the CC system that will have impact on pathogenesis for any given virus. Indeed, for at least four viral pathogens, as well as for a variety of other bacterial and fungal pathogens, QTL have been identified that contribute to differential disease responses and pathogenesis. Taken as a whole, these QTL typically have shown modest effects on pathogenic traits (e.g., a summary of the results of [38, 40] show most QTL explaining 25 % of phenotypic variance for each trait). Furthermore, it is likely that transgressive segregation operates within the recombined genomes of the CC. That is, alleles driving extreme responses may come from a founder strain(s) that exhibits a mild or suppressed phenotype. Thus, only when the genetic structure of the founder strains has been rearranged will the true effects of alleles be identifiable. Lastly, it is likely that once QTL are identified, it will be possible to identify a set of high priority SNPs, based on the founder strain sequences, which act as the causative variants. Additional pathological and molecular phenotyping will be required for validation, but the integration of multiple allele effects, as well as sequence data, is a substantial improvement over classical positional cloning for identifying causal variants.

4 Lessons Learned

The use of the Collaborative Cross and related populations in studying infectious diseases is still in its nascent stages. Nevertheless, there are several important considerations, gleaned from the studies to date, that can specifically inform future studies and analysis of determinants of infectious disease susceptibility.

One clear lesson learned from virus infection studies in the CC and preCC so far is that phenotypic correlations present within any set of characterized founder strains or knockouts are likely to be broken apart within the CC, unless there are strong causal relationships between the correlated phenotypes. For example, a complete disassociation was seen between different aspects of SARS-coronavirus (SARS-CoV) induced pathology and disease within the preCC population [40]. Furthermore, QTL mapping will often show that distinct loci affect individual, distinct pathologic traits, as seen in both the influenza preCC and SARS-CoV preCC studies [38, 40]. These results highlight one main impetus for utilizing GRPs such as the CC (i.e., the discovery of novel phenotypic relationships and distinct genetic markers), but they also point to a critical consideration in the design and analysis of studies in these systems. Namely, the assessment of a wide variety of phenotypes, even those classically thought to be redundant, will be highly useful and enable a better understanding of disease pathogenesis.

It is well known that susceptibility and resistance genes of major effect are predominant within host–pathogen systems. These genes of major effect include, for example *Cmv1* for murine cytomegalovirus [46, 47]; *Oas1b* for flaviviruses [48]; and *Mx1* for influenza [49, 50]. Indeed, both functional and defective *Oas1b* and *Mx1* alleles circulate within the CC/DO [38, 39]. Given the genetic diversity present within the CC and DO, it is likely that other genes of major effect for specific pathogens will be found segregating within these populations. Although the presence of genes and alleles of major effect may appear to be an obstacle for discovery of novel regulators of disease, obscuring the contribution of genes or alleles of lesser effect size, new biological insight can still be obtained in the presence of these large effect alleles. For example, given the potential for up to eight alleles segregating within the CC at any locus, there may be several alleles, isoforms and/or transcriptional variants at a given causal locus. This was observed clearly in the influenza challenge of the preCC, where the antiviral and clinically protective effects of *Mx1* were disassociated via the presence of three unique alleles at the *Mx1* locus [38].

Furthermore, epistasis and transgressive segregation are at work within the CC population. Such segregation can most commonly reveal previously "hidden" genetic variation. For example, in the preCC study of SARS, the wild-derived founder strains (CAST, PWK, and WSB) all die of SARS-CoV infection at low doses [40]. In contrast, for the preCC lines that survived SARS-CoV infection, causative alleles at a variety of QTL are driven by wild derived parental alleles, and were therefore hidden in the context of super-susceptible parent founders. Thus, the allelic variants that affect immunopathology, viral replication, and immune infiltration were identified only in the context of disruption of founder haplotypes through recombination. Alternately, recombination driving reassortment of alleles may cause emergent phenotypes by introducing evolutionarily distinct allelic combinations. For example, it is only via this genetic reassortment across the CC that a severe Ebola virus (EBOV)-induced hemorrhagic fever was identified in mice, as this phenotype was not present in any of the CC founder strains [37].

There are several reasons for emphasizing the thoughtful use of mock controls in infectious disease studies in the CC. Firstly, given the novelty of the genetic backgrounds generated in the CC, the response to mock treatment in some lines may differ substantially from that of common inbred lines. Additionally, genetic loci that regulate baseline immune phenotypes may be quite distinct from those that regulate immune phenotypes after infection-induced pathways are upregulated or downregulated, hence QTL may be mapped for untreated or mock-treated animals as a complement to QTL mapped for infection response.

Finally, it should be noted that characterizing the variability or variance in a disease phenotype, both within-strain and between-strain, is worthwhile and may be critical for identifying genetic causes of differential disease. Identifying a strain or set of strains with increased variance may lead you to identifying a novel genetic factor or latent environmental variable that causes a substantial change in the phenotype of interest [51]. During the characterization of within-strain variation, you may be able to identify experimental issues that ought to be modeled or corrected (e.g., batch effects), or rare de novo genetic variants that substantially modulate your phenotype, which can be identified with additional genotyping or sequencing. In one recent study, using a diallel of the wild-derived CC founder mice and their F1 reciprocal crosses, gene expression was substantially altered in two mice, including one which was found to have a de novo duplication [25]. Thus, having well-characterized within- and between-strain variance estimates are critical for identifying novel genetic variants, for estimating statistical power, and for successful experimental design and analysis in the CC.

5 Further Considerations and Limitations

Although systems genetics approaches and genetically diverse study populations provide a powerful combination of tools to identify host genetic variants driving infectious disease, there are several caveats that ought to be considered in optimizing study design and analysis approaches: namely appropriate molecular phenotyping, disentangling complex phenotypic networks, and mechanistic insight into variant loci.

Omics analysis (transcriptomics, proteomics, metabolomics, etc.) is a cornerstone of the systems biology approach to research. One strong caveat for omics analysis is the dependence of these approaches on accurate assessment of genome sequences for utilized strains. The C57BL/6J genome has formed the backbone of mouse sequence analysis and annotation, however we know that the other CC founder strains, and therefore the CC themselves contain large numbers of polymorphisms, and more importantly structural variants and large insertion/deletions [52] (http://www.sanger.ac.uk/science/data/mouse-genomes-project). As described in more detail in this volume in a chapter by Green et al., integration of imputed genome sequences (pseudogenomes) for CC lines or DO animals [27] will substantially improve integration of these omics data within these GRPs.

A variety of factors, such as prior immune history, opportunistic coinfection, and microbiome influences on the immune system [53, 54] can influence host responses to viral infections. Furthermore, there is evidence for genetic variants within the CC affecting basal variation in immune populations [34]. Given the

potential for host genetic variation to impact a variety of immune phenotype and the microbiome, it is likely that there will be complex causal networks underlying variation in the direct viral pathogen traits of interest of a researcher. While dissection of these networks may provide many years of fruitful study, they may present daunting obstacles to study within the CC. Careful design of experiments (e.g., cohousing animals from different CC lines; antibiotic pretreatment to limit bacterial coinfection) can help to ameliorate and control some of these effects and improve the ability to identify genetic variants directly affecting host responses to viral pathogens of interest.

Finally, we note that identification of genetic variants with a GRP affecting host responses to viral infection do little to identify the mechanisms and processes through which these variants act. While integration of a variety of phenotypes (e.g., pathological, immunological, and molecular responses) can help to highlight mechanisms and pathways of activity, it is only through classical (and phenotype-specific) manipulation and experimentation that a true understanding of these variants can be elucidated. Such approaches, often deemed "reductionist," are critically useful in transitioning broad systems-based responses with clear and actionable mechanistic processes.

6 Outlook

Small animal models for the host response to infectious disease pathogens are critical tools for the study of human susceptibility to disease, as well as for the development of novel prophylactics and therapeutics. Indeed, the utility of these systems for studying host–pathogen interactions appears to be persistent and critical. Notably, by varying the host genetic background in the study of infectious disease, we enable the detection of genetic variants that are important for disease across a population of genetically diverse individuals, improving our chances that variants are reproducible across experiments and, it is hoped, across species. Importantly, not only can these systems be used for identifying genetic susceptibility loci, but they can also be used to identify and develop of novel infectious disease models, using specific strains of CC mice as new resources for understanding severe disease, such as has been done in the recent development of CC mouse models of Ebola virus pathogenesis [37].

The CC is also useful for better understanding the genetic architecture of the host response to infection. It has been recognized that nonadditive genetic effects, such as dominance, epistasis, and parent-of-origin effects, may contribute substantially to quantitative traits, including the host immune system and infectious disease responses. In order to estimate, quantify, and explore

such complex genetic interactions, and to quantify broad and narrow-sense heritability, future directions include characterizing infection phenotypes in F1 reciprocal crosses of the eight founder lines and of the CC lines (using CC-F1s). Such experiments will add to our knowledge about how disease susceptibility and resistance may be expressed and transmitted from parents to offspring, and this work may reveal important genetic complexities, hard to uncover in human studies. These complex genetic effects may be responsible for inhibiting our ability, at present, to identify candidate genes through GWAS and linkage mapping studies, which less often include rigorous screens for nonadditive effects. Finally, the experimental designs and phenotypic data sets that are being generated for systems genetics in the Collaborative Cross lend themselves to innovative statistical and quantitative genetics models. These new models and quantitative tools advance our understanding of human disease, and complement the variety of experimental tools being developed for the CC. Thus, infectious disease research in CC promises to advance our knowledge about complex host–pathogen interactions, and to enhance our ability to unravel and interpret increasingly complex biological networks in order to improve human health.

Acknowledgments

We acknowledge U19AI100625 to M.T.F. and 5T32AI007419-23 to P.L.M. for support.

References

1. Srivastava B, Błazejewska P, Hessmann M, Bruder D, Geffers R, Mauel S, Gruber AD, Schughart K (2009) Host genetic background strongly influences the response to influenza A virus infections. PLoS One 4(3):4857
2. Boon AC, deBeauchamp J, Hollmann A, Luke J, Kotb M, Rowe S, Finkelstein D, Neale G, Lu L, Williams RW, Webby RJ (2009) Host genetic variation affects resistance to infection with a highly pathogenic H5N1 influenza A virus in mice. J Virol 83(20):10417–10426
3. Bouvier NM, Lowen AC (2010) Animal models for influenza virus pathogenesis and transmission. Viruses 2:1530–1563
4. Boivin GA, Pothlichet J, Skamene E, Brown EG, Loredo-Osti JC, Sladek R, Vidal SM (2012) Mapping of clinical and expression quantitative trait loci in a sex-dependent effect of host susceptibility to mouse-adapted influenza H3N2/HK/1/68. J Immunol 188(8):3949–3960
5. Boon AC, Finkelstein D, Zheng M, Liao G, Allard J, Klumpp K, Webster R, Peltz G, Webby RJ (2011) H5N1 influenza virus pathogenesis in genetically diverse mice is mediated at the level of viral load. mBio 2(5):pii:e00171-11
6. Wei W-H, Hemani G, Haley CS (2014) Detecting epistasis in human complex traits. Nat Rev Genet 15(11):722–733
7. Lenz TL, Deutsch AJ, Han B, Hu X, Okada Y, Eyre S, Knapp M, Zhernakova A, Huizinga TW, Abecasis G, Becker J, Boeckxstaens GE, Chen WM, Franke A, Gladman DD, Gockel I, Gutierrez-Achury J, Martin J, Nair RP, Nöthen MM, Onengut-Gumuscu S, Rahman P, Rantapää-Dahlqvist S, Stuart PE, Tsoi LC, van Heel DA, Worthington J, Wouters MM, Klareskog L, Elder JT, Gregersen PK, Schumacher J, Rich SS, Wijmenga C, Sunyaev SR, de Bakker PI, Raychaudhuri S (2015) Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. Nat Genet 47(9):4–7
8. Galson JD, Trück J, Fowler A, Clutterbuck EA, Münz M, Cerundolo V, Reinhard C, van

der Most R, Pollard AJ, Lunter G, Kelly DF (2015) B-cell repertoire responses to varicella-zoster vaccination in human identical twins. Proc Natl Acad Sci 112(2):500–505

9. Shin D-L, Hatesuer B, Bergmann S, Nedelko T, Schughart K (2015) Protection from severe influenza infections in mice carrying the *Mx1* influenza resistance gene strongly depends on genetic background. J Virol 89(19):01305–01315
10. Turcotte K, Gauthier S, Mitsos L-M, Shustik C, Copeland NG, Jenkins NA, Fournet J-C, Jolicoeur P, Gros P (2004) Genetic control of myeloproliferation in BXH-2 mice. Blood 103(6):2343–2350
11. Marquis J-F, LaCourse R, Ryan L, North RJ, Gros P (2009) Disseminated and rapidly fatal tuberculosis in mice bearing a defective allele at IFN regulatory factor 8. J Immunol 182(5):3008–3015
12. Berghout J, Langlais D, Radovanovic I, Tam M, MacMicking JD, Stevenson MM, Gros P (2013) Irf8-regulated genomic responses drive pathological inflammation during cerebral malaria. PLoS Pathog 9(7):e1003491
13. Hassan MA, Jensen KD, Butty V, Hu K, Boedec E, Prins P, Saeij JP (2015) Transcriptional and linkage analyses identify loci that mediate the differential macrophage response to inflammatory stimuli and infection. PLoS Genet 11(10):1005619
14. Nedelko T, Kollmus H, Klawonn F, Spijker S, Lu L, Heßman M, Alberts R, Williams RW, Schughart K (2012) Distinct gene loci control the host response to influenza H1N1 virus infection in a time-dependent manner. BMC Genomics 13(1):411
15. Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J (2004) The collaborative cross, a community resource for the genetic analysis of complex traits. Nat Genet 36(11):1133–1137
16. Churchill GA, Gatti DM, Munger SC, Svenson KL (2012) The diversity outbred mouse population. Mamm Genome 23(9-10):713–718
17. Svenson KL, Gatti DM, Valdar W, Welsh CE, Cheng R, Chesler EJ, Palmer AA, McMillan L, Churchill GA (2012) High-resolution genetic mapping using the mouse diversity outbred population. Genetics 190(2):437–447
18. Aylor DL, Valdar W, Foulds-Mathes W, Buus RJ, Verdugo RA et al (2011) Genetic analysis of complex traits in the emerging collaborative cross. Genome Res 21(8):1213–1222
19. Mott R, Talbot CJ, Turri MG, Collins AC, Flint J (2000) A method for fine mapping quantitative trait loci in outbred animal stocks. Proc Natl Acad Sci 97(23):12649–12654
20. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, Furlotte NA, Eskin E, Nellåker C, Whitley H, Cleak J, Janowitz D, Hernandez-Pliego P, Edwards A, Belgard TG, Oliver PL, McIntyre RE, Bhomra A, Nicod J, Gan X, Yuan W, van der Weyden L, Steward CA, Bala S, Stalker J, Mott R, Durbin R, Jackson IJ, Czechanski A, Guerra-Assunção AJ, Donahue LR, Reinholdt LG, Payseur BA, Ponting CP, Birney E, Flint J, Adams DJ (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477(7364):289–294
21. Adams DJ, Doran AG, Lilue J, Keane TM (2015) The mouse genomes project: a repository of inbred laboratory mouse strain genomes. Mamm Genome 26(9-10):403–412
22. Arends D, Prins P, Jansen RC, Broman KW (2010) R/qtl: high-throughput multiple QTL mapping. Bioinformatics 26(23):2990–2992
23. Zhang Z, Wang W, Valdar W (2014) Bayesian modeling of haplotype effects in multiparent populations. Genetics 198(1):139–156
24. Gatti SKL, Shabalin A, Wu LY, Valdar W, Simecek P, Goodwin N, Cheng R, Pomp D, Palmer A, Chesler EJ, Broman KW, Churchill GA (2014) Quantitative trait locus mapping methods for diversity outbred mice. G3 (Bethesda, MD) 4(9):1623–1633
25. Crowley JJ, Zhabotynsky V, Sun W, Huang S, Pakatci IK, Kim Y, Wang JR, MorganAP CJD, Aylor DL, Yun Z, Bell TA, Buus RJ, Calaway ME, Didion JP, Gooch TJ, Hansen SD, Robinson NN, Shaw GD, Spence JS, Quackenbush CR, Barrick CJ, Nonneman RJ, Kim K, Xenakis J, Xie Y, Valdar W, Lenarcic AB, Wang W, Welsh CE, Fu CP, Zhang Z, Holt J, Guo Z, Threadgill DW, Tarantino LM, Miller DR, Zou F, McMillan L, Sullivan PF, de Villena FP-M (2015) Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat Genet 47(4):353–360
26. Zou F, Sun W, Crowley JJ, Zhabotynsky V, Sullivan PF, de Villena FP-M (2014) A novel statistical approach for jointly analyzing RNA-seq data from F1 reciprocal crosses and inbred lines. Genetics 197(1):389–399
27. Munger SC, Raghupathy N, Choi K, Simons AK, Gatti DM, Hinerfeld DA, Svenson KL, Keller MP, Attie AD, Hibbs MA, Graber JH, Chesler EJ, Churchill GA (2014) RNA-seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. Genetics 198(1):59–73
28. Morgan AP, Welsh CE (2015) Informatics resources for the collaborative cross and

related mouse populations. Mamm Genome 26(9):521–539

29. Mulligan MK, Williams RW (2015) Systems genetics of behavior: a prelude. Curr Opin Behav Sci 2:108–115
30. Schäfer A, Baric RS, Ferris MT (2014) Systems approaches to coronavirus pathogenesis. Curr Opin Virol 6(1):61–69
31. Civelek M, Lusis AJ (2014) Systems genetics approaches to understand complex traits. Nat Rev Genet 15(1):34–48
32. Kelada SNP, Aylor DL, Peck BCE, Ryan JF, Tavarez U, Buus RJ, Miller DR, Chesler EJ, Threadgill DW, Churchill GA, de Villena FP-M, Collins FS (2012) Genetic analysis of hematological parameters in incipient lines of the collaborative cross. G3 (Bethesda, MD) 2(2):157–165
33. Kelada SNP, Carpenter DE, Aylor DL, Chines P, Rutledge H, Chesler EJ, Churchill GA, de Villena FP-M, Schwartz DA, Collins FS (2014) Integrative genetic analysis of allergic inflammation in the murine lung. Am J Respir Cell Mol Biol 51(3):436–445
34. Phillippi J, Xie Y, Miller DR, Bell TA, Zhang Z, Lenarcic AB, Aylor DL, Krovi SH, Threadgill DW, de Villena FP-M, Wang W, Valdar W, Frelinger JA (2014) Using the emerging collaborative cross to probe the immune system. Genes Immun 15(1):38–46
35. Rutledge H, Aylor DL, Carpenter DE, Peck BC, Chines P, Ostrowski LE, Chesler EJ, Churchill GA, de Villena FP-M, Kelada SNP (2014) Genetic regulation of zfp30, cxcl1, and neutrophilic inflammation in murine lung. Genetics 198(2):735–745
36. Kirk RE (2009) The SAGE handbook of quantitative methods in psychology. Sage, Thousand Oaks, CA
37. Rasmussen AL, Okumura A, Ferris MT, Green R, Feldmann F, Kelly SM, Scott DP, Safronetz D, Haddock E, LaCasse R, Thomas MJ, Sova P, Weiss JM, Carter VS, Miller DR, Shaw GD, Korth MJ, Heise MT, Baric RS, de Villena FP-M, Feldmann H, Katze MG (2014) Host genetic diversity enables Ebola hemorrhagic fever pathogenesis and resistance. Science 346(6212):987–991
38. Ferris MT, Aylor DL, Bottomly D, Whitmore AC, Aicher LD, Bell TA, Bradel-Tretheway B, Bryan JT, Buus RJ, Gralinski E, Haagmans BL, McMillan L, Miller DR, Rosenzweig E, Valdar W, Wang J, Churchill GA, Threadgill DL, McWeeney SK, Katze MG, de Villena FP-M, Baric RS, Heise MT (2013) Modeling host genetic regulation of influenza pathogenesis in the Collaborative Cross. PLoS Pathog 9(2):e1003196
39. Graham JB, Thomas S, Swarts J, McMillan AA, Ferris MT, Suthar MS, Treuting PM, Ireton R, Gale M, Lund M (2015) Genetic diversity in the collaborative cross model recapitulates human West Nile virus disease outcomes. mBio 6(3):1–11
40. Gralinski LE, Ferris MT, Aylor DL, Whitmore AC, Green R, Frieman MB, Deming D, Menachery VD, Miller DR, Buus RJ, Bell TA, Churchill GA, Threadgill DW, Katze MG, McMillan L, Valdar W, Heise MT, de Villena FP-M, Baric RS (2015) Genome-wide identification of SARS-CoV susceptibility loci using the Collaborative Cross. PLoS Genet 11(10):e1005504
41. Blackwell JM, Jamieson SE, Burgner D (2009) HLA and infectious diseases. Clin Microbiol Rev 22(2):370–385
42. Sellers RS, Clifford CB, Treuting PM, Brayton C (2012) Immunological variation between inbred laboratory mouse strains: points to consider in phenotyping genetically immunomodified mice. Vet Pathol 49(1):32–43
43. Francis M, Festing W (1979) Notes on genetic analysis (Chapter 7). In: Inbred strains in biomedical research. Macmillan, New York, NY, pp 80–98
44. Bottomly D, Ferris MT, Aicher LD, Rosenzweig E, Whitmore A, Aylor DL, Haagmans BL, Gralinski LE, Bradel-Tretheway BG, Bryan JT, Threadgill DV, de Villena FP-M, Baric RS, Katze MG, Heise M, McWeeney SK (2012) Expression quantitative trait loci for extreme host response to influenza A in pre-Collaborative Cross mice. G3 (Bethesda, MD 2(2):213–221
45. Rogala AR, Morgan AP, Christensen AM, Gooch TJ, Bell TA, Miller DR, Godfrey VL, Villena FP-M (2014) The Collaborative Cross as a resource for modeling human disease: CC011/Unc, a new mouse model for spontaneous colitis. Mamm Genome 25(3-4):95–108
46. Scalzo AA, Fitzgerald NA, Simmons A, La Vista AL, Shellam GR (1990) Cmv-1, a genetic locus that controls murine cytomegalovirus replication in the spleen. J Exp Med 171(5):1469–1483
47. Scalzo AA, Yokoyama M (2008) Cmv1 and natural killer cell responses to murine cytomegalovirus infection. Curr Top Microbiol Immunol 321:101–122
48. Scherbik SV, Kluetzman K, Perelygin AA, Brinton MA (2007) Knock-in of the Oas1b(r) allele into a flavivirus-induced disease

susceptible mouse generates the resistant phenotype. Virology 368(2):232–237

49. Arnheiter H, Skuntz S, Noteborn M, Chang S, Meier E (1990) Transgenic mice with intracellular resistance to influenza. Cell 62:51–61
50. Staeheli P, Grob R, Meier E, Sutcliffe JG, Haller O (1988) Influenza virus-susceptible mice carry Mx genes with a large deletion or a nonsense mutation. Mol Cell Biol 8(10):4518–4523
51. Rönnegård L, Valdar W (2011) Detecting major genetic loci controlling phenotypic variability in experimental crosses. Genetics 188(2):435–447
52. Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan X, Nellåker C, Goodstadt L, Nicod J, Bhomra A, Hernandez-Pliego P, Whitley H, Cleak J, Dutton R, Janowitz D, Mott R, Adams DJ, Flint J (2011) Sequence-based characterization of structural variation in the mouse genome. Nature 477(7364):326–329
53. Ichinohe T, Pang IK, Kumamoto Y, Peaper DR, Ho JH, Murray TS, Iwasaki A (2011) Microbiota regulates immune defense against respiratory tract influenza A virus infection. Proc Natl Acad Sci 108(13):5354–5359
54. Zhang D, Chen G, Manwani D, Mortha A, Xu C, Faith JJ, Burk RD, Kunisaki Y, Jang JE, Scheiermann C, Merad M, Frenette PS (2015) Neutrophil ageing is regulated by the microbiome. Nature 525(7570):528–532