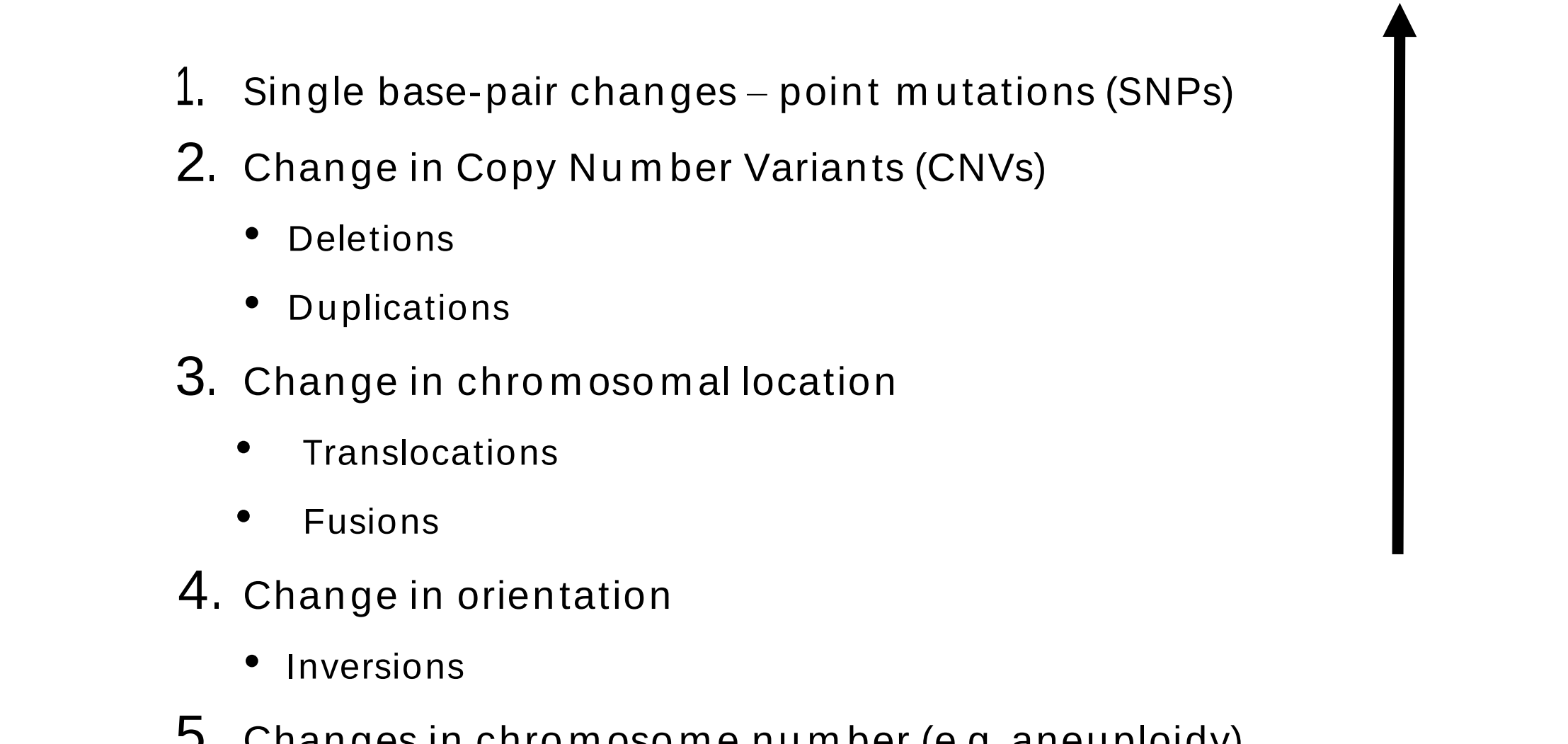
Detecting and analysing genomic structural variants

Prepared with A. Tigano & M. Wellenreuther

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 2020: many tools available with advantages and drawbacks

  - Short-reads (illumina)

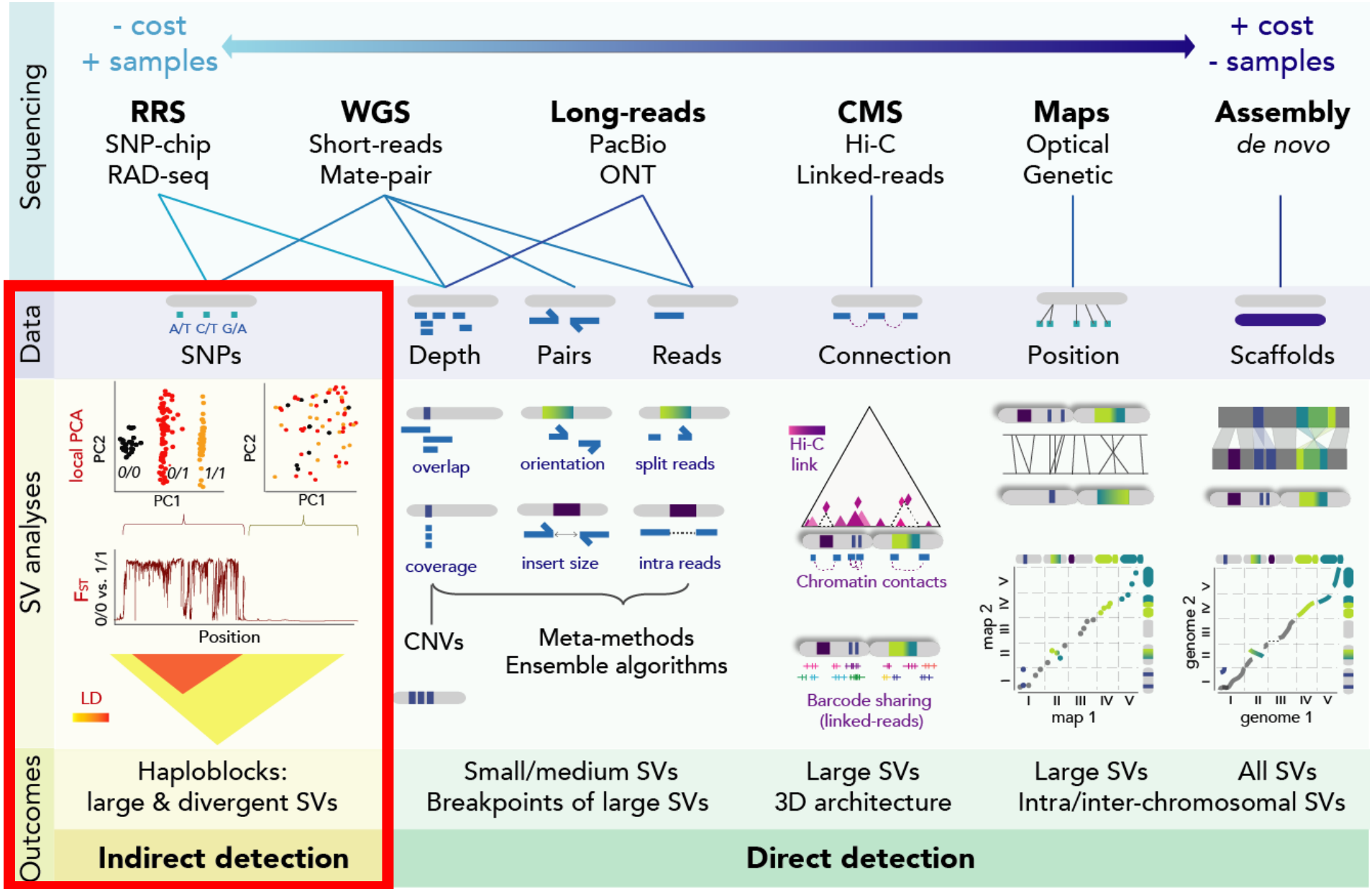
- high single-nucleotide accuracy & paired-end
- underrepresentation of high-GC regions

  - Long-reads (PacBio/Nanopore)

- Higher error rate (~15%) and single-end
- Longer sequences (~1-50kb)

  - Emerging technologies (Hi-C, 10x, optical mapping)

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

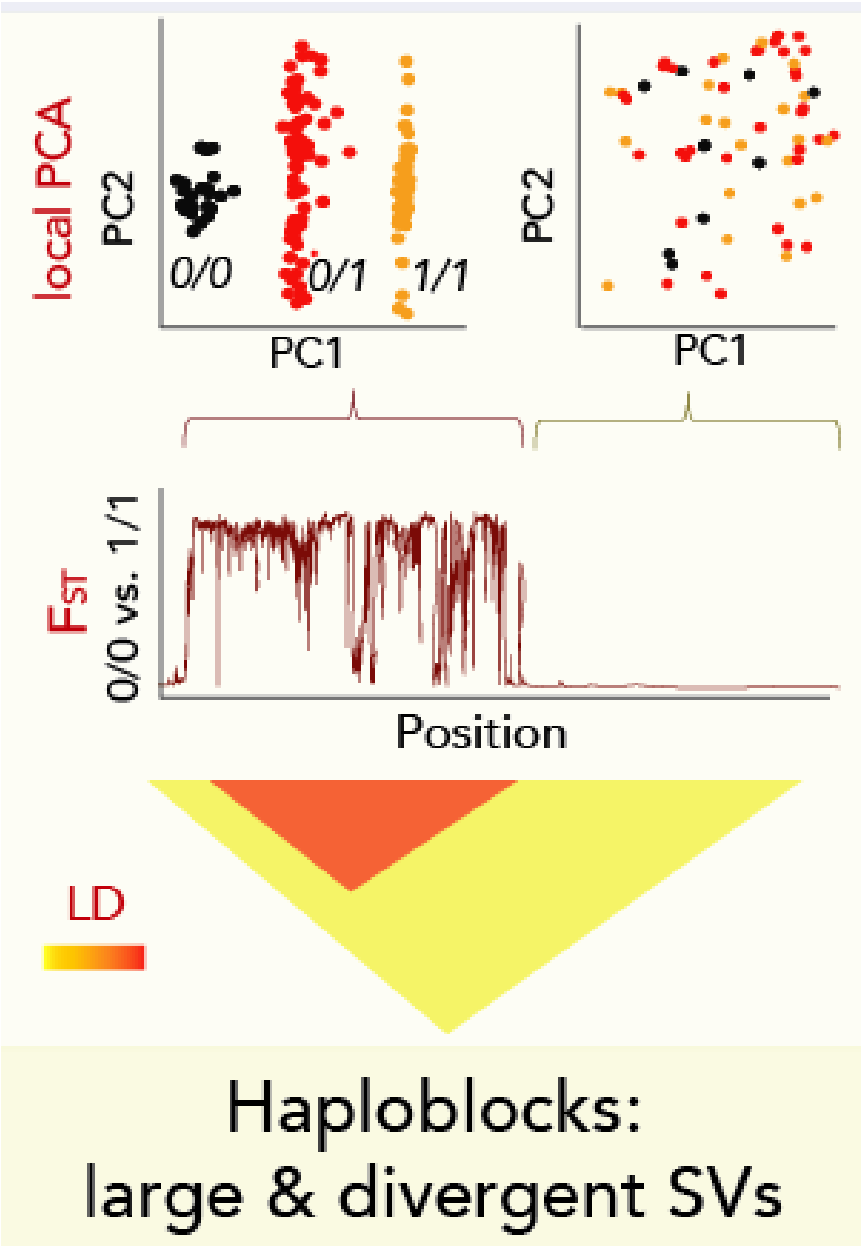

Sequencing
- cost
+ samples
+ cost
- samples
RRS
SNP-chip
RAD-seq
WGS
Short-reads
Mate-pair
Long-reads
PacBio
ONT
CMS
Hi-C
Linked-reads
Maps
Optical
Genetic
Assembly
de novo
Data
A/T C/T G/A
SNPs
Depth
Pairs
Reads
Connection
Position
Scaffolds
SV analyses
local PCA
PC2
PC1
0/0
0/1
1/1
FST
0/0 vs. 1/1
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....)

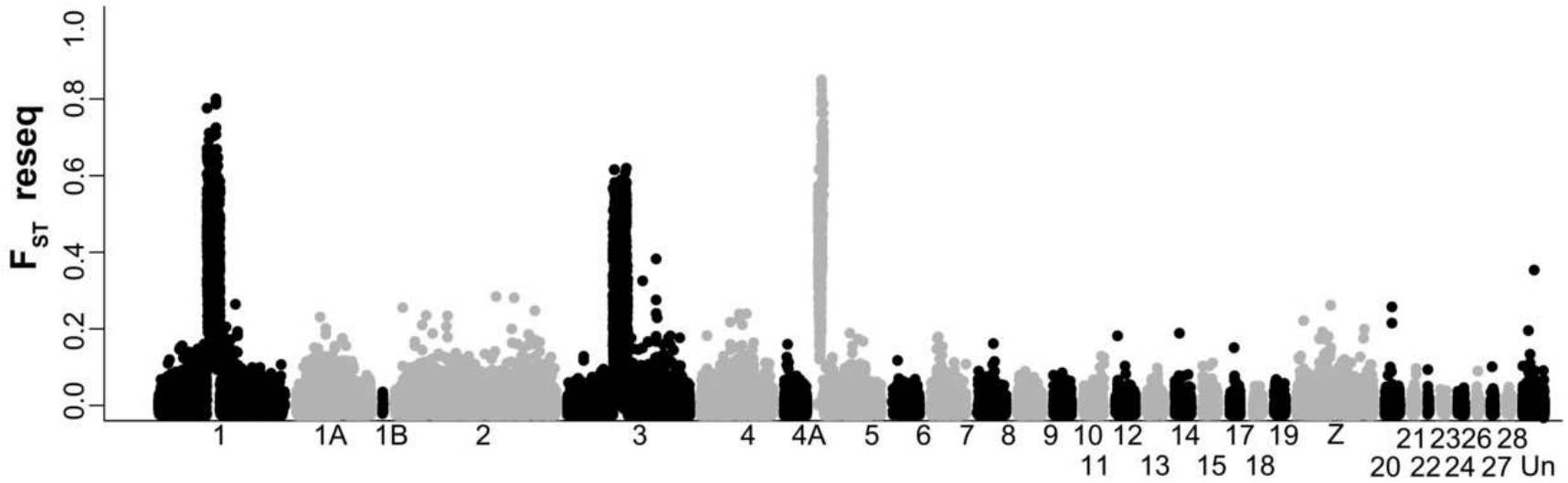
- Able to detect chromosomal rearrangements if they are:
  - Large (> 100 kb)
  - Polymorphic
  - Divergent

⇒ 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?

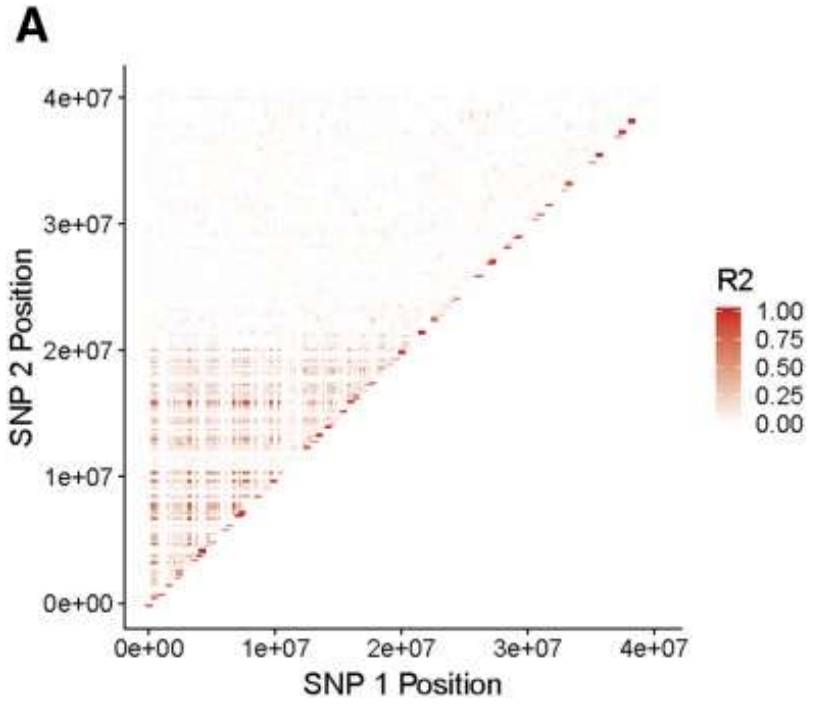
Lundberg et al. Evolution Letters 2017

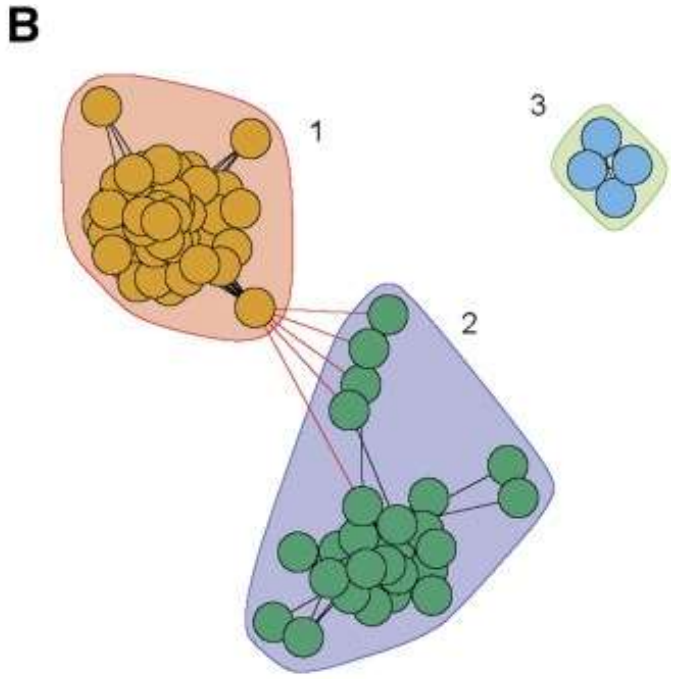
Indirect detection : LD networks

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





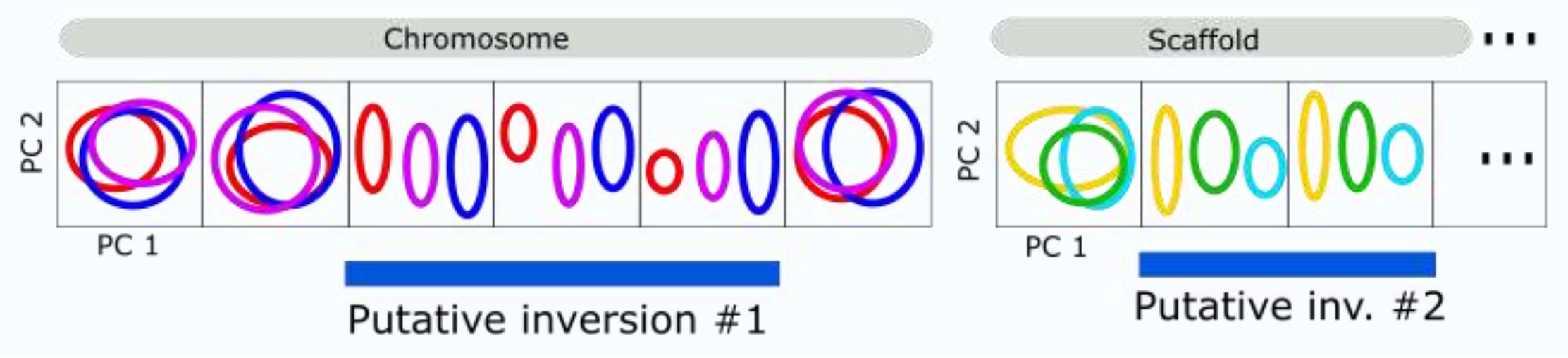
McKinney et al 2020. *G3, 10*(5), 1553–1561. https://doi.org/10.1534/g3.119.400972

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 aninversion 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

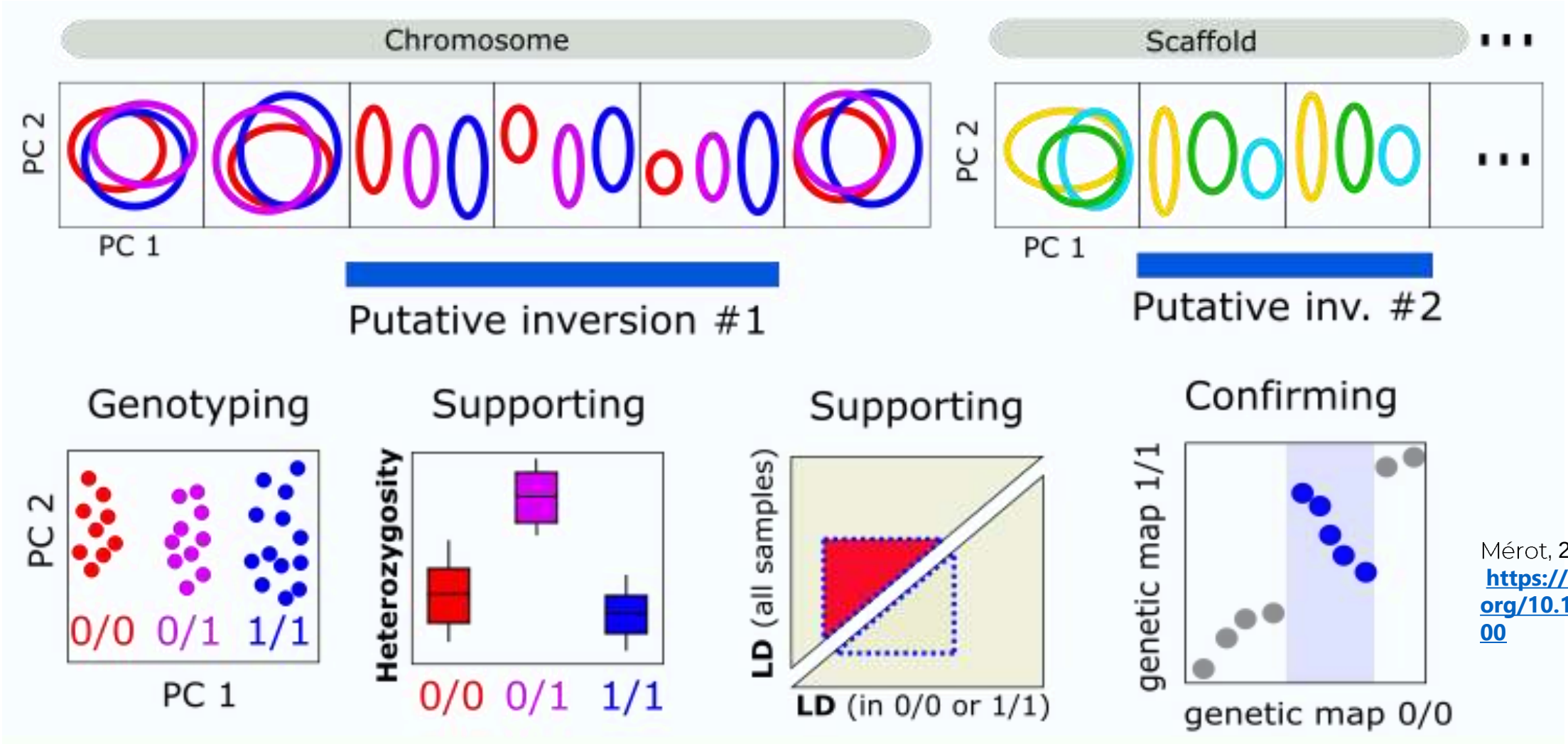
⇒ Perhaps less likely when blocks are very large (>1MB)

- Low-recombination regions?

⇒ LD should be observed in all clusters

Indirect detection :

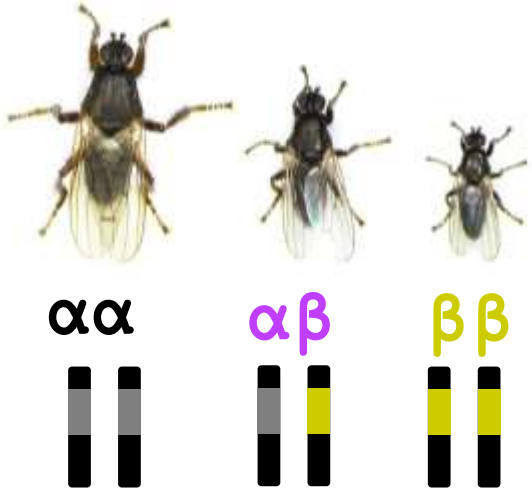
How can we support that an haploblock is an inversion?

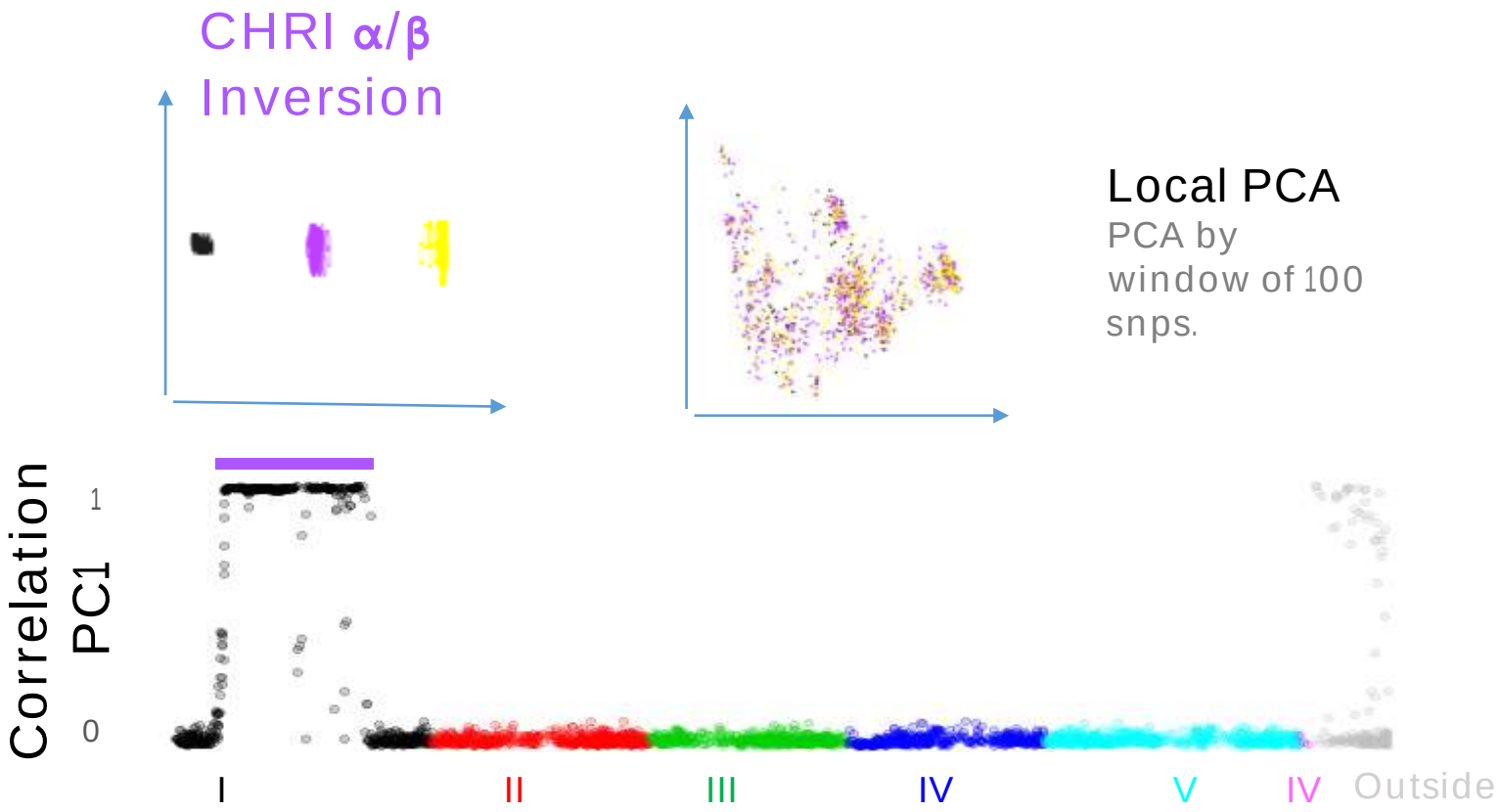


Mérot, 2020, MolEcol https://doi-org/10.1111/mec.15500

Indirect detection: Case study in the seaweed fly Coelopa frigida

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



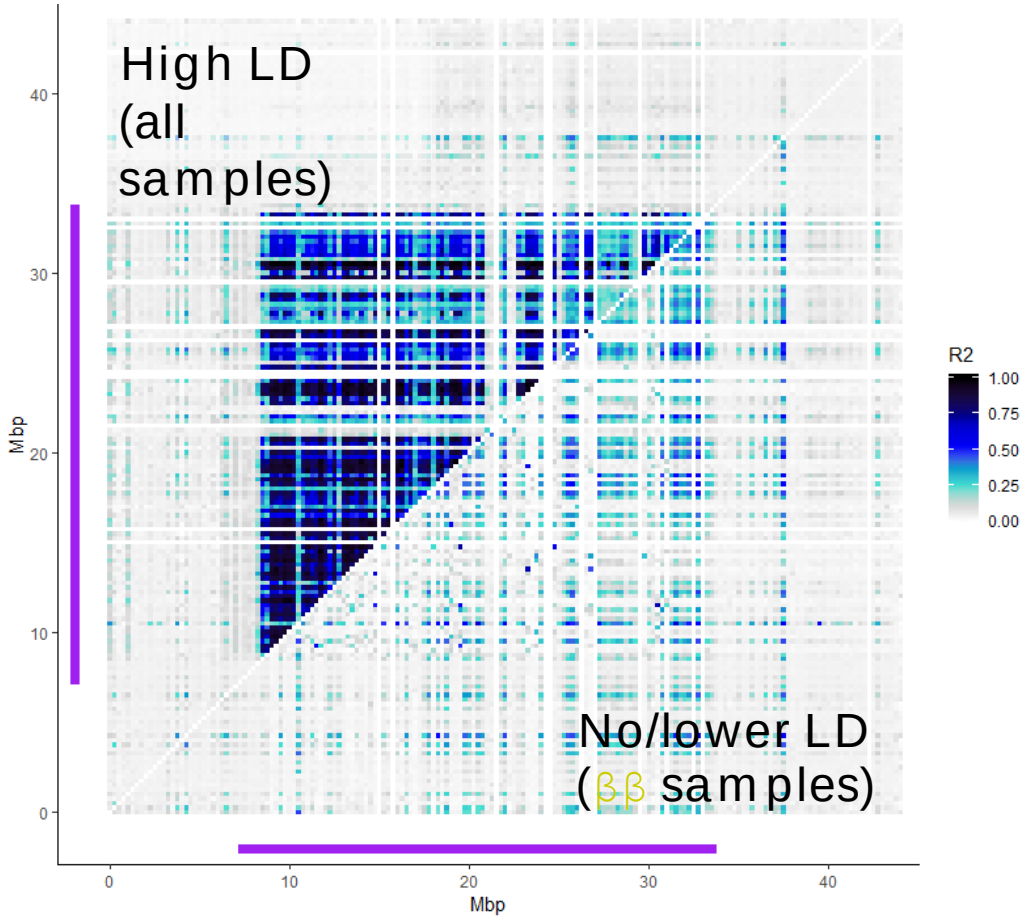


CHR-I inversion
27Mb
11% genome
16,5% of SNPs
1500 genes

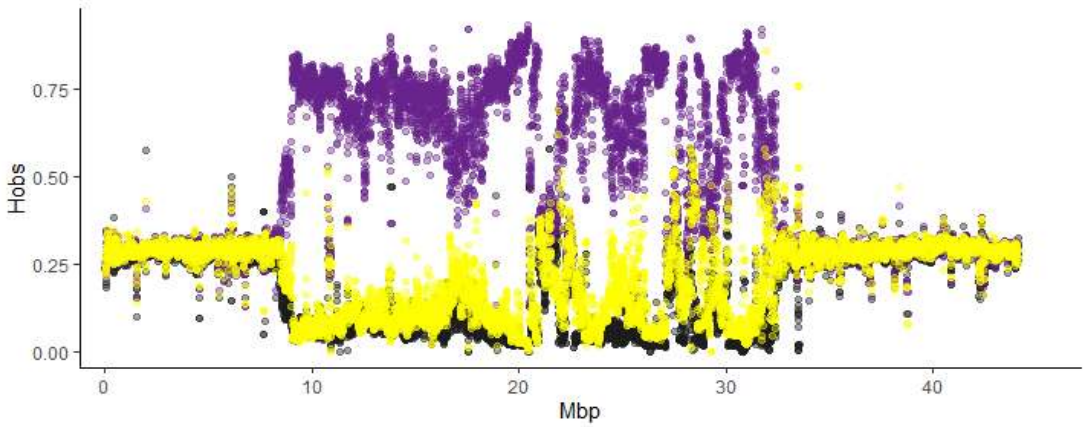
Indirect detection: Case study in the seaweed fly Coelopa frigida

Exploration of the haploblock/inversion

Chr I

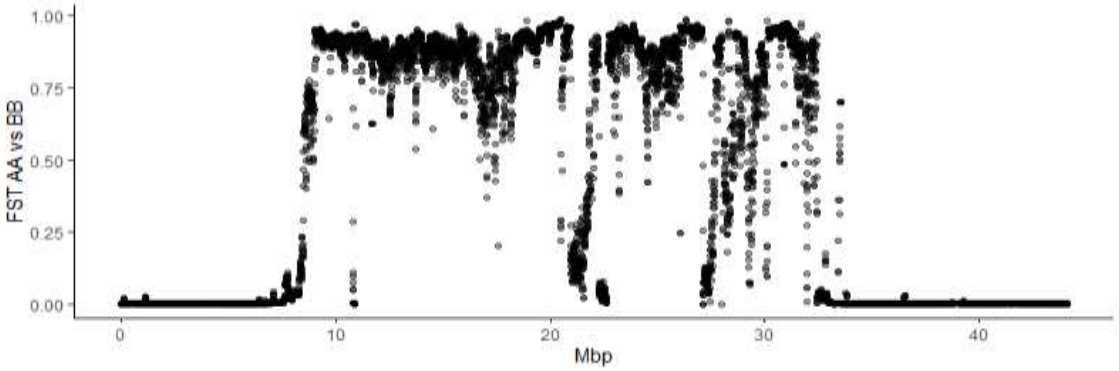


Higher observed heterozygosity in αβ than in αα or ββ



Chr I

High FST differentiation between αα and ββ



Chr I

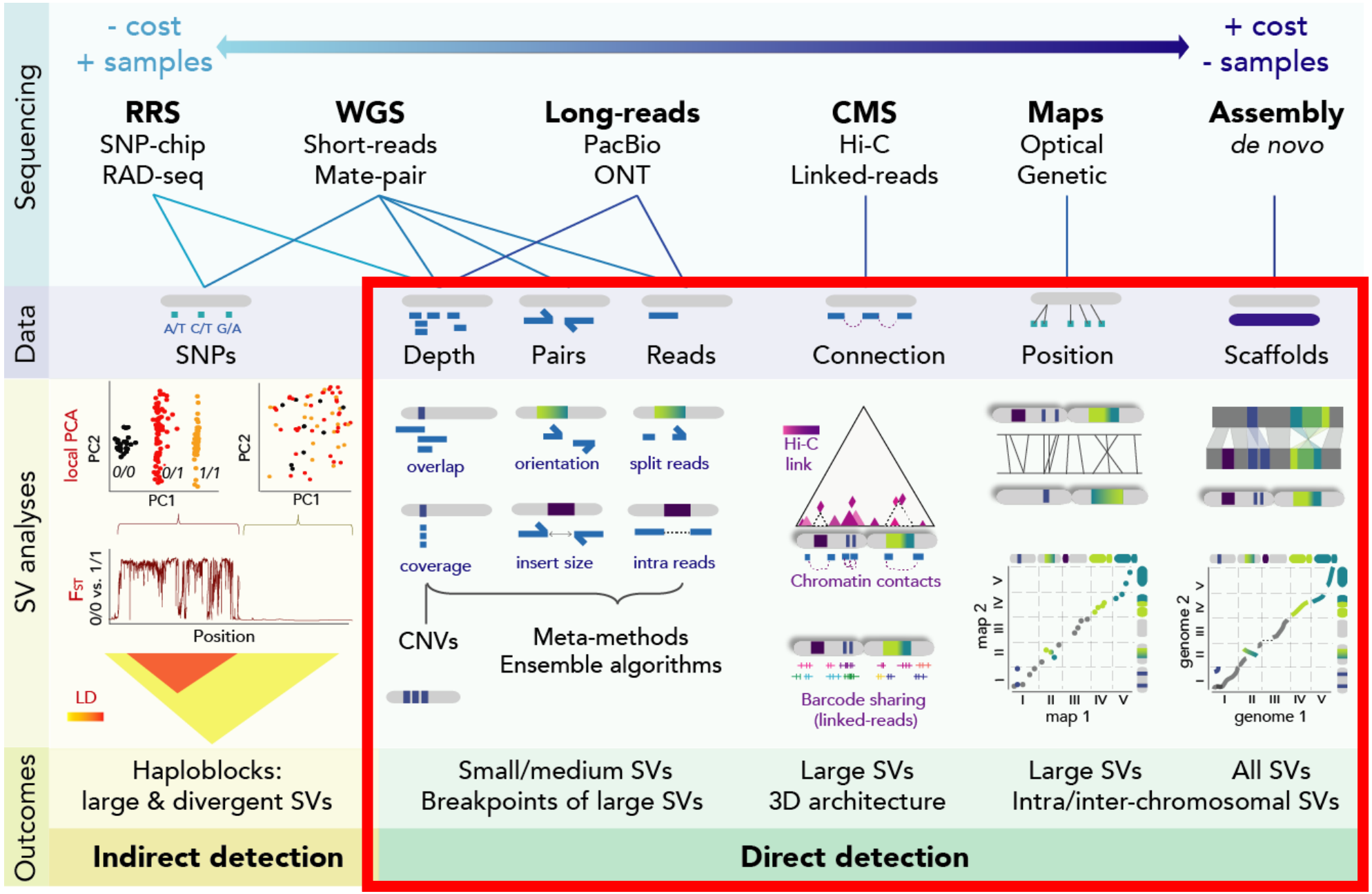
Indirect detection of SV :

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


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
I
II
III
IV
V
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

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

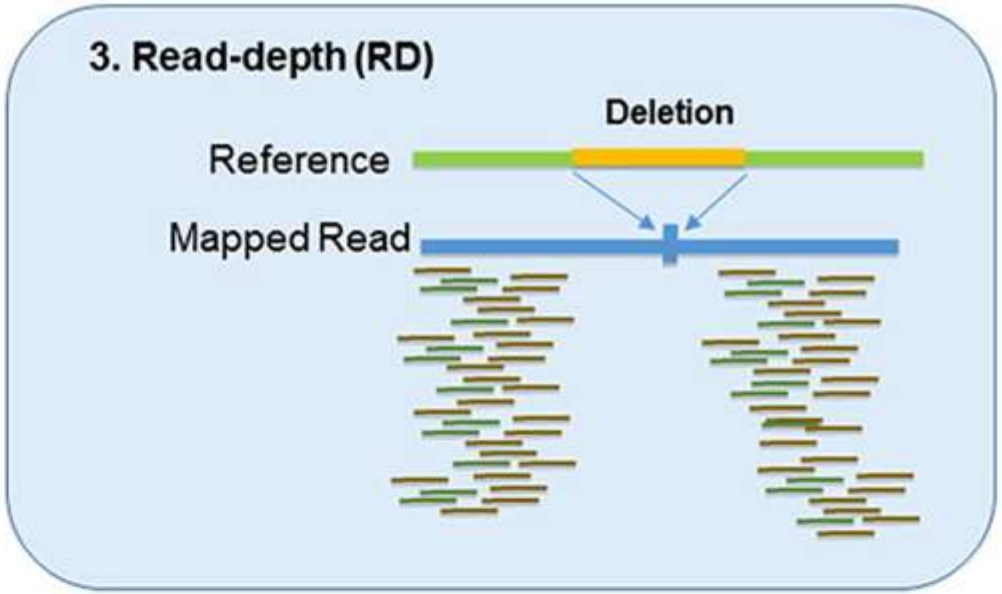
- Short-reads (100-150 bp) single or paired-end

⇒ 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)



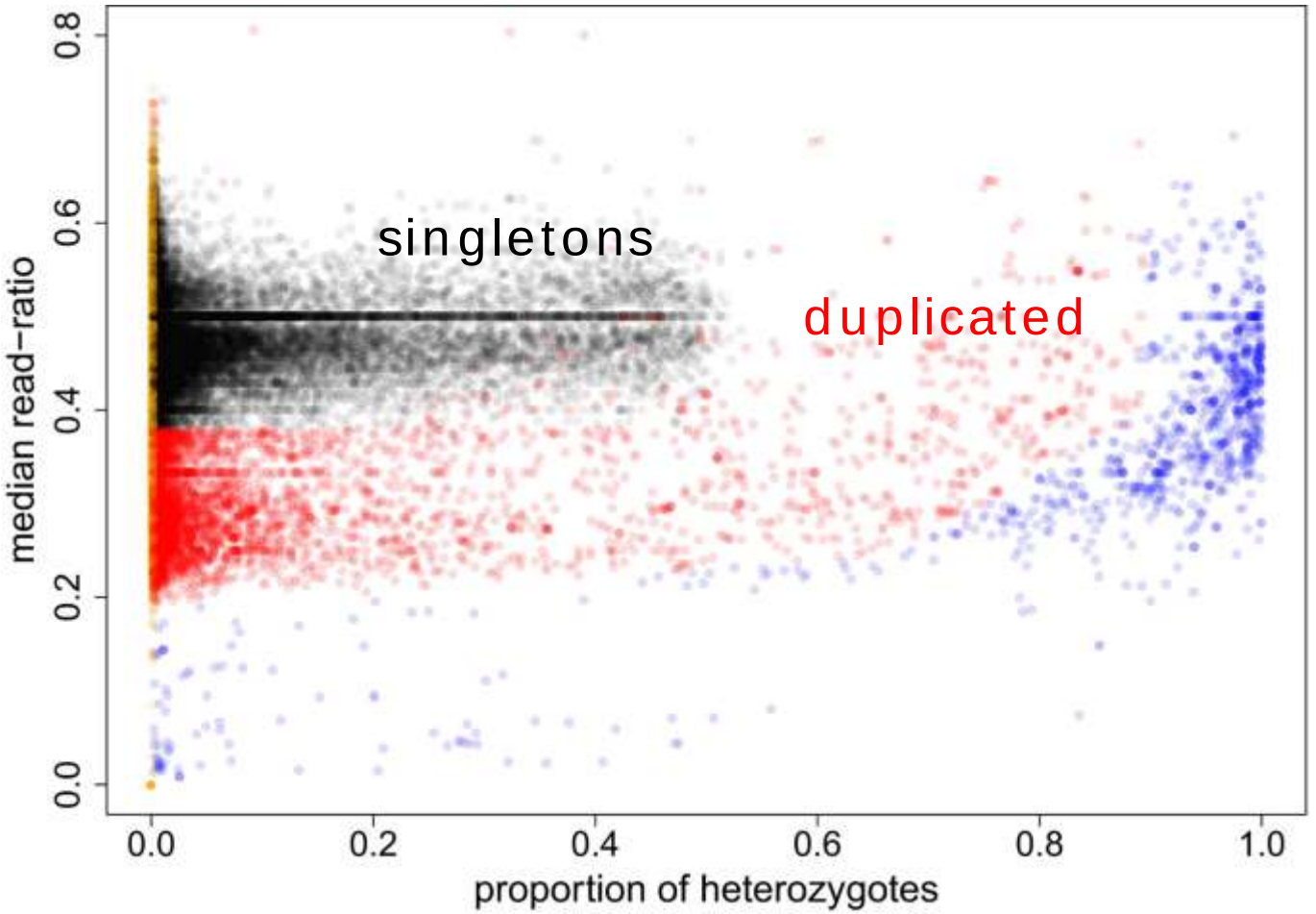
Direct detection : with read depth

Adding allelic information and heterozygote information...

⇒ Detect duplicated loci in RAD-seq

-> Filter them out for regular analysis

-> Keep them apart to analyse CNVs

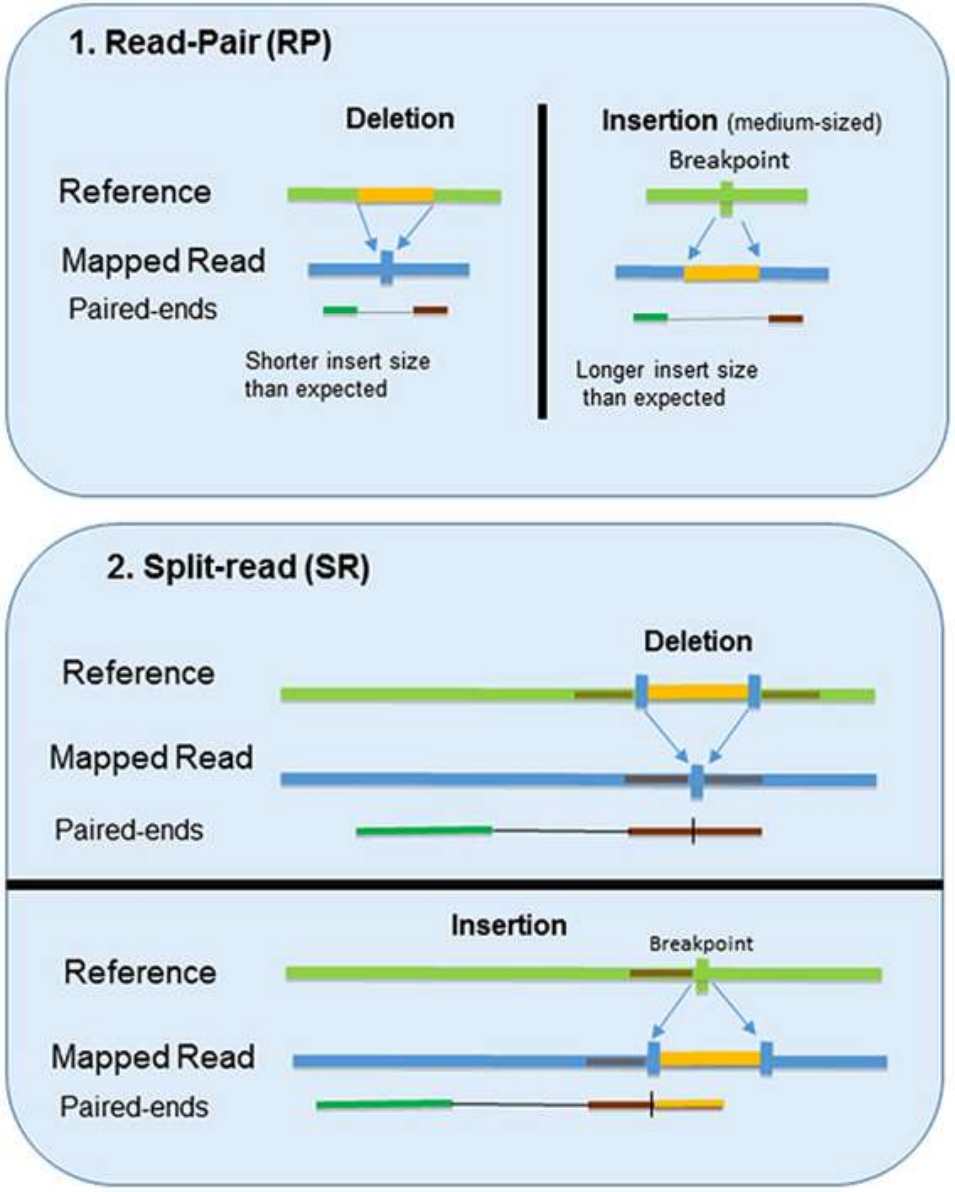


Dorant et al 2020. MolEcol https://doi.org/10.1111/mec.15565
McKinney,et al. 2017 MolEcol Ressources. https://doi.org/10.1111/1755-0998.12613

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

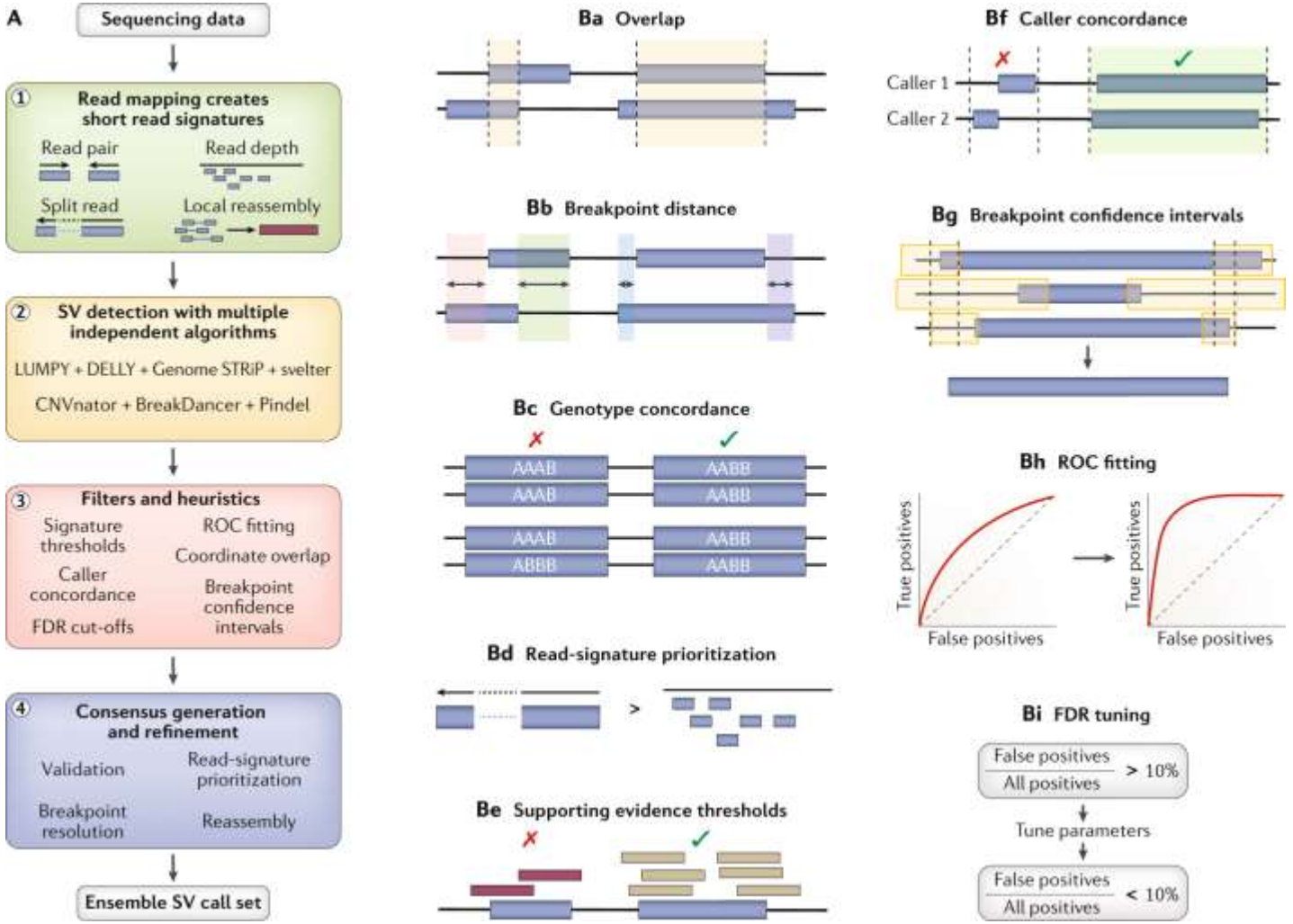
This will detect shorts indels and breakpoints of duplications, translocations or inversions

Most-used tools:
Delly, Manta, 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



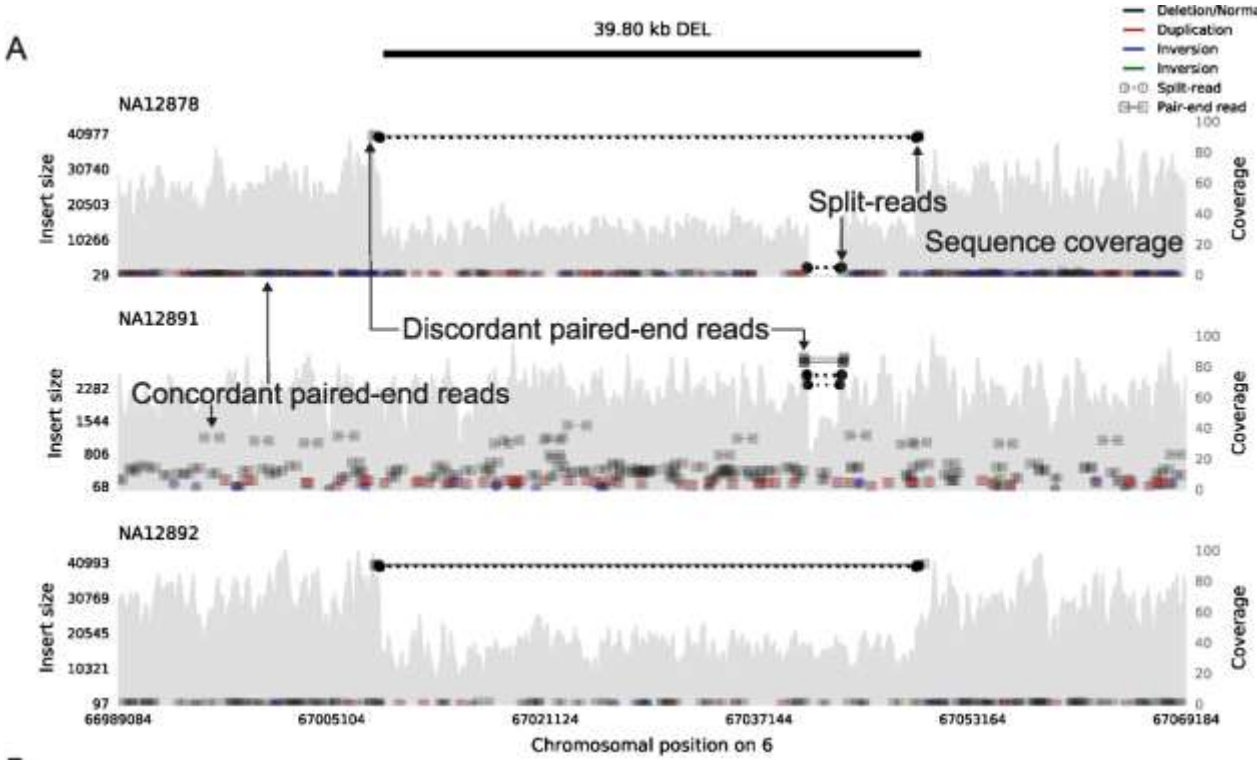
Ho, S.S., Urban, A.E. & Mills, R.E. Structural variation in the sequencing era. *Nat Rev Genet* **21,** 171–189 (2020). https://doi.org/10.1038/s41576-019-0180-9

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:

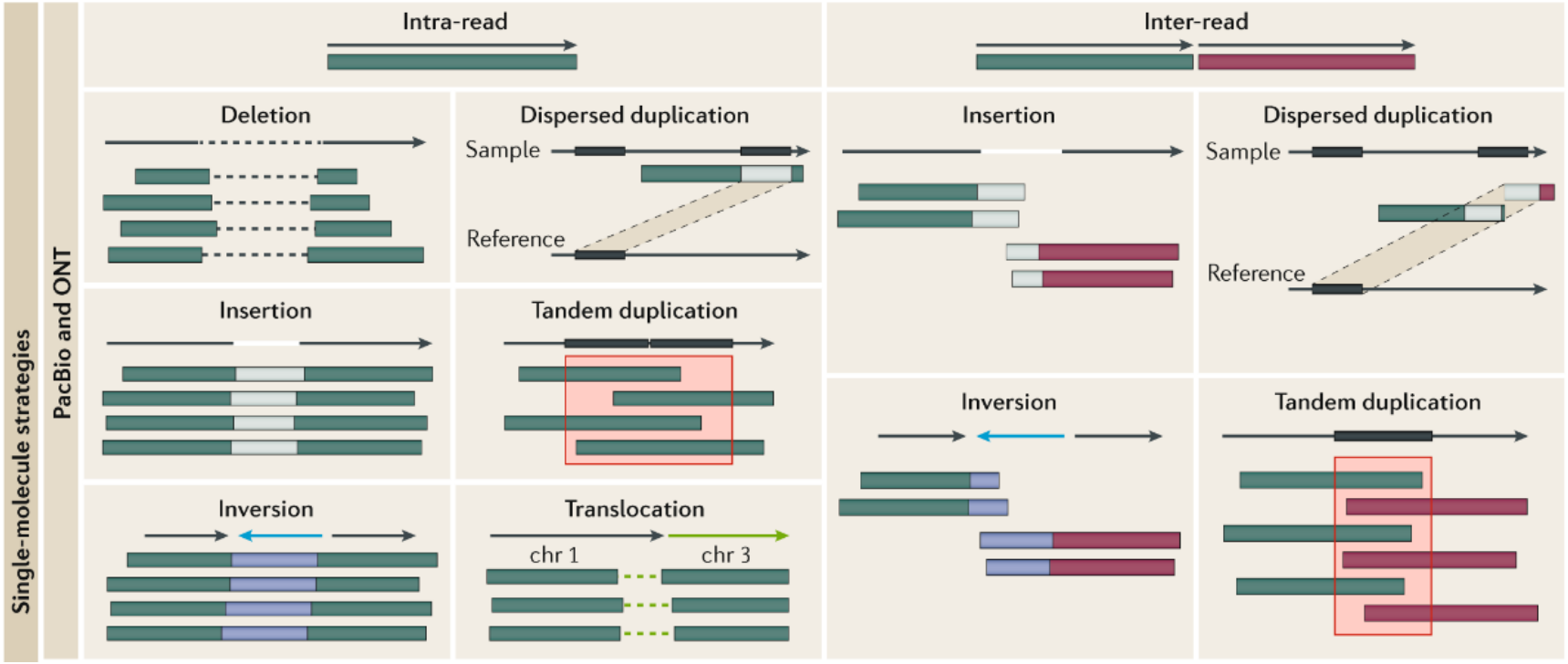
- graph-based approaches
- population-scale genotyping of SV

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 genotype SV at population scale…



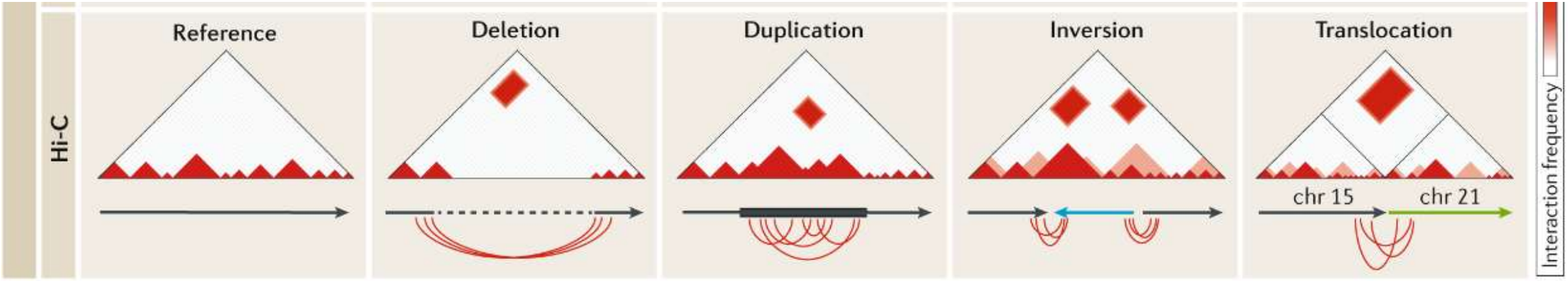
H et al . *Nat Rev Genet* (2020). https://doi.org/10.1038/s41576-019-0180-9

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

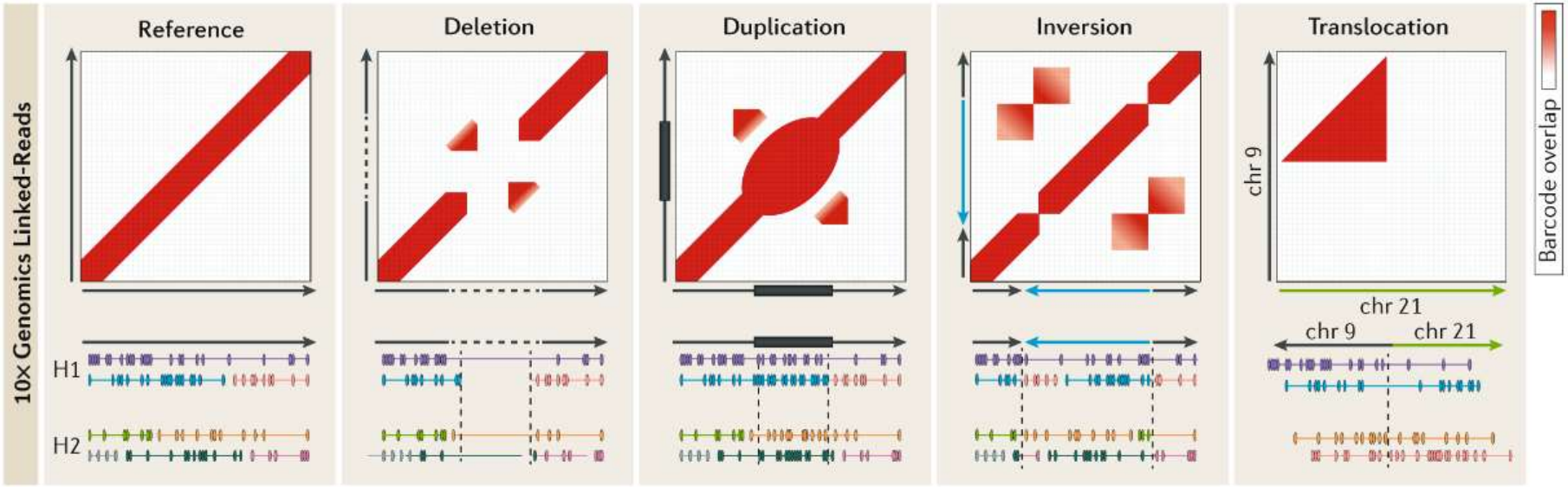
Direct detection: Connected-molecule strategies

Linked-reads (10xGenomics, Emerging in-house haplotagging)

Meier et al bioRxiv 2020 https://doi.org/10.1101/2020.05.25.113688

- 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

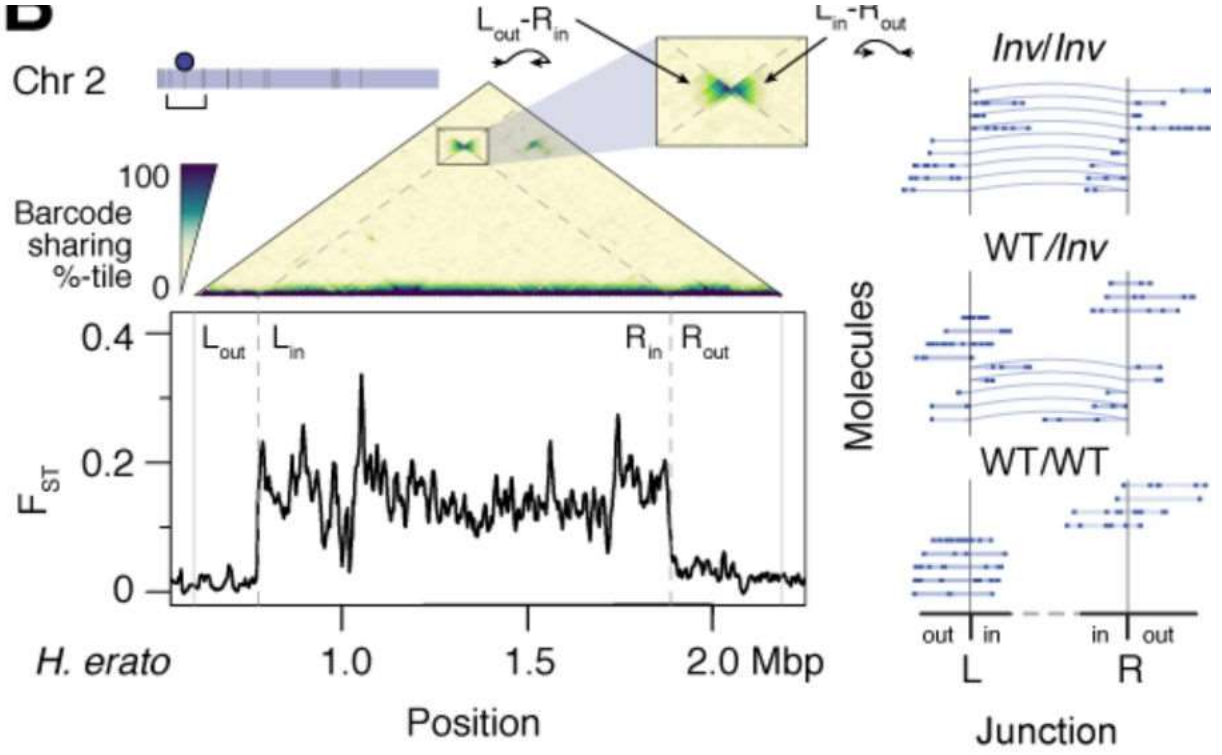
Direct detection: Connected-molecule strategies

Linked-reads

⇒ Medium and large inversions & indels

- Example:

Inversion detection in Heliconius butterflies



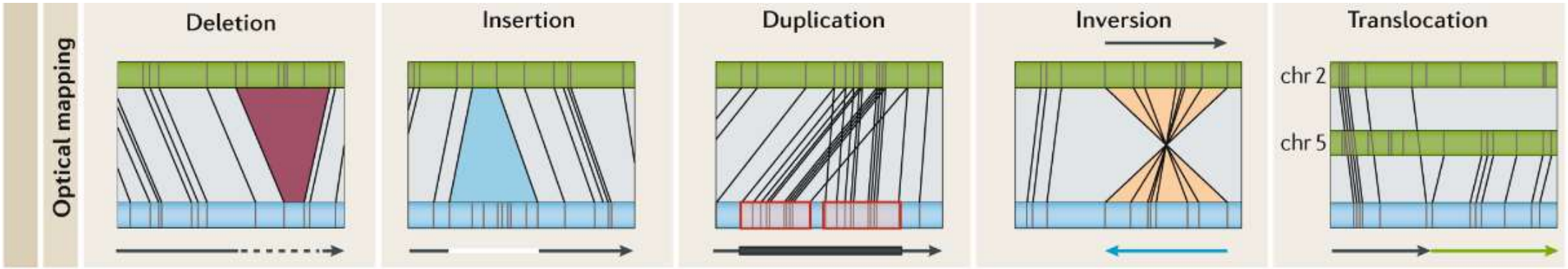
Meier et al bioRxiv 2020
https://doi.org/10.1101/2020.05.25.113688

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)

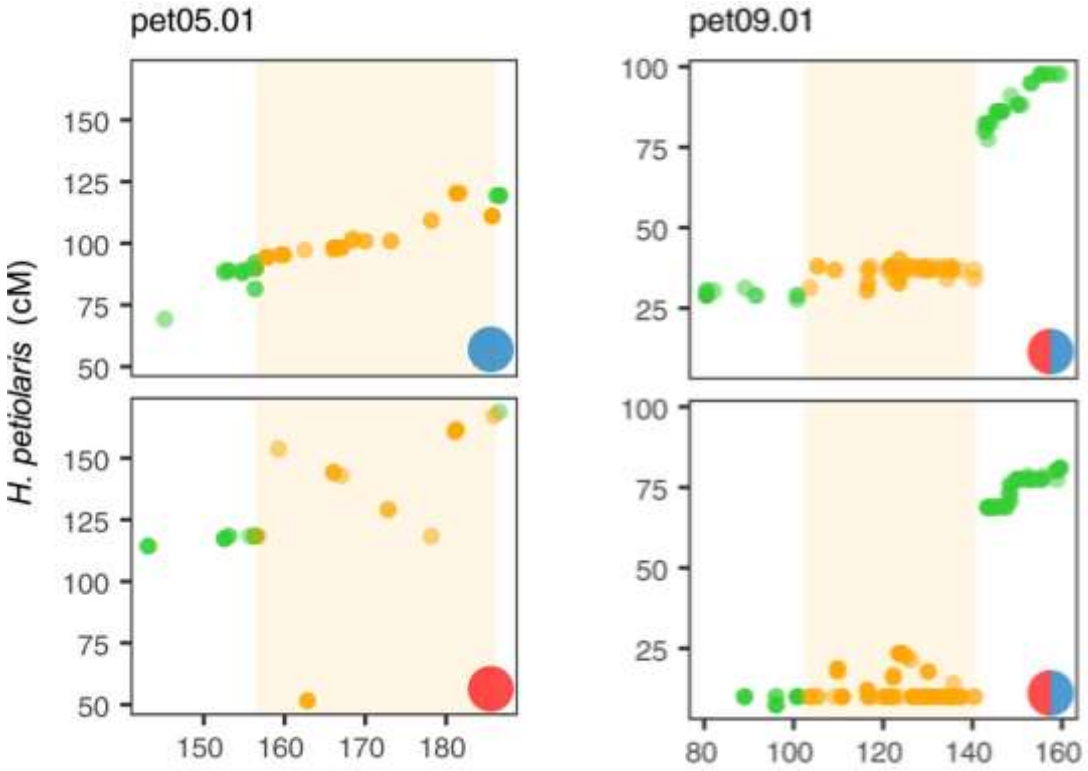
- compare marker position between families or between one family and reference genome

- 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

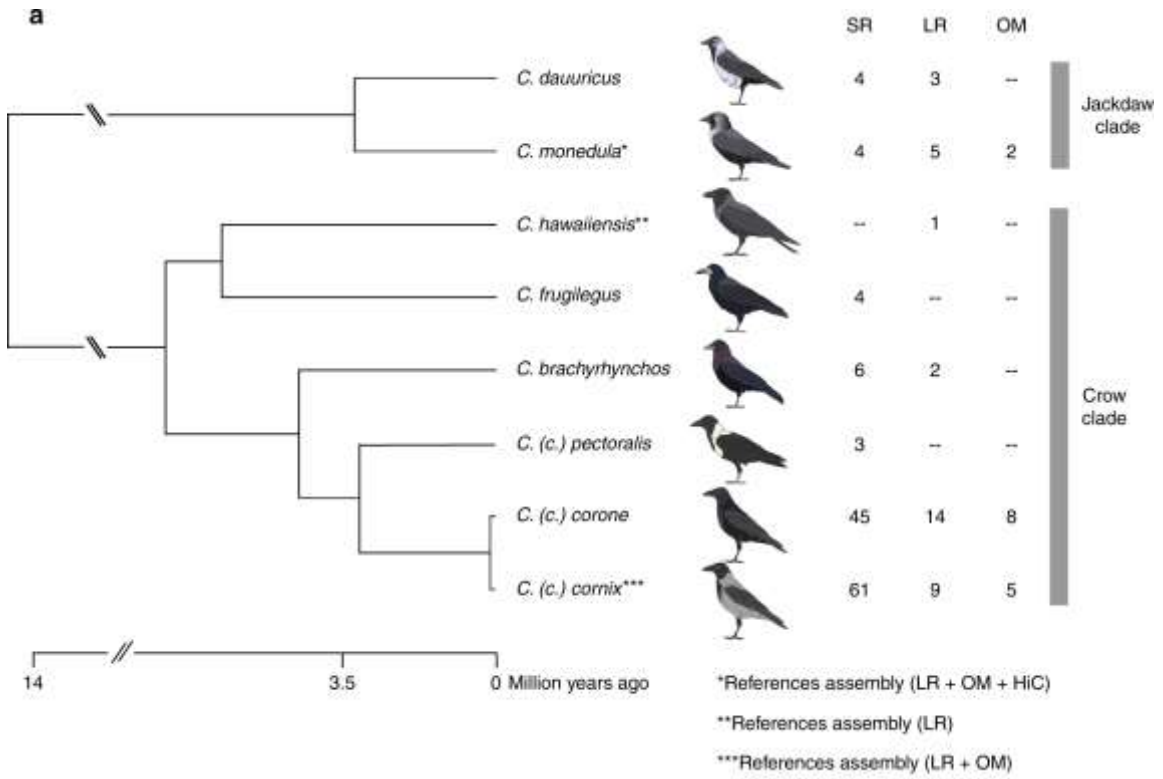
Direct detection: genome comparison

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 (for instance 1 sample per species)

Direct detection : combining platforms

Short-reads Long-reads Optical Mapping



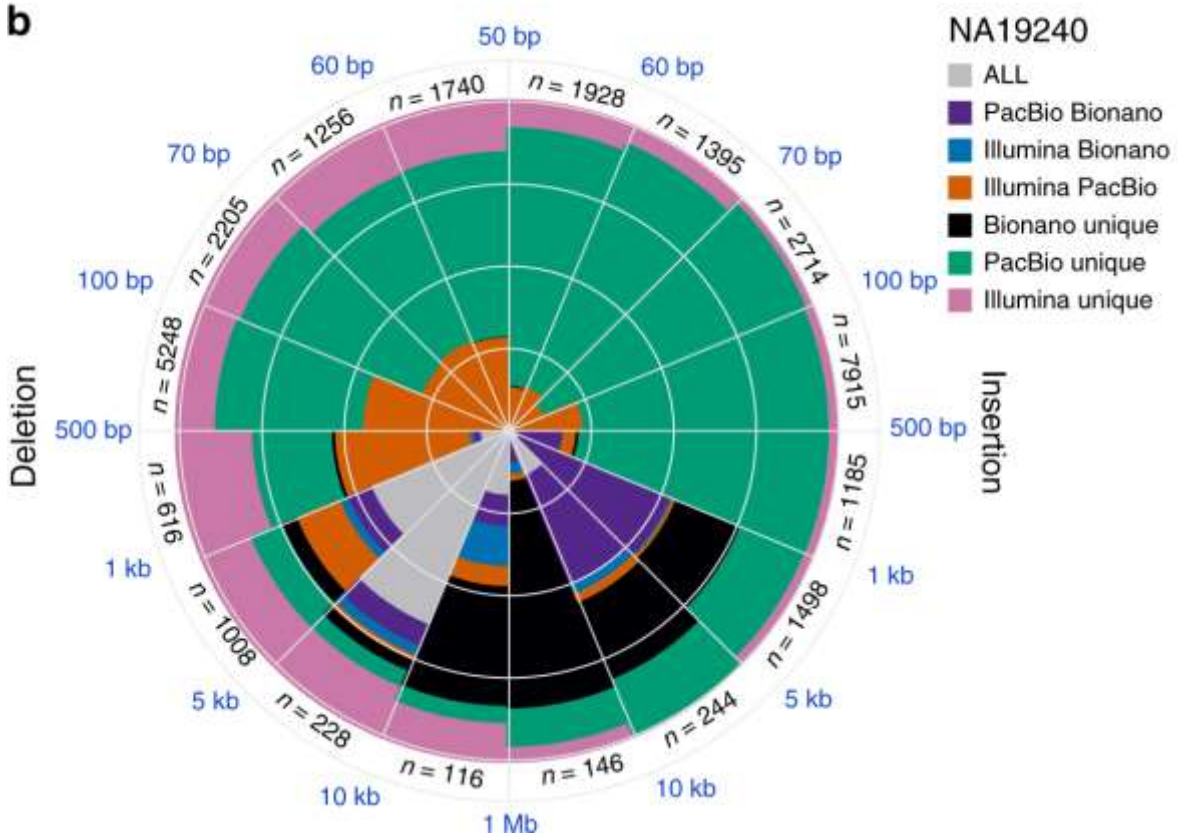
Long-reads/optical mapping
-> a few individuls per species

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

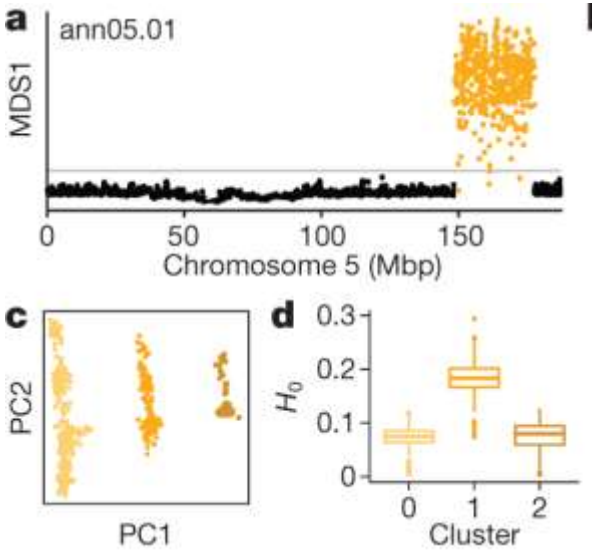
10kb->1MB: Bionano

20bp -> 1kb illumina + PacBio

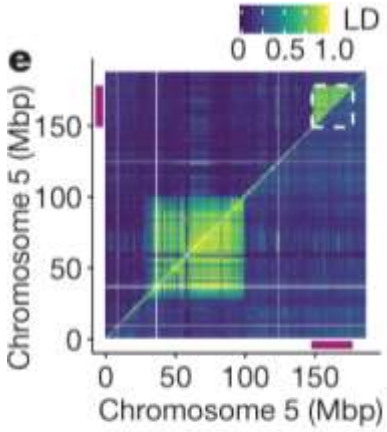
Short-reads only : just a fraction of Sv, more deletions tahn insertions

Direct detection : combining platforms

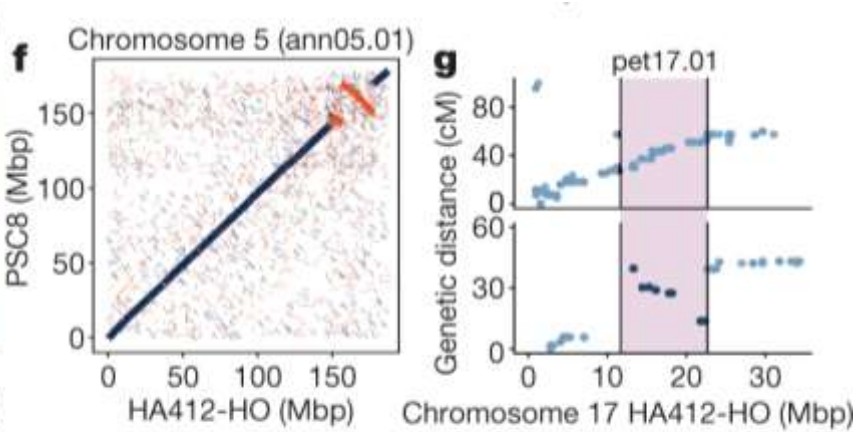
Indirect detection (local PCA)



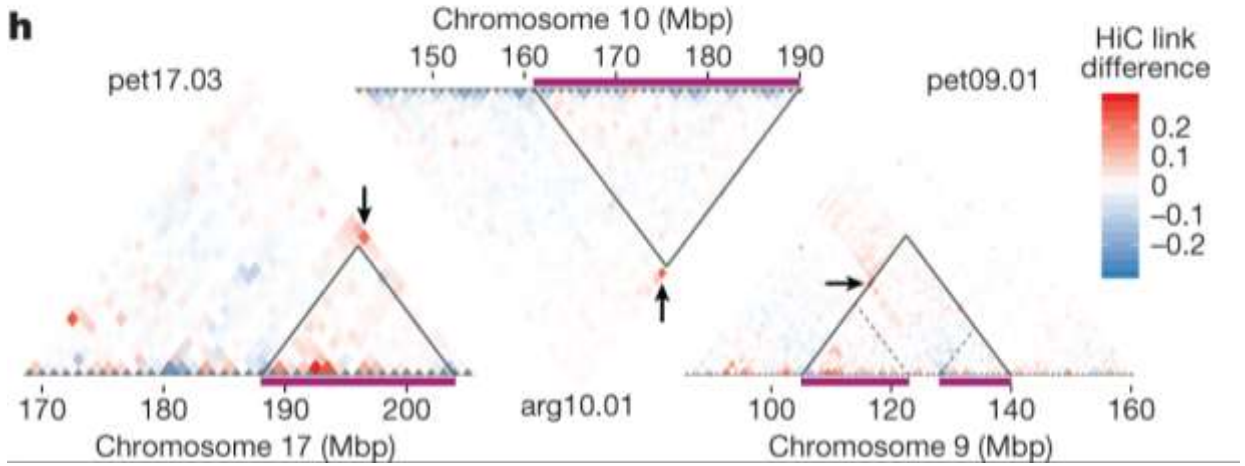
Indirect detection (LD)



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

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 rearrangement can drive artefactual population structure
  - Not the same interpretation if an islands of divergence is an inversion or not...

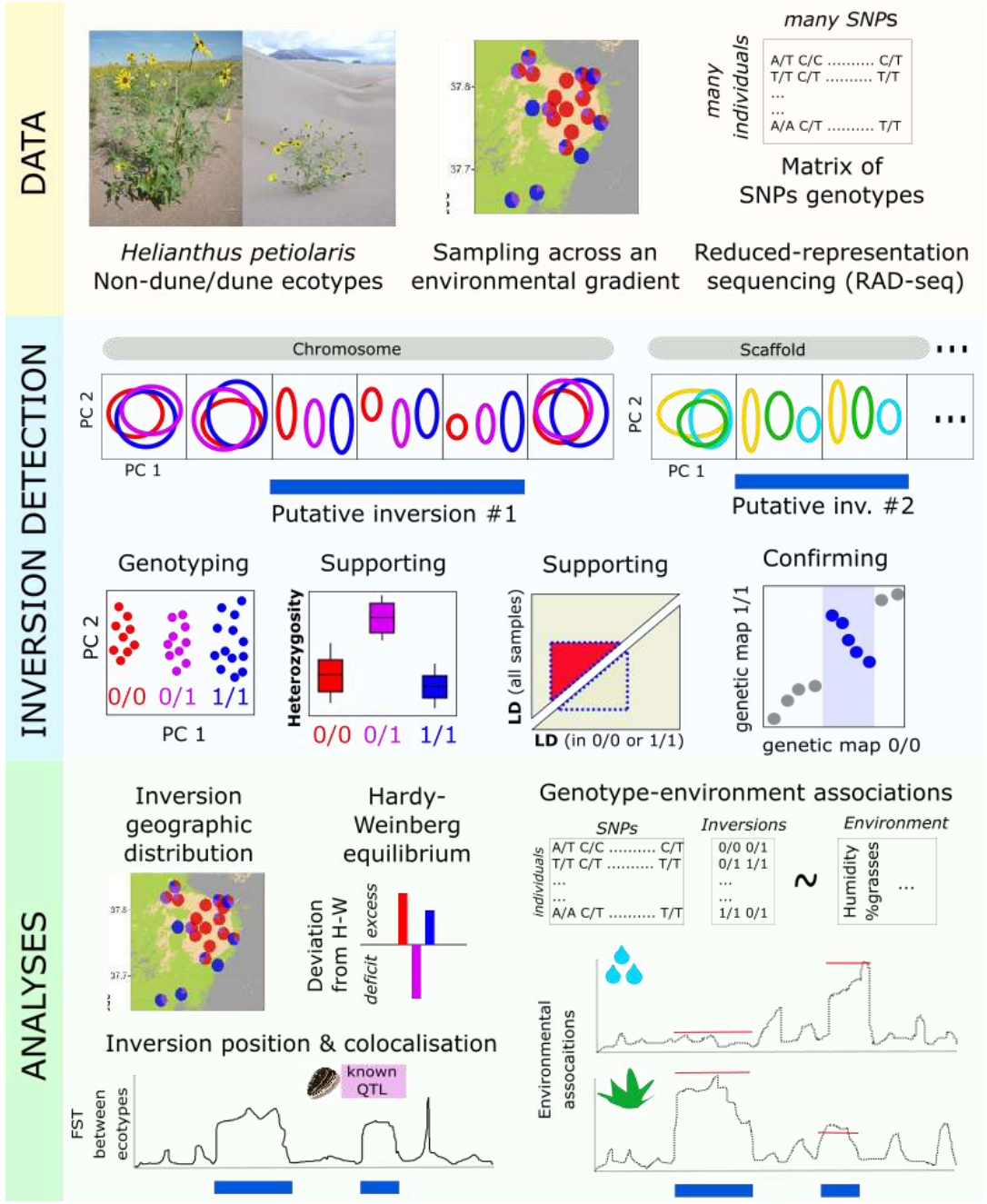
- Test the role of SV in adaptation
  - Evidence of adaptive SV are anecdotical...
  - Can we test which SV are putatively adaptive as we did on SNPs?

⇒ 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?



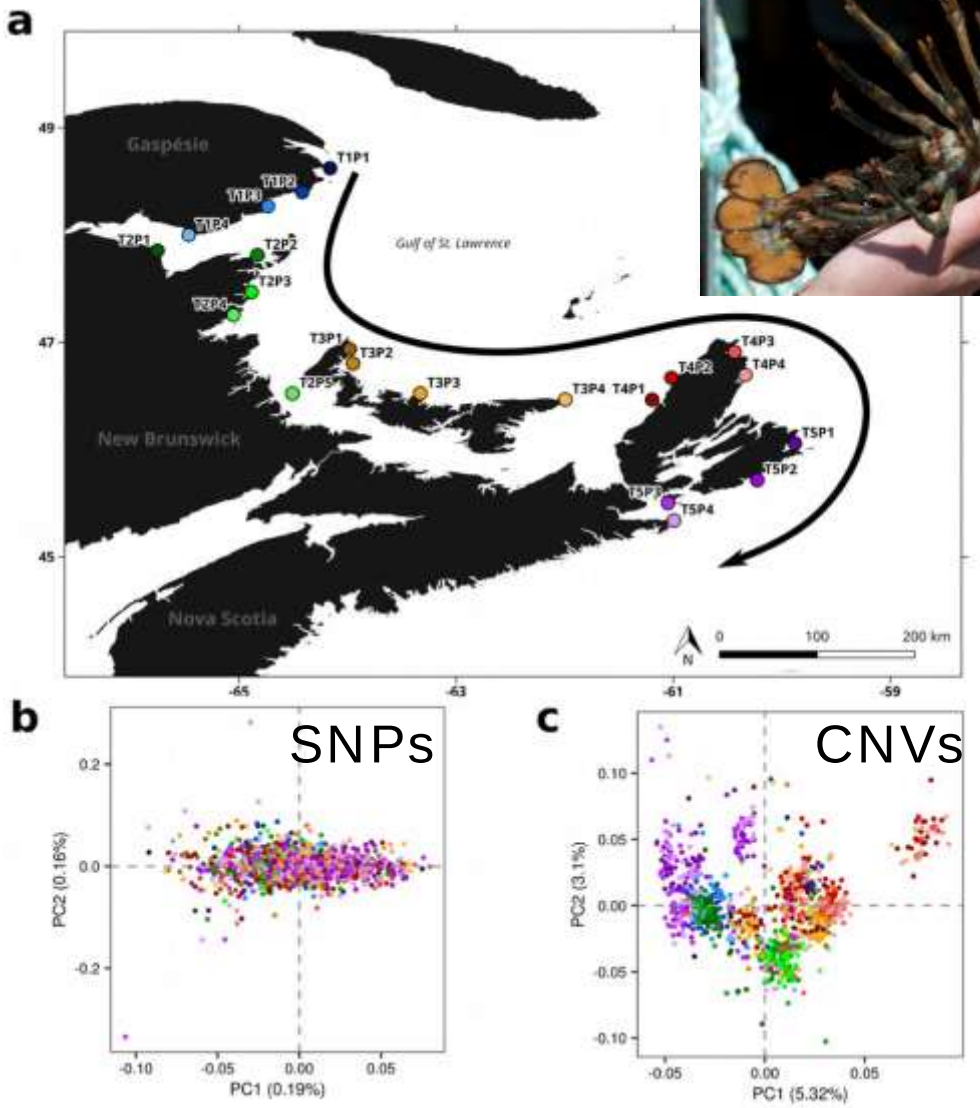
Mérot, 2020, Mol Ecol

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

Remaining challenges

- Large repetitive regions remain inaccessible due to constraints of read length and sequence composition
- Statistical tools for population genomics, adaptation genomics, ecological genomics are based on SNPs