Copy Number Variation and Adaptive Evolutionary Radiations across the African Cichlid phylogeny

Karl Menzel
Biology Department
Reed College
Portland, OR
menzelk@reed.edu

Suzy C. P. Renn
Biology Department
Reed College
Portland, OR
srenn@reed.edu

Anna Ritz
Biology Department
Reed College
Portland, OR
aritz@reed.edu

Keywords

Copy Number Variation; Phylogeny; African cichlid

1. INTRODUCTION

There are wide ranges of variation between rates of divergence; some lineages diverge very quickly while others do not. When rapid divergence leads to speciation events that fill multiple niches, it is termed *adaptive radiation*. Examples of this include the Galapagos finches [2], the African cichlids [3], and at a broader level the Cambrian radiation [4]. Gene duplication and other genome rearrangement events have been proposed to contribute to (or result from) these explosions of diversification. Copy number variation is associated with diversification by changing gene dosage effects, creating novel gene functions, or by allowing for loss of part of the function of the ancestral gene [6].

African cichlids lend themselves towards research in rapid divergence because these fish consist of closely related clades, some of which have undergone adaptive radiation while others have not. This allows for a large number species to be compared across multiple radiation events. Looking at copy number variation and maintenance of this variation across many cichlid species will allow us begin to infer where and when genome rearrangement events occurred in the evolutionary tree and what effect they might have had. This will increase our understanding of copy number patterns in cichlids as well as other rapidly diversifying species.

Previous studies of adaptive radiation in cichlids have used arrays [5] as well as whole genome sequencing [1] to study copy number variation. Both approaches focused on five species within the cichlid lineage. High depth sequencing is expensive which makes it difficult to sequence a large number of individuals. Array-based experiments are cheaper and more feasible for studies with many individuals. However, specific arrays may not be available for non-model organisms such as cichlids.

BCB '16, October 2–5, 2016, Seattle, WA, USA Copyright 2016 ACM 978-1-4503-4225-4/16/10.
http://dx.doi.org/10.1145/2975167.2985656

In this genome-wide study, we quantify copy number variation across a broad phylogenetic range of 50 African cichlids. Specifically, we count genes whose copy number status ("absent","present", or "present more than once") have been maintained in lineages. We use array Comparative Genomic Hybridization (aCGH) to infer copy number status for each of the species as well as shared status among related clades.

A major challenge involved with inferring copy number status in such a large study is that cichlid genomes are not well-annotated. *Oreochromis niloticus* (Nile tilapia) is the best annotated cichlid genome, yet the *O. niloticus* assembly remains organized into linkage groups rather than chromosomes. Four other cichlid genomes have been sequenced, but they are sparsely assembled and are missing linkage group information. Genomic resources for the remaining cichlid species (>2000) are virtually non-existent. We make three contributions that address this challenge, making a study of this scale feasible.

1. We designed an oligonucleotide array using a consensus of the five sequenced cichlid genomes, making it more robust to species-specific sequence divergence than arrays designed using only one species.
2. We performed a focused study with multiple hybridizations of the five sequenced species using two species as references for comparison.
3. We used the customized array to competitively hybridize 50 species against *A. burtoni*. These species were selected to represent multiple adaptive radiations as well as lineages that have undergone little speciation.

2. METHODS

Array design.

The oligonucleotide array used in these experiments was specifically designed to be robust to species-specific sequence divergence. It was designed using a consensus genome sequence that was biased toward the genome of *O. niloticus*, the cichlid with the most complete genome assembly, while considering sequence similarity with the genomes available from the four other sequenced cichlids. Three probes for each annotated gene as well as intergenic probes were selected from highly conserved sequences across the majority of the five sequenced genomes. The resulting array contains $m = 123,068$ total probes: $72,084$ probes representing $20,614$ highly-conserved *O. niloticus* genes and $50,984$ probes evenly-spaced within intergenic regions.

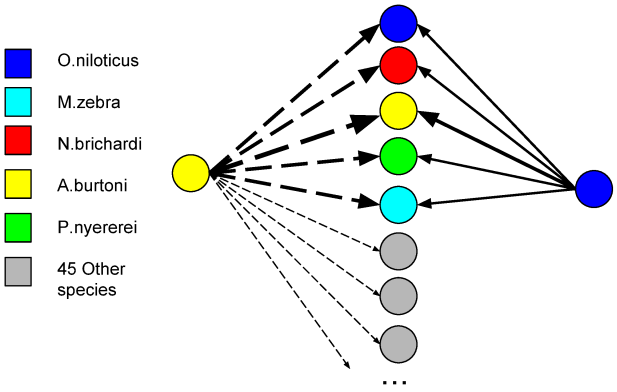

Figure 1: Hybridizations performed in this study. Circles represent species and directed edges denote reference-test pairs for the aCGH arrays. Edge thickness corresponds to number of replicates.

Inferring Copy Number Status.

Consider a set of arrays $A = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ corresponding to n hybridizations. Each array $\mathbf{a}_i$ is a vector of length m denoting the $\log_2$ fold ratio of test-to-reference DNA for each probe. The arrays are independently normalized and segmented using the `DNAcopy` and `snapCGH` R modules to produce "smoothed" copy number profiles $\tilde{A}$. Given the copy number profiles $\tilde{A}$, we wish to establish an n-by-m matrix C where

$$C_{ij} = \begin{cases} 1 & \text{if there are } \geq 2 \text{ copies at probe } j \text{ in array } \tilde{\mathbf{a}}_i \\ 0 & \text{if there is 1 copy at probe } j \text{ in array } \tilde{\mathbf{a}}_i \\ -1 & \text{if there are 0 copies at probe } j \text{ in array } \tilde{\mathbf{a}}_i \ , \end{cases}$$

where the presence/absence is relative to the reference genome used as comparison. To compute C, we "smooth" the segmented profiles $\tilde{A}$ across the n arrays by merging overlapping intervals.

False Deletion Detection.

Despite the array's careful design, divergence among orthologs will have the same signature on the array results as deletions. In the five-species study, we use information in the sequenced genomes by aligning each probe to each genome and and quantify alignment scores within genes. We use the alignment score to adjust our confidence of the copy number states C and estimate a proportion of "false deletions."

3. RESULTS

In order to identify a core set of reliable probes for phylogenetic analysis we first performed 22 hybridizations using *O. niloticus* as a reference and 32 hybridizations using *A. burtoni* as reference in comparison to multiple individuals of the other three sequenced species (Figure 1). Probes that preformed consistently will give reliable information across a broad phylogenetic range. By using multiple individuals, we estimate the range of proportion of copy number variant that are expected to vary within a species. Our results show that some species have greater variance and fail to cluster. For example in the 22-array *O. niloticus* reference set, *O. niloticus* and *P. nyererei* cluster reliably while *M. zebra*, *A. burtoni*, and *N. brichardi* do not (Figure 2).

We also have the unique opportunity to quantify the technical error, since we used two species as different reference genomes. We compare the we compared the results obtained when *O. niloticus* was hybridized to *A.burtoni*. Dye artifacts and and other technical variance will be revealed though a lack of negative correlation.

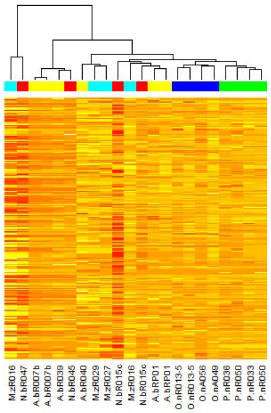

Figure 2: Heatmap of 22 hybridizations with *O. niloticus* as reference species. Columns represent aCGH arrays, colors directly below dendrogram indicate species (color legend in Figure 1).

Finally, with this core set of reliable probes, estimated reliability, and quantification of technical error rate, we apply our pipeline to aCGH data from 50 cichlid genomes. We map the copy number status to the current mitochondrial phylogeny for the 50 species in order to determine the evolutionary history of copy number variation among cichlids and its relationship to adaptive radiation.

4. REFERENCES

[1] D. Brawand et al. The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518):375–381, Sep 2014.

[2] C. Darwin. *On the origin of species*. New York :D. Appleton and Co.,. http://www.biodiversitylibrary.org/bibliography/28875.

[3] G. Fryer. Evolution of species flocks of cichlid fishes in african lakes. *Journal of Zoological Systematics and Evolutionary Research*, 15(2):141–165, 1977.

[4] S. Gould. *Wonderful Life: The Burgess Shale and the Nature of History*. A Norton paperback. W. W. Norton, 1990.

[5] H. E. Machado, G. Jui, D. A. Joyce, C. R. Reilly, D. H. Lunt, and S. C. Renn. Gene duplication in an African cichlid adaptive radiation. *BMC Genomics*, 15:161, 2014.

[6] S. Ohno, U. Wolf, and N. B. Atkin. Evolution from fish to mammals by gene duplication. *Hereditas*, 59(1):169–187, 1968.