Reconstructing *Anopheles* ancestral gene orders

Report, 2.1, April 27, 2014

Cedric Chauve[1] and Ashok Rajaraman[1,2]

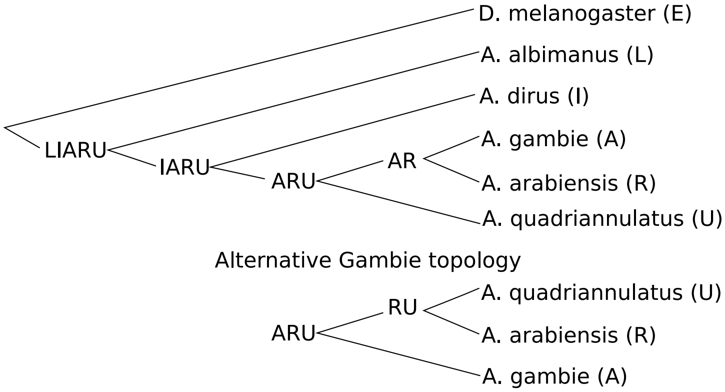
[1] Department of Mathematics, Simon Fraser University, Vancouver, BC, Canada
[2] International Graduate Training Center in Mathematical Biology, Pacific Institute of Mathematical Sciences, Vancouver, BC, Canada

Abstract. We describe a computationally reconstructed Contiguous Ancestral Regions (CARs) for several *Anopheles* ancestral genomes. At this point this work is purely a resource work, that makes available to the community a useful resource for the analysis ogf *Anopheles* genome evolution.

1 Preliminaries

The initial data of our experiments are composed of

- homologous gene families for the *Diptera* clade, obtained from Robert Waterhouse website at MIT [3];
- gene coordinates obtained from GFF files, again from Robert Waterhouse website at MIT[4];
- the following *Diptera* species tree, including a variant of the phylogeny of the gambiae complex.



We reconstructed the gene orders of the all ancestral *Anopheles* ancestors in this phylogeny, respectively denoted by AR, RU, ARU, $IARU$, $LIARU$ to be consistent with the notations of [1]. We used *Drosophila melanogaster* as outgroup.

[3] `http://people.csail.mit.edu/waterhouse/AGCC/Orthology/MOZ2-DEC2013/ODBMOZ2_Diptera_tabtext.gz`
[4] `http://people.csail.mit.edu/waterhouse/AGCC/Orthology/MOZ2-DEC2013/GFF/`

2 Methods and results

Method. We considered two methods and datasets:

- ANGES reconstructing CARs based on the 4,976 one-to-one orthologous genes families present in exactly one copy in each of the 6 genomes, using the method ANGES [2,4]; for this dataset, in order to be consistent with [1] who studied a similar dataset, we replaced each gene by its median exon to avoid filtering for overlapping genes.
- FPMAG reconstructing CARs from the genes families belonging to at least two of the 6 genomes using a variant FPMAG (unpublished) of the method FPSAC [3]. We filtered families containing overlapping genes and discarded gene families with an estimated ancestral copy number above 5, which resulted in roughly 11,000 gene families for each ancestor, providing thus a much broader gene coverage than the ANGES experiment.

Note that we considered the same datasets without *Drosophila melanogaster*, but unlike [1], we did not see a significant difference.

Results. The main result is thus a set of CARs for each of the considered ancestors. As shown in Table 1, we can notice an increase of the number of CARs with divergence time, with relatively well defined ancestors in the *Gambia* complex. As expected, the experiment ANGES, based on a smaller set of gene families, produces less CARs. One can also notice the difference between the two competing hypothesis *AR* and *RU* for the most recent *Gambia* ancestor: *AR* is much less fragmented than *RU*, that has similar charasteristics than *ARU*, which provides a stronger support for the *AR* topology over the *RU* topology.

Table 1. Statistics on the fragmentation of extant and ancestral genomes

# Segments/CARs	A	R	U	I	L	AR	RU	ARU	IARU	LIARU
ANGES (4,976 gene families)	6	111	320	117	39	39	76	67	249	1,201
FPMAG (17,461 gene families)	6	301	585	261	52	436	411	443	774	1,427

Tables 2 and 3 provide a more precise description of the fragmentation. For the ANGES experiment, for *Gambia* ancestors, most genes are in a few long CARs (i.e CARs containing at least 100 genes), a phenomenon that vanishes with older ancestors, although most genes of the *Cellia* ancestor ($IARU$) are still in roughly 100 CARs. This phenomenon is not apparent anymore in the FPMAG experiment, where roughly half of the genes are in CARs of size $11 - 50$.

Table 2. Statistics on the CARs content of ancestral genomes, ANGES experiment

# CARs/genes in CARs	AR	RU	ARU	IARU	LIARU
CARS of size 1	8/8	25/25	10/10	47/47	353/353
CARs of size 2-5	8/23	12/34	9/24	63/206	386/1,1195
CARs of size 6-10	0/0	0/0	3/22	42/306	177/1,342
CARs of size 11-50	4/137	11/344	20/491	68/1,831	119/1,947
CARs of size 51-100	8/549	13/917	10/688	23/1,594	2/139
CARs of size > 100	11/4,259	15/3,656	15/3,741	6/992	0/0

Table 3. Statistics on the CARs content of ancestral genomes, FPMAG experiment

# CARs/genes in CARs	AR	RU	ARU	IARU	LIARU
CARs of size 2-5	87/269	83/261	108/343	293/976	905/2,802
CARs of size 6-10	61/486	61/482	74/589	153/1,226	337/2,592
CARs of size 11-50	229/5,676	209/5,370	197/4,960	305/6,363	184/2,896
CARs of size 51-100	47/3,046	48/3,292	56/3,776	21/1,306	0/0
CARs of size > 100	11/1,376	9/1,203	7/1,049	1/162	0/0

Finally, we can address the robustness of the obtained results. With both methods, we can observe a very strong support for the provided results. Indeed both methods rely on the detection of conserved, and thus putatively ancestral, syntenic features (oriented adjacencies and intervals) under a Dollo parsimony criterion, that are then processed to be ordered in linear structures (CARs), by trying to minimize the number of syntenic features that need to be discarded to do so. As shown in Tables 4 and 5, in both experiments, the number of such discarded features is low. This is especially true for the ANGES experiment, pointing at a strong syntenic signal supporting the proposed CARs. The ratio of discarded intervals in the FPMAG experiment points at issues with ancestral copy number prediction.

Table 4. Syntenic support for computed CARs, ANGES experiment

# Syntenic features	AR	RU	ARU	IARU	LIARU
Adjacencies	4,940	4,904	4,909	4,665	3,729
Discarded	11	6	16	12	9
Intervals	2,183	2,358	2,393	2,438	1,917
Discarded	8	5	10	13	5

Table 5. Syntenic support for computed CARs, FPMAG experiment

# Syntenic features	AR	RU	ARU	IARU	LIARU
Adjacencies	11,392	10,997	11,367	9,959	7,247
Discarded	629	318	801	392	81
Intervals	717	728	723	510	312
Discarded	99	113	124	69	31

3 Conclusion

The main contribution of this work is a resource in the form of sets of well supported CARs for several ancestral *Anopheles* genomes that might prove useful for comparative and evolutionary studies of these genomes.

The main issue is the fragmentation of these CARs, that parallels the fragmentation of the available extant genomes.

All sets of CARs are available at : `http://paleogenomics.irmacs.sfu.ca/ANOPHELES`.

References

1. S. Aganezov Jr., M.A. Alekseyev Initial analysis of Anopheles evoltuion. Unpublished report. March 2014.
2. B. Jones, A. Rajaraman, E. Tannier, C. Chauve. ANGES: reconstructing ANcestral GEnomeS maps. *Bioinformatics*, 28:2388–2390. 2012. DOI: 10.1093/bioinformatics/bts457.
3. A. Rajaraman, E. Tannier, C. Chauve. FPSAC: Fast Phylogenetic Scaffolding of Ancient Contigs. *Bioinformatics*, 29:2987–2994. 2013. DOI: 10.1093/bioinformatics/btt527.
4. C. Chauve, E.Tannier. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput. Biol.*, 4(11):e1000234. 2008. DOI: 10.1371/journal.pcbi.1000234.