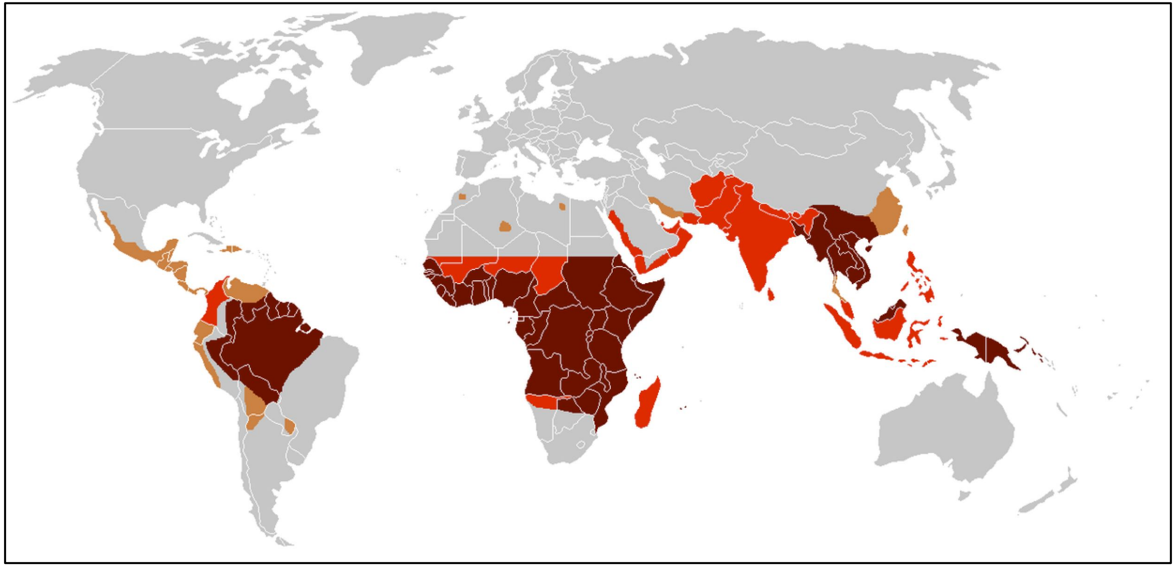
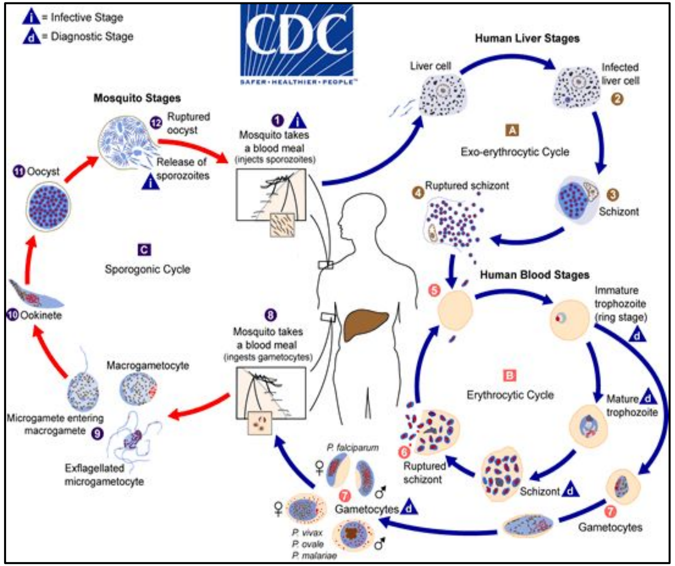
Comparative genomics of *Anopheles* mosquitoes genomes

Cedric Chauve
Department of Mathematics
Simon Fraser University

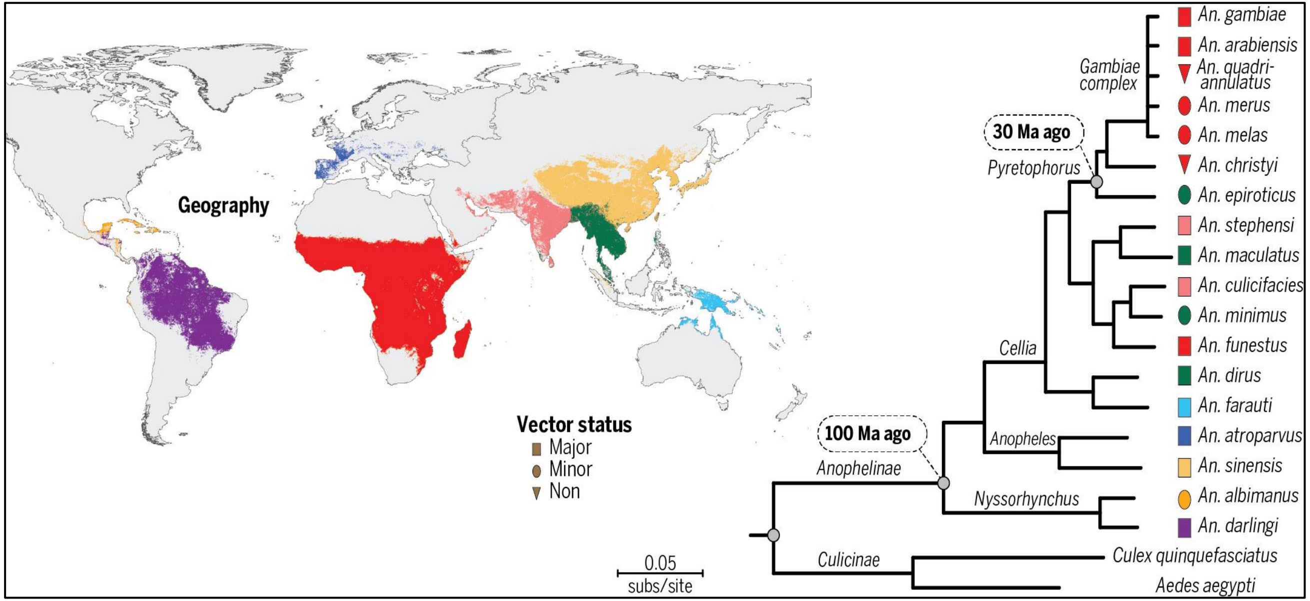
Anopheles mosquitoes, vector of the malaria

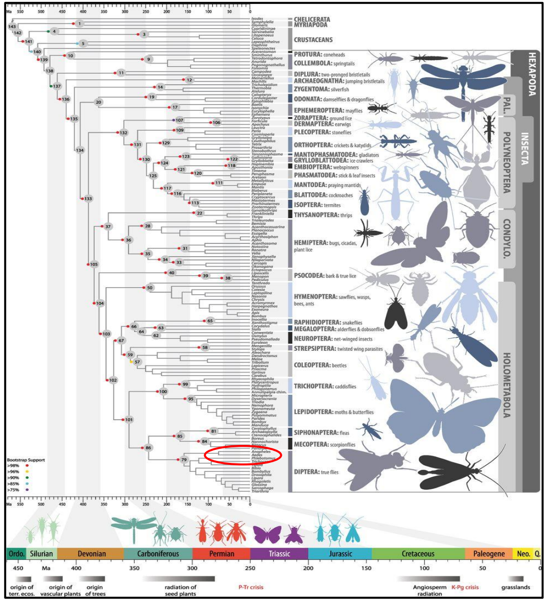




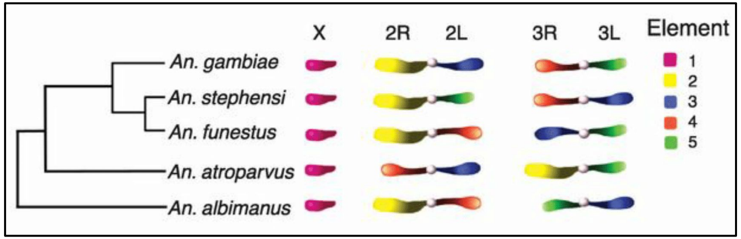
Malaria is a disease caused by the parasite *Plasmodium falciparum* whose are female mosquitoes of the genus *Anopheles*. The WHO estimates that in 2010 there were 219 million cases of malaria resulting in 660,000 deaths. The majority of cases occur in children under 15 years old. Controlling malaria transmission is a major public health issue.

Anopheles mosquitoes genomes: phylogeny





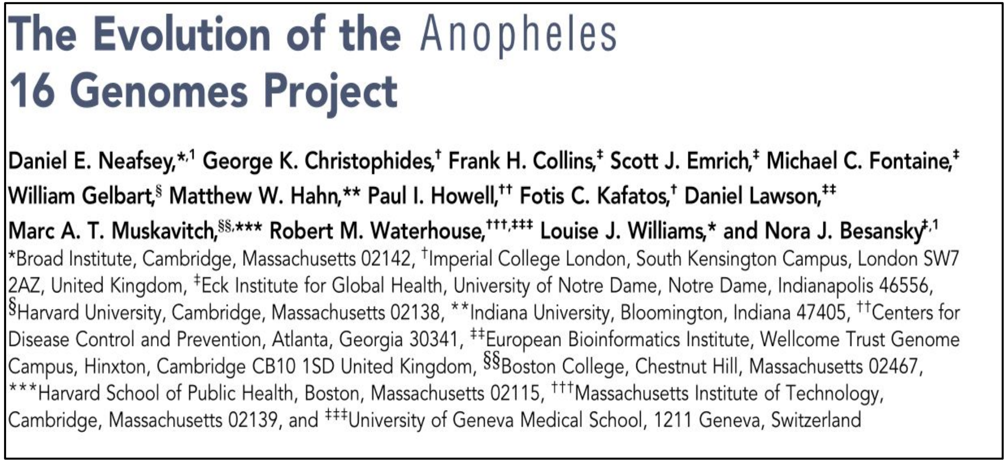
Misof *et al*, Science (2014)



Waterhouse *et al*, Science (2015)

- ❑ Highly variable vectorial capacities
- ❑ 100M years of evolution (recent insect group)
- ❑ Well resolved phylogeny (but for the Gambiae complex)
- ❑ Five chromosomal arms, ~280Mb, ~13,000 coding genes

The *Anopheles* Genome project



The Evolution of the Anopheles 16 Genomes Project

Daniel E. Neafsey,[*,1] George K. Christophides,[†] Frank H. Collins,[‡] Scott J. Emrich,[‡] Michael C. Fontaine,[‡] William Gelbart,[§] Matthew W. Hahn,[**] Paul I. Howell,[††] Fotis C. Kafatos,[†] Daniel Lawson,[‡‡] Marc A. T. Muskavitch,[§§,***] Robert M. Waterhouse,[†††,‡‡‡] Louise J. Williams,[*] and Nora J. Besansky[‡,1]

*Broad Institute, Cambridge, Massachusetts 02142, [†]Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom, [‡]Eck Institute for Global Health, University of Notre Dame, Notre Dame, Indianapolis 46556, [§]Harvard University, Cambridge, Massachusetts 02138, **Indiana University, Bloomington, Indiana 47405, [††]Centers for Disease Control and Prevention, Atlanta, Georgia 30341, [‡‡]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD United Kingdom, [§§]Boston College, Chestnut Hill, Massachusetts 02467, ***Harvard School of Public Health, Boston, Massachusetts 02115, [†††]Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, and [‡‡‡]University of Geneva Medical School, 1211 Geneva, Switzerland

The biological basis for **variable vectorial capacity** surely lies in poorly understood differences in mosquito physiology, molecular biology, and / or behavior.

...

The publication of the African vector *Anopheles gambiae* genome in 2002 was a landmark for the field of malaria vector research, but **gaining a better understanding of vectorial capacity clearly requires a comparative framework.**

Once **assemblies** and gene predictions are finalized, the **community analysis** will begin in earnest. Members of the vector community and wider genomics research community interested in the comparative analyses are invited to contact the project organizers. Major analysis themes will include speciation, molecular evolution, chemoreception, circadian rhythm, development, immunity, insecticide resistance, metabolism, repetitive elements, reproduction, the sialome, **inversions and chromosomal architectures**, neuropeptides, blood / sugar digestion, and transcriptional regulation.

Neafsey *et al*, G3 (2013))

Comparative genomics of *Anopheles* mosquitoes

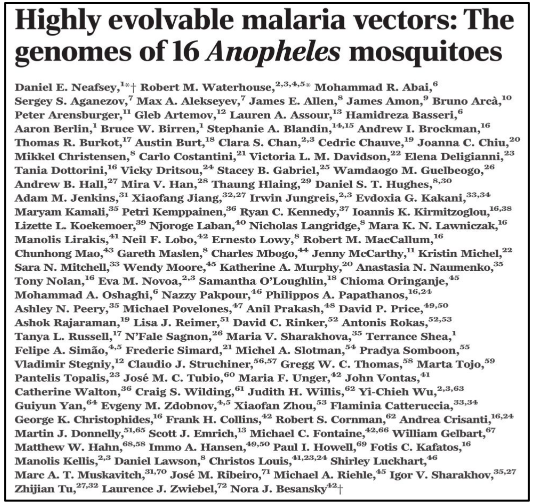
Fantastic!

We have genomes, annotated genes, gene families, a phylogeny, so in theory we have all we need to do a comparative study of these insects, with a focus on **gene family, genome and gene order evolution**. Can we just plug-in these data into our published and tested methods?

We did it.

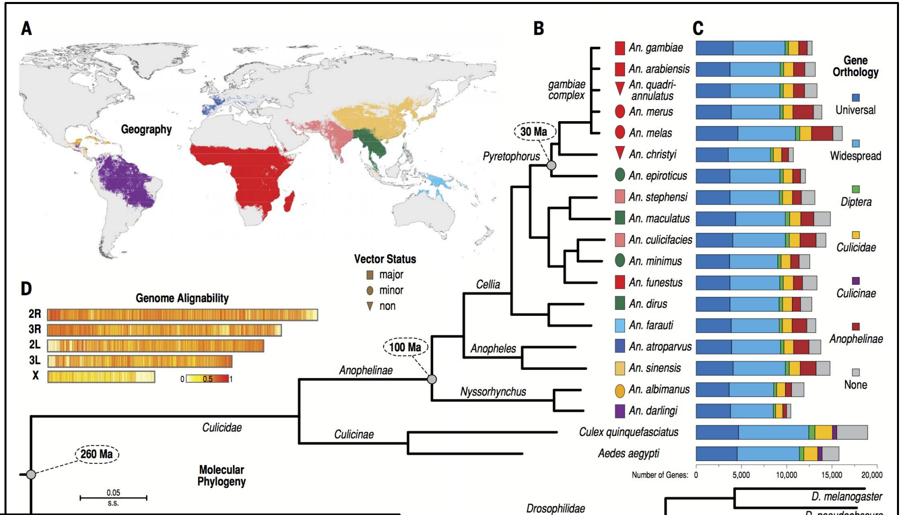
- ❏ Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. *Science*, 2015.
- ❏ Phylogenetic signal from rearrangements in 18 *Anopheles* species by joint scaffolding extant and ancestral genomes. *BMC Genomics*, 2018.
- ❏ Detecting Introgression in *Anopheles* Mosquito Genomes Using a Reconciliation-Based Approach. *LNBI,* 2018 (RECOMB-CG).
- ❏ Evolutionary superscaffolding and chromosome anchoring to improve *Anopheles* genome assemblies. *Biorxiv,* 2018 (in revision).

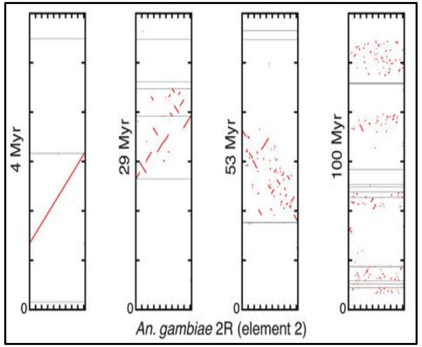
The *Anopheles* Science papers (2015)



Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes

Daniel E. Neafsey,[1*†] Robert M. Waterhouse,[2,3,4,5*] Mohammad R. Abai,[6] Sergey S. Aganezov,[7] Max A. Alekseyev,[7] James E. Allen,[8] James Amon,[9] Bruno Arcà,[10] Peter Arensburger,[11] Gleb Artemov,[12] Lauren A. Assour,[13] Hamidreza Basseri,[6] Aaron Berlin,[1] Bruce W. Birren,[1] Stephanie A. Blandin,[14,15] Andrew I. Brockman,[16] Thomas R. Burkot,[17] Austin Burt,[18] Clara S. Chan,[2,3] Cedric Chauve,[19] Joanna C. Chiu,[20] Mikkel Christensen,[8] Carlo Costantini,[21] Victoria L. M. Davidson,[22] Elena Deligianni,[23] Tania Dottorini,[16] Vicky Dritsou,[24] Stacey B. Gabriel,[25] Wamdaogo M. Guelbeogo,[26] Andrew B. Hall,[27] Mira V. Han,[28] Thaung Hlaing,[29] Daniel S. T. Hughes,[8,30] Adam M. Jenkins,[31] Xiaofang Jiang,[32,27] Irwin Jungreis,[2,3] Evdoxia G. Kakani,[33,34] Maryam Kamali,[35] Petri Kemppainen,[36] Ryan C. Kennedy,[37] Ioannis K. Kirmitzoglou,[16,38] Lizette L. Koekemoer,[39] Njoroge Laban,[40] Nicholas Langridge,[8] Mara K. N. Lawniczak,[16] Manolis Lirakis,[41] Neil F. Lobo,[42] Ernesto Lowy,[8] Robert M. MacCallum,[16] Chunhong Mao,[43] Gareth Maslen,[8] Charles Mbogo,[44] Jenny McCarthy,[11] Kristin Michel,[22] Sara N. Mitchell,[33] Wendy Moore,[45] Katherine A. Murphy,[20] Anastasia N. Naumenko,[35] Tony Nolan,[16] Eva M. Novoa,[2,3] Samantha O'Loughlin,[18] Chioma Oringanje,[45] Mohammad A. Oshaghi,[6] Nazzy Pakpour,[46] Philippos A. Papathanos,[16,24] Ashley N. Peery,[35] Michael Povelones,[47] Anil Prakash,[48] David P. Price,[49,50] Ashok Rajaraman,[19] Lisa J. Reimer,[51] David C. Rinker,[52] Antonis Rokas,[52,53] Tanya L. Russell,[17] N'Fale Sagnon,[26] Maria V. Sharakhova,[35] Terrance Shea,[1] Felipe A. Simão,[4,5] Frederic Simard,[21] Michel A. Slotman,[54] Pradya Somboon,[55] Vladimir Stegniy,[12] Claudio J. Struchiner,[56,57] Gregg W. C. Thomas,[58] Marta Tojo,[59] Pantelis Topalis,[23] José M. C. Tubio,[60] Maria F. Unger,[42] John Vontas,[41] Catherine Walton,[36] Craig S. Wilding,[61] Judith H. Willis,[62] Yi-Chieh Wu,[2,3,63] Guiyun Yan,[64] Evgeny M. Zdobnov,[4,5] Xiaofan Zhou,[53] Flaminia Catteruccia,[33,34] George K. Christophides,[16] Frank H. Collins,[42] Robert S. Cornman,[62] Andrea Crisanti,[16,24] Martin J. Donnelly,[51,65] Scott J. Emrich,[13] Michael C. Fontaine,[42,66] William Gelbart,[67] Matthew W. Hahn,[68,58] Immo A. Hansen,[49,50] Paul I. Howell,[69] Fotis C. Kafatos,[16] Manolis Kellis,[2,3] Daniel Lawson,[8] Christos Louis,[41,23,24] Shirley Luckhart,[46] Marc A. T. Muskavitch,[31,70] José M. Ribeiro,[71] Michael A. Riehle,[45] Igor V. Sharakhov,[35,27] Zhijian Tu,[27,32] Laurence J. Zwiebel,[72] Nora J. Besansky[42†]





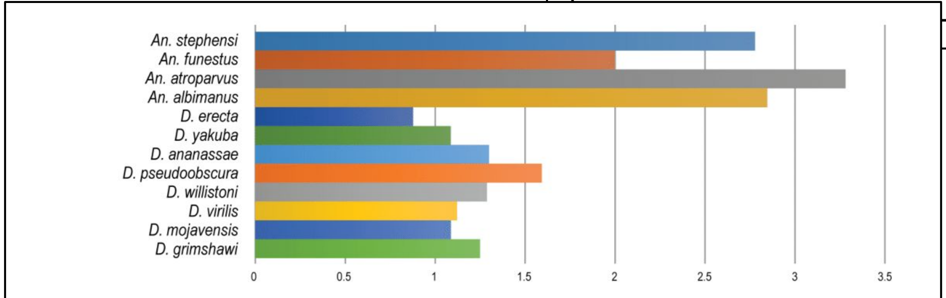


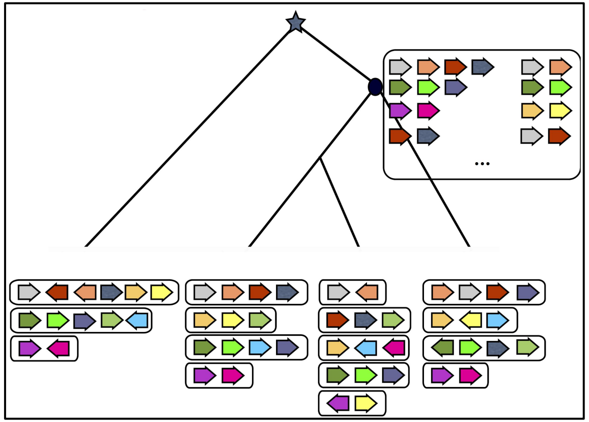
Figure S8. X versus autosome rearrangement rates.

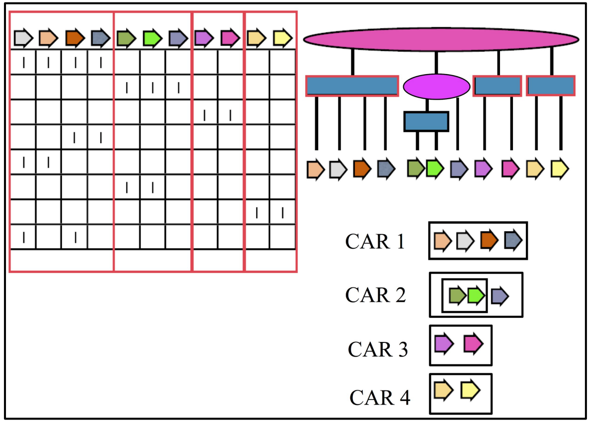
The ratio of X chromosome evolutionary rate to the total rate of rearrangement in anophelines and drosophilids.

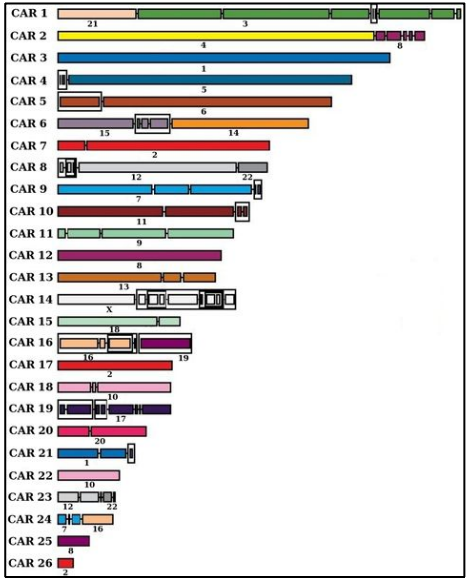
- ❑ Synteny conservation at arm level
- ❑ Gene order shuffling
- ❑ Striking difference of rearrangement rates between sex chromosomes and autosomes, compared to other dipteran genomes.

Waterhouse *et al*, Science (2015)

Reconstructing ancestral genomes from unique genes: methods (AnGeS)



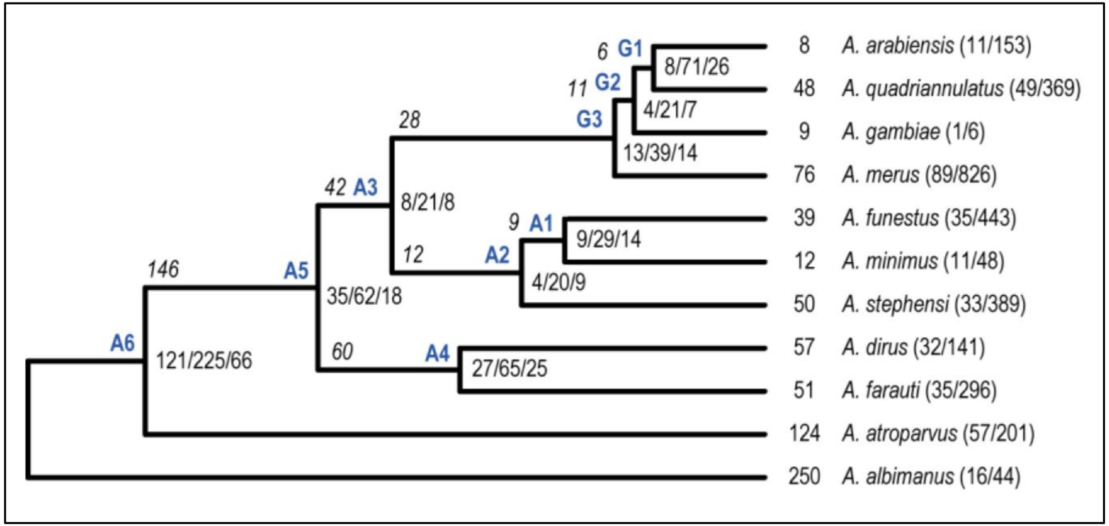




- Detect group of co-localized unique genes conserved in several genomes (Dollo criterion)
- Weight such groups in based on their conservation pattern
- Extract a maximum-weight subset that is compatible with a linear order
- CARs: Contiguous Ancestral Regions

Chauve & Tannier, PLoS Comp. Biol. (2008)
Ouangraoua *et al*, J. Comp. Biol. (2010)
Gavranovic *et al*, Bioinformatics (2011)
Wittler *et al*, CPM (2011)
Jones *et al*, Bioinformatics (2012)

Reconstructing ancestral genomes from unique genes: results



X / W / N90:

1. X = number of X chromosome CARs
2. W = number of WG CARs
3. N90 = number of CARs covering 90% of the 5,343 genes

Branches length: rearrangement (DCJ) distance

- 5,343 one-to-one orthologous genes
- ANGeS applied independently to each internal node
- Rearrangement distance corrected to account for fragmented genomes assemblies / reconstruction

- Very low level of syntenic conflict
- A few number of CARs cover most genes
- In 11 out of 17 branches of the A5 subtree, the ratio of the DCJ distance for the X chromosome and the autosomes is at least 2.
- The X chromosome reconstruction supports the new Gambiae complex phylogeny

Waterhouse *et al*, Science (2015)

Comparative genomics of *Anopheles* mosquitoes

Fantastic!

We have genomes, annotated genes, gene families, a phylogeny, so in theory we have all we need to do a comparative study of these insects, with a focus on **gene family, genome and gene order evolution**. Can we just plug-in these data into our published and tested methods?

Actually, we should not.

The work we did on the Science paper was quite unsatisfying and did not address properly many of the issues we need tod eal with on such data;

- ❑ The gene families are likely quite noisy.
- ❑ The genomes are poorly assembled.
- ❑ The phylogeny of the recently diverged *Gambia complex* is controversial.
- ❑ The evolution of closely related mosquitoes involve introgression
- ❑ ...

We need to look more into how these data have been obtained, where are some possible issues, how we can address them, before we can do an actual biologically meaningful analysis of the data.

Goal of the seminar

The goal of this seminar course is to go through some of the preliminary steps required to do comparative genomics with such “real-world large-scale data”, to look into the actual methods used in these projects, reflect on the encountered issues, discuss the solutions we proposed, and why not propose alternative solutions or ideas.

This is a classical seminar course, centered on bioinformatics papers, but the data are easily available and we can think about experiments toward research results.

Overview of the analysis pipeline (1)

Starting data. Sequencing data, species phylogeny

1: Genome assembly.
Contigs assembly, Scaffolding, Assembly metrics

2: Genes.
Gene annotation (skipped)

3: Genomic markers.
Orthology, Gene families, Synteny blocks

These steps provide the material required to start work on genome rearrangements and genome evolution.

Overview of the analysis pipeline (2)

Starting data: species phylogeny, gene families, extant gene orders

1: Gene families.

Copy-number evolution

Reconciled gene trees

2: Ancestral Gene Orders (AGO).

Single-copy markers

Gene adjacency evolution, SPP with duplicated markers

Joint AGO/Scaffolding

The *Anopheles* genomes assemblies

Data.

- ❏ HiSeq2000 libraries from a single female for each species:
  - ❏ 101bp, 180bp insert, 30-170 Depth of Coverage (DoC)
  - ❏ 101bp, 1.5kb jump, 50-145 DoC
- ❏ Fosill libraries (Williams *et al*, Genome Res. (2012)) from hundred females, 101bp, 35kb jump, 3-11 DoC, 11 species.

Software/strategy.

- ❏ ALLPATHS-LG (Gnerre et al, PNAS (2011)) + Pilon (Walker et al, PLoS One (2014))
- ❏ *An. gambiae* used as a reference for the Gambia complex

Species	*An. alb.*	*An. atr.*	*An. chr.*	*An. col.*	*An. dar.*	*An. dir.*	*An. far.*	*An. fun.*	*An. mac.*	*An. mel.*	*An. mer.*	*An. min.*	*An. qua.*
Size	170Mb	224	172	224	134	216	181	225	141	227	251	201	283
Gaps	6.9Kb	35Mb	2.6Mb	15Mb	27Kb	18Mb	5Mb	35Mb	10Mb	20Mb	33Mb	15Mb	75Mb
Scf. N50	18Mb	9.2Mb	9Kb	4.4Mb	115Kb	7Mb	1,2Mb	672Kb	4Kb	18Kb	342Kb	1.5Mb	1.6Mb
#Scf.	204	1,371	30K	10K	2,160	1,266	550	1,392	47K	20K	2,753	678	2,823

Waterhouse *et al*, Science (2015)

The *Anopheles* genomes assemblies: papers

ALLPATHS: High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *PNAS*, 2011.

PILON: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One*, 2014.

Mostly of historical interest, because used in this project.

The *Anopheles* genomes assemblies: papers

The most interesting element of the assembly in this project was the use of the *An. gambia* genome as a reference for the *Gambia complex*. There are interesting methods for comparative scaffolding (a topic we will discuss later):

- RAGOUT 2: Chromosome assembly of large and complex genomes using multiple references. *Genome Research,* 2018.
- RACA: Reference-assisted chromosome assembly. *PNAS,* 2013.

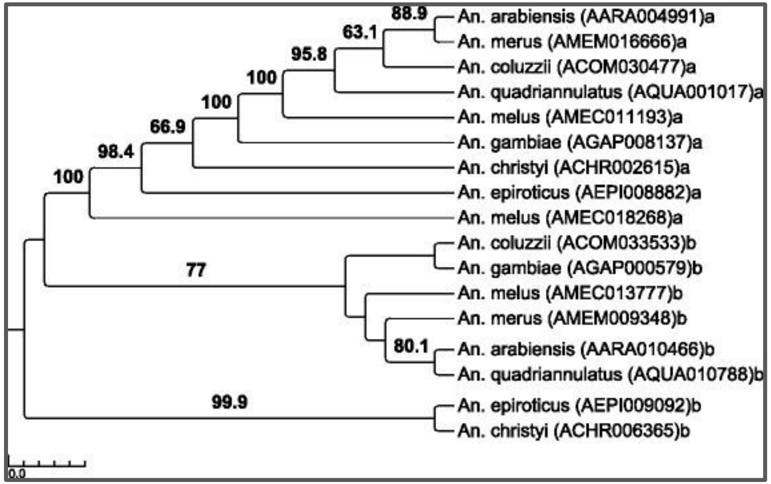
In recent *Anopheles* papers, there has also be some attempts to improve scaffolding using *transcriptomics* data

- AGOUTI: improving genome assembly and annotation using transcriptome data. *GigaScience*, 2016.
  Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLoS Comp. Biol,* 2014.

The *Anopheles* genes families

Once we have assembled our genomes the best we can using the available data, we need to define *families of genomic markers*, that will for the basic input to study genome evolution.

The natural markers, that speak to biologists, are *genes*, forming *homologous gene families*.


88.9
63.1
95.8
100
100
66.9
98.4
100
77
80.1
99.9
An. arabiensis (AARA004991)a
An. merus (AMEM016666)a
An. coluzzii (ACOM030477)a
An. quadriannulatus (AQUA001017)a
An. melus (AMEC011193)a
An. gambiae (AGAP008137)a
An. christyi (ACHR002615)a
An. epiroticus (AEPI008882)a
An. melus (AMEC018268)a
An. coluzzii (ACOM033533)b
An. gambiae (AGAP000579)b
An. melus (AMEC013777)b
An. merus (AMEM009348)b
An. arabiensis (AARA010466)b
An. quadriannulatus (AQUA010788)b
An. epiroticus (AEPI009092)b
An. christyi (ACHR006365)b
0.0

The *Anopheles* genes families: OrthoDB

Gene families are usually defined in terms of *sequence homology*: an all-against-all comparison of all coding sequences serves as basic input and then various methods can be used from this input to cluster genes into families.

For the *Anopheles* dataset, we used OrthoDB:

❑ OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *NAR*, 2015

OrthoDB has a sister method that uses *universal orthologs* to assess the completeness of assemblies.

❑ BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *GBE,* 2017.

The *Anopheles* genes families: other methods

There are many other ways to compute gene families and orthologous groups. Recent examples:

- ❑ OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 2015.
- ❑ GETHOGS: Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. *Bioinformatics*, 2017.
- ❑ HyPPO: Accurate prediction of orthologs in the presence of divergence after duplication. *Bioinformatics*, 2017.

The *Anopheles* genes families: synteny

Next, one interesting aspect is to account also for the conservation of *synteny*, i.e. genome co-localization:

❏ Quantitative synteny scoring improves homology inference and partitioning of gene families. *BMC Bioinformatics*, 2013.
GenFamClust: an accurate, synteny-aware and reliable homology inference algorithm. *BMC Evol. Biol.* 2016.

❏ SYNFUSE: Synteny guided fusion of *Anopheles* gene families. SFU Honours Thesis 2017.
This last paper is actually pretty open for research: the initial idea seems to be very effective, but several issues in methods design and implementation are open.

The *Anopheles* genes families: synteny blocks

Last, genes are not the only possible genomics markers. There exist other ways to define such markers, especially *synteny blocks*, that have never be tried on *Anopheles* data:

- ❏ Evaluating synteny for improved comparative studies. *Bioinformatics*, 2014.
  Sequence-Based Synteny Analysis of Multiple Large Genomes. *Methods in Molecular Biology*, 2018.