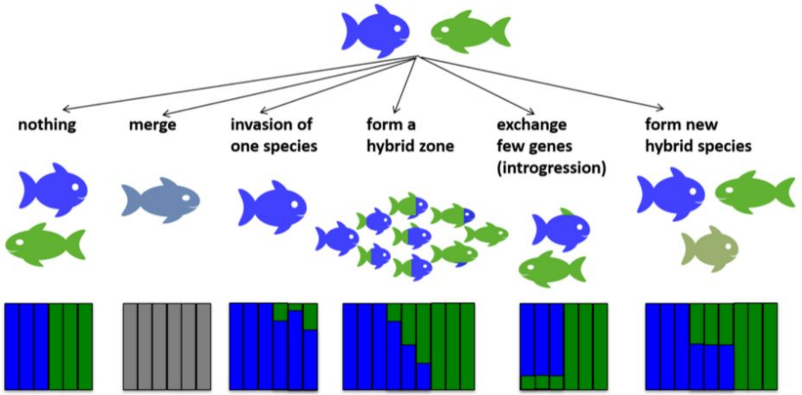
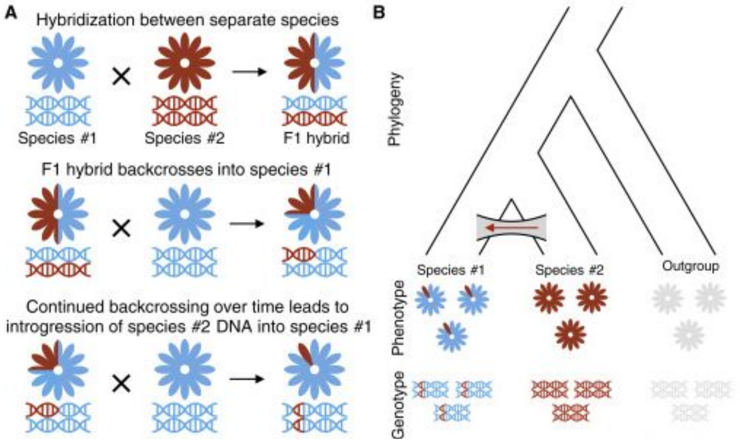
# Inversions and barrier loci in a complex of butterfly species

Fabrice Legeai
Journées PEPI IBIS - 15/10/2025

# Introgression and gene flow

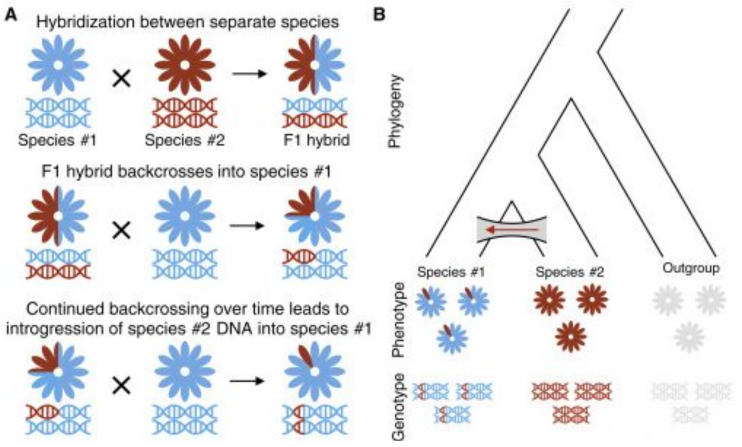




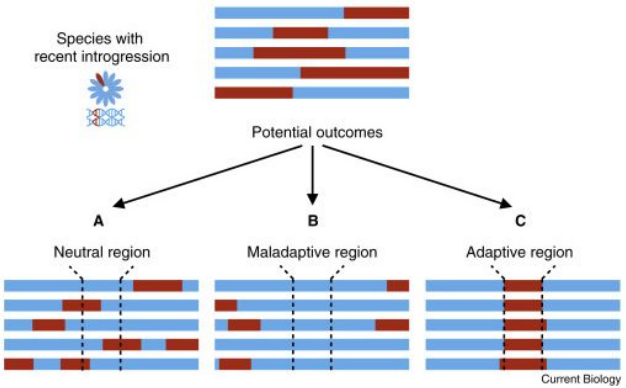
Aguillon et al. Current Biology 2022

# Barrier loci

**Loci resisting to introgression** = Regions of the genome that prevent the exchange of genetic material (gene flow) between different populations or species.



Aguillon et al. Current Biology 2022

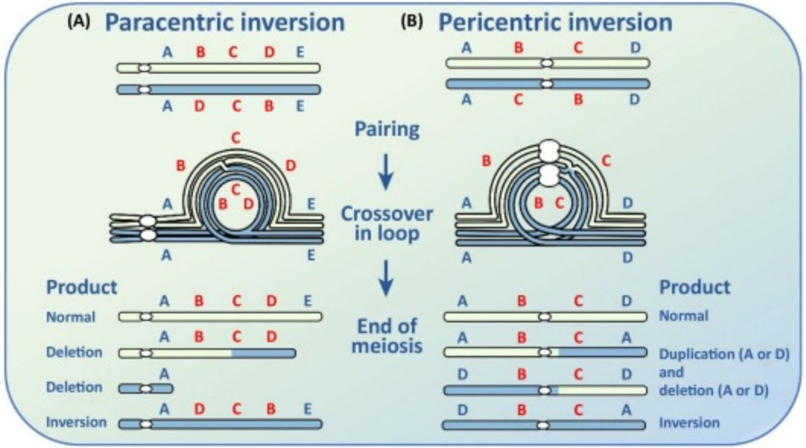


**Local adaptation genes:** Genes that are advantageous in one environment but harmful in another, so selection prevents their spread between populations.

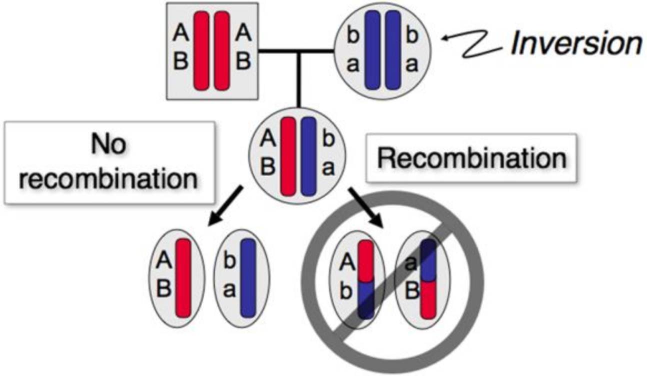
**Chromosomal rearrangements:** Inversions or translocations that suppress recombination in certain regions.

# Impact of large inversions

inversions are in the heterozygous state in a diploid genome => a single crossover during meiosis within the inversion generates unbalanced gametes that contain duplication and/or deletion.



Wellenreuther and Bernatchez Trends in Ecol and Evol
https://doi.org/10.1016/j.tree.2018.04.002

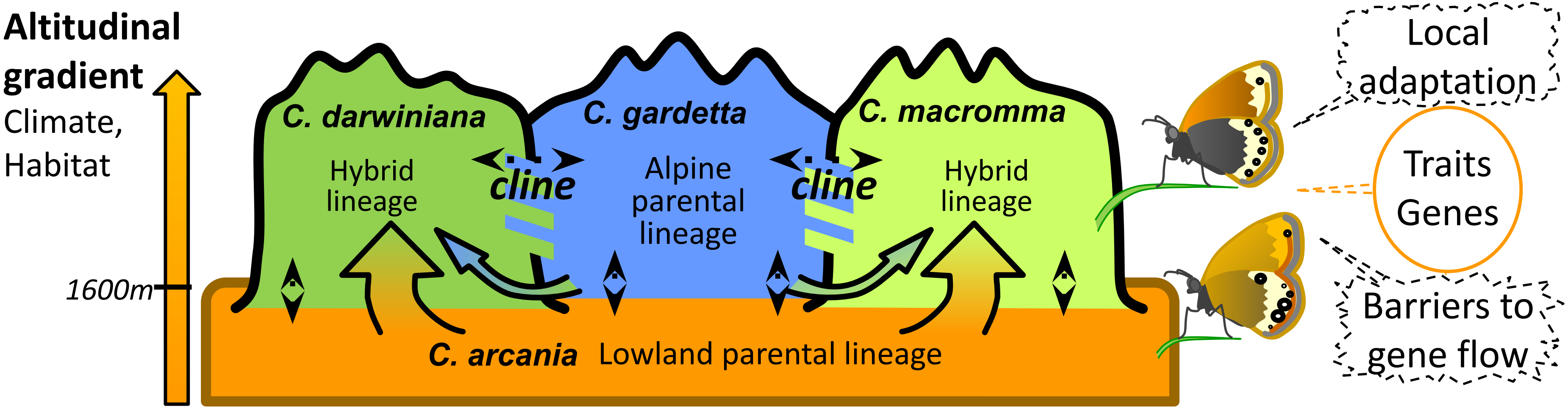


Kirkpatrick M.. *PLoS Biol*. 2010;8(9):e1000501.
doi:10.1371/journal.pbio.1000501

large inverted non-recombining stretches of DNA provides a way for clusters of adaptive genes or genes involved in reproductive isolation to avoid recombination with sister or parental lineages

# The Divalps project

identify genomic regions (i.e. barrier loci and inversions) associated to local adaptation and barriers to gene flow



**Altitudinal gradient**

Climate,

Habitat

*1600m*

***C. darwiniana***

Hybrid lineage

***C. gardetta***

Alpine parental lineage

***C. macromma***

Hybrid lineage

**cline**

**cline**

***C. arcania*** Lowland parental lineage

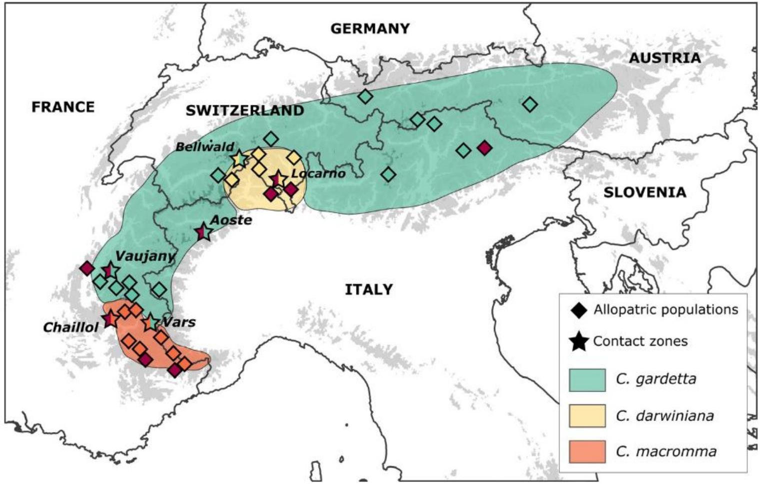
Local adaptation

Traits

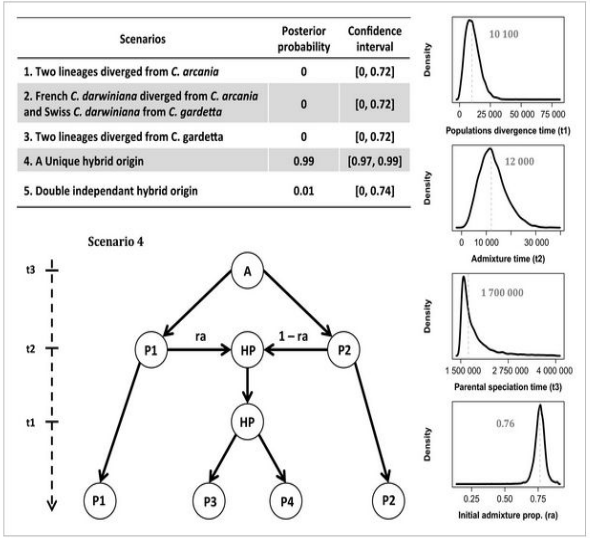
Genes

Barriers to gene flow

# The Coenonympha complex species

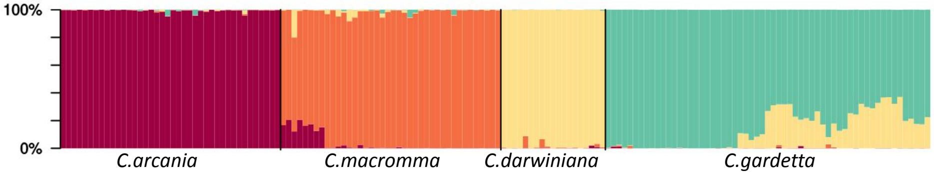


Capblancq T et al. Ecol Evol. 2019 May 3;9(11):6444-6457. doi: 10.1002/ece3.5220.

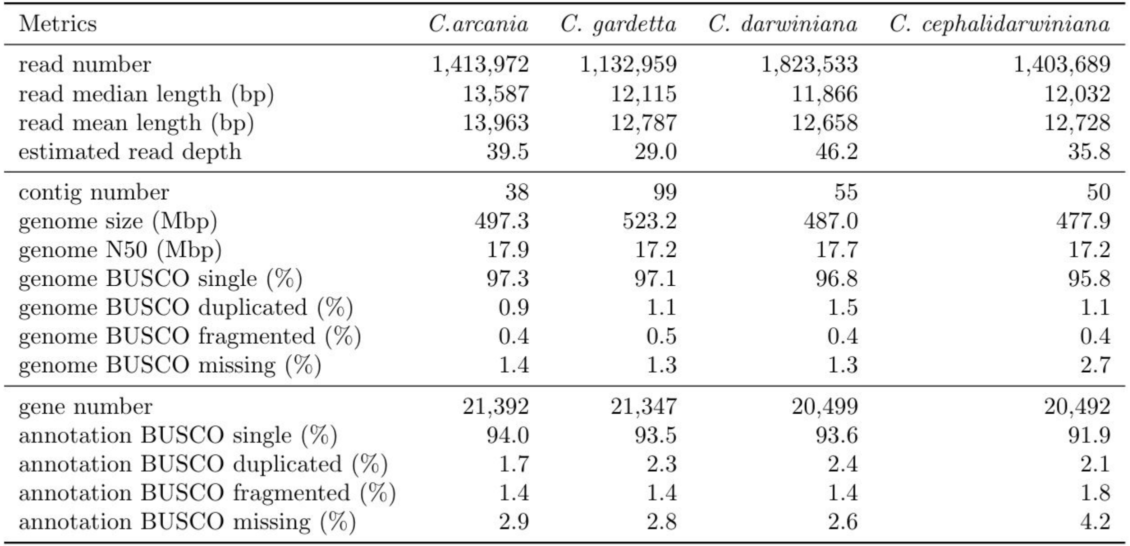


| Scenarios | Posterior probability | Confidence interval |
|---|---|---|
| 1. Two lineages diverged from *C. arcania* | 0 | [0, 0.72] |
| 2. French *C. darwiniana* diverged from *C. arcania* and Swiss *C. darwiniana* from *C. gardetta* | 0 | [0, 0.72] |
| 3. Two lineages diverged from C. gardetta | 0 | [0, 0.72] |
| 4. A Unique hybrid origin | 0.99 | [0.97, 0.99] |
| 5. Double independant hybrid origin | 0.01 | [0, 0.74] |

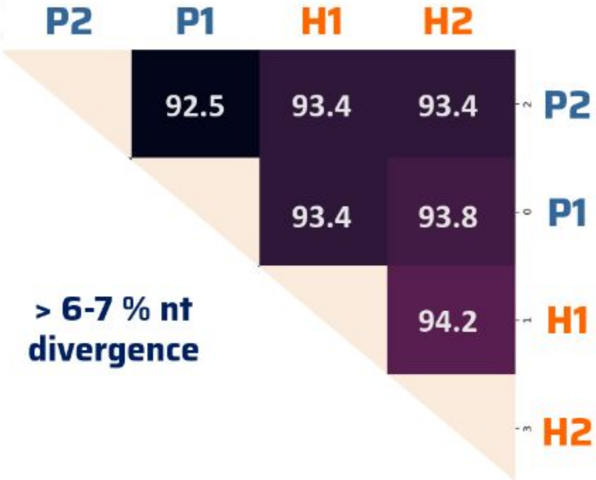
Capblancq T et al. Mol. Ecol.. 2015, doi: 10.1111/mec.13479



# Genome assemblies



| Metrics | *C.arcania* | *C. gardetta* | *C. darwiniana* | *C. cephalidarwiniana* |
|---|---|---|---|---|
| read number | 1,413,972 | 1,132,959 | 1,823,533 | 1,403,689 |
| read median length (bp) | 13,587 | 12,115 | 11,866 | 12,032 |
| read mean length (bp) | 13,963 | 12,787 | 12,658 | 12,728 |
| estimated read depth | 39.5 | 29.0 | 46.2 | 35.8 |
| contig number | 38 | 99 | 55 | 50 |
| genome size (Mbp) | 497.3 | 523.2 | 487.0 | 477.9 |
| genome N50 (Mbp) | 17.9 | 17.2 | 17.7 | 17.2 |
| genome BUSCO single (%) | 97.3 | 97.1 | 96.8 | 95.8 |
| genome BUSCO duplicated (%) | 0.9 | 1.1 | 1.5 | 1.1 |
| genome BUSCO fragmented (%) | 0.4 | 0.5 | 0.4 | 0.4 |
| genome BUSCO missing (%) | 1.4 | 1.3 | 1.3 | 2.7 |
| gene number | 21,392 | 21,347 | 20,499 | 20,492 |
| annotation BUSCO single (%) | 94.0 | 93.5 | 93.6 | 91.9 |
| annotation BUSCO duplicated (%) | 1.7 | 2.3 | 2.4 | 2.1 |
| annotation BUSCO fragmented (%) | 1.4 | 1.4 | 1.4 | 1.8 |
| annotation BUSCO missing (%) | 2.9 | 2.8 | 2.6 | 4.2 |



| | P2 | P1 | H1 | H2 |
|---|---|---|---|---|
| P2 | | 92.5 | 93.4 | 93.4 |
| P1 | | | 93.4 | 93.8 |
| H1 | | | | 94.2 |
| H2 | | | | |

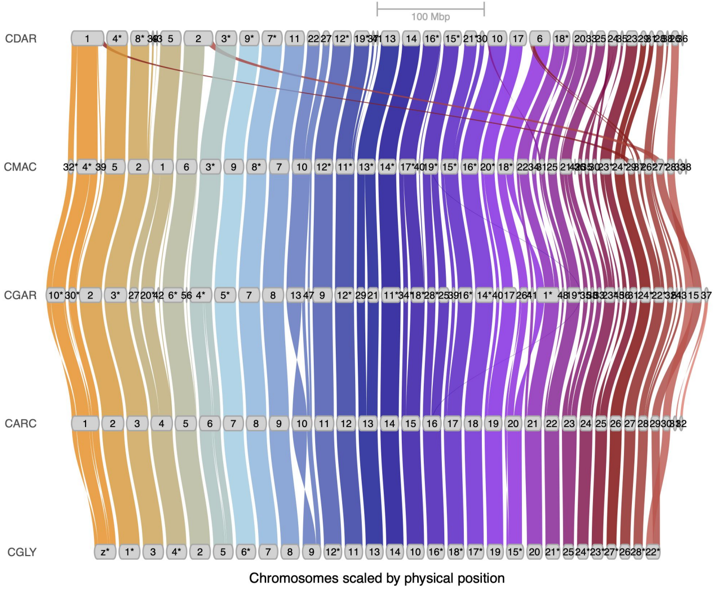
**> 6-7 % nt divergence**

Matrix of estimated ANI (% average nucleotidic identity) in the complex

calculated with skani [Shaw & Yu 2023] (only aligned 20-30% of the genomes)

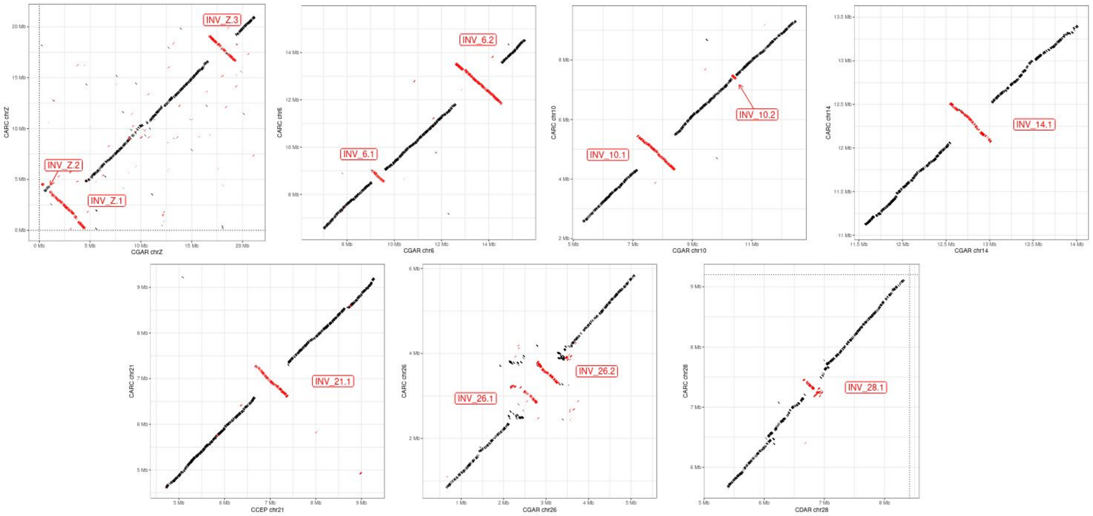
# Genome assemblies

| Metrics | *C.arcania* | *C. gardetta* | *C. darwiniana* | *C. cephalidarwiniana* |
|---|---|---|---|---|
| read number | 1,413,972 | 1,132,959 | 1,823,533 | 1,403,689 |
| read median length (bp) | 13,587 | 12,115 | 11,866 | 12,032 |
| read mean length (bp) | 13,963 | 12,787 | 12,658 | 12,728 |
| estimated read depth | 39.5 | 29.0 | 46.2 | 35.8 |
| contig number | 38 | 99 | 55 | 50 |
| genome size (Mbp) | 497.3 | 523.2 | 487.0 | 477.9 |
| genome N50 (Mbp) | 17.9 | 17.2 | 17.7 | 17.2 |
| genome BUSCO single (%) | 97.3 | 97.1 | 96.8 | 95.8 |
| genome BUSCO duplicated (%) | 0.9 | 1.1 | 1.5 | 1.1 |
| genome BUSCO fragmented (%) | 0.4 | 0.5 | 0.4 | 0.4 |
| genome BUSCO missing (%) | 1.4 | 1.3 | 1.3 | 2.7 |
| gene number | 21,392 | 21,347 | 20,499 | 20,492 |
| annotation BUSCO single (%) | 94.0 | 93.5 | 93.6 | 91.9 |
| annotation BUSCO duplicated (%) | 1.7 | 2.3 | 2.4 | 2.1 |
| annotation BUSCO fragmented (%) | 1.4 | 1.4 | 1.4 | 1.8 |
| annotation BUSCO missing (%) | 2.9 | 2.8 | 2.6 | 4.2 |



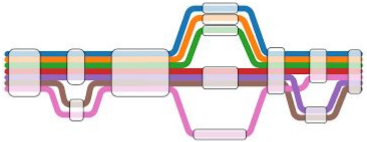
# Characterization of large inversions

*C. arcania* genome was compared with the other three genomes of the complex and *C. glycerion* using minimap2+chroder+Syri+merging+manual curation

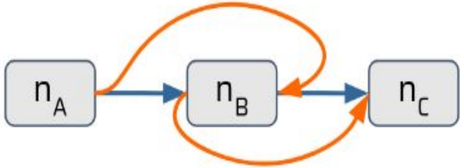


12 inversions
> 100kbp

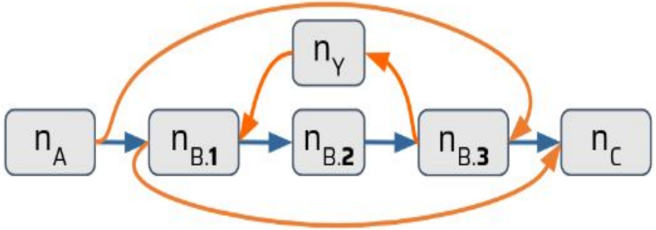
# (Identification of the inversions in pangenome graph)



| Tool | Resolution | Scope |
|---|---|---|
| **Minigraph** [Li *et al.* 2020] | ≥ 50 bp (SV) | intra-species |
| **Cactus** [Armstrong *et al. 2020*] | ≥ 1 bp (SNP, indel, SV) | inter-species |
| **Minigraph-Cactus** [Hickey *et al.* 2023] | | intra-species |
| **PGGB** [Garrison *et al.* 2023, preprint] | | intra- to inter-species? |

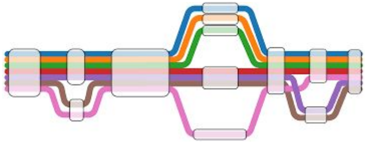


Simplest form

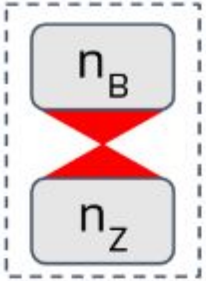


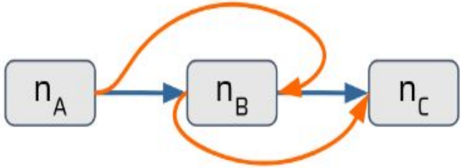
With divergence

# (Identification of the inversions in pangenome graph)

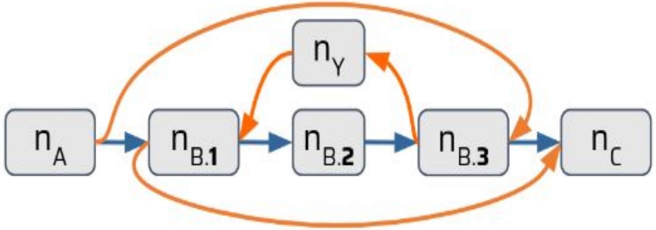


| Tool | Resolution | Scope |
|---|---|---|
| **Minigraph** [Li *et al.* 2020] | ≥ 50 bp (SV) | intra-species |
| **Cactus** [Armstrong *et al. 2020*] | ≥ 1 bp (SNP, indel, SV) | inter-species |
| **Minigraph-Cactus** [Hickey *et al.* 2023] | | intra-species |
| **PGGB** [Garrison *et al.* 2023, preprint] | | intra- to inter-species? |

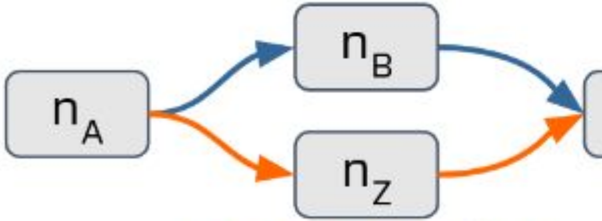




Simplest form

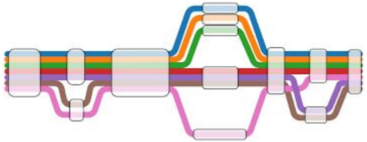


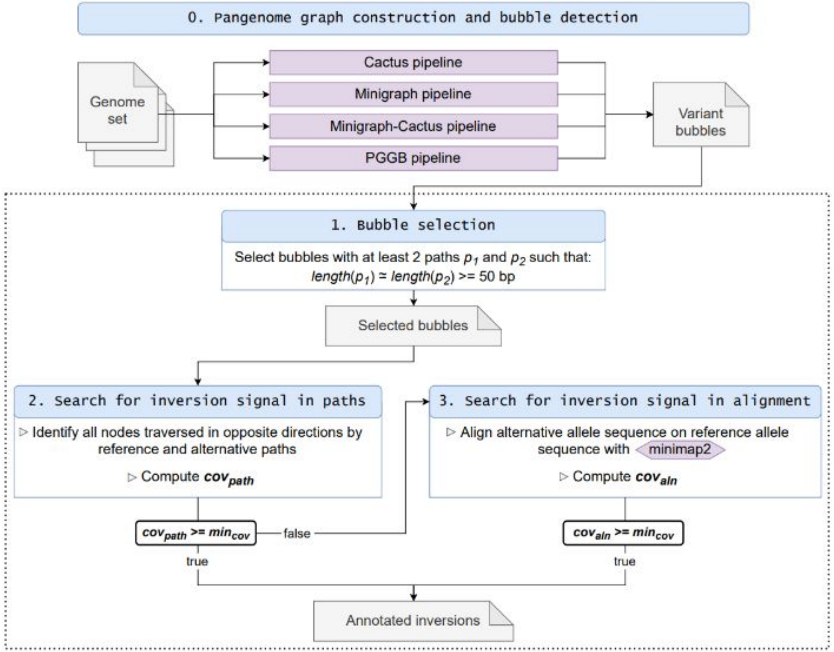
With divergence



Represented as unrelated sequences

# (Identification of the inversions in pangenome graph)





`0. Pangenome graph construction and bubble detection`

Genome set

Cactus pipeline

Minigraph pipeline

Minigraph-Cactus pipeline

PGGB pipeline

Variant bubbles

`1. Bubble selection`

Select bubbles with at least 2 paths $p_1$ and $p_2$ such that:
$length(p_1) \simeq length(p_2) >= 50$ bp

Selected bubbles

`2. Search for inversion signal in paths`

▷ Identify all nodes traversed in opposite directions by reference and alternative paths

▷ Compute $cov_{path}$

`3. Search for inversion signal in alignment`

▷ Align alternative allele sequence on reference allele sequence with minimap2

▷ Compute $cov_{aln}$

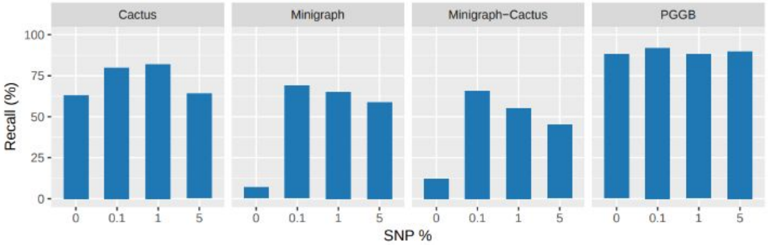
$cov_{path} >= min_{cov}$ — false

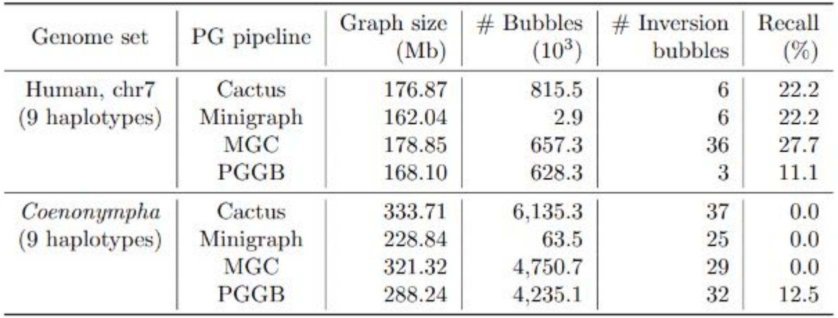
true

$cov_{aln} >= min_{cov}$

true

Annotated inversions





| Genome set | PG pipeline | Graph size (Mb) | # Bubbles ($10^3$) | # Inversion bubbles | Recall (%) |
|---|---|---|---|---|---|
| Human, chr7 (9 haplotypes) | Cactus | 176.87 | 815.5 | 6 | 22.2 |
| | Minigraph | 162.04 | 2.9 | 6 | 22.2 |
| | MGC | 178.85 | 657.3 | 36 | 27.7 |
| | PGGB | 168.10 | 628.3 | 3 | 11.1 |
| *Coenonympha* (9 haplotypes) | Cactus | 333.71 | 6,135.3 | 37 | 0.0 |
| | Minigraph | 228.84 | 63.5 | 25 | 0.0 |
| | MGC | 321.32 | 4,750.7 | 29 | 0.0 |
| | PGGB | 288.24 | 4,235.1 | 32 | 12.5 |

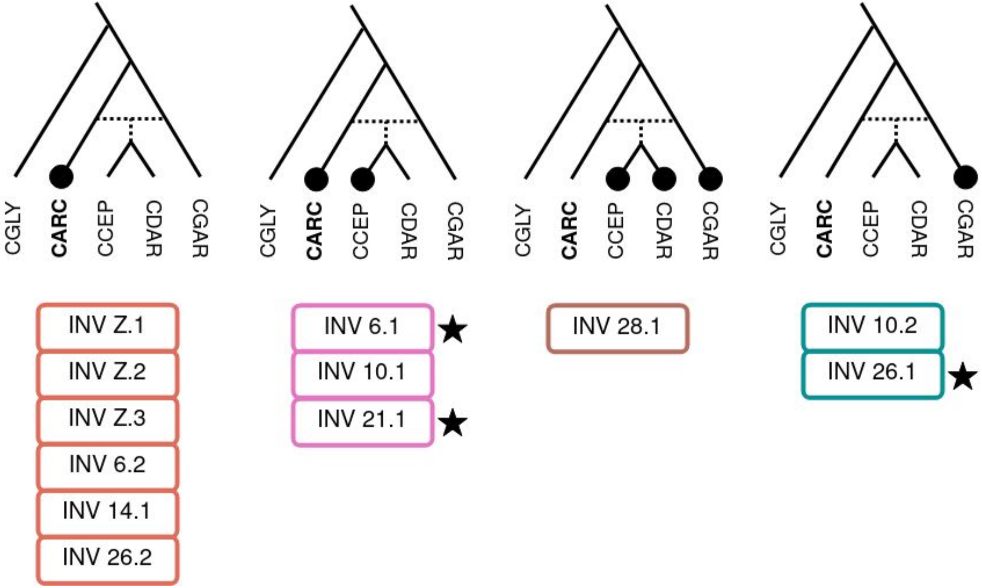
**Investigating the topological motifs of inversions in pangenome graphs**

Sandra Romain, Siegfried Dubois, Fabrice Legeai, Claire Lemaitre

**doi:** https://doi.org/10.1101/2025.03.14.643331

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

# Inversion’s presence-absence patterns

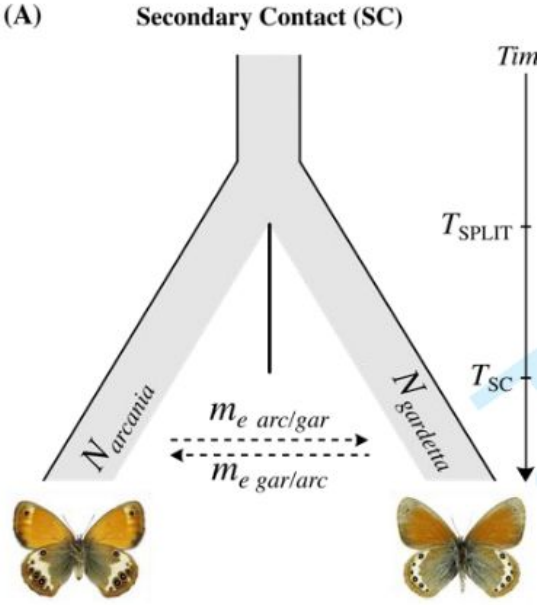


The structural form inherited by the hybrid species was more frequently transmitted by *C. gardetta (pattern 1 & 3)*.

Pattern 2 identified in only one of the parents were inherited differently between the two hybrid species.

Inversions 6.1 and 21.1 were found heterozygous in *C. macromma (CCEP)*, and inversion 26.1 was found heterozygous in *C. gardetta*.

# Barriers to gene flow between the parental lines



**(B)**

| | Fastsimcoal *(95% CI)* | | DILS *(95% CI)* | |
|---|---|---|---|---|
| ***Best scenario*** | **SC** | $\Delta_{AIC}$ AM = 313,230<br>$\Delta_{AIC}$ IM = 302,806<br>$\Delta_{AIC}$ SI = 3,980,332 | **SC** | proba migration = 0.982<br>proba SC vs IM = 0.903 |
| *Divergence times* | | | | |
| $T_{SPLIT}$ | 1,793,459 *(1,694,115 - 1,892,802)* | | 1,930,105 *(1,688,758 - 2,182,893)* | |
| $T_{SC}$ | 360,240 *(293,122 - 427,358)* | | 498,562 *(381,457 - 602,492)* | |
| *Effective sizes* | | | | |
| $N$ arcania | 2,955,432 *(2,913,267 - 2,997,597)* | | 2,659,735 *(2,296,304 - 3,189,278)* | |
| $N$ gardetta | 1,748,050 *(1,737,751 - 1,758,348)* | | 1,257,299 *(795,278 - 1,857,227)* | |
| *Migration rates* | | | | |
| $m_e$ arc/gar | 3.99e-07 *(0 - 1.2e-05)* | | 7.9e-07 *(6.37e-07 - 9.29e-07)* | |
| $m_e$ gar/arc | 9.07e-08 *(0 - 1.2e-04)* | | 2.9e-07 *(1.87e-07 - 4.33e-07)* | |

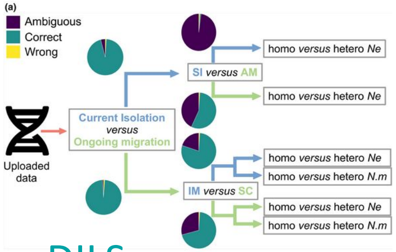
global pattern of gene-flow

# Barriers to gene flow between the parental lines

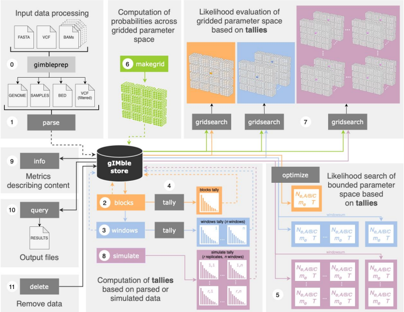
17 individuals of C.arcania

19 individuals of C. gardetta

(13-25.2x - avg : 18x)



DILS



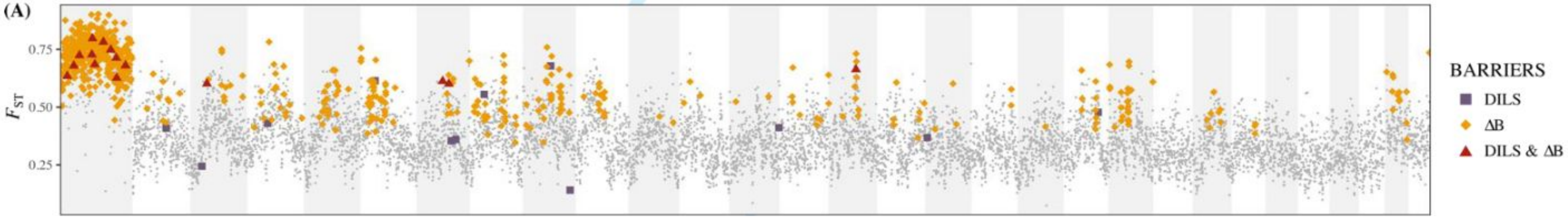
gIMble ΔB

654 genomic loci

6.6 % of the genome is impermeable to gene flow

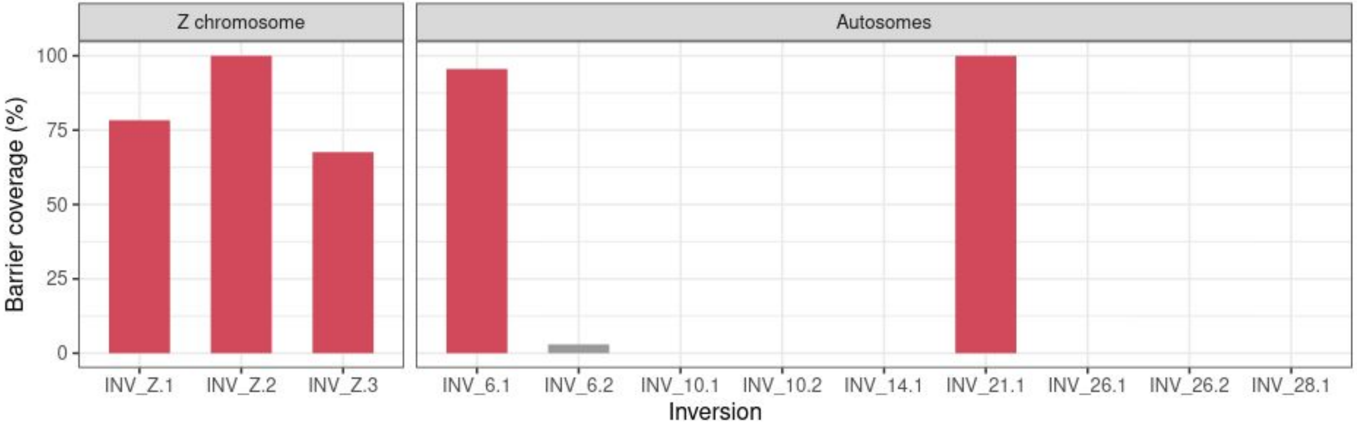
Many of the barriers are on the Z chromosome spanning 75% of its length.

substantial difference between the number of regions identified as barrier by the ΔB method (642 windows) and DILS (27 windows) with only 15 common windows



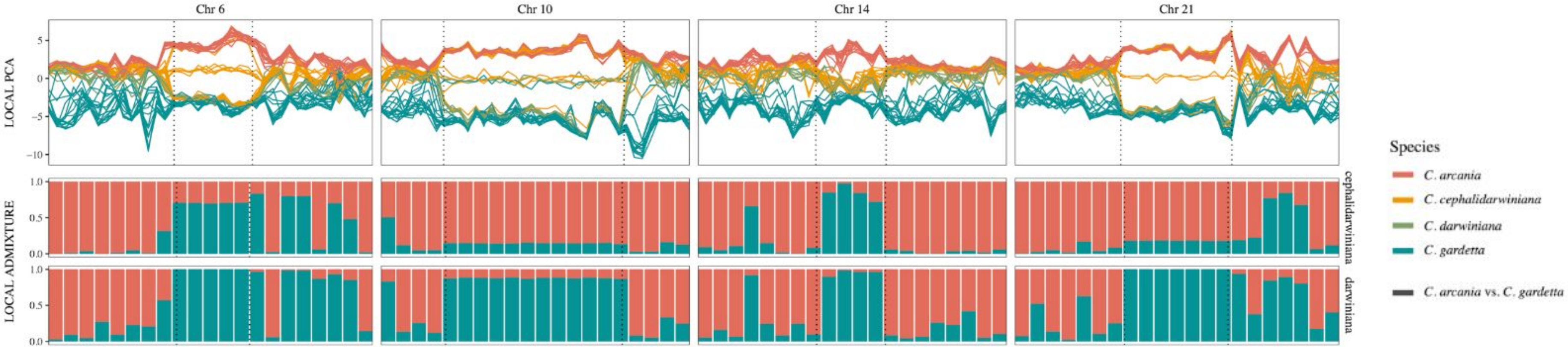
# Large inversions and barrier loci

- all three inversions located on the Z chromosome show more than 60 % of genomic barriers.
- Autosomal inversions 6.1 and 21.1, both previously reported as heterozygous in C. cephalidarwiniana
- other inversions and other autosomal loci being almost all bare of such barriers



# Genetic inheritance of the inversion

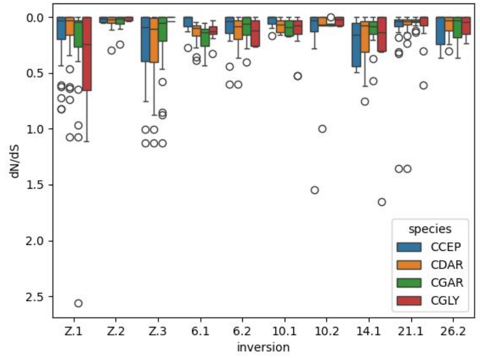
## Use of the previous resequencing of *C. arcania* and *C.gardetta* +16 C. *darwiniana* and 9 *C.macromma*.

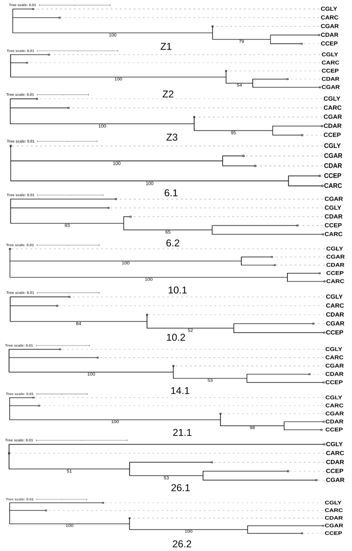


Globally, about 75% inherited from *arcania* and 25% from gardetta. Z is almost completely inherited from C.gardetta, while autosomes are inherited from C.arcania

Autosomal CDAR inversions 6.1, 10.1, 10.2, 14.1, 21.1 are inherited from gardetta
Autosomal CCEP inversions 6.1, 10.2 and 14.1 are inherited from gardetta

# And the functions





- High quality gene annotation of *C. arcania* was done using EGAPx with 2 RNASeq libraries
- Projection of the annotation achieved with TOGA (with make-lastz-chains)
- Alignments with MACSE + ClipKit

- No particular difference in dN/dS, neither in the trees.
- Lack of candidate gene but one manually validated  Olfactory receptor

Need to integrate more data!

# Acknowledgments

GENSCALE

Genscale INRIA Rennes
*Sandra Romain*
Claire Lemaitre

LECA CNRS Grenoble
*Thibaut Capblancq*
Laurence Després

CENTRE D'ECOLOGIE
FONCTIONNELLE
& EVOLUTIVE

CEFE CNRS Montpellier
Mathieu Joron

DivAlps

cnrs

inria

INRAE
la science pour la vie, l'humain, la terre

anr