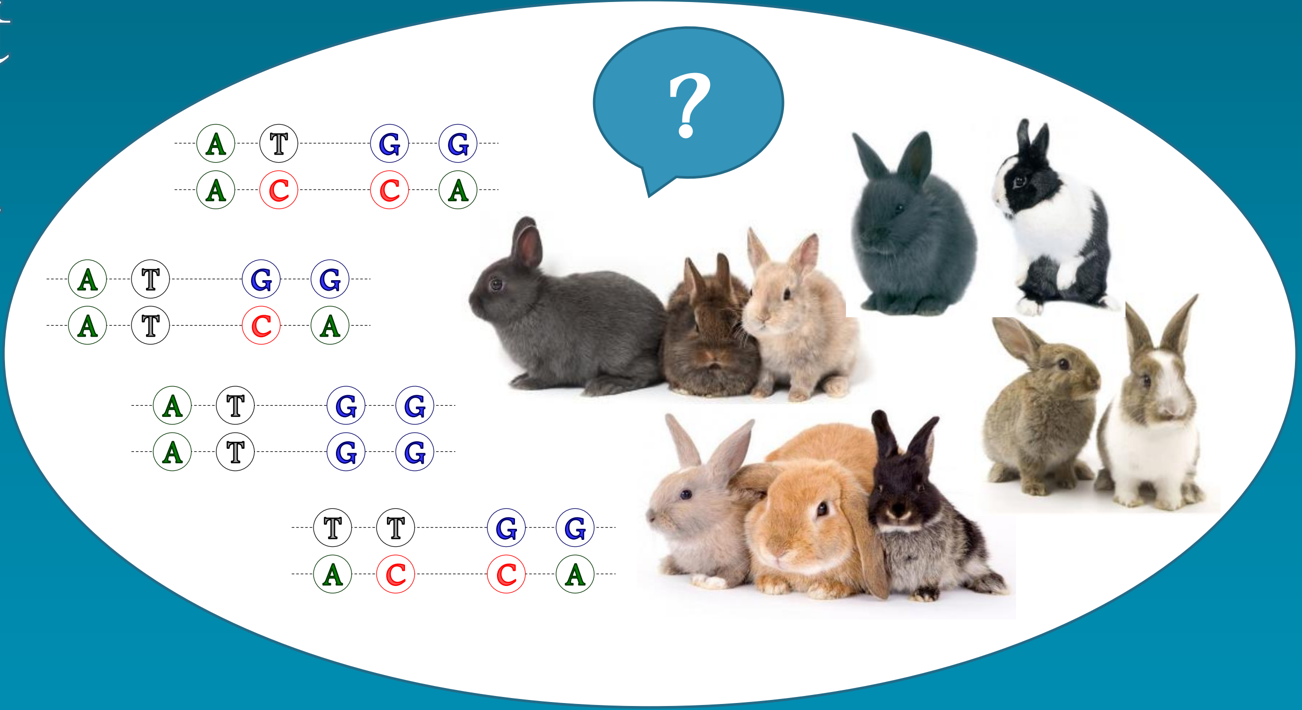
Phasing haplotypes in rabbit using long reads technology

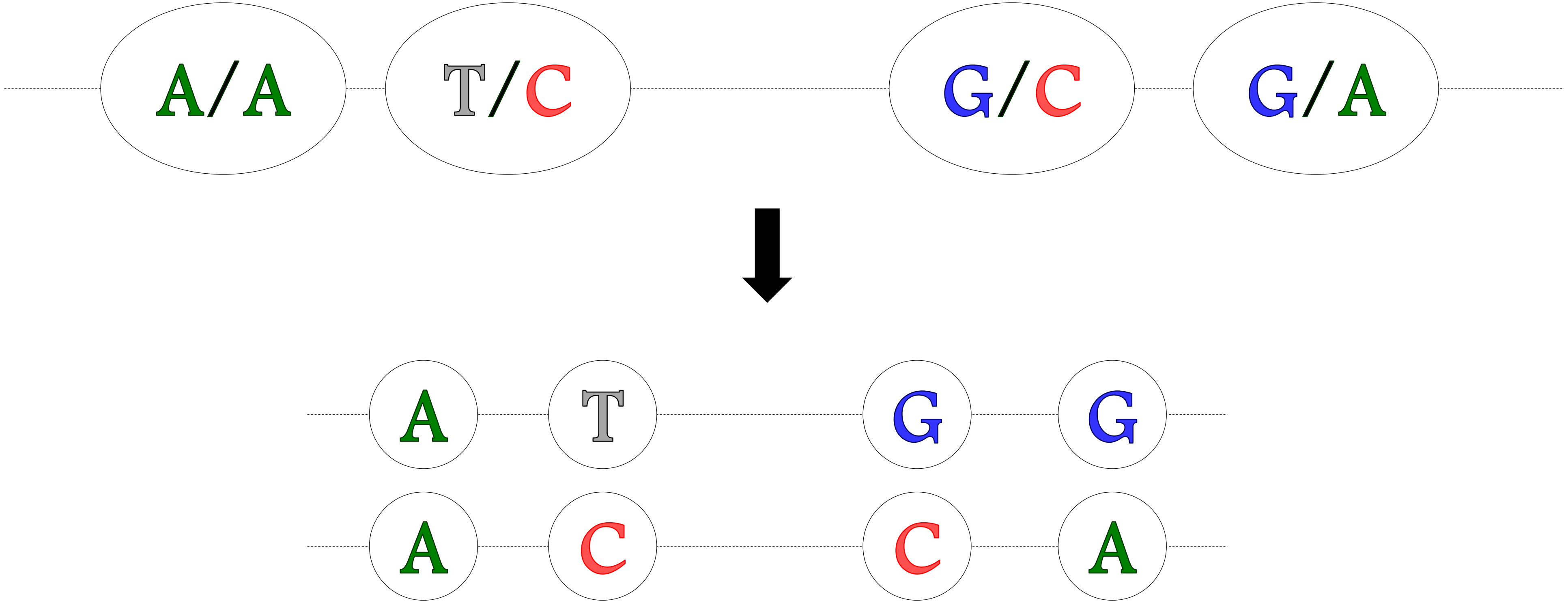


Julie Demars
GenPhySE-Genetics, Physiology and Livestock Systems

Long reads : Dream or Reality ?
28th of November 2017

http://get.genotoul.fr
get@genotoul.fr
@get_genotoul

Genome phasing



Fundamental aspect of genetics that is relevant in many applied problems

Interest of genome phasing

- **Allele-specific expression or methylation**

Genomic imprinting

- **Highly heterozygous regions**

HLA genes

- **Population genetics and genome-wide association studies**

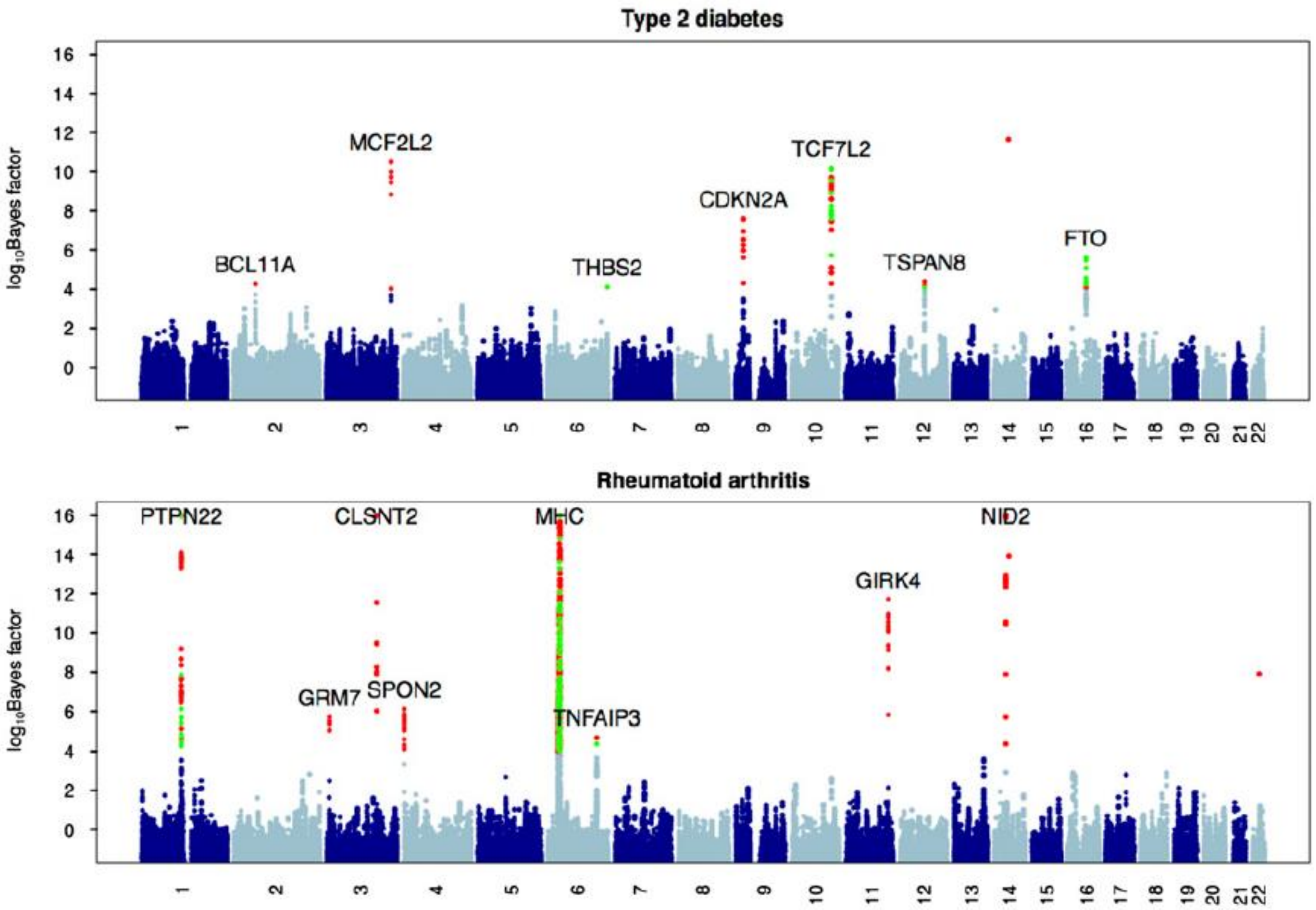
Polygenic traits, Allelic heterogeneity

Interest of genome phasing

Genetics. 2014 Jul;197(3):823-38. doi: 10.1534/genetics.114.164814. Epub 2014 May 8.

Detecting local haplotype sharing and haplotype association.

Xu H[1], Guan Y[2].

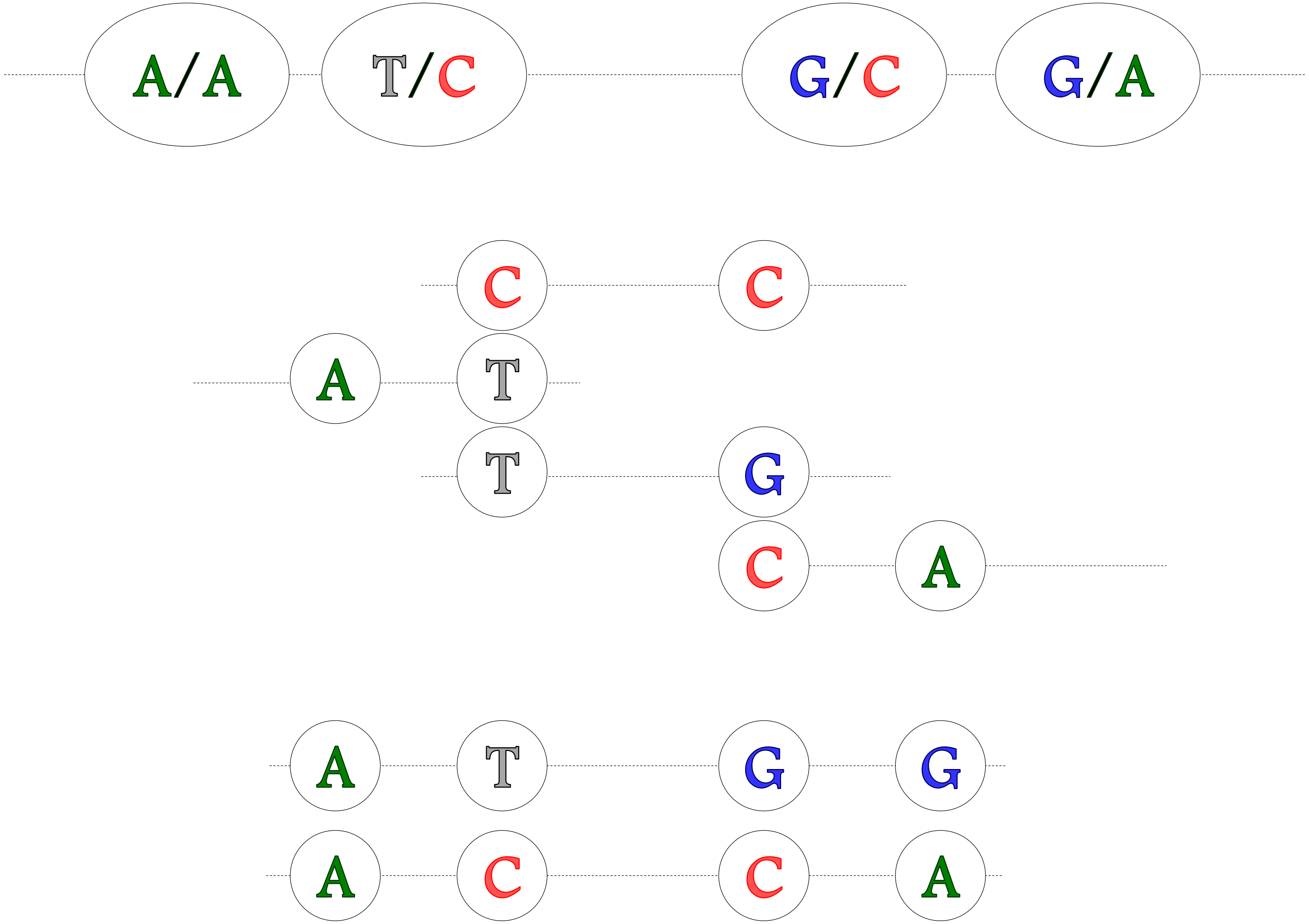


General principle of genome phasing

Unphased genome

Sequence reads

Phased results



Genome phasing : methods and developments

- **Scale of datasets**

From SNP beadchips to whole genome

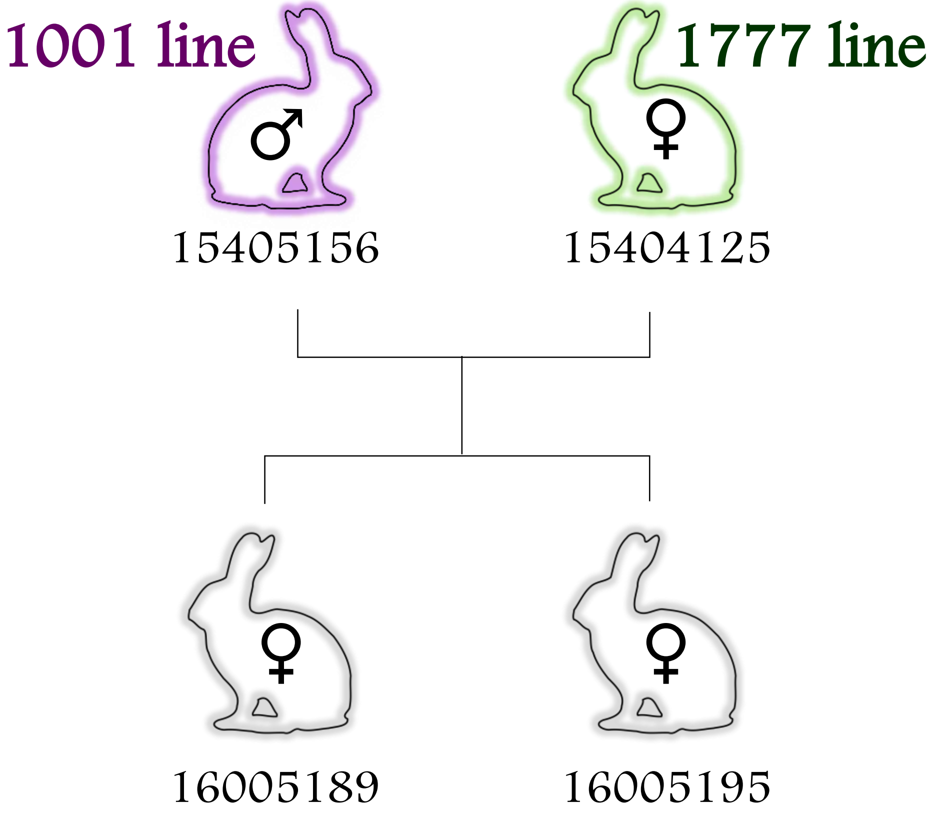
- **Statistical models and computational approaches**

Time consuming : balance quality of phasing/computational cost

- **Laboratory-based experimental methods**

Long reads produce virtual multi-kilobases reads on regular sequencers

Evaluation of rabbit genome phasing

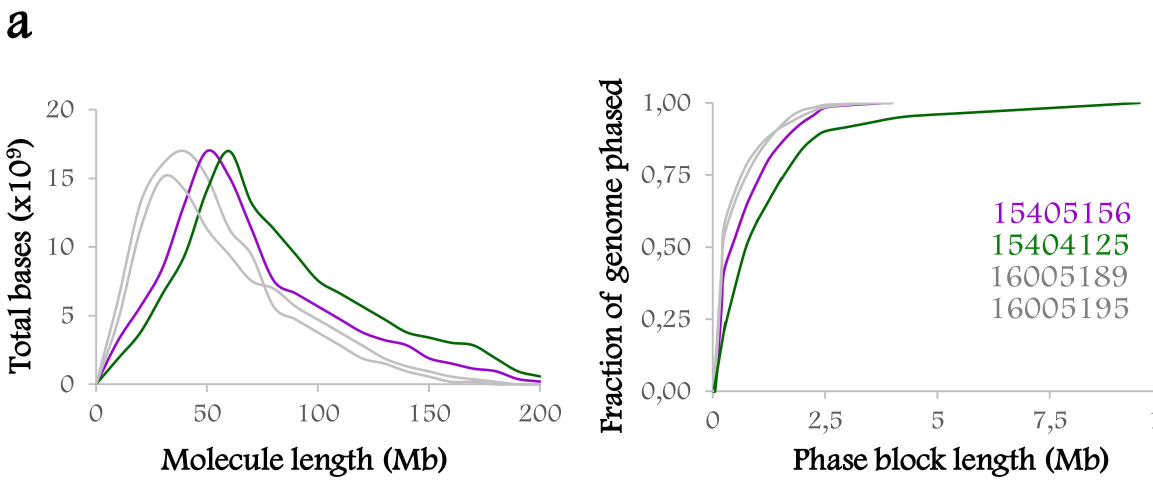


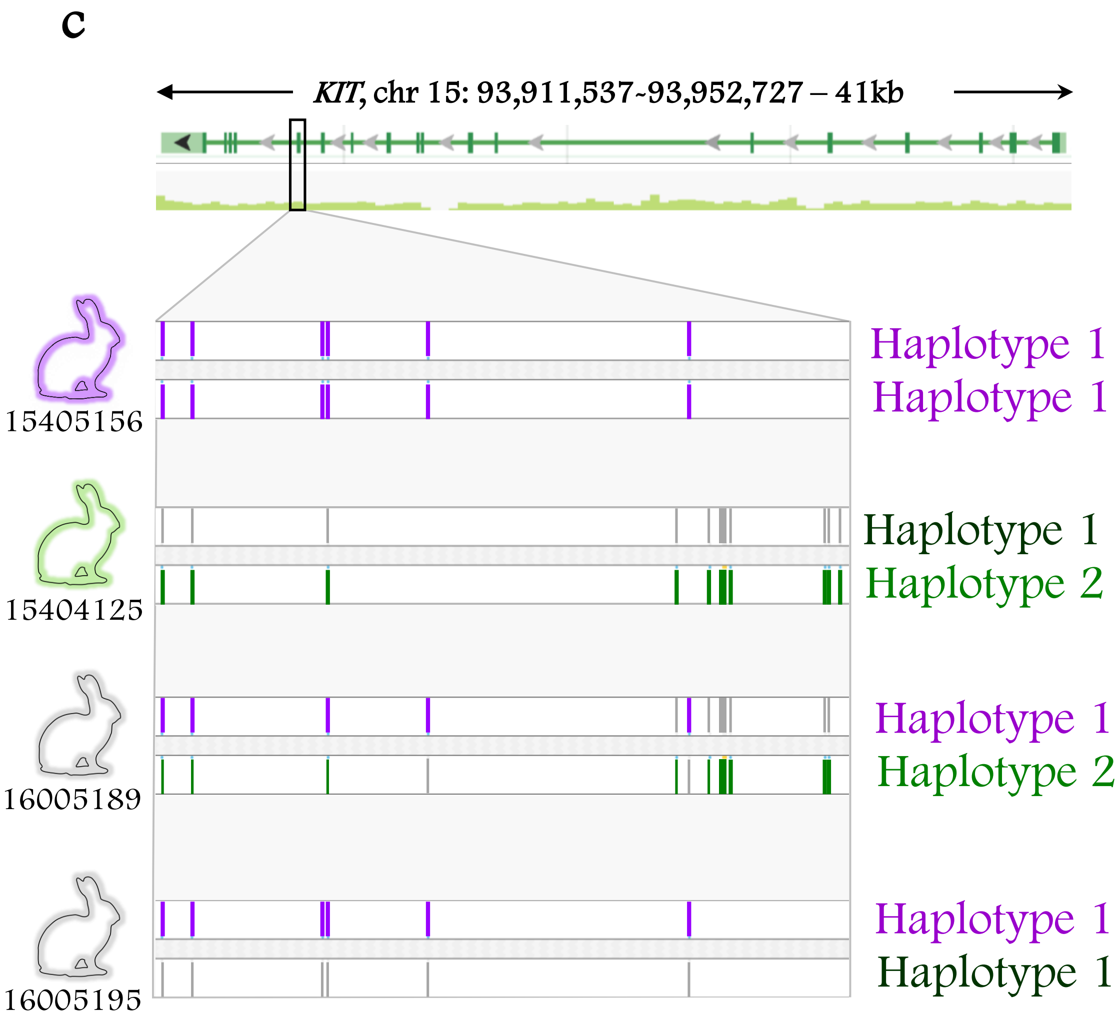
Rabbit Reference Genome (Oryctolagus cuniculus)	
Version	OryCun2 (GCA_000003625.1), 1st assembly
Reference rabbit	New Zealand
Sequencing Depth	7X
Chromosomes	21
Size (Mb)	2247,75
Genes	20999
Scaffolds	3318
Size (Mb)	489,69
Genes	8099

Summary of phasing results

Phased WGS	15405156 Male Father	15404125 Female Mother	16005189 Female Offspring	16005195 Female Offspring
Fold coverage	32	28	26	28
% aligned	81	80	80	77
% duplication	4,42	6,09	6,30	6,66
Relative genomic equivalents per partition	0,362	0,321	0,342	0,508
Number of molecules (millions)	1,59	1,59	1,40	1,34
Length-weighted mean molecule length (kb)	63,17(±41,95)	73,90(±48,02)	45,73(±30,30)	54,06(±43,98)
% SNPs phased	97	97	98	97
N50 phase block (kb)	875,49	1086,73	609,92	838,05
Longest phase block (kb)	4 417,55	9 930,61	3 725,93	4 599,55
Long structural variants (number)	162	183	197	155
Short structural variants (number)	49 846	45 901	46 265	47 723

Phase the « full » spectrum of called variants

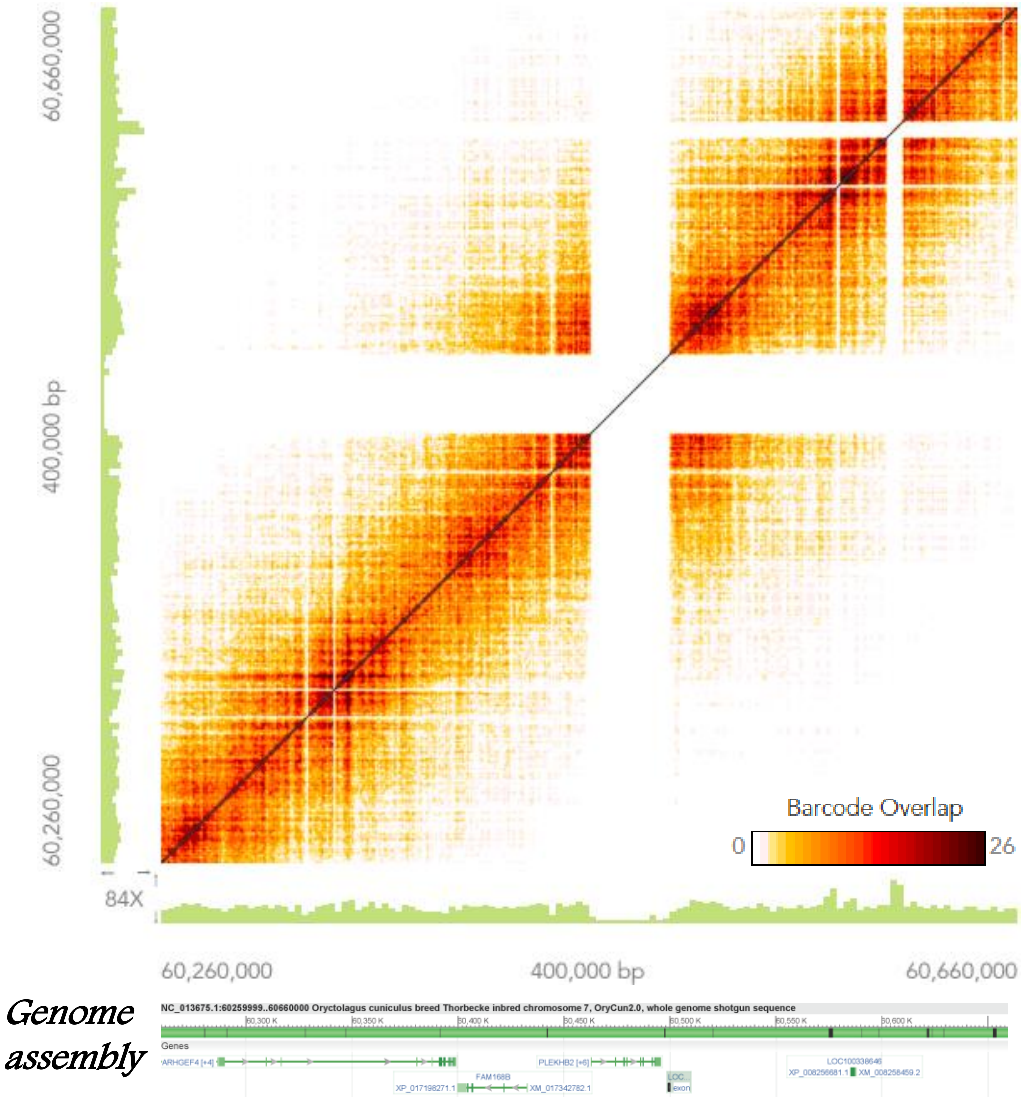


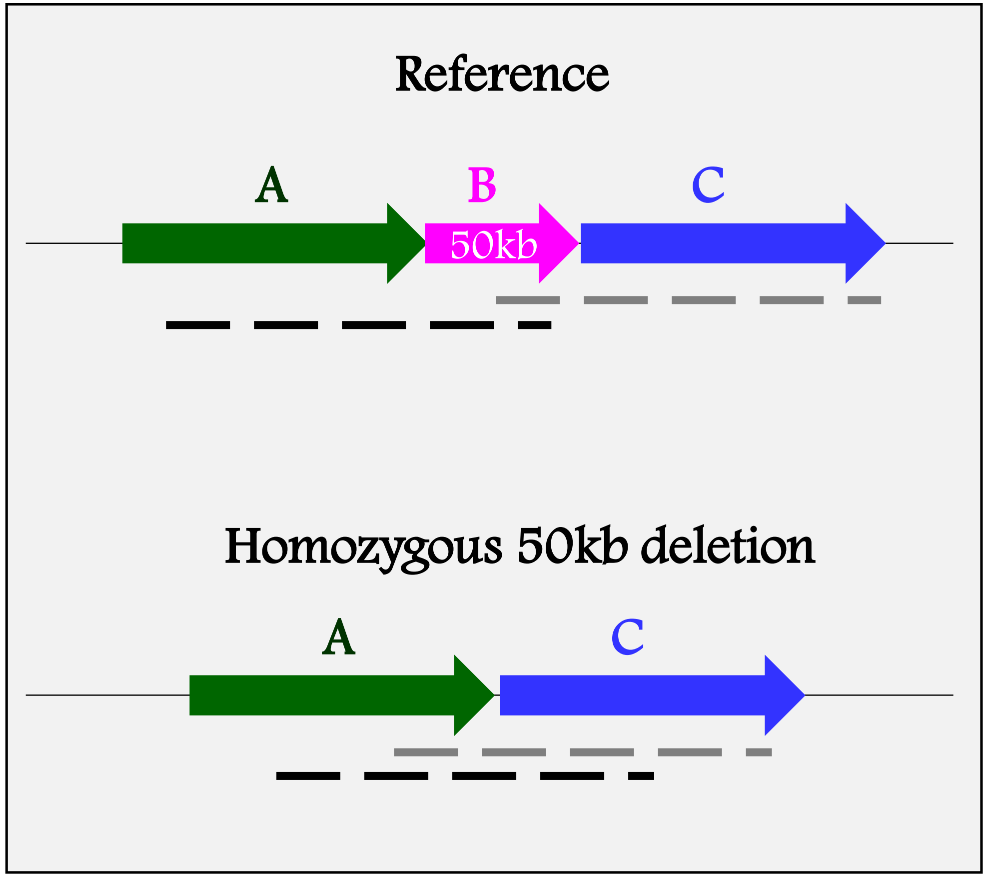


Phase structural variants : large deletion

Homozygous deletion

chr7



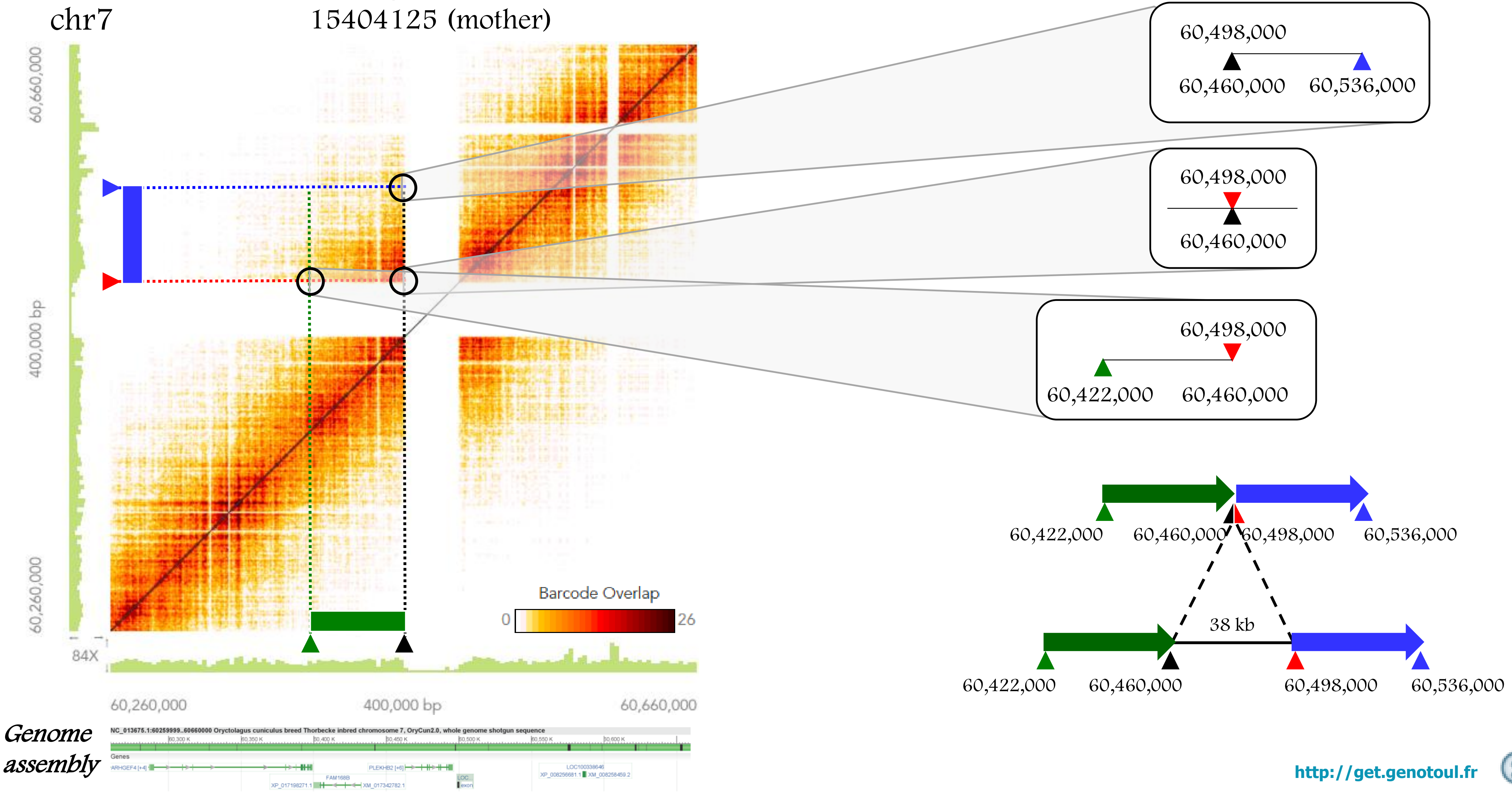


Phase structural variants : large deletion

Homozygous deletion

chr7

15404125 (mother)



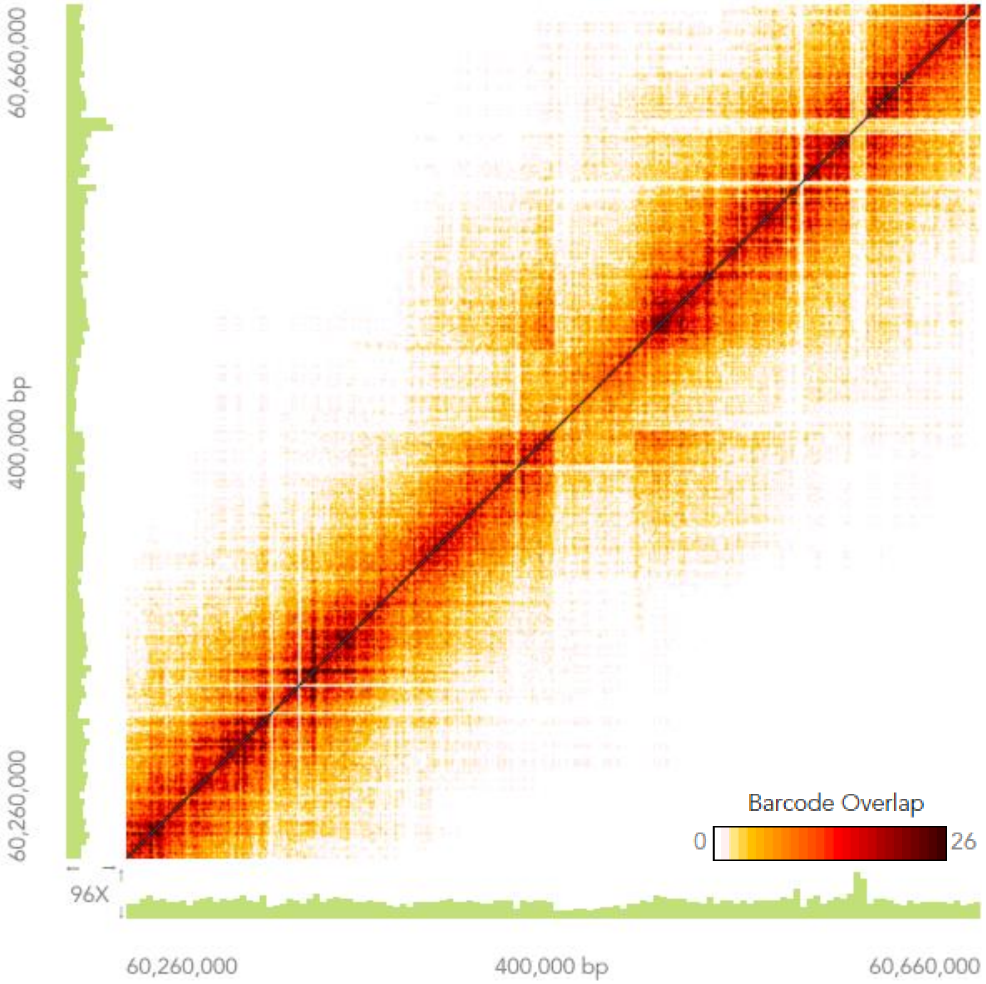
Genome assembly

Phase structural variants : large deletion

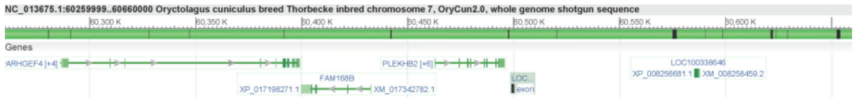
Heterozygous deletion

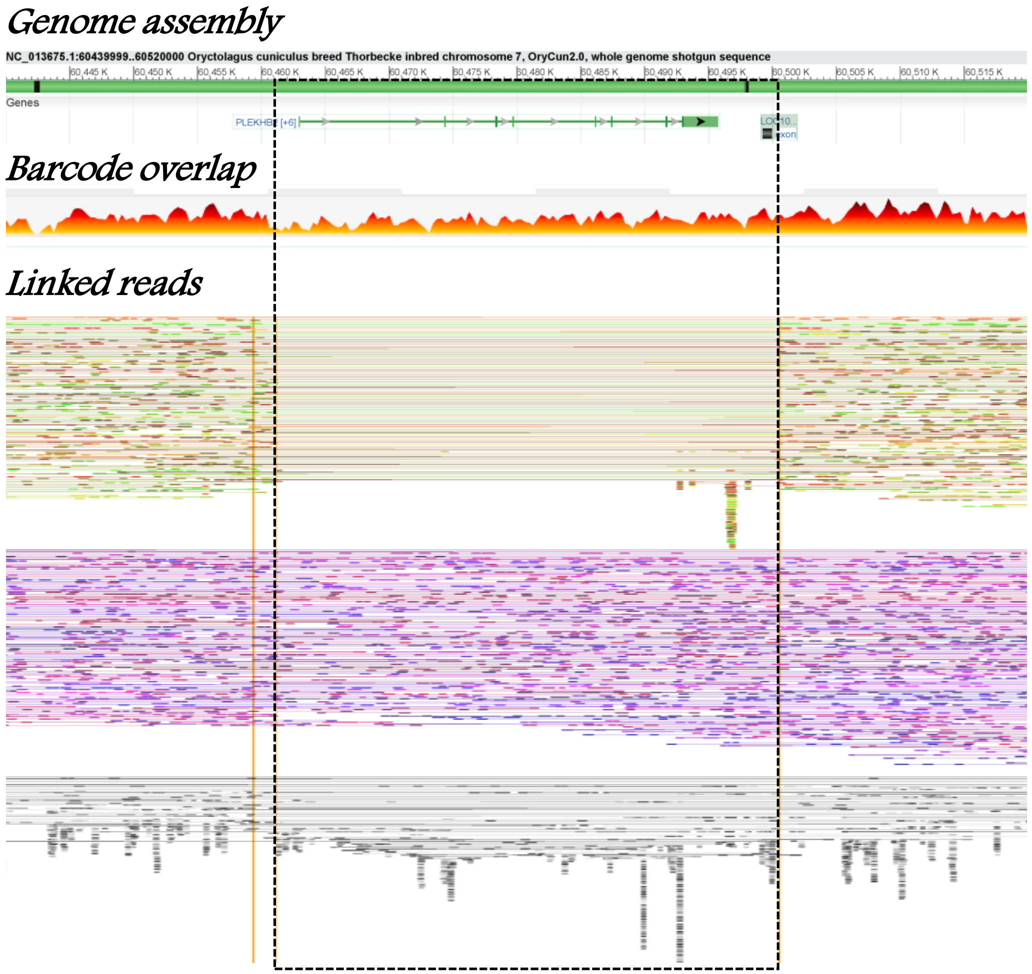
chr7

15405156 (father)



Genome assembly



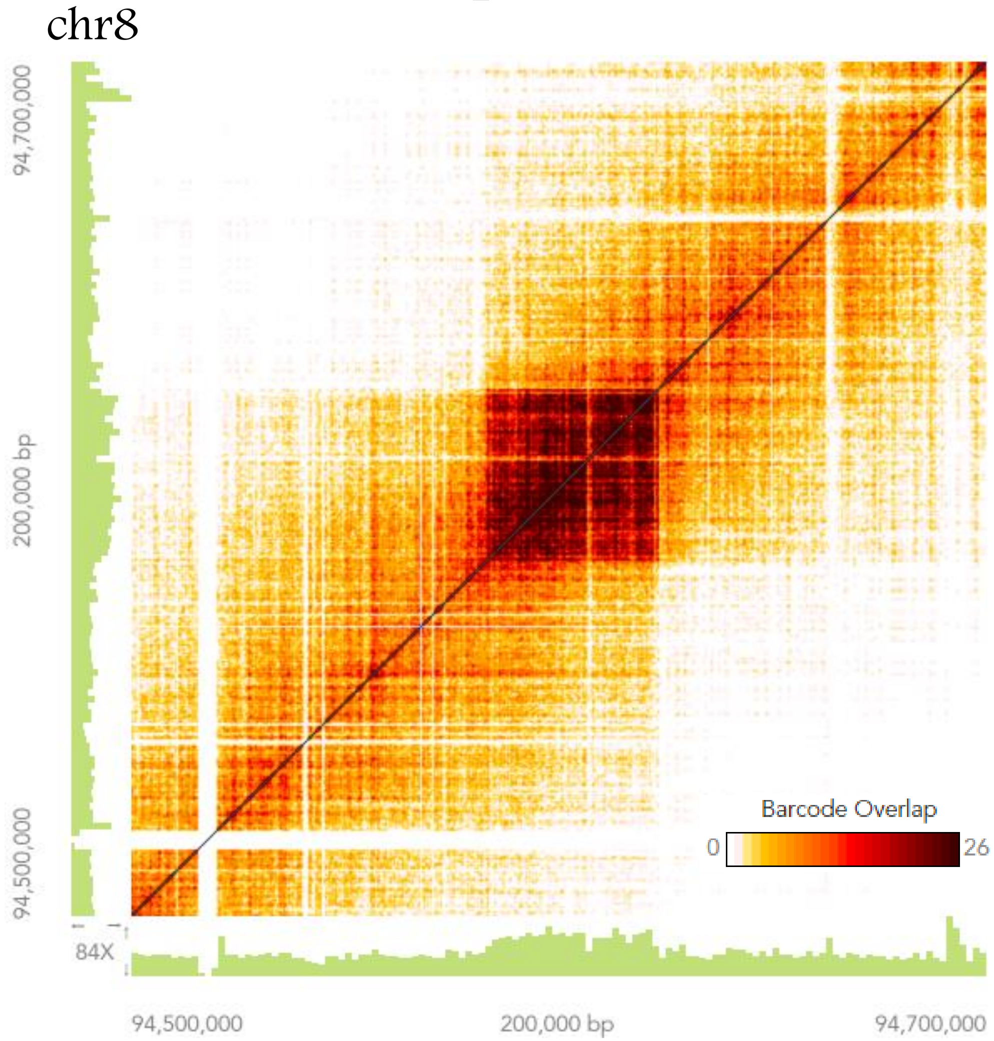

Genome assembly

Barcode overlap

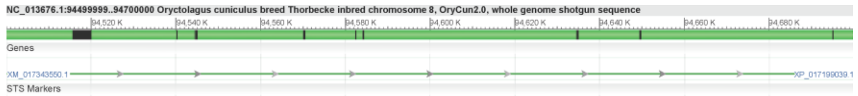
Linked reads

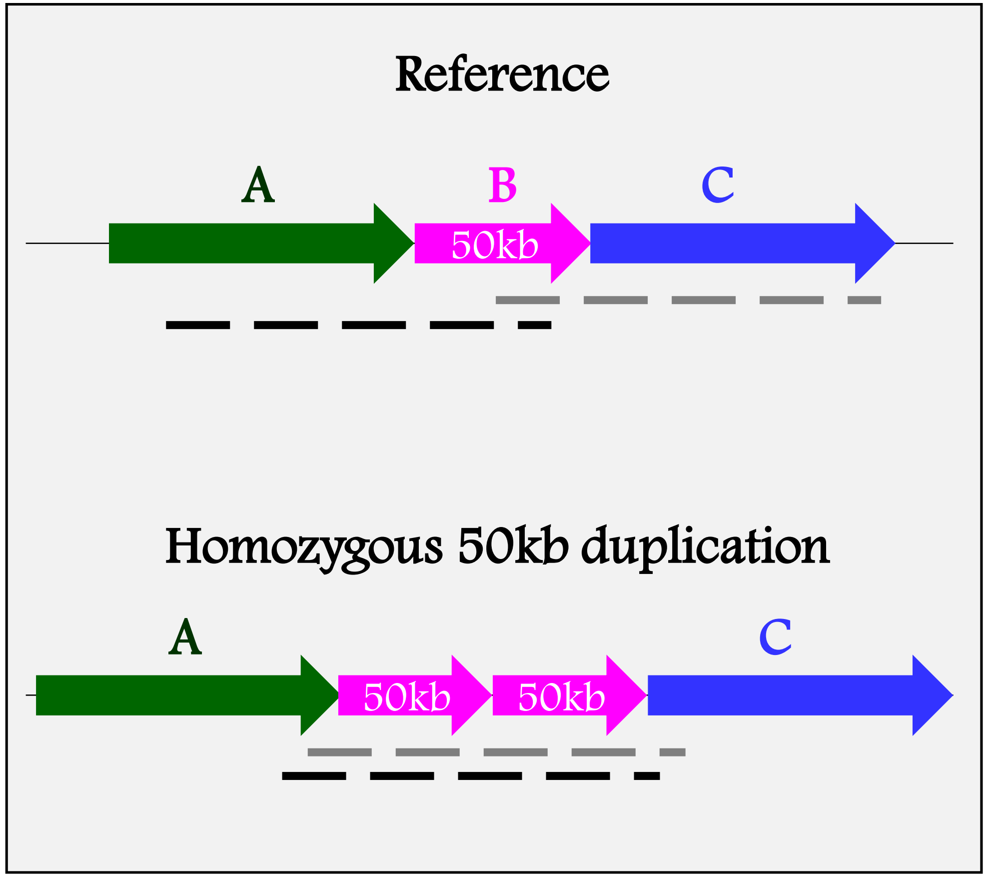
Phase structural variants : tandem duplication

Duplication



Genome assembly

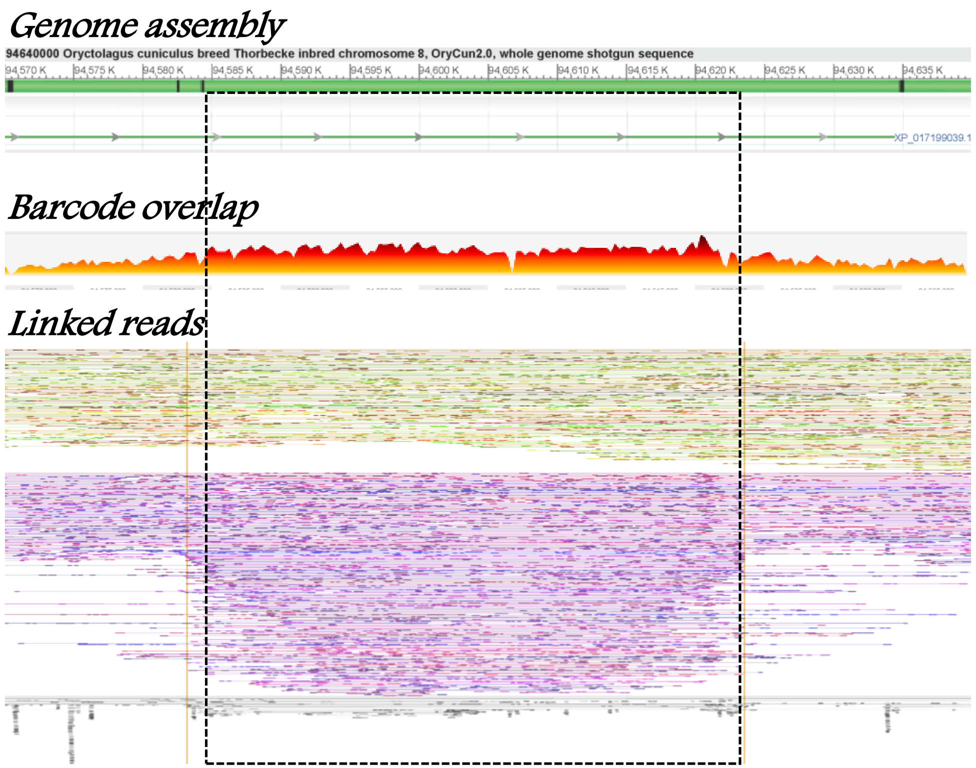


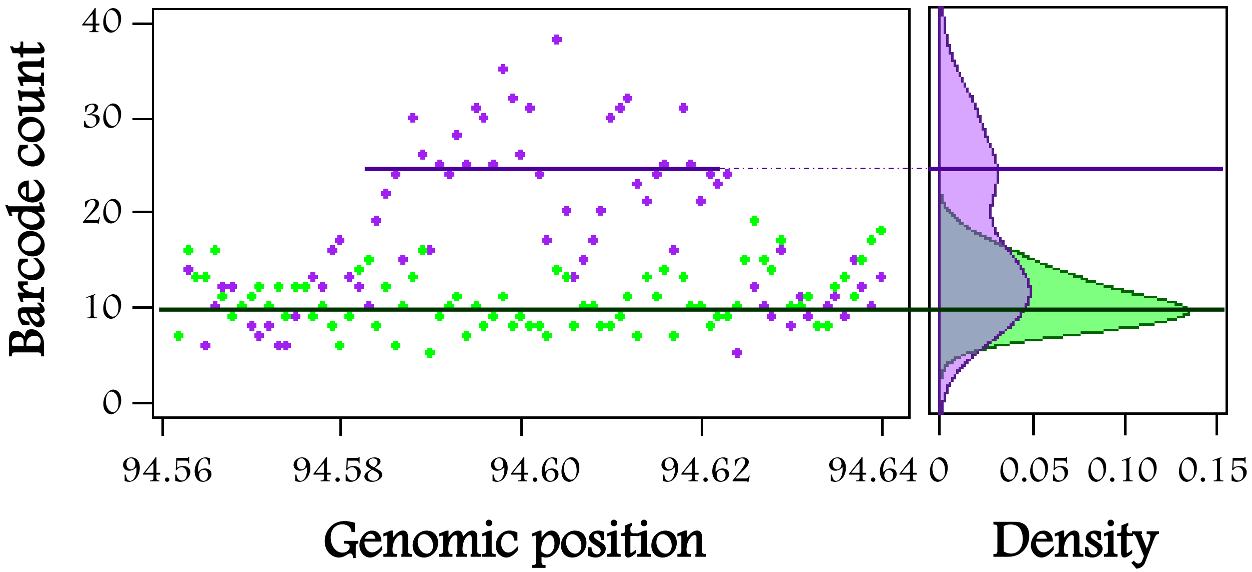


Phase structural variants : tandem duplication

Heterozygous duplication

15404125 (mother)

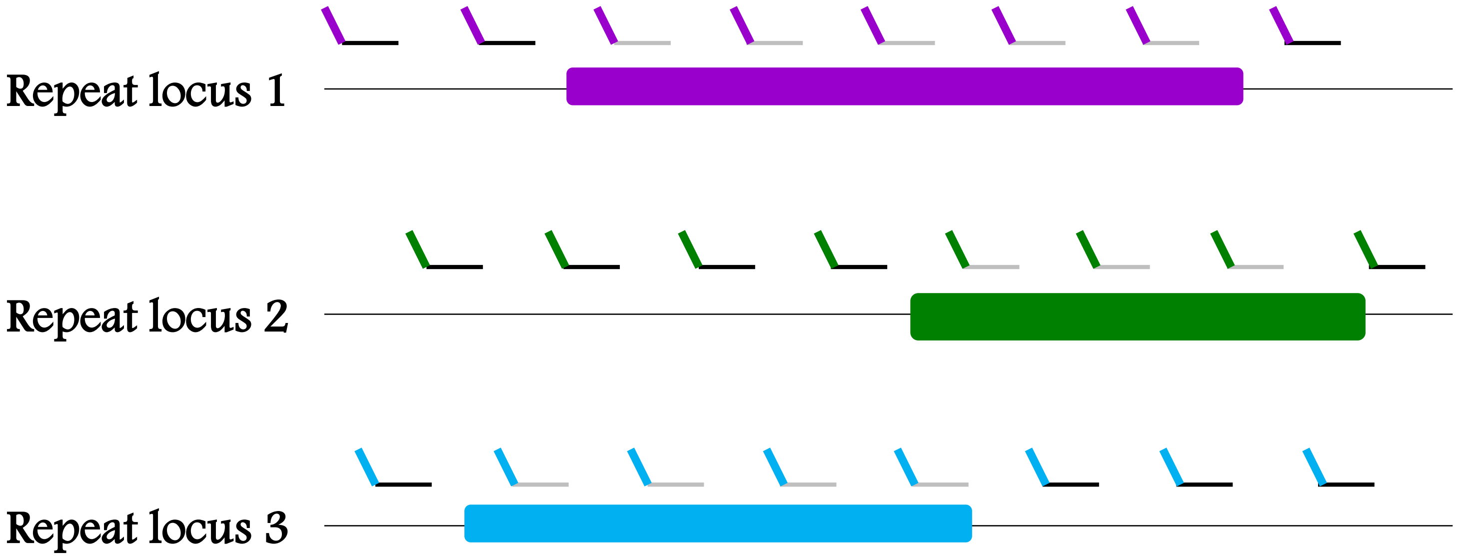




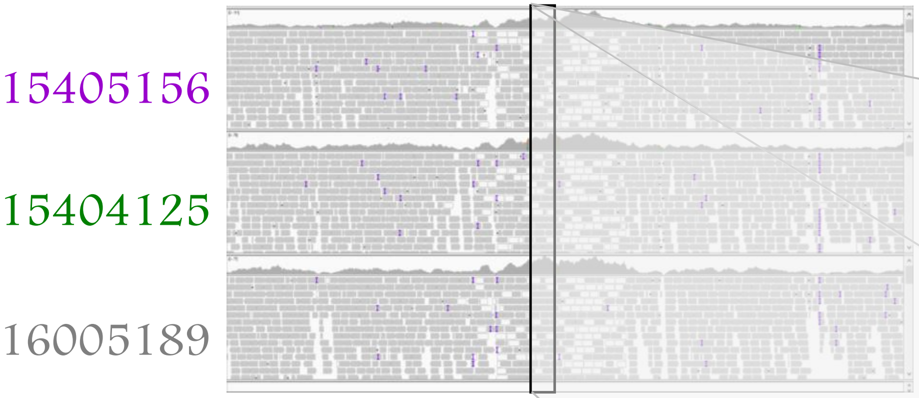
Rescue repetitive regions

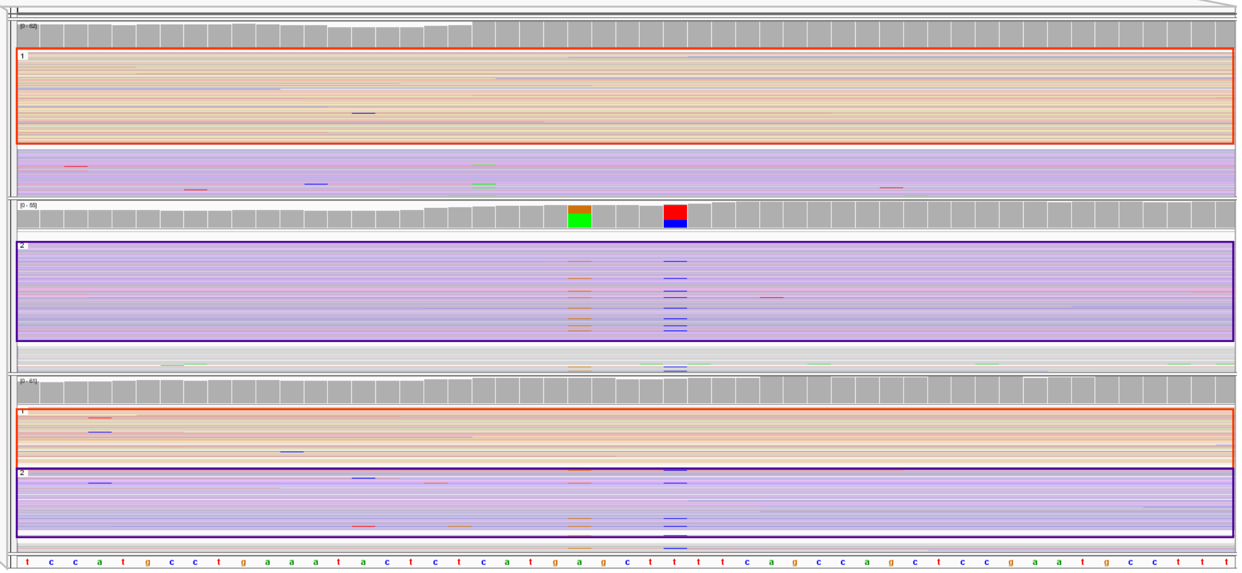
- Problem to assign reads in repetitive elements

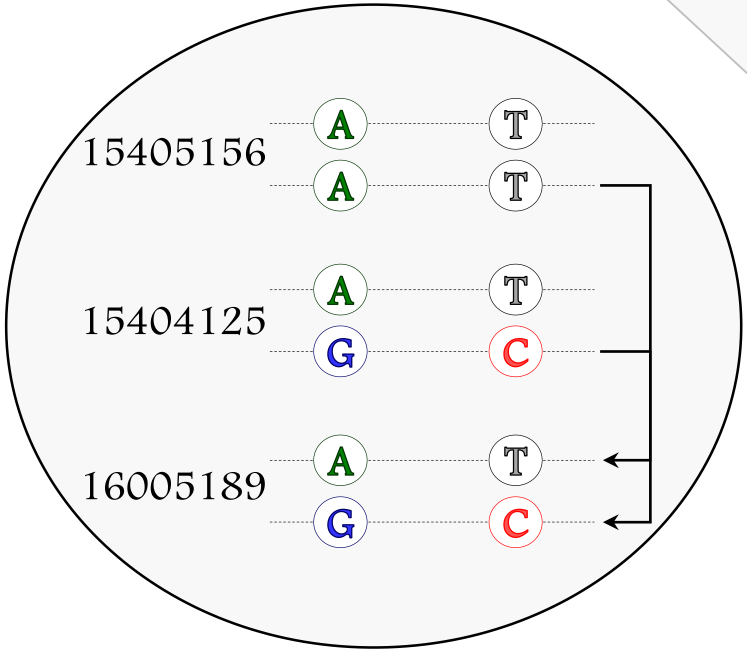
- Information from 10X barcodes rescue unmapped reads



Recover variants in repetitive regions







Linked reads and the 10X technology

- **Resolve the genome into long (megabases) phase blocks**

Phase the “full” spectrum of called variants

- **Identify structural variants and breakpoints**

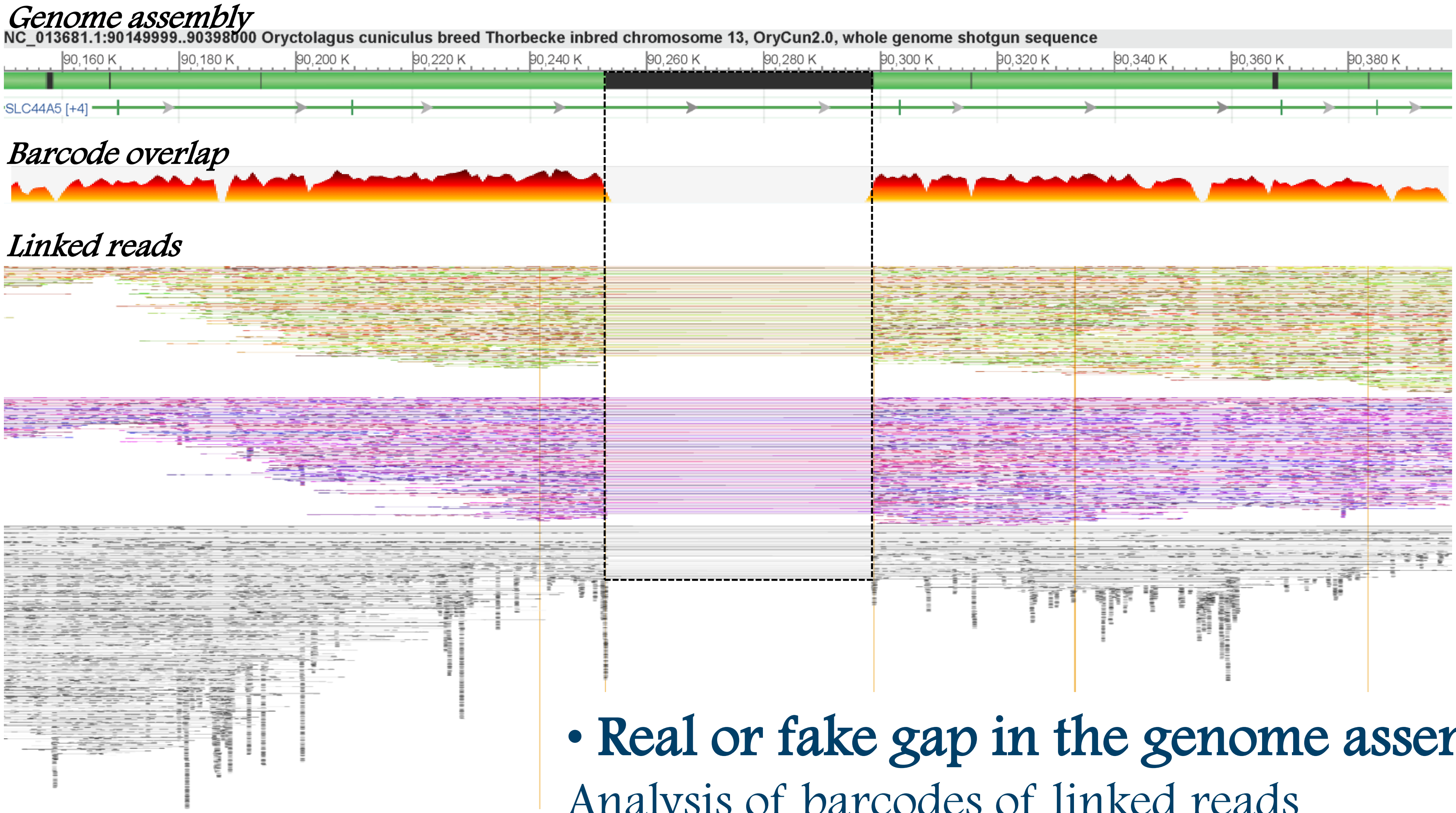
Insertions, deletions, duplications, translocations…

- **Rescue variants in inaccessible parts of the genome**

Confidently map reads in repetitive regions

- **Improve genome assembly**

Improve genome assembly : gap resolution



- **Real or fake gap in the genome assembly ?**

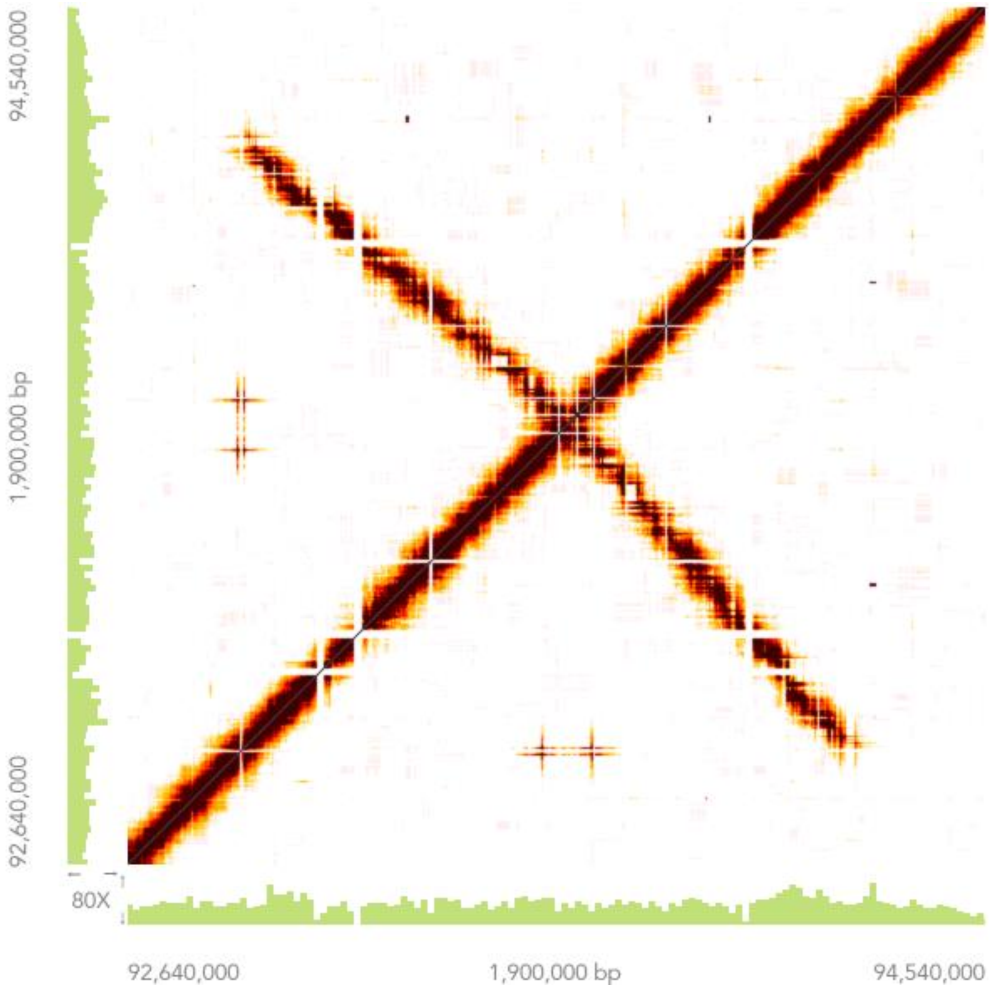
Analysis of barcodes of linked reads

Improve genome assembly : inversion identification

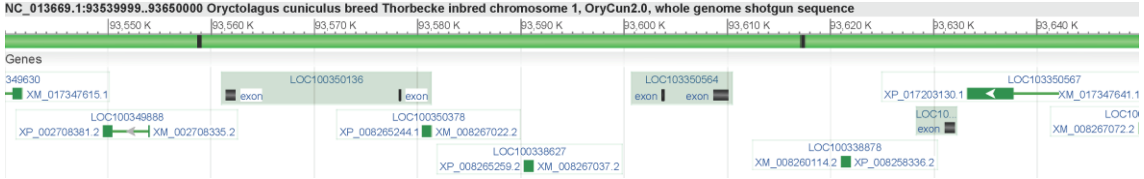
Complex rearrangement

including inversion

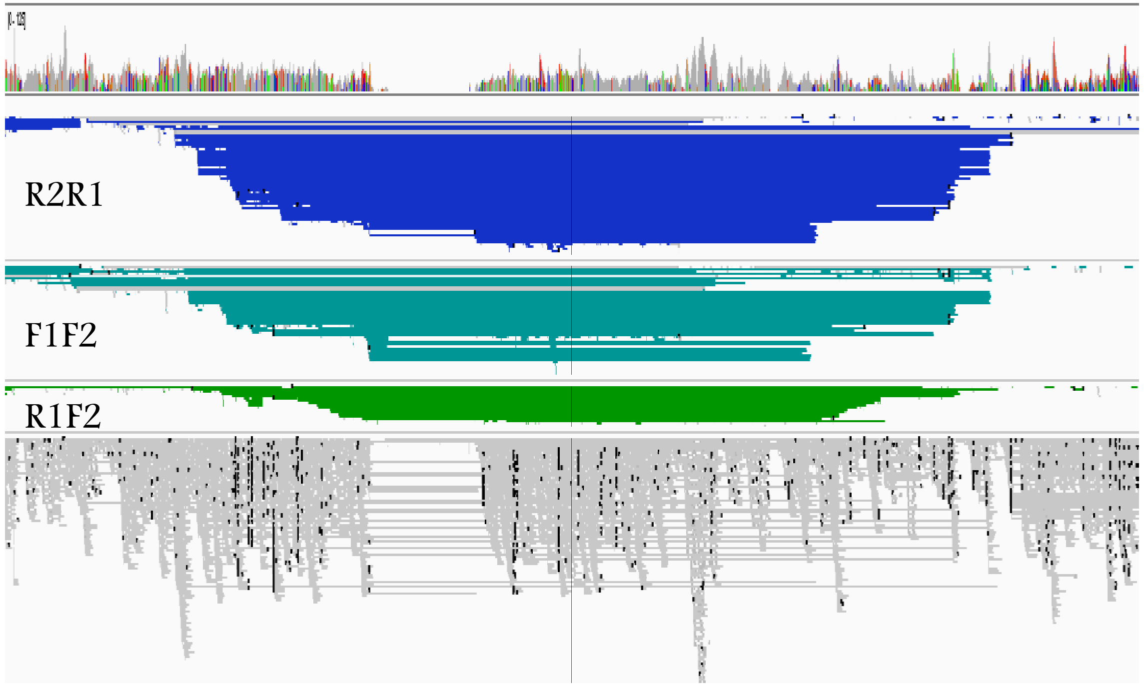
chr1



Genome assembly


NC_013669.1:93539999..93650000 Oryctolagus cuniculus breed Thorbecke inbred chromosome 1, OryCun2.0, whole genome shotgun sequence

IGV window (pair orientation)



Acknowledgements

- Genotoul GeT-PlaGe

Olivier Bouchez
Cécile Donnadieu
Pauline Heuillard
Claire Kuchly
Sophie Valière

- GenPhySE

Isabelle Hochu
Sophie Leroux
Juliette Riquet
Julien Ruech

- PECTOUL

Patrick Aymard
Elodie Balmisse
Virginie Helies