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

Variations de structure et longues lectures

Colloque INRAE Genomics

Thomas Faraut

Orléans, 17 mai 2022

Outline

- 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

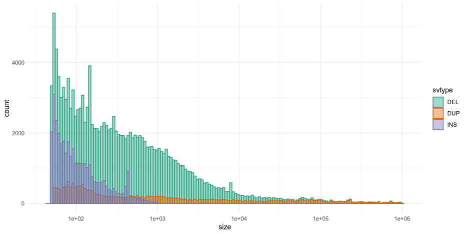
From short reads to long reads and back

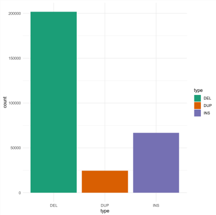
- The variation graph

Projet cofinancé par le Fonds Européen de Développement Régional

SV: the case of short reads

1000 goats (Vargoats project)

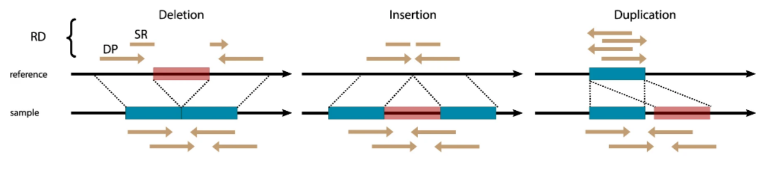




- Deletions clearly outnumber insertions

SV: the case of short reads

Structural variation signatures



RD: Read depth

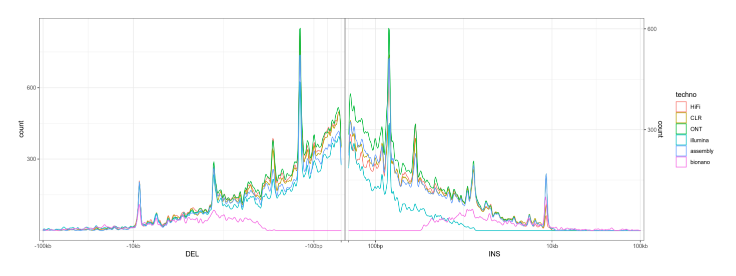
DP : Discordant pairs

SR : Split reads

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

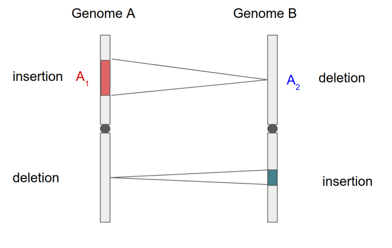
The symmetry of insertions and deletions

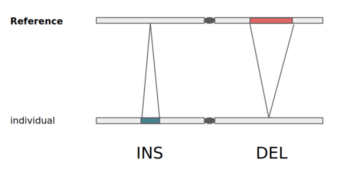
Trio2 heifer



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

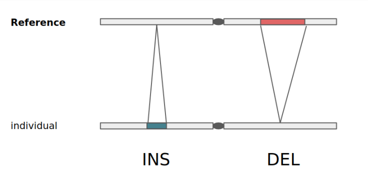
The symmetry of insertions and deletions

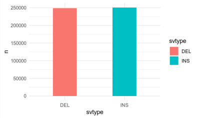




- Structural variation type INS and DEL are somehow ill-defined
- For small insertions/deletions the term indel preserves this ambiguity

The symmetry of insertions and deletions

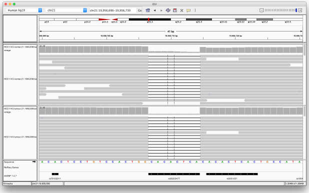


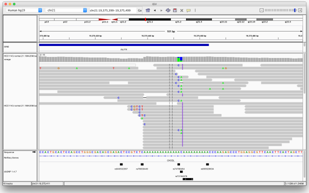


- When comparing two genomes we expect to see a symmetric pattern, the same number of INS and of DEL

Indels detection

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

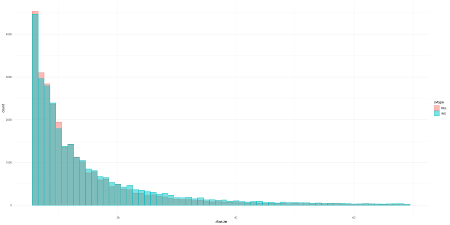


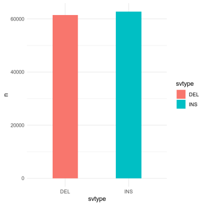


- Indels are detected directly from alignments

Indel size distribution: PacBio

Trio2 heifer

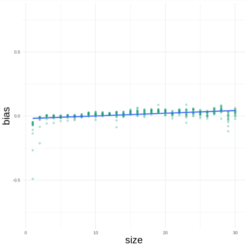




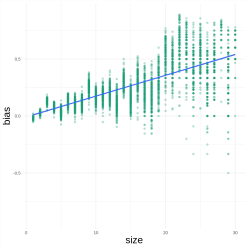
- With PacBio HiFi, the distribution of indel sizes exhibit a clear symmetric pattern

Indel INS/DEL bias

PacBio HiFi



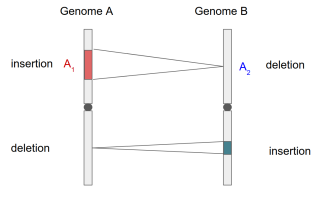
Illumina

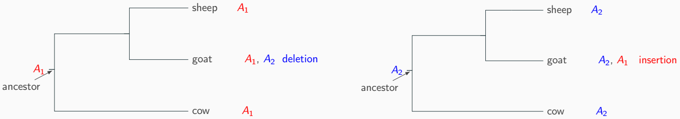


$$bias = \frac{n_d - n_i}{n_d + n_i}$$

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

Ancestral allele



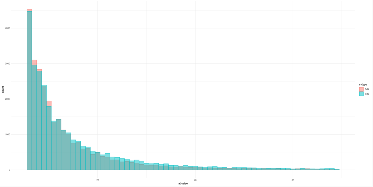


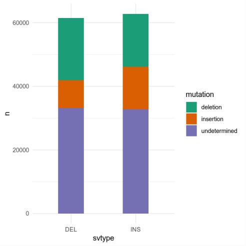
- The identification of the ancestral state is mandatory to identify the mutation, deletion of insertion for a given variant

Indels ancestral allele

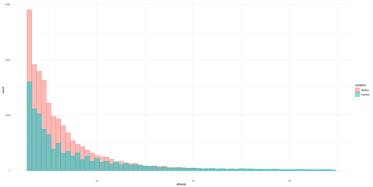
Heifer

Variation





Mutation



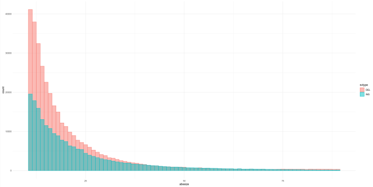
- Short deletions clearly outnumber short insertions

SeqOccIn

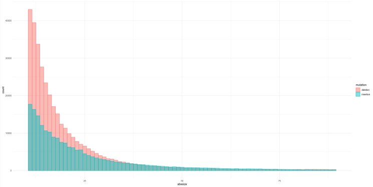
Indels at the population level

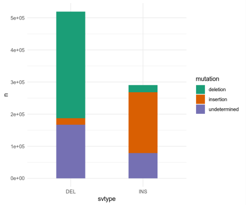
1000 Bull genomes project

Variation



Mutation



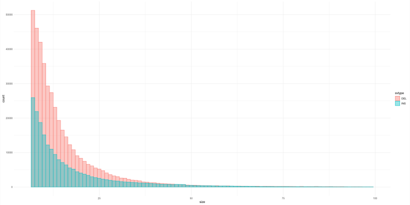


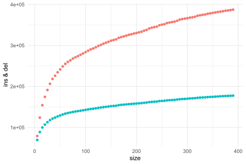
- A large deletion bias and the symmetry is lost

Indels at the population level

Vargoats project

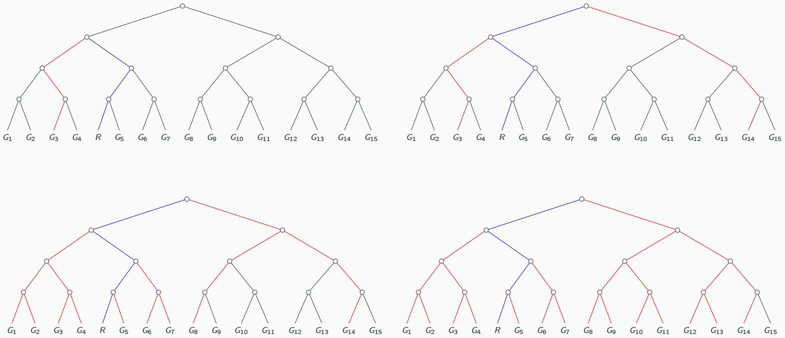
1000 goats



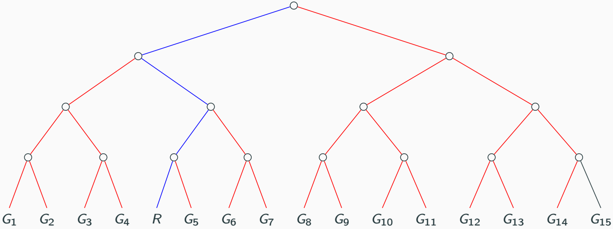


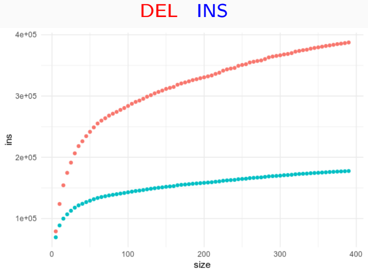
Decrease of the ins/del with an increasing population size

Sample size and number of variants

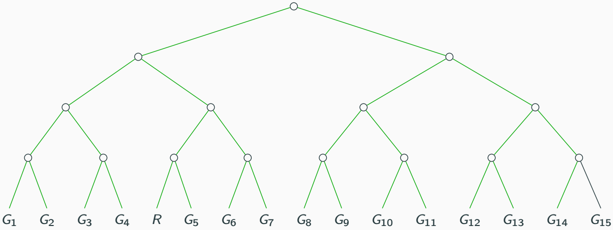

G_1 G_2 G_3 G_4 R G_5 G_6 G_7 G_8 G_9 G_{10} G_{11} G_{12} G_{13} G_{14} G_{15}
G_1 G_2 G_3 G_4 R G_5 G_6 G_7 G_8 G_9 G_{10} G_{11} G_{12} G_{13} G_{14} G_{15}
G_1 G_2 G_3 G_4 R G_5 G_6 G_7 G_8 G_9 G_{10} G_{11} G_{12} G_{13} G_{14} G_{15}
G_1 G_2 G_3 G_4 R G_5 G_6 G_7 G_8 G_9 G_{10} G_{11} G_{12} G_{13} G_{14} G_{15}

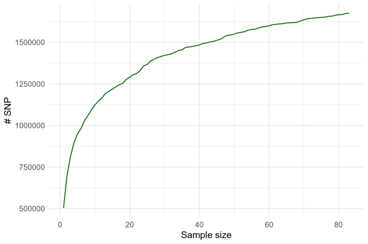
Sample size and number of variants





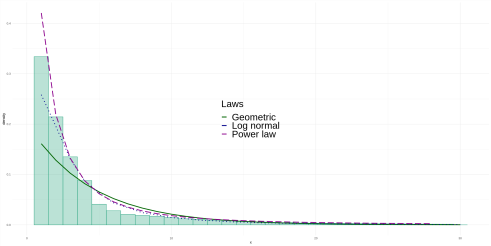
Number of Segregating sites: S_n





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

Deletion size distribution



Affine gap costs

Bwa mem (Smith-Waterman 1981) $G_c = \alpha + \beta(k-1)$

Minimap2 (Gotoh 1990) $G_c = \min\{q + |l|, \tilde{q} + |\tilde{l}|\}$

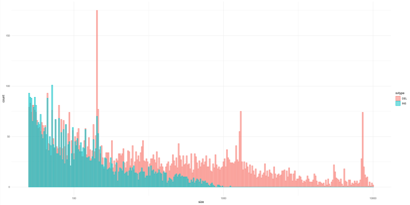
Indels summary

- 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 litterature)
- Ancestral state reconstruction provides information of the specific mutation for each variant

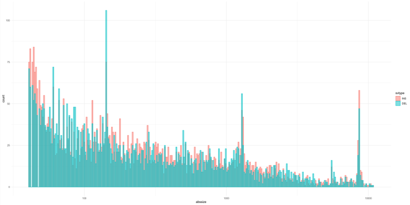
SV detection short/long reads

OTEDOR

Illumina



CLR



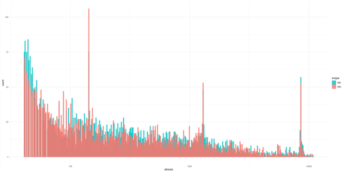
- Illumina fails to detect a large proportion of insertions

SeqOccIn

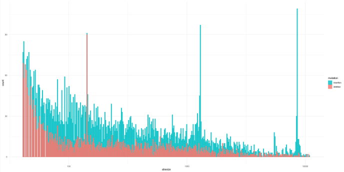
Ancestral State

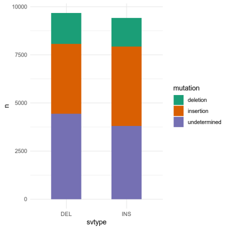
OTEDOR

Variation



Mutation

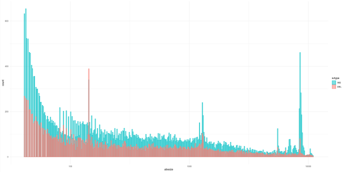




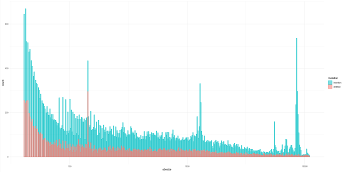
Ancestral State

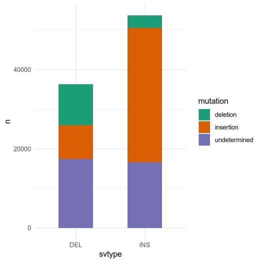
100 SeqOccIn CLR bulls

Variation



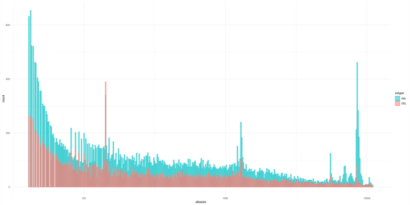
Mutation

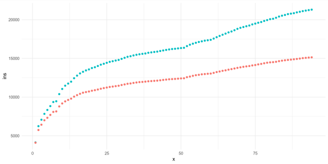


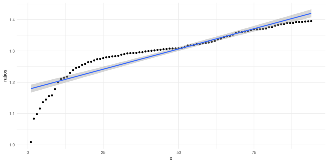


SeqOccIn Population approach

100 SeqOccIn CLR bulls







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

SV summary

- 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

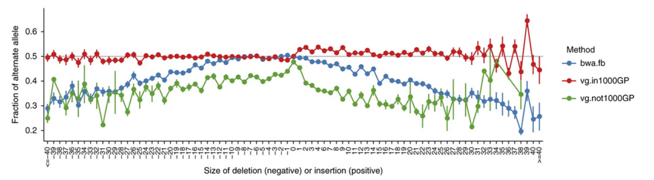
Pangenomes and variation graph

- 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

Reference bias

Giab son (HG002)

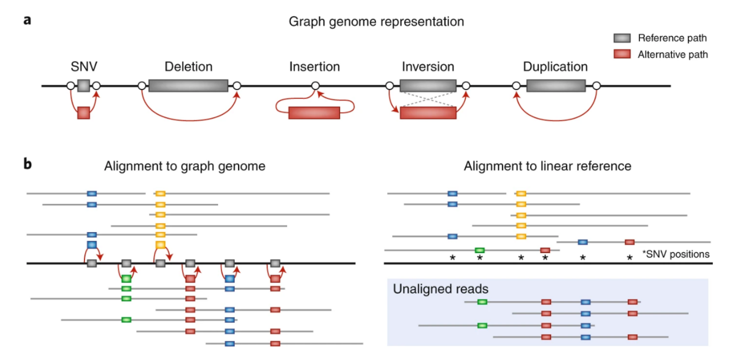
Reference alleles map better



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

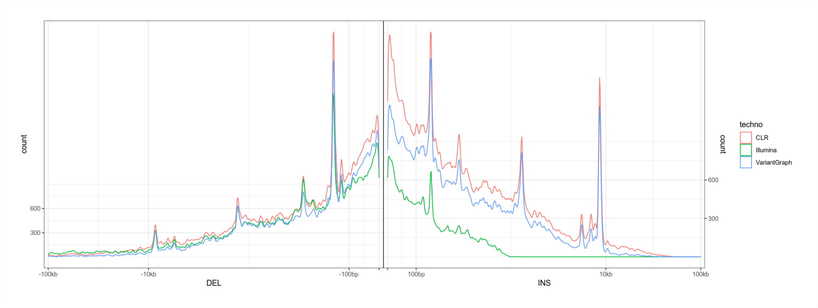
Garrison *et al*. Nat Biot **36**:875 (2018). https://doi.org/10.1038/nbt.4227

Variation graph



Goodbye reference, hello genome graphs Nat Biotechnol 37, 866–868 (2019)
https://doi.org/10.1038/s41587-019-0199-7

Variation graph



- 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

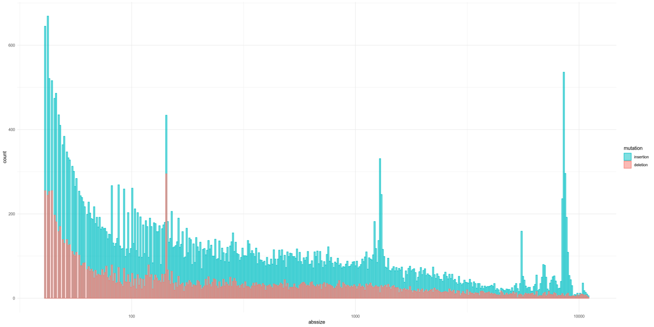
Summary

- 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

Famille BovA2

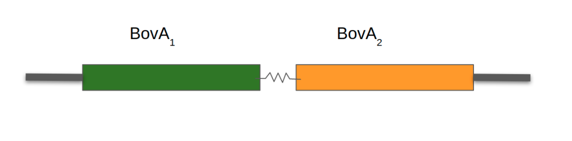
Chronicle of a Disparition Foretold

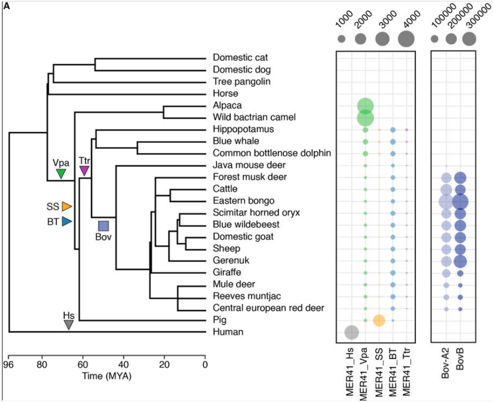
BovA2 family



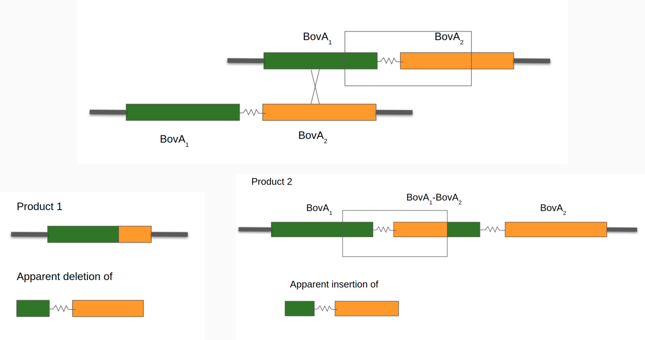
- There is a sharp peak at 143bp made of insertions and deletions
- This is approximatively half the size of a BovA2 SINE

The 143bp peak



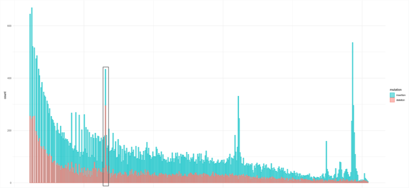


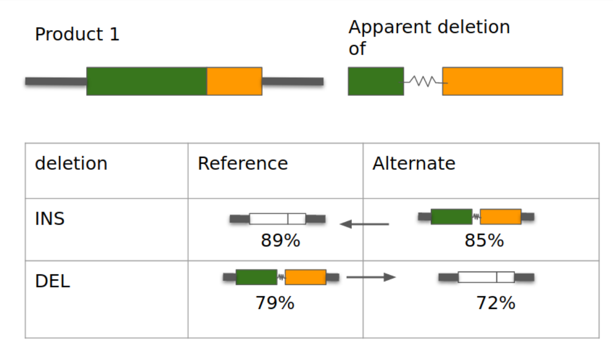
BovA2 recombination



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

BovA2 recombination

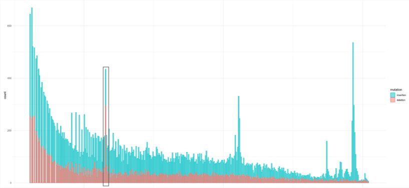


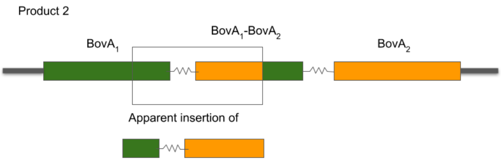

Product 1

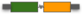
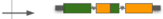
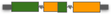
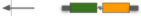
Apparent deletion of

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

BovA2 recombination





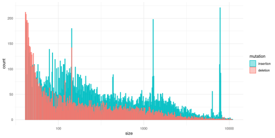
insertion	Reference	Alternate
INS	 50%	 54%
DEL	 63%	 73%

BovA2 summary

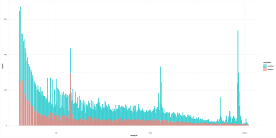
- BovA2 copies in the bovine (ruminants) genome experience an intensive erosion mechanism due to recombination

Genomes are breathing

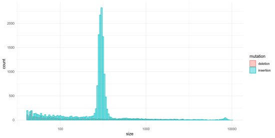
Goat



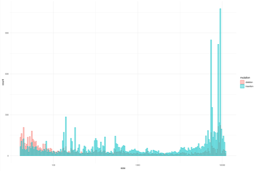
cow



Pig



Maize



la science pour la vie, l'humain, la terre
Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
L'EUROPE S'ENGAGE
L'OCCITANIE AGIT
Projet cofinancé par le Fonds Européen de Développement Régional

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