Coevolution of Genome Architecture and Social Behavior

Dustin R. Rubenstein,1,*,@ J. Arvid Ågren,2,† Lucia Carbone,3,4,† Nels C. Eldredge,5,† Hopi E. Hoekstra,2,6,‡ Karen M. Kapheim,7,*,† Laurent Keller,8 Corrie S. Moreau,9,‡ Amy L. Toth,10,‡ Sam Yeaman,11,‡ and Hans A. Hofmann12,*,†

Although social behavior can have a strong genetic component, it can also result in selection on genome structure and function, thereby influencing the evolution of the genome itself. Here we explore the bidirectional links between social behavior and genome architecture by considering variation in social and/or mating behavior among populations (social polymorphisms) and across closely related species. We propose that social behavior can influence genome architecture via associated demographic changes due to social living. We establish guidelines to exploit emerging whole-genome sequences using analytical approaches that examine genome structure and function at different levels (regulatory vs structural variation) from the perspective of both molecular biology and population genetics in an ecological context.

Bidirectional Relationships between Social Behavior and Genome Architecture

Behavioral traits can be difficult to study because their development is often only poorly understood, they can change rapidly in response to the (social or ecological) environment, and they can be flexible in their expression. Social behavior (see Glossary), or the behavioral interactions between two or more individuals typically of the same species, can be particularly challenging to tackle mechanically, not only because it encompasses a wide range of complex behavioral actions that emerge from a suite of less complex correlated displays or phenotypic traits, but also because the fitness of one individual depends on the interactions with others [1]. In addition, social traits often show high levels of behavioral flexibility [2], which we hypothesize can select for genomic characteristics that enhance phenotypic plasticity, such as increasing complexity of gene regulation or an increased reliance on epigenetic mechanisms. Intraspecific variation in social behavior is often associated with the evolution of alternative social states within a species, such as different mating strategies or forms of social organization. These social polymorphisms may themselves be plastic and related to differences in environmental factors [3,4] or fixed throughout an organism’s lifetime and related to features of the genome such as chromosomal inversions [5,6] or gene duplications [7,8]. Social interactions can also generate emergent ‘group-level’ phenotypes, such as dominance hierarchies or caste systems, which in turn can favor gene duplication, neofunctionalization, or the evolution of novel genes [8,9].

Most studies examining the genomic underpinnings of social behavior [10–12] are typically unidirectional, in that they seek to identify genes that underlie behavior, but fail to consider how social behavior might affect genome structure and function and thereby influence the evolution of the genome itself (Figure 1). Importantly, the genome is not simply a sequence of nucleotides or a collection of genes. Rather, the genome has an intricate architecture, including a complex regulatory machinery, mobile elements, and chemical modifications, all of which potentially influence – and can be influenced by – behavioral phenotypes in underappreciated ways. For example, genome evolution (i.e., rates of molecular evolution) in ants has been shown to have been shaped by mu-
ualistic associations with plants [13], illustrating how complex social behavior can influence genome sequence. The ecological conditions underlying social behavior, as well as the unique demographic attributes of social living, may also influence genome architecture via the interplay between natural selection and genetic drift. Operational sex ratio, age structure, and reproductive skew (as a consequence of social organization) can lead to a reduced effective population size \( (N_e) \), which in turn can affect genome architecture via changes in the efficacy of natural selection and the increased chance of genetic drift. For example, high reproductive skew in eusocial insect colonies results in a low \( N_e \), despite often having large numbers of individuals [14]. Demography-related changes in genome architecture may include recombination rate, mutation rate, transposable element (TE), or simple sequence repeat (SSR) accumulation, and/or replication error [15]. In addition to influencing sequence variation, these genomic changes can also affect ploidy levels, gene copy number variation (CNV), chromosomal inversions, and novel and orphan gene formation. All of these processes have been suggested to potentially influence – or be influenced by – social evolution, although a comprehensive framework for the study of the coevolution of social behavior and genome architecture is still lacking.

Here we hypothesize that social behavior can profoundly affect the processes that drive the evolution of genome architecture, mediated through the effects of demographic changes on natural selection via the response to interactions with ecological factors. To fully understand the complex and bidirectional interrelationships between social behavior and genome architecture,
researchers must examine genome structure and function at different levels (i.e., regulatory vs structural variation) and consider how socially and ecologically driven changes in demography shape genome architecture. The rise of ever more powerful genomic tools provides an opportunity to explore and generate new questions and hypotheses linking sociality and the genome. In what follows, we review the literature on these connections, develop a predictive framework for thinking about these bidirectional relationships, and outline avenues of future research about where we expect to see links between intra- and interspecific variation in social behavior and genome features.

Attributes of Genome Architecture
As more whole-genome data from individuals of the same species have become available, it is clear that there is an unexpected amount of individual and population variability in genome structure that is associated with variation in phenotypes. Mutations affecting single base pairs (SNPs) are the most studied form of genetic variation (in either the coding or regulatory regions), although small insertions or deletions (indels), gene CNVs and gene family expansions, and even large chromosomal rearrangements (e.g., inversions) can also commonly be found. Changes in genome architecture can occur through a variety of mechanisms, with different effects depending on the scale and frequency of these events, as numerous processes often work together to dynamically alter the genome over time.

At the most global level, genome size can be highly variable across eukaryotic species, ranging from 2.3 Mbp to 150 giga base pairs (Gbp), mostly driven by the accumulation of repetitive sequences such as TEs or SSRs [16]. Variation in genome size affects numerous cellular properties (e.g., size, proliferation, differentiation) and as a consequence can affect brain [17] and body [18] size. Increases in genome size also are correlated with genomic changes that can influence genome structure, such as decreases in recombination rate and increases in mutation rate [19, 20]. Whether changes in genome size affect the evolution of behavioral traits per se is less clear. However, recent evidence suggests that genome size variation underlies in part the extensive diversity of social structures in Synalpheus snapping shrimps [21].

Many other structural attributes of genome architecture can vary and change within as well as across species, often as a consequence of DNA replication or recombination (Figure 2 and Table 1). It has recently been shown that mutation rates can be elevated in specific genomic loci in ways that can substantially affect evolution under realistic demographic conditions. Specifically, repeated independent deletions of the enhancer of the Pitx1 gene have been causally linked to the loss of pelvic hind fins in stickleback fish (Gasterosteus aculeatus), as different populations of this species adapted to freshwater habitats multiple times independently [22]. Notably, increased DNA fragility in this enhancer is the consequence of SSRs (TG dinucleotide repeats), which increase the frequency of double-strand breaks and deletions during DNA replication [23].

Studies linking genome architecture to behavioral variation directly or indirectly through changes in development and physiology have been successful at identifying the genetic underpinnings of a variety of social characteristics, ranging from alternative reproductive tactics [7] to differences in social organization in eusocial species [5,6] (Figure 1). However, the converse – where mating system and/or social behavior influence the structure of the genome – is also likely to occur (e.g. [24]). Thus, we argue (and discuss in detail below) that a species’ mating and social system can alter its genome architecture through dynamic processes such as recombination, gene duplication, inversion, and TE accumulation as a consequence of effects on demography and subsequent shifts in the efficacy of natural selection relative to genetic drift (Figure 1). Below we highlight three attributes of genome architecture that have been shown to be related

---

**Glossary**

- **Alternative reproductive tactics:** differences in suites of life history, morphological, physiological, and behavioral traits that affect nearly every aspect of their reproductive biology.
- **Chromosomal inversion:** rearrangement in which a segment of a chromosome is reversed end to end.
- **Dosage overcompensation:** adaptive outcome of gene duplication when both copies retain the original function but increase gene product output.
- **Effective population size (N_e):** number of breeding individuals in the population.
- **Epigenetic mechanisms:** processes (e.g., DNA methylation, histone modification, noncoding RNA regulation) that determine chromatin structure, accessibility of genetic loci to transcriptional machinery, and gene expression levels.
- **Gene copy number variation (CNV):** additional copies of DNA sequence and losses of genetic material at an intermediate scale (i.e., DNA segments 1 kb to 5 Mb in length). The number of repeated sections can vary between individuals.
- **Gene duplication:** any duplication of a stretch of DNA that contains a gene.
- **Gene order:** organization of genes within a genome.
- **Genome architecture:** nonrandom structural and spatial attributes of the genome.
- **Genome size:** total amount of DNA contained in one copy of a single genome.
- **Genomic imprinting:** parent-of-origin effect on a gene’s expression during embryo or endosperm formation.
- **Neofunctionalization:** outcome of gene duplication when one paralog assumes a new, and adaptive, function.
- **Novel and orphan genes:** genes lacking detectable orthologs due to de novo origin, duplication and divergence, or horizontal gene transfer.
- **Odorant receptor (OR):** proteins expressed in olfactory receptor neurons that bind odors and are critical for the sense of smell.
- **Operational sex ratio:** the ratio of sexually competing males to sexually competing females that are ready to mate.
- **Phenotypic plasticity:** the ability of a genotype to produce more than one phenotype when exposed to different environments.
- **Ploidy:** number of complete sets of chromosomes in a cell.
bidirectionally to social behavior within (gene CNV, supergenes) or across (gene duplications/ gene family expansions) vertebrate and invertebrate species.

**Supergenes**

Supergenes are clusters of tightly linked loci with related phenotypic effects, characterized by reduced recombination, often also involving a chromosomal inversion event [25]. TEs have been suggested to play a role in generating inversions as well as relocating genes into supergenes [26–28]. One of the best examples linking supergenes to social behavior comes from variation in the number of queens per colony in eusocial insects, which is frequent in ants and is typically associated with differences in a suite of reproductive and life history traits including queen phenotype, queen lifespan, breeding strategy, mode of dispersal, and mode of colony founding [3]. A recent study revealed that variation in social organization in the fire ant *Solenopsis invicta* is controlled by a supergene comprising a large nonrecombining chromosomal region with more than 500 genes in tight linkage [5]. The lack of recombination between the two existing haplotypes allows the two social forms to express fixed differences in several behavioral and life history traits that are essential to social evolution [29], including the mode of dispersal, body size, and chemical odor of individuals. Interestingly, variation in social organization in another ant species, the alpine silver ant *Formica selysi*, also was found to be under the control of a single supergene or ‘social chromosome’ [6], which evolved independently from the supergene in *S. invicta*. In addition to differences in social structure, other social polymorphisms in which multiple traits are linked together also are controlled by supergenes in plants [30], cichlid fishes [31], birds [32–34], and insects [35, 36].

Because supergenes allow the cosegregation of adaptive variation within species, they may facilitate the spread of complex phenotypes across species boundaries and be more common in social species than currently realized [25]. Theoretical analyses suggest that supergenes should readily emerge when there is coevolution between dispersal and social traits (e.g., mediated by variation in physiological or locomotor traits) because there will be selection for more benevolent individuals to preferentially interact with relatives (i.e., disperse less) and self-serving individuals to interact with nonrelatives (i.e., disperse more) [37]. As a result, linkage between loci responsible for dispersal and social behavior should be favored under a wide spectrum of conditions. In line with this prediction, dispersers across a diversity of organisms have been reported to differ in their social tendencies from nondispersers [37].

**Gene Copy Number**

CNVs appear to account for much of the genomic variation within many eukaryotic species [38]. Recent research on a plumage color polymorphism in the seabird the common murre (*Uria aalge*) demonstrated how a much more complex CNV (potentially mediated by TEs) of a 60-kb genomic region acts as a ‘supergene’ that maintains the association between three candidate genes affecting two traits – plumage color and thermal physiology – despite random mating [39]. Although it is not yet known whether this CNV also affects social behavior, it underscores how different structural attributes of the genome can be related to each other. There is, nevertheless, accumulating evidence that CNV may play a role in social behavior, including for humans in terms of social disorders [40] and for species with alternative reproductive tactics, where several individual traits can be functionally integrated to produce divergent reproductive phenotypes (reviewed in [4]). For example, in freshwater *Xiphophorus* swordtail fish, large males are deeper bodied, have higher dorsal fins, exhibit longer swords (extended rays of the caudal fin), and perform stereotyped courtship displays to entice females to mate, whereas small males are slender, have relatively smaller dorsal fins, lack prominent swords, and chase females to force copulation (reviewed in [41]). Lampert et al. [7] showed that variation in male size – and thus in reproductive tactic – in...
Figure 2. Attributes of Genome Architecture That Can Generate Phenotypic Variation across Individuals and Species. How these attributes affect social behavior is described in Table 1. (A) Structural variation in the 3D genome and (B) variation in ploidy are likely to affect all aspects of organismal function broadly. Variation in genome structure limited to specific chromosomes, such as (C) chromosomal inversions (which can result in supergenes) and (E) gene order, may have intermediate effects on trait variability. More local changes as a consequence of gene duplication events – such as (D1) neofunctionalization, (D2) subfunctionalization, (D3) evolution of gene families, and (G) gene copy number variations – or (D4) de novo evolution of genes are likely to have more specific phenotypic effects, depending on the extent of pleiotropy. Altered gene function due to (F) variation in the length of short tandem repeats or (H) random insertion of transposable elements (TEs) may affect phenotypes in more subtle ways. Importantly, all of these attributes can be sources of or be affected by variation in social behavior. Note that the two most common forms of genetic variation, SNPs and small insertions or deletions (indels), are not shown here, as their role in creating phenotypic (including behavioral) variation has been well studied.
X. nigrensis and X. multilineatus is correlated with Y-linked CNV of the ‘B’ allele of the mc4r gene, which encodes a melanocortin receptor subtype that binds α-melanocyte-stimulating hormone (α-MSH). Importantly, while mc4r CNV directly regulates growth and size at maturity, it controls behavioral tactics only indirectly [42].

**Gene Duplication and Gene Family Expansion**

Functional divergence of paralogs after gene duplication events has long been suggested to generate phenotypic variation and novel traits [43]. Possible outcomes of gene duplication include neofunctionalization, subfunctionalization, pseudogenization, and dosage overcompensation. Duplications of genes that encode ion channels have been critical to the evolution of phenotypic diversity in many taxa. A compelling example illustrating how gene duplication can influence the evolution of different social behaviors is the convergent evolution of two sodium
channel paralogs (Scn4aa and Scn4ab) in two divergent lineages of weakly electric fishes that have independently evolved electric organs, whose discharges are used for active sensing and communication [44]. Although Scn4aa is downregulated in the ancestors of both electric lineages, expression of this paralog has shifted from skeletal muscle to a novel communication function in the muscle-derived electric organ in a remarkable example of convergent evolution. Intriguingly, a similar shift in expression away from ancestral muscle cells to an evolutionarily novel muscle-derived sonic organ occurred for the same paralog in the sonic fish Porichthys notatus, in which males advertise their availability to females by characteristic sounds [45].

Another example of how gene duplication can influence social behavior is the Malvolio (Mvl) gene, a member of the natural resistance-associated macrophage protein (Nramp) family, which has been duplicated in Nicrophorus vespilloides, a subsocial beetle that exhibits advanced parental care behavior [46]. Although the two gene copies show differing expression profiles across many tissues, Mvl1 increases in expression during resource preparation and feeding offspring whereas Mvl2 decreases in the same states. This result suggests that following its duplication, Mvl has experienced sub- or neofunctionalization in N. vespilloides in ways relevant to the regulation of parental care behavior.

Mechanisms for recognition, the ability to distinguish friend from foe or kin from nonkin, are critical to most social interactions [47]. In many eusocial insects, for example, recognition is mediated via chemical signals that are detected and discriminated among by chemosensory receptors. The odorant receptor (OR) gene family, which arose soon after the evolution of life on land, encodes proteins that detect odorants with high specificity and sensitivity [48]. In Hymenoptera, eusocial species have several-fold more ORs than solitary ones [49]. In both bees [50] and ants [51], OR gene evolution is the result of lineage-specific gene duplications that underwent independent expansions. In ants in particular, where all of the more than 13,000 species are eusocial, a single lineage of the OR gene family underwent a massive expansion in the ancestor of all extant species [52], resulting in up to nearly 400 putative functional OR genes in some ants [51,53], largely due to tandem array expansion [54]. Similar dynamic expansions and contractions are seen among olfactory receptor gene families in vertebrates [55], although they have not yet been examined in relation to social behavior per se.

**Avenues for Future Research in an Age of Comparative Genomics**

As our ability to sequence large numbers of individuals rapidly increases, researchers will inevitably move beyond studies of transcriptomic variation to those of genomic variation both within and across species. In particular, complete genome sequencing of multiple individuals within a single species, coupled with population-genetic analyses, will enable researchers to identify the genomic architectural polymorphisms that give rise to intraspecific variation in social behavior. At the interspecific level, newly developed phylogenetic comparative methods, along with emerging long-read sequencing technologies, will also allow researchers to identify key architectural features of the genome that are shared among species with similar social or mating systems and to examine not just cooperation but also conflict in the genome (Box 1). Below we discuss three areas ripe for future study.

**Linking Transposable Elements and Regulatory Variation in Genome Architecture in Social Species**

One potentially powerful source of variation in regulatory relationships is the movement of TEs. First described by Barbara McClintock [56] as ‘controlling elements’, TEs are DNA sequences that move from one location in the genome to another and in this way provide a continuous source of genetic innovation. By inserting within genes or their regulatory elements, for example, TEs can directly disrupt gene transcription [57,58]. TEs may become exapted into novel
regulatory elements, thus shaping the evolution of gene regulatory networks [59,60], or spread sequences readily modified by epigenetic marks to neighboring protein-coding genes and repress gene expression [61]. There is also evidence in support of the hypothesis that regulatory features can be derived from TEs (e.g., miRNAs [62,63], transcription factor-binding sites [64,65]), although to date few of these have been explicitly linked to novel behavioral phenotypes, possibly due to an ascertainment bias. One exception comes from a study of ORs in ants. Although TE islands were depauperate in ORs of the clonal raider ant Ooceraea biroi, these ORs were enriched for TEs relative to the rest of the genome, potentially facilitating gene duplication by increasing unequal crossing over, promoting ectopic recombination, or via retrotransposition [54]. Ultimately, understanding how TEs might affect behavioral phenotypes will require high-quality genomes, which are often still lacking for nontraditional model organisms, and studies of their possible effects on gene expression rather than simple descriptions of their abundance, type, and distribution.

**Linking Demography and Structural Variation in Genome Architecture in Social Species**

Intraspecific variation in social behavior may be caused by plasticity and/or genetic variation maintained by balancing selection. In the case of complex (polygenic) behavioral traits, we can hypothesize that selection acts through gene linkage relationships or gene regulation. The tension between recombination and balancing selection on linkage relationships favors spatially integrated architectures, which can arise through chromosomal rearrangements that move behaviorally relevant loci into close proximity on a chromosome or inversions that reduce the rate of recombination [66]. As noted previously, chromosomal inversions can contain numerous genes underlying behavioral, morphological, and physiological traits that are all associated with social polymorphisms. Similarly, we hypothesize that rearrangements that move coexpressed loci into close proximity on a chromosome can be adaptive, if gene expression can be more efficiently or precisely coordinated this way.

If selection favors such rearrangement mutations, the rate at which they accumulate should vary as a function of the effective population size $N_e$, the frequency $\mu$ at which such mutations occur,
and the realized strength $s$ of selection on the rearrangement [15]. Although $N_e$ can be readily estimated from genomic data, the variables $\mu$ and $s$ are often unknown. In the case of selection on linkage, the benefit that accrues from a given rearrangement will be proportional to the reduction in recombination and the alleviation of fitness costs that would have resulted in recombinant offspring [67]. In the case of selection on gene regulation, fitness benefits would depend on the regulatory interactions of the genes involved and resulting changes in energetic efficiency. Any benefits would also be offset by costs associated with the rearrangement due to disruption of existing genes or regulatory machinery. While accumulation rates have been measured for different kinds of rearrangements (e.g., [68]), their rates of occurrence are largely unknown, although they may correlate with overall TE abundance [69]. Thus, we hypothesize that species with either a suitable type of TE or a high abundance of TEs (e.g., as a consequence of reduced $N_e$ in social species, in species where $N_e$ fluctuates over time) will have more flexible genomes and can therefore evolve more readily to optimize gene order in response to selection on linkage or gene expression regulation. Such increased evolvability would, in turn, allow species to more readily adapt to novel or changing environments [70,71]. More generally, we predict tightly linked gene clusters (e.g., superfamilies) to occur in species with: (i) a long history of sociality, where polymorphisms are maintained by strong frequency-dependent selection or migration–selection balance; (ii) abundant and/or mobile TEs; (iii) traits determined by tandemly duplicated (and thus already tightly linked) genes.

**Identifying Mechanisms of Behavioral Plasticity in the Genomes of Social Species**

The opportunities and constraints imposed by social structure are likely to influence the degree of intraspecific variation observed among individuals within populations and the genomic mechanisms underlying this plasticity. Social living can lead to increases in behavioral plasticity by releasing individual group members from certain behavioral and ecological constraints, while also adding a social component to genotype-by-environment (G × E) interactions. For example, reproductive development in many animals is influenced by nutritional feedback on endocrine pathways, which in social insects is typically modulated by pheromones and dominance interactions [72]. In species showing rudimentary forms of eusociality (e.g., many paper wasps and sweat bees), workers are not only totipotent and maintain reproductive capabilities, but also more plastic in that they can dramatically change their roles in society (e.g., switching castes) in response to cues from the social environment [73]. Recent studies suggest that the early stages of eusociality, and the associated increase in individual plasticity, are shaped by genetic variants [11], a high diversity and abundance of TEs, and changes in the regulatory architecture of key gene networks [74,75]. As social complexity increases further, reproductive skew intensifies as fewer and fewer individuals reproduce (due to division of labor). We hypothesize that the resulting small $N_e$ will cause changes in TE content, which in turn will facilitate further evolution of genome structure. It is intriguing in this context that in bumblebees (*Bombus* spp.), which are highly socially plastic, genes experiencing positive selection are predominately expressed by reproductive foundresses, whereas in honeybees, *Apis mellifera*, worker-biased genes are more likely to experience adaptive evolution, suggesting that the loss of queen totipotency may change how selection acts on the social insect genome [76].

**Concluding Remarks**

Social behavior can affect selection regimes and demography, which in turn influence genome evolution in important and underappreciated ways. We have illustrated some of the bidirectional links and feedbacks between social behavior and genome architecture, emphasizing the role that population genetics can play in understanding social evolution. Because social traits often emerge from a suite of less complex correlated behavioral patterns or phenotypic traits [1], genetic relationships underlying behaviors may emerge less frequently than other, simpler traits.
Conversely, because numerous genes contribute to complex social behavior, they may have profound impacts on genome architecture. For example, the accumulation of traits related to social polymorphisms on chromosomal inversions in the form of supergenes may be a common, but underappreciated, mechanism underlying alternative reproductive strategies and driving differences in social structures.

As genomic tools and resources become more readily accessible, opportunities to study the co-evolution of social behavior and genome architecture abound. Moreover, advances in digital evolution approaches provide novel avenues for in silico testing of hypotheses about the causal relationship between sociality and genome structure [77]. However, we caution that a better general understanding of the evolution of genome architecture is necessary. Vertebrate genomes, for example, have been dynamic over evolutionary time, often the result of TE expansion and deletion [78]. Moreover, it remains to be seen what the null expectation is for the evolutionary dynamics associated with genome architecture in social species. If demographic changes related to social behavior influence genome architecture, comparing social and nonsocial species with similar demographic histories might help to isolate the role of social behavior on genome evolution and further define null expectations for future studies. Ultimately, by simultaneously considering both the genomic causes and the consequences of behavioral variation within and across species, we will generate a richer and more integrated view of social evolution (see Outstanding Questions).

Acknowledgments

This work is the result of a symposium and workshop on Social Evolution and Genome Complexity held at Columbia University in March 2017. The event was organized by D.R.R. and H.A.H. with support from Columbia University Seminars and the Research Coordination Network in Sociogenomics, which was funded by the US National Science Foundation (IOS-1256839). We thank Dr W. Wilczynski for his generous support and Nicole Elmer for assistance with the preparation of Figure 2.

References

84. Ågren, J.A. et al. (2014) Mating system shifts and transposable element evolution in the plant genus *Capsella*. BMC Genomics 15, 602