Lovell *et al. BMC Res Notes (2018) 11:309*
https://doi.org/10.1186/s13104-018-3402-x

BMC Research Notes

DATA NOTE

Open Access

Curation of microarray oligonucleotides and corresponding ESTs/cDNAs used for gene expression analysis in zebra finches

Peter V. Lovell[1], Nicole A. Huizinga[1], Abel Getachew[1], Brianna Mees[1], Samantha R. Friedrich[1], Morgan Wirthlin[1,2] and Claudio V. Mello[1*]

Abstract

Objectives: Zebra finches are a major model organism for investigating mechanisms of vocal learning, a trait that enables spoken language in humans. The development of cDNA collections with expressed sequence tags (ESTs) and microarrays has allowed for extensive molecular characterizations of circuitry underlying vocal learning and production. However, poor database curation can lead to errors in transcriptome and bioinformatics analyses, limiting the impact of these resources. Here we used genomic alignments and synteny analysis for orthology verification to curate and reannotate ~ 35% of the oligonucleotides and corresponding ESTs/cDNAs that make-up Agilent microarrays for gene expression analysis in finches.

Data description: We found that: (1) 5475 out of 43,084 oligos (a) failed to align to the zebra finch genome, (b) aligned to multiple loci, or (c) aligned to Chr_un only, and thus need to be flagged until a better genome assembly is available, or (d) reflect cloning artifacts; (2) Out of 9635 valid oligos examined further, 3120 were incorrectly named, including 1533 with no known orthologs; and (3) 2635 oligos required name update. The resulting curated dataset provides a reference for correcting gene identification errors in previous finch microarrays studies, and avoiding such errors in future studies.

Keywords: Molecular, Speech and language, Birdsong, cDNA microarray, Oligo array, Gene expression, Brain, Vocal learning

Objective

Zebra finches represent a major model organism for studying vocal learning [1–6], a trait that provides a basis for spoken language acquisition in humans. Studies in finches have led to insights into the molecular machinery that underlies learned vocalizations [7–19], including the transcriptome of the vocal control circuitry [7, 8, 11–16, 18–25] and the identification of convergent molecular specializations of the vocal control systems of birds and humans [7]. Such studies were largely based on the Songbird array v2 [16], a ~ 44,000 60-mer oligonucleotide array designed with eArray 5.4 (Agilent Technologies) and sequences from three cDNA collections [11, 16, 23]. Initial cDNA annotations were made before the zebra finch genome was available through BLAST searches of annotated cDNA/EST databases. Later efforts aligned oligo and EST sequences to the zebra finch genome (Taegut1; [26]), and assigned Ensembl model annotations to oligos that mapped to within 5 kb (or ESTs within 3 kb) of those models [7, 25]. However, this effort did not take into account strand information, did not detect ESTs/oligos intronic to gene models, or assigned ESTs/oligos to models that were incorrectly annotated. Other oligos were derived from cDNA cloning artifacts, or erroneous sequence selection. By removing and correcting these errors, we generated what we consider the most thorough and accurate constitutive transcriptome of the

*Correspondence: melloc@ohsu.edu
[1] Department of Behavioral Neuroscience, OHSU, Portland, OR 97221, USA
Full list of author information is available at the end of the article

zebra finch song control system [27]. We describe this curation effort below.

Data description

We retrieved the full set of oligos (60-mers) from the Agilent-021323 Zebra Finch Oligoarray (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL18442), removing redundancies and controls. For 43,084 non-redundant oligos we applied a similar curation effort as described in [9, 28, 29]. Table 1 provides links to a summary of our curation effort (Table 2), and relevant datasets (Tables 3–13). The complete collection of datasets can be found at https://doi.org/10.6084/m9.figshare.c.4081835 [30]. We first aligned all oligos to the finch genome (Taegut1) using BLAT [31] with stringent parameters (minScore = 30; minIdentity = 0). 2792 oligos (6%) failed to align to Taegut1 (alignment score < 25; Table 3), 503 (1%) only aligned to Chr_Un (i.e., chromosome unknown; Table 4), a concatenation of unassembled regions and allelic variants, and 1952 (5%) aligned to multiple loci on different chromosomes (Table 5). All cases above were removed from further analyses, as one cannot determine specificity or establish gene orthology based on synteny. We retained oligos with high scoring (> 95%) secondary alignments to Chr_Un, since these correspond to allelic variants. Another 228 oligos (< 1%) were in opposite orientation to ESTs sequenced from the 5′ end of the cDNAs, or in the same orientation as ESTs sequenced from the 3′ end (Table 6). These were also removed as they represent antisense strands of short ESTs with T-stretches at both ends due to second-strand oligo-dT priming and non-directional cDNA cloning.

For 27,974 out of the 37,609 oligos (74%) that passed initial filters we provide the consensus gene symbol as in previous efforts [7, 32] (Table 7), based on the Human Genome Nomenclature Consortium (HGNC; 2018). For the remaining 9635 oligos (26%) that define the constitutive transcriptome of the finch song system [27], we inspected alignments against Taegut1, and annotated sequences based on association with a gene model (Ensembl or finch-/xeno-RefSeqs on the correct strand). For ESTs corresponding to 3′-UTRs, we BLAT-aligned sequences to chicken (*Galgal5*) to try to connect them to chicken gene models by 'walking' the extensive chicken ESTs/mRNAs collection. In total, 3750 oligos annotations were confirmed by direct inspection (Table 8), and an additional 130 oligos further confirmed by synteny (Table 9), which required additional alignments and neighbor gene comparisons with other avian (e.g. chicken, Tibetan tit, other finches, budgerigar, starling, falcon) and non-avian (i.e., alligator, lizard, mouse, human) genomes. We provided correct annotations for 1529 unannotated or misannotated oligos (Table 10), including cases of improper Ensembl model assignment (e.g. wrong strand) or intronic location to a model, determining orthology for another 58 oligos (Table 11). 1533 oligos associated with loci with no orthologs in other organisms (Table 12) were named unknown. Lastly, we updated 2635 oligos to an HGNC symbol, or a consensus NCBI:Gene name (Table 13).

Our findings highlight the need for accurate curations to avoid propagating errors in gene identification and bioinformatics. This partial curated dataset (~ 35% of oligos on this array) serves as a reference for correcting errors from previous studies, and a roadmap for future oligo curations. We anticipate for the 27,975 oligos not

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Table 2	MS Excel file (.xlsx)	https://doi.org/10.6084/m9.figshare.6189485
Data file 1	Table 3	MS Excel file (.xlsx)	https://doi.org/10.6084/m9.figshare.6189482
Data file 1	Table 4	MS Excel file (.xlsx)	https://doi.org/10.6084/m9.figshare.6189479
Data file 1	Table 5	MS Excel file (.xlsx)	https://doi.org/10.6084/m9.figshare.6189476
Data file 1	Table 6	MS Excel file (.xlsx)	https://doi.org/10.6084/m9.figshare.6189470
Data file 1	Table 7	MS Excel file (.xlsx)	https://doi.org/10.6084/m9.figshare.6189467
Data file 1	Table 8	MS Excel file (.xlsx)	https://doi.org/10.6084/m9.figshare.6189461
Data file 1	Table 9	MS Excel file (.xlsx)	https://doi.org/10.6084/m9.figshare.6189452
Data file 1	Table 10	MS Excel file (.xlsx)	https://doi.org/10.6084/m9.figshare.6189446
Data file 1	Table 11	MS Excel file (.xlsx)	https://doi.org/10.6084/m9.figshare.6189440
Data file 1	Table 12	MS Excel file (.xlsx)	https://doi.org/10.6084/m9.figshare.6189437
Data file 1	Table 13	MS Excel file (.xlsx)	https://doi.org/10.6084/m9.figshare.6189431

examined here, 32% will require further curation, and 27% will require updated gene symbols (Table 7).

Limitations

- In our experience, accurate orthology assignment requires synteny verification, however there are no adequate computational methods for large scale analyses, and manual assessment of a large gene set is beyond a reasonable scope of effort. We recommend that caution should be exerted and direct synteny verification be applied whenever deciding to focus on one or a few genes from microarray screenings. This is particularly important in cases of suspected paralogy or sequence cross-alignments to close family members.
- The HGNC annotation step is important since most bioinformatics pipelines use these approved symbols. Here we downloaded the entire set of HGNC gene symbols along with any older gene symbols or synonyms and cross-referenced the lists to verify that the gene symbols of our curated oligo sets were approved terms by HGNC. In most cases, we were able to update older gene symbols or synonyms to a current HGNC gene symbol. In some cases, however, particularly when the zebra finch gene does not have a human ortholog, there is no approved HGNC gene symbol. In these cases, we consulted NCBI:Gene and assigned the gene symbol most commonly shared amongst multiple non-human vertebrates (e.g. mouse, anole lizard, chicken, frog). These NCBI:Gene names are listed as 'Not Approved' under the column heading "HGNC Symbol Status" in Tables 7–13 and are not valid entries for bioinformatics applications based on approved human gene terms.

Abbreviations

BLAT: BLAST-like alignment tool; BLAST: Basic Local Alignment Search Tool; cDNA: DNA synthesized from a single stranded mRNA; chr_Un: chromosome unknown; EST: expressed sequence tags; HGNC: Human Genome Nomenclature Consortium; NCBI: National Center for Biotechnology Information; Xeno-/Refseqs: annotated and curated nucleotide sequences (DNA, RNA).

Authors' contributions

PVL, CVM, MW, SRF, NAH, BM: Initial conceptual and experimental design of the study. NAH, AG, PVL, BM, SRF, CVM, MW: Analysis of microarray data, analysis of genomic alignments to determine probe specificity, curation and reannotation of oligonucleotides. PVL, CVM, NAH, BM, MW, SRF, and NH: Interpretation of data, key discussions on principal findings, and preparation of tables, figures and their legends. PVL and CVM: Presentation of tables and manuscript writing. All authors read and approved the final manuscript.

Author details

[1] Department of Behavioral Neuroscience, OHSU, Portland, OR 97221, USA. [2] Present Address: Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data materials

The datasets generated during and/or analyzed during the current study are available in the figshare repository: https://doi.org/10.6084/m9.figshare.c.4081835.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

This research was supported by the National Institutes of Health (R24_GM120464 and R21_DC014432) and by the National Science Foundation (NSF-143602). These funding bodies had no role in the design of the study, collection, analysis, and interpretation of data, or in writing the manuscript.

Publisher's Note

Received: 26 March 2018 Accepted: 7 May 2018
Published online: 18 May 2018

References

1. Wilbrecht L, Nottebohm F. Vocal learning in birds and humans. Ment Retard Dev Disabil Res Rev. 2003;9(3):135–48.
2. Doupe AJ, Kuhl PK. Birdsong and human speech: common themes and mechanisms. Annu Rev Neurosci. 1999;22:567–631.
3. Zeigler HP, Marler P, editors. Behavioral neurobiology of birdsong: annals of the New York academy of science. New York: New York academy of science; 2004.
4. Zeigler HP, Marler P, editors. Neuroscience of birdsong. Cambridge: Cambridge University Press; 2008.
5. Bolhuis JJ, Everaert M. Birdsong, speech, and language. In: Bolhuis JJ, Everaert M, editors. Exploring the evolution of mind and brain. Cambridge: MIT Press; 2013. p. 542.
6. Mello CV. The zebra finch, Taeniopygia guttata: an avian model for investigating the neurobiological basis of vocal learning. Cold Spring Harb Protoc. 2014;2014(12):1237–42.
7. Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, Roulhac PL, Howard JT, Wirthlin M, Lovell PV, Ganapathy G, et al. Convergent transcriptional specializations in the brains of humans and song-learning birds. Science. 2014;346(6215):1256846.
8. Lovell PV, Clayton DF, Replogle KL, Mello CV. Birdsong "transcriptomics": neurochemical specializations of the oscine song system. PLoS ONE. 2008;3(10):e3440.
9. Lovell PV, Carleton JB, Mello CV. Genomics analysis of potassium channel genes in songbirds reveals molecular specializations of brain circuits for the maintenance and production of learned vocalizations. BMC Genomics. 2013;14(1):470.
10. Wada K, Sakaguchi H, Jarvis ED, Hagiwara M. Differential expression of glutamate receptors in avian neural pathways for learned vocalization. J Comp Neurol. 2004;476(1):44–64.
11. Li X, Wang XJ, Tannenhauser J, Podell S, Mukherjee P, Hertel M, Biane J, Masuda S, Nottebohm F, Gaasterland T. Genomic resources for songbird research and their use in characterizing gene expression during brain development. Proc Natl Acad Sci USA. 2007;104(16):6834–9.
12. Thompson CK, Meitzen J, Replogle K, Drnevich J, Lent KL, Wissman AM, Farin FM, Bammler TK, Beyer RP, Clayton DF, et al. Seasonal changes in patterns of gene expression in avian song control brain regions. PLoS ONE. 2012;7(4):e35119.

13. Stevenson TJ, Replogle K, Drnevich J, Clayton DF, Ball GF. High throughput analysis reveals dissociable gene expression profiles in two independent neural systems involved in the regulation of social behavior. BMC Neurosci. 2012;13:126.
14. Hilliard AT, Miller JE, Horvath S, White SA. Distinct neurogenomic states in basal ganglia subregions relate differently to singing behavior in songbirds. PLoS Comput Biol. 2012;8(11):e1002773.
15. Hilliard AT, Miller JE, Fraley ER, Horvath S, White SA. Molecular microcircuitry underlies functional specification in a basal ganglia circuit dedicated to vocal learning. Neuron. 2012;73(3):537–52.
16. Wada K, Howard JT, McConnell P, Whitney O, Lints T, Rivas MV, Horita H, Patterson MA, White SA, Scharff C, et al. A molecular neuroethological approach for identifying and characterizing a cascade of behaviorally regulated genes. Proc Natl Acad Sci USA. 2006;103(41):15212–7.
17. Hara E, Rivas MV, Ward JM, Okanoya K, Jarvis ED. Convergent differential regulation of parvalbumin in the brains of vocal learners. PLoS ONE. 2012;7:e29457.
18. Kubikova L, Wada K, Jarvis ED. Dopamine receptors in a songbird brain. J Comp Neurol. 2010;518(6):741–69.
19. Kato M, Okanoya K. Molecular characterization of the song control nucleus HVC in Bengalese finch brain. Brain Res. 2010;1360:56–76.
20. Lombardino AJ, Hertel M, Li XC, Haripal B, Martin-Harris L, Pariser E, Nottebohm F. Expression profiling of intermingled long-range projection neurons harvested by laser capture microdissection. J Neurosci Methods. 2006;157(2):195–207.
21. Dong S, Replogle KL, Hasadsri L, Imai BS, Yau PM, Rodriguez-Zas S, Southey BR, Sweedler JV, Clayton DF. Discrete molecular states in the brain accompany changing responses to a vocal signal. Proc Natl Acad Sci USA. 2009;106(27):11364–9.
22. London SE, Dong S, Replogle K, Clayton DF. Developmental shifts in gene expression in the auditory forebrain during the sensitive period for song learning. Dev Neurobiol. 2009;69(7):437–50.
23. Replogle KL, Arnold AP, Ball GF, Band M, Bensch S, Brenowitz EA, Dong S, Drnevich J, Ferris M, George JM, et al. The Songbird Neurogenomics (SoNG) Initiative: community-based tools and strategies for study of brain gene function and evolution. BMC Genomics. 2008;9(1):131.
24. Drnevich J, Replogle KL, Lovell P, Hahn TP, Johnson F, Mast TG, Nordeen E, Nordeen K, Strand C, London SE, et al. Impact of experience-dependent and -independent factors on gene expression in songbird brain. Proc Natl Acad Sci USA. 2012;109(Suppl 2):17245–52.
25. Whitney O, Johnson F. Motor-induced transcription but sensory-regulated translation of ZENK in socially interactive songbirds. J Neurobiol. 2005;65(3):251–9.
26. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S, et al. The genome of a songbird. Nature. 2010;464(7289):757–62.
27. Lovell PV, Huizinga NA, Friedrich SR, Wirthlin M, Mello CV. The constitutive differential transcriptome of a brain circuit for vocal learning. BMC Genomics. 2018;19(1):231.
28. Wirthlin M, Lovell PV, Jarvis ED, Mello CV. Comparative genomics reveals molecular features unique to the songbird lineage. BMC Genomics. 1082;2014:15.
29. Lovell PV, Wirthlin M, Wilhelm L, Minx P, Lazar NH, Carbone L, Warren WC, Mello CV. Conserved syntenic clusters of protein coding genes are missing in birds. Genome Biol. 2014;15(12):565.
30. Lovell PV, Huizinga NA, Friedrich SR, Wirthlin M, Mello CV. The constitutive differential transcriptome of a brain circuit for vocal learning. figshare. https://doi.org/10.6084/m9.figshare.c.4081835.
31. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
32. Whitney O, Pfenning AR, Howard JT, Blatti CA, Liu F, Ward JM, Wang R, Audet JN, Kellis M, Mukherjee S, et al. Core and region-enriched networks of behaviorally regulated genes and the singing genome. Science. 2014;346:1256780.