Post-genomic behavioral genetics: From revolution to routine

David G. Ashbrook, Megan K. Mulligan, and Robert W. Williams
University of Tennessee Health Sciences Center, College of Medicine, Department of Genetics, Genomics and Informatics, Memphis, TN, USA

Abstract

What was once expensive and revolutionary—full genome sequence—is now affordable and routine. Costs will continue to drop, opening up new frontiers in behavioral genetics. This shift in costs from the genome to the phenome is most notable in large clinical studies of behavior and associated diseases in cohorts that exceed hundreds of thousands of subjects. Examples include the Women’s Health Initiative (www.whi.org), the Million Veterans Program (www.research.va.gov/MVP), the 100,000 Genomes Project (genomicsengland.co.uk), and commercial efforts such as those by deCode (www.decode.com) and 23andme (www.23andme.com). The same transition is happening in experimental neuro- and behavioral genetics, and samples sizes of many hundreds of cases are becoming routine (www.genenetwork.org, www.mousephenotyping.org).

There are two major consequences of this new affordability of massive omics data sets: (1) It is now far more practical to explore genetic modulation of behavioral differences and the key role of gene-by-environment interactions. Researchers are already doing the hard part—the quantitative analysis of behavior. Adding the omics component can provide powerful links to molecules, cells, circuits, and even better treatment. (2) There is an acute need to highlight and train behavioral scientists in how best to exploit new omics approaches. This review addresses this second issue and highlights several new trends and opportunities that will be of interest to experts in animal and human behaviors.

Keywords

reverse genetics; QTL mapping; PheWAS; GWAS; quantitative trait locus; complex traits; systems genetics; omics; behavior; genomics

Introduction

Genetic analysis of behavior has involved two competing and almost antithetical approaches. The older of the two, sometimes called forward or classical genetics, starts with interesting and heritable differences in behavior among individuals and attempts to define gene loci, sequence variants, and systems that contribute to these differences (reviewed in Mulligan and Williams, 2015; Baud and Flint, 2017). Quantitative trait locus (QLT) mapping and genome-wide association studies (GWASs) are common examples of this systematic downward approach—from differences in behavior to a subset of causal loci, haplotypes, alleles, or even specific DNA variants. The other approach, sometimes called reverse genetics (although upward makes more sense), starts with defined sequence differences or mutations (knock-outs, knock-ins, transposon or transgene insertions) and then works from
the polymorphism up through molecular and cellular cascades that influence (but rarely
determine) brain function and behavior (Takahashi et al. 1994; Li et al. 2010; Wang et al.
2016).

The two communities of researchers have worked back-to-back on many of the same
problems for years; often without listening to each other. On the one hand, the forward
genetics community enthusiastically uses ever more sophisticated methods to fine-map
multitudes of variants with small effects on behavior (e.g., Philip et al. 2010; reviewed in
Baud & Flint 2017). On the other hand, the reverse genetics community enthusiastically uses
ever more sophisticated molecular methods to modify DNA sequence and define the impact
on behaviors—almost always on a single genetic background. Both approaches build causal
models, but the forward method has the advantage of direct relevance to natural populations
and common sequence variants that are responsible for most human neurological and
psychiatric disease. The reverse method has the advantage of much more direct linkage to
mechanism.

Over the past decade forward and reverse approaches have begun to merge, and are now
happy partners thanks to high throughput genomic methods, in particular genome and
transcriptome sequencing. One example of the hybridization of these two methods is that the
final phase of QTL mapping now involves an almost obligatory analysis of sequence
variants of large numbers of candidate genes. Polymorphisms with high potential biological
impact (e.g., missense and nonsense mutations) rank well as candidates. In effect, the last
stage of QTL mapping now involves a reverse genetic assessment. There are many other
interesting ways that genomics and transcriptomics are changing the landscape of behavioral
neurogenetics, as well as the analysis of neurological and psychiatric diseases. Here we
provide several examples of the impact that high-throughput omics is having on our field.
We are now able to ask and answer questions using new methods. While we focus on mouse
models, there are almost always direct applications of these same methods to other species,
and more specifically, to human populations.

Results and Discussion

1. QTL mapping of behavior using sequence data

As mentioned above, the analysis of genome sequence data is now almost an obligatory last
step of QTL analyses (e.g., Mozhui et al. 2008; Chintalapudi et al. 2016; Neuner et al. 2016;
Delprato et al. 2017). After a QTL has been mapped with sufficient precision (a topic to
which we return), it is necessary to sort through and evaluate the strength of candidates
within the 1.5–2.0 LOD confidence interval based on numbers and types of sequence
variants in each locus or gene or conserved element in a particular population or cross.
While an interval may be large, encompassing hundreds of genes, if we are highly confident
that there are no sequence variants in or around those genes, we can discount (not eliminate)
their causal role. While those of us who map for a living obsess about the size of our QTL
confidence intervals with the goal of getting down to 1 Mb or less, this is actually an old-
fashioned metric (Williams & Williams 2017). A much more important metric in a post-
sequence world is the number of genes and classes of sequence variants within an interval,
almost irrespective of its absolute length. For example, using reduced complexity crosses, it
is practical to “clone” causal variants with a QTL that is formally well over 10 Mb in size (see section below, Kumar et al. 2013 and Bryant et al. 2017).

The most obvious and interesting variants are nonsense and missense “mutations” that have a serious impact on protein structure and function. When full sequence data are available for parents of a cross, it is now easy to assemble lists of missense variants in each gene and to rank all candidates by their suspected biological impact on protein function. It is also becoming practical to evaluate and rank some types of non-coding variants—those in promoters, UTRs, introns, enhancers, and suspected regulatory elements.

Two very important caveats: (1) Sequence data even for mouse, human, rat, and drosophila is still far from perfect. SNPs, small indels, and copy number variants are typically not a serious problem, although they do require a simple validation step involving re-sequencing using the gold-standard dideoxy chain termination method. But inversions and longer repeat polymorphisms (200 to 2000 bp) are still hard to detect. This means that some intervals that appear devoid of SNPs or indels, may contain cryptic variants with large effects (e.g. Li et al. 2010). (2) The assignment of variants to particular genes is not always straightforward. Genome-wide studies in human have highlighted SNPs with impressive LOD scores that are far from known coding genes or even from putative regulatory elements. A variant with a behavioral impact may not be within a gene that encodes a protein at all, but may be in a non-coding transcript or pre-microRNA sequence that modulates expression of a set of genes and proteins. Polymorphic gene regulatory elements are not always contained within a canonical gene region. These elements can be interleaved across regions of up to 1–2 megabases (Mb). For example, an apparently causal SNP within the human FTO gene (within intron 1) that predicts obesity risk is apparently a variant within an enhancer that affects expression of the neighboring IRX3 transcription factor (Smemo et al. 2014). While we have highlighted these analytic complexities, our own perhaps premature impression is that problems of these types will be comparatively infrequent. Further, as we learn more about the functional impact of control elements in the genome and the complex configuration of chromatin loops, it should become possible to rank even these intergenic and intronic variants with better assurance.

To summarize this section: the last steps in QTL mapping are dependent on sequence data. Once an interval has been defined, the analysis quickly takes on the character of reverse genetics, evaluating potential biological impact of particular variants. However, unlike genetically engineered knockout lines and transgene insertions, there is usually a large set of polymorphisms that are causal contenders. This is why large-scale genetic and omics data sets for multiple species, diseases, organs, and cells (Chesler et al. 2005; Mozhui et al. 2008; Williams et al. 2016) are so valuable in efficient evaluation of the candidacy of genes in any interval.

Finally, there really are no longer any limitations as to the type of cross or population or species that will benefit from sequence data. Any experiment involving an intercross or backcross should add a sequencing prelude, whether the cross is between outbred, inbred, or wild-derived individuals. Obviously, this is not required if the parents have been recently sequenced, but if not, this smart prelude may require some expertise in genome assembly.
and comparison. At a cost of only about $1000 per genome, this component is likely to be a minor expense compared to the cost and time involved in phenotyping. Even if the parental lines have been sequenced previously, genetic drift may have occurred in the intervening time, so resequencing may well be useful to account for this.

2. Phenome-wide association or reverse complex trait analysis

With full sequence data for a cohort or cross it is possible to perform reverse genetic analysis starting with any segregating sequence variant (that is, as described above, moving from a known sequence variant to the phenotypes it influences). The analysis can be simple—just compute the correlation between a SNP marker in or near a sequence variant and one or more phenotypes. The null hypothesis is that correlations of this type should be close to zero if traits are not modulated by the index gene or the region around that gene.

An initial question is what phenotypes are available for testing these gene-to-phenotype (G-to-P) associations? How many cases or strains do we need to have reasonable power to detect a G-to-P link? An important statistical question is how do we control for false discoveries and how do we control for the effects of linkage between neighboring sequence variants in the same or different genes? All of these questions are still challenging.

Phenome-wide association using human cohorts and using experimental rodent populations, such as the BXDs, is at early stage. There are just two key paper using animals: Li et al. (2010) and Wang et al. (2016). However, phenome-wide association has great promise because it scales well—across entire genomes of variants—and because it addresses the important and hard problem of pleiotropic effects of gene variants on surprisingly different phenotypes (e.g. Williams et al. 2014). The method also has practical promise because there is one large and open resource (www.genenetwork.org; see Mulligan et al. 2017) that can be used to test-drive this new method.

Here is a simple illustration of the process, the promise, and the problems of phenome-wide association. First the process: in the supplement of Wang et al. (2016) there are many tables of sequence differences that segregate in all crosses between two common inbred strains of mice—C57BL/6J and DBA/2J. These strains differ at about 5 million distinct sequence sites, of which about 12,000 are associated with missense variants. A much smaller number are associated with nonsense and frameshift mutations (Wang et al. 2016, Supplement Table 5).

The phenome-wide analysis starts by selecting a small number of interesting sequence variants and genes. For example, in the BXD family there are genes with validated nonsense variants that are likely to damage protein function including: Abca3, Peli3, Samhd1, Zfp277, Kif17, C15orf52, Etnk2, Proser3, Kcna10, Fam166b, Iqcc, Cela3a, Oog3, Gm5111, Cpmad8, and Dlgap5. Those genes listed in bold have reasonably high expression levels in brain (see Mulligan et al. 2017 for methods to evaluate expression efficiently). The gene, Kif17, already has known function in learning, memory, and schizophrenia (Wong et al. 2002; Tarabeux et al. 2010). Let’s select a novel gene of unknown function—C15orf52—to perform the type of phenome-wide scan described in Wang et al. (2016) and to see what kinds of problems we run into and what kind of hypotheses we can generate.

C15orf52 (also known as A43105L19Rik) is located on chromosome (Chr) 2 at about 118.75 Mb. There are a wide variety of systems genetics tools available (Schughart &
Williams 2017), and databases containing previously known information. To give but a few examples from each category: general overviews can be found at NCBI entrez (https://www.ncbi.nlm.nih.gov/gene), GeneCards (http://www.genecards.org) or GeneWiki (https://en.wikipedia.org/wiki/Portal:Gene_Wiki); phenotype and disease relationships can be retrieved from Online Mendelian Inheritance in Man (OMIM; https://www.omim.org) or Mouse Genome Informatics (MGI; http://www.informatics.jax.org); known expression of genes or their protein product are available at the Allen Brain atlas (http://www.brain-map.org), GenePaint (http://www.genepaint.org) or BioGPS (http://biogps.org).

This gene (C15orf52) is located in the middle of tight pack of 40 other genes, all within a 2 Mb interval. Most of these genes are highly polymorphic in crosses between the parental strains—C57BL/6J (B for short) and DBA/2J (D for short). The problem is that while we can easily look up phenotypes and “reverse map” them to the C15orf52 region using the large assemblage of BXD phenotypes, it will be hard to attribute function strictly to the D nonsense mutation in C15orf52. Let’s try this anyway.

**Step 1.** Find a SNP located close to the mutation that has been genotyped in all of the BXD strains. SNP rs27440124 is a good choice and will be used as a positional surrogate for the actual mutation.

**Step 2.** Compute the correlation between SNP genotypes, encoded in this case as −1 if the genotype is BB, 0 if BD, and +1 if DD. The top trait that covaries with this SNP is related to a marked difference in the degree of locomotor activation after a single treatment with 1.5 g/kg of ethanol. The G-to-P correlation is 0.65, and the D haplotype is associated with a robust increase in activity, whereas the B allele is associated with minimal change or even a moderate decrease (BXD Phenotypes 10129, 10791, and 10792 in www.genenetwork.org). The nominal p value associated with this correlation is <0.0005, and this seems to be a reasonably compelling p value, but perhaps less so in the context of several thousands of tests that we performed to extract the top hits. As a result, there is still a high risk that this is a false discovery. We can address, perhaps even redress, this problem by estimating the false discovery rate (FDR) following methods of Benjamini and Hochberg (1995) and we can also now reverse the analysis and ask just how strongly this trait maps to Chr 2 near C15orf52.

**Step 3.** We now use the forward methods to affirm that the G-to-P correlation is associated with a significant QTL. This is not really necessary, but one point of this article is to make it clear that reverse and forward methods are merging (Williams & Auwerx 2015). In this particular case, all of the acute ethanol activation traits map precisely to Chr 2 with a peak LOD of about 3 between 116.6 and 118.9 Mb (Demarest et al. 1999; see latest mapping in GeneNetwork). Demarest and colleagues also generated a large F2 intercross between the same parental strains (n = 900 cases) to validate and improve the linkage to Chr 2 (LOD of 5.3).

Our first conclusion is that we can be quite sure that there is at least one interesting trait associated with the mutation in C15orf52 in mouse—namely acute ethanol locomotor activity. To validate or refute this hypothesis that the C15orf52 protein modulates the
response to ethanol we would want to use genetic engineering techniques to introduce a deletion in the gene using a strain with the \( B \) haplotype, or better yet, to repair the nonsense mutation in a strain with the \( D \) haplotype, and show pronounced modulation of ethanol-induced locomotor activity. This would validate both G-to-P causality and put us on the path to understanding the genetic mechanisms of alcohol’s action on brain function and behavior.

What about potential pleiotropic actions of the mutation in \( C15orf52 \)? We can use our reference SNP to nominate candidate phenotypes (as opposed to candidate genes). Two blood chemistry traits also covary well (\( p < 0.001 \)) and also map to Chr 2 (Trait 17896 from Williams \textit{et al.} 2016, platelet volume; and Trait 12526 from Jellen \textit{et al.} 2012, hemoglobin concentration). If we were to exploit genetically modified animals in the next phase of work, it would make good sense to also check these phenotypes.

Genetics studies of brain and behavior are gradually becoming larger and more systematic. In the case of genome-wide studies of human behavioral diseases, sample size is king, and studies now are creeping up from tens of thousands to millions. The driver for this process is the small effect sizes of most sequence variants on behavioral traits—even those with very high heritability such as autism and schizophrenia. But there is another size dimension that is arguably even more important and that gets far less attention—the size of the phenome acquired for each subject. This has been a huge blindspot in biomedical research, brought about by our cottage industry approach to diseases. Few investigators have the expertise or budget to take a holistic approach to phenotyping humans, mice, rats, or drosophila, especially in non-inbred populations that include individuals with unique genomes and finite life-spans. For these populations, a large number of tests would need to be carried out in each individual and there may be carryover effects, where earlier testing influences the results of later testing. Even if these carryover effects are small, they may have cumulative effects when an entire phenome is being tested. These can be thought of as environmental effects, and may contribute to some of the lack of replicability between different cohorts. In inbred populations this can be overcome: the tests can be carried out in different individuals with identical genomes. In human cohorts and outbred model populations, such as the diversity outbred (DO) mice, these must be modelled as environmental interactions, which will require measurement of e.g. the effect of test order. This will also depend upon the relatively size of the carryover effect compared with all other environmental variation: we might expect that in large human cohorts, where there will be vastly different environments between participants, the carryover effect may well be drowned out by the rest of the environmental noise. In model organism populations, where other environmental variables can be tightly controlled, this carryover effect may contribute a much larger proportion to the environmental variability. A relatively simple way to overcome this would be to conduct the tests in different orders in different individuals and include the testing order as a covariate, however, this will require larger samples sizes to maintain power.

Over the next 50 years we are going to rapidly need to become comfortable with whole-phenome analysis if we expect to make genuine progress toward predictive (aka precision) health care. But there have been at least a few notable successes in aggregating large phenome collections for specific cohorts and crosses of both human and mouse (Denny 2014; Sudlow \textit{et al.} 2015; Wang \textit{et al.} 2016).
3. Behavioral analysis coupled to whole-transcriptomes, whole-epigenomes and other ‘omes’

In the previous section we describe how sequencing technology can be used to link sequence variants at the whole genome level with phenotypes. However, we need to understand how these variants alter the phenotype. It is not only important to find linkage between genomic variants and phenotypes, but we also need to understand the biological networks between levels of analysis. In this section, we outline some of the ways in which sequencing technologies have been used to tackle this problem, and how they could be used in the future.

Many of the significant GWAS hits that have been found in humans have been in non-coding regions (Civelek & Lusis 2014; Gusev et al. 2014; Albert & Kruglyak 2015; Zhang & Lupski 2015)—sites which are unlikely to alter the sequence of a protein product, but are much more likely to alter gene expression or isoform use: either how much, where, or when different versions of a protein are expressed.

Microarrays have traditionally been used to examine mRNA expression (Jansen & Nap 2001; Chesler et al. 2005; Mozhui et al. 2008), but RNA-seq has now become the method of choice for many investigators (reviewed by Hitzemann et al. 2013, 2014). We may give the impression here that microarrays are an obsolete technology, which has been overtaken by sequencing but this is not the case, as in a number of situations microarrays can be preferable: they are quick, cheap and bioinformatically very easy to interpret. RNA-seq is not without its challenges and problems, but offers several unique advantages over microarrays (Mortazavi et al. 2008; Mane et al. 2009; Tang et al. 2009; Walter et al. 2009; de Klerk & ’t Hoen 2015). For example, RNA-seq includes genotype data, and therefore can be used to identify allelic imbalance (e.g. Yeo et al. 2016), parent-of-origin effects, and imprinting (Hager et al. 2008).

For expression microarrays, the sequence of the RNA of interest should ideally be known. Generating accurate transcriptome data is difficult in individuals for whom potential sequence variants are undefined. This is exemplified in several studies (Peirce et al. 2006; Huang et al. 2009; Walter et al. 2009; Ciobanu et al. 2010), that have shown how sequence variants (as found by genomic sequencing), altered the apparent expression levels (i.e. because of poor hybridization or sequence alignment due to previously unknown SNPs).

Therefore, one potential advantage of RNA-seq over arrays for linking gene expression and behavioral phenotypes is that it allows testing of wild-type and wild-type derived populations (e.g., the Collaborative Cross), but even here it is important to ensure that the alignment procedure is not confounded by sequence differences and the sole use of a “canonical” genome assembly. Again, the solution is simply to sequence the genome of a subset of cases to produce a smart genome assembly that incorporates variants. Done properly, RNA-seq and related methods are beginning to enable the analysis of a broader range of behaviors in natural populations. This also introduces new challenges: wild-derived populations do not behave in the same way as inbred or domesticated laboratory populations. For example, wild-derived animals show climbing behavior during the tail suspension test, something which is rarely seen in traditional laboratory strains (Miller et al. 2010; Logan et al. 2013).

*Genes Brain Behav.* Author manuscript; available in PMC 2019 March 01.
Gene expression data and many other molecular data sets can be treated as microphenotypes, and QTLs can be mapped in much the same way as described above for behaviors, enabling us to identify relations between variants and expression. When these expression-QTL (eQTL) map to the same location as a QTL for a phenotype, then this supports the hypothesis that molecular expression variation is an intermediate step (and endophenotype) between the genotype and the behavioral phenotype. This is especially the case when the eQTL is local, e.g., expression is altered by a genomic variant close to the cognate gene (Qin et al. 2004; Doss et al. 2005; Mehrabian et al. 2005; Baud & Flint 2017).

Early studies using arrays were able to find links between QTL, gene expression and behavioral traits, e.g. expression of Drd2 and ethanol-induced conditioned place preference (Chesler et al. 2005), and including gene expression data became a useful method of identifying candidate genes within a QTL. Resources such as GeneNetwork (http://www.genenetwork.org; Mulligan et al. 2017) have been particularly useful, and contains molecular data for inbred populations, allowing expression to be used to identify candidates without expense to users.

More recent studies have used RNA-seq for transcriptomics. Using a population of Carworth Farms White (CFW) mice, along with genotyping by sequencing (GBS) and RNA-seq, Parker et al. (2016) took full advantage of sequencing technologies to examine behavioral and physiological phenotypes. They were able to identify a number of phenotypes where a SNP influenced both behavior and gene expression: for example, methamphetamine sensitivity mapped to a region on chromosome 9, tagged by rs46497021, containing two genes Cmc1 and Azi2. Both showed cis-eQTLs, but rs46497021 was most strongly correlated with Azi2 expression. Therefore, they are able to identify Azi2 as a novel candidate for a drug-abuse-relevant trait (Parker et al. 2016).

Another illustration of how sequencing is allowing greater insight is Kogelman et al. (2015), who took a multi-omics approach, combining RNA-seq with genotyping and protein data to investigate obesity. They integrate many different resources, including both human and animal data: the next step for us, of course, would be to add in behavioral phenotypes on top of this multi-omics data.

A caveat of transcriptomic studies which has been re-emphasized recently is age: the cellular transcriptome alters over time, with distinct ages having distinct expression profiles (Skene et al. 2017). This needs to be considered carefully during study design for several reasons. First, if a phenotype changes with age, a change in gene expression with age could be causative of the phenotype or both gene expression and phenotype could be independently correlated with age. Secondly, certain ages appear to be transcriptome trajectory turning points (TTTPs; Skene et al. 2017), when gene expression may reverse direction or plateau, rather than following a simple gradient. The most prominent of these in mice is around 156 days for male and 165 days for females, equivalent to 26.0 years and 27.5 years in human males and females respectively. Skene and colleagues were able to show that these TTTPs influenced the age of onset of schizophrenia, and it is probable that that they influence other disorders and behaviors. Further, interventions before or after these TTTPs may have different effects on phenotype, due to different levels of transcripts and their protein.
products. These TTTPs can be studies themselves (e.g. identifying the genetic basis of the timing of TTTPs), but they may also add unaccounted-for differences in a cohort if testing occurs during TTTPs. This study also shows the utility of large, publicly available datasets, as the authors were able to use a number of complementary datasets to test their initial results, and add confidence to their conclusion (Skene et al. 2017).

Although methylation is often thought of as an epigenetic mark altered by the environment, genetic variants can have a significant impact, and this has been repeatedly demonstrated in the human brain (Gibbs et al. 2010; Gamazon et al. 2013; Hannon et al. 2016). As with the transcriptome, microarrays have been traditionally used to detect methylation changes, but sequencing technologies are becoming more common, such as reduced representation bisulfite sequencing (RRBS) and whole-genome bisulfite sequencing (WGBS or BS-seq) (de Vega et al. 2017). Similar to the genome and transcriptome, sequencing of the methylome allows a potentially greater depth of coverage (not limited by the number of sites on a microarray) and the discovery of novel methylation sites.

Methylation can also act as an intermediate step between genomic variants and altered gene expression. For example, Bonder et al. (2017) used genome, transcript, methylation, and histone data in combination, finding many novel links between the different levels of regulation. They were able to show that the minor allele of rs3774937, associated with increased risk of ulcerative colitis, increased the expression of NFKB1 in cis, and that this increased NFKB1 bound elsewhere in the genome, decreasing DNA methylation at those sites and therefore causing trans-mQTLs (Bonder et al. 2017). As they acknowledge, their information about methylation is sparse, due to the use of microarrays, and therefore the use of methylation sequencing would allow the further elucidation of these complex biological interactions. Additionally, being able to make these associations relied on ChIP-seq data, demonstrating that whole-epigenome sequencing could offer exciting insights into the regulation of phenotypes.

We have shown evidence of the advantages of whole-transcriptome and whole-epigenome analyses above, how these have been used, and are confident that these will become commonplace. Further, new methods such as whole-proteome and whole-metabolome analyses are coming closer to fruition, which will allow an additional layer of biology to be put on top of these, bringing us nearer to the final behavioral phenotype. Proteomics is being significantly supported by the availability of genome and transcriptome data, acting as references for potential proteins, and many of the challenges are being overcome (reviewed by Low & Heck, 2016). Proteomics is already being carried out in the BXD population with integrated RNA-seq (Wu et al. 2014), and this approach is ripe to be applied to behavioral phenotypes.

4. Reduced Complexity Crosses: a new method dependent on full genome sequence

Mapping gene variants onto complex traits has typically required highly genetically diverse crosses or populations. The dogma has been that more variation is better in two ways—in providing higher levels of phenotypic difference, and in providing large numbers of underlying sequence differences. Two cases in point: the major community resources of behavioral genetics now include the BXD family of mice that segregate from about 5 million
variants, the CFW outbred population that segregated for ~7 million SNPs (Nicod et al. 2016), and the collaborative cross that segregates for an unprecedented ~50 million variants (about four times that of common alleles in all of humanity; Srivastava et al. 2017). As described in this section, fewer variants can actually be much more efficient, especially when it comes to isolating the specific sequence variants that modify behavioral phenotypes.

A reduced complexity cross (RCC) is generated by mating two or even more isogenic or closely related substrains (e.g., C57BL/6J, C57BL/6NJ, C57BL/6ByJ, and C57BL/6JEiJ). F1 hybrids are crossed to either generate F2 progeny or backcrosses (reviewed in Bryant et al., 2017). The resulting RCC segregates for variants that underlie behavioral differences between progenitor substrains. Genotyping followed by genetic mapping in a conventional intercross between common inbred strains will typically result in QTL intervals that contain hundreds of polymorphic genes and several thousand variants. Candidate intervals in the RCC, in contrast, are similarly sized but contain orders of magnitude fewer candidate gene variants. As an example, consider a cross between C57BL/6 (B6) substrains.

B6 substrains are used for medical research worldwide. Separation of colonies at independent and isolated facilities has occurred since the 1950s and led to fixation of colony-specific genetic mutations (genetic drift) and the creation of substrains. Commonly used B6 lines include C57BL/6J (B6J, maintained at the Jackson Laboratory) and C57BL/6NJ (B6NJ, used to generate genetically engineered knockout mice as part of the International Mouse Knockout Consortium). These two substrains, separated at the 32nd generation of inbreeding, differ at roughly 10,000 SNPs (Keane et al. 2011) and ~50 polymorphisms have been validated as coding variants (Simon et al. 2013). Several of these variants have a functional impact on trait variation, including mutations in Nnt (Freeman et al. 2006), Crb1 (Mattapallil et al. 2012), and Cyfip2 (Kumar et al. 2013; Kirkpatrick et al. 2017). In addition, these substrains differ in an astonishing number of behavioral traits; phenotypic differences have been reported for alcohol consumption and preference (Mulligan et al. 2008), cocaine-and amphetamine-induced locomotor sensitization (Kumar et al. 2013), binge eating (Kirkpatrick et al. 2017), anxiety-like behavior (Matsuo et al. 2010; Simon et al. 2013), response to a thermal stimulus (Matsuo et al. 2010), acoustic startle and pre-pulse inhibition (Matsuo et al. 2010; Simon et al. 2013), depression-like behavior in the forced swim test (Matsuo et al. 2010), fear-conditioned response (Bryant et al. 2008), and response to chronic stress (Sturm et al. 2015). Connecting the dots between genetic variation and behavioral variation in this particular RCC (an F2 cross between B6J and B6NJ) is greatly facilitated by the low genetic complexity. For example, consider two crosses. Cross A is a RCC between the B6J and B6NJ substrains. Cross B is a typical F2 cross between the highly divergent inbred strains B6J and DBA/2J. In cross A there are roughly 20,000 segregating variants compared to a little over 5 million in cross B. Thus, identically sized behavioral QTLs will contain 250x fewer variants in cross A (the RCC) compared to cross B (a typical F2 cross), greatly facilitating candidate gene or candidate variant identification. In fact, the RCC strategy has already been used successfully for crosses between B6 substrains in order to identify a missense mutation in the Cyfip2 gene that drives substrain differences in binge eating (Kirkpatrick et al. 2017) and the behavioral response to cocaine (Kumar et al. 2013). This approach has also been used to identify a gene variant driving differences in alcohol consumption in inbred selected rat lines (Zhou et al. 2013). Advances in sequencing
now facilitate the RCC approach for any species for which substrains or nearly isogenic lines are available. Of course, there are caveats with the use of these and other isogenic or inbred lines that result from >100 years of domestication. Inbreeding and maintenance under laboratory conditions results in a loss of allelic and phenotypic diversity (Chesler 2014). Therefore, they may not express the full range of phenotypes seen in the wild population, for example activity and aggression are often reduced in domesticated population (Connor 1975), although there is some evidence of greater extremes of behavior in inbred lines (Fonio et al. 2006).

The recipe for a successful RCC design (see Bryant et al. 2017) involves: (1) the existence of heritable, large effect-size, and reproducible behavioral variation between two nearly isogenic strains (these may not exist for all behaviors), (2) generation of an F2 cross or backcross between the two nearly isogenic parental inbred strains, (3) accurate measurement of the behavior of interest in the resulting recombinant progeny, (4) genotyping of the RCC progeny, and (5) genetic mapping and/or fine-mapping to resolve QTLs and quantitative trait genes underlying trait variation.

The genotyping strategies in step 4 are an essential part of the design of RCC experiments, and this is the step that has been greatly facilitated by the ease of generating high quality genomes and comprehensive lists of polymorphisms. It is practical to do this for any pair of substrains of mice or rats. It is even possible to do this for spontaneous mutations that arise within colonies of inbred strains. For example, Rosen and colleagues (2013) discovered a heritable neocortical heterotopia in one of two coisogenic lines of BXD29. They backcrossed the two lines—BXD29-\textit{Tlr4}^{pos-2/J} and BXD29/TyJ—and using a simple rtPCR genotype assay showed that the heterotopia is not caused by a mutation in \textit{Tlr4}. A logical next step would be to sequence both strains at more than 30X coverage to extract a list of potential SNP markers that could be used for mapping and to generate a complete list of genes that may be involved in migration of neocortical neurons.

5. Genotyping strategies

Many genotyping methods are available for mapping RCCs as well as the other types of population mentioned above, such as recombinant inbred lines or outbred populations. Selection of the genotyping method depends on characteristics of the population, such as whether or not the progenitor strains have been sequenced, the cost of genotyping in terms of number of genotypic markers per individual, and computational considerations. A typical strategy that depends on the availability of sequenced genomes, has low computational requirements, is relatively affordable, and can provide ample marker coverage (typical spacing for an F2 cross in mice is ~10 to 20 cM or 20 to 40 Mb) is genotyping by array. For example, the latest version of the Mouse Universal Genotyping Array platform (MUGA; GigaMUGA, ~$100 per array) was designed for genotyping crosses from common inbred strains and the Collaborative Cross and Diversity Outcross populations derived from eight progenitor strains. The GigaMUGA (Illumina Infinium HD platform) also contains 96 to 391 informative markers that differentiate multiple murine substrains, including C57BL/6J and C57BL/6N, 129S1/SvImJ and 129S6/SvEvTac, A/J and A/WySnJ, BALB/cJ and BALB/cByJ, C3H/HeJ and C3HeB/FeJ, and finally, DBA/2J and DBA/2DeJ (Morgan et al. 2015).
Another approach involves selection of markers and utilization of multiplex quantitative PCR methods and platforms such as the Genotyping Dynamic Array Integrated Fluidic Circuits (IFCs) offered by Fluidigm Inc. This method has been used to map B6-derived RCCs (Kumar et al. 2013; Kirkpatrick et al. 2017) has low computational requirements and is relatively inexpensive, but is only moderately high-throughput. Current formats (samples x genotypes) are 192 × 24, 48 × 48, and 96 × 96. We highlight some additional higher-throughput strategies involving next-generation sequencing applications: targeted high-throughput genotyping and genotyping by sequencing.

Higher throughput approaches to genotyping include the use of multiplex PCR to generate hundreds to thousands of allele-specific products from a targeted set of marker loci that can then be identified by DNA-seq. In this approach pioneered by K. Lamour and colleagues at FloodLight Genomics, multiplexed and targeted amplification of genomic loci containing SNP markers in every sample is followed by sequencing of the resulting libraries, and alignment and variant calling to identify genotypes. Cost varies greatly depending on: (1) the number of locus-specific primers needed and (2) the number of individuals to be genotyped. As with any NGS application, targeted resequencing of marker loci can be computationally intensive if not commercially outsourced. Also, this approach depends on a priori knowledge of marker position and sequence. However, this approach allows for massive multiplexing of samples (as many as 96 per run) and genotypes (as many as 500) to generate dense marker panels for an RCC.

In contrast to the design-based genotyping platforms discussed thus far, genotyping by sequencing (GBS) leverages NGS technology for simultaneous variant discovery and genotyping. It does not require genomic sequencing of progenitor strains nor does it require a priori knowledge of positions or genotypes of markers. Thus, this approach is useful for RCCs that involve inbred selected lines or other cases in which sequence data are not available for parental strains. In a typical GBS approach, combinations of restriction enzymes are used to create a reduced representation of the genome. Instead of sequencing the entire genome, the focus is limited to restriction sites where polymorphisms that impact restriction sites or adjacent regions, can be detected (Parker et al. 2016). Sequencing coverage is determined by the combination of restriction enzymes that are used. This can result in greater coverage at polymorphic regions compared to random whole-genome sequencing but there will also be large gaps in the sequence where there are no restriction sites. An advantage of this approach is that new markers are discovered and can subsequently be used for mapping. This approach does have some design flaws when paired with a typical RCC, which will segregate at very few variants. Low numbers decrease the likelihood that a variant will be located near a restriction site. It may therefore be difficult to generate a dense marker panel using this approach.

Another GBS approach involves RNA sequencing of tissue from F1 or F2 individuals. RNA-seq of F1 individuals can be used to analyze allele specific expression in order to identify gene variants associated with cis-eQTLs (Castel et al. 2015). RNA-seq can also be extended to the F2 in order to identify variants that can subsequently be used for genotyping. These variants are usually in coding regions, although non-coding variants can be detected as read depth increases and/or if the intron has atypically high expression. The advantage of these
approaches is that they allow for simultaneous analysis of gene expression and genotyping. The resulting analysis of gene expression can improve identification of candidate QTL genes (through identification of cis-eQTLs) and provide a deeper understanding of the networks and brain regions involved in behavior. However, caution should again be used when applying this approach to the RCC because the total number of coding variants segregating in RCCs is generally low. Far fewer variants are expected to be located in exons, and thus most variants will be inaccessible to sequencing. Recall that only ~50 validated coding variants have been identified between B6 substrains. This number would be sufficient as a first pass for genotyping in this cross, but additional markers would be needed to further define QTL intervals.

Substrains and nearly isogenic lines arise unintentionally as a natural consequence of genetic drift. Variants that are fixed in these lines sometimes produce alterations in behavior. In the context of highly identical genomes, these mutations are typically highly penetrant and produce large effects because they are not muted by the presence of numerous background QTLs. In this context of reduced genetic complexity, RCCs facilitate identification of gene variants that have a large impact on behavior. Advances in sequencing greatly increase the utility of the RCC design. These approaches, especially targeted high-throughput genotyping using massively multiplexed samples and targeted resequencing of marker loci, facilitate the essential step of genotyping by allowing for higher throughput. Some of these approaches do not require genomic sequencing of the progenitor substrains or nearly isogenic lines, such as GBS, however, these should be applied cautiously to the RCC design. Finally, as sequencing prices fall it may soon be feasible to sequence all individuals at moderate coverage (40x or higher). The depth of coverage needed will be dependent upon the application, and will be different between e.g. genotyping and differential expression profiling. This has been reviewed by Sims et al. (2014) and a referenced summary table is available at https://genohub.com/recommended-sequencing-coverage-by-application. Integration of complete genome profiles with multi-tissue or even single cell RNA-seq and epigenetic profiles will greatly enhance not only identification of variants that modulate behavior, but also underlying regulatory networks and molecular mechanisms.

6. Frontiers in behavioral genetics and the impact of sequencing

A. Social genetics—The focus thus far has been on direct genetic effects (DGEs): the effect of an individual’s genotype on the same individual’s phenotype (Geldermann 1975). However, social genetic effects, also known as indirect genetic effects (IGEs), have become increasingly recognized (Moore et al. 1997; Wolf et al. 1998; Agrawal et al. 2001; Wolf 2003; Mutic & Wolf 2007; McGlothlin & Brodie 2009; Wilson et al. 2009; Bijma 2010; Teplitsky et al. 2010; Trubenová & Hager 2014). Social genetic effects refer to the genotype of one individual altering the phenotype of a second individual, and these effects are notable in any group of interacting individuals (Moore et al. 1997). For a social genetic effect to occur, the focal phenotype must be plastic in response to the interacting individual’s phenotype (Donohue 2003) and this indirect effect must have a genetic basis (Kirkpatrick & Lande 1989; Wolf et al. 1998; Petfield et al. 2005).
Social genetic effects are not just important for understanding the behavioral phenotype of an individual, but also for the evolution of behaviors, as they represent an environment which is composed of the genotype of interacting individuals; therefore it is an environment that can respond to selection (Moore et al. 1997, 2002; Wolf & Brodie 1998; Bijma & Wade 2008; McGlothlin et al. 2010). Indeed, some behaviors cannot exist without social interactions, for example courtship (Whitney et al. 1973; Doty 1974; Petfield et al. 2005; Neunuebel et al. 2015), play (Cox & Rissman 2011; Lukas & Wöhr 2015), aggression (Wilson et al. 2009; Anholt & Mackay 2012; Camerlink et al. 2013) and parental care (Kölliker et al. 2000; Kölliker & Richner 2001; Hunt & Simmons 2002; Smiseth et al. 2012; Ashbrook & Hager 2017). Some of the earliest social genetic effects identified were maternal genetic effects (MGE), and these were studied extensively in livestock (Dickerson 1947; Willham 1963, 1972; Hanrahan 1976; Ellen et al. 2014), although some work was carried out in rodents (Hanrahan & Eisen 1973). Social genetic effects have now been identified for a broad range of phenotypes in a broad range of species (Aggrey & Cheng 1993; Smith & Wettermark 1995; Neser et al. 2001; McAdam et al. 2002; Moore et al. 2002; Rauter & Moore 2002a, 2002b; Wolf et al. 2002; Hunt & Simmons 2002; Petfield et al. 2005; Miller & Moore 2007; Mutic & Wolf 2007; Bijma et al. 2007; Kent et al. 2008; Wilson et al. 2009, 2011; Bleakley & Brodie 2009; Teplitsky et al. 2010; Head et al. 2012; Moore 2013; Sartori & Mantovani 2013; Camerlink et al. 2013; Bailey & Hoskins 2014; Ellen et al. 2014; Greenwood & Peichel 2015). As with many environmental effects, social genetic effects can have their long-lasting action through altering the epigenome of the target individual. As we have discussed above, next generation sequencing is allowing investigation of the whole-epigenome, and social genetic effects will be an important aspect of this.

Although it has now been shown that a variety of behaviors expressed in mice are influenced by the genotype of interacting individuals (e.g. Hughes 1989; Cheverud & Moore 1994; Wolf & Brodie 1998; Baud et al. 2017), there have been few studies to identify the genetic loci influencing these traits (Wolf et al. 2002, 2011; Wilson et al. 2004; Casellas et al. 2009; Wolf & Cheverud 2012). Recently, Baud et al. (2017) used RNA-seq to show that social genetic effects can influence gene expression and health-related traits in a conspecific. To do this, they cohoused mice of the C57BL/6J (B) and DBA/2J (D) strains in pairs of either B/B, B/D or D/D, and after six weeks collected a panel of phenotypes, including anxiety and stress behaviors. Eleven out of 50 recorded phenotypes showed social genetic effects, and there was an interaction between the genotype of the target individual and the interacting individual. Ashbrook et al. (2015a, 2017; Ashbrook & Hager 2017) have begun to map social genetic effect QTL for a number of interactions between mothers, offspring and siblings. To do this, cross-fostering was used: for social genetic effects on maternal behavior, the genotype of the mother was kept constant between all litters, whereas the genotype of the offspring was differed (different lines of the BXD population). Therefore, any variation in the behavior of the mothers could be mapped to the genotype of their foster offspring. In both of these experimental designs, it has been shown that for some phenotypes, social genetic effects can explain more of the phenotypic variation than direct genetic effects (Ashbrook et al. 2017; Baud et al. 2017).
We hope that we have now hammered home that social genetic effects are important and require careful consideration. The literature above has shown that social genetic effects can influence phenotypes due to cage mates (Baud et al. 2017), maternal care (Ashbrook et al. 2015a), and early life sibling interactions (Ashbrook et al. 2017), and therefore potentially influence every study performed in model organisms. In most lab experiments the social genetic effect is thought to be included in the variance due to cage or litter effects, however Baud et al. (2017) showed that modeling cage effects does not completely mitigate the social genetic effects. In inbred lines, it is not strictly possible to distinguish between a QTL due to DGEs or social genetic effects: the mother and siblings are the same genotype, and therefore any QTL mapped in the focal individual could in fact by due to a social genetic effect of mother or sibling. We should emphasize that phenotypes here include gene expression and methylation: no phenotypic level is immune to this, and therefore it may have an unrecognized influence on sequencing data.

We will give a few examples of how the social genetic effect studies above could be extended to a whole genome level. We have previously shown that cross-fostering can be used to investigate maternal, offspring and sibling effects during early life (Ashbrook et al. 2015a): the technique is described in depth in Ashbrook & Hager (2017) and this could be extended into later life phenotypes, or, for example, transcriptome mapping. It has been clearly demonstrated that differences in maternal care can influence offspring behavior and gene expression during adult life (e.g. Francis et al. 1999; Weaver et al. 2006; Gracia-Rubio et al. 2016), but identification of genetic variants in mothers that can then be directly associated with offspring adult gene expression or behavior has not, yet, been carried out.

The co-housing design of Baud et al. (2017) could be extended such that the genotype of one cage-mate is constant in all experiments, and the genotype of the second cage-mate is different in each experiment. Therefore, differences in the phenotype of individuals of the constant genotype will be due to differences in the genotype of their cage-mates, and these can be mapped to the genome. The low cost of sequencing means that this can now, potentially, be carried out in any population. A further step beyond this would be to map ‘epistasis between genomes’. The phenotypic outcome of a social genetic effect is not just dependent on the genotype underlying the social genetic effect, but also on the genotype of the individual displaying the phenotype. Let us take aggression as an example: an individual whose genotype predisposed to increased aggressive behavior (individual 1) may cause an increase in submissive behavior in an interacting individual (individual 2: a social genetic effect). However, if individual 2 also has a genotype predisposing to aggressive behavior, then individual 1s behavior may lead to an increase in aggressive behavior from individual 2 (still a social genetic effect, but in the opposite direction). Therefore, the phenotype of individual 2 is dependent upon the combination of its own genotype and individual 1s genotype. Only large scale genomic studies, enabled by high-throughput sequencing, with careful monitoring of the interactions between individuals will allow mapping of this effect. This could be done using another extension of the designs suggested above, but such that an individual of every genotype is fostered or housed with an individual of every other genotype. Again, to be able to map this epistasis between genomes, the genotype of all interacting individuals must be known.
An alternative extension is to not just interactions between two individuals, but the more natural setting of interactions between many individuals. Home-cage monitoring of several animals in the same cage is now becoming feasible (e.g. Bains et al. 2016). Therefore, if each individual in the cage has their genome sequenced, and all of them have their behavior monitored 24/7, we can potentially detect many direct and social genetic effects: however, having more than two genotypes can be expected to make analysis and mapping exponentially more complicated.

B. Forward and reverse behavioral genetics of wild populations—As described above, next-generation sequencing has allowed the sequencing of wild populations as it does not require the genotype to be known in advance, as microarray methods do, although having a reference genome for a related species certainly makes the process easier. The availability of long-read sequences vastly improves our ability to assemble new genomes, and new technologies, such as PacBio and Nanopore, are now able to give reads between 100–1000 kilobases, an order of magnitude longer than current Illumina sequencing (Shendure et al. 2017). Combining these new technologies with ‘traditional’ short read NGS will greatly improve our ability to assemble de novo genomes (see, for example Bickhart et al. 2017, for a de novo assembly of a large mammalian genome). With this revolution in sequencing it will again be the phenotype which becomes important. Here we highlight a few examples of ‘wild’, or at least non-model, populations and how they are now being used for behavioral genetics.

Using a RAD-tag approach, Weber et al. (2013; reviewed by Hu & Hoekstra 2017) were able to map the genetic architecture of burrowing behavior in a cross between two species, *Peromyscus polionotus* and *P. maniculatus*. They showed that different aspects of the burrow (length and presence of an escape tunnel) are under the control of different genetic loci. This demonstrates that new sequencing technologies can be used to investigate behaviors in non-model organisms, and give insight into the evolution of behavioral traits, which may not be amenable to study in model organisms. In a closely related species, *P. leucopus*, a combination of next-generation genomic-sequencing and transcriptomic-sequencing was used to show that transcriptomics could be carried out in a wild population without a reference genome (Harris et al. 2015), again giving insight into evolution, and into the divergence of populations.

In foxes, *Vulpes vulpes*, genotype-by-sequencing was carried out by Johnson et al. (2015) to determine the genetic architecture of tameness and aggressiveness, in a long-term study of domestication. They were able to detect QTL and candidate SNPs for this important feature of domestication, a process that comes with a wide-variety of behavioral adaptations. It also clearly shows how a genomics aspect is now being added to many behavioral studies, and can be potentially added retrospectively to add value to already long-running experiments.

Another species long studied, but that has only recently been the subject of genomic analyses is the domestic dog (*Canis lupus familiaris*). Sequencing has been used to examine a range of behaviors from stereotypically canid behaviors, such as pointing (Akkad et al. 2015), to behaviors which may give insight into human disorders, such as obsessive-compulsive behavior (Tang et al. 2014). Under the common assumption that similar

---

*Genes Brain Behav. Author manuscript; available in PMC 2019 March 01.*
behavioral phenotypes have a similar genetic basis across species (a subject we come back to), these obsessive-compulsive behaviors could be of interest for obsessive compulsive disorder, and the aggression related genes in foxes could be of interest for antisocial personality disorder.

All of these studies are individually important for the behavior studied, however, the SNPs discovered and the transcriptome constructed can be applied to other studies in that species, as long as data are openly shared. For example, future studies in Peromyscus species now have reference genomes and transcriptomes, as well as known genetic polymorphisms, meaning that every sequencing study carried out increases our knowledge of the species as a whole.

C. Behavioral genetics, brain structure, and connectome—In the bulk of this review we have concentrated on cellular and subcellular phenotypes, such as gene and protein expression, as intermediates between genetic variation and behavior. However, cellular changes do not influence behavior in isolation. Behavior is the outcome of the integration of activities of many cells, synapse, and hormones of different types working together in brain and body. Polymorphisms that influence structure or connectivity are necessary to understand behavior, and especially pathological behavior (Shepherd 2013). Recent developments in imaging technologies (Badea et al. 2009, 2012; Shibata et al. 2015) have made it possible to image in large numbers of living subjects, allowing novel insight into the link between genes, brain and behavior.

Natural variation in the size and connectivity of brain structure are heritable traits and thus can be mapped in the same way as the behavioral traits described throughout this review. Furthermore, they also provide an additional layer of biological insight: in much the same way that we can use an eQTL as an intermediate phenotype between a genetic variant and behavior, we can use structural changes as an intermediate phenotype between a genetic variant and behavior to tell us more about the functional pathway between the two.

We are beginning to understand the genetic architecture of subcortical brain structure in both mouse (Hager et al. 2012) and human (Badea et al. 2009; Hibar et al. 2015; Satizabal et al. 2017; Thompson et al. 2017), with individual genes being identified which influence the volume of subcortical areas. Similarly to many other phenotypes, it has been found that individual genomic variants have only a small impact on the total variation in brain structure size, and so large consortia have been established for the meta-analysis of brain imaging GWASs, such as the ENIGMA consortium (Thompson et al. 2016, 2017).

Connections between these physiological changes and behavioral, especially pathological, phenotypes are being made (e.g. Schmaal et al. 2016, 2017; Satizabal et al. 2017; Thompson et al. 2017). For example, Satizabal et al. (2017) observed strong genetic correlations between reduced nucleus accumbens and caudate nucleus volumes and risk of bipolar disorder. More of these studies are being carried out, including for schizophrenia, depression and ADHD (Thompson et al. 2016), with results forthcoming.
It is not just the size of brain regions which influence behavior, but the connectivity between brain regions (Shibata et al. 2015). Heritability has been established for a number of measures of brain connectivity (Glahn et al. 2010; Smit et al. 2010; Fornito et al. 2011; Jahanshad et al. 2013; Bohlken et al. 2014; Thompson et al. 2016), therefore making this amenable for mapping, and use as an intermediate phenotype. The mapping of every connection in the nervous system has been named the connectome, analogous to the mapping of every gene in the genome. This idea of a ‘connectome’ was first outlined by Sporns et al. (2005) and Hagmann et al. (2007), although the first of what we would now call a connectome was published by White et al. in 1986 for Caenorhabditis elegans (Emmons 2015). White et al were also the first to take the next step: identifying how genetic mutations altered the connectome, and therefore altered behavior (Miller et al. 1992; White et al. 1992; Emmons 2015). They showed that a specific mutation in C. elegans altered synaptic input to a specific subset of motor neurons, and therefore specific changes to locomotor behavior (White et al. 1992).

The connectivity between brain regions has being increasingly used to investigate the pathology of brain disorders, the symptoms of which include behavioral changes (Cao et al. 2014; Fornito & Bullmore 2015; Narr & Leaver 2015). Macroscale connectomes (connections linking large-scale neuronal populations, ~ mm resolution) have been mapped for both human (Van Essen et al. 2013; Glasser et al. 2016) and mouse (Calabrese et al. 2015; Kuan et al. 2015), and techniques for this have been reviewed by Fornito & Bullmore (2015) and Shibata et al. (2015). However, complete microscale connectomes (a map of each and every synapse) are still a hope rather than a reality for any mammal, as these connectomes will be orders of magnitude larger than the C. elegans connectome mentioned above: there are around 300 neurons in C. elegans, and an estimated \(10^8\) in humans (Emmons 2015).

D. Directly homologous behaviors in model organisms and humans—Genomic sequencing has allowed the discovery that the vast majority of genes have homologs between rodents and humans, and analogous variants influence analogous behaviors. However, a common criticism of behavioral genetics in model organisms is that it does not have direct relevance to human phenotypes (e.g. low face validity; van der Staay et al. 2009), especially when they are ‘advertised’ as disease models (e.g. the tail suspension test for depression). This criticism is often countered by the fact that traits may share a common biological or pharmacological basis, even if the resulting phenotype is dissimilar. Nevertheless, there has been an increasing interest in developing behavioral tests which can be applied with minimal differences in both model organisms and humans, the argument being that it allows us to distinguish between homologous and analogous behaviors (that is, behaviors which are inherited from a common ancestor, and similar behaviors which have developed independently; Nithianantharajah & Grant 2013). A number of previous attempts have been made to find shared genetic influences on phenotypes which are analogous in model organisms and human (de Mooij-van Malsen et al. 2009; Koutnikova et al. 2009; Leduc et al. 2011; Poot et al. 2011; Schofield et al. 2012; Ashbrook et al. 2014, 2015b; Wang et al. 2016), however, these have not specifically tested if traits are homologous. With
genomic sequencing’s ability to give us precise resolution of homology amongst genomes, we can now move on to precise resolution of homology amongst phenotypes.

One of the first to do this were Nithianantharajah & Grant (2013; Nithianantharajah et al. 2013), who used touchscreen approaches to analyses cognitive domains in both mouse and human. They have the advantage that a battery of tests can be carried out in the same environment, with minimal experimenter contact, providing standardization and robustness (Nithianantharajah & Grant 2013). They were able to show that subjects with a DLG2 mutation, which has been associated with schizophrenia, and mice with a Dlg2 mutation has a strikingly similar cognitive profile, which is distinct from mice carrying mutations of Dlg2 paralogues (Nithianantharajah et al. 2013). The approach is, therefore, cross-species reverse genetics, and as such would not be possible without genomic sequencing. This high-throughput ‘aligned mouse–human touchscreen battery’, combined with high-throughput next generation sequencing could allow the rapid identification of homologous variants underlying homologous disease relevant phenotypes.

Earlier studies had been designed to try and developed human equivalents to traditional animal paradigms, e.g., human versions of the open field test (Young et al. 2007; Perry et al. 2009) and Morris water maze (Shipman & Astur 2008). This approach has been described as ‘reverse-translational’, since they were developing a human model of rodent behavior (Malkesman et al. 2009). Although these initial studies did not carry out genetic analysis, it should be clear how their paradigms could be used for genetic screening, as motor and exploratory behavior can be non-invasively examined, and combined with genomic data from the human and mouse subject to find homologous genes associated with homologous behaviors.

One consideration here is the difference between innate (those which arise spontaneously, or are instinctive; such as exploratory behavior above) and learned behaviors (e.g. the behaviors taught in touchscreen tests). We may perhaps assume that innate behaviors are under greater genetic control, as they arise spontaneously, whereas learned behaviors are not just dependent upon the individual’s ability to perform the task, but also their capacity and motivation to learn, which in turn can be influenced by the environment. However, it is not the case that innate behaviors are easier to map, due to these behaviors often being vital to the survival of the organism. Therefore, these traits are well buffered with no individual variant having a large effect, as they would be quickly selected against (imagine a variant that significantly disrupted feeding or mating behavior).

The ability to map variants underlying differences between innate and learned behaviors will depend on the experimental design: a well-controlled laboratory experiment, where great care has been taken to minimize environmental variables will have greater power than a measurement in a wild population, where confounding variables could not be measured, independent of the difference between learned and innate behaviors. Take, for example, touchscreen tests which often provide phenotypes for which environmental effects can be minimized, delivering ‘clean’ traits which can be efficiently mapped (Nithianantharajah & Grant 2013; Phillips et al. 2017).
Theoretically, large phenome datasets in either models or humans may provide us with the ability to test the assumption of different genetic architectures of innate and learned behaviors. If all behaviors were classified as either learned or innate, the average genetic effect size of each could be calculated and compared. Unfortunately, in practical terms this is unlikely to be the case, as the noise level of current phenotypic assays is large, and may well drown out any real difference.

In either case, homologous behaviors in humans and model organisms are of great interest and utility for identifying shared genetic and biological underpinnings of phenotypes, especially those with direct relevance to human health. If we can show that a homologous phenotype is associated with a homologous gene across species, it gives us far greater confidence in testing interventions in model organisms.

**Conclusion**

**It’s the Phenotype, stupid**

What was once horrifically expensive—full genome sequence—is now affordable and will soon be cheap. A complete analysis of a 30X genome currently costs about $1000 per genome, and Illumina claims cost will soon drop close to $100 (Illumina 2017). In contrast, the relative expense of phenotypes continues to rise. QTL studies and GWASs are usually limited by sample size and the cost of generating reliable qualitative and quantitative experimental and clinical phenotypes across cohorts of hundreds to millions of subjects. This is true whether dealing with yeast, corn, fish, mice, rats, humans, *Arabidopsis*, nematodes, pigs, or fruit flies. GWAS projects are beginning to look more like phenotyping projects than genome projects. While GWASs are still not easy, researchers are getting much better at genotyping logistics and analyses, and over the next ten years we can expect the cost and effort to drop quickly on the genome side of the genotype-to-phenotype equation.

The Precision Medicine Initiative does involve a massive amount of sequencing, but what sets this and other incipient projects apart is the depth of phenotype data for normal control subjects that they will be generating over longer period of time (https://allofus.nih.gov). This is also true of the Human Connectome Project (www.humanconnectome.org), the UK Biobank (www.ukbiobank.ac.uk), the Million Veterans Program (www.research.va.gov/MVP/default.cfm), the 100,000 Genomes Project (genomicsengland.co.uk), and many programs in large hospital systems (e.g., BioVu at Vanderbilt, www.vumc.org/dbmi/biovu), and many commercial efforts such as those by deCode (www.decode.com) and 23andme (www.23andme.com).

At this point any large systematic phenotyping project that retains tissue is likely to have a comparatively inexpensive genetics component “bolted on”. For example, investigators funded by NIH National Institute on Aging have been working assiduously for years using a complex 4-way mouse F2 intercross to understand how a multitude of interventions affect longevity (www.nia.nih.gov/research/dab/interventions-testing-program-itp). Scientists at three sites have systematically phenotyped ~15,000 animals using a range of diets, pharmaceuticals, and supplements. Fortunately, they had the foresight to save tails, and the genetic component is now being bolted on at a cost of about $20/animal—a bargain given the cost of obtaining longevity data. The same is also true for the Framingham Heart Study.
that began as an epidemiological analysis of heart disease, but which in its third generation has become a major genetics and genomics study with sequence data for ~4000 subjects (www.nhlbiwgs.org/group/fhs).

We will soon have comprehensive catalogs of common and rare sequence variants of all types for many species

This is already almost true now for human, mouse, and drosophila. In the case of human, it is possible to browse a first draft list of ~85 million short variants segregating in ~2500 sequenced humans from 26 different groups (http://phase3browser.1000genomes.org). In the case of mouse, it is possible to browse a list of ~45 million SNPs defined by whole genome sequencing of 36 strains by region, gene, and strain contrast (phenome.jax.org/snp/retrievals; Keane et al. 2011). The web browsers may not yet be easy to use, but interfaces and coverage will be improving over the next few years (for example, see the attractive ClinGen site at www.clinicalgenome.org). Even a novice will soon be able to extract a list of intriguing missense variants (common or rare) in any gene or protein domain in humans and several model organisms. It will also soon be practical to extract lists of sequence variants (long and short) that affect epigenetic state expression level, and splicing of transcripts. From the geneticist's perspective we will be able to see the tips of the roots of causality. Whether we can understand the networks that produce the many leaves of phenotypes is an open question.

The phenome is a mess

There really can be no argument: the phenome is central to the future of biology and what we audaciously or presumptuously call precision medicine. But the study of phenomes for any organism is still a backwater of a backwater. This is true in both model organisms and humans: behavioral phenotypes are measured with different methods, and yet still given the same name; disorders in humans, especially psychiatric disorders, can have very different symptoms and yet are lumped together. While we have become good at identifying homologous genes and variants across species, we struggle much more at identifying homologous phenotypes. While we will soon have complete and precise genomes, the next and much hard step will be to gain complete and precise phenomes.

Acknowledgments

We thank the support of the UT Center for Integrative and Translational Genomics, and funds from the UT-ORNL Governor’s Chair, NIDA grant P30DA044223, NIAAA U01 AA013499 and U01 AA016662. We would also like to thank Melloni N. Cook, Mike E. Hook, and the anonymous reviewers for their helpful comments on the manuscript.

References


Genes Brain Behav. Author manuscript; available in PMC 2019 March 01.


Chesler EJ. Out of the bottleneck: The Diversity Outcross and Collaborative Cross mouse populations in behavioral genetics research. Mamm Genome. 2014


*Genes Brain Behav. Author manuscript; available in PMC 2019 March 01.*


*Genes Brain Behav*. Author manuscript; available in PMC 2019 March 01.


Genes Brain Behav. Author manuscript; available in PMC 2019 March 01.


