BAC-Based Sequencing of Behaviorally-Relevant Genes in the Prairie Vole

Lisa A. McGraw, Emory University
Jamie K. Davis, Emory University
Pamela J. Thomas, National Human Genome Research Institute
Larry Young, Emory University
James W. Thomas, Emory University

Journal Title: PLoS ONE
Volume: Volume 7, Number 1
Publisher: Public Library of Science | 2012-01-06, Pages e29345-e29345
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1371/journal.pone.0029345
Permanent URL: https://pid.emory.edu/ark:/25593/s5ppz

Final published version: http://dx.doi.org/10.1371/journal.pone.0029345

Copyright information:
This is an open-access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.
This is an Open Access work distributed under the terms of the Creative Commons Universal : Public Domain Dedication License (http://creativecommons.org/publicdomain/zero/1.0/).

Accessed April 16, 2019 8:10 PM EDT
BAC-Based Sequencing of Behaviorally-Relevant Genes in the Prairie Vole

Lisa A. McGraw1,4, Jamie K. Davis2, Pamela J. Thomas3, NISC Comparative Sequencing Program3, Larry J. Young1,4, James W. Thomas2

1 Center for Translational Social Neuroscience and Yerkes National Primate Research Center, Emory University, Atlanta, Georgia, United States of America, 2 Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia, United States of America, 3 Genome Technology Branch and NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America, 4 Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, Georgia, United States of America

Abstract
The prairie vole (Microtus ochrogaster) is an important model organism for the study of social behavior, yet our ability to correlate genes and behavior in this species has been limited due to a lack of genetic and genomic resources. Here we report the BAC-based targeted sequencing of behaviorally-relevant genes and flanking regions in the prairie vole. A total of 6.4 Mb of non-redundant or haplotype-specific sequence assemblies were generated that span the partial or complete sequence of 21 behaviorally-relevant genes as well as an additional 55 flanking genes. Estimates of nucleotide diversity from 13 loci based on alignments of 1.7 Mb of haplotype-specific assemblies revealed an average pairwise heterozygosity \((8.4 \times 10^{-3})\). Comparative analyses of the prairie vole proteins encoded by the behaviorally-relevant genes identified \(>100\) substitutions specific to the prairie vole lineage. Finally, our sequencing data indicate that a duplication of the prairie vole AVPR1A locus likely originated from a recent segmental duplication spanning a minimum of 105 kb. In summary, the results of our study provide the genomic resources necessary for the molecular and genetic characterization of a high-priority set of candidate genes for regulating social behavior in the prairie vole.

Introduction
The prairie vole (Microtus ochrogaster) is a North American microtine rodent that has become a premier animal model for the study of social behavior and has proven useful for discovering gene-brain-behavior relationships [1,2]. Unlike the majority of mammalian species, prairie voles are highly social, often forming lifelong partnerships with their mates (pair bonds) and both parents take part in rearing offspring [3]. Contrary to the socially monogamous prairie vole, other closely related Microtine vole species (i.e., M. montanus and M. pennsylvanicus) are promiscuous, largely asocial, do not form pair bonds and only females contribute to offspring care [4]. The unique differences in the social repertoires of these species has allowed for comparative studies that have led to substantial contributions to our understanding of the neural and molecular circuitry involved in behaviors such as social attachment, parental behavior, addictive behavior, effects of early life experience and social influences on physiological traits [1]. Delineating genomic characteristics that potentially differentiate the social prairie voles from other asocial rodent species may therefore provide important insights as to how the genome contributes to gene expression patterns in the brain and ultimately to both between- and within-species variation in behaviors. Further, the genetic and neurobiological mechanisms discovered to regulate prairie vole social behavior have also been found to contribute to human social cognition (reviewed in [5]).

To date, brain-expressed genes that contribute to the behavioral diversity in prairie voles, DNA sequence resources for studying cis-regulatory and/or transcriptional profiles are available for the arginine vasopressin receptor 1a (Avpr1a) [6], the oxytocin receptor (Oxt) [7], arginine vasopressin (Avp, GenBank Accession Number: DQ269208) and estrogen receptor-\(\alpha\) (Esra, [8]). Thus, detailed genetic and molecular studies focused on behaviorally-relevant genes in the prairie vole, like those that have associated differential distribution of AVPR1A in the brain with affiliative behavior (reviewed in [4]), are currently limited by a lack of gene and genomic sequences for this species. In this study, we selected and fully sequenced BAC clones containing 21 brain-expressed genes falling into five functional classes that are known to or likely to play a role in affiliative behavior (Table 1).

The neurohypophysial peptides, oxytocin (OXT) and vasopressin (AVP) along with their receptors, OXTR and AVPR1A, respectively, have long been known to regulate species-specific social behaviors including pair bonding, parental care, social recognition and aggression by acting within the reward circuitry regions of the brain (reviewed in [2]). Although there is little
difference in the distribution of these receptors between sexes, pharmacological and transgenic manipulations have demonstrated that OXTR within the nucleus accumbens plays an important role in pair bonding in females, while AVPR1A within the ventral pallidum and lateral septum contributes to pair bond formation in males (reviewed in [4]). Like the neurohypophysial peptides and their receptors, the dopaminergic system, acting primarily within the nucleus accumbens, also plays a role in pair bonding in prairie voles. While D2 receptors (DRD2) are essential in the formation of pair bonds in males, D1 receptors (DRD1A) appear to be inhibitory [9,10]. In voles and other species, other genes within the dopaminergic system also contribute to aspects of pair bonding such as learning and memory, parental behavior, sexual behavior, social choice and olfaction (reviewed in [2]).

The hypothalamic-pituitary-adrenal (HPA) axis which plays a prominent role in the stress response has also been implicated in social bond formation in prairie voles. Corticotrophin releasing (CRHR) receptors within the nucleus accumbens facilitate pair bonding in males [11] and when a male loses his partner, CRHR receptors facilitate passive stress-coping behavior much akin to depressive behavior in our own species [12]. The effects are mediated by both CRHR1 and CRHR2 receptors, and the ligands that are potentially involved in this process are CRH, and the urocortins (UCN, UCN2, and UCN3). Sex steroid hormones are also known to contribute to the expression of affiliative behaviors. For example, in prairie voles, social affiliation is influenced by estrogen receptor alpha (ESR1) within the amygdala and the bed nucleus of the striata terminalis and by estrogen receptor beta (ESR2) within the paraventricular nucleus of the hypothalamus [13,14,15]. Finally, while genes involved in synaptic plasticity have not been directly implicated in affiliative behaviors within prairie voles, there is substantial potential for these genes to regulate aspects of social learning based on social experiences [16]. For example, when BDNF is knocked-down in the nucleus accumbens of mice, males can be rescued from developing an aversion to social contact after experiencing long bouts of aggression from another animal [17].

Here, we report the targeted bacterial artificial chromosome (BAC)-based sequencing and accompanying analyses of the 21 behaviorally-relevant genes and flanking regions in the prairie vole.

**Materials and Methods**

**BAC sequencing, assembly and annotation**

Targeted BAC-based sequencing was used to assemble 6.4 Mb of non-redundant or haplotype-specific sequence from 22 chromosomal segments that contain or immediately flank 21 behaviorally-relevant genes (Table 2). In addition to the targeted genes of interest, 55 flanking genes and a single microRNA were at least partially spanned by the sequence assemblies. With the exception of ANKK1, which we predict may be a pseudogene in the prairie vole, and the absence of a prairie vole ortholog of Calm5, the gene order, orientation and content was the same in the prairie vole as that observed in the mouse (data not shown).

Prairie vole BACs from the CHORI-232 library were selected for sequencing based on probe-content and restriction-enzyme fingerprint contigs constructed from clones isolated from the targeted regions of interest [18,19]. When possible, aligned BAC-end sequences were used to select pairs of clones from the autosomal loci that represented the two alternative haplotypes.
present in the library using the strategy described in [20]. Individual BAC clones were either Sanger shotgun sequenced and assembled as described in [21], or pooled and shotgun sequenced using Roche 454 single-end reads. Note that the two clones pooled and sequenced using the Roche 454 platform were from different target regions and that the haplotype-specific assemblies were restricted to the Sanger sequencing of individual clones. Multi-BAC assemblies were generated from clones representing the same haplotype. Genes were annotated primarily based on alignments between mouse cDNAs and the prairie vole genomic sequence, and when available prairie vole cDNAs, using Spidey [22]. The gene annotation is available in the GenBank records listed in Table S1.

Sequence alignments and identification of genetic variation

Genomic sequence assemblies representing alternative haplotypes were aligned with blastz [23] and used as the basis to identify SNPs and indels. Prior to the identification of SNPs, the alignments were masked to exclude low quality sites (phred score <50) as well as simple and low complexity sequence. All the identified SNPs have been deposited in dbSNP. Prairie vole protein coding regions representing the alternative haplotypes were aligned with ClustalX [24] excluding codons with one or more site with a phred score <50. Non-synonymous and synonymous SNPs were identified using PAML [25]. Amino acid sequences were also aligned with ClustalX. Orthologous proteins from other species were downloaded from GenBank or publicly available genome assemblies and are provided in File S1. Amino acid replacements unique to the prairie vole lineage were inferred using simple parsimony and represent a conservative number of changes that occurred in the prairie vole lineage. Radical amino acid substitutions were defined as those that changed at least two out of the three properties for the amino acids outlined in [26], i.e., charge, polarity, and polarity/volume, whereas conservative amino acid substitutions resulted in a change in at most one of those properties.

Results and Discussion

SNP and indel frequency

In order to survey the frequency and type of genetic variation present in the individual from which the BAC library was constructed, we aligned the genomic sequence assemblies derived from BAC clones representing alternative haplotypes (see Table 2 and Table 3). Pair-wise heterozygosity (p) based on single-nucleotide polymorphisms (SNPs) at the 13 sampled loci ranged from 3.6–11.0×10⁻³ with the average being 8.4×10⁻³. Insertions and deletions (indels) polymorphisms were on the order of 5-fold less abundant than the SNPs (Table 3). Similar to what has been observed in other mammals (for example see [27]), the indel length distribution was heavily skewed toward the smaller size range with 1-bp indels being the most common.

<table>
<thead>
<tr>
<th>Target gene(s)</th>
<th>Other genes</th>
<th>Sequence* (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ar***</td>
<td></td>
<td>234</td>
</tr>
<tr>
<td>Avpr1a</td>
<td></td>
<td>140</td>
</tr>
<tr>
<td>Avpr1a truncated</td>
<td></td>
<td>149</td>
</tr>
<tr>
<td>Bdnf</td>
<td></td>
<td>141</td>
</tr>
<tr>
<td>Cfr</td>
<td>Dnaj5b, Trim55**</td>
<td>208 (155)</td>
</tr>
<tr>
<td>Cfr1**</td>
<td>4933407p14rik, Map7, 170008111rik**</td>
<td>150 (161)</td>
</tr>
<tr>
<td>Cfr2</td>
<td>Nod1**, Gcg**, Gar**, Inmt, Fam188b</td>
<td>253</td>
</tr>
<tr>
<td>Drd1A</td>
<td>Sfn1, Hrh2**</td>
<td>206 (182)</td>
</tr>
<tr>
<td>Drd2</td>
<td>Ankk1, Tct12**,Ncam**</td>
<td>282 (156)</td>
</tr>
<tr>
<td>Esr1**</td>
<td>Syne1**</td>
<td>303 (442)</td>
</tr>
<tr>
<td>Esr2**</td>
<td>Syne2**, Tex21**</td>
<td>106 (177)</td>
</tr>
<tr>
<td>Mc4r</td>
<td>-</td>
<td>141</td>
</tr>
<tr>
<td>Nr3c1</td>
<td>Arghap26**</td>
<td>169 (170)</td>
</tr>
<tr>
<td>Nr4k2**</td>
<td>-</td>
<td>166</td>
</tr>
<tr>
<td>Oxt, Avp</td>
<td>Ppina, Mrpi26, Ubox5, FastkdS</td>
<td>160 (145)</td>
</tr>
<tr>
<td>Otot</td>
<td>Lmc11**, D630042p16nik, Cav3**, Rad18</td>
<td>176 (309)</td>
</tr>
<tr>
<td>Slc6a2</td>
<td>Lpcat2**, Capns2**, Gm4976, Es1, Ces3**</td>
<td>213 (137)</td>
</tr>
<tr>
<td>Slc6a3</td>
<td>Lpcat1**, Clptm11**, Ter**, Slc6a18, Slc6a19</td>
<td>114 (148)</td>
</tr>
<tr>
<td>Th</td>
<td>-</td>
<td>86</td>
</tr>
<tr>
<td>Ucn</td>
<td>Slc5a6, 0610007c21rik, Cad, Slc3oa3, Dnaj5c***, Trim54</td>
<td>154</td>
</tr>
<tr>
<td>Ucn2***</td>
<td>Ip6k2, Nckipsd, Cels3, Slc26a6, Tnmem89, Uqrc1, Col7a1***, Mir711</td>
<td>155 (154)</td>
</tr>
<tr>
<td>Ucn3</td>
<td>Calm3, Calm4, Net1, Tubal3**</td>
<td>158 (178)</td>
</tr>
</tbody>
</table>

*Size of the alternative haplotype assembly is shown in parentheses.
**Partial sequence.
***Sequence assembly is adjacent to the target gene and does not include the target gene itself.

doi:10.1371/journal.pone.0029345.t002

Table 2. Summary of BAC sequencing.
Table 3. Nucleotide diversity between alternative haplotypes.

<table>
<thead>
<tr>
<th>Region*</th>
<th>Aligned sites**</th>
<th>SNPs</th>
<th>$\pi (\times 10^{-8})$</th>
<th>Indels</th>
<th>Indel size range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cdh</td>
<td>138,103</td>
<td>1,518</td>
<td>11.0</td>
<td>278</td>
<td>1–1,054 bp</td>
</tr>
<tr>
<td>Cgr1</td>
<td>72,493</td>
<td>604</td>
<td>8.3</td>
<td>108</td>
<td>1–3,066 bp</td>
</tr>
<tr>
<td>Drd1A</td>
<td>117,457</td>
<td>1,156</td>
<td>9.8</td>
<td>235</td>
<td>1–773 bp</td>
</tr>
<tr>
<td>Drd2</td>
<td>132,468</td>
<td>1,313</td>
<td>9.9</td>
<td>192</td>
<td>1–6,070 bp</td>
</tr>
<tr>
<td>Esl1</td>
<td>265,109</td>
<td>2,665</td>
<td>10.1</td>
<td>503</td>
<td>1–13,667 bp</td>
</tr>
<tr>
<td>Esa2</td>
<td>56,409</td>
<td>476</td>
<td>8.4</td>
<td>100</td>
<td>1–1,069 bp</td>
</tr>
<tr>
<td>Nrc3c1</td>
<td>159,447</td>
<td>934</td>
<td>5.9</td>
<td>219</td>
<td>1–680 bp</td>
</tr>
<tr>
<td>Ost, Avp</td>
<td>138,034</td>
<td>931</td>
<td>6.7</td>
<td>271</td>
<td>1–1,091 bp</td>
</tr>
<tr>
<td>Ostr</td>
<td>151,640</td>
<td>1,353</td>
<td>8.9</td>
<td>256</td>
<td>1–1,241 bp</td>
</tr>
<tr>
<td>Slc6a2</td>
<td>127,323</td>
<td>1,288</td>
<td>10.1</td>
<td>206</td>
<td>1–4,188 bp</td>
</tr>
<tr>
<td>Slc6a3</td>
<td>100,371</td>
<td>745</td>
<td>7.4</td>
<td>142</td>
<td>1–1,681 bp</td>
</tr>
<tr>
<td>Ucn2</td>
<td>144,027</td>
<td>518</td>
<td>3.6</td>
<td>119</td>
<td>1–338 bp</td>
</tr>
<tr>
<td>Ucn3</td>
<td>145,967</td>
<td>1,198</td>
<td>8.2</td>
<td>231</td>
<td>1–1,776 bp</td>
</tr>
<tr>
<td>Total</td>
<td>1,748,848</td>
<td>14,699</td>
<td>8.4</td>
<td>2,860</td>
<td>1–13,667 bp</td>
</tr>
</tbody>
</table>

$\pi$ represents the pair-wise heterozygosity (substitutions/site) between the alternative haplotypes.

*Regions are referred to by the target gene of interest.

**Represents aligned sites after excluding simple and low complexity sequence, and sites with low sequence quality.

doi:10.1371/journal.pone.0029345.t003

in [28]). Thus, though sampling bias, both in term of the individuals included in a study and position in the genome, can have a strong effect on estimates of $\pi$, it is nonetheless of interest to compare the estimate of $\pi$ we observed in the prairie vole to those reported for other mammals. For example, in three other rodents [the field vole (Microtus agrestis), wild mouse (Mus musculus), and deer mouse (Peromyscus maniculatus)] sequence-based estimates of $\pi$ for the nuclear genome were reported to be $0.8 \times 10^{-8}$, $1.3 \times 10^{-8}$, and $2.9 \times 10^{-8}$, respectively [29,30,31]. The values of $\pi$ we observed in a single prairie vole of $3.6 \times 10^{-8}$–$1.1 \times 10^{-8}$, average $= 8.4 \times 10^{-9}$, are therefore within the range previously observed in rodents, but are higher than the nucleotide diversity observed in other mammals such as the panda ($1.3 \times 10^{-8}$) [32], chimpanzee ($0.8 \times 10^{-9}$–$1.9 \times 10^{-9}$) [33], and humans ($0.6 \times 10^{-9}$–$0.9 \times 10^{-9}$) [34] and references therein). Future studies estimating the genetic diversity in prairie voles based on multiple individuals and additional loci will be needed to determine if the level of nucleotide diversity observed in this study is truly representative of the species.

Intra- and interspecific gene and amino acid sequence comparisons

The protein coding region of the 43 prairie vole genes sequenced on both haplotypes were aligned to identify synonymous and nonsynonymous SNPs. In total we identified 201 synonymous ($dS = 9.9 \times 10^{-9}$) and 75 nonsynonymous ($dN = 1.5 \times 10^{-9}$) SNPs between the two haplotypes sampled at each locus. Within the 12 behaviorally-relevant genes that were sequenced on both haplotypes there were 39 synonymous and 6 nonsynonymous SNPs. No SNPs were observed in Ostr or Ucn3, synonymous but no nonsynonymous SNPs were present in seven genes (Ded1a, Esl1, Evi2, Nrc3c1, Ost, Slc6a2, Slc6a3), and both synonymous and nonsynonymous SNPs were observed in three genes (Aqp, Cgr1 And Dru2). To evaluate the potential functional consequence of the nonsynonymous SNPs in Aqp, Cgr1 And Dru2, and to identify amino acid replacements that were specific to the prairie vole lineage in all of the behaviorally-relevant genes, we aligned the predicted prairie vole protein sequences to orthologous proteins from other rodents: mouse (Mus musculus), rat (Rattus norvegicus), and guinea pig (Cavia porcellus), as well as rabbit (Oryctolagus cuniculus) (see Methods and File S1).

The proteins encoded by the behaviorally-relevant prairie vole genes ($n = 21$) were on average 99% identical (range of 94%–99%) to their mouse/rat orthologs. A total of 127 unique amino acid replacements in these proteins could be assigned by parsimony to the prairie vole lineage, of which 32 were classified as radical substitutions (Fig. 1 and Table S2). The potential functional impact of the nonsynonymous SNPs in Aqp, Cgr1 And Dru2 was predicted using evolutionary conservation using the program SIFT [35]. Based on this metric three of the nonsynonymous changes were predicted to affect protein function while the remaining changes were predicted to be tolerated (Table S3).

The prairie vole is considered a valuable rodent model for social behavior due to the phenotypes observed in this species that are uncommon in other rodents, such as pair-bonding [1]. While differential distribution of AVPR1A in the brain has been correlated with variation in social behavior [36], lineage-specific changes that alter the regulation or proteins of other genes related to social behavior that distinguish the prairie vole from other rodents may also be functionally relevant. We therefore consider the >100 amino acid changes in proteins relevant to behavior we identified in the prairie vole lineage candidates for altering the activity of these proteins. However, since the prairie vole lineage has been evolving independently from the other rodent lineages for at least 25 million years, we anticipate that most of these lineage specific changes may have accumulated by chance and will not be functionally relevant. Future comparative studies will be needed to determine if in fact the prairie vole proteins do exhibit any differences in activity compared to other rodents and which specific changes are responsible for such functional alterations.

Segmental duplication of the Avpr1a locus

Previous cloning and sequencing efforts of the prairie vole Aqp1a gene detected the presence of a duplicate copy that encoded a truncated protein [6]. To gain further insight into this duplication, we sequenced BAC clones containing either the functional and truncated prairie vole Aqp1a loci (Table 1). Alignment of the resulting sequences revealed a duplication of $\geq 105$ kb spanning the Aqp1a loci and flanking regions. The

Figure 1. Amino acid substitutions in the prairie vole lineage.

The evolutionary relationship of the prairie vole to other rodents and rabbit is illustrated as a phylogenetic tree. Divergence times are represented by the branch lengths of the tree based on [39,40]. The numbers above the terminal branch leading to the prairie vole represent the number of conservative/radical amino acid substitutions in the behaviorally-relevant proteins that were inferred by parsimony to have occurred in that lineage and unique to the prairie vole. MYA refers to millions of years ago.

doi:10.1371/journal.pone.0029345.g001
differences between the duphicons was 0.0177+/-0.0004 substitutions/site (87,614 sites, Kimura 2-parameter distance [37]), suggesting the duplication likely occurred relatively recently. As was reported previously [6], the truncated Avpr1a locus included a ~700 bp indel upstream of the gene and frame-shift mutations within the processed region (c.597delCC, c.927delCT, c.927insCC, and c.890_894delGTGTCAGCAGC, where the positions refer to the protein coding sequence for the prairie vole Avpr1a annotated in GenBank Ac# AF069304).

A previous study reported that the Avpr1a locus was duplicated in the prairie vole but not in the montane vole (Microtus montanus) [6]. The low divergence between the duplicated Avpr1a loci we observed in this study and the size of the duplicated region (2230 bp) is therefore consistent with a recent segmental duplication of this region having occurred in the prairie vole lineage. The frameshift mutations in the truncated copy of Avpr1a suggests that it is now a pseudogene, which is a common evolutionary fate for newly duplicated genes [38]. Characterization of Avpr1a in additional species will be needed to better reconstruct the history of this duplication and the phylogenetic distribution and fate of the duplicated copy of this gene in other voles.

Conclusions

The ability to study genes and their molecular and genetic correlates with behavior is dependent in part on the availability of genetic and sequence resources. In this study we have generated genomic sequence, the predicted cDNA and protein sequences for Avpr1a in a number of closely related species. Combined, these data can be used as a starting platform for future studies focused on characterizing the role of these genes in behavioral phenotypes in the prairie vole, such as genetic association studies, quantification of gene transcript levels and expression patterns, as well as scans for cis-regulatory elements. In addition, our results provided novel information as to the genetic diversity within the prairie vole and candidate lineage-specific changes to a number of behaviorally-relevant proteins.

Supporting Information

File S1 Sequences used in the analyses of the prairie vole proteins.

Table S1 GenBank accession numbers for assembled and annotated prairie vole sequences.

Table S2 Amino acid substitutions specific to the prairie vole lineage.

Table S3 Nonsynonymous variants identified in the behaviorally-relevant prairie vole proteins.

Acknowledgments

The authors wish to acknowledge Greg K. Tharp for computational support, the Georgia Research Alliance Genomics Core for the 454 sequencing, and members of the NIH Intramural Sequencing Center, including E. D. Green, R. Blakesley, G. Bouffard, J. Mullikin, and J. McDowell.

Author Contributions

Conceived and designed the experiments: LAM LJY JWT. Performed the experiments: JKD NISC. Analyzed the data: LAM PJT JWT. Wrote the paper: LAM JWT.

References