



**HAL**  
open science

## Nanopore adaptive sampling to identify the NLR-gene family in melon (*Cucumis melo* L.)

Javier Belinchon-Moreno, Aurelie Berard, Aurelie Canaguier, Veronique Chovelon, Corinne Cruaud, Stéfan Engelen, Rafael Feriche-Linares, Isabelle Le-Clainche, William Marande, Vincent Rittener-Ruff, et al.

### ► To cite this version:

Javier Belinchon-Moreno, Aurelie Berard, Aurelie Canaguier, Veronique Chovelon, Corinne Cruaud, et al.. Nanopore adaptive sampling to identify the NLR-gene family in melon (*Cucumis melo* L.). 2024. hal-04670606

**HAL Id: hal-04670606**

**<https://hal.inrae.fr/hal-04670606v1>**

Preprint submitted on 28 Aug 2024

**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.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

# 1 Nanopore adaptive sampling to identify the NLR-gene family in 2 melon (*Cucumis melo* L.)

3 Javier Belinchon-Moreno<sup>1,2</sup>, Aurelie Berard<sup>1</sup>, Aurelie Canaguier<sup>1</sup>, Véronique Chovelon<sup>2</sup>,  
4 Corinne Cruaud<sup>3</sup>, Stéfan Engelen<sup>4</sup>, Rafael Feriche-Linares<sup>2</sup>, Isabelle Le-Clainche<sup>1</sup>, William  
5 Marande<sup>5</sup>, Vincent Rittener-Ruff<sup>2</sup>, Jacques Lagnel<sup>2</sup>, Damien Hinsinger<sup>1</sup>, Nathalie Boissot<sup>2</sup>,  
6 Patricia Faivre Rampant<sup>1</sup>

7

8 <sup>1</sup>Université Paris-Saclay, Centre INRAE Île-de-France Versailles-Saclay, EPGV, F-91057  
9 Evry, France

10 <sup>2</sup>INRAE, Génétique et Amélioration des Fruits et Légumes, 84143 Montfavet, France

11 <sup>3</sup>Genoscope, Institut de Biologie François-Jacob, Commissariat à l’Energie Atomique (CEA),  
12 Université Paris-Saclay, 2 Rue Gaston Crémieux, 91057 Evry, France

13 <sup>4</sup>Génomique Métabolique, Genoscope, Institut François Jacob, Commissariat à l’Energie  
14 Atomique (CEA), CNRS, Univ. Evry, Université Paris-Saclay, 91057 Evry, France.

15 <sup>5</sup>INRAE, Centre National de Ressources Génomiques Végétales, 31326 Castanet-Tolosan,  
16 France

17

## 18 Corresponding authors

19 Patricia Faivre Rampant, [patricia.favre-rampant@inrae.fr](mailto:patricia.favre-rampant@inrae.fr)

20

## 21 Keywords

22 Melon; Nanopore adaptive sampling; targeted sequencing; resistance genes; NLR; genome  
23 assembly

24

25

26

27

28

29

30

31

32 **Abstract**

33 **Background**

34 Nanopore Adaptive Sampling (NAS) offers a promising approach for assessing genetic  
35 diversity in targeted genomic regions. Herein, we design and validate an experiment to enrich  
36 a set of resistance genes in several melon cultivars as a proof of concept.

37 **Results**

38 We showed that each of the 15 regions we identified in two newly assembled melon genomes  
39 (subspecies *melo*) were successfully and accurately reconstructed as well as in a third cultivar  
40 from the *agrestis* subspecies. We obtained a fourfold enrichment, independently from the  
41 samples, but with some variations according to the enriched regions. In the *agrestis* cultivar,  
42 we further confirmed our assembly by PCR. We discussed parameters that can influence  
43 enrichment and accuracy of assemblies generated through NAS.

44 **Conclusions**

45 Altogether, we demonstrated NAS as a simple and efficient approach to explore complex  
46 genomic regions. This approach finally unlocks the characterization of resistance genes for a  
47 large number of individuals, as required for breeding new cultivars responding to the  
48 agroecological transition.

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

## 64 **Background**

65 The correct assembly of complex and highly repeated genome regions, especially notorious in  
66 plants, remains a challenge. Despite being widely used on whole genome sequencing (WGS)  
67 approaches due to their cost-effectiveness, short-read sequencing methods prove ineffective in  
68 these complex regions as their length cannot allow a proper assessment of copy-number  
69 variation and duplication events (Lee & Chae, 2020). Long-read technologies, such as those  
70 provided by Oxford Nanopore Technologies (ONT, Oxford, UK), demonstrated their potential  
71 in accurately resolving complex regions by spanning long repetitive elements or areas with  
72 tandemly repeated genes (Mohamed et al., 2020; Lieberman et al., 2022). Nevertheless, whole-  
73 genome long-read sequencing remains cost-ineffective for most studies, particularly those  
74 requiring the sequencing of numerous genotypes for a few specific regions of interest.

75 Targeted sequencing approaches are a valuable alternative for characterizing specific genomic  
76 regions while reducing sequencing and data storage costs compared to WGS (Hook & Timp,  
77 2023). Current targeted sequencing protocols have been adapted for long-read sequencing and  
78 predominantly employ hybridization capture (Witek et al., 2016), PCR amplification ((Norris  
79 et al., 2016), Cas9-assisted targeting methods (Gilpatrick et al., 2020), or microfluidic-based  
80 droplet sorting procedures (Madsen et al., 2020). However, these approaches require substantial  
81 experimental and design efforts, along with a high prior knowledge of the sequence to be  
82 enriched and its genotypic diversity (Gilpatrick et al., 2020; Hook & Timp, 2023). Particularly,  
83 PCR-based techniques are prone to introducing biases in the enriched sequences, and long  
84 amplicons are difficult to consistently amplify (Hook & Timp, 2023). Hybridization-based  
85 methods require the construction of complex RNA libraries together with very specific  
86 hybridization and capture conditions (Hook & Timp, 2023). Cas9-based methods like Nanopore  
87 Cas9-targeted sequencing (nCATS) (Gilpatrick et al., 2020) or Cas9-Assisted Targeting of  
88 Chromosome segments (CATCH) (Gabrieli et al., 2017) require the design of multiple guide-  
89 RNAs, a task that may be challenging for complex and repetitive genome regions. Finally,  
90 microfluidic-based methods like Xdrop (Madsen et al., 2020) are highly complex and require  
91 specialized microfluidic equipment (Hook & Timp, 2023).

92 The Nanopore Adaptive Sampling (NAS) approach recently implemented by ONT overcome  
93 these limitations. NAS was first suggested in 2016 (Loose et al., 2016) and has been  
94 implemented with different algorithms since the end of 2019 (Edwards et al., 2019; Kovaka et  
95 al., 2021; Payne, Holmes, et al., 2021; Weilguny et al., 2023). It takes advantage of the ability  
96 of the pores to control the directional flow of the DNA strand that is being sequenced depending  
97 on the applied current's polarity. By combining live calling of sequenced bases with real-time  
98 mapping to a set of DNA sequences provided by the user for enrichment, the DNA strand is  
99 dynamically either discarded or fully sequenced based on the similarity of its initial first few  
100 hundred bases to the provided reference (Loose et al., 2016). NAS requires a standard library  
101 preparation, eliminates the need for DNA amplification, circumvents laborious or expensive  
102 experimental design or probes synthesis, and offers real-time selective enrichment (Martin et  
103 al., 2022; Miyatake et al., 2022). NAS has been used in clinical settings and for the enrichment  
104 of metagenomic samples (Kipp et al., 2021; De Meulenaere et al., 2022; Martin et al., 2022;  
105 Greer et al., 2023; Hewel et al., 2023; Su et al., 2023; Wrenn & Drown, 2023). Therefore, NAS  
106 emerges as a promising approach for studying target regions, especially those that are highly  
107 complex, such as disease-associated repeat loci in humans (Miyatake et al., 2022; Stevanovski  
108 et al., 2022).

109 In plants, immunity is encoded by resistance genes (R genes), frequently organized in complex  
110 regions (Liu et al., 2007). Among R genes, Nucleotide-binding site leucine-rich repeat

111 resistance genes (NLRs) form the largest family (Barragan & Weigel, 2021). These genes  
112 encode intracellular receptors that play a central role in the so-called effector-triggered  
113 immunity (ETI) against pathogens. NLR genes exhibit a highly conserved structure with three  
114 main domains (Barragan & Weigel, 2021; Zhang et al., 2022): the N-terminal domain, the  
115 central domain, and the C-terminal domain. The N-terminal domain can be a Toll/Interleukin-  
116 1 receptor (TIR), a Coiled-coil (CC), or a resistance to Powdery Mildew 8-like (RPW8) domain.  
117 The central domain, the most conserved one, is a nucleotide-binding adaptor (NB-ARC), also  
118 named as NBS (nucleotide-binding site) domain. This domain plays a crucial role in signal  
119 transduction. Finally, the C-terminal domain is often composed of leucine-rich repeats (LRR)  
120 with ligand-binding functions. A clustered genomic arrangement is a common characteristic of  
121 NLR genes (Van Wersch & Li, 2019). These clusters often result from unequal crossing overs,  
122 tandem duplications, or intra-cluster rearrangements (Barragan & Weigel, 2021). In this  
123 context, NAS, combining long-read sequencing and target enrichment, should allow the  
124 accurate characterization of NLR clusters in plants.

125 To investigate the ability of NAS to efficiently retrieve the sequence of the complete set of NLR  
126 clusters into a species (or NLRome), we selected melon (*Cucumis melo* L.) as a model. Melon  
127 genome presents a i/ small genome size; ii/ NLR content estimated at  $\approx 1\%$  of the genome  
128 (González et al., 2013) aligning with ONT target size recommendations (Nanopore  
129 Community, 2023); and iii/ finely characterized, highly variable, and complex NLR cluster, *Vat*  
130 (Chovelon et al., 2021; Boissot et al., 2023), suitable for benchmarking. Among the well-  
131 characterized accessions for the *Vat* region, we chose Anso77 (ssp. *melo*) for its highest number  
132 of functional *Vat* genes (Chovelon et al., 2021). We also selected Doublon (ssp. *melo*) as an  
133 accession with a contrasting *Vat* region structure compared to Anso77 (Chovelon et al., 2021).  
134 Therefore, we assembled and annotated their whole genomes and we established Anso77 as the  
135 reference to identify the regions of interest (ROIs) for NAS. We assessed the performance of  
136 the method in capturing the set of NLR clusters on Anso77 and Doublon. Furthermore, we  
137 extended our assessment to an accession belonging to a different subspecies (Chang-Bougi, ssp.  
138 *agrestis*) for which a genome was publicly available (Shin et al., 2019).

139

## 140 **Materials and Methods**

### 141 **BIOLOGICAL MATERIAL**

142 We selected melon cultivars Anso77, Doublon and Chang-Bougi to develop a proof of concept  
143 for the NLRome adaptive sampling experiment, with Anso77 serving as the reference. The  
144 origin of these cultivars is located in Spain, France and Korea, respectively. Anso77 and  
145 Doublon were chosen as sp. *melo* lines belonging to the *inodorus* and *cantalupensis* botanical  
146 groups. Chang-Bougi, belonging to the *agrestis* subspecies and specifically to the *makuwa*  
147 botanical group, was selected as a distantly related cultivar compared to Anso77 and Doublon.  
148 This choice aimed to validate the NAS procedure with cultivars significantly differing from the  
149 provided reference. Chang-Bougi belongs to the *agrestis* subspecies and more specifically to  
150 the *makuwa* botanical group. Additionally, a draft genome assembly of Chang-Bougi,  
151 constructed with Illumina HiSeq reads, was readily available (Shin et al., 2019).

152 We obtained the seeds from the INRAE Centre for Vegetable Germplasm in Avignon (Salinier  
153 et al., 2022) and cultivated them under greenhouse conditions at INRAE GAFL, Avignon,  
154 France.

## 155 **ANSO77 AND DOUBLON *DE NOVO* WHOLE GENOME SEQUENCING,** 156 **ASSEMBLY, AND ANNOTATION**

157 We produced whole *de novo* genome assemblies of Anso77 and Doublon by combining long-  
158 read sequencing: ONT for Anso77 and ONT combined with Pacific Biosciences (PacBio,  
159 Menlo Park, CA, USA) for Doublon. Raw reads were already deposited in the NCBI database  
160 under the following Bioproject accession numbers: PRJNA662717 and PRJNA662721  
161 (Chovelon et al., 2021). Bionano optical maps (BioNano Genomics, San Diego, CA, USA); 10x  
162 Linked-Reads (Pleasanton, CA, USA) for Anso77; Illumina Novaseq short-read sequencing  
163 (Illumina, San Diego, CA, USA); and linkage map information were developed and used to  
164 construct genome assemblies. Fully detailed methods and parameters employed for the  
165 assemblies and annotations are provided in Additional files: Supplementary Methods.

166

## 167 **NAS ENRICHMENT PANEL DEFINITION AND EXPERIMENTAL DESIGN**

168 We used the cultivar Anso77 as the reference for constructing the target regions for the NAS  
169 approach. We predicted the presence of NLR-related genes using NLGenomeSweeper (Toda et  
170 al., 2020) with default parameters. This tool approximates the presence of NLR genes by the  
171 identification of the well-conserved NBS domain. We defined the regions of interest (ROIs) by  
172 grouping predicted NBS domains separated by regions shorter than 1 Mb. To ensure robust read  
173 depth coverage on the selected ROIs, we added a 20 kb buffer zone flanking the ROIs to  
174 constitute the initial target regions.

175 We performed a REs annotation within the initial target regions using the CENSOR tool from  
176 the curated giri Repbase website (Kohany et al., 2006). Predicted REs longer than 200 bp, as  
177 well as sequences shorter than 500 bp located between them, were excluded from the initial  
178 target regions. Figure 1 illustrates the definition of the ROIs, target regions, and target regions  
179 without REs.

180 We provided these target regions without REs in bed format and the reference genome of  
181 Anso77 in fasta format to the MinKNOW software (ONT, Oxford, UK). These files were used  
182 to determine the acceptance or rejection of reads. If the initial ~500 bps of the DNA strands  
183 matched the target regions without REs, they underwent complete sequencing; otherwise, they  
184 were rejected from the pore.

185

## 186 **DNA EXTRACTION, ADAPTIVE SAMPLING SEQUENCING, AND BASE-CALLING**

187 Plant leaves were harvested and immediately frozen in liquid nitrogen for subsequent DNA  
188 extraction. Genomic DNA was extracted using the NucleoSpin Plant II kit (Macherey-Nagel,  
189 Germany) following the manufacturer's protocol. DNA quantity and quality assessment were  
190 conducted using Qubit4® 1x dsDNA BR Assay Kit (Invitrogen, Carlsbad, CA, USA) and  
191 Agilent 2200 TapeStation (Agilent Technologies, Santa Clara, CA, USA).

192 We multiplexed and sequenced DNAs of Anso77 and Doublon on a single PromethION *R10.4.1*  
193 flowcell (ONT, Oxford, UK). Half the channels were used as control channels in which no  
194 adaptive sampling was performed. Moreover, we sequenced Chang-Bougi using one-tenth of a  
195 PromethION *R10.4.1* flowcell to assess the flexibility and scalability of NAS. We prepared the  
196 sequencing libraries using the Native Barcoding SQK-NBD114.24 (ONT, Oxford, UK) and

197 following ONT guidelines with some modifications. One microgram of genomic DNA from  
198 each sample was repaired and end-prepped with an incubation at 20°C for 20 minutes followed  
199 by a heat-inactivation of the enzymes at 65 °C for an additional 20 minutes. DNA was purified  
200 and barcodes were individually ligated to each of the purified DNA samples. Barcodes NB01  
201 and NB02 were used for Anso77 and Doublon. NB22 was used for Chang-Bougi. Barcoded  
202 samples were purified using AMPure XP beads (Beckman Coulter Inc., Brea, CA, USA) at a  
203 ratio of 0.4:1 beads-to-barcoding mix, keeping each barcoded sample in an independent  
204 Eppendorf tube. Finally, purified barcoded DNA samples were pooled at an equimolar  
205 concentration to a total volume of 30 µl. Adapters were ligated to the pooled samples. After  
206 purification, DNA size and concentration of the barcoded pools were quantified using Agilent  
207 2200 TapeStation and Qubit4® 1x dsDNA HR Assay Kit (Invitrogen, Carlsbad, CA, USA),  
208 respectively. Final libraries were adjusted to a volume of 32 µl containing 10-20 fmol of DNA.  
209 All incubations shorter than 10 minutes were extended to 10 minutes.

210 We completed libraries by adding the Sequencing Buffer (ONT, Oxford, UK) and Library  
211 Loading Beads (ONT, Oxford, UK), and subsequently loaded them into R10.4.1 PromethION  
212 flowcells for 96-hour runs in the case of Anso77 and Doublon, or 120-hours runs in the case of  
213 Chang-Bougi. A library reloading (washing flush) was performed in all the experiences when  
214 the percentage of sequencing pores dropped to 10-15%. NAS was performed using channels 1-  
215 1500 of the PromethION flowcells for Anso77 and Doublon, keeping the rest of the channels  
216 as control. The sequencing speed was set to 260 bps (accuracy mode) for Anso77 and Doublon,  
217 and the quality score threshold was set to 10. For Chang-Bougi, NAS was performed on the  
218 whole flowcell and the sequenced speed was modified to 400 bps (default mode) because the  
219 260 bps option has been deprecated from MinKNOW version 23.04.

220 Raw ONT FAST5 files were live base-called during the PromethION run with Guppy (ONT,  
221 London, UK) v. 6.3.9 for Anso77 and Doublon and Guppy v. 6.5.7 for Chang-Bougi in “super  
222 accurate base-calling” mode. Barcodes were automatically trimmed using the “trim barcodes”  
223 option of the MinKNOW software v. 22.10.7 for Anso77 and Doublon, and v. 23.04.5 for  
224 Chang-Bougi. For each run, the automatically generated “sequencing\_summary.txt” file and  
225 the FASTQ files of the samples were retained for further processing.

226

## 227 **NAS DATA PROCESSING AND ENRICHMENT CALCULATION**

228 For Anso77 and Doublon, we split the reads by channel generating two FASTQ files per  
229 sample: one with the reads sequenced on channels 1-1500 for NAS, and another with the reads  
230 generated on channels 1501-3000 for WGS. No splitting by channel was performed for Chang-  
231 Bougi as the totality of the channels were used for NAS. We retained for downstream analyses  
232 reads with a “PASS” flag, meaning that their quality score was greater than 10. We identified  
233 the rejected reads by NAS based on their “end reason” in the sequencing summary file. This  
234 file includes the classification of the generated reads based on their “end reason”. In this way,  
235 rejected reads by adaptive sampling are labeled as “Data Service Unblock Mux Change”. Reads  
236 labeled as “Unblock Mux Change”, “Mux Change” and “Signal Negative” were also filtered  
237 out. Afterwards, we filtered the generated FASTQ files by size, keeping only reads longer than  
238 1 kb. We assessed statistics on these FASTQ files using seqkit stats v. 2.4.0 (Shen et al., 2016).

239 We computed sequence depth statistics for Anso77 and Doublon by aligning the reads to their  
240 reference whole genome assemblies with minimap2 v. 2.24-r1122 (Li, 2018) and using  
241 mosdepth v. 0.3.3 (Pedersen & Quinlan, 2018) with a bed file containing the coordinates of the

242 15 target regions. For Chang-Bougi, sequence depth statistics were calculating by aligning the  
243 reads to its assembled target regions. The split flowcell setup allowed the calculation of  
244 enrichment in Anso77 and Doublon by comparing read depth generated in NAS and WGS. We  
245 assessed the efficiency of NAS using two measures of enrichment. The enrichment by yield,  
246 denoted as the ratio of the on-target sequence depth (NLR cluster+20 kb flanking) with NAS to  
247 that with WGS, was assessed as follows:

$$248 \quad \text{Enrichment by yield} = \frac{\text{depth}_{\text{region\_NAS}}}{\text{depth}_{\text{region\_WGS}}} \quad (1)$$

249 where  $\text{depth}_{\text{region\_NAS}}$  and  $\text{depth}_{\text{region\_WGS}}$  represent the on-target sequence depth in the NAS and  
250 WGS experiments, respectively.

251 The enrichment by selection, as the ratio of the relative selection of the target regions between  
252 NAS and WGS, measures how much NAS can alter the abundance of the given target regions  
253 in the context of a complete genome, considering the sequencing behavior of each ROI. It was  
254 calculated as follows:

$$255 \quad \text{Enrichment by selection} = \frac{\text{depth}_{\text{region\_NAS}} / \text{depth}_{\text{chr\_NAS}}}{\text{depth}_{\text{region\_std}} / \text{depth}_{\text{chr\_std}}} \quad (2)$$

256 where  $\text{depth}_{\text{region\_NAS}}$  and  $\text{depth}_{\text{chr\_NAS}}$  represent, respectively, the sequence depth on-target  
257 (NLR cluster+20 kb flanking) and on the rest of the chromosome in the adaptive sampling  
258 approach, while  $\text{depth}_{\text{region\_WGS}}$  and  $\text{depth}_{\text{chr\_WGS}}$  represent the depth coverage on-target (NLR  
259 cluster+20 kb flanking) and on the rest of the chromosome in the WGS approach. If no bias  
260 exists that causes the target regions to be differentially enriched compared to the rest of the  
261 genome, the relative selection with WGS should be equal to one.

262 We calculated average enrichment by yield between all target regions as the ratio of the average  
263 sequence depth on-target in NAS and WGS. Similarly, we assessed the average enrichment by  
264 selection between all target regions as the ratio of the average relative frequency of target  
265 regions in NAS and WGS. The average relative frequencies of target regions were calculated  
266 as the ratio of average sequence depth on-target and off-target. Only chromosomes containing  
267 target regions were considered in calculating the off-target average sequence depth.

268

## 269 **TARGET REGIONS ASSEMBLY, NLR ANNOTATION AND QUALITY CONTROLS**

270 We tested a set of assemblers tailored for ONT data including Canu (Koren et al., 2017), Flye  
271 (Kolmogorov et al., 2019), Shasta (Shafin et al., 2020), Necat (Chen et al., 2021), Raven (Vaser  
272 & Šikić, 2021) and SMARTdenovo (Liu et al., 2021), for assembling the NAS reads (data not  
273 shown). SMARTdenovo was primarily selected due to its superior assembly metrics (contiguity  
274 and assembly errors) of the target regions within a short time and with low memory usage. For  
275 Chang-Bougi, one target region was selected from the Canu assembly, as SMARTdenovo failed  
276 to collapse a repeat region generating two contigs instead of a single one. We used default  
277 parameters and added the “generate consensus” option for SMARTdenovo. Canu was executed  
278 with the options “genomesize=7m –corrected –trimmed –nanopore”.

279 For each assembly, we filtered the contigs retaining only those including at least one predicted  
280 NBS domain or matching more than 15 kb with at least 45% identity to any of the 15 target

281 regions of Anso77. We assessed NBS domain prediction using NLGenomeSweeper. We used  
282 the nucmer and delta-filter commands from MUMmer's v. 4.0.0rc1 (Marçais et al., 2018) to  
283 select contigs matching the target regions of Anso77. Nucmer was used with the option  $-l$  100,  
284 keeping the rest of the parameters as default. Hits reported by nucmer were filtered with delta-  
285 filter with the options  $-r -q -l$  15000  $-i$  45. We ran QUAST v. 5.0.2 to assess the basic statistics  
286 of the generated filtered assemblies.

287 Assembly errors were analysed focusing on the well-studied *Vat* region. We performed manual  
288 annotations of the *Vat* regions following (Chovelon et al., 2021). To analyze the accuracy of  
289 the *Vat* homologs, we used two PCR markers: Z649 FR which indicates the number of R65aa  
290 motifs in the *Vat* homologs (Chovelon et al., 2021), and Z1431 FR which is specific to a *Vat*  
291 homolog with four R65aa motifs (Boissot et al., 2023).

292 Figure 2 depicts a workflow diagram summarizing the different steps involved in data  
293 processing and target regions assembly. All statistical tests were performed using R v. 4.1.1 (R  
294 Core Team, 2021).

295

## 296 **Results**

### 297 **ANSO77 AND DOUBLON *DE NOVO* GENOME ASSEMBLIES AND ANNOTATION**

298 Metrics of the different steps during the assembly process are detailed in Additional files:  
299 Supplementary Data 1. Key metrics of the final assemblies are summarized in Table 1.

300 We annotated 21,928 genes in the Anso77 genome, and we identified a functional annotation  
301 in 20,946 (95.5%) of them. In the Doublon assembly, we annotated 21,553 genes, with 20,561  
302 (95.4 %) being functionally annotated. Genes had an average length of 4,297 and 4,265 bp, with  
303 6.09 and 6.03 exons in average per gene in Anso77 and Doublon, respectively.

304 We predicted 84 and 76 NBS domains in the genomes of Anso77 and Doublon, respectively.  
305 These numbers fall within the values obtained in other previously published melon genomes  
306 (Additional file 1: Table S1). Based on InterProScan (Hunter et al., 2009) domain identification  
307 in the 10 kb flanking sequence on both sides of the NBS domain, potential genes containing the  
308 predicted NBS domains were classified into different categories (Table 1).

309 The accuracy of the assemblies concerning NLR genes was assessed by inspecting the accuracy  
310 of the *Vat* homologs, whose cDNA sequence was previously obtained by Sanger sequencing  
311 (Chovelon et al., 2021). For Anso77, the homologs *AN-Vat2*, *AN-Vat3* and *AN-Vat5* fully  
312 matched the assembly here generated, while *AN-Vat1* and *AN-Vat4* contained one SNP each.  
313 For Doublon, all the three *Vat* homologs fully matched the assembled sequence.

314

### 315 **NAS TARGET REGIONS CONSTRUCTION**

316 We arranged the 84 NLR predicted domains on Anso77 into 15 groups comprising nine ROIs  
317 with 2-28 NBS domains and six ROIs with isolated NBS domains. We also found 15 groups  
318 and similar physical positions of the NLR genes on Doublon and on the previously published  
319 melon genomes. After adding the 20 kb flanking zones, the 15 target regions varied in sizes

320 from ~41 to ~1,378 kb, representing a total length of ~6.16 Mb of the ~370 Mb Anso77 genome  
321 (~1.68%) (Table 2).

322 After masking REs, the final file provided to the PromethION sequencer comprised 935 target  
323 regions, ranging from 502 pb to 67.609 kb (Additional file 2: Supplementary Data). They  
324 accounted for a total of ~5.23 Mb of the ~370 Mb Anso77 genome (~1.41%).

325

## 326 **NAS TARGET REGIONS VALIDATION: EFFECTIVE ENRICHMENT OF NLR** 327 **CLUSTERS IN MELON**

328 Over 8.62 million reads and 19.86 Gb (42.71%) belonged to Anso77, and over 11.97 million  
329 reads and 26.63 Gb (57.27%) belonged to Doublon after barcodes trimming. We compared  
330 NAS to WGS of Anso77 and Doublon in terms of general metrics. For Anso77, further read  
331 splitting by channel and filtering of this data by quality, “end reason” and length resulted in  
332 110.06 K reads resulting on 1.14 Gb coming from the NAS half-flowcell. The other half-  
333 flowcell (WGS) yielded 1.12 million reads with a cumulative size of 11.93 Gb for Anso77.  
334 Using the same processing for Doublon, 163.84 K reads generating 1.56 Gb from were assigned  
335 to the NAS half-flowcell, while 15.81 Gb from 1.70 million reads belonged to the WGS part.  
336 For both Anso77 and Doublon, the N50 value from the filtered NAS reads was very similar to  
337 the filtered WGS ones. All information about the generated datasets is grouped in Table 3.

338 The length distribution of NAS-generated reads peaked at around 500 bp corresponding to  
339 rejected reads by adaptive sampling (91.20% and 92.20% of the total “pass” reads for Anso77  
340 and Doublon) (Figure 3A). When rejected reads were out-filtered, a similar length distribution  
341 profile was observed for reads obtained in WGS and in NAS for both cultivars (Figure 3B, C).

342 Focusing on Anso77, we evaluated the read depth on the target regions and on the rest of the  
343 chromosome in the NAS approach. The sequence depth on the target regions at the end of the  
344 experience was vastly higher than on the rest of the chromosome (Figure 4), with an average  
345 frequency of the target regions of 63.37. Sequence depth remained stable throughout the entire  
346 ROIs, even more so when the ROIs had a smaller size (Figure 4A-C). The standard deviation  
347 of the sequence depth on the ROIs ranged from 1.38 for region 03 to 20.19 for region 07. It  
348 presented values of 10.29, 16.15 and 1.80 for regions 01, 08 and 12, respectively. Actually, the  
349 increase of sequence depth was gradual on the 20 kb flanking, obtaining the highest depth on  
350 the ROIs. This increase exhibited a similar behavior regardless of the ROI’s size (Figure 4D).

351 The half-flowcell design allowed us to calculate the enrichment obtained in NAS compared to  
352 a WGS approach. We obtained an enrichment by yield for Anso77 variable among target  
353 regions, ranging from 2.45 to 5.18 at the end of the run (Figure 5A). Considering the enrichment  
354 by selection, we observed an increase ranging from 45.56 to 102.91 (Figure 5B). On average,  
355 we obtained an enrichment by yield and an enrichment by selection of 3.96 and 78.38,  
356 respectively (Table 4). Target regions were sequenced at a lower rate than the rest of the genome  
357 in WGS, with a relative frequency of 0.81 (Table 4). Regarding enrichment over time, we  
358 observed that both the enrichment by yield and enrichment by selection followed very similar  
359 patterns. This enrichment reached its maximum at the beginning of the run for most of the  
360 regions when most of the flowcell channels were actively sequencing, and it decreased over  
361 time. This fact reflects that channel inactivation occurred faster on the NAS half-flowcell.

362 Region 10 exhibited a very particular behavior, being extremely enriched at the beginning of  
363 the run (Figure 5A, B). As shown in Additional files: Figure S1A, this high enrichment during  
364 the first hours of the run corresponded to a poor sequencing in the WGS approach. Furthermore,  
365 the NAS approach provided in just 10 hours a sequence depth comparable to that achieved in  
366 the entire WGS run (Additional files: Figure S1A). The washing flush contributed to an increase  
367 in pore activity in both NAS and WGS approaches (Additional files: Figure S1A).

368 To confirm the applicability of NAS in targeting the entire spectrum of NLR clusters in melon,  
369 we extended its use to Doublon, a cultivar within the same subspecies as Anso77 but belonging  
370 to a distinct botanical group. Similar to Anso77, the NAS sequence depth on the target regions  
371 at the end of the run always exceeded (2.25 to 4.65X) that of the WGS approach (Figure 5C,  
372 D). The least (region 06) and most (region 10) enriched regions were the same for both cultivars.  
373 Notably, the enrichment by yield presented a Kendall's coefficient of concordance (W) of 0.87  
374 between Anso77 and Doublon (p-value of 0.04), suggesting region-specific patterns rather than  
375 cultivar-based differences.

376 Target regions of Doublon were sequenced and mapped with an identical ratio (0.98) to the rest  
377 of the genome in WGS (Table 4). Regarding the enrichment by selection in Doublon, we  
378 obtained values ranging from 43.52 to 83.89 fold, maintaining a significant correlation with the  
379 values previously observed for Anso77 (W=0.89; p-value=0.04). Overall, we demonstrated an  
380 average enrichment by yield between all the target regions of 3.73 and an average enrichment  
381 by selection of 69.92 (Table 4). In terms of enrichment over time, the results were consistent  
382 with those obtained for Anso77, encompassing both the enrichment by yield and enrichment by  
383 selection (Figure 5C,D).

384

## 385 **ONT ADAPTIVE SAMPLING ALLOWS THE CORRECT ASSEMBLY OF NLR** 386 **CLUSTERS IN ANSO77 AND DOUBLON**

387 The assembly of NAS-enriched reads provided very contiguous and accurate assemblies of the  
388 target regions in both Anso77 and Doublon. Each cultivars presented a single contig for each  
389 target region. NAS assembly metrics are grouped in Table 5. Notably, the size of all contigs  
390 was over the size of their corresponding target region.

391 We generated dot plots for Anso77 and Doublon comparing each assembled contig with the  
392 corresponding region in the whole genomes assembly we produced (Additional files: Figure  
393 S2). A perfect collinearity appeared in the dot plot of each target region, showing the fidelity  
394 of the NAS assemblies to the reference genomes. Notably, the same number of NBS domains  
395 and at the same positions were predicted in the NAS assembly compared to the reference  
396 genome for both cultivars, emphasizing the reliability of the NAS assemblies.

397 To further assess the accuracy of the NAS assemblies, we focused our attention on the well-  
398 known *Vat* region. Dot plots representing the *Vat* regions are shown in Figure 6. The dot plots  
399 heightened the complexity of this area with numerous duplicated sequences, but a perfect  
400 diagonal appeared between the reference and the NAS-assembled sequences. Moreover, we  
401 checked the sequence of the *Vat* homologs previously sequenced by Sanger sequencing (cDNA  
402 sequencing). Among the five homologs of Anso77, *AN-Vat1*, *AN-Vat2*, *AN-Vat3* and *AN-Vat5*  
403 fully matched the assembly generated with the NAS library, while *AN-Vat4* contained one SNP  
404 in the first exon (G/T on position 402710). This overcomes the reference assembly here  
405 presented, which contained the SNP in *AN-Vat4* but also one SNP in *AN-Vat1*. For Doublon,

406 the three *Vat* homologous presented 100% DNA sequence similarity with the NAS assembly.  
407 Altogether, we demonstrated that NAS produces very contiguous and accurate assemblies in  
408 highly complex clusters of resistance genes.

409

## 410 **TOWARDS AN ESTABLISHED PROCEDURE: ENRICHMENT AND ASSEMBLY OF** 411 **NLR CLUSTERS FROM A DISTANT CULTIVAR PROVIDE VERY VALUABLE** 412 **STRUCTURAL INFORMATION**

413 We obtained 3.32 million reads and 3.45 Gb for Chang-Bougi in a 1/10 flowcell. After  
414 eliminating the reads rejected by adaptive sampling and filtering by quality and 1 kb length,  
415 96.62 K target reads and 0.80 Gb were available for further processing. These reads exhibited  
416 an N50 of 13.75 kb, a measure comparable to that obtained for Anso77 and Doublon. Rejected  
417 reads had an average length of ~790.52 bp, longer than that obtained for Anso77 and Doublon,  
418 due to the updated sequencing speed (400 bp/s).

419 The sequence depth on the target regions, mapped on the assembled contigs, averaged 41.82X.  
420 As illustrated in Figure 7, this depth was variable between regions, with the highest depth in  
421 region 05 (50.37X) and the lowest in region 03 (21.08X). However, the obtained sequenced  
422 depth between regions kept a significant concordance with that obtained for Anso77 and  
423 Doublon (Additional files: Figure S1-left) ( $W=0.78$ ;  $p=0.003$ ).

424 For Chang-Bougi, the assembly of NAS-enriched reads with SMARTdenovo resulted in 17  
425 contigs. Notably, we observed an inversion of  $\approx 100$  kb on region 01 compared to Anso77  
426 (Figure 8A). Region 13 and region 15 were fragmented into two contigs. Among the tested  
427 assemblers, Canu achieved a contiguous assembly of region 13 in a single contig (Additional  
428 files: Figure S3). Consequently, we retained region 13 from the Canu assembly. No assembler  
429 succeeded in reconstructing region 15 into a single contig. We then investigated why region 15  
430 was fragmented into two contigs. Due to the high degree of fragmentation of the Illumina-based  
431 published assembly, we could not conclude after contig alignment. As Chang-Bougi belongs to  
432 the *makuwa* botanical group, we mapped the two contigs to the publicly available genomes  
433 from this group: Early Silver Line, Ohgon and Sakata's Sweet (Oren et al., 2022). We identified  
434 a very large and size-conserved insertion (ranging from 862 to 871 kb) at the breakpoint  
435 between the two contigs obtained for Chang-Bougi (Figure 8A). No NBS domain was predicted  
436 in this insertion on the genomes of Early Silver Line, Ohgon and Sakata's Sweet. We could not  
437 recover this insertion in Chang-Bougi as it was not present in the provided reference. The total  
438 size of the assembly was 6.68 Mb. Table 4 shows the detailed NAS assembly metrics.

439 In the draft genome of Chang-Bougi generated by (Shin et al., 2019), we identified 81 NBS  
440 domains spread across 18 contigs. These contigs matched the previously identified 15 ROIs of  
441 Anso77 with no extra clusters (Figure 8B). In the NAS assembly, we identified 83 NBS  
442 domains. The two extra NBS domains predicted in the NAS assembly were located in region  
443 08, one in the *Vat* region and the other outside. We manually annotated the *Vat* region of both  
444 Chang-Bougi assemblies (NAS and published) revealing some discrepancies in the complex  
445 and repetitive area between *Vat1* and *VatRev* (Figure 9A). In the NAS assembly, we identified  
446 the extra NBS domain within the *Vat* region as a *Vat* homolog with four R65aa motifs (Figure  
447 9A). We confirmed the presence and the structure of this *Vat* gene with four R65aa motifs as  
448 well as the presence of the *Vat* genes with three R65aa through PCR using the published primers  
449 Z649FR and Z1431FR (Figure 9B). Finally, the presence of long reads encompassing the pairs  
450 of genes *Vat1:Vat2* and *Vat2:Vat3* confirmed the NAS-assembled structure.

## 451 Discussion

452 Herein, we highlighted NAS as a promising approach for studying the polymorphisms of  
453 complex genomic ROIs. We chose melon as a model for which we selected highly diverse ROIs  
454 in terms of size and NLR gene content.

455 Several factors can influence the efficiency of NAS. Among them, we set up two key  
456 considerations before launching the NAS experiments. First, we hypothesized a direct  
457 relationship between the ideal size of sequenced fragments and the size of the ROIs (Figure 10).  
458 Given that ROIs sizes varied here from  $\approx 41$  to  $\approx 1378$  kb, we used standard DNA extractions  
459 (10-30 kb) as they were expected to produce a more stable sequencing depth on ROIs than ultra-  
460 long reads (100-300 kb) for the same yield. We believe that this approach reduces off-target  
461 sequencing and avoids channels blockage when rejecting very long reads, as established in ONT  
462 recommendations (Community Nanopore, 2023). Second, REs cover an important portion of  
463 the melon genome (Castanera et al., 2020), and they are especially frequent inside the NLR  
464 gene clusters (Chovelon et al., 2021). To prevent sequencing off-target REs with high sequence  
465 similarity to those within the initial target regions, we assumed that masking repetitive elements  
466 within the provided target regions would diminish the quantity of off-target data, contributing  
467 to increased enrichment. Among the 6.16 Mb, we masked 0.93 Mb of repetitive sequences.  
468 Masking REs in genomes prior to NAS was previously suggested (Zhang et al., 2021).

469 We investigated the interest of NAS compared to WGS. One key factor that largely affects final  
470 yield in ONT sequencing runs is the number of active channels at the beginning of the run and  
471 their lifespan, which may significantly differ from one flowcell to another. Therefore, we  
472 implemented a half-flowcell design to compare NAS and WGS eliminating biases that may  
473 arise when using two different flowcells, as previously settled in many studies (De Meulenaere  
474 et al., 2022; Martin et al., 2022). Results showed that for both Anso77 and Doublon, NAS  
475 produced about four times more on-target data than WGS (Table 3), while generating about ten  
476 times less total data (see filtered NAS and WGS in Table 2). Nevertheless, the sequence depth  
477 between target regions was more variable in NAS than in WGS, although this did not  
478 compromise the accurate assembly of the ROIs in all the cultivars. We found no correlation  
479 between target size and sequence depth (Additional files: Figure S4A) but there was a moderate  
480 correlation between percentage of masking and sequence depth (additional files: Figure S4B).

481 We proposed two measures of enrichment adapted from previous studies that have tested NAS  
482 on metagenomics samples or panels of many small regions of the genome (Hogers et al., 2020;  
483 Martin et al., 2022): the enrichment by yield, a widespread and simple metric; and the  
484 enrichment by selection, a metric not biased by the sequencing behavior of each target region.  
485 These enrichment measures make sense in our study because our goal is to increase coverage  
486 in the complex ROIs to generate more accurate assemblies. We obtained an enrichment by yield  
487 up to 3.7 times on average and an enrichment by selection up to 69 times on average, even when  
488 the reference was genetically distant from the sequenced accession. These findings are  
489 comparable to the best results previously obtained enriching individuals in metagenomics  
490 samples (De Meulenaere et al., 2022; Martin et al., 2022) and outperformed the enrichment  
491 values previously obtained with loci panels (Hogers et al., 2020). The successful enrichment  
492 with NAS could be linked to the percentage of the genome targeted here ( $\sim 1.41\%$ ), the size of  
493 the targets, or the size of the DNA fragments, as they have been demonstrated to be main factors  
494 of enrichment rate (Martin et al., 2022; Community Nanopore, 2023). In addition, a late  
495 nuclease flush performed when the percentage of sequencing pores was around 10% might have  
496 contributed to the good performance, instead of doing it at a fixed time (Payne et al., 2021;  
497 Martin et al., 2022; Nakamura et al., 2023). The enrichments over time of the different target

498 regions were higher and more variable at the beginning of the run but globally stabilized after  
499 70h (Figure 5). Actually, channel inactivation occurred faster on the NAS half-flowcell, either  
500 due to the repetitive potential flipping to reject off-target sequences or simply because the  
501 likelihood of channel clogging is statistically related to the number of sequenced molecules  
502 (Kovaka et al., 2021; Martin et al., 2022).

503 Previous studies have used NAS to enrich specific species in metagenomics samples (Kipp et  
504 al., 2021; Martin et al., 2022; Ulrich et al., 2022) and relatively small sequences within an  
505 organism, such as exon panels or panels of loci of key variants (Hogers et al., 2020; Filser et  
506 al., 2023; Nakamura et al., 2023). Here, we demonstrated the power of NAS as a tool for  
507 enriching ROIs that represent isolated NLR genes or complex clusters of NLRs in a plant crop  
508 species. The correct assembly of these complex regions typically requires long reads as those  
509 provided by ONT or dedicated laborious approaches such as the R gene enrichment sequencing  
510 (RenSeq) method (Witek et al., 2016; Van de Weyer et al., 2019; Huang et al., 2022; Vendelbo  
511 et al., 2022; Adams et al., 2023). In fact, NLR genes may have been miss-predicted, especially  
512 when short-read sequencing technologies were used. When deciphering the *Vat* region of  
513 DHL92 (Garcia-Mas et al., 2012), 2 to 4 functional NLR genes and some pseudo NLRs were  
514 found (Chovelon et al., 2021). After looking for these NLRs in the early released genome of  
515 DHL92, it appeared they were misassembled. Only when a high-quality genome (with long  
516 reads, optical maps or HiC) was released (Castanera et al., 2020), the *Vat* genes were finally  
517 congruent with Sanger sequencing from long-range PCR (Chovelon et al., 2021). Moreover, in  
518 the present study, we compared the *Vat* cluster in the cultivar Chang-Bougi derived from short  
519 WGS (Shin et al., 2019) and long NAS reads. We showed that the WGS assembly was  
520 erroneous in terms of the number and sequences of homologous genes (Figure 9) and that the  
521 NAS assembly accurately reconstructed the region.

522 NLR genes are encompassed within the dispensable portion of the genome (Barragan & Weigel,  
523 2021; Shang et al., 2022), and therefore the NLR reference used for NAS should be carefully  
524 selected when targeting the NLRome of a species. Our prediction of the number of NBS  
525 domains in Anso77 and Doublon, along with all previously published melon genome  
526 assemblies, consistently yielded similar values (Additional files: Table S1). Across all cases,  
527 we did not find more than 15 groups of NLR genes regardless of the subspecies addressed, *melo*  
528 or *agrestis*. These findings indicate a well-conserved number and location of NLR genes in  
529 melon. We chose Anso77, a Spanish cultivar belonging to the subspecies *melo* and the botanical  
530 group *inodorus*, as the reference for the NAS approach because it contained the highest number  
531 of *Vat* homologs within the *Vat* region used for benchmarking (Chovelon et al., 2021). The  
532 results here obtained with Doublon and Chang-Bougi suggested that this strategy was judicious.  
533 Doublon is a French melon line, belonging to the subspecies *melo*, and the botanical group  
534 *cantalupensis*. Using NAS without any short-read polishing we obtained a *Vat* cluster almost  
535 identical to the one derived from an assembly using HW-DNA, PacBio and ONT long  
536 sequences, Illumina short sequences, and optical maps. Chang-Bougi is a Korean melon line,  
537 belonging to the subspecies *agrestis*, and the botanical group *makuwa*. The *Vat* cluster we  
538 obtained using NAS was highly consistent with the *Vat* cluster of PI 161375 (Chovelon et al.,  
539 2021), a Korean line belonging to *agrestis* subspecies. However, a limitation appeared for very  
540 large SVs not present in the reference, as exemplified by the one identified in the chromosome  
541 11 of Chang-Bougi that turned out to be present in the oriental melon clade. Using different  
542 high-quality reference genomes or even combining them into an “artificial” reference genome  
543 could address this limitation. Existing software such as BOSS-RUNS (Weilguny et al., 2023)  
544 already enables dynamically updating the decision strategies during the run, thereby allowing  
545 a better balance in the depth of target regions or multiplexed samples. Nonetheless, it remains  
546 uncertain whether NAS would have been able to discover extra NLR gene clusters if they

547 existed. This should be possible if the additional NBS domains are conserved enough to match  
548 the provided reference.

549 Altogether, our study provides a blueprint for the selective capture of the NLRome in melon  
550 and could be extended to other important crop species. NLR gene numbers are generally low in  
551 the Cucurbitaceae family (Baggs et al., 2017; Barragan & Weigel, 2021), and they represent an  
552 ideal percentage of the genome to be targeted using NAS. However, this ideal situation does  
553 not correspond to reality in other species, as the number of NLR genes is highly variable  
554 between plant species independently of their genome size (Barragan & Weigel, 2021). To adapt  
555 the NAS procedure implemented here for NLR-rich plant species, certain adjustments should  
556 be made to align with the ideal targeted percentage of the genome. First, a reduction in the  
557 length of flanking regions surrounding the ROIs is recommended. Subsequently, employing a  
558 more rigorous definition of NLR clusters would help to reduce the percentage of targeted  
559 genome. Finally, a focus with NAS could be strictly done on the clustered NLRs, recovering  
560 the isolated NLRs with the low-pass rejected reads by NAS.

561

## 562 **Conclusions**

563 NAS offered a flexible, real-time enrichment of selected clusters of NLR genes while reducing  
564 costs compared to a WGS approach. Such target enrichment did not require any laborious or  
565 expensive library preparation nor probes design and synthesis unlike previously developed  
566 target sequencing methods. This is particularly advantageous for researchers who may not have  
567 access to special molecular biology techniques or seek to conduct in-field experiments. NAS  
568 only requires an ONT sequencing device (that can be the low-cost MinION device), a reference  
569 genome, and one or several ROIs. In addition, the fast enrichment observed here may be of  
570 crucial interest when time is a critical factor. Moreover, we evidenced the ability of NAS to  
571 reduce the high off-target volume of data produced by WGS, addressing the growing challenges  
572 of data management and storage in the field of bioinformatics and genomics. This methodology,  
573 validated here on three melon cultivars, holds promise for its application across a large number  
574 of accessions. This is particularly relevant for breeding purposes as it opens avenues for creating  
575 multi-resistant varieties by tapping into the NLRome diversity.

576

## 577 **Acknowledgements**

578 We thank Isabelle Dufau (INRAE-CNRGV) for furnishing high-quality HMW DNA and  
579 optical maps for Anso77 and Doublon. Moreover, we are grateful to Valerie Barbe for advice  
580 concerning the manuscript.

581

## 582 **Ethics approval and consent to participate**

583 Not applicable.

584

585

## 586 **Consent for publication**

587 Not applicable.

588

## 589 **Competing interests**

590 The authors declare that they have no competing interests.

591

## 592 **Funding**

593 This work was partly funded by the French *Ministère de l'agriculture et de la souveraineté*  
594 *alimentaire* (Vat&Co project - CASDAR – 2017-2021) and the French National Research  
595 Institute for Agriculture, Food and Environment, INRAE. The doctoral position of Javier  
596 Belinchon-Moreno is co-funded by the INRAE BAP Department and the EUR Implanteus of  
597 Avignon University, France.

598

## 599 **Authors' contributions**

600 J.B.M. performed the sequencing experiments with NAS, conducted the bioinformatics and  
601 statistical analyses, and carried out the NAS assemblies. P.F.R., N.B. and D.H. conceived the  
602 study. J.L., V.C., A.C. designed and identified the ROIs for NAS. A.B. and I.L. generated the  
603 ONT, Illumina and 10x genomic data for the whole genome assemblies. W.M. generated  
604 Bionano data for the whole genome assemblies, and participated in the hybrid scaffolding. J.L.  
605 and R.F.L. performed the whole genome assemblies. S.E. provided expertise and bioinformatics  
606 support. C.C. provided expertise and experimental support. V.R.R. performed manual  
607 annotation of the *Vat* cluster and did the PCR experiments. J.B.M., P.F.R., N.B. and D.H. wrote  
608 the manuscript. J.B.M. and A.C. did the data submission. All authors read and approved the  
609 final manuscript.

610

## 611 **References**

612

613 Adams, T. M., Smith, M., Wang, Y., Brown, L. H., Bayer, M. M., & Hein, I. (2023). HISS: Snakemake-  
614 based workflows for performing SMRT-RenSeq assembly, AgRenSeq and dRenSeq for the  
615 discovery of novel plant disease resistance genes. *BMC Bioinformatics*, 24(1), 204.

616 <https://doi.org/10.1186/s12859-023-05335-8>

- 617 Baggs, E., Dagdas, G., & Krasileva, K. (2017). NLR diversity, helpers and integrated domains: Making  
618 sense of the NLR IDentity. *Current Opinion in Plant Biology*, 38, 59–67.  
619 <https://doi.org/10.1016/j.pbi.2017.04.012>
- 620 Barragan, A. C., & Weigel, D. (2021). Plant NLR diversity: The known unknowns of pan-NLRomes. *The*  
621 *Plant Cell*, 33(4), 814–831. <https://doi.org/10.1093/plcell/koaa002>
- 622 Boissot, N., Chovelon, V., Rittener-Ruff, V., Giovinazzo, N., Mistral, P., Pitrat, M., Charpentier, M.,  
623 Troadec, C., Bendahmane, A., & Dogimont, C. (2023). A highly diversified NLR cluster in  
624 melon contains homologs that confer powdery mildew and aphid resistance. *Horticulture*  
625 *Research*, uhad256. <https://doi.org/10.1093/hr/uhad256>
- 626 Castanera, R., Ruggieri, V., Pujol, M., Garcia-Mas, J., & Casacuberta, J. M. (2020). An Improved Melon  
627 Reference Genome With Single-Molecule Sequencing Uncovers a Recent Burst of  
628 Transposable Elements With Potential Impact on Genes. *Frontiers in Plant Science*, 10.  
629 <https://www.frontiersin.org/articles/10.3389/fpls.2019.01815>
- 630 Chen, Y., Nie, F., Xie, S.-Q., Zheng, Y.-F., Dai, Q., Bray, T., Wang, Y.-X., Xing, J.-F., Huang, Z.-J., Wang,  
631 D.-P., He, L.-J., Luo, F., Wang, J.-X., Liu, Y.-Z., & Xiao, C.-L. (2021). Efficient assembly of  
632 nanopore reads via highly accurate and intact error correction. *Nature Communications*,  
633 12(1), Art. 1. <https://doi.org/10.1038/s41467-020-20236-7>
- 634 Chovelon, V., Feriche-Linares, R., Barreau, G., Chadoeuf, J., Callot, C., Gautier, V., Le Paslier, M.-C.,  
635 Berad, A., Faivre-Rampant, P., & Lagnel, J. (2021). Building a cluster of NLR genes conferring  
636 resistance to pests and pathogens: The story of the *Vat* gene cluster in cucurbits. *Horticulture*  
637 *Research*, 8.
- 638 *Community Nanopore* (2023). *Adaptive sampling methodology best practices*. Retrieved December  
639 20, 2023, from [https://community.nanoporetech.com/docs/plan/best\\_practice/adaptive-](https://community.nanoporetech.com/docs/plan/best_practice/adaptive-sampling/v/ads_s1016_v1_revi_12nov2020)  
640 [sampling/v/ads\\_s1016\\_v1\\_revi\\_12nov2020](https://community.nanoporetech.com/docs/plan/best_practice/adaptive-sampling/v/ads_s1016_v1_revi_12nov2020)

- 641 De Meulenaere, K., Cuypers, W. L., Rosanas-Urgell, A., Laukens, K., & Cuypers, B. (2022). *Selective*  
642 *whole-genome sequencing of Plasmodium parasites directly from blood samples by Nanopore*  
643 *adaptive sampling* [Preprint]. *Microbiology*. <https://doi.org/10.1101/2022.11.29.518068>
- 644 Edwards, H. S., Krishnakumar, R., Sinha, A., Bird, S. W., Patel, K. D., & Bartsch, M. S. (2019). Real-Time  
645 Selective Sequencing with RUBRIC: Read Until with Basecall and Reference-Informed Criteria.  
646 *Scientific Reports*, *9*(1), Art. 1. <https://doi.org/10.1038/s41598-019-47857-3>
- 647 Filser, M., Schwartz, M., Merchadou, K., Hamza, A., Villy, M.-C., Decees, A., Frouin, E., Girard, E.,  
648 Caputo, S. M., Renault, V., Becette, V., Golmard, L., Servant, N., Stoppa-Lyonnet, D., Delattre,  
649 O., Colas, C., & Masliah-Planchon, J. (2023). Adaptive nanopore sequencing to determine  
650 pathogenicity of BRCA1 exonic duplication. *Journal of Medical Genetics*, *60*(12), 1206–1209.  
651 <https://doi.org/10.1136/jmg-2023-109155>
- 652 Gabrieli, T., Sharim, H., Michaeli, Y., & Ebenstein, Y. (2017). *Cas9-Assisted Targeting of CHromosome*  
653 *segments (CATCH) for targeted nanopore sequencing and optical genome mapping*  
654 [Preprint]. *Genomics*. <https://doi.org/10.1101/110163>
- 655 Garcia-Mas, J., Benjak, A., Sanseverino, W., Bourgeois, M., Mir, G., González, V. M., Hénaff, E.,  
656 Câmara, F., Cozzuto, L., Lowy, E., Alioto, T., Capella-Gutiérrez, S., Blanca, J., Cañizares, J.,  
657 Ziarso, P., Gonzalez-Ibeas, D., Rodríguez-Moreno, L., Droege, M., Du, L., ... Puigdomènech,  
658 P. (2012). The genome of melon (*Cucumis melo* L.). *Proceedings of the National Academy of*  
659 *Sciences of the United States of America*, *109*(29), 11872–11877.  
660 <https://doi.org/10.1073/pnas.1205415109>
- 661 Gilpatrick, T., Lee, I., Graham, J. E., Raimondeau, E., Bowen, R., Heron, A., Downs, B., Sukumar, S.,  
662 Sedlazeck, F. J., & Timp, W. (2020). Targeted nanopore sequencing with Cas9-guided adapter  
663 ligation. *Nature Biotechnology*, *38*(4), Art. 4. <https://doi.org/10.1038/s41587-020-0407-5>
- 664 González, V. M., Aventín, N., Centeno, E., & Puigdomènech, P. (2013). High presence/absence gene  
665 variability in defense-related gene clusters of *Cucumis melo*. *BMC Genomics*, *14*(1), 782.  
666 <https://doi.org/10.1186/1471-2164-14-782>

- 667 Greer, S. U., Botello, J., Hongo, D., Levy, B., Shah, P., Rabinowitz, M., Miller, D. E., Im, K., & Kumar, A.  
668 (2023). Implementation of Nanopore sequencing as a pragmatic workflow for copy number  
669 variant confirmation in the clinic. *Journal of Translational Medicine*, *21*(1), 378.  
670 <https://doi.org/10.1186/s12967-023-04243-y>
- 671 Hewel, C., Schmidt, H., Runkel, S., Kohnen, W., Schweiger-Seemann, S., Michel, A., Bikar, S.-E.,  
672 Plachter, B., Hankeln, T., Linke, M., & Gerber, S. (2023). *Nanopore adaptive sampling of a*  
673 *metagenomic sample derived from a human monkeypox case* [Preprint]. *Microbiology*.  
674 <https://doi.org/10.1101/2023.03.21.533647>
- 675 Hogers, R., Wittenberg, A., & Roelofs, D. (2020). *Adaptive sequencing in crop species*.
- 676 Hook, P. W., & Timp, W. (2023). Beyond assembly: The increasing flexibility of single-molecule  
677 sequencing technology. *Nature Reviews Genetics*, *24*(9), Art. 9.  
678 <https://doi.org/10.1038/s41576-023-00600-1>
- 679 Huang, Z., Qiao, F., Yang, B., Liu, J., Liu, Y., Wulff, B. B. H., Hu, P., Lv, Z., Zhang, R., Chen, P., Xing, L., &  
680 Cao, A. (2022). Genome-wide identification of the NLR gene family in *Haynaldia villosa* by  
681 SMRT-RenSeq. *BMC Genomics*, *23*(1), 118. <https://doi.org/10.1186/s12864-022-08334-w>
- 682 Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U.,  
683 Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E.,  
684 Laugraud, A., Letunic, I., Lonsdale, D., ... Yeats, C. (2009). InterPro: The integrative protein  
685 signature database. *Nucleic Acids Research*, *37*(Suppl\_1), D211-215.  
686 <https://doi.org/10.1093/nar/gkn785>
- 687 Kipp, E., Lindsey, L., Khoo, B., Faulk, C., Oliver, J., & Larsen, P. (2021). *Enabling metagenomic*  
688 *surveillance for bacterial tick-borne pathogens using nanopore sequencing with adaptive*  
689 *sampling*. <https://doi.org/10.1101/2021.08.17.456696>
- 690 Kohany, O., Gentles, A. J., Hankus, L., & Jurka, J. (2006). Annotation, submission and screening of  
691 repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, *7*(1),  
692 474. <https://doi.org/10.1186/1471-2105-7-474>

- 693 Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using  
694 repeat graphs. *Nature Biotechnology*, 37(5), Art. 5. [https://doi.org/10.1038/s41587-019-](https://doi.org/10.1038/s41587-019-0072-8)  
695 0072-8
- 696 Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu:  
697 Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat  
698 separation. *Genome Research*, 27(5), 722–736. <https://doi.org/10.1101/gr.215087.116>
- 699 Kovaka, S., Fan, Y., Ni, B., Timp, W., & Schatz, M. C. (2021). Targeted nanopore sequencing by real-  
700 time mapping of raw electrical signal with UNCALLED. *Nature Biotechnology*, 39(4), Art. 4.  
701 <https://doi.org/10.1038/s41587-020-0731-9>
- 702 Lee, R. R. Q., & Chae, E. (2020). Variation Patterns of NLR Clusters in Arabidopsis thaliana Genomes.  
703 *Plant Communications*, 1(4), 100089. <https://doi.org/10.1016/j.xplc.2020.100089>
- 704 Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–  
705 3100. <https://doi.org/10.1093/bioinformatics/bty191>
- 706 Lieberman, N. A. P., Armstrong, T. D., Chung, B., Pfalmer, D., Hennelly, C. M., Haynes, A., Romeis, E.,  
707 Wang, Q.-Q., Zhang, R.-L., Kou, C.-X., Ciccarese, G., Conte, I. D., Cusini, M., Drago, F.,  
708 Nakayama, S., Lee, K., Ohnishi, M., Konda, K. A., Vargas, S. K., ... Greninger, A. L. (2022). High-  
709 throughput nanopore sequencing of *Treponema pallidum* tandem repeat genes arp and  
710 tp0470 reveals clade-specific patterns and recapitulates global whole genome phylogeny.  
711 *Frontiers in Microbiology*, 13.  
712 <https://www.frontiersin.org/articles/10.3389/fmicb.2022.1007056>
- 713 Liu, H., Wu, S., Li, A., & Ruan, J. (2021). SMARTdenovo: A de novo assembler using long noisy reads.  
714 *GigaByte*, 2021, gigabyte15. <https://doi.org/10.46471/gigabyte.15>
- 715 Liu, J., Liu, X., Dai, L., & Wang, G. (2007). Recent Progress in Elucidating the Structure, Function and  
716 Evolution of Disease Resistance Genes in Plants. *Journal of Genetics and Genomics*, 34(9),  
717 765–776. [https://doi.org/10.1016/S1673-8527\(07\)60087-3](https://doi.org/10.1016/S1673-8527(07)60087-3)

- 718 Loose, M., Malla, S., & Stout, M. (2016). Real-time selective sequencing using nanopore technology.  
719 *Nature Methods*, 13(9), Art. 9. <https://doi.org/10.1038/nmeth.3930>
- 720 Madsen, E. B., Höijer, I., Kvist, T., Ameer, A., & Mikkelsen, M. J. (2020). Xdrop: Targeted sequencing  
721 of long DNA molecules from low input samples using droplet sorting. *Human Mutation*,  
722 41(9), 1671–1679. <https://doi.org/10.1002/humu.24063>
- 723 Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4:  
724 A fast and versatile genome alignment system. *PLOS Computational Biology*, 14(1),  
725 e1005944. <https://doi.org/10.1371/journal.pcbi.1005944>
- 726 Martin, S., Heavens, D., Lan, Y., Horsfield, S., Clark, M. D., & Leggett, R. M. (2022). Nanopore adaptive  
727 sampling: A tool for enrichment of low abundance species in metagenomic samples. *Genome*  
728 *Biology*, 23(1), 11. <https://doi.org/10.1186/s13059-021-02582-x>
- 729 Miyatake, S., Koshimizu, E., Fujita, A., Doi, H., Okubo, M., Wada, T., Hamanaka, K., Ueda, N., Kishida,  
730 H., Minase, G., Matsuno, A., Kodaira, M., Ogata, K., Kato, R., Sugiyama, A., Sasaki, A.,  
731 Miyama, T., Satoh, M., Uchiyama, Y., ... Matsumoto, N. (2022). Rapid and comprehensive  
732 diagnostic method for repeat expansion diseases using nanopore sequencing. *Npj Genomic*  
733 *Medicine*, 7(1), Art. 1. <https://doi.org/10.1038/s41525-022-00331-y>
- 734 Mohamed, M., Dang, N. T.-M., Ogyama, Y., Burlet, N., Mugat, B., Boulesteix, M., Mérel, V., Veber, P.,  
735 Salces-Ortiz, J., Severac, D., Péliesson, A., Vieira, C., Sabot, F., Fablet, M., & Chambeyron, S.  
736 (2020). A Transposon Story: From TE Content to TE Dynamic Invasion of Drosophila Genomes  
737 Using the Single-Molecule Sequencing Technology from Oxford Nanopore. *Cells*, 9(8), 1776.  
738 <https://doi.org/10.3390/cells9081776>
- 739 Nakamura, W., Hirata, M., Oda, S., Chiba, K., Okada, A., Mateos, R. N., Sugawa, M., Iida, N., Ushiyama,  
740 M., Tanabe, N., Sakamoto, H., Kawai, Y., Tokunaga, K., Consortium, N. C. W., Tsujimoto, S.,  
741 Shiba, N., Ito, S., Yoshida, T., & Shiraishi, Y. (2023). *A comprehensive workflow for target*  
742 *adaptive sampling long-read sequencing applied to hereditary cancer patient genomes* (p.  
743 2023.05.30.23289318). medRxiv. <https://doi.org/10.1101/2023.05.30.23289318>

- 744 Norris, A. L., Workman, R. E., Fan, Y., Eshleman, J. R., & Timp, W. (2016). Nanopore sequencing  
745 detects structural variants in cancer. *Cancer Biology & Therapy*, *17*(3), 246–253.  
746 <https://doi.org/10.1080/15384047.2016.1139236>
- 747 Oren, E., Dafna, A., Tzuri, G., Halperin, I., Isaacson, T., Elkabetz, M., Meir, A., Saar, U., Ohali, S., La, T.,  
748 Romay, C., Tadmor, Y., Schaffer, A. A., Buckler, E. S., Cohen, R., Burger, J., & Gur, A. (2022).  
749 Pan-genome and multi-parental framework for high-resolution trait dissection in melon  
750 (*Cucumis melo*). *The Plant Journal*, *112*(6), 1525–1542. <https://doi.org/10.1111/tpj.16021>
- 751 Payne, A., Holmes, N., Clarke, T., Munro, R., Debebe, B. J., & Loose, M. (2021). Readfish enables  
752 targeted nanopore sequencing of gigabase-sized genomes. *Nature Biotechnology*, *39*(4), Art.  
753 4. <https://doi.org/10.1038/s41587-020-00746-x>
- 754 Payne, A., Munro, R., Holmes, N., Moore, C., Carlile, M., & Loose, M. (2021). *Barcode aware adaptive*  
755 *sampling for GridION and PromethION Oxford Nanopore sequencers* [Preprint]. Genomics.  
756 <https://doi.org/10.1101/2021.12.01.470722>
- 757 Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: Quick coverage calculation for genomes and  
758 exomes. *Bioinformatics*, *34*(5), 867–868. <https://doi.org/10.1093/bioinformatics/btx699>
- 759 R Core Team (2021). R: A language and environment for statistical computing. R Foundation for  
760 Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/> .
- 761 Salinier, J., Lefebvre, V., Besombes, D., Burck, H., Causse, M., Daunay, M.-C., Dogimont, C.,  
762 Goussopoulos, J., Gros, C., Maisonneuve, B., McLeod, L., Tobal, F., & Stevens, R. (2022). The  
763 INRAE Centre for Vegetable Germplasm: Geographically and Phenotypically Diverse  
764 Collections and Their Use in Genetics and Plant Breeding. *Plants*, *11*(3), Art. 3.  
765 <https://doi.org/10.3390/plants11030347>
- 766 Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K.,  
767 Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J.,  
768 Kilburn, D., Sorensen, M., Munson, K. M., ... Paten, B. (2020). Nanopore sequencing and the

- 769 Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature*  
770 *Biotechnology*, 38(9), Art. 9. <https://doi.org/10.1038/s41587-020-0503-6>
- 771 Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., Lin, H., Hu, M., Zhao, F., Zhang, C., Li, Y., Gao, H.,  
772 Wang, T., Liu, X., Zhang, H., Zhang, Y., Cao, S., Yu, X., Zhang, B., ... Qian, Q. (2022). A super  
773 pan-genomic landscape of rice. *Cell Research*, 32(10), Art. 10.  
774 <https://doi.org/10.1038/s41422-022-00685-z>
- 775 Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File  
776 Manipulation. *PLOS ONE*, 11(10), e0163962. <https://doi.org/10.1371/journal.pone.0163962>
- 777 Shin, A.-Y., Koo, N., Kim, S., Sim, Y. M., Choi, D., Kim, Y.-M., & Kwon, S.-Y. (2019). Draft genome  
778 sequences of two oriental melons, *Cucumis melo* L. var. Makuwa. *Scientific Data*, 6(1), Art. 1.  
779 <https://doi.org/10.1038/s41597-019-0244-x>
- 780 Stevanovski, I., Chintalaphani, S. R., Gamaarachchi, H., Ferguson, J. M., Pineda, S. S., Scriba, C. K.,  
781 Tchan, M., Fung, V., Ng, K., Cortese, A., Houlden, H., Dobson-Stone, C., Fitzpatrick, L.,  
782 Halliday, G., Ravenscroft, G., Davis, M. R., Laing, N. G., Fellner, A., Kennerson, M., ... Deveson,  
783 I. W. (2022). Comprehensive genetic diagnosis of tandem repeat expansion disorders with  
784 programmable targeted nanopore sequencing. *Science Advances*, 8(9), eabm5386.  
785 <https://doi.org/10.1126/sciadv.abm5386>
- 786 Su, J., Lui, W. W., Lee, Y., Zheng, Z., Siu, G. K.-H., Ng, T. T.-L., Zhang, T., Lam, T. T.-Y., Lao, H.-Y., Yam,  
787 W.-C., Tam, K. K.-G., Leung, K. S.-S., Lam, T.-W., Leung, A. W.-S., & Luo, R. (2023). Evaluation  
788 of Mycobacterium tuberculosis enrichment in metagenomic samples using ONT adaptive  
789 sequencing and amplicon sequencing for identification and variant calling. *Scientific Reports*,  
790 13(1), Art. 1. <https://doi.org/10.1038/s41598-023-32378-x>
- 791 Toda, N., Rustenholz, C., Baud, A., Le Paslier, M.-C., Amselem, J., Merdinoglu, D., & Faivre-Rampant,  
792 P. (2020). NLGenomeSweeper: A tool for genome-wide NBS-LRR resistance gene  
793 identification. *Genes*, 11(3), 333.

- 794 Ulrich, J.-U., Lutfi, A., Rutzen, K., & Renard, B. Y. (2022). ReadBouncer: Precise and scalable adaptive  
795 sampling for nanopore sequencing. *Bioinformatics*, *38*(Supplement\_1), i153–i160.  
796 <https://doi.org/10.1093/bioinformatics/btac223>
- 797 Van de Weyer, A.-L., Monteiro, F., Furzer, O. J., Nishimura, M. T., Cevik, V., Witek, K., Jones, J. D. G.,  
798 Dangl, J. L., Weigel, D., & Bemm, F. (2019). A Species-Wide Inventory of NLR Genes and  
799 Alleles in *Arabidopsis thaliana*. *Cell*, *178*(5), 1260-1272.e14.  
800 <https://doi.org/10.1016/j.cell.2019.07.038>
- 801 Van Wersch, S., & Li, X. (2019). Stronger When Together: Clustering of Plant NLR Disease resistance  
802 Genes. *Trends in Plant Science*, *24*(8), 688–699.  
803 <https://doi.org/10.1016/j.tplants.2019.05.005>
- 804 Vaser, R., & Šikić, M. (2021). Time- and memory-efficient genome assembly with Raven. *Nature*  
805 *Computational Science*, *1*(5), Art. 5. <https://doi.org/10.1038/s43588-021-00073-4>
- 806 Vendelbo, N. M., Mahmood, K., Steuernagel, B., Wulff, B. B. H., Sarup, P., Hovmøller, M. S., Justesen,  
807 A. F., Kristensen, P. S., Orabi, J., & Jahoor, A. (2022). Discovery of Resistance Genes in Rye by  
808 Targeted Long-Read Sequencing and Association Genetics. *Cells*, *11*(8), Art. 8.  
809 <https://doi.org/10.3390/cells11081273>
- 810 Weilguny, L., De Maio, N., Munro, R., Manser, C., Birney, E., Loose, M., & Goldman, N. (2023).  
811 Dynamic, adaptive sampling during nanopore sequencing using Bayesian experimental  
812 design. *Nature Biotechnology*, *41*, 1018–1025. <https://doi.org/10.1038/s41587-022-01580-z>
- 813 Witek, K., Jupe, F., Witek, A. I., Baker, D., Clark, M. D., & Jones, J. D. G. (2016). Accelerated cloning of  
814 a potato late blight–resistance gene using RenSeq and SMRT sequencing. *Nature*  
815 *Biotechnology*, *34*(6), Art. 6. <https://doi.org/10.1038/nbt.3540>
- 816 Wrenn, D. C., & Drown, D. M. (2023). *Nanopore Adaptive Sampling Enriches for Antimicrobial*  
817 *Resistance Genes in Microbial Communities* [Preprint]. *Bioinformatics*.  
818 <https://doi.org/10.1101/2023.06.27.546783>

819 Zhang, H., Li, H., Jain, C., Cheng, H., Au, K. F., Li, H., & Aluru, S. (2021). Real-time mapping of  
820 nanopore raw signals. *Bioinformatics*, 37(Supplement\_1), i477–i483.

821 <https://doi.org/10.1093/bioinformatics/btab264>

822 Zhang, W., Yuan, Q., Wu, Y., Zhang, J., & Nie, J. (2022). Genome-Wide Identification and  
823 Characterization of the CC-NBS-LRR Gene Family in Cucumber (*Cucumis sativus* L.).  
824 *International Journal of Molecular Sciences*, 23(9), Art. 9.

825 <https://doi.org/10.3390/ijms23095048>

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

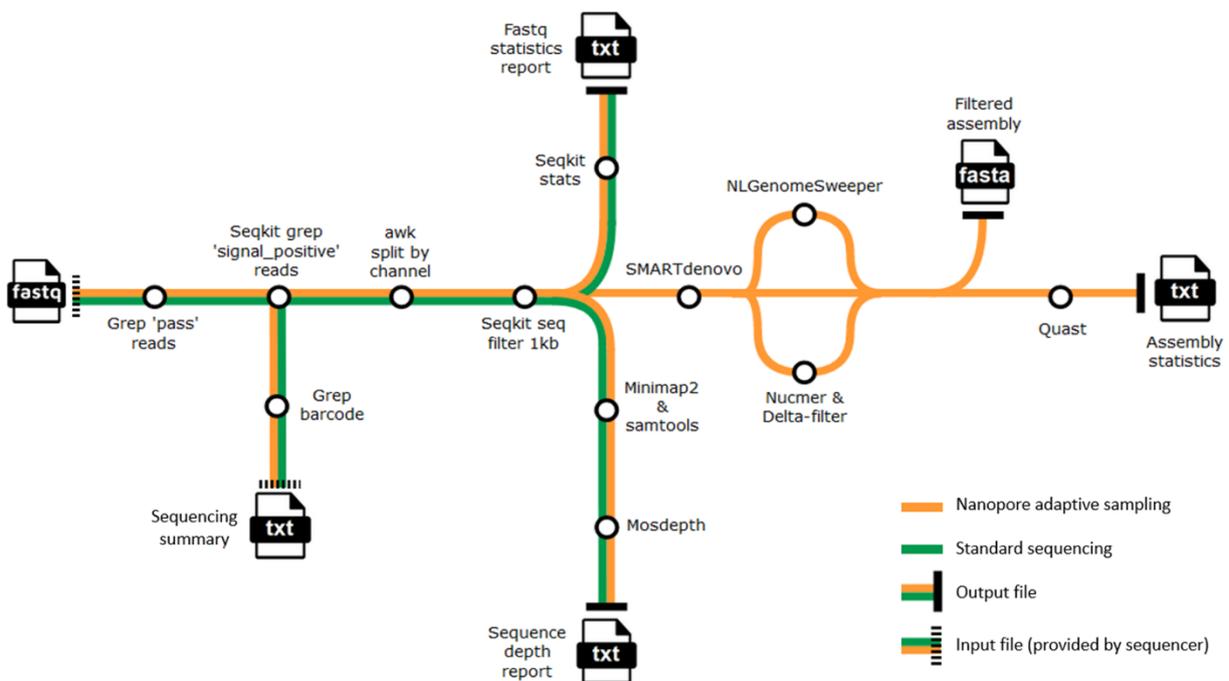
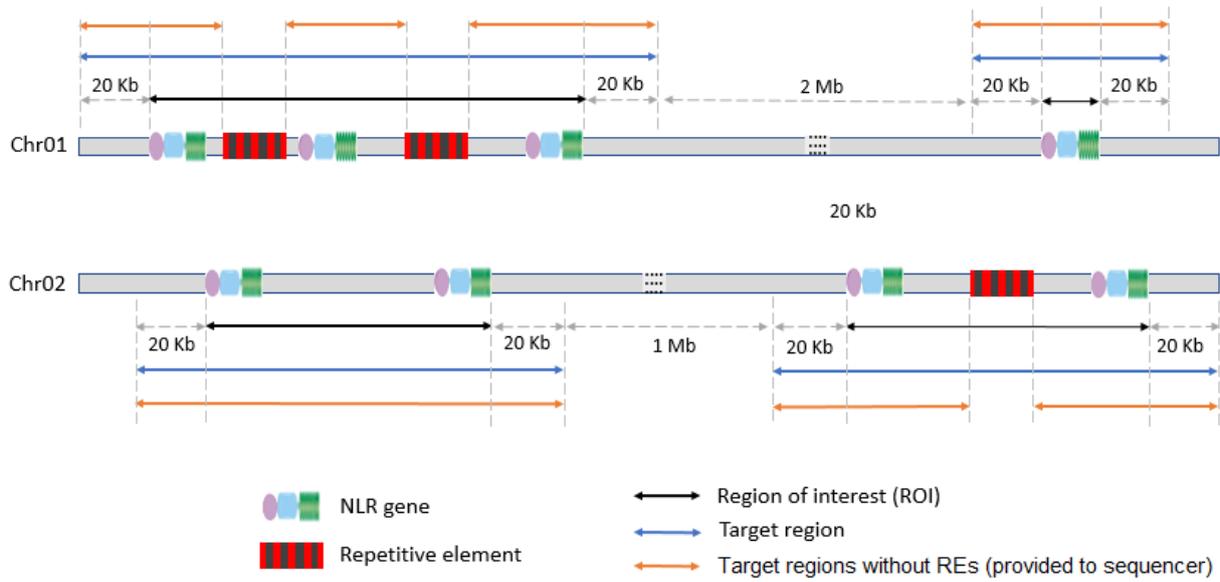
842

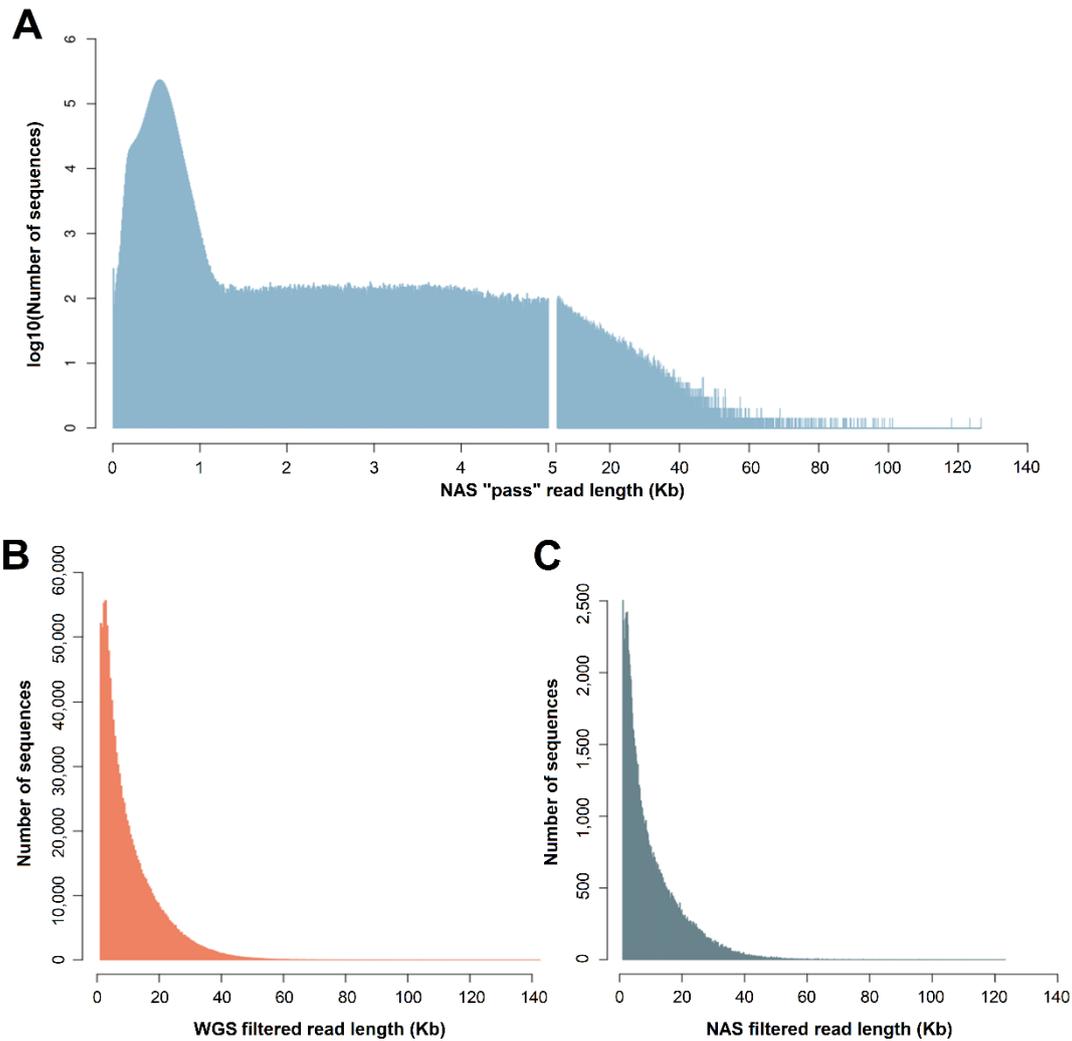
843

844

845

846 **Figures**

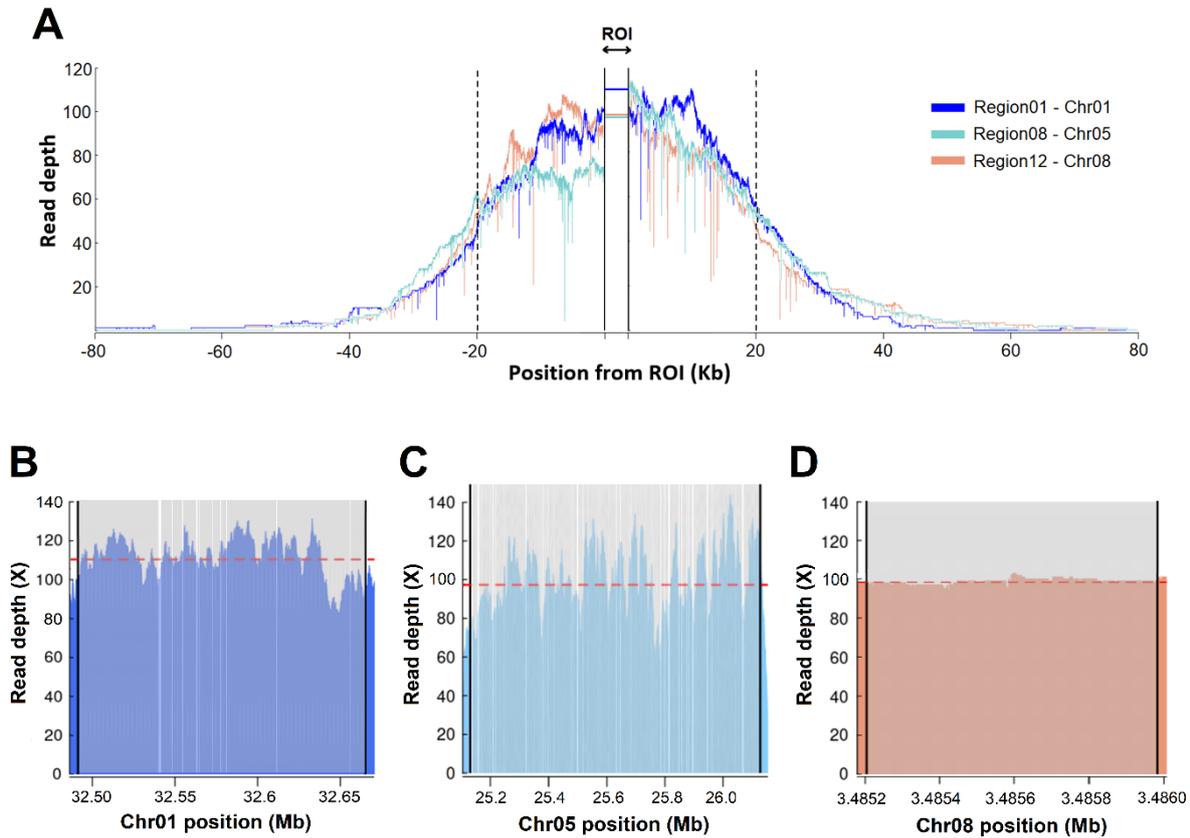




856

857 **Figure 3.** A) Length distribution of “PASS”-tagged NAS reads. The number of reads was log-  
858 transformed. B) Length distribution of WGS reads after filtering by “end reason”, quality and  
859 length. C) Length distribution of NAS reads after filtering by “end reason”, quality and length.

860



861

862 **Figure 4.** A) NAS sequencing depth on three target regions of different sizes compared to the  
863 rest of the chromosome. Target regions (ROI + 20kb buffer) are represented between black  
864 dotted bars, while ROIs are collapsed and represented between black solid bars. B) NAS  
865 sequencing depth on the ROI of region 01 ( $\approx 173$  kb). C) NAS sequencing depth on the ROI of  
866 region 08 ( $\approx 998$  kb) D) NAS sequencing depth on the ROI of region 12 ( $\approx 1$  kb). Region 08  
867 contains the well-studied *Vat* cluster. For B, C and D, vertical colored bars represent the  
868 enriched regions, while vertical white bars represent masked repetitive elements.

869

870

871

872

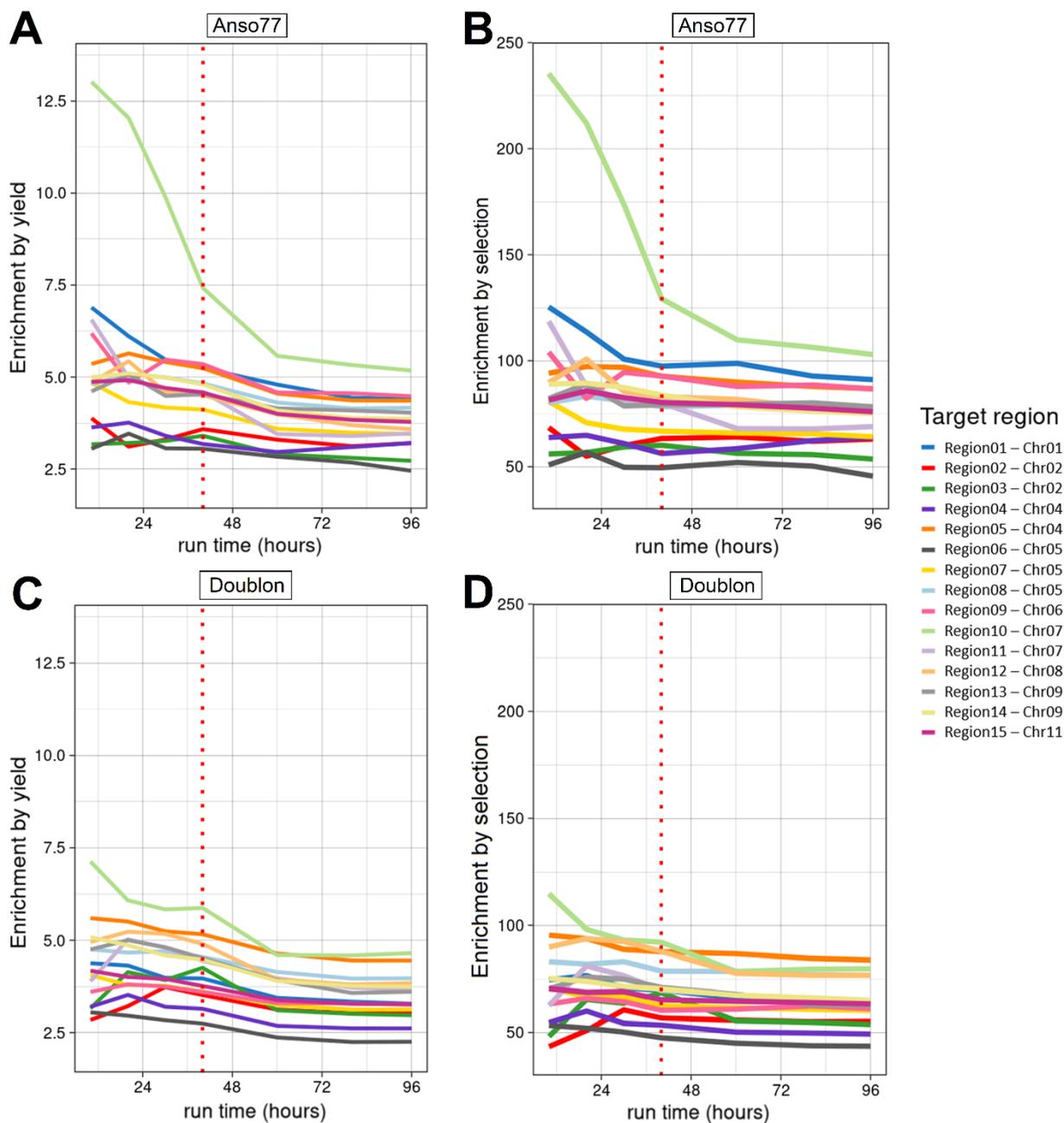
873

874

875

876

877



878

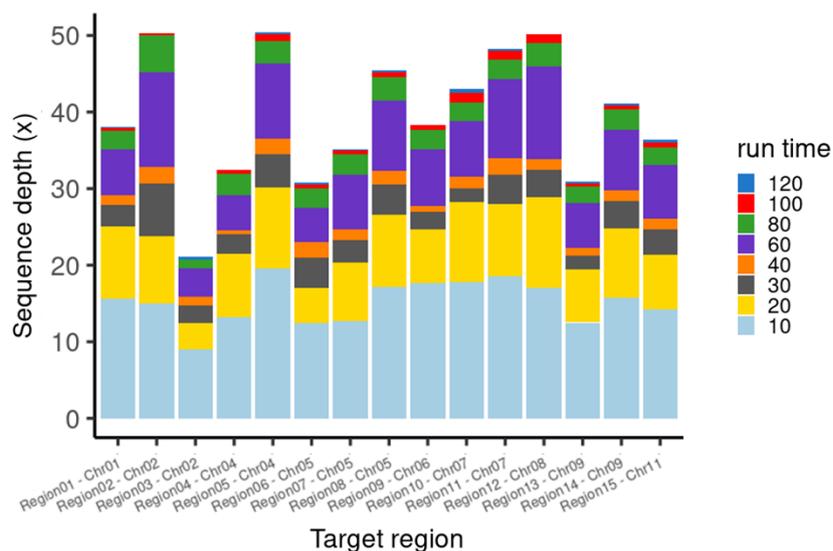
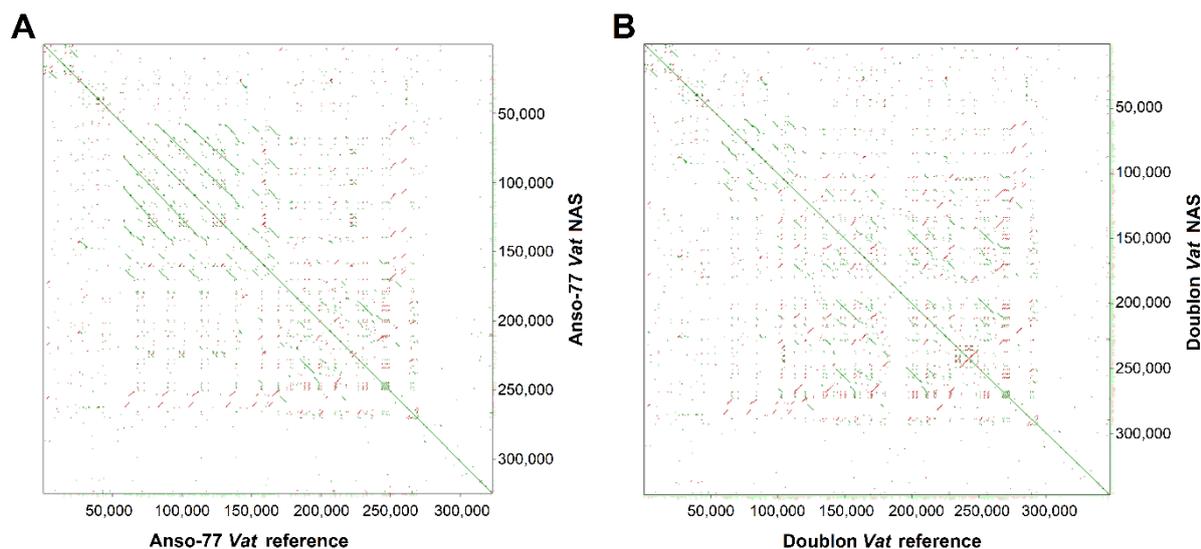
879 **Figure 5.** Enrichment by yield (A, C) and enrichment by selection (B, D) of the 15 target  
880 regions from Anso77 (A, B) and Doublon (C, D). Vertical red-dotted bars denote the flowcell  
881 washing flush time.

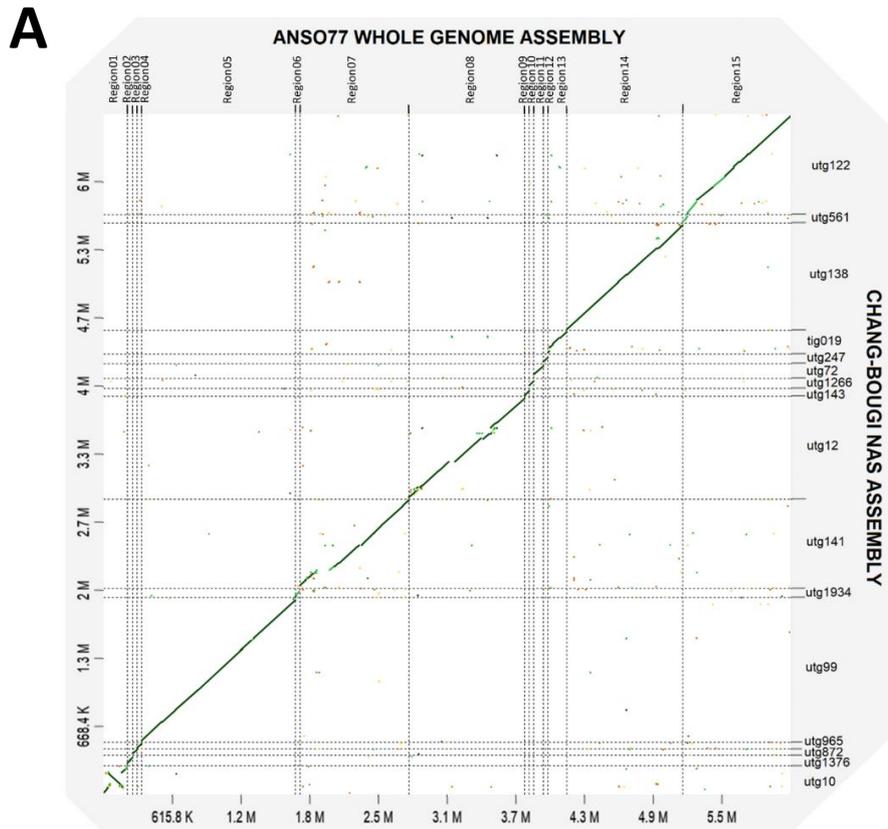
882

883

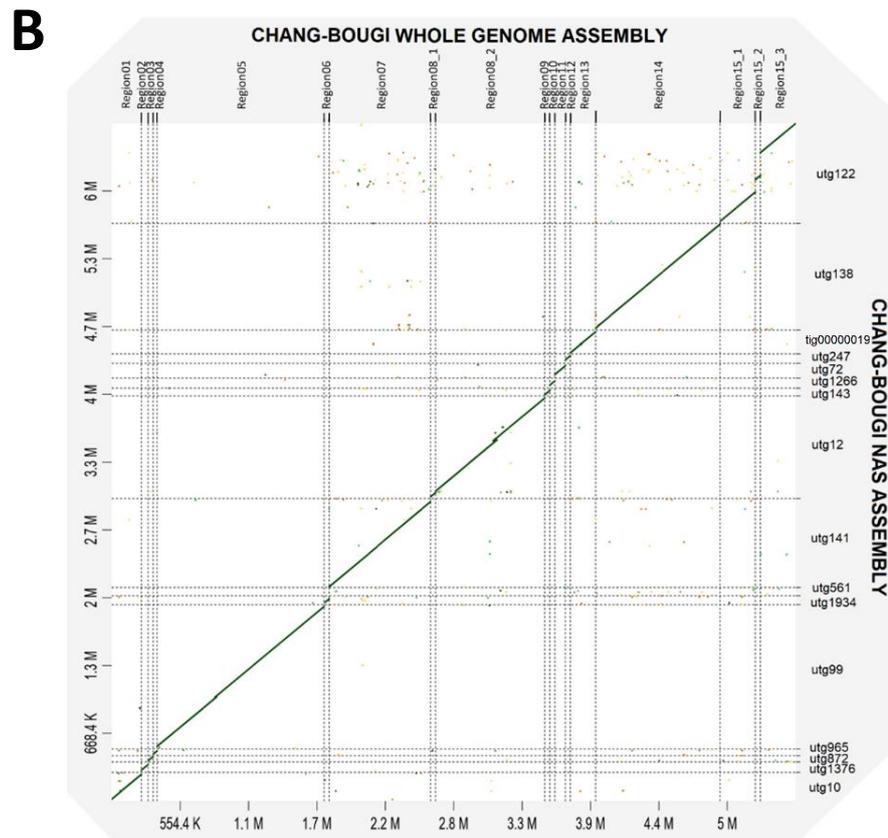
884

885



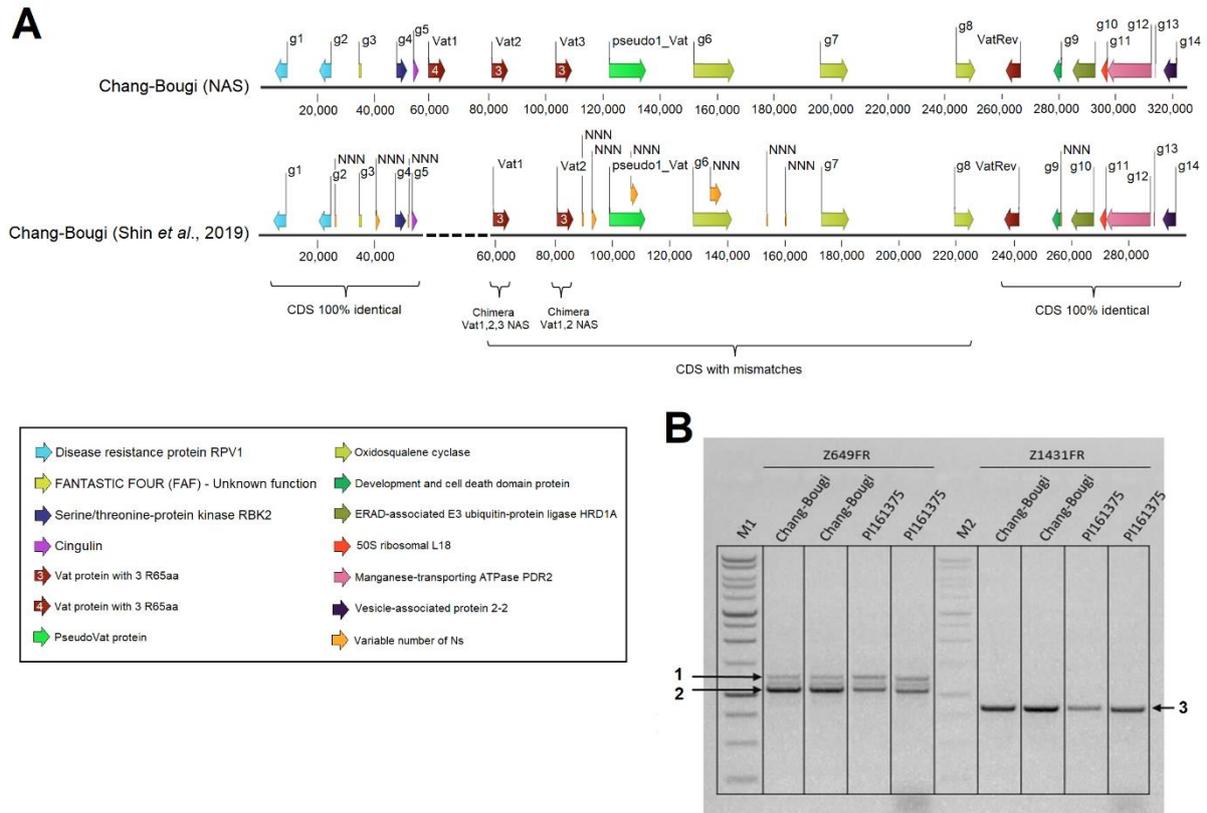


899



900

901 **Figure 8.** Dot plots representing the NAS filtered assembly of Chang-Bougi (y-axis) against  
902 the 15 target regions from Anso77 (A) and the 18 NLR clusters from Chang-Bougi (B).



903

904 **Figure 9.** A) Genes identified after manual annotation within the *Vat* regions of Chang-Bougi.  
 905 The sequence above was obtained from the NAS assembly, while the sequence below was  
 906 recovered from the publicly available genome assembly (Shin *et al.*, 2019). B) Agarose gel  
 907 electrophoresis of PCR products obtained using primers Z649FR and Z1431FR. Lanes M1 and  
 908 M2 are the two 1 kb DNA ladders (Promega, Madison, WI, USA). PI161375 was used as a  
 909 control having a *Vat1* with four R65aa motifs and a *Vat2* with three R65aa motifs. Bands  
 910 pointed with arrows represent an amplicon of four R65aa motifs (1), an amplicon of three R65aa  
 911 motifs (2), and a specific amplicon of four R65aa motifs (3).

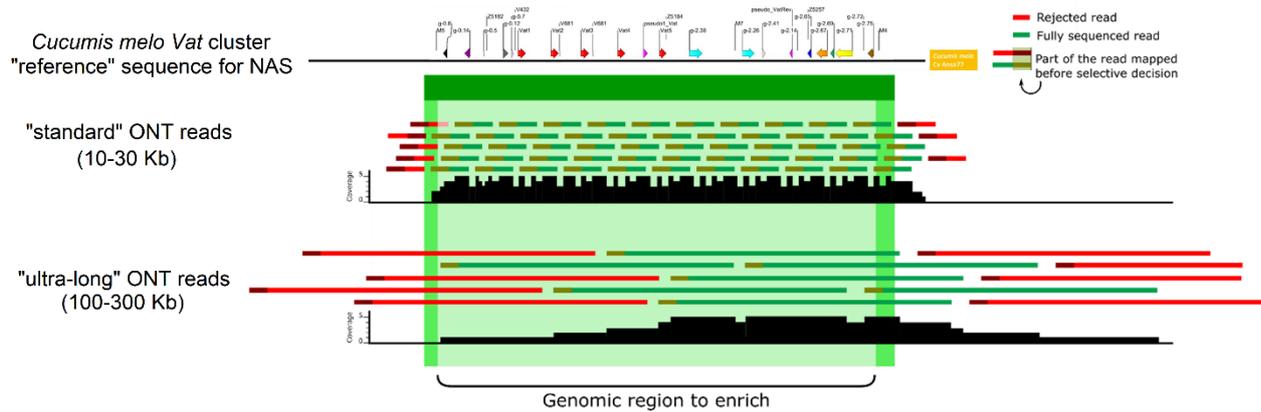
912

913

914

915

916



917

918 **Figure 10.** Diagram illustrating the difference in coverage and extent of the area outside the  
919 region to be enriched for standard size fragments (10-30kb) and for ultra-high molecular weight  
920 fragments (100-300kb) on the *Vat* (melon) cluster. For the same yield (illustrated here by an  
921 arbitrary overall depth of 5X), standard fragments make it possible to achieve a depth more  
922 concentrated on the area to be enriched and to sequence less outside this area. For convenience,  
923 only reads oriented from 5' to 3' are represented.

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940 **Tables**

941 **Table 1.** Summary metrics of the Anso77 and Doublon hybrid genome assemblies and  
 942 annotations.

		<b>Anso77</b>	<b>Doublon</b>
<b>Chromosome assembly</b>	Number of chromosomes	12	12
	Number of scaffolds	24	22
	Min. scaffold length (Mb)	3.4	4.6
	Max. scaffold length (Mb)	29.0	29.0
	Total size of scaffolds (Mb)	356.6	349.7
<b>Chromosome 0</b>	Size (Mb)	9.1	9.5
	GC (%)	34	34
<b>Whole genome</b>	BUSCO Complete genes	2275 (97.8%)	2182 (93.8%)
	BUSCO Duplicated genes	38 (1.6%)	33 (1.4%)
	BUSCO Fragmented genes	12 (0.5%)	12 (0.5%)
	BUSCO Missing genes	39 (1.7%)	132 (5.7%)
	Number of predicted genes	21,928	21,553
	Number of predicted NBS domains	84	76
	Number of predicted CC-NBS-LRRs	18	15
	Number of predicted TIR-NBS-LRRs	24	23
	Number of predicted RPW8-NBS-LRRs	1	1
	Number of predicted NBS-LRRs	31	30
	Number of predicted TIR-NBSs	4	1
	Number of predicted NBSs	6	5

943

944

945

946 **Table 2.** Detailed information about the 15 target regions of Anso77. The *Vat* region is included  
947 within region 08.

<b>Target region</b>	<b>Chr</b>	<b>Start position</b>	<b>End position</b>	<b>Size</b>	<b>Predicted NBS domains</b>
<b>Region 01</b>	01	32,471,836	32,684,765	212,929	14
<b>Region 02</b>	02	979,011	1,026,074	47,063	2
<b>Region 03</b>	02	15,072,686	15,113,435	40,749	1
<b>Region 04</b>	04	5,452,467	5,493,242	40,775	1
<b>Region 05</b>	04	26,516,353	27,894,199	1,377,846	7
<b>Region 06</b>	05	14,002,942	14,043,763	40,821	1
<b>Region 07</b>	05	17,364,466	18,340,871	976,405	2
<b>Region 08</b>	05	25,108,836	26,146,869	1,038,033	28
<b>Region 09</b>	06	5,837,171	5,877,998	40,827	1
<b>Region 10</b>	07	2,550,415	2,591,165	40,750	1
<b>Region 11</b>	07	24,326,063	24,411,978	85,915	4
<b>Region 12</b>	08	3,465,204	3,505,984	40,780	1
<b>Region 13</b>	09	645,676	815,779	170,103	10
<b>Region 14</b>	09	6,558,754	7,597,791	1,039,037	4
<b>Region 15</b>	11	6,635,285	7,601,705	966,420	7

948

949

950

951

952

953

954

955

956 **Table 3.** Anso77, Doublon and Chang-Bougi sequencing metrics. Anso77 and Doublon were  
 957 sequenced using both NAS and WGS, while Chang-Bougi was only sequenced using NAS.  
 958 Total WGS “pass” and total NAS “pass” represent the reads having a “PASS” flag (quality over  
 959 10) and assigned to the WGS and NAS half-flowcell, respectively. Filtered WGS and filtered  
 960 NAS represent the two sets of reads just mentioned after filters of “end reason” and length.

<b>Dataset</b>	<b>Cultivar</b>	<b>Sequences number</b>	<b>Size (bp)</b>	<b>Mean length (bp)</b>	<b>Max. length (bp)</b>	<b>N50</b>	<b>Q20 (%)</b>	<b>Q30 (%)</b>
<b>Total dataset</b>	<b>Anso77</b>	8,628,783	19,860,863,026	2,302	174,519	13,263	82.17	72.04
	<b>Doublon</b>	11,976,866	26,639,849,880	2,224	193,630	11,968	83.65	73.58
	<b>Chang-Bougi</b>	3,320,000	3,454,320,986	1,041	467,827	883	76.23	58.39
<b>Total WGS “pass”</b>	<b>Anso77</b>	1,423,908	13,579,603,804	9,536	143,366	16,713	87.55	77.18
	<b>Doublon</b>	2,033,986	18,285,705,305	8,990	137,509	15,261	87.55	77.33
<b>Total NAS “pass”</b>	<b>Anso77</b>	6,634,798	4,807,120,593	725	126,590	617	87.89	77.41
	<b>Doublon</b>	9,407,579	6,779,872,639	721	112,867	614	88.01	77.78
	<b>Chang-Bougi</b>	3,056,000	3,115,590,337	1,020	232,918	878	82.81	64.01
<b>Filtered WGS</b>	<b>Anso77</b>	1,122,605	11,939,444,337	10,636	143,366	16,772	88.00	77.75
	<b>Doublon</b>	1,704,828	15,807,284,512	9,272	137,509	15,120	88.08	78.00
<b>Filtered NAS</b>	<b>Anso77</b>	110,061	1,147,896,611	10,430	123,373	16,911	88.17	78.03
	<b>Doublon</b>	163,843	1,561,078,862	9,528	112,867	15,182	88.09	77.99
	<b>Chang-Bougi</b>	96,626	803,580,422	8,316	122,918	13,749	81.40	66.60

961

962

963

964

965

966

967 **Table 4.** Summary statistics of the NAS and WGS runs for Anso77 and Doublon. Only the  
968 chromosomes including target regions were considered for calculating the average off-target  
969 depth.

	<b>WGS Anso77</b>	<b>NAS Anso77</b>	<b>WGS Doublon</b>	<b>NAS Doublon</b>
<b>Average on-target depth</b>	22.79	90.18	31.74	118.28
<b>Average off-target depth</b>	28.19	1.42	32.52	1.73
<b>Relative frequency of target regions</b>	0.81	63.37	0.98	68.21
<b>Average enrichment by yield</b>		3.96		3.73
<b>Average enrichment by selection</b>		78.38		69.92

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986 **Table 5.** Anso77 and Doublon NAS assembly metrics. Regions of the NAS reference file  
 987 represent the size (bp) of the target regions provided to the sequencer. Regions of Anso77,  
 988 Doublon and Chang-Bougi represent the size (bp) of the contigs matching those regions.

	<b>NAS reference file</b>	<b>Anso77</b>	<b>Doublon</b>	<b>Chang-Bougi</b>
<b>Number of sequences</b>	15	15	15	16
<b>Assembly total size</b>	6,158,453	7,022,836	7,137,869	6,684,108
<b>%GC</b>	33	33	33	33
<b>Region 01</b>	212,929	267,987	288,544	283,266
<b>Region 02</b>	47,063	111,515	145,599	103,652
<b>Region 03</b>	40,749	92,325	109,144	62,015
<b>Region 04</b>	40,775	88,396	91,470	65,369
<b>Region 05</b>	1,377,846	1,442,885	1,469,924	1,422,416
<b>Region 06</b>	40,821	80,095	94,491	87,574
<b>Region 07</b>	976,405	1,030,491	1,002,692	877,386
<b>Region 08</b>	1,038,033	1,109,994	1,047,373	1,011,255
<b>Region 09</b>	40,827	93,035	116,040	76,550
<b>Region 10</b>	40,750	95,015	127,865	97,965
<b>Region 11</b>	85,915	132,117	154,578	146,542
<b>Region 12</b>	40,780	127,510	110,582	93,626
<b>Region 13</b>	170,103	218,950	239,877	233,624
<b>Region 14</b>	1,039,037	1,106,869	1,072,236	1,052,862
<b>Region 15</b>	966,420	1,025,652	1,067,454	82,354 & 987,652

989