



# Variations de structure et longues lectures

Thomas Faraut

## ► To cite this version:

Thomas Faraut. Variations de structure et longues lectures. Colloque INRAE Genomics, INRAE Genomics, May 2022, Orléans, France. hal-04220082

**HAL Id: hal-04220082**

**<https://hal.inrae.fr/hal-04220082>**

Submitted on 27 Sep 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Variations de structure et longues lectures

Colloque INRAE Genomics

---

Thomas Faraut

Orléans, 17 mai 2022

- Landscape of structural variations with short and long reads

What is the relative contribution of insertions and deletions to the structural variability of genomes ?

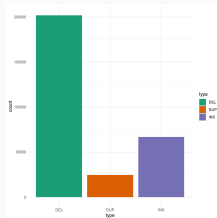
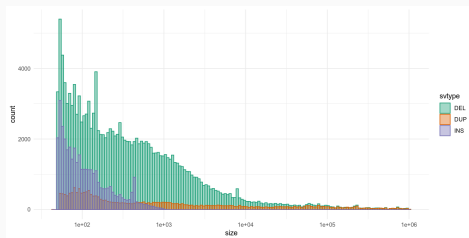
- The case of indels
- The case of large insertions and deletions

From short reads to long reads and back

- The variation graph

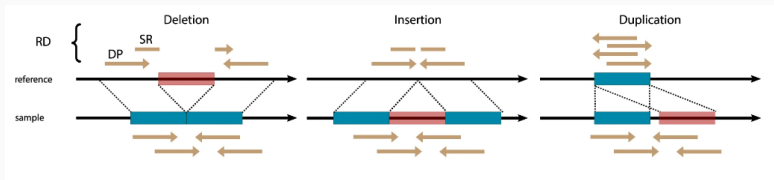
# SeqOCCln SV: the case of short reads

1000 goats (Vargoaats project)



- Deletions clearly outnumber insertions

## Structural variation signatures

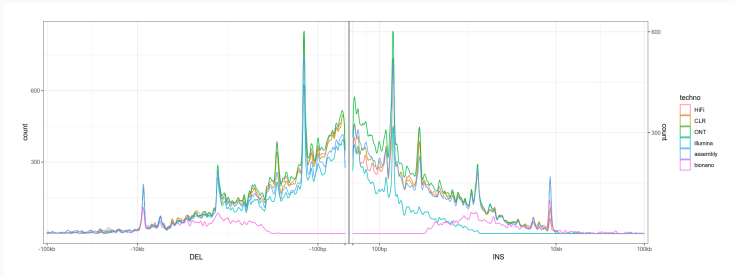


RD: Read depth

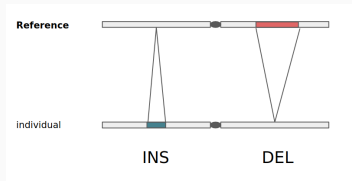
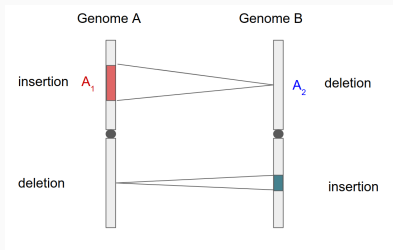
DP : Discordant pairs

SR : Split reads

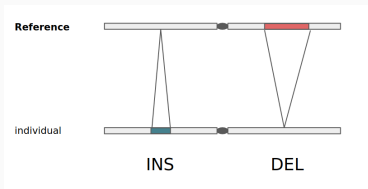
van Belzen *et al.* npj Precis Onc 5 (2021). <https://doi.org/10.1038/s41698-021-00155-6>



- Symmetry of the size distribution of insertion/deletions is a sign of good health of the detected variants



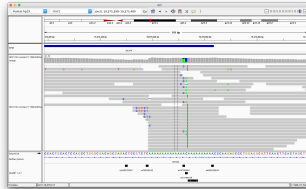
- Structural variation type INS and DEL are somehow ill-defined
- For small insertions/deletions the term indel preserves this ambiguity



- When comparing two genomes we expect to see a symmetric pattern, the same number of INS and of DEL

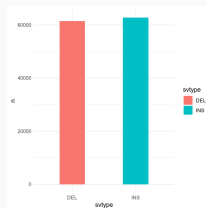
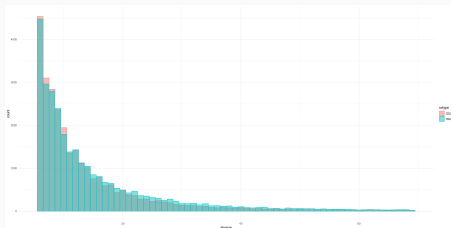


```
AGGCTCCTCCGTCCATGCGATT - GCAGGCC
||| | | | | | | | | | | | | | |
AGTC - - - TCCGTCCATGGGATTTTGCAGGCC
```



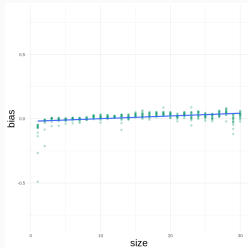
- Indels are detected directly from alignments

## Trio2 heifer

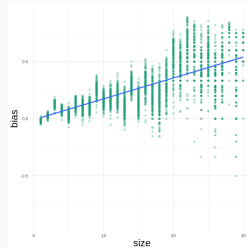


- With PacBio HiFi, the distribution of indel sizes exhibit a clear symmetric pattern

PacBio HiFi

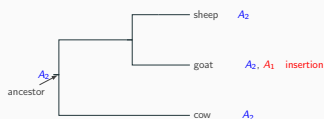
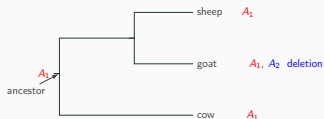
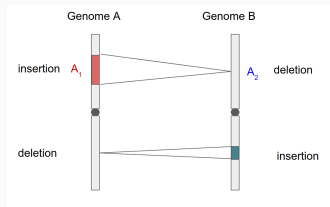


Illumina



$$bias = \frac{n_d - n_i}{n_d + n_i}$$

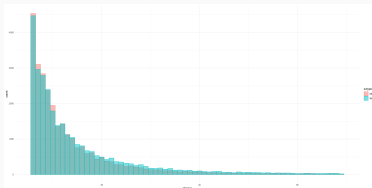
- PacBio HiFi seems unbiased, while a clear bias is observed for Illumina



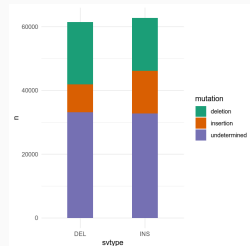
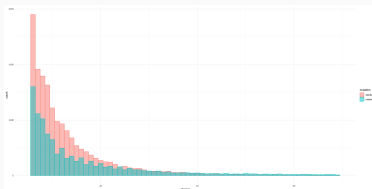
- The identification of the ancestral state is mandatory to identify the mutation, deletion or insertion for a given variant

## Heifer

## Variation



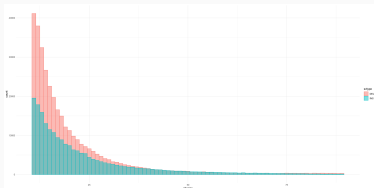
## Mutation



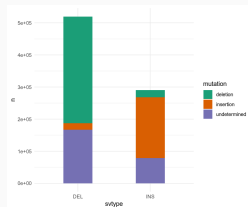
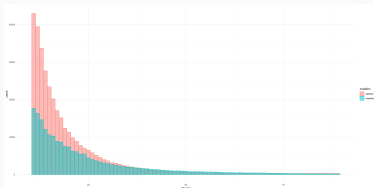
- Short deletions clearly outnumber short insertions

## 1000 Bull genomes project

Variation



Mutation

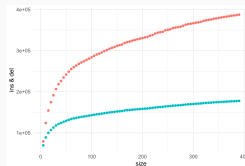
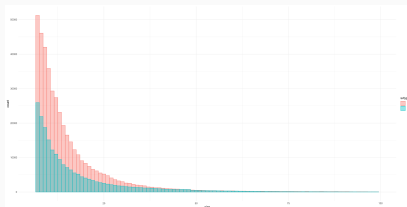


- A large deletion bias and the symmetry is lost

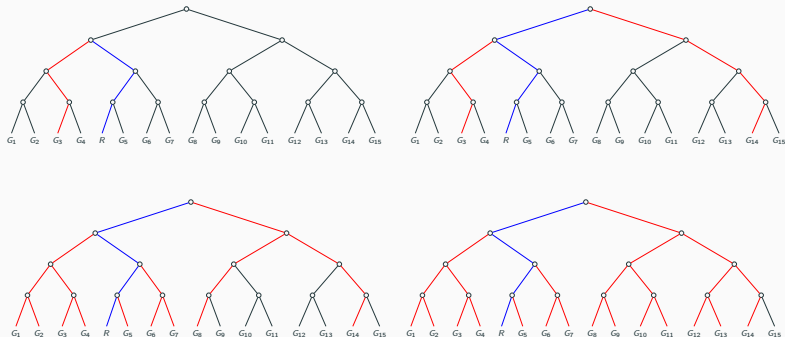


## Vargoats project

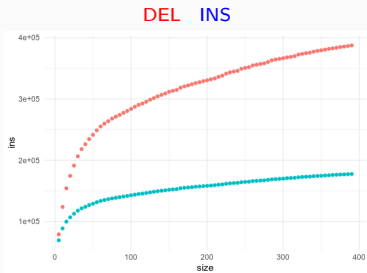
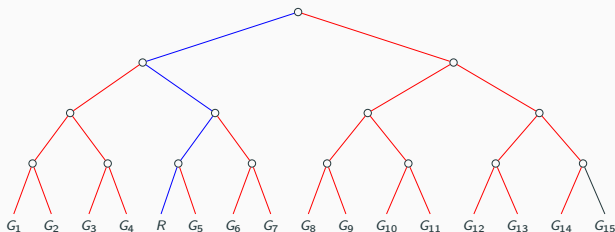
1000 goats

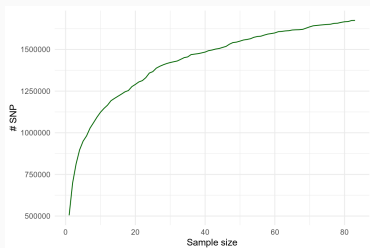
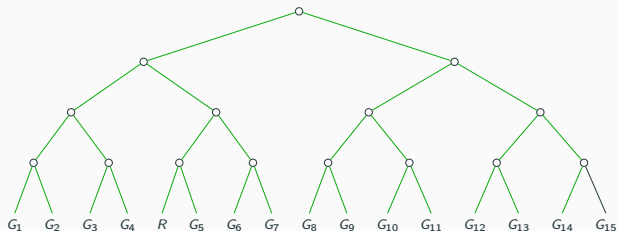


Decrease of the ins/del with an increasing population size

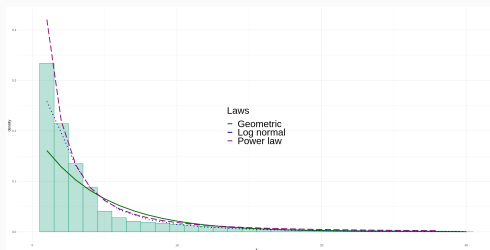








$$S_n = \theta \sum_{i=0}^n \frac{1}{i}$$



## Affine gap costs

Bwa mem (Smith-Waterman 1981)

Minimap2 (Gotoh 1990)

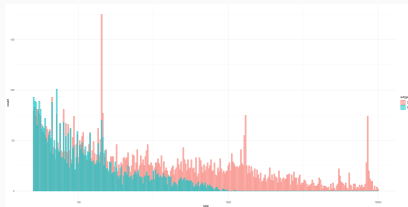
$$G_c = \alpha + \beta(k - 1)$$

$$G_c = \min\{q + |l|, \tilde{q} + |\tilde{l}|\}$$

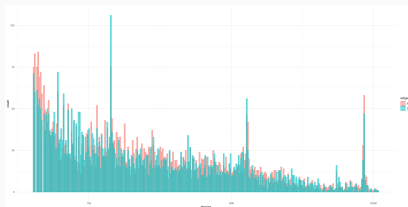
- Population data provides information on the dominant mutational mechanism for indel
- For small insertion/deletions (indels), deletions are about two times more frequent than insertions (also documented in the literature)
- Ancestral state reconstruction provides information of the specific mutation for each variant

## OTEDOR

Illumina



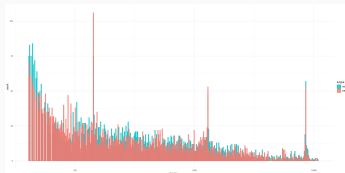
CLR



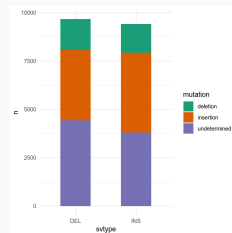
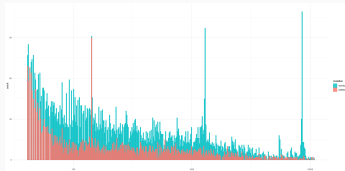
- Illumina fails to detect a large proportion of insertions

## OTEDOR

Variation

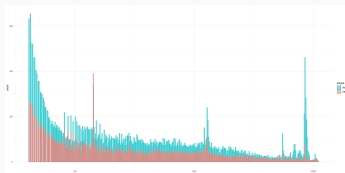


Mutation

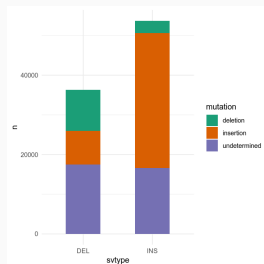
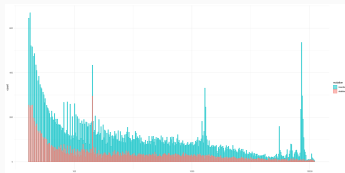


## 100 SeqOccln CLR bulls

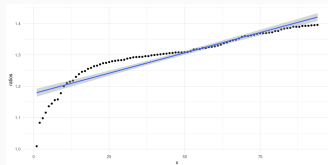
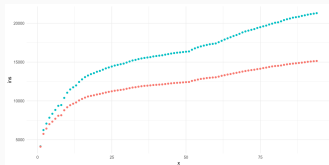
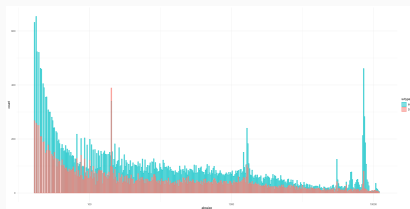
Variation



Mutation



## 100 SeqOocln CLR bulls



Increase of the ins/del ratio with the number of samples

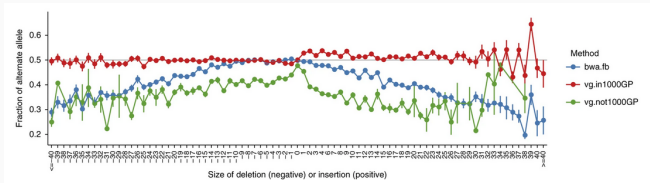


- In contrast to indels, for medium to large variations, insertion seems to be the predominant mutation mechanism
- A junction is needed between indel and SV catalogues to confirm this potential switch

- All technologies exhibit different kind of bias
- This is especially true for short reads in the context of structural variant detection
- A major source of bias is the use of a single reference genome

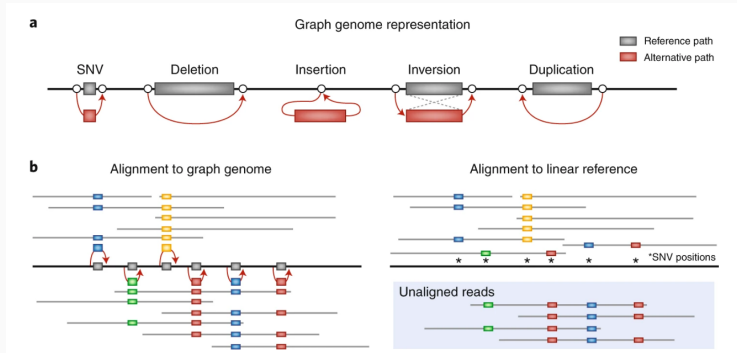
## Giab son (HG002)

Reference alleles map better



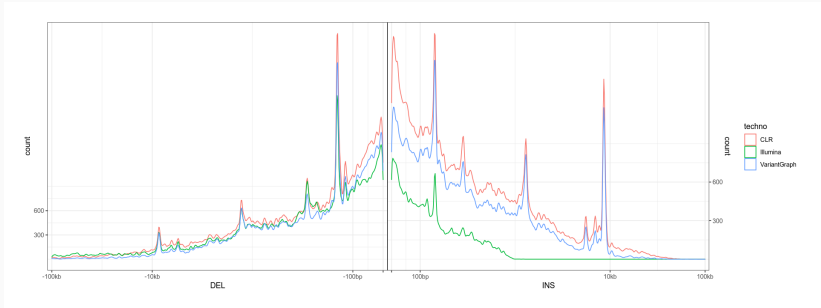
- Fraction of alternate allele when aligned to the variation graph (red) with vg or to the reference (blue for bwa, or green with vg)

Garrison *et al.* Nat Biot **36**:875 (2018). <https://doi.org/10.1038/nbt.4227>



Goodbye reference, hello genome graphs Nat Biotechnol 37, 866–868 (2019)

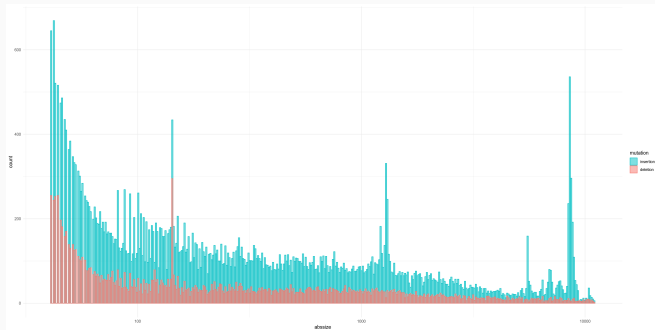
<https://doi.org/10.1038/s41587-019-0199-7>



- Variants detected in 100 bulls (CLR) are used to construct a variation graph
- Illumina data from these same 100 bulls sample are used to genotype them on the variation graph

- For small insertion/deletions (indels), deletions are about two times more frequent than insertions
- In contrast to indels, for medium to large variations, insertion seems to be the predominant mutation mechanism
- Goodbye reference, hello genome graphs

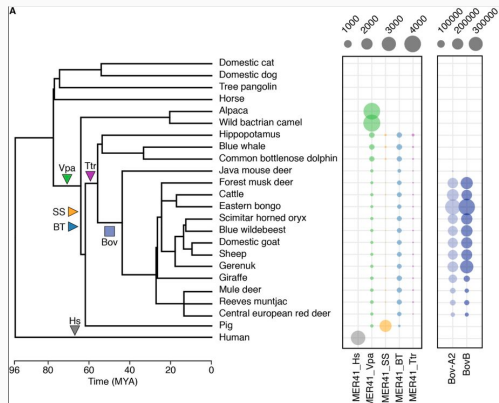
# Chronicle of a Disparition Foretold

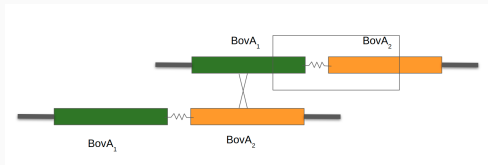


- There is a sharp peak at 143bp made of insertions and deletions
- This is approximatively half the size of a BovA2 SINE



# The 143bp peak





Product 1



Apparent deletion of



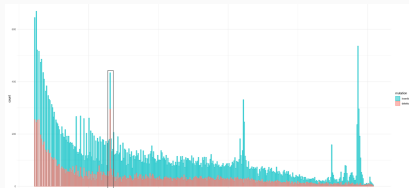
Product 2



Apparent insertion of



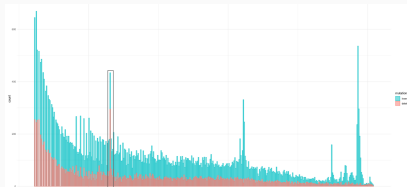
- The high similarity between the two copies can lead to unequal crossing-over



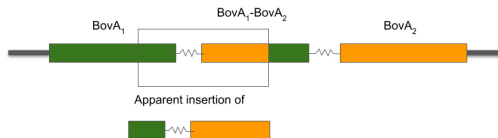
Product 1

Apparent deletion  
of

deletion	Reference	Alternate
INS	 89%	 85%
DEL	 79%	 72%



Product 2



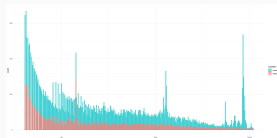
insertion	Reference	Alternate
INS	 50%	 54%
DEL	 63%	 73%

- BovA2 copies in the bovine (ruminants) genome experience an intensive erosion mechanism due to recombination

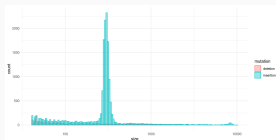
Goat



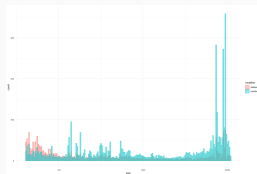
cow



Pig



Maize



# Remerciements

## SeqOccIn

### Coordination

Cécile Donnadieu  
Christine Gaspin  
Carole Lampietro  
Denis Milan

### Plus spécifiquement impliqués dans les travaux présentés aujourd'hui

Quentin Boone  
Mekki Boussaha  
Mathieu Charles  
Arnaud Di Franco  
Thomas Faraut  
Christophe Klopp

