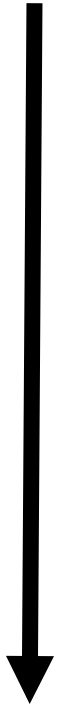
Detecting and analysing genomic structural variants

Prepared by Claire Mérot
with A. Tigano & M. Wellenreuther
Physalia Courses

RECAP: Forms of genetic variation

Sequence

1. Single base-pair changes – point mutations (SNPs)
2. Change in Copy Number Variants (CNVs)
   - Deletions
   - Duplications
3. Change in chromosomal location
   - Translocations
   - Fusions
4. Change in orientation
   - Inversions
5. Changes in chromosome number (e.g. aneuploidy)



Cytogenetics

Using sequencing to detect SV

- Massive parallel sequencing drastically reduced costs and enabled population-wide sequencing

- In 2020s: many tools available with advantages and drawbacks

  - Short-reads (illumina)

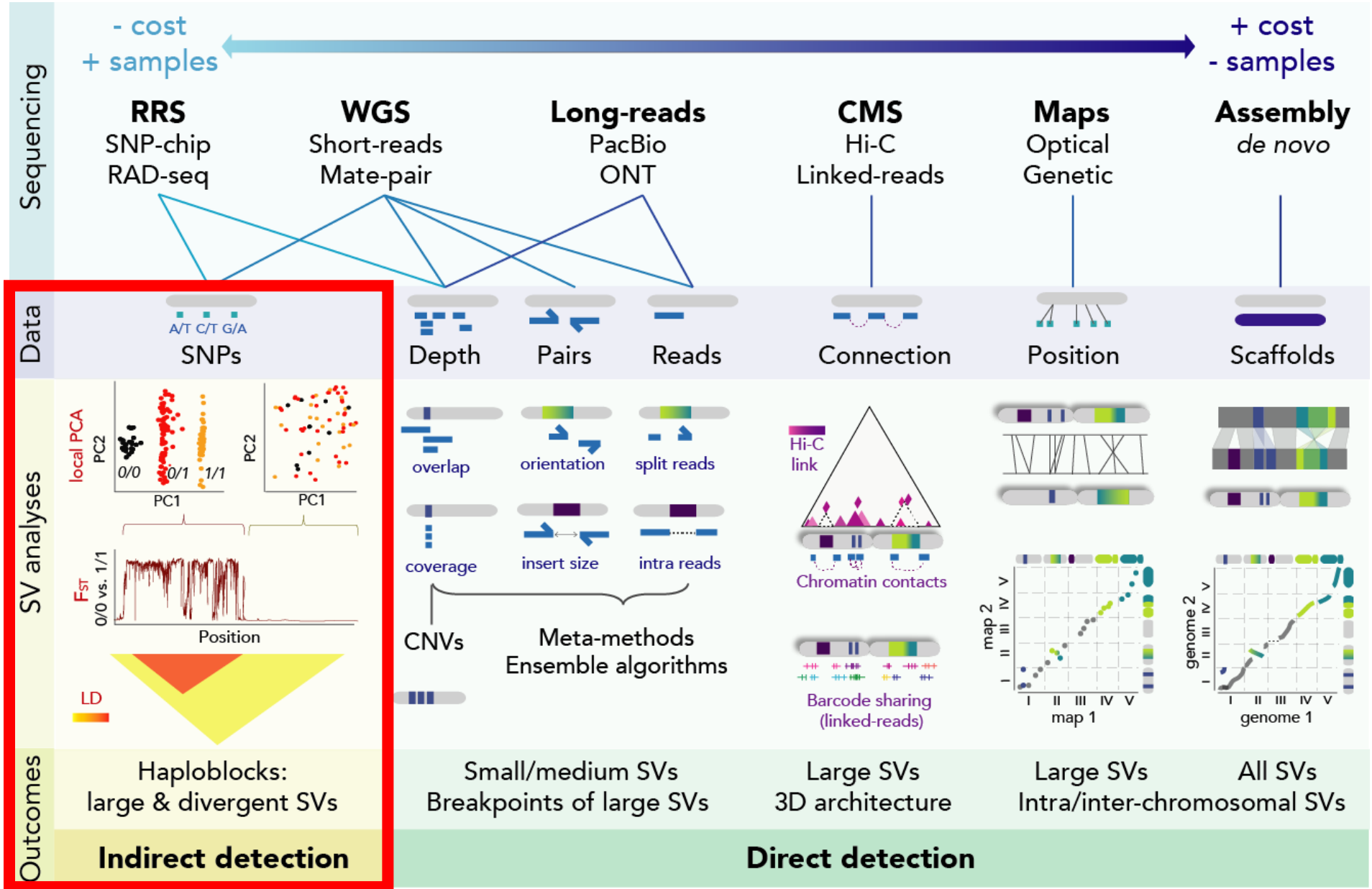
- high single-nucleotide accuracy & paired-end
- underrepresentation of high-GC regions

  - Long-reads (PacBio/Nanopore)

- Higher error rate (~15%) and single-end (but see PacBio Hi-Fi! & improvement of ONT)
- Longer sequences (~1-50kb >> 100s of kb sometimes for nanopore)

  - Emerging technologies (Hi-C, 10x/synthetic long-reads, optical mapping)

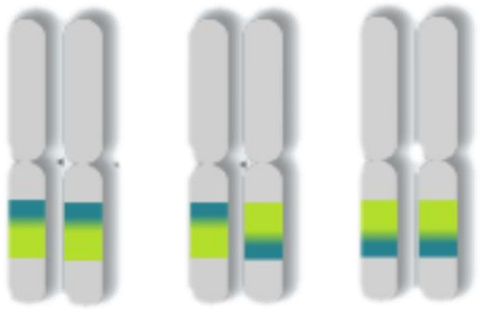
  ⇒ How can we exploit this amazing resource to detect SV?


Sequencing
- cost
+ samples
+ cost
- samples
RRS
SNP-chip
RAD-seq
WGS
Short-reads
Mate-pair
Long-reads
PacBio
ONT
CMS
Hi-C
Linked-reads
Maps
Optical
Genetic
Assembly
de novo
Data
A/T C/T G/A
SNPs
Depth
Pairs
Reads
Connection
Position
Scaffolds
SV analyses
local PCA
PC2
PC1
0/0
0/1
1/1
FST
0/0 vs. 1/1
Position
LD
overlap
orientation
split reads
coverage
insert size
intra reads
CNVs
Meta-methods
Ensemble algorithms
Hi-C
link
Chromatin contacts
Barcode sharing
(linked-reads)
map 2
map 1
genome 2
genome 1
Outcomes
Haploblocks:
large & divergent SVs
Small/medium SVs
Breakpoints of large SVs
Large SVs
3D architecture
Large SVs
Intra/inter-chromosomal SVs
All SVs
Indirect detection
Direct detection

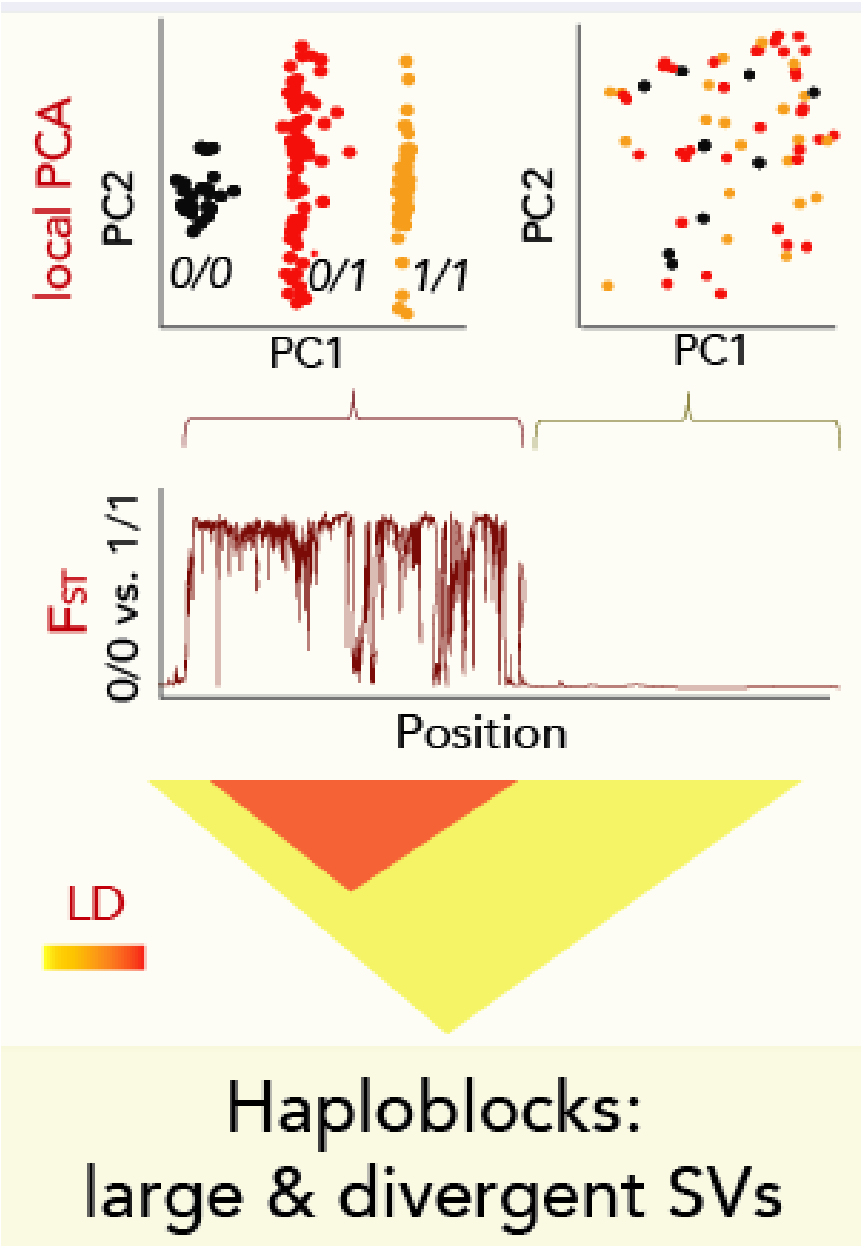
Indirect detection

It is based on the idea that large rearrangements (like an inversion) block recombination.

Hence when they are polymorphic in a species, they appear as large non-recombining haploblocks with two (or more) divergent haplotypes.



0/0 0/1 1/1



Indirect detection

- Using population genomics data:
  - Many samples
  - Many SNPs (from short-reads, SNPchip, RAD-seq….)

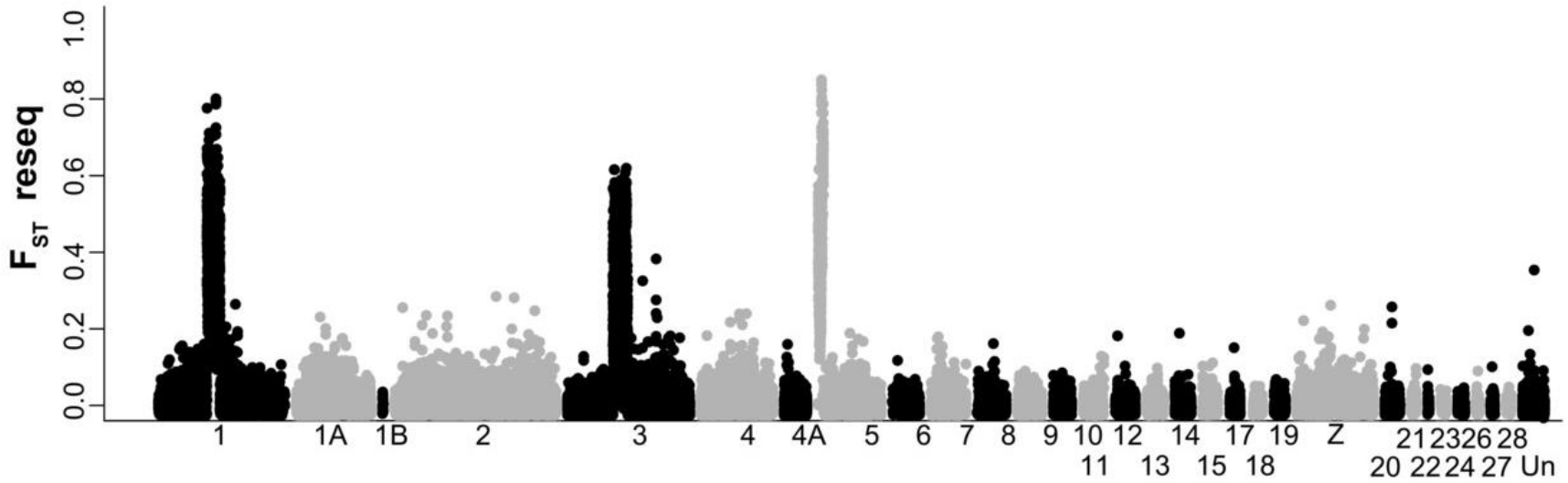
- Able to detect chromosomal rearrangements if they are:
  - Large (> 100 kb)
  - Polymorphic
  - Divergent

⇒ Typically good to detect large inversions (or fusions, large blocks without recombination)…

- Tools:
  - Fst - Linkage disequilibrium - PCA & clustering

Indirect detection : Fst/islands of divergence

Genetic differences between willow warbler migratory phenotypes are few and cluster in large haplotype blocks



-> chromosomal rearrangements preventing recombination?

-> linked selection? Hitch-hikng around specific loci?

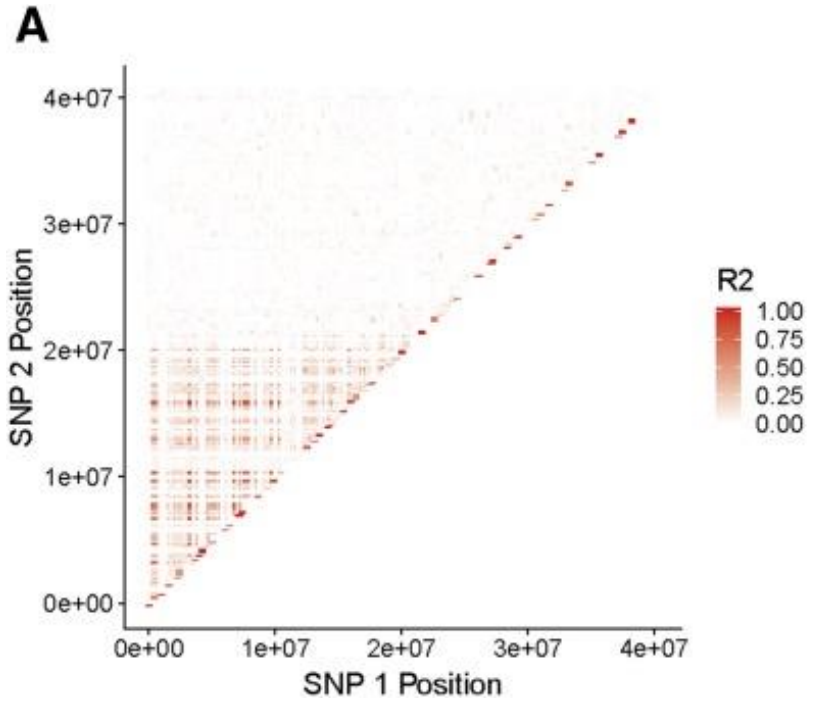
Lundberg et al. Evolution Letters 2017

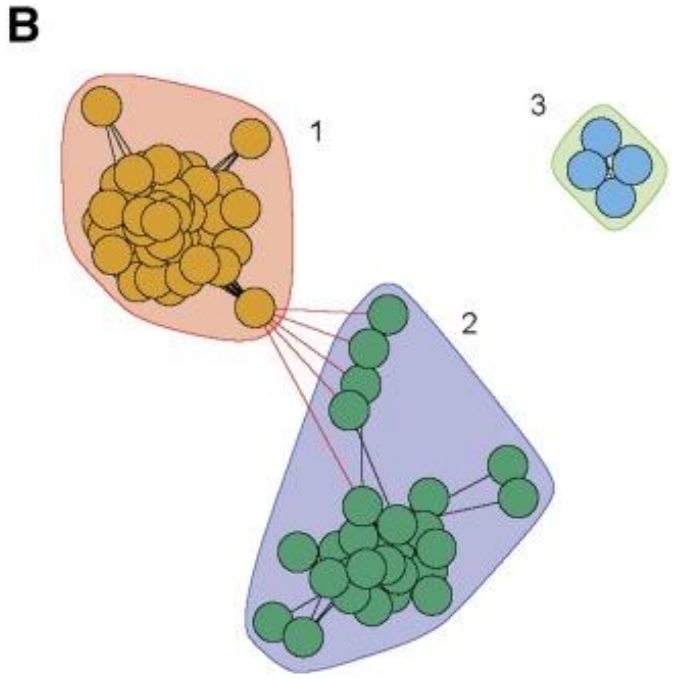
Indirect detection : LD networks

SNPs within an inversion will be in high linkage disequilibrium and belong to one cluster of LD

-> can be applied without reference genome

-> any methods to get SNPs





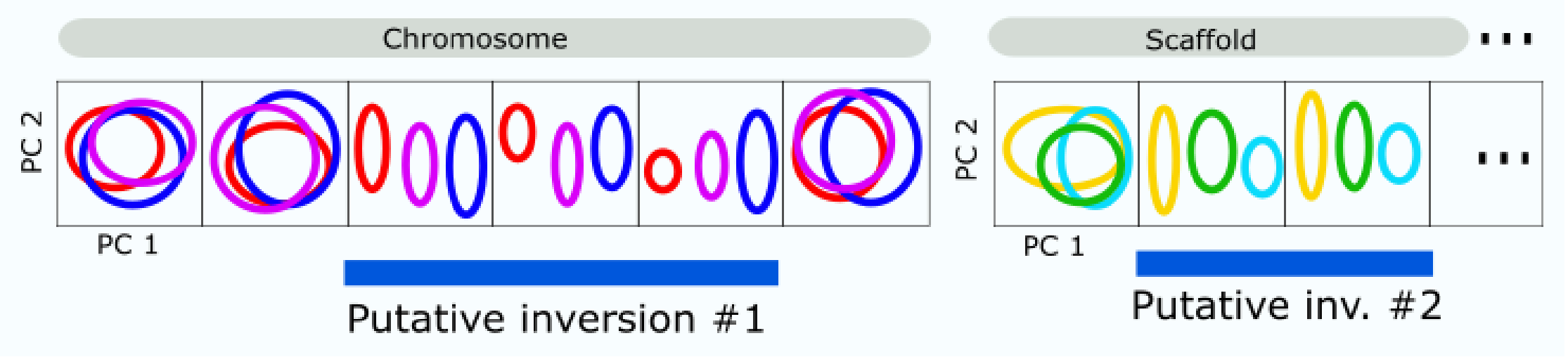
McKinney et al 2020. *G3, 10*(5), 1553–1561. https://doi.org/10.1534/g3.119.400972

Ldna Package:
Kemppainen P, Knight CG, Sarma DK, et al. *Mol Ecol Resour*. 2015;15(5):1031-1045. https://doi.org/10.1111/1755-0998.12369

Detection of 17 inversions in Littorina:
Faria et al. *Mol Ecol*. 2019; 28: 1375– 1393. https://doi.org/10.1111/mec.14972

Indirect detection : Local PCA

A PCA performed on SNPs belonging to an inversion will usually display three clusters while PCA outside will show no clustering



Lostruct Package:
Li & Ralph. 2019 Genetics https://doi.org/10.1534/genetics.118.301747

Detection of 7 inversions in Helianthus with Rad-seq data:
Huang et al. *Mol Ecol*. 2020. https://doi.org/10.1111/mec.15428

Indirect detection :

Indirect methods typically identifies non-recombining blocks of haplotypes which may or may not be due to an inversion.

What else can haploblocks be?

- Recent introgression?
- Linked selection?

⇒ Breakpoints should start eroding with gene flow

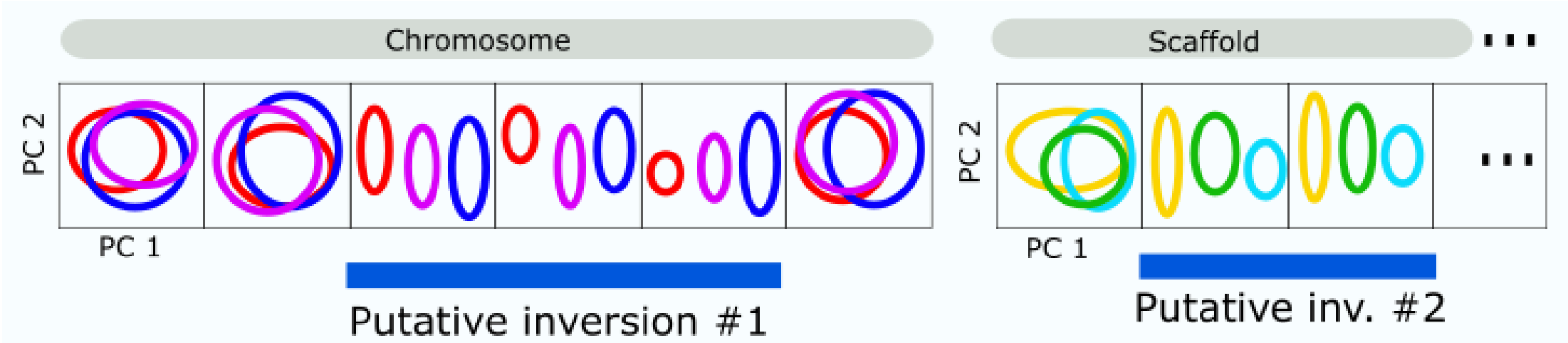
⇒ Perhaps less likely when blocks are very large (>1MB)

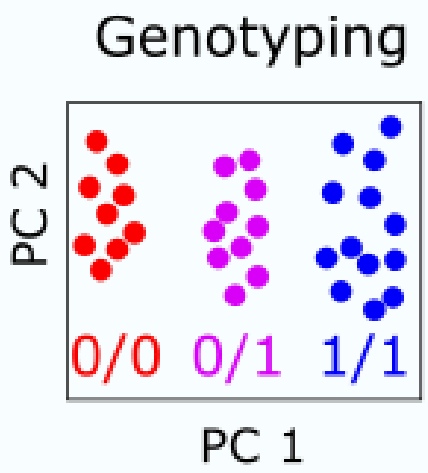
- Low-recombination regions?

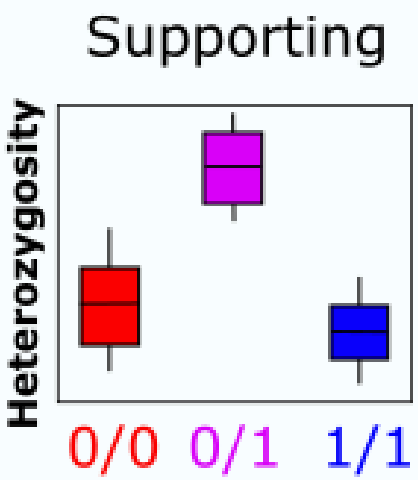
⇒ LD should be observed in all clusters

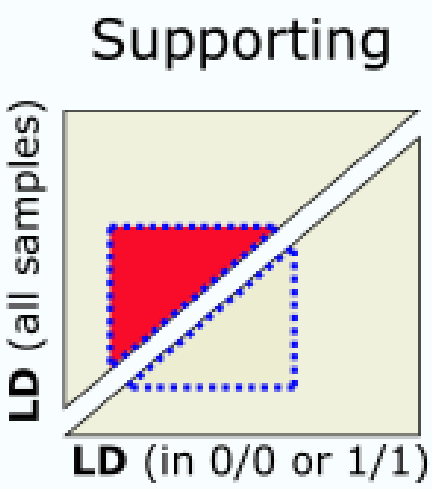
Indirect detection :

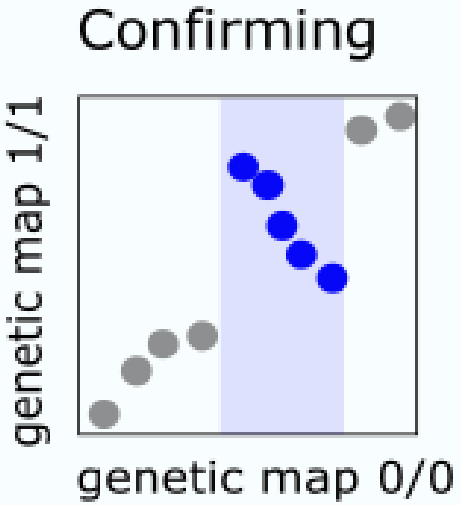
How can we support that an haploblock is an inversion?







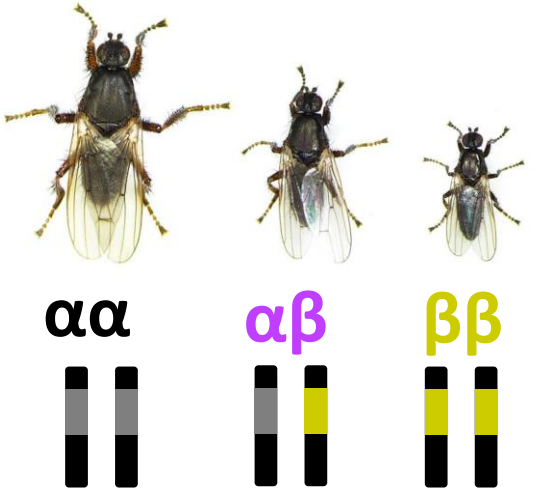


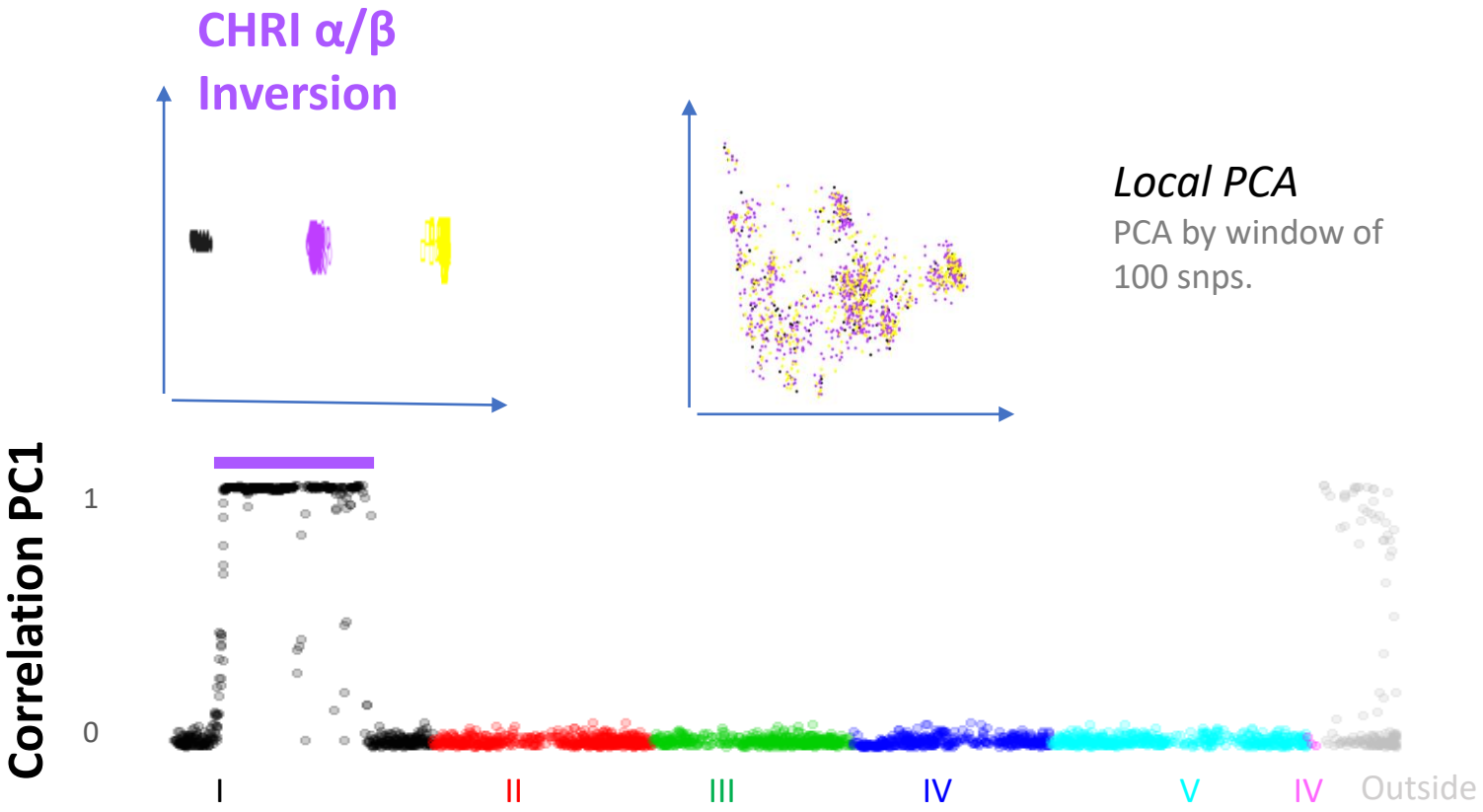


Mérot, 2020, MolEcol https://doi-org/10.1111/mec.15500

Indirect detection: Case study in the seaweed fly *Coelopa frigida*

- Whole-genome sequencing at low coverage for 1,446 flies



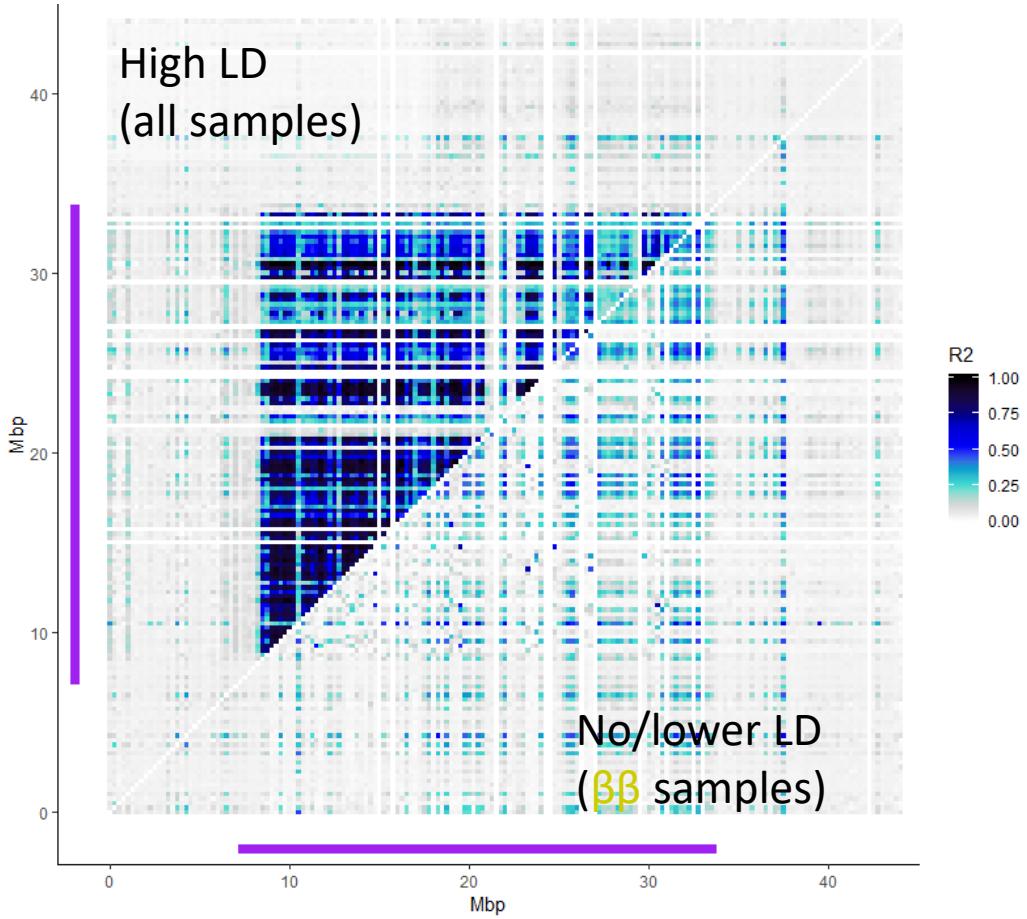


CHR-I inversion
27Mb
11% genome
16,5% of SNPs
1500 genes

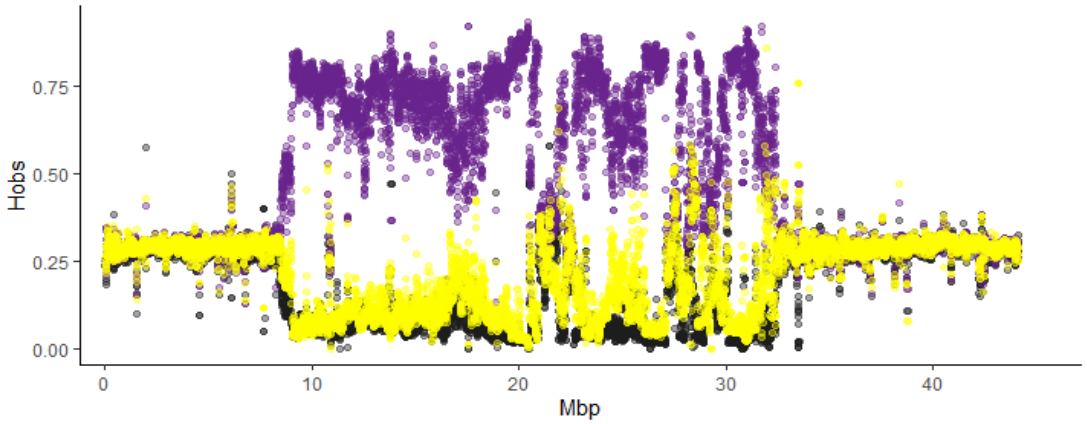
Indirect detection: Case study in the seaweed fly *Coelopa frigida*

Exploration of the haploblock/inversion

Chr I

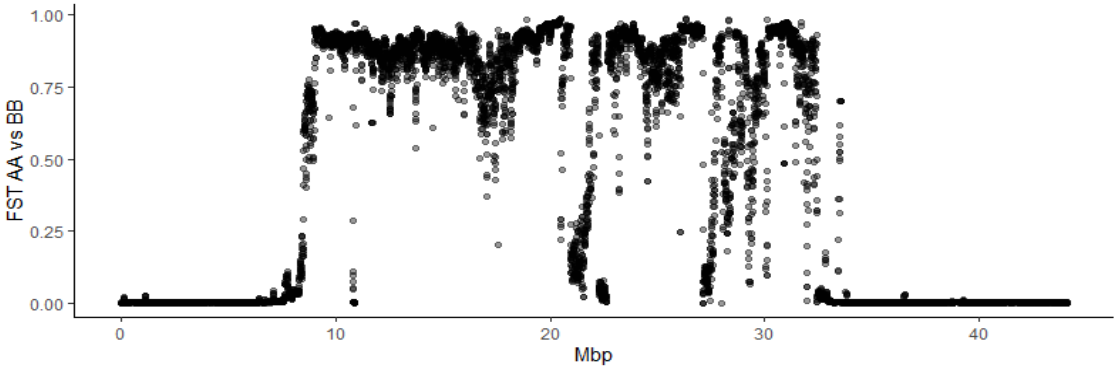


Higher observed heterozygosity in αβ than in αα or ββ



Chr I

High FST differentiation between αα and ββ



Chr I

Indirect detection of SV :

Advantages:

- Same data as population genomics (even RAD-seq)
- Genotyping inversions accross large datasets

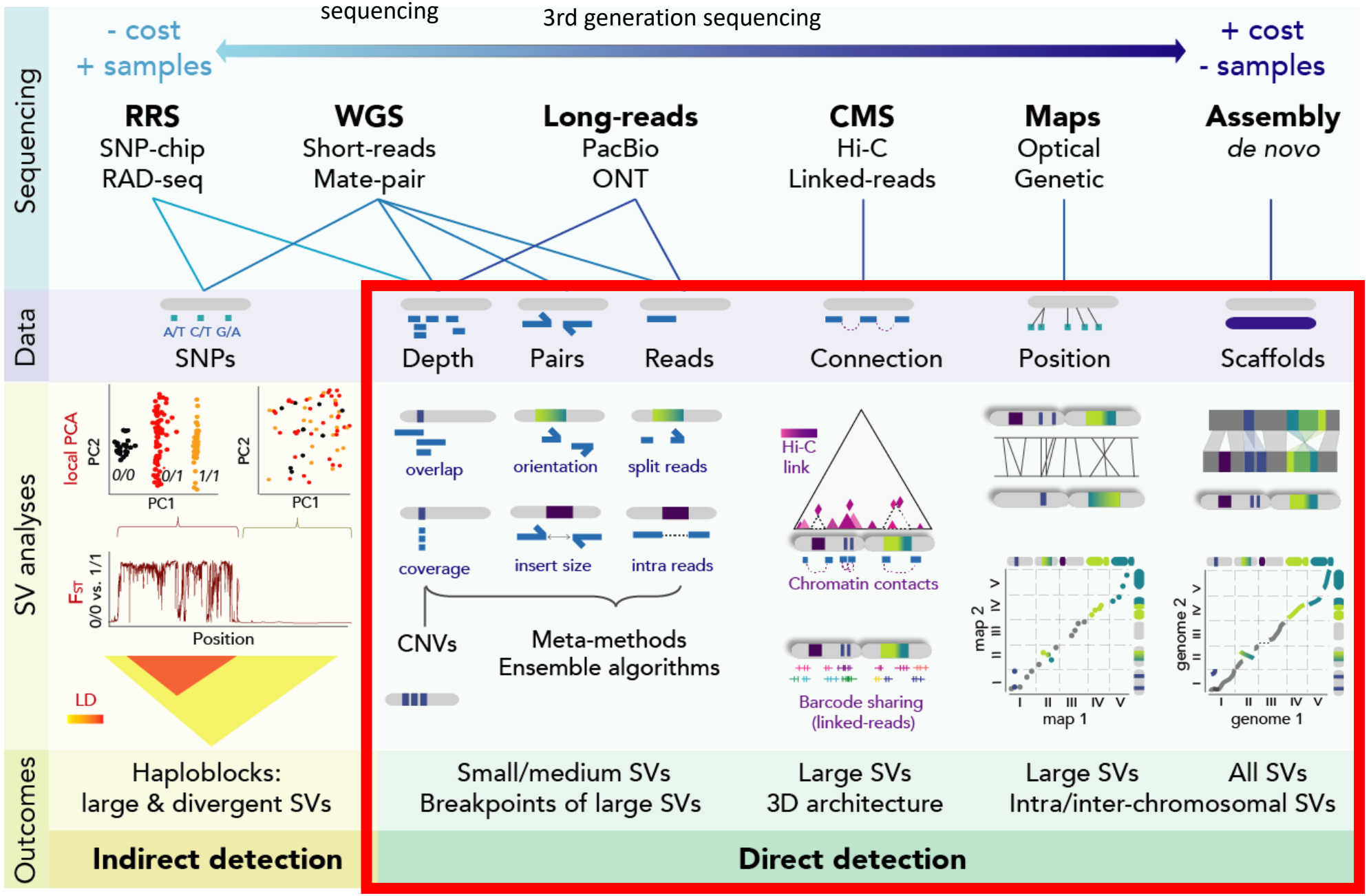
Drawbacks:

- Better confirmed with direct detection methods (cytogenetics or sequence analysis)
- Easier with a reference genome

2nd generation sequencing

CMS= connected molecule strategy

3rd generation sequencing



2nd generation sequencing : Short-reads (illumina)

- SVs are usually inferred indirectly from aberrant short-read alignments, such as an unexpected depth of coverage or inconsistent orientation or distance between the alignment of paired-end reads

- Low costs of short-reads allow population-wide sequencing

⇒ SV can be genotyped in many individuals

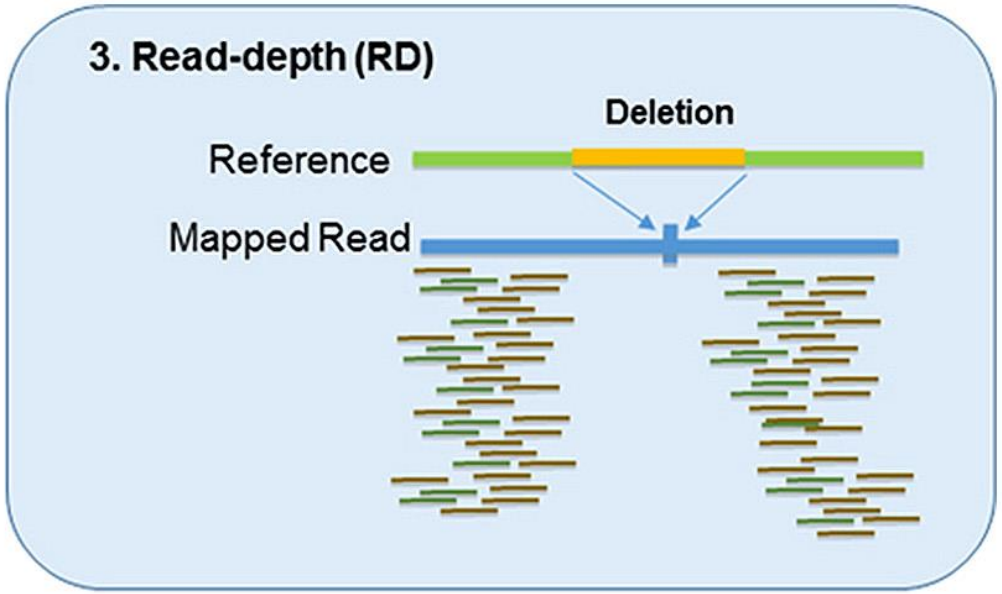
- Short-reads (100-150 bp) single or paired-end

⇒ Limited range of Sv that can be detected by this technology

Direct detection : with read depth

- Detect CNVs (duplications, indels)

- Applicable to SNP-chip, RAD-seq, WGS (short & long reads)



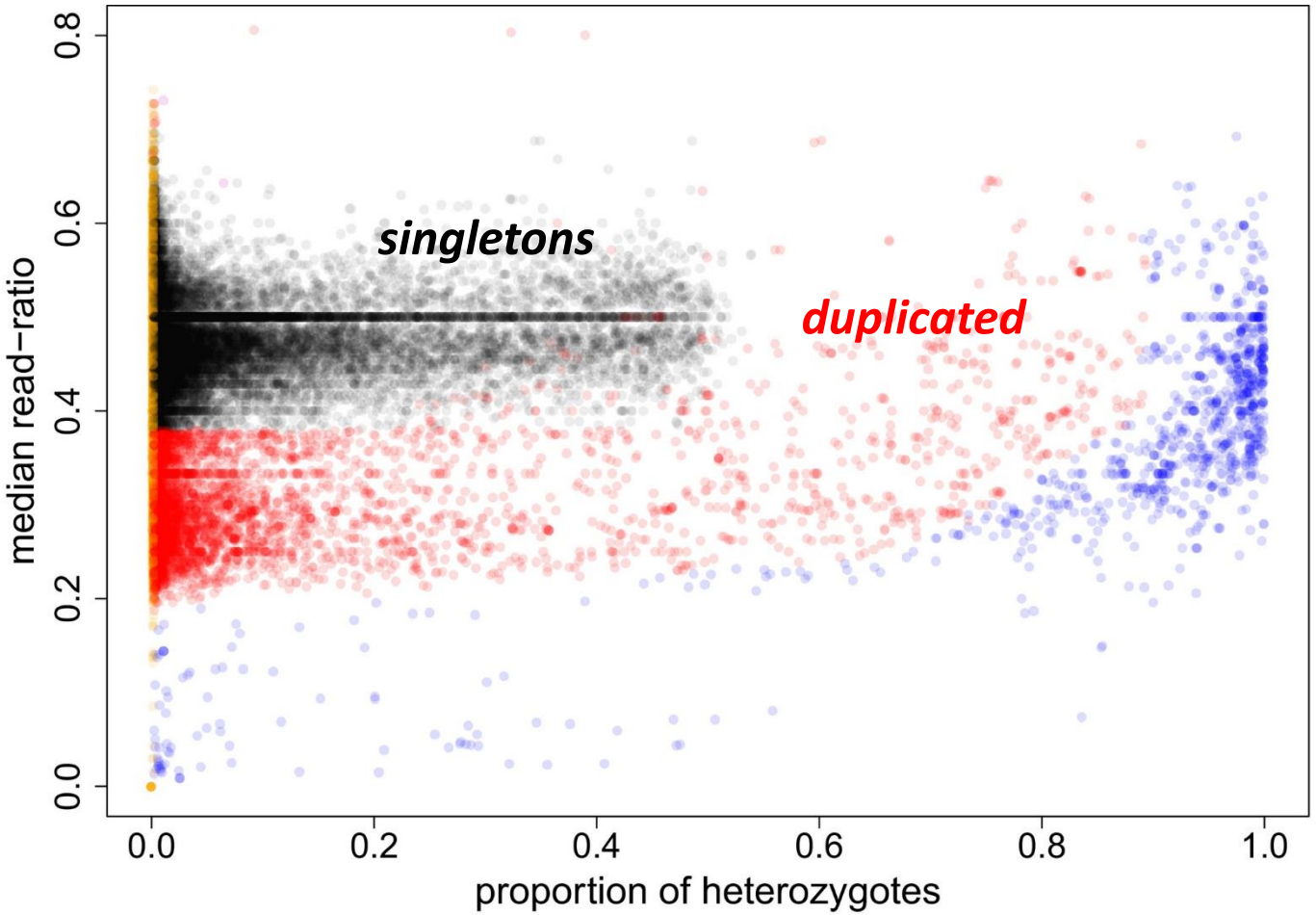
Direct detection : with read depth

Adding allelic information and heterozygote information…

⇒ Detect duplicated loci in RAD-seq

-> Filter them out for regular analysis

-> Keep them apart to analyse CNVs

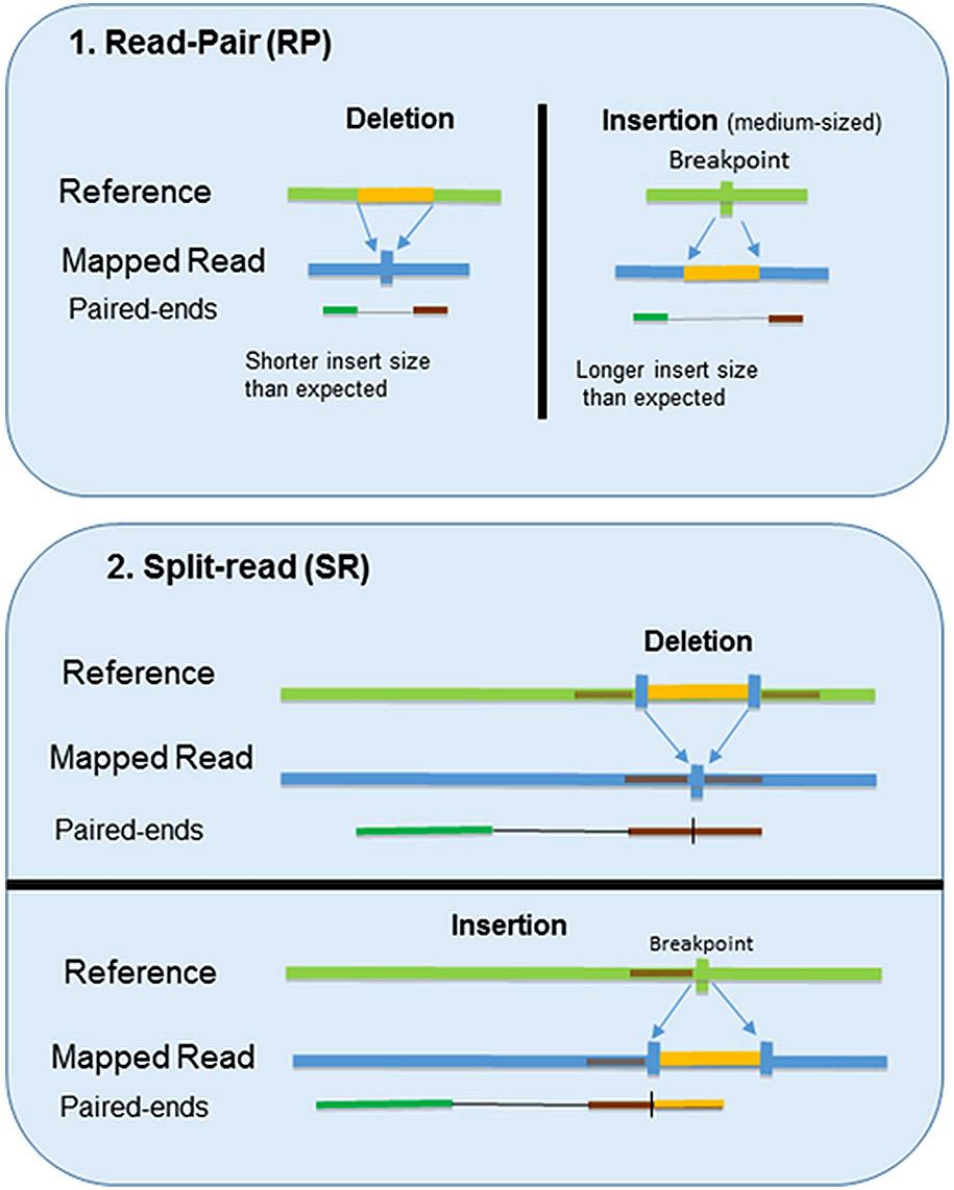


Dorant et al 2020. MolEcol https://doi.org/10.1111/mec.15565
McKinney,et al. 2017 MolEcol Ressources. https://doi.org/10.1111/1755-0998.126

Direct detection : with paired-read orientation & split-reads

This will detect shorts indels and breakpoints of duplications, translocations or inversions

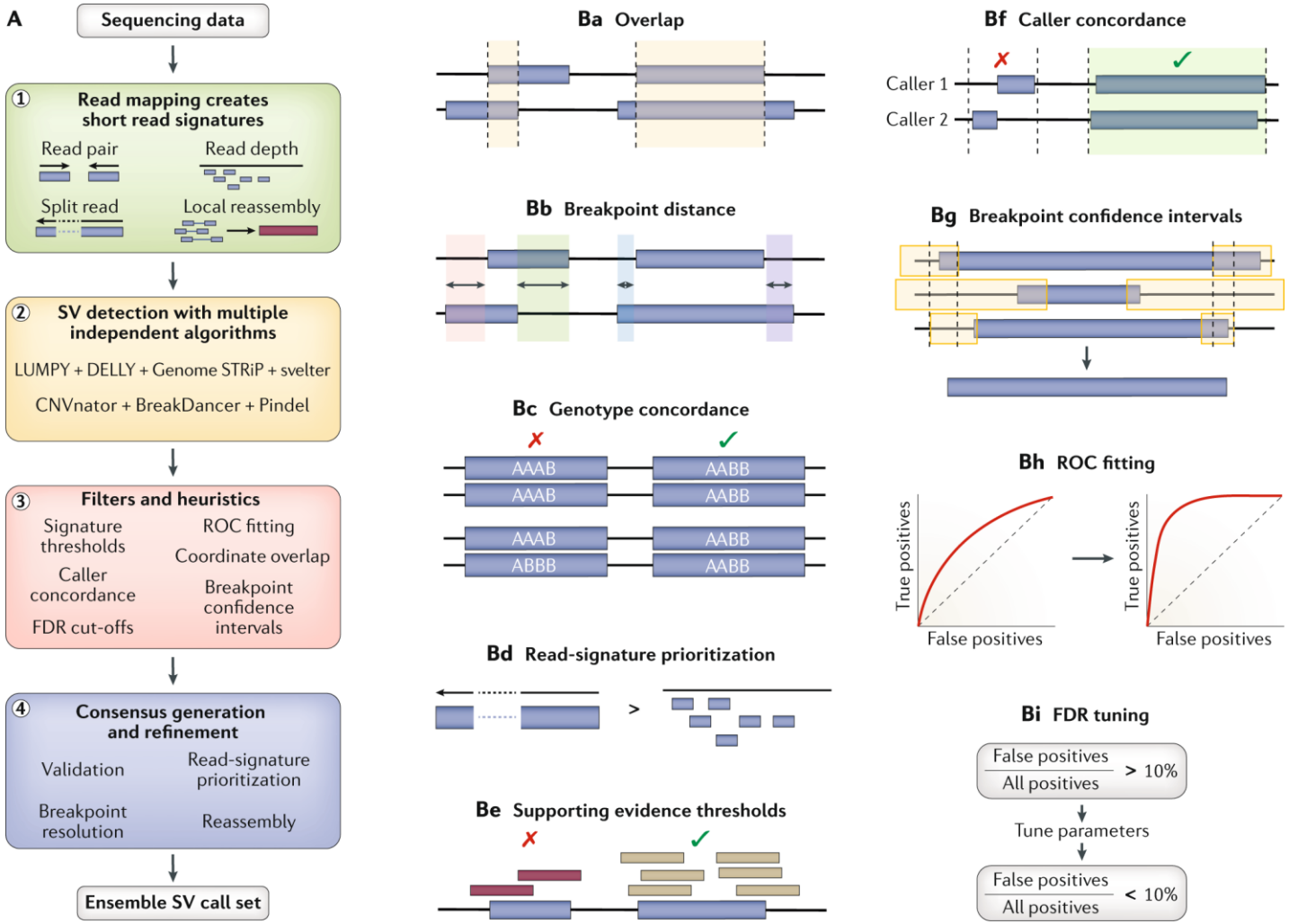
Most-used tools:
Delly, Manta, Lumpy,GRIDSS



Direct detection : Ensemble methods

- Combining
- read depth,
- paired-reads distance
- paired-end orientation
- split-reads.

- Merge the output of several tools to improve confidence



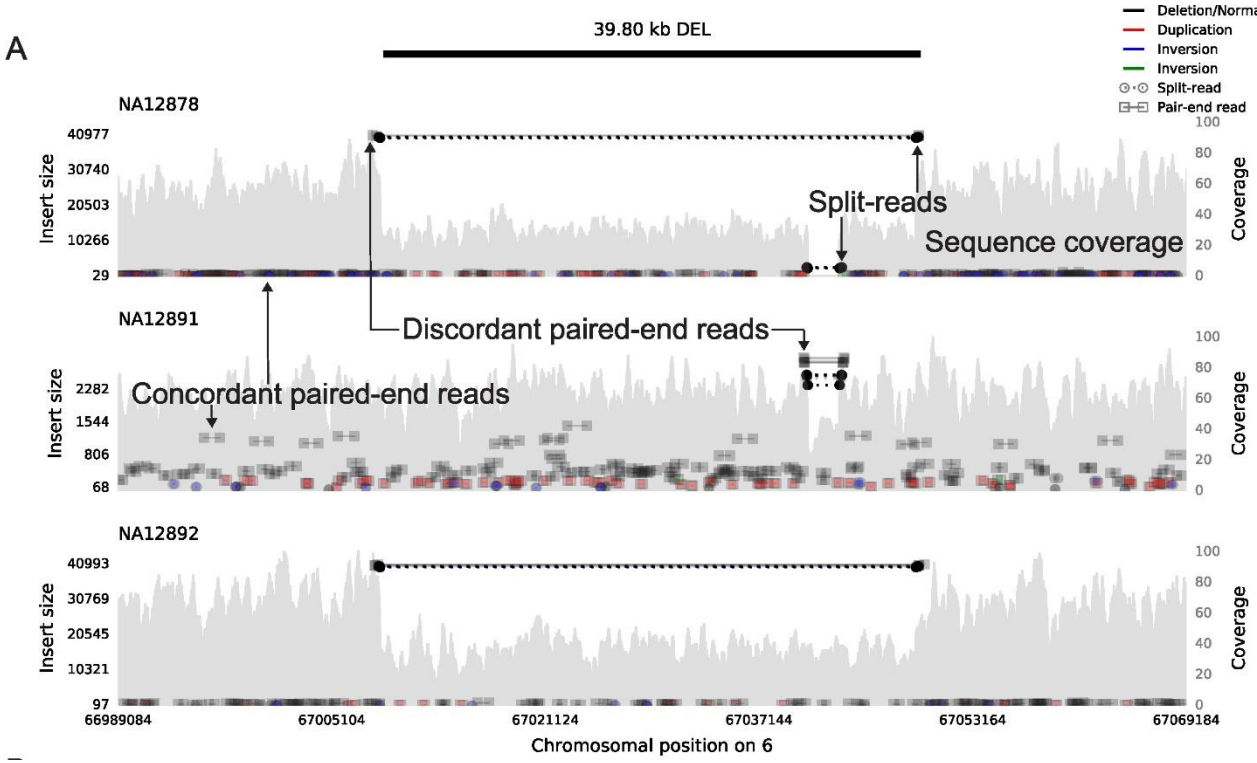
Ho, S.S., Urban, A.E. & Mills, R.E. Structural variation in the sequencing era. *Nat Rev Genet* **21,** 171–189 (2020). https://doi.org/10.1038/s41576-019-0180-9

Direct detection : based on short-reads

Lots of false positive!!

Manual curation with SV-plaudit in 492 Atlantic Salmon

« *The overall estimated false discovery rate was **0.91** with 149,491 out of 65,116 of calls which had low confidence* »
Bertolotti et al, 2020 BioRxiv https://doi.org/10.1101/2020.05.16.099614



Belyeu et al, 2018
GigaScience https://doi.org/10.1093/gigascience/giy064

Recent improvements:

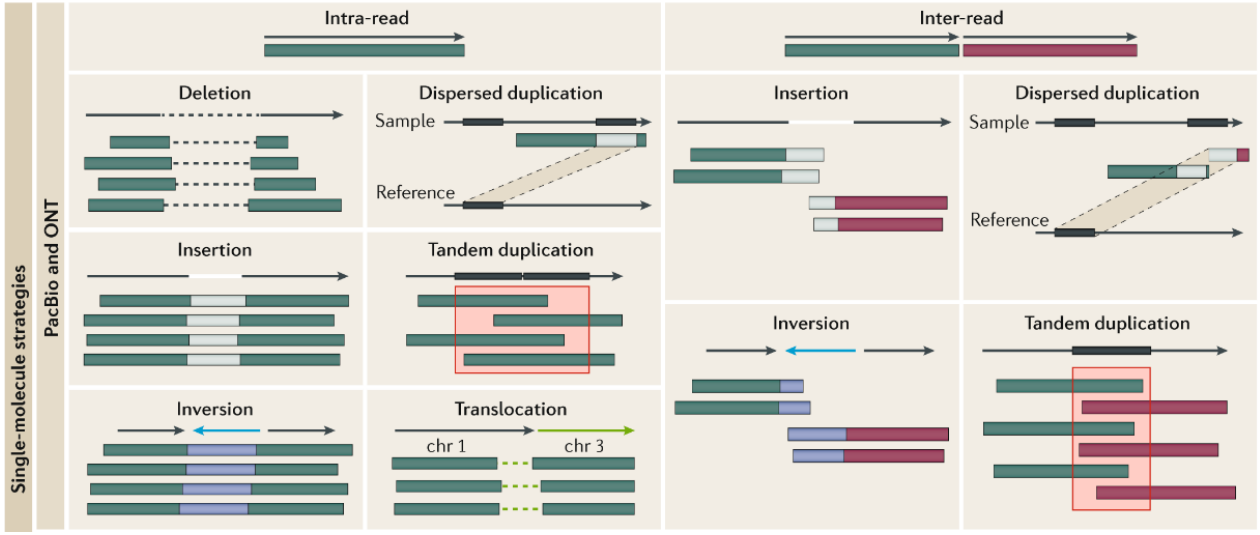
- graph-based approaches
- population-scale genotyping of SV

Eggertsson *et al. Nat Commun* **10,** 5402 (2019).
https://doi.org/10.1038/s41467-019-13341-9

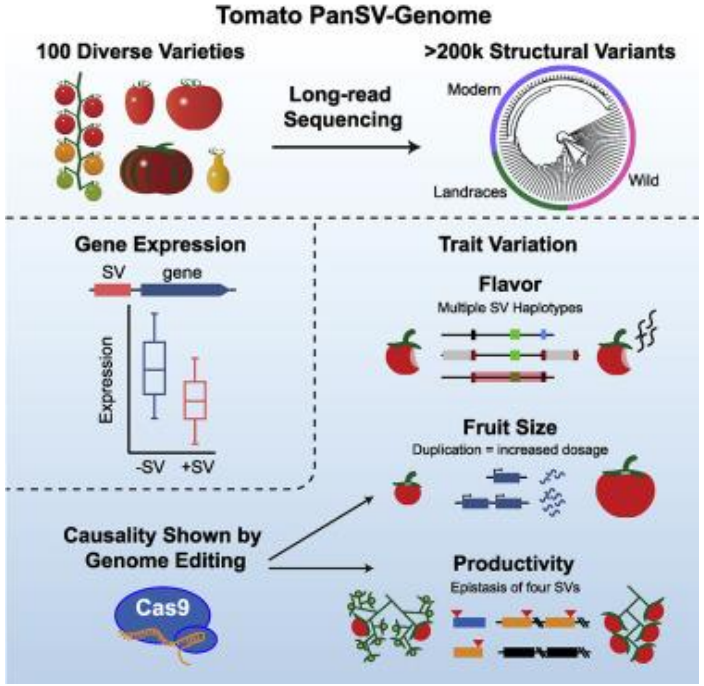
Direct detection : using long-reads

- Long reads will allow to detect longer SV, will cover the highly-repetitive regions at breakpoints, etc.

- But they are expensive, we cannot (routinely) genotype SV at population



Ho et al . *Nat Rev Genet* (2020). https://doi.org/10.1038/s41576-019-0180-9



https://doi.org/10.1016/j.cell.2020.05.021

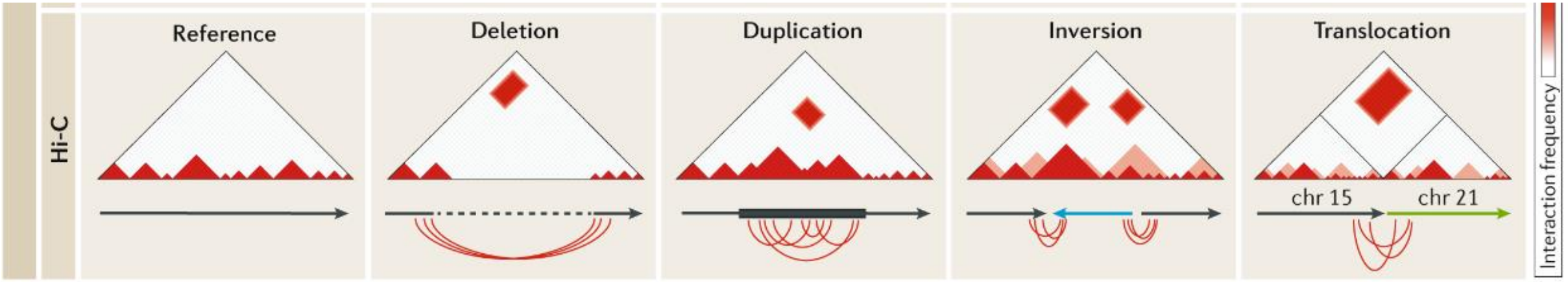
Alonge et al 2020 Cell

Direct detection: Connected-molecule strategies

Hi-C (DoveTail)

- Analyze the spatial organization of chromatin in a cell
- Output the interactions between fragments of DNA

⇒ Allows detecting medium to large rearrangements



H et al . *Nat Rev Genet* (2020). https://doi.org/10.1038/s41576-019-0180-9

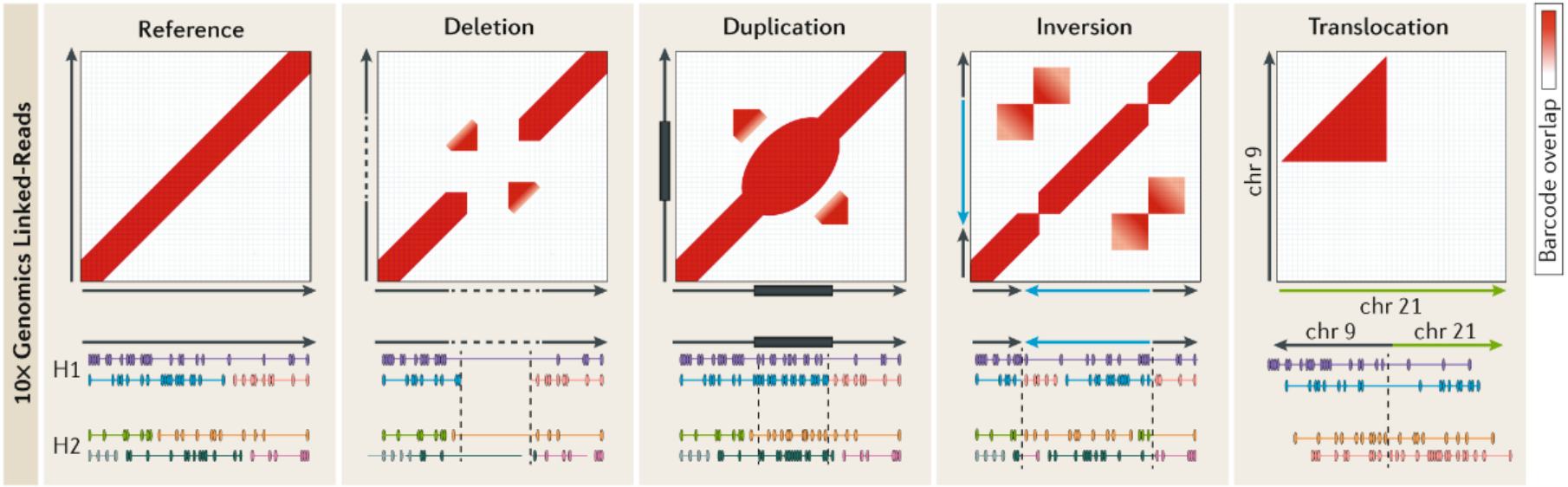
Direct detection: Connected-molecule strategies

« Linked-reads » - « synthetic long-reads » - « haplotagging »
(10xGenomics (deprecated), Tell-seq, Stlfr, Emerging in-house)

Meier et al PNAS 2021
https://doi.org/10.1073/pnas.2015005118

- Long DNA fragments (50kb-100kb) are barcoded before short –reads sequencing

⇒ Sequences that are physically close share the same barcodes



H et al . *Nat Rev Genet* (2020).
https://doi.org/10.1038/s41576-019-0180-9

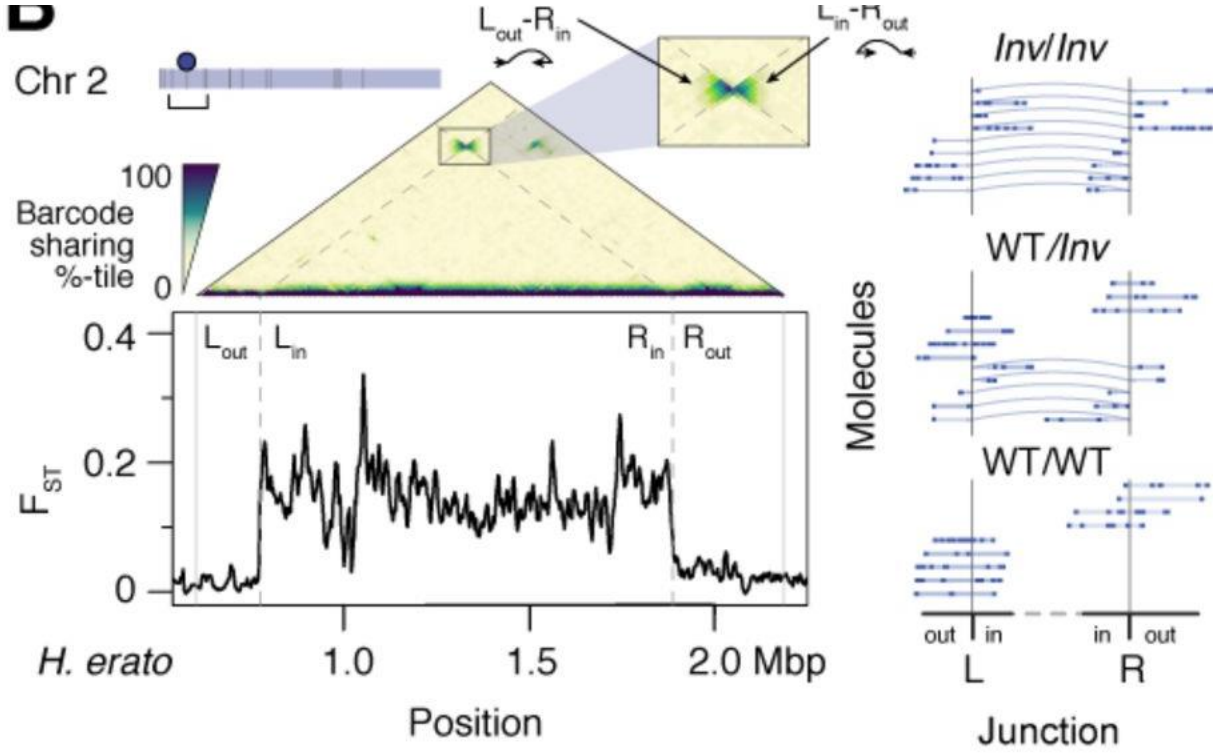
Direct detection: Connected-molecule strategies

Linked-reads

⇒ Medium and large inversions & indels

- Example:

Inversion detection in *Heliconius* butterflies



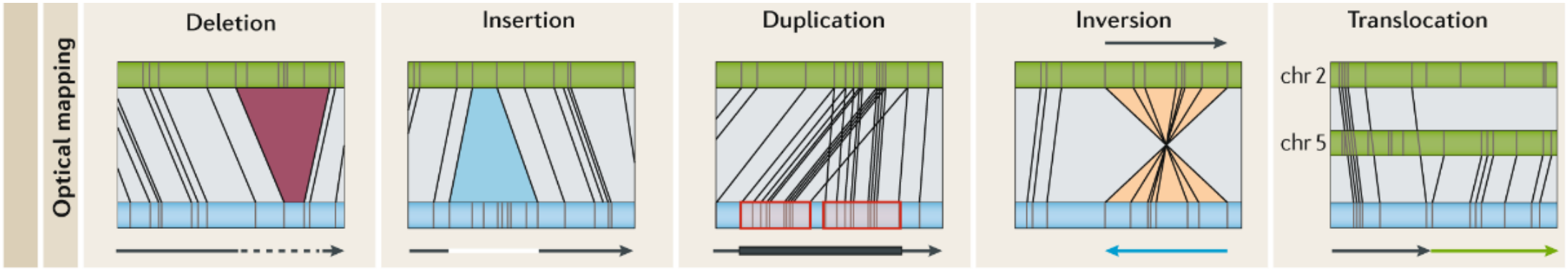
Meier et al PNAS 2021
https://doi.org/10.1073/pnas.2015005118

Direct detection: genetic maps

Optical maps (BioNano)

- Maps the location of restriction enzyme sites along the chromosomes

⇒ Good for detecting large rearrangements encompassing several sites



H et al . *Nat Rev Genet* (2020). https://doi.org/10.1038/s41576-019-0180-9

Direct detection: genetic maps

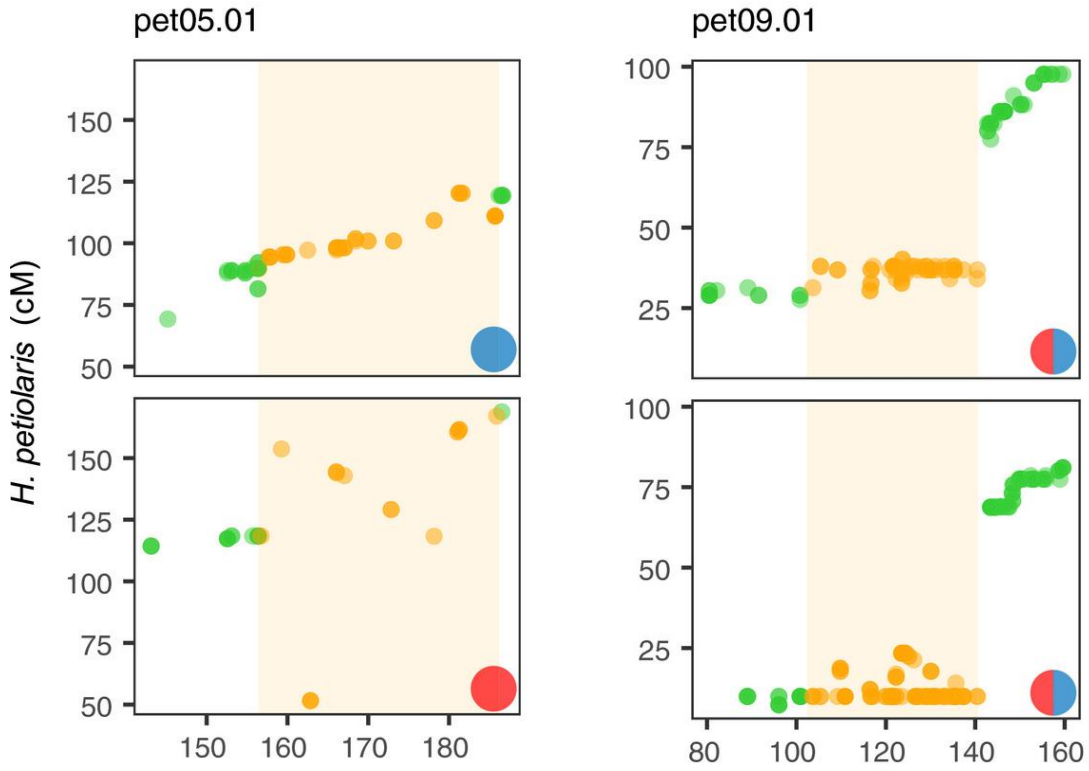
Linkage maps (based on families)

- compare marker position between families or between one family and reference genome
- Easy even on very divergent species

⇒ will detect large rearrangements, including inter-chromosomal fusion, translocation, etc

Homozygotypic parents
-> order is inversed

Heterozygotypic parents
-> no recombination



Huang et al 2020 MolEcol
https://doi-org/10.1111/mec.15428

Direct detection: genome comparison

Except for highly repetitive regions, assembly-based SV identification is accurate but expensive due to the requirement of high sequence coverage.

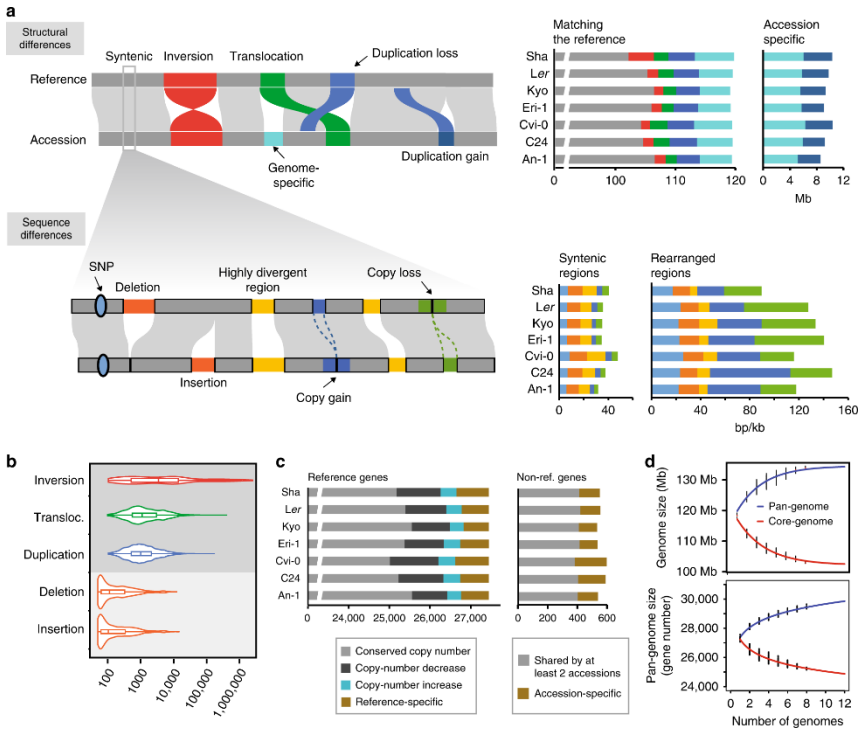
⇒ Will typically be done only on a limited number of samples

⇒ A field in progress!!

62 high-quality genome assemblies of Cacao trees

Hamala et al, 2021 PNAS
https://doi.org/10.1073/pnas.2102914118

7 high-quality genome assemblies of Arabidopsis thaliana

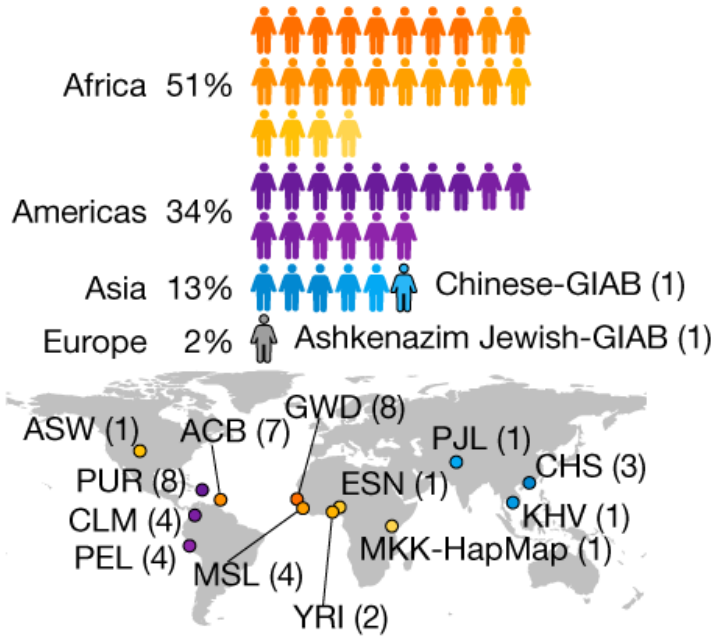


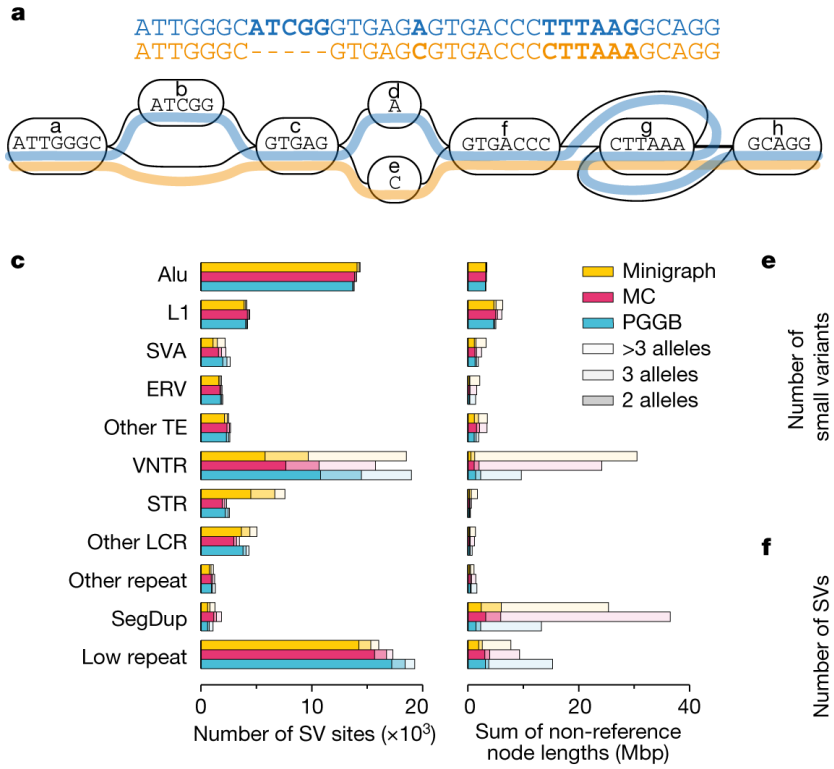
https://doi.org/10.1038/s41467-020-14779-y
Jiao & Schneeberger 2020 NatCom

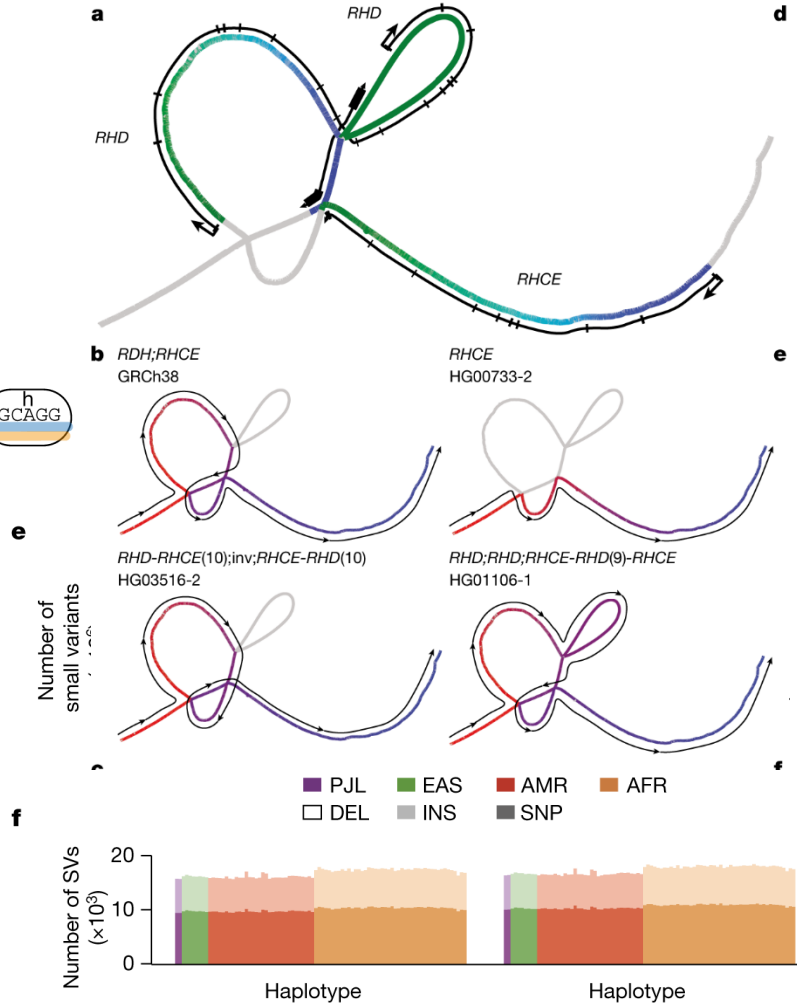
Direct detection: genome comparison

The new power of pan-genomes

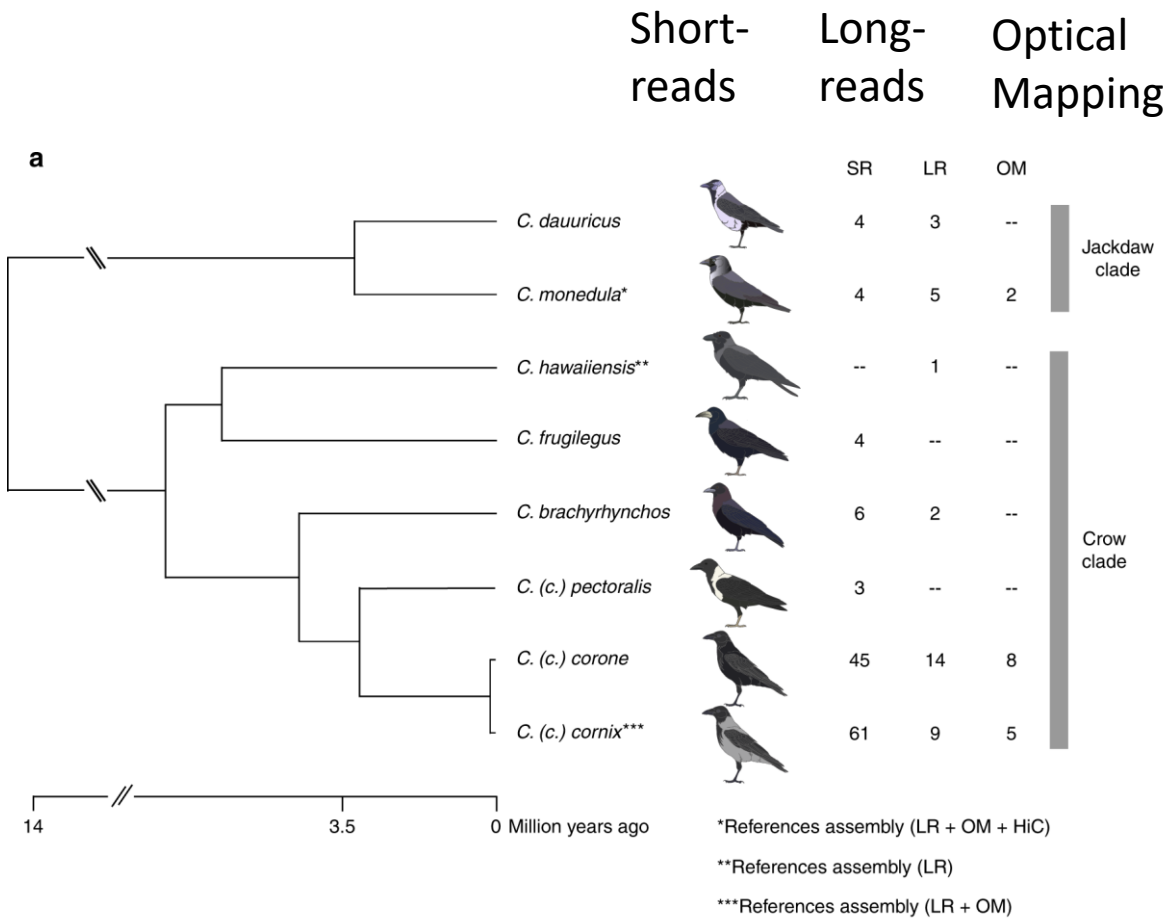
Liao, WW., Asri, M., Ebler, J. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023). https://doi.org/10.1038/s41586-023-05896-x







Direct detection : combining platforms



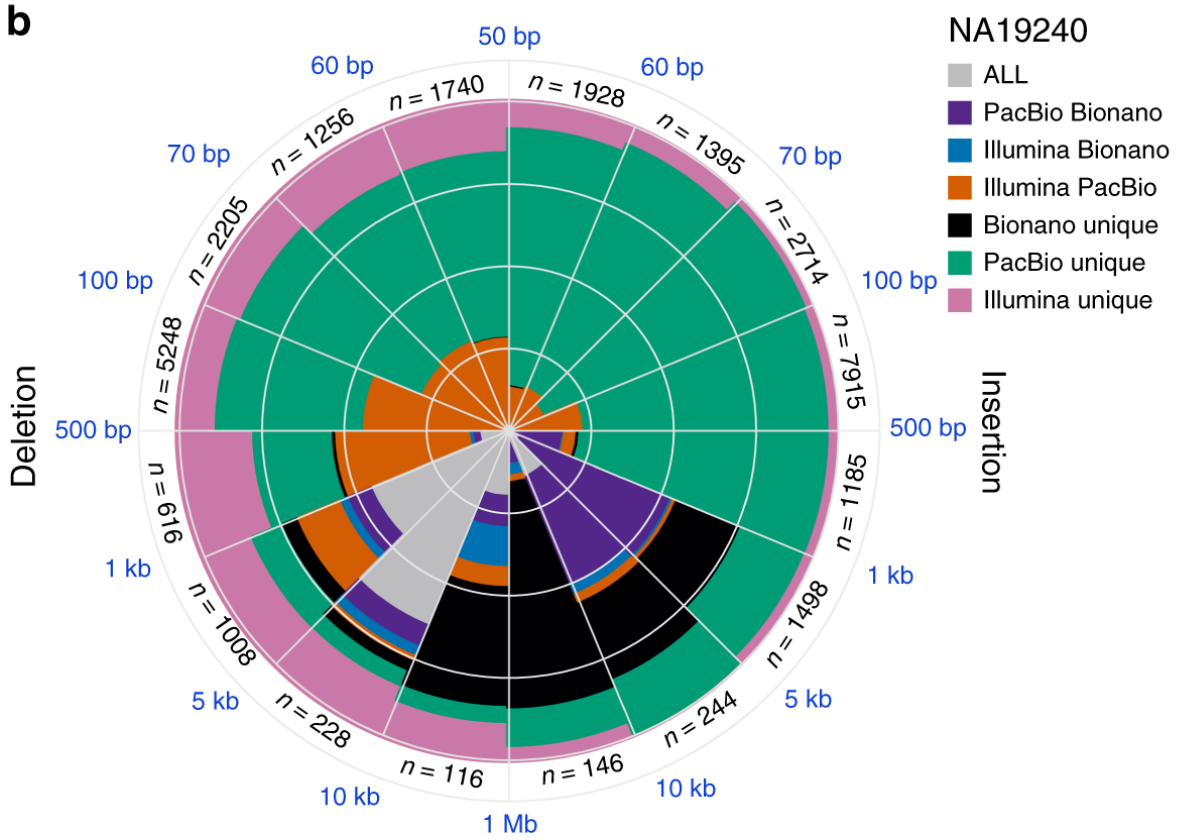
Long-reads/optical mapping
-> a few individuls per species

Short-reads
-> many individuals (pop genomics)

Weissensteiner et al *Nat Comm* https://doi.org/10.1038/s41467-020-17195-4

Direct detection : combining platforms

Different platforms detect indels of different sizes



In humans
Chaisson et al, 2019, Nat Comm
https://doi.org/10.1038/s41467-018-08148-z

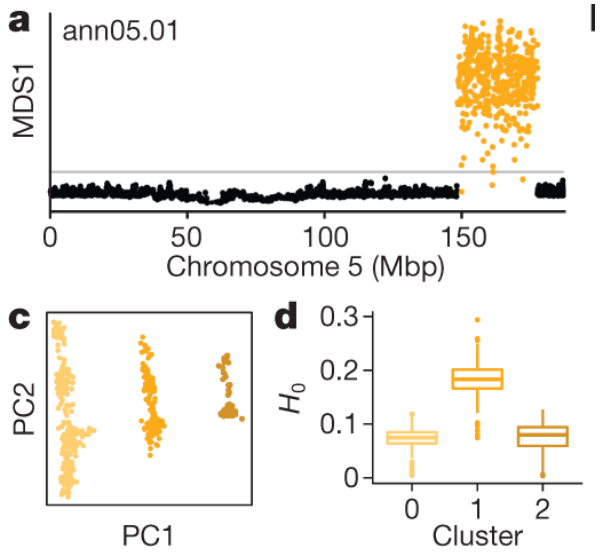
10kb->1MB: Bionano

20bp -> 1kb illumina + PacBio

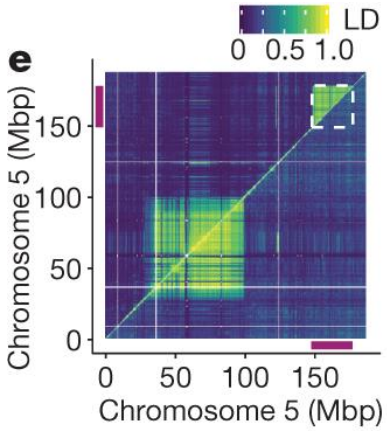
Short-reads only : just a fraction of Sv, more deletions tahn insertions

Direct detection : combining platforms

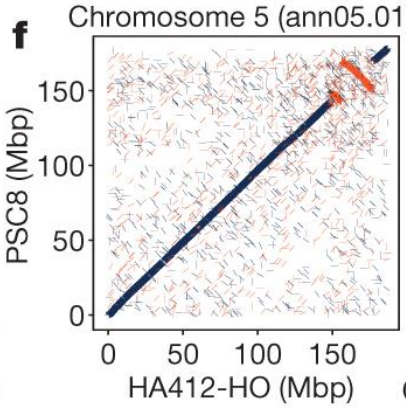
Indirect detection
(local PCA)

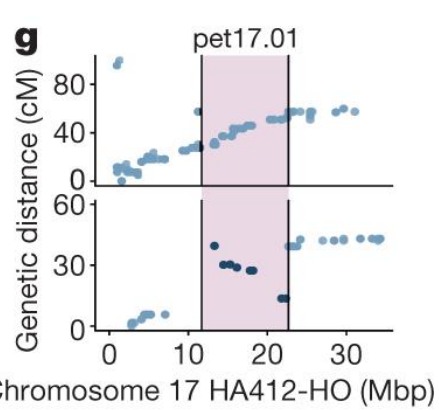


Indirect detection
(LD)

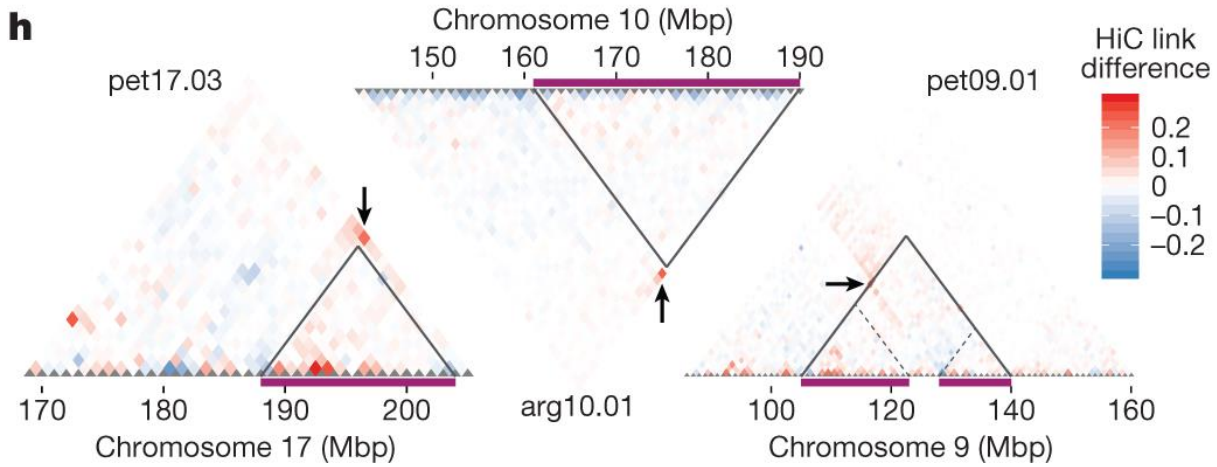


Direct detection
(genome comparison)





Direct detection
(genetic maps)



Direct detection
(Hi-C)

In Sunflowers
Todesco et al, 2020, Nature
https://doi.org/10.1038/s41586-020-2467-6

Characterize SVs in the Lake Whitefish?

1 Normal sp.

1 Dwarf sp.

Nanopore long-reads

Nanopore long-reads

Genome assembly (reference)

Genome assembly

Align assembly to the reference

Detect SVs with 3 tools & filter

SVIM-asm+ Minigraph-bubble + Syri

⇒A total of 89,909 SVs in the Dwarf assembly relative to the Normal

Mérot et al, MolEcol, 2022

https://doi.org/10.1111/mec.16468

Assembly
de novo

Scaffolds

Characterize SVs in the Lake Whitefish?

Norwegian University of Life Sciences

Mérot et al, MolEcol, 2022

https://doi.org/10.1111/mec.16468

Kristina S R Stenløkk

1 Normal sp.

1 Dwarf sp.

Nanopore long-reads

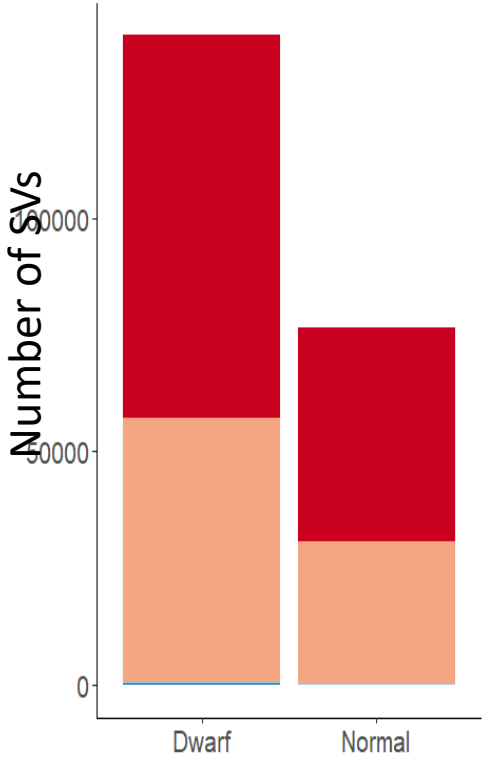
Nanopore long-reads

Genome assembly (reference)

Align LR to the reference

Detect SVs with 3 tools & Filter

Sniffles+ SVIM + nanovar



⇒A total of 194,861 SVs

⇒In the Dwarf and/or heterozygotes in the Normal

Characterize SVs in the Lake Whitefish?

Mérot et al,
MolEcol, 2022

https://doi.org/10.1111/mec.16468

1 Normal sp.

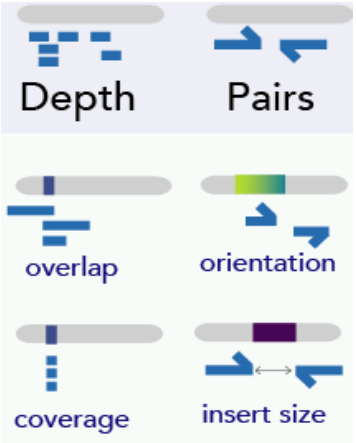
Nanopore long-reads

Genome assembly (reference)

Illumina short-reads (4-6x)

Align SR to the reference

	Normal sp.	Dwarf sp.
Cliff Lake	8	8
Indian Lake	7	9



Detect SVs with 3 tools
& Filter
Manta+Delly+Lumpy

⇒A total of 28,579 SVs

⇒Population-level variability

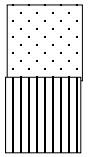
Characterize SVs in the Lake Whitefish?

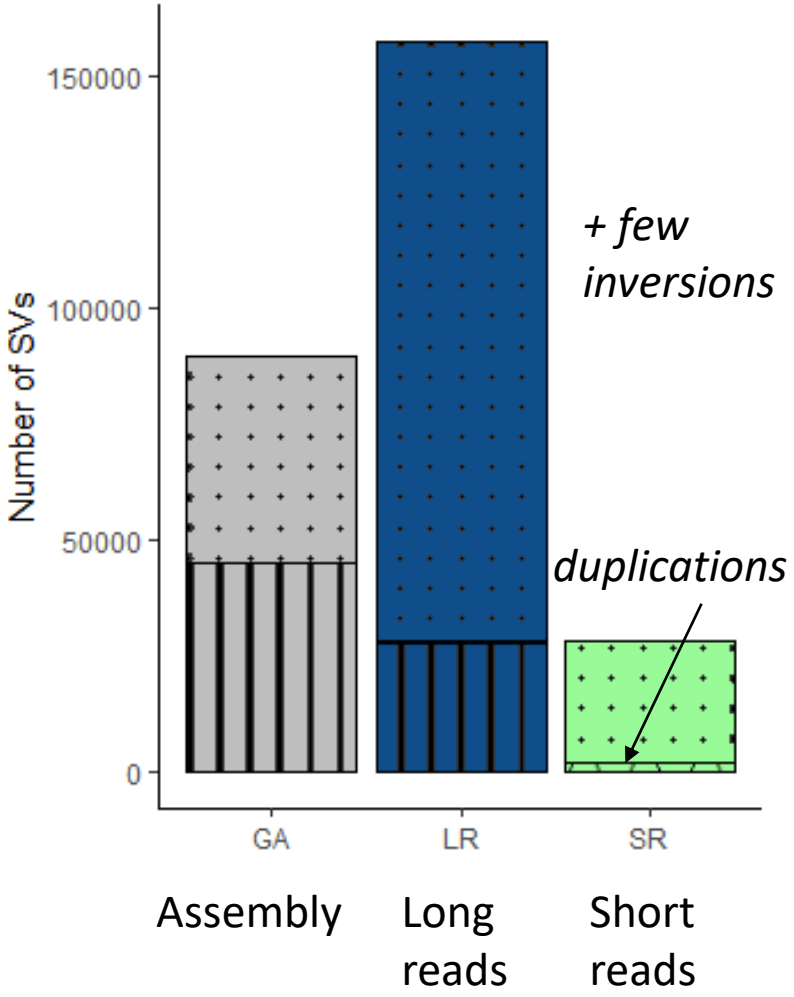
Mérot et al, MolEcol, 2022

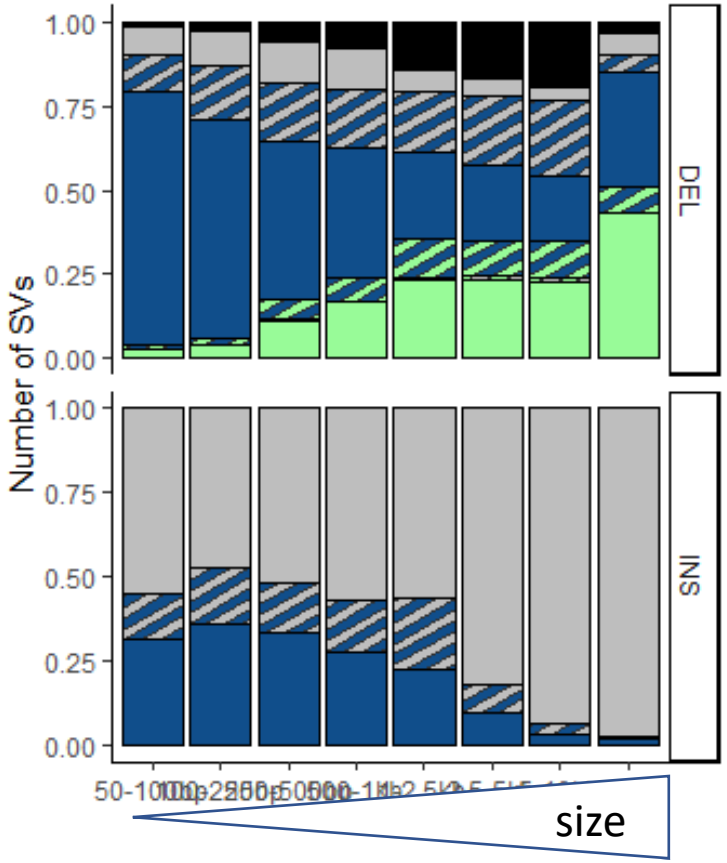
https://doi.org/10.1111/mec.16468

deletions

insertions







⇒A large majority are insertions & deletions

⇒Long-reads detect more SVs and provide sequences of INS

⇒Limited overlap: Different tools detect different types/length of SVs

Summary

- Structural variation has been systematically missed
- Previous technologies missed most of the SVs due to technical limitations.
- The majority of SVs are novel and rare variants, implicating that structural variation databases are not saturated yet

We can detect SV… now what?!
=> Why does it matter to understand adaptation?

- Avoid misinterpretation:
  - Large rearrangements can drive artefactual population structure
  - Not the same interpretation if an islands of divergence is an inversion or not…

- Test the role of SV in adaptation
  - Evidence of adaptive SV are anecdotical…
  - Can we test which SV are putatively adaptive as we did on SNPs?

⇒ Need of methodological development

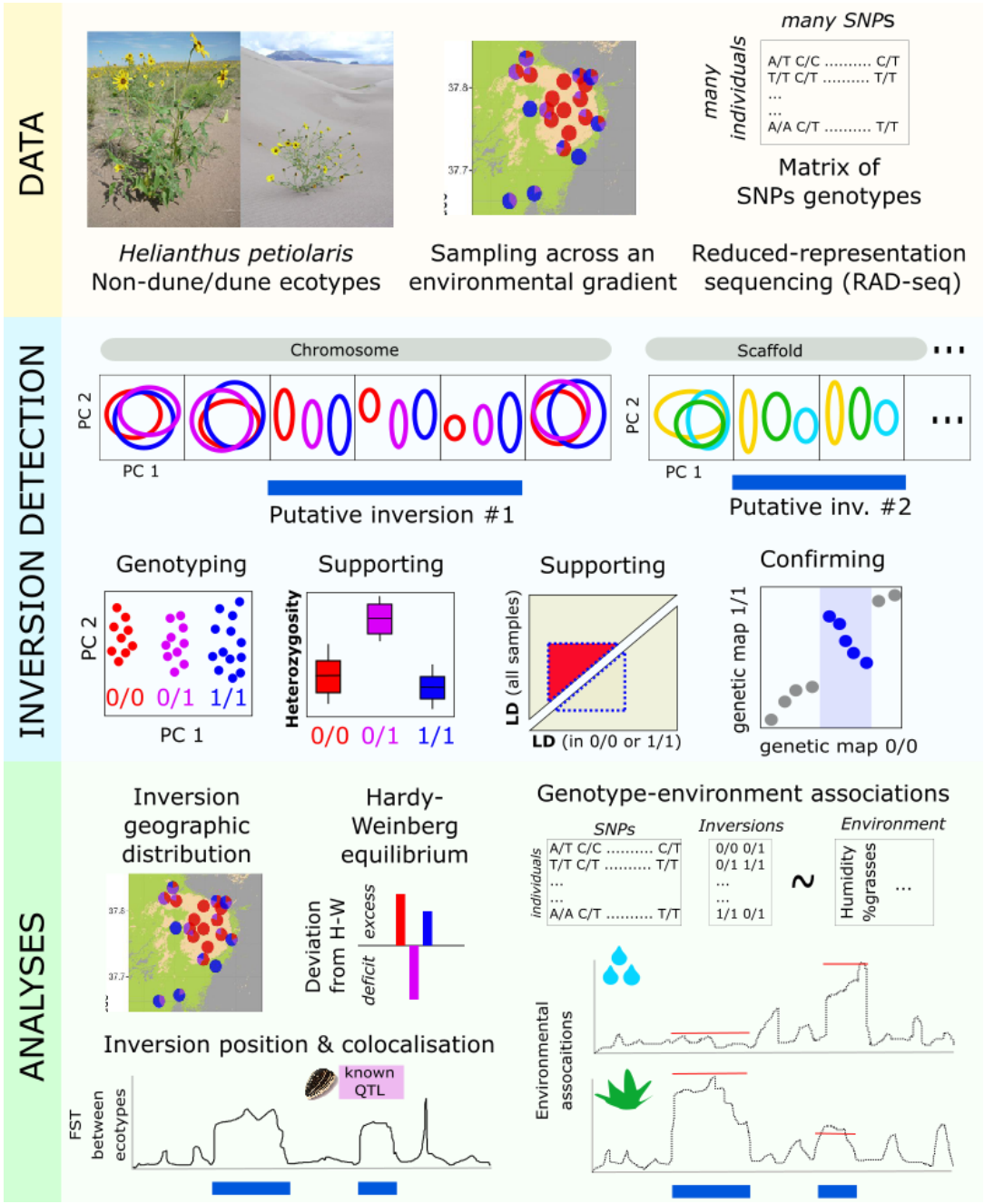
SV and adaptation genomics

Previously identified « islands of divergence »…
are now identified as inversions

Analyse SV within population genomics
or landscape genomics frameworks?

Mérot, 2020,
Mol Ecol

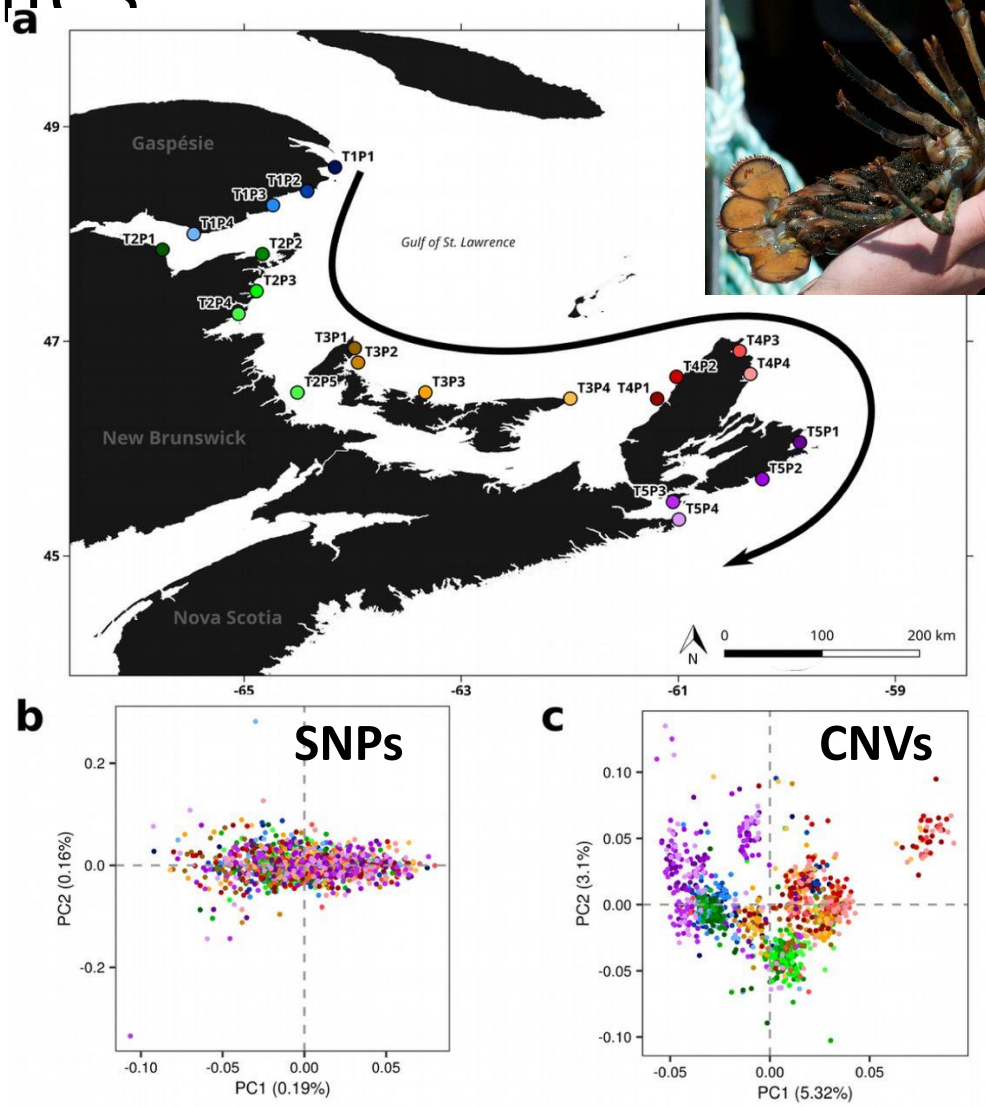
Huang *et al*
2020, Mol Ecol



SV and adaptation genomics

Use SV as a different kind of markers?

In the American Lobster, fine-scale structure and adaptation are better described by CNVs than by SNPs

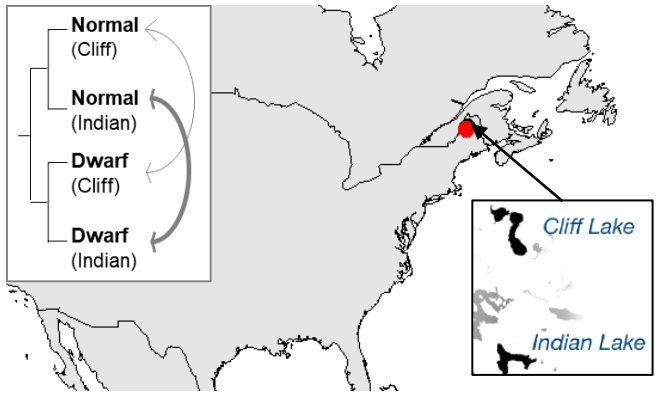


Dorant et al. (2020) Mol Ecol

SV and speciation genomics

Mérot et al, MolEcol, 2022

https://doi.org/10.1111/mec.16468



	Normal sp.	Dwarf sp.
Cliff Lake	8	8
Indian Lake	7	9

sp. Normal

sp. Dwarf

Nanopore long-reads (1 Cliff sample - 23x)

Nanopore long-reads (1 Cliff sample - 23x)

Genome assembly (reference)

Genome assembly

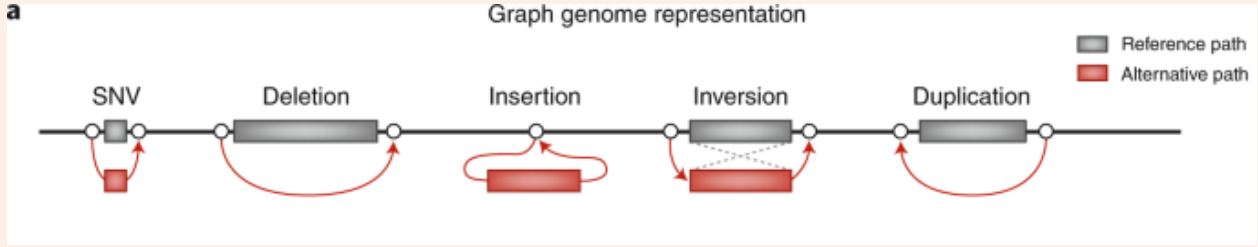
Illumina short-reads (5x)

SV detection short-reads

SV detection long-reads

SV detection 2 assemblies

Genome-graph representation
of SVs within and between Normal and Dwarf species



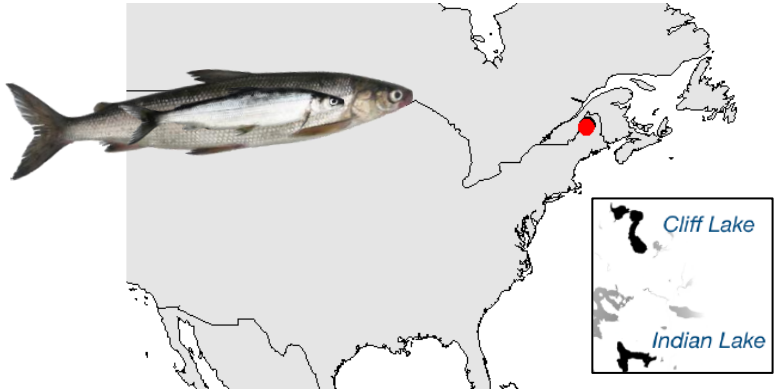
⇒ VG-tools + Giraffe

⇒Genotyping 210,455 SVs

SV and speciation genomics

Mérot et al,
MolEcol, 2022

https://doi.org/10.1111/mec.16468



	Normal sp.	Dwarf sp.
Cliff Lake	8	8
Indian Lake	7	9

Illumina short-reads (5x)

->SVs:

maf > 5%

& missing data <50%

103,857 SVs

⇒Total length ***66,5 Mb***

-> *SNPs:*

maf > 5%

& missing data <50%

12,886,292 SNPs

⇒***12.9 Mb***

⇒SV cover about 5 times more bp than SNPs

⇒ Capture a larger fraction of genetic diversity

SV and speciation genomics

Mérot et al,
MolEcol, 2022

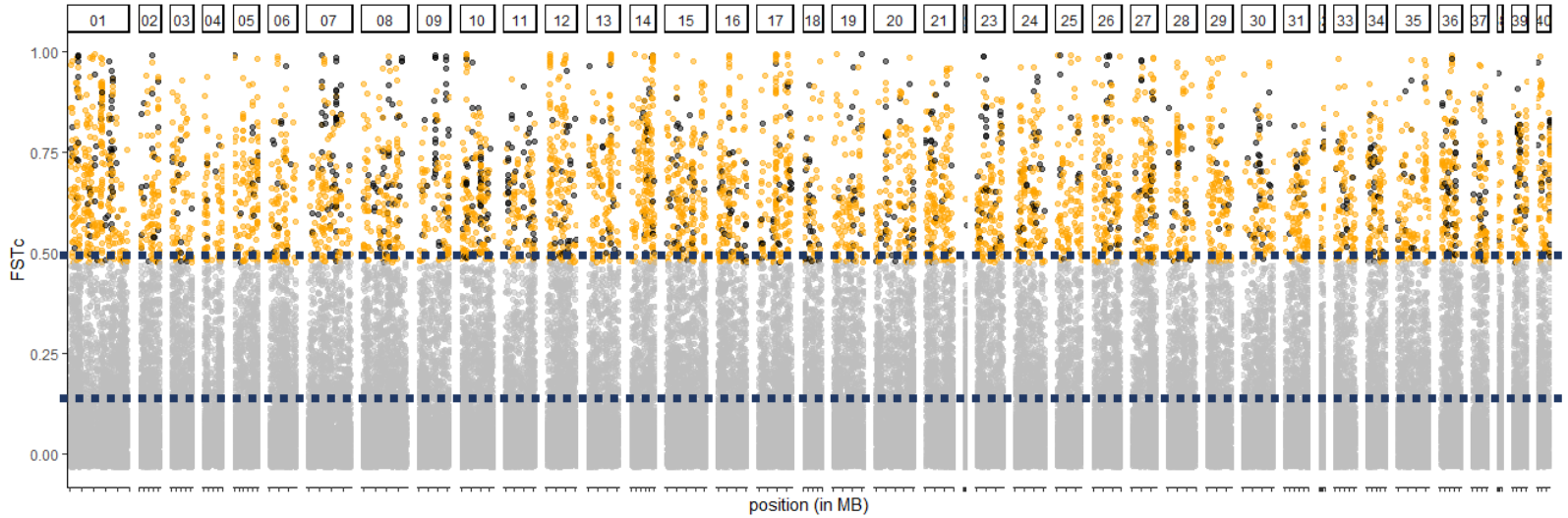
https://doi.org/10.1111/mec.16468

Normal vs. Dwarf species

(based on 103,857 SVs)

Cliff Lake

F_{ST}= 0.14



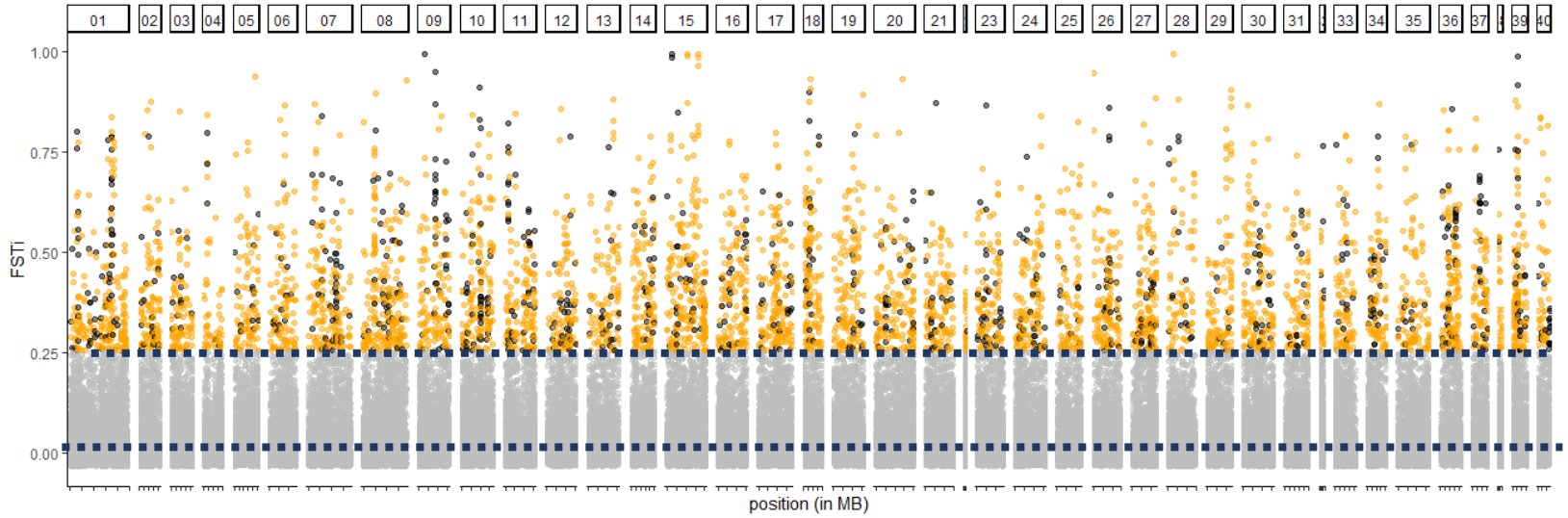
95% quantile

Genome-wide Fst = 0.14

Outliers in 1 lake

Indian Lake

F_{ST}= 0.06



Outliers in 2 lakes

95% quantile

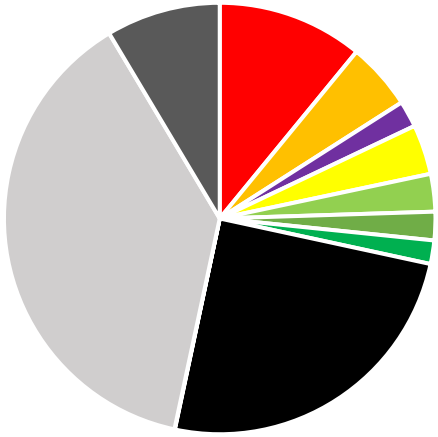
Genome-wide Fst = 0,06

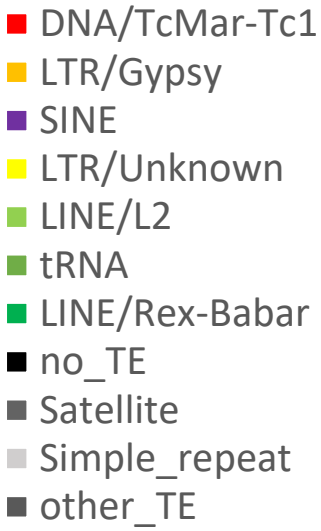
SV and speciation genomics

Mérot et al,
MolEcol, 2022

https://doi.org/10.1111/mec.16468

TE in SV





TE in outliers SV

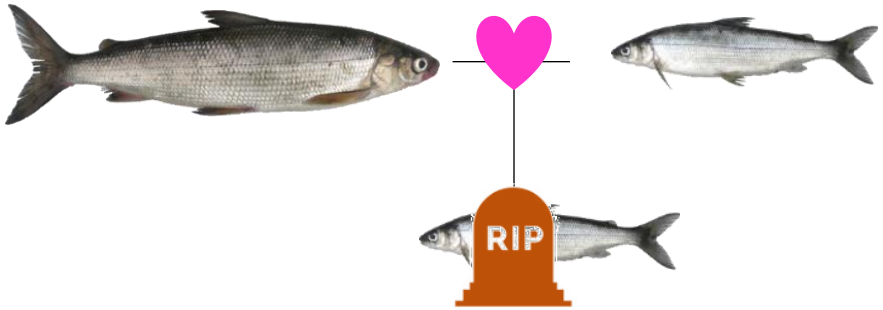
TE	N_SV	prop_SV	N_outlier	Prop_outlier	OR_fish
DNA/TcMar-Tc1	11358	11%	227	24%	2.2
LTR/Gypsy	5216	5%	100	11%	2.1
SINE	2059	2%	51	5%	2.7
LTR/Unknown	3880	4%	62	7%	1.8
LINE/L2	2927	3%	50	5%	1.9
tRNA	2215	2%	36	4%	1.8
LINE/Rex-Babar	1828	2%	30	3%	1.8
no_TE	25912	25%	225	24%	1.0
Satellite	33	0%	NA	NA	0.0
Simple_repeat	39529	38%	37	4%	0.1
other_TE	8900	9%	118	14%	

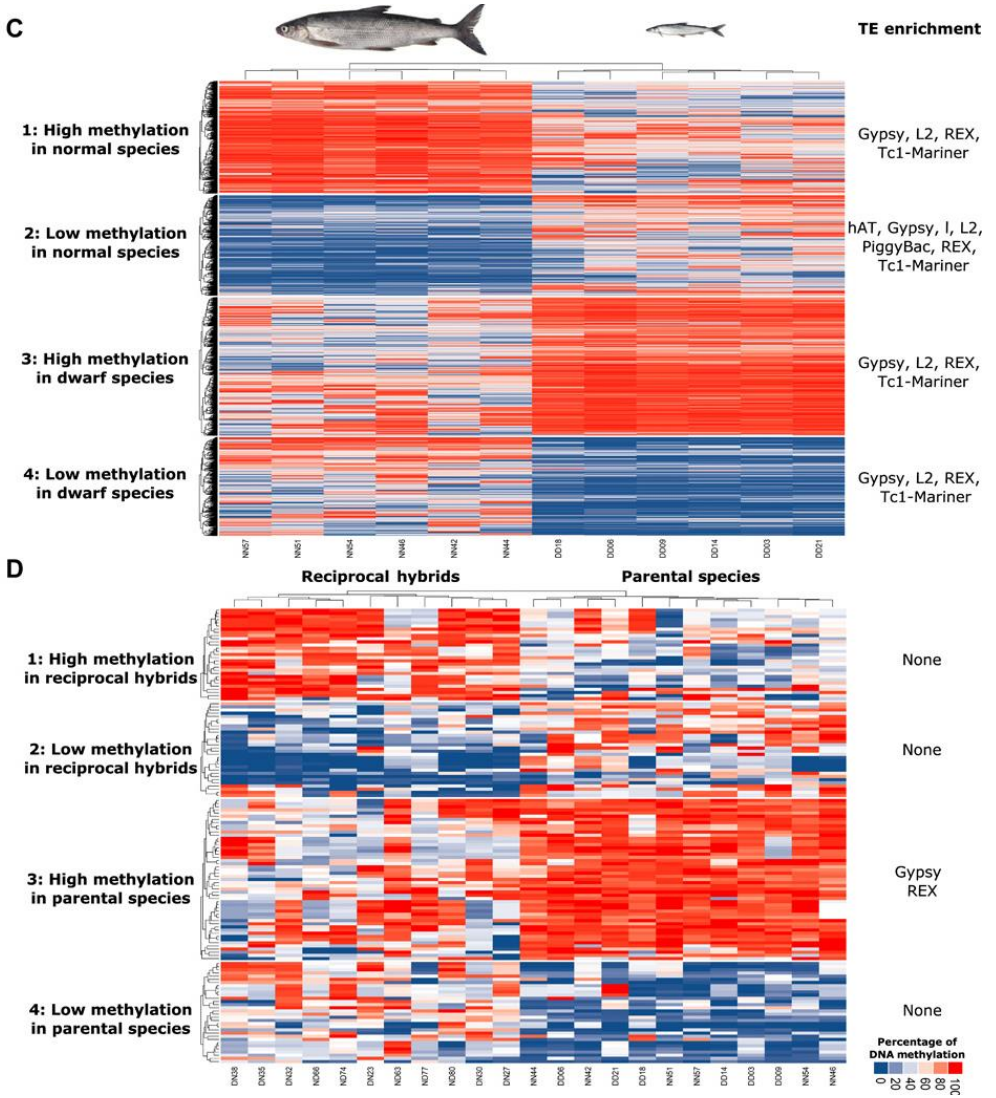
Significant enrichmer
(fdr<0,05)

-> deficit

SV and speciation genomics

Dion-Côté et al, 2014
Laporte et al, 2019





M. Laporte

TEs = reactivation shock
⇒ post-zygotic breakdown

Remaining challenges

- Large repetitive regions remain inaccessible due to constraints of read length and sequence composition
- No gold standard for SV detection and validation
- Statistical tools for population genomics, adaptation genomics, ecological genomics are based on SNPs