Biodiversity analysis in plant genomes :
An innovative capture approach to characterize targeted region of interest

Caroline CALLOT[1], Carine SATGE[1], Margaux-Alison FUSTIER[1], Stéphane CAUET[1], William MARANDE[1], Roberto BACILIERI[2], Matthieu CHABANNES[3,4], Audrey GUICHEMERRE[3,4], Arnaud BELLEC[1] and Sonia VAUTRIN[1*]

[1] *French Plant Genomic Center (CNRGV) – INRA, 24 Chemin de Borde Rouge, 31326 Castanet-Tolosan, France*
[2] *INRAE-CIRAD-SupAgro-Université Montpellier - UMR 1334 AGAP - Bât. 21 - 2, Place P. Viala, 34060 Montpellier, France*
[3] CIRAD, UMR AGAP Institut, F-34398 Montpellier, France
[4] UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France
** To whom correspondence should be addressed. Email: sonia.vautrin@inrae.fr*

Introduction

Access to plant genome assemblies allows a better understanding of the diversity in plant species. Although evolution of Next-Generation Sequencing (NGS) has made it possible to access to genome information, the **complexity of some plants** (very high genome sizes, repetitive element contents, polyploidy level) remains challenging.
Exploring **intra-species variability of a region of interest** by using targeted enrichment methods is one of the strategic approaches for biodiversity analysis. This method offers precise and reliable information to link a genomic region to a trait of interest carried by a specific genotype.

Here, we have adapted an innovative approach[1,2] based on a modified CRISPR/Cas9 system **to capture and sequence large genomic regions of interest from plant genomes.** This method coupled with SMRT sequencing generate long reads with high consensus accuracy. They allow comparing **polymorphisms and structural variations** for large genomic regions of interest between **several genotypes**.

Captured Method Workflow to Target a Genomic Region of Interest

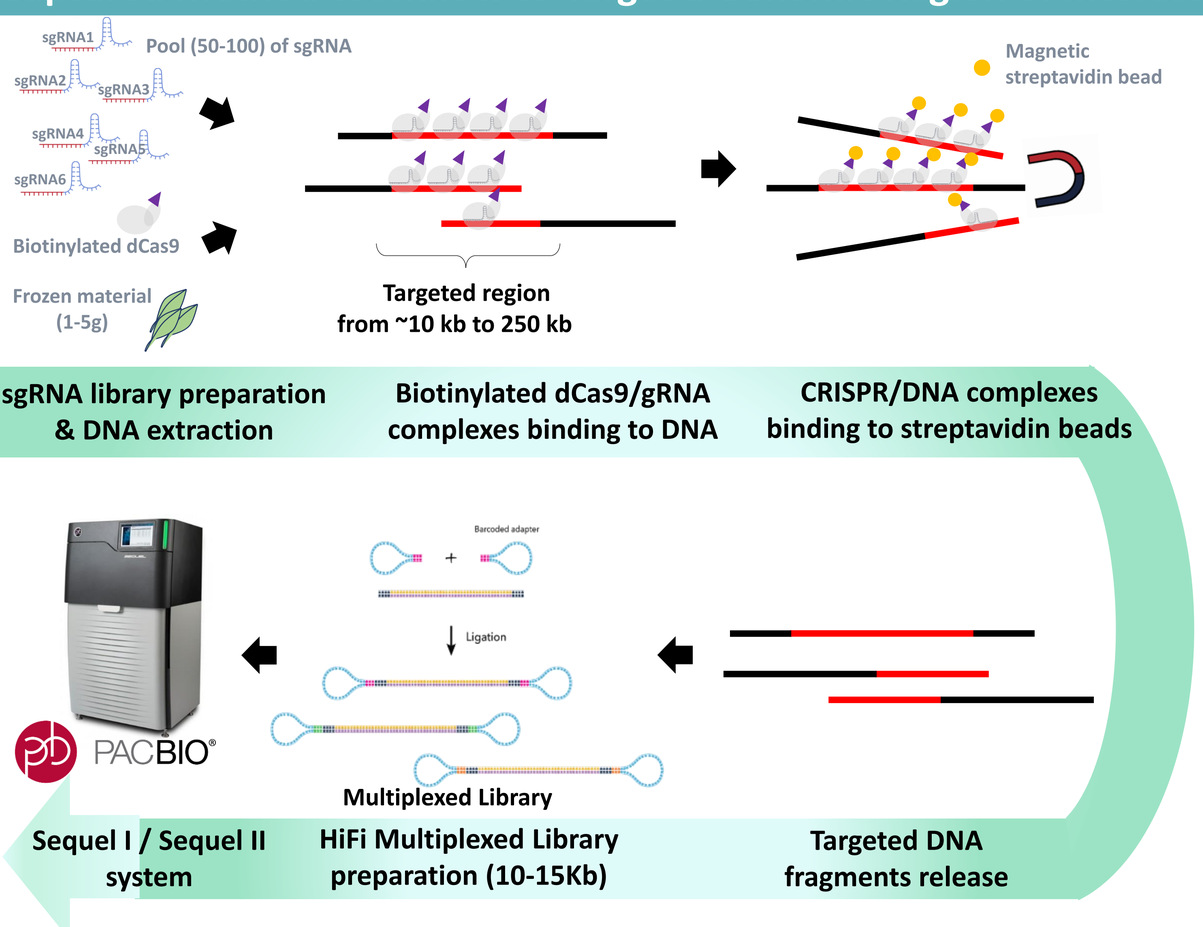

Figure 1. The Captured Method Workflow. HMW-DNA is extracted from frozen leaves. A pool of sgRNA is prepared and incubated with biotinylated dCas9. CRISPR complexes bind to the target region. CRISPR/DNA complexes are captured with magnetic streptavidin beads and target DNA is released from the complexes. A library is performed from 100 ng to 250 ng of captured DNA and sample is sequenced on PacBio Sequel II system. Depending on genome size and region size, **up to 12 target regions can be multiplexed** per SMRT Cell.

Case study 1 : Characterizing complex viruses insertion in banana genome

The *Musa balbisiana* banana genome harbours several integrated sequences of Banana Streak Virus (eBSV). Some of them are still functional and can trigger systemic infection of the plant.
As a proof of concept, we captured the 3 eBSV species (Imove, GoldFinger and Obino L'Ewai) which were published by Chabannes et al., 2013 in the diploid *Musa bal*bisiana Pisang Klutuk Wulung (PKW) genotype[4].

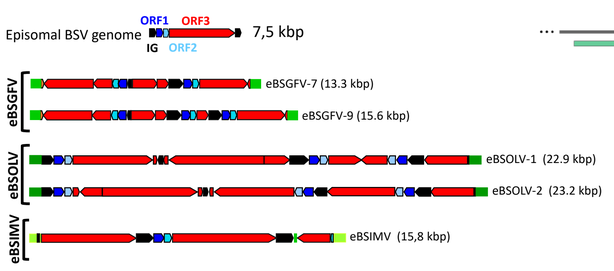

Figure 3. Diversity of eBSV structure in PKW plant [4]. GoldFinger (GFV) and Obino l'Ewai (OLV) species present 2 allelic forms whereas Imove (IMV) is monoallelic in PKW banana plant.

Sequencing system: Sequel I	Metrics
Raw data (Gb)	14,1
Read number	1 554 392
Mean read length (kb)	**9,1**
N50 (kb)	**14,2**
Reads on-target	**1902 (0,12%)**

Table 1. Raw data metrics for a Multiplex Low-Input Library on 1 SMRTcell 1M (Sequel I).

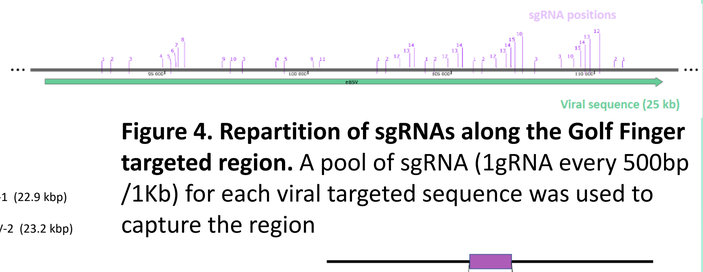

Figure 4. Repartition of sgRNAs along the Golf Finger targeted region. A pool of sgRNA (1gRNA every 500bp /1Kb) for each viral targeted sequence was used to capture the region

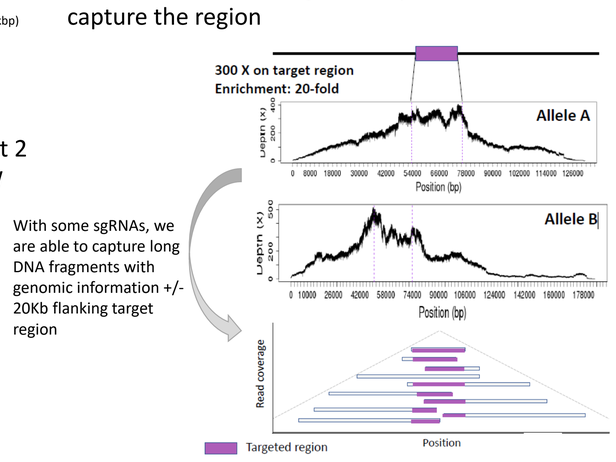

Figure 5.Coverage analysis on the enriched target region. By *de novo* assembly, we obtained 1 contig for each allelic form of virus sequence (minimum size of 50Kb)

✓ **Capture on other genotypes is in progress**

Perspectives : Capture method with Xdrop technology (Samplix®)

Objective: Sequencing large targeted regions of interest with few prior knowlegde on large panels

Prerequisite :
✓ Ultra specific PCR markers every 80-100 Kb
✓ Design on conserved sequences to allow biodiversity analysis

Pilot project in progress on Sunflower and Olive tree

Bio-informatic Analysis Pipeline

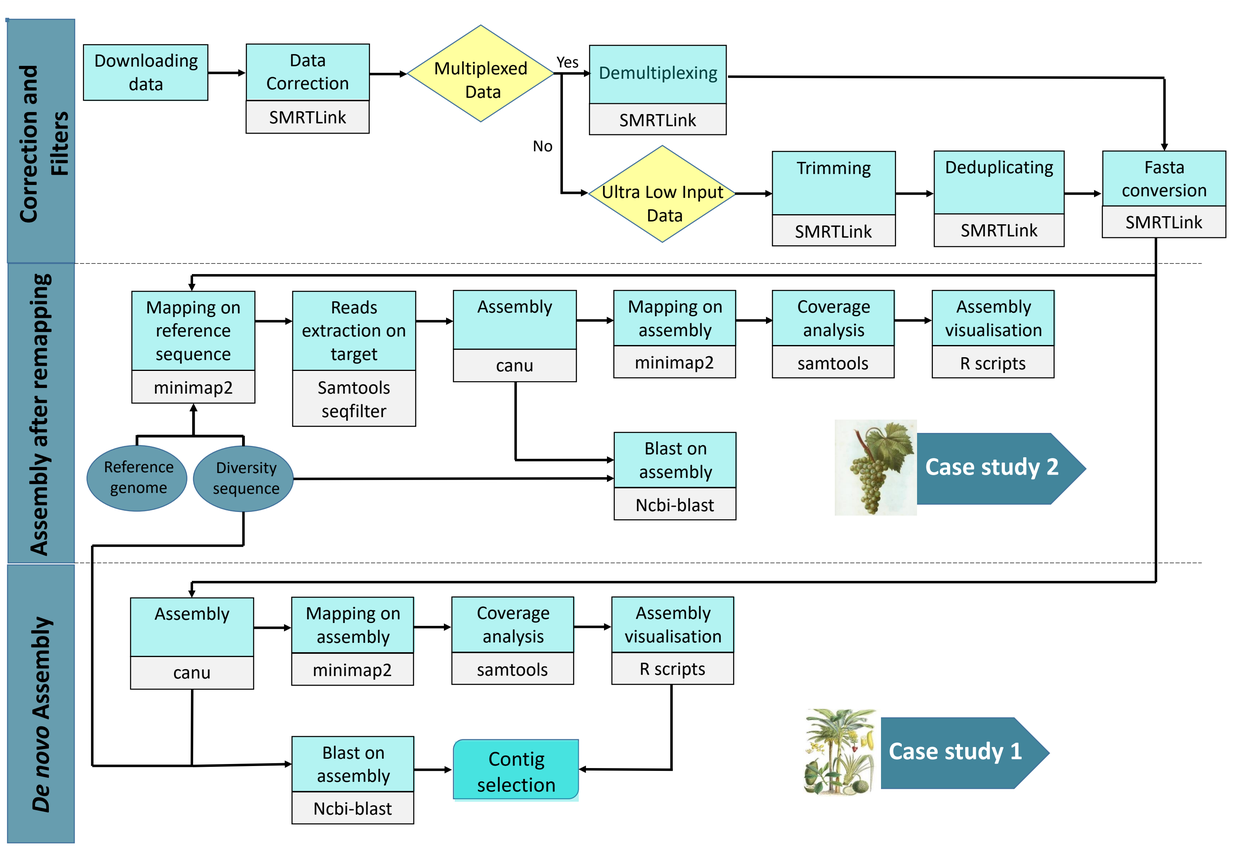

Figure 2 : Capture analysis simplified workflow. The workflow is divided in 3 parts. In the correction and filters part the workflow depends on the design experiment.
Two assemblies are done in parallel : assembly after remapping and *de novo* assembly. Selection on the best assembly is done after assemblies comparison.

Case study 2 : Determining diversity of target region from 11 grapevine genotypes

Grapevine is an important crops for fruit and wine production. To better understand phenotypic differences, there is a growing interest for identifying genomic variations and their functional effects at the intra-species level. Here, we captured the separate haplotypes a of the XY-like sex locus in a biodiversity collection of wild and cultivated grapevine plants[5].
We used 15 µg of HMW DNA and a pool of 60 sgRNA to capture target region and perform a multiplexed Low-Input PacBio library[3].
From only one SMRTCell, we enriched each target regions 126 fold on average.

Sequencing system Sequel II	Metrics
Raw data (Gb)	355
Read number	4 821 087
Mean read length (kb)	**14**
N50 (kb)	**19**

Table 2. Raw data metrics for a Multiplex Low-Input Library on 1 SMRTcell 8M (Sequel II).

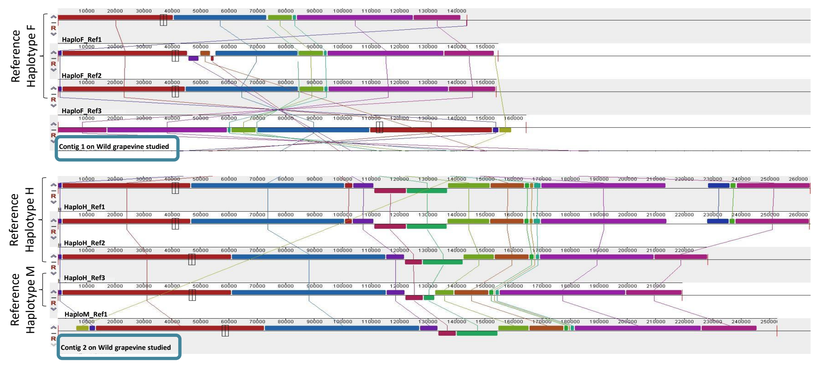

Figure 4. Identification of the sex locus haplotypes on a wild grapevine species. By using assembly after remapping, we have obtained a competed assembly for each haplotype male (M) and female (F). The structure was compared to haplotypes structure known. Gene annotation is in progress

✓**Haplotype differenciation**
✓**possibility to capture the entire region even if there are large insertion (40Kb)**

Conclusion

❖ **Capture method presented offer an efficient solution to target large genomic region of interest in complex plant genomes**

✓ Need **Low-input (100ng) or Ultra-Low Input (20ng) amount DNA**
✓ Provide an **accurate and reliable genomic information** for the region of interest
✓ Allow a **rapid comparison** of a region of interest between several genotypes

References

[1] *In situ* Capture of Chromatin Interactions by Biotinylated dCas9. Liu et al., *Cell*, 2017, http://dx.doi.org/10.1016/j.cell.2017.08.003
[2] CRISPR-Cap: multiplexed double-stranded DNA enrichment based on the CRISPR system, Lee et al., *Nucleic Acids Research*, 2018, *doi: 10.1093/nar/gky820*
[3] PacBio Documentation, https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-HiFi-Libraries-from-Low-DNA-Input-Using-SMRTbell-Express-Template-Prep-Kit-2.0.pdf
[4] Three infectious viral species lying in wait in the banana genome, Chabannes et al., *Journal of Virology*, 2013, doi.org/10.1128/JVI.00899-13
[5] The wild grape genome sequence provides insights into the transition from dioecy to hermaphroditism during grape domestication, Badouin, H., Velt, A., Gindraud, F. *et al, Genome Biology,* 2020, doi: 10.1186/s13059-020-02131-y