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

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

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

Javier Belinchon-Moreno[1,2], Aurélie Berard[1], Aurélie Canaguier[1], Véronique Chovelon[2], Corinne Cruaud[3], Stéfan Engelen[4], Rafael Feriche-Linares[2], Isabelle Le-Clainche[1], William Marande[5], Vincent Rittener-Ruff[2], Jacques Lagnel[2], Damien Hinsinger[1], Nathalie Boissot[2] and Patricia Faivre-Rampant[1*]

Abstract

Background Nanopore adaptive sampling (NAS) offers a promising approach for assessing genetic diversity in targeted genomic regions. Here we designed and validated an experiment to enrich a set of resistance genes in several melon cultivars as a proof of concept.

Results Using the same reference to guide read acceptance or rejection with NAS, we successfully and accurately reconstructed the 15 regions in two newly assembled ssp. *melo* genomes and in a third ssp. *agrestis* cultivar. We obtained fourfold enrichment regardless of the tested samples, but with some variations according to the enriched regions. The accuracy of our assembly was further confirmed by PCR in the *agrestis* cultivar. We discussed parameters that could influence the enrichment and accuracy of NAS generated assemblies.

Conclusions Overall, we demonstrated that NAS is a simple and efficient approach for exploring complex genomic regions, such as clusters of Nucleotide-binding site leucine-rich repeat (NLR) resistance genes. These regions are characterized by containing a high number of copy number variations, presence-absence polymorphisms and repetitive elements. These features make accurate assembly challenging but are crucial to study due to their central role in plant immunity and disease resistance. This approach facilitates resistance gene characterization in a large number of individuals, as required when breeding new cultivars suitable for the agroecological transition.

Keywords Melon, Nanopore adaptive sampling, Targeted sequencing, Resistance genes, NLR, Genome assembly

*Correspondence:
Patricia Faivre-Rampant
patricia.faivre-rampant@inrae.fr
[1]Université Paris-Saclay, Centre INRAE Île-de-France Versailles-Saclay, EPGV, Evry 91057, France
[2]INRAE, Génétique et Amélioration des Fruits et Légumes, Montfavet 84143, France
[3]Commissariat à l'Energie Atomique (CEA), Genoscope, Institut de Biologie François-Jacob, Université Paris-Saclay, 2 Rue Gaston Crémieux, Evry 91057, France
[4]Génomique Métabolique, Institut François Jacob, Commissariat à l'Energie Atomique (CEA), CNRS, Univ. Evry, Université Paris-Saclay, Genoscope, Evry 91057, France
[5]INRAE, Centre National de Ressources Génomiques Végétales, Castanet-Tolosan 31326, France

Background

The correct assembly of complex and highly repeated genome regions, which are especially prevalent in plants, remains a challenge. Short-read sequencing is a cost-effective method that is widely used in whole genome sequencing (WGS) approaches, yet it is ineffective in these regions as the read length is not sufficient for proper assessment of repeated elements, copy-number variations and duplication events [1]. Long-read technologies, such as those developed by Oxford Nanopore Technologies (ONT, Oxford, UK), have demonstrated their potential in accurately resolving complex regions as they are able to span long repetitive elements or areas

Belinchon-Moreno *et al. BMC Genomics*   (2025) 26:126

Page 2 of 18

with tandemly repeated genes [2, 3]. Whole genome long-read sequencing is nevertheless still too costly for most studies, particularly when numerous genotypes have to be sequenced to highlight a few specific regions of interest [4].

Targeted sequencing approaches offer a valuable alternative for characterizing specific genomic regions while reducing sequencing and data storage costs as compared to WGS [5]. Current targeted sequencing protocols have been adapted for long-read sequencing and mainly involve hybridization capture [6], PCR amplification [7], Cas9-assisted targeting [8] or microfluidic-based droplet sorting [9]. However, these approaches require substantial experimental and design efforts, along with substantial prior knowledge of the sequence to be enriched and its genotypic diversity [5, 8]. PCR-based techniques are particularly prone to introducing bias in the enriched sequences, and long amplicons are hard to consistently amplify [5]. Hybridization-based methods require the construction of complex RNA libraries alongside very specific hybridization and capture conditions [5]. Cas9-based methods, e.g. Nanopore Cas9-targeted sequencing (nCATS) [8] or Cas9-Assisted Targeting of CHromosome segments (CATCH) [10], require the design of multiple guide-RNAs, which may be a challenging task when dealing with complex and repetitive genome regions. Finally, microfluidic-based methods like Xdrop [9] are highly complex and require specialized microfluidic equipment [5].

The Nanopore adaptive sampling (NAS) approach recently developed by ONT overcomes these limitations. NAS was first suggested in 2016 [11] and has been implemented with different algorithms since late 2019 [12–15]. This strategy takes advantage of the ability of the pores to control the directional flow of the DNA strand being sequenced through alterations in the applied current polarity. By combining live calling of sequenced bases with real-time mapping to a set of DNA sequences provided by the user for enrichment, the DNA strand may be dynamically discarded or fully sequenced based on the similarity of its initial first few hundred bases to the provided reference [11]. NAS just requires standard library preparation, while overcoming the need for DNA amplification, laborious and expensive experimental design or probe synthesis, and it offers real-time selective enrichment [16, 17]. NAS has been used in clinical settings and for metagenomic sample enrichment [16, 18–23]. NAS is therefore a promising approach for studying target regions, especially those that are highly complex, such as disease-associated repeat loci in humans [17, 24].

In plants, immunity is encoded by resistance genes (R genes), frequently organized in complex regions [25]. Among R genes, Nucleotide-binding site leucine-rich repeat resistance genes (NLRs) form the largest family

[26]. These genes encode intracellular receptors that play a central role in the so-called effector-triggered immunity (ETI) against pathogens. NLR genes exhibit a highly conserved structure with three main domains [26, 27]: the N-terminal domain, the central domain, and the C-terminal domain. The N-terminal domain can be a Toll/Interleukin-1 receptor (TIR), a Coiled-coil (CC), or a resistance to Powdery Mildew 8-like (RPW8) domain. The central domain, the most conserved one, is a nucleotide-binding adaptor (NB-ARC), also named as NBS (nucleotide-binding site) domain. This domain plays a crucial role in signal transduction. Finally, the C-terminal domain is often composed of leucine-rich repeats (LRR) with ligand-binding functions. A clustered genomic arrangement is a common characteristic of NLR genes [28]. These clusters often result from unequal crossing overs, tandem duplications, or intra-cluster rearrangements [26]. In this context, NAS, combining long-read sequencing and target enrichment, should allow the accurate characterization of NLR clusters in plants.

We selected melon (*Cucumis melo* L.) as a model to investigate the ability of NAS to efficiently sequence a complete set of NLR clusters within a species (or NLRome). The melon genome features: i/ a small genome size; ii/ an NLR content estimated to account for ≈1% of the genome [29], which is in line with ONT target size recommendations [30]; and iii/ a finely characterized, highly variable complex NLR cluster, *Vat* [31, 32], that is suitable for benchmarking. Among the accessions that are well characterized with regard to the *Vat* region, we chose Anso77 (ssp. *melo*) because it features the highest number of functional *Vat* genes [31]. We also selected Doublon (ssp. *melo*) as an accession whose *Vat* region structure contrasts with that of Anso77 [31]. We assembled and annotated their whole genomes and selected Anso77 as reference for identifying regions of interest (ROIs) for NAS. The performance of the method in capturing a set of NLR clusters in Anso77 and Doublon was assessed. Furthermore, we extended our assessment to an accession of a different subspecies (Chang-Bougi, ssp. *agrestis*) for which open-access genomic data are available [33].

Materials and methods
Biological material
We selected Anso77, Doublon and Chang-Bougi melon cultivars to develop a proof of concept for the NLRome adaptive sampling experiment, with Anso77 serving as the reference cultivar. These cultivars originated from Spain, France and Korea, respectively. Anso77 and Doublon were chosen as ssp. *melo* lines belonging to the *inodorus* and *cantalupensis* botanical groups. Chang-Bougi, belonging to ssp. *agrestis* and specifically to the *makuwa* botanical group, was selected as a cultivar

distantly related to Anso77 and Doublon. The aim of this choice was to validate the NAS procedure with cultivars differing markedly from the selected reference. Moreover, a draft genome assembly for Chang-Bougi constructed via Illumina HiSeq reads was readily available [33].

We obtained the seeds from the INRAE Centre for Vegetable Germplasm in Avignon, France [34], and grew them under greenhouse conditions at the INRAE GAFL research unit, Avignon, France.

Anso77 and Doublon de novo whole genome sequencing, assembly and annotation

We produced whole *de novo* Anso77 and Doublon genome assemblies using long-read sequencing: ONT for Anso77 and ONT combined with PacBio (Pacific Biosciences, Menlo Park, CA, USA) for Doublon. Raw reads were already deposited in the NCBI database under the following Bioproject accession numbers: PRJNA662717 and PRJNA662721 [31]. BioNano optical maps (BioNano Genomics, San Diego, CA, USA), 10x Linked-Reads (Pleasanton, CA, USA) for Anso77, Illumina Novaseq short-read sequencing (Illumina, San Diego, CA, USA), and linkage map information were developed and used to construct the genome assemblies. We performed gene prediction on the assembled genomes using two independent procedures: A combination of *ab initio*, homology-based and transcriptome-based methods implemented with EuGene v. 4.3 [35]; and an ab initio approach using deep-learning implemented with Helixer [36]. Functional annotation of predicted genes was performed using the EggNOG-mapper v. 2.1.12 suite [37]. The fully detailed

methods and parameters used for the assemblies and annotations are provided in Additional file 1: Supplementary Methods.

NAS enrichment panel definition and experimental design

We used Anso77 cultivar as reference for constructing the target regions for the NAS approach. We predicted the presence of NLR-related genes using NLGenomeSweeper [38] with default parameters. This tool approximates the presence of NLR genes via identification of the well-conserved NBS domain. We defined the regions of interest (ROIs) by grouping the predicted NBS domains separated by regions < 1 Mb. We ensured that there would be robust read depth coverage on the selected ROIs by adding a 20 kb buffer zone flanking the ROIs to constitute the initial target regions.

We performed repetitive elements (RE) annotation in the initial target regions using the CENSOR tool available on the curated GIRI Repbase website [39]. Predicted REs > 200 bp were excluded from the initial target regions. Moreover, sequences < 500 bp located between them were also excluded, as adaptive sampling requires ≈ 500 bp to accept or reject the DNA strand. Figure 1 provides definitions of the ROIs, target regions and target regions without REs.

We input these target regions without REs in bed format (Additional file 3) and the reference genome of Anso77 in fasta format in the MinKNOW software platform (ONT, Oxford, UK). These files were used for read acceptance or rejection. If the initial ~ 500 bp of the DNA strands matched the target regions without REs they
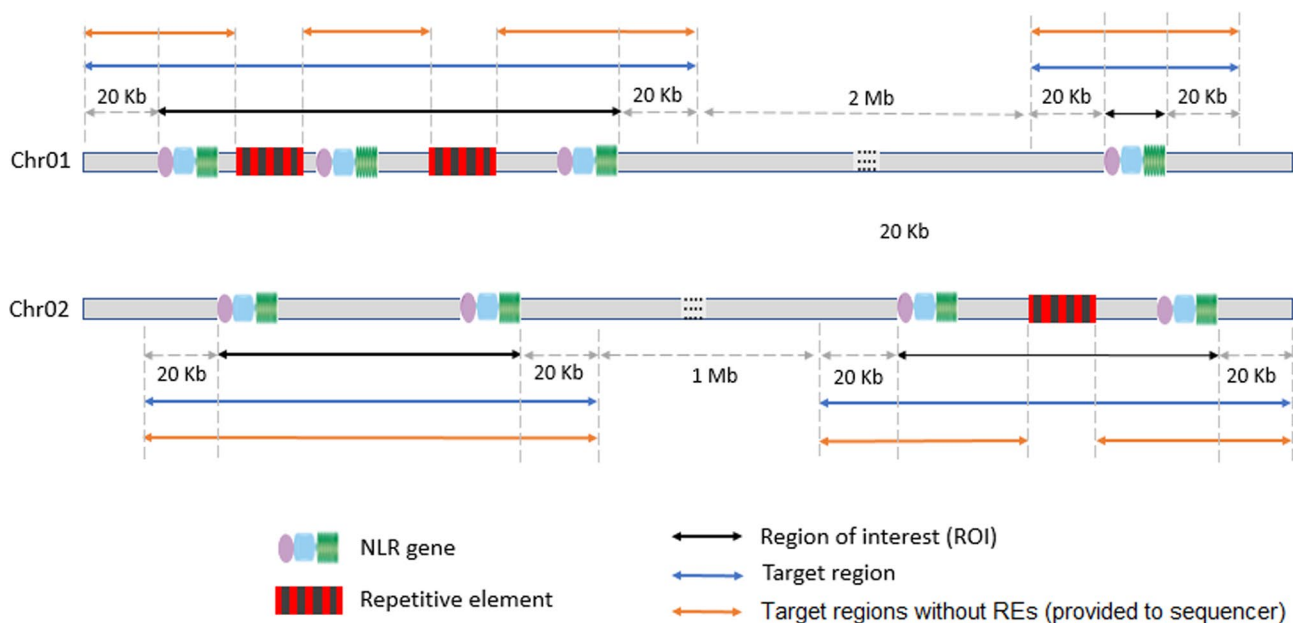
Fig. 1 Schematic representation of the definition of ROIs, target regions and target regions without repetitive elements (REs). Target regions without REs were provided to the MinKNOW software. Regions were defined in the reference genome for NAS, Anso77

underwent complete sequencing, otherwise they were ejected from the pore.

DNA extraction, adaptive sampling sequencing and base calling

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

We multiplexed and sequenced Anso77 and Doublon DNA on a single PromethION *R10*.4.1 flowcell (ONT, Oxford, UK). Half the channels were used as control channels in which no adaptive sampling was performed. Moreover, we sequenced Chang-Bougi DNA using one tenth of a PromethION *R10*.4.1 flowcell to assess the NAS flexibility and scalability. We prepared the sequencing libraries using the Native Barcoding Kit 24 V14 (SQK-NBD114.24) (ONT, Oxford, UK) according to the ONT guidelines with some modifications. One microgram of genomic DNA from each sample was repaired and end-prepped with 20 min incubation at 20 °C followed by a heat-inactivation of the enzymes at 65 °C for an additional 20 min. DNA was purified and barcodes were individually ligated to each of the purified DNA samples. NB01 and NB02 barcodes were used for Anso77 and Doublon, and NB22 was used for Chang-Bougi. Barcoded samples were purified using AMPure XP beads (Beckman Coulter Inc., Brea, CA, USA) at a ratio of 0.4:1 beads-to-barcoding mix, while keeping each barcoded sample in an independent Eppendorf tube. Finally, purified barcoded DNA samples were pooled at equimolar concentration to obtain a total volume of 30 µl. Adapters were ligated to the pooled samples. After purification, DNA sizes and concentrations in the barcoded pools were quantified using the Agilent 2200 TapeStation system and Qubit4® 1x dsDNA HS Assay Kit (Invitrogen, Carlsbad, CA, USA), respectively. Final libraries were adjusted to a volume of 32 µl containing 10–20 fmol of DNA. All incubations < 10 min were extended to 10 min.

We supplemented the libraries by adding the Sequencing Buffer (ONT, Oxford, UK) and Library Loading Beads (ONT, Oxford, UK), and subsequently loaded them into R10.4.1 PromethION flowcells for 96 h runs for Anso77 and Doublon, or 120 h runs for Chang-Bougi. Library reloading (washing flush) was performed in all of the experiments when the percentage of sequencing pores dropped to 10–15%. NAS was performed using PromethION flowcell channels 1-1500, while the remaining channels served as control. The sequencing speed was set at 260 bp (accuracy mode) for Anso77 and Doublon, and the quality score threshold was set at 10. For Chang-Bougi, NAS was performed on the whole flowcell and the sequencing speed was modified to 400 bp (default mode) because the 260 bp option was deprecated from MinKNOW version 23.04.

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

NAS data processing and enrichment calculation

For Anso77 and Doublon, we split the reads by channel, thereby generating two FASTQ files per sample: one with the reads sequenced on channels 1-1500 for NAS, and another with the reads generated on channels 1501–3000 for WGS. No splitting by channel was performed for Chang-Bougi as all of the channels were used for NAS. Reads with a "PASS" flag, i.e. their quality score was > 10, were retained for downstream analysis. We identified the NAS-rejected reads based on their "end reason" in the sequencing summary file. This file included the generated read classifications based on their "end reason". In this way, reads rejected by the adaptive sampling were labeled as "Data Service Unblock Mux Change". Reads labeled as "Unblock Mux Change", "Mux Change" and "Signal Negative" were also filtered out. Thereafter we filtered the generated FASTQ files by size, keeping reads > 1 kb. We assessed statistics on these FASTQ files using seqkit stats v. 2.4.0 [40].

We computed sequence depth statistics for Anso77 and Doublon by aligning the reads to their reference whole genome assemblies with minimap2 v. 2.24-r1122 [41] and using mosdepth v. 0.3.3 [42], with a bed file containing the coordinates of the 15 target regions. For Chang-Bougi, sequence depth statistics were calculated by aligning the reads to their assembled target regions. The split flowcell setup allowed us to calculate the extent of enrichment in Anso77 and Doublon by comparing the read depths generated in NAS and WGS. We assessed the NAS efficiency using the two following enrichment measurements.

Enrichment by yield, i.e. the ratio of the on-target sequence depth (NLR cluster + 20 kb flanking) with NAS to that with WGS, was assessed as follows:

$$Enrichment\ by\ yield = \frac{depth_{region\_NAS}}{depth_{region\_WGS}} \quad (1)$$

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

Enrichment by selection, i.e. the ratio of the relative selection of the target regions between NAS and WGS, measures the extent to which NAS can alter the abundance of the given target regions within a complete genome, while considering the sequencing behavior of each ROI. This was calculated as follows:

$$Enrichment\ by\ selection = \frac{depth_{region\_NAS}\Big/depth_{chr\_NAS}}{depth_{region\_WGS}\Big/depth_{chr\_WGS}} \quad (2)$$

where $depth_{region\_NAS}$ and $depth_{chr\_NAS}$ represent, respectively, the sequence depth on-target (NLR cluster + 20 kb flanking) and on the rest of the chromosome in the adaptive sampling approach, while $depth_{region\_WGS}$ and $depth_{chr\_WGS}$ represent the depth coverage on-target (NLR cluster + 20 kb flanking) and on the rest of the chromosome in the WGS approach. The relative selection with WGS should be equal to one if there is no bias that could cause the target regions to be differentially enriched compared to the rest of the genome.

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

Target regions assembly, NLRs annotation and quality control

We tested a set of assemblers tailored for ONT sequencing data, including Canu [43], Flye [44], Shasta [45], Necat [46], Raven [47] and SMARTdenovo [48], for assembling the NAS reads (data not shown). SMARTdenovo was primarily selected due to its superior target region assembly metrics (contiguity and assembly errors), achieved within a short time and with low memory usage. For Chang-Bougi, one target region was selected from the Canu assembly, as SMARTdenovo failed to collapse a repeat region generating two contigs instead of a single one. We used default parameters and added the "generate consensus" option for SMARTdenovo v. 2018.2.19. Canu v. 2.2 was executed with the

"genomesize = 7m − corrected − trimmed − nanopore" options, and using reads > 8 kb.

For each assembly, we filtered the contigs and only kept those that included at least one predicted NBS domain or matched more than 15 kb with at least 45% identity to any of the 15 target regions of Anso77. We assessed NBS domain prediction using NLGenomeSweeper. We used the nucmer and delta-filter commands of MUMmer v. 4.0.0rc1 [49] to select contigs matching the Anso77 target regions. Nucmer was used with the −l 100 option, keeping the rest of the parameters as default. Hits reported by nucmer were filtered with delta-filter with the -r -q -l 15,000 -i 45 options. We ran QUAST v. 5.0.2 [50] to assess the basic statistics of the generated filtered assemblies.

Assembly errors were analysed by focusing on the well-studied *Vat* region. We performed manual annotations of the *Vat* regions as described by Chovelon et al. [31]. The accuracy of the *Vat* homologs was analysed using two PCR markers: Z649 FR, which indicates the number of R65aa motifs in *Vat* homologs [31], and Z1431 FR which is specific to a *Vat* homolog with four R65aa motifs [32].
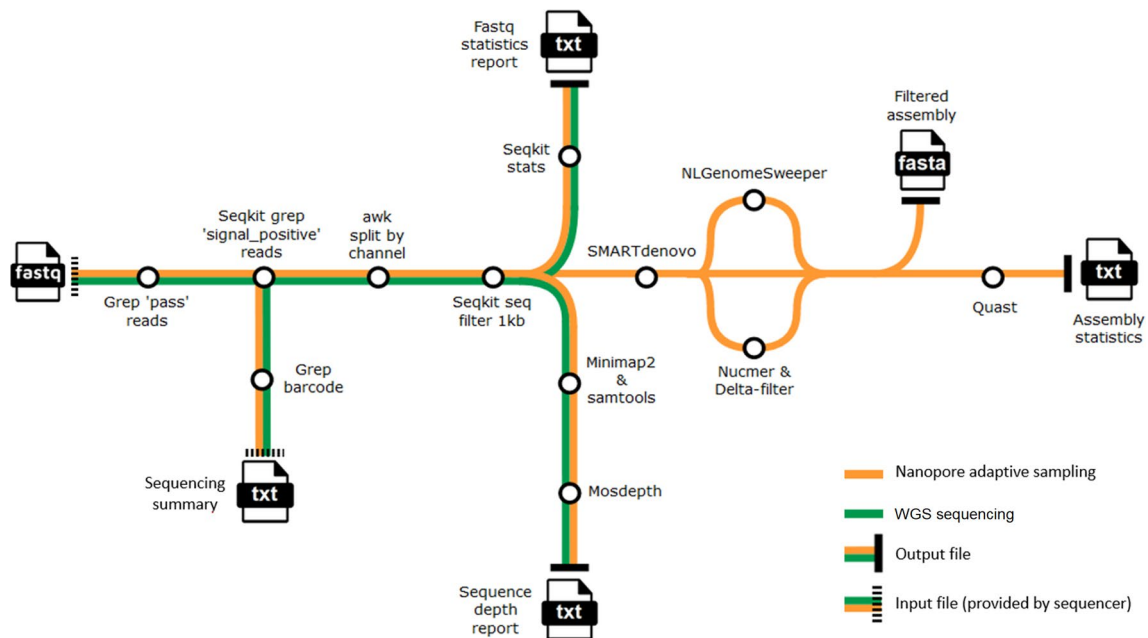
Figure 2. Workflow diagram summarizing the different steps involved in data processing and target regions assembly. For Chang-Bougi, only the NAS way was followed, and the split by channel step was omitted.

Results

Anso77 and Doublon de novo genome assemblies and annotation

The metrics of the different genome assembly steps are detailed in Additional file 1: Supplementary Data. Key metrics of the final assemblies are summarized in Table 1. In summary, initial contig assemblies provided 159 and 186 contigs for Anso77 and Doublon, with long N50 values of 8.9 and 15.2 Mb. Both genomes were organized in 12 chromosomes and presented a total size of 369.47 and 362.63 Mb. From this size, 12.84 and 12.92 Mb corresponded to unscaffolded contigs or unoriented scaffolds. BUSCO results showed that 98.5 and 94.4% of the single-copy orthologs were completely present in the Anso77 and Doublon assemblies. In addition, Anso77 and Doublon assemblies presented Merqury QVs of 31.21 and 42.52, proving their great quality and completeness.

We predicted 32,714 and 33,404 genes in the Anso77 and Doublon assemblies using the EuGene annotation software. Using Helixer, we predicted 21,692 genes for Anso77 and 21,125 for Doublon. As the EuGene results are more consistent with those previously observed in the literature in terms of number of predicted genes (Additional file 2: Table S1), we kept this annotation as a basis. However, we noticed that Helixer works better in predicting NLR genes structure, since it annotated more precisely the intron-exon structure of the previously

Fig. 2 shows a workflow diagram summarizing the different steps involved in the data processing and target region assembly process. All statistical analyses were performed using R v. 4.1.1 [51]

validated NLR genes (*Vat* homologs, *Fom-1* and *Fom-2*). For this reason, we selected the Helixer annotation for those genome regions where NLGenomeSweeper indicated the presence of NBS domains. In addition, we removed those ncRNA genes that did not include Rfam information. Therefore, our final annotation contained 31,152 genes for Anso77, and 31,865 genes for Doublon. From them, 29,714 and 30,198 were protein-coding genes. The average gene length was 3,398 bp for both accessions, with 4.46 and 4.29 exons/gene on average in Anso77 and Doublon, respectively. 28,627 genes were functionally annotated for Anso77, and 29,144 for Doublon. The gene models captured 95.2 and 91.6% of the BUSCOs for Anso77 and Doublon. Only 3.3 and 3% of the genes were found fragmented, and 1.5 and 5.4% were missing.

We predicted 84 and 76 NBS domains in the Anso77 and Doublon genomes, respectively, which was within the range of values previously reported for melon genomes (Additional file 2: Table S1). Based on Inter-ProScan [52] domain identification in the 10 kb flanking sequence on both sides of the NBS domain, potential genes containing the predicted NBS domains were classified in different categories (Table 1).

The accuracy of the NLR gene assemblies was assessed on the basis of the accuracy of the *Vat* homologs, whose cDNA sequence was previously obtained by Sanger sequencing [31]. For Anso77, the *AN-Vat2*, *AN-Vat3* and *AN-Vat5* homologs fully matched the assembly generated here, while *AN-Vat1* and *AN-Vat4* contained one SNP

each. For Doublon, all three *Vat* homologs fully matched our assembled sequence.

NAS target region construction

We arranged the 84 NLR predicted domains on Anso77 into 15 groups encompassing nine ROIs with 2–28 NBS domains and six ROIs with isolated NBS domains. We also found 15 groups and similar physical positions of the NLR genes on Doublon and on the previously published melon genomes. After adding the 20 kb flanking zones, the sizes of the 15 target regions ranged from ~41 to ~1,378 kb, representing a total length of ~6.16 Mb of the ~370 Mb Anso77 genome (~1.68%) (Table 2).

After masking the REs, the final file input in the PromethION sequencer included 935 target regions, ranging from 502 bp to 67.609 kb in size (Additional file 3). They accounted for a total of ~5.23 Mb of the ~370 Mb Anso77 genome (~1.41%).

NAS target region validation: effective enrichment of NLR clusters in melon

Over 8.62 million reads and 19.86 Gb (42.71%) were assigned to Anso77, and over 11.97 million reads and 26.63 Gb (57.27%) were assigned to Doublon after barcode trimming. We compared NAS to WGS of Anso77 and Doublon in terms of their general metrics. For Anso77, further read splitting by channel and filtering of this data by quality, "end reason" and length resulted in 110.06 K reads and 1.14 Gb derived from the NAS half-flowcell. The other half-flowcell (WGS) yielded 1.12 million reads, with a cumulative size of 11.93 Gb for Anso77.

Table 1 Summary metrics of the Anso77 and Doublon hybrid genome assemblies and annotations

		Anso77	Doublon
Chromo-some assembly	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
Unplaced	Number of scaffolds	6	6
	Number of contigs (> 15 Kb)	57	60
	Size of unplaced scaffolds (Mb)	9.1	9.4
	Size of unplaced contigs (Mb)	3.7	3.4
Whole genome	GC (%)	34	34
	Merqury QV	31.2	42.5
	BUSCO Complete genes	1,591 (98.5%)	1,524 (94.4%)
	BUSCO Duplicated genes	20 (1.2%)	16 (1.0%)
	BUSCO Fragmented genes	11 (0.7%)	12 (0.7%)
	BUSCO Missing genes	12 (0.8%)	78 (4.9%)
	Number of predicted genes	31,152	31,865
	Number of functionally annotated genes	28,627	29,144
	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

Table 2 Detailed information about the 15 target regions of Anso77

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

The *Vat* region is included within region 08

Using the same processing for Doublon, 163.84 K reads generating 1.56 Gb from were assigned to the NAS half-flowcell, while 15.81 Gb from 1.70 million reads were assigned to the WGS part. For both Anso77 and Doublon, the N50 value from the filtered NAS reads was very similar to that of the filtered WGS reads. All information on the generated datasets is presented in Table 3.

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

When focusing on Anso77, we evaluated the read depth on the target regions and on the rest of the chromosome in the NAS approach. The sequence depth on the target regions at the end of the experiment was much higher than on the rest of the chromosome

(Fig. 4), with an average frequency of target regions of 63.37. The sequence depth remained stable throughout the entire ROIs, even more so when the ROIs had a smaller size (Fig. 4A-C). The standard deviation of the sequence depth in the ROIs ranged from 1.38 for region 03 to 20.19 for region 07, with values of 10.29, 16.15 and 1.80 obtained for regions 01, 08 and 12, respectively. The increase in sequence depth was gradual on the 20 kb flanking regions, with the highest depth obtained in the ROIs. The increase had a similar pattern regardless of the ROIs size (Fig. 4D).

The half-flowcell design allowed us to calculate the enrichment obtained in NAS compared to the WGS approach. We obtained enrichment by yield for Anso77 that varied among the target regions, i.e. ranging from 2.45 to 5.18 at the end of the run (Fig. 5A). Otherwise, we obtained increased enrichment by selection ranging from 45.56 to 102.91 (Fig. 5B). On average, we obtained enrichment by yield and by selection of 3.96 and 78.38, respectively (Table 4). Target regions were sequenced at a lower rate than the rest of the genome in WGS, with a relative frequency of 0.81 (Table 4). We also noted that both enrichment by yield and by selection followed very similar temporal patterns. This enrichment peaked at the beginning of the run for most of the regions when most of the flowcell channels were actively sequencing, and then it decreased over time. This trend implies that channel inactivation occurred faster on the NAS half-flowcell.

Region 10 exhibited a very particular behavior, i.e. it was extremely enriched at the beginning of the run (Fig. 5A, B). As shown in Additional file 1: Figure S1A, this high enrichment during the first hours of the run corresponded to poor sequencing in the WGS approach.

Table 3 Anso77, Doublon and Chang-Bougi sequencing metrics

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

Anso77 and Doublon were sequenced using both NAS and WGS, while Chang-Bougi was only sequenced using NAS. Total WGS "pass" and total NAS "pass" represent the reads having a "PASS" flag (quality over 10) and assigned to the WGS and NAS half-flowcell, respectively. Filtered WGS and filtered NAS represent the two sets of reads just mentioned after filters of "end reason" and length
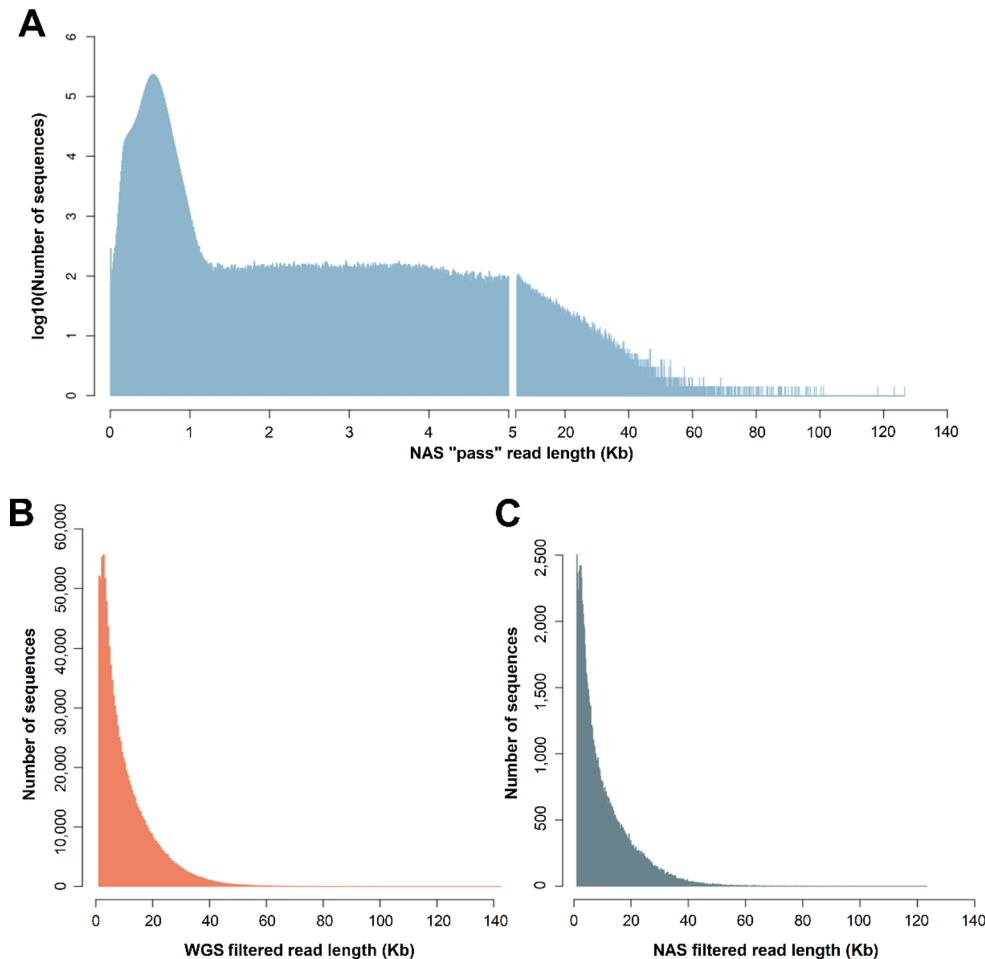
Fig. 3 Read length distribution of the different generated datasets. (**A**) Length distribution of "PASS"-tagged NAS reads. The number of reads was log-transformed. (**B**) Length distribution of WGS reads after filtering by "end reason", quality and length. (**C**) Length distribution of NAS reads after filtering by "end reason", quality and length
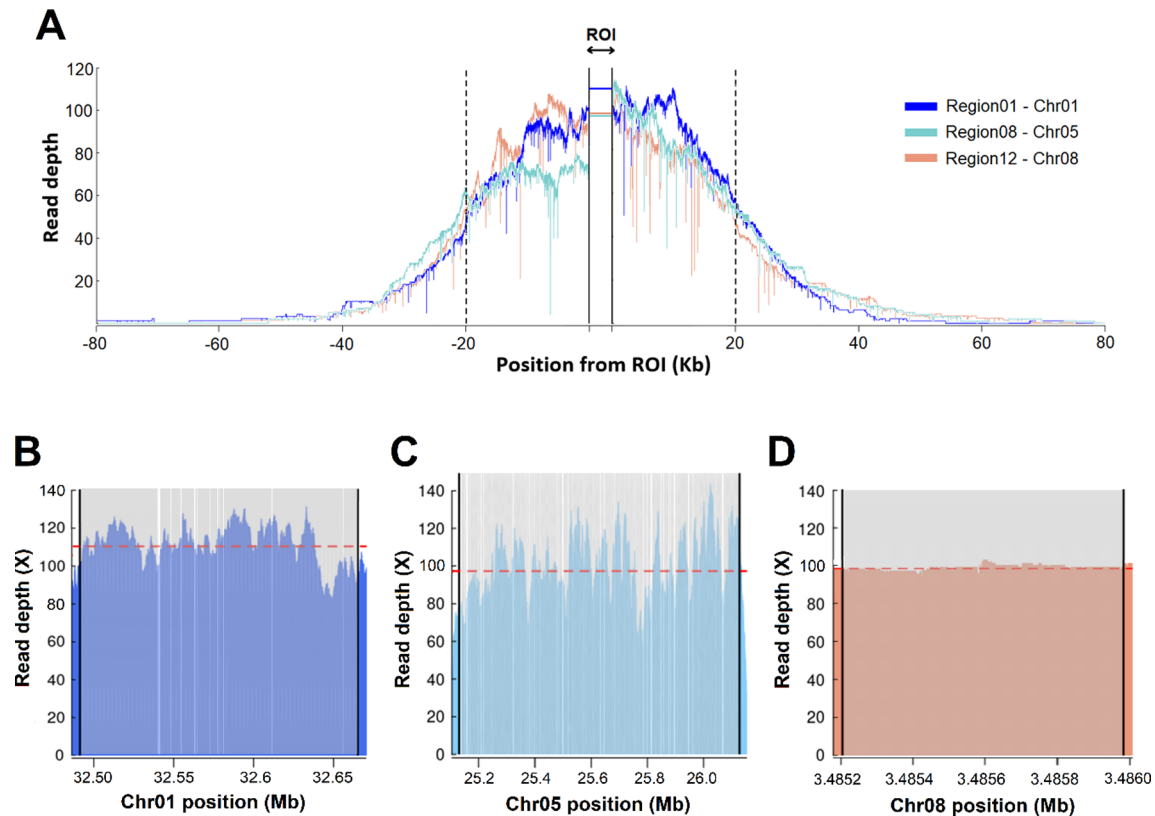
Fig. 4 NAS sequencing depth on three representative target regions of different sizes. (**A**) NAS sequencing depth on the three target regions compared to the rest of the chromosome. Target regions (ROI + 20 kb buffer) are represented between black dotted bars, while ROIs are collapsed and represented between black solid bars. (**B**) NAS sequencing depth on the ROI of region 01 of chromosome 1 (≈173 kb). (**C**) NAS sequencing depth on the ROI of region 08 of chromosome 5 (≈998 kb) (**D**) NAS sequencing depth on the ROI of region 12 of chromosome 8 (≈1 kb). Region 08 contains the well-studied *Vat* cluster. For B, C and D, vertical colored bars represent the enriched regions, while vertical white bars represent masked repetitive elements

Furthermore, within a period of just 10 h, the NAS approach provided a sequence depth comparable to that achieved in the entire WGS run (Additional file 1: Figure S1A). In addition, the washing flush contributed to an increase in pore activity in both the NAS and WGS approaches (Additional file 1: Figure S1A).

We sought to confirm the applicability of NAS in targeting the entire spectrum of NLR clusters in melon by extending its use to Doublon, i.e. a cultivar of the same subspecies as Anso77 but belonging to a distinct botanical group. Similar to Anso77, the NAS sequence depth on the target regions at the end of the run was always 2.25 to 4.65-fold greater than that obtained with the WGS approach (Fig. 5C, D). The least (region 06) and most (region 10) enriched regions were the same for both cultivars. Notably, the enrichment by yield presented a Kendall's coefficient of concordance (W) of 0.87 when comparing Anso77 and Doublon (p 0.04), suggesting region-specific patterns rather than cultivar-based differences.

Target regions of Doublon were sequenced and mapped with a ratio (0.98) identical to that of the rest of the genome in WGS (Table 4). Regarding enrichment by

selection in Doublon, we obtained values ranging from 43.52- to 83.89-fold, representing a significant correlation with the values previously obtained for Anso77 (W = 0.89; p 0.04). Overall, we demonstrated an average enrichment by yield of 3.73 for all the target regions and an average enrichment by selection of 69.92 (Table 4). In terms of the temporal enrichment patterns, the results were consistent with those obtained for Anso77 regarding both enrichment by yield and by selection (Fig. 5C, D).

Nanopore adaptive sampling facilitates the correct assembly of NLR clusters in Anso77 and Doublon genomes

NAS-enriched reads provided very contiguous and accurate assemblies of the target regions in both Anso77 and Doublon. Each cultivar presented a single contig per target region. The NAS assembly metrics are outlined in Table 5. Notably, the size of all contigs was over that of their corresponding target regions.

We generated dot plots for Anso77 and Doublon comparing each assembled contig with the corresponding region in the whole genome assemblies we produced (Additional file 1: Figure S2). There was perfect collinearity in the dot plot of each target region, thereby
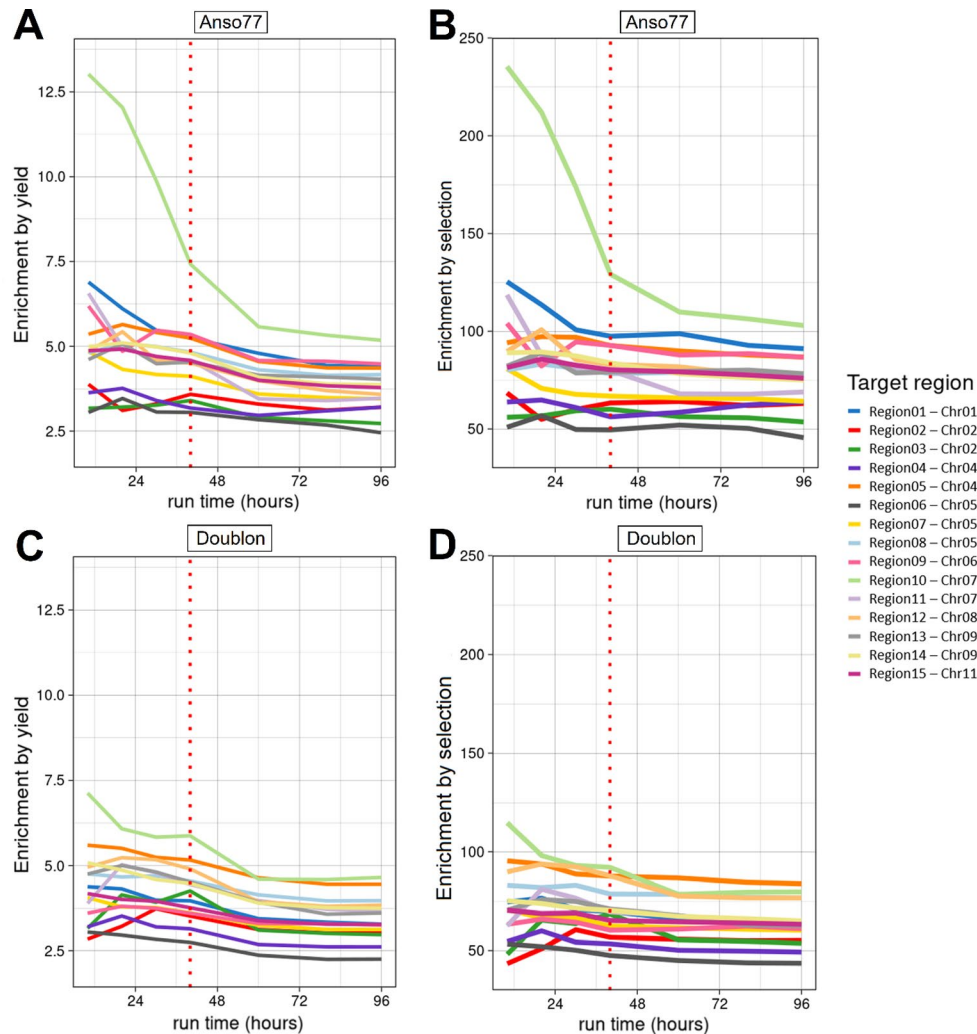
Fig. 5 Enrichment by yield (**A**, **C**) and enrichment by selection (**B**, **D**) of the 15 target regions from Anso77 (**A**, **B**) and Doublon (**C**, **D**). Vertical red-dotted bars denote the flowcell washing flush time

Table 4 Summary statistics of the NAS and WGS runs for Anso77 and Doublon

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

Only the chromosomes including target regions were considered for calculating the average off-target depth

confirming the accuracy of the NAS assemblies with regard to the reference genomes. Notably, the same number of NBS domains at the same positions were predicted in the NAS assembly compared to the reference genome for both cultivars.

We focused on the well-known *Vat* region to further assess the accuracy of the NAS assemblies. Dot plots representing the *Vat* regions are shown in Fig. 6. The dot plots highlighted the complexity of this area with numerous duplicated sequences, but a perfect diagonal was noted between the reference and NAS-assembled sequences. Moreover, we checked the sequence of the *Vat* homologs previously sequenced by Sanger sequencing (cDNA sequencing). Among the five Anso77 homologs, *AN-Vat1*, *AN-Vat2*, *AN-Vat3* and *AN-Vat5* fully matched the assembly generated from the NAS library, while *AN-Vat4* contained one SNP in the first exon (G/T on position 402710). This overcomes the reference assembly presented here, which contained one SNP in *AN-Vat4* as well as one SNP in *AN-Vat1*. For Doublon, the three *Vat* homologs presented 100% DNA sequence similarity with the NAS assembly. Overall, we demonstrated that NAS produced very contiguous and accurate assemblies in highly complex resistance gene clusters.

Table 5 Anso77 and Doublon NAS assembly metrics

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

Regions of the NAS reference file represent the size (bp) of the target regions provided to the sequencer. Regions of Anso77, Doublon and Chang-Bougi represent the size (bp) of the contigs matching those regions

Towards a benchmark procedure: enrichment and assembly of NLR clusters from a distant cultivar provide very valuable structural information

We obtained 3.32 million reads and 3.45 Gb for Chang-Bougi in a 1/10 flowcell. After eliminating the reads rejected by adaptive sampling and filtering by quality and 1 kb length, 96.62 K target reads and 0.80 Gb were available for further processing. These reads exhibited an N50 of 13.75 kb, i.e. comparable to that obtained for Anso77 and Doublon. Rejected reads had an average length of ~790.52 bp, which was longer than that obtained for

Anso77 and Doublon due to the updated sequencing speed (400 bp/s) in the last MinKNOW version.

The sequence depth mapped on the assembled contigs of the target regions averaged 41.82X. As illustrated in Fig. 7, this depth varied between regions, with the highest depth obtained in region 05 (50.37X) and the lowest in region 03 (21.08X). However, the sequence depth obtained between regions kept a significant concordance with that obtained for Anso77 and Doublon (Additional file 1: Figure S1-left) (W = 0.78; $p = 0.003$).

For Chang-Bougi, the NAS-enriched read assemblies obtained with SMARTdenovo resulted in 17 contigs. Notably, we observed an inversion of ≈100 kb in region 01 compared to Anso77 (Fig. 8A). Region 13 and region 15 were fragmented into two contigs. Among the tested assemblers, Canu generated a contiguous assembly of region 13 in a single contig (Additional file 1: Figure S3). Consequently, we retained region 13 from the Canu assembly. No assembler succeeded in reconstructing region 15 into a single contig. We then investigated why region 15 was fragmented into two contigs, but we were unable to conclude after contig alignment to the published Illumina-based assembly due to its high degree of fragmentation. As Chang-Bougi belongs to the *makuwa* botanical group, we mapped the two contigs to the genomes in this group for which open-access data were available: Early Silver Line, Ohgon and Sakata's Sweet [53]. We identified a very large and size-conserved insertion (ranging from 862 to 871 kb) at the breakpoint between the two contigs obtained for Chang-Bougi (Fig. 8A). No NBS domain was predicted in this insertion on the Early Silver Line, Ohgon and Sakata's Sweet genomes. We could not recover this insertion in Chang-Bougi as it was not present in the provided reference. The total assembly size was 6.68 Mb. Table 4 shows the detailed NAS assembly metrics.
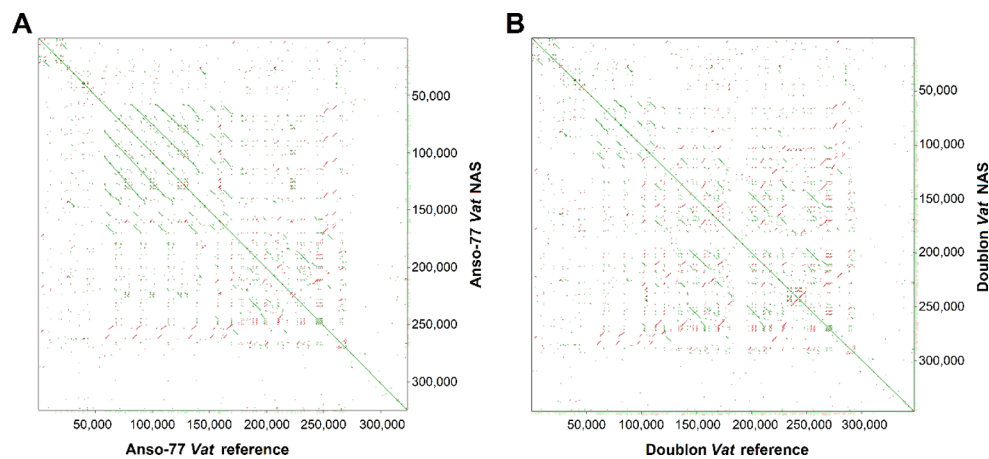
Fig. 6 Dot plots representing the *Vat* region for Anso77 (**A**) and Doublon (**B**). Reference *Vat* region is represented on the x-axis, while the NAS-reconstructed *Vat* region is represented on the y-axis
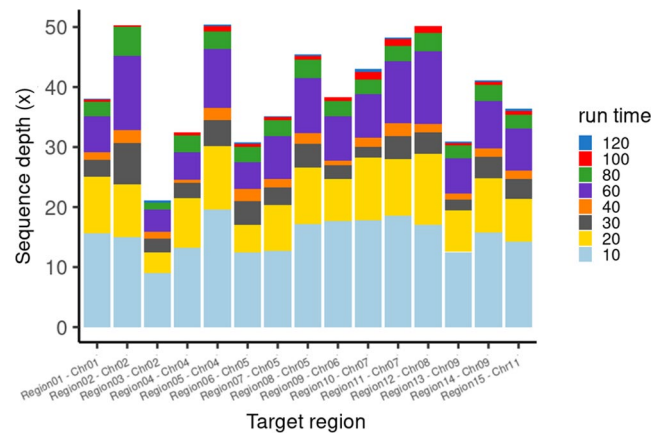
Fig. 7 Sequence depth by time in the NAS experience of the 15 target regions from Chang-Bougi

In the Chang-Bougi draft genome generated by Shin et al. [33], we identified 81 NBS domains spanning 18 contigs. These contigs matched the previously identified 15 ROIs of the NAS assembly with no extra clusters (Fig. 8B). We identified 83 NBS domains in the NAS assembly. The two extra NBS domains predicted in the NAS assembly were located in region 08, one in the *Vat* region and the other outside. We manually annotated the *Vat* region of both Chang-Bougi assemblies (NAS and published) and some discrepancies were detected in the complex and repetitive area between *Vat1* and *VatRev* (Fig. 9A). In the NAS assembly, we identified the extra NBS domain within the *Vat* region as being a *Vat* homolog with four R65aa motifs (Fig. 9A). We confirmed the presence and structure of this *Vat* gene with four R65aa motifs, as well as the presence of *Vat* genes with three R65aa through PCR using the published Z649FR and Z1431FR primers (Fig. 9B). Finally, the presence of long reads encompassing the *Vat1:Vat2* and *Vat2:Vat3* gene pairs confirmed the NAS-assembled structure.

Discussion

The findings of the present study highlighted that NAS is a promising approach for studying polymorphism in complex genomic ROIs. We used melon as a model to select highly diverse ROIs in terms of size and NLR gene content. We generated two *de novo* assemblies and used one of them as reference for the adaptive sampling experiment. We selected the two accessions based on their different responses to the most studied pathogens at INRAE GAFL. Both *de novo* assemblies presented great quality and completeness, comparable to the best published melon assemblies to date [54]. Similarly, predicted gene content values were in the range of previously melon published assemblies (Additional file 2: Table S1).

Several factors can influence the NAS efficiency, two of which we took into account prior to launching the NAS experiments. First, we hypothesized that there is a

direct relationship between the ideal size of sequenced fragments and the ROI size (Fig. 10). Given that here the ROI sizes ranged from ≈41 to ≈1,378 kb, we used standard DNA extraction (10–30 kb), which we expected would lead to a more stable sequencing depth in ROIs than ultra-long reads (100–300 kb) for the same yield. We felt that this approach would reduce off-target sequencing and avoid channel blockage when rejecting very long reads, as outlined in the ONT recommendations [30]. Second, REs span a major portion of the melon genome [55], and are especially common in NLR gene clusters [31]. To avoid sequencing off-target REs with high sequence similarity to those within the initial target regions, we assumed that masking repetitive elements within the provided target regions would reduce the quantity of off-target data, thereby enhancing enrichment. We masked 0.93 Mb of repetitive sequences over the entire 6.16 Mb target length. Masking REs in genomes prior to NAS has been previously suggested [56].

We assessed the efficacy of NAS compared to WGS. One key factor that may have a major effect on the final yield in ONT sequencing runs is the number of active channels at the beginning of the run and their lifespan, which may differ markedly between flowcells. Therefore we implemented a half-flowcell design to compare NAS and WGS with respect to eliminating biases that may arise when using two different flowcells, as previously revealed in many studies [16, 23]. The results showed that for both Anso77 and Doublon, NAS produced about fourfold more on-target data than WGS (Table 4), while generating about tenfold less total data (see filtered NAS and WGS in Table 3). Yet the sequence depth between target regions was more variable in NAS than in WGS, but this did not compromise accurate ROI assembly in all the cultivars. We found no correlation between target size and sequence depth (Additional file 1: Figure S4A) but there was a moderate correlation between
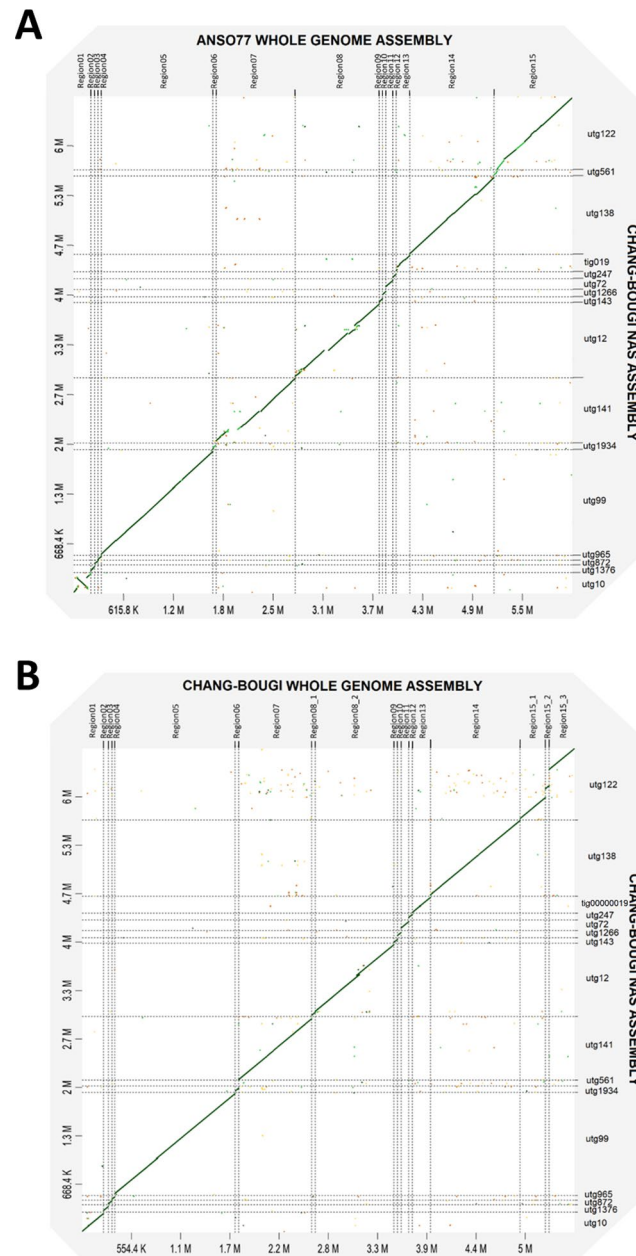
Fig. 8 Dot plots representing the NAS filtered assembly of Chang-Bougi (y-axis) against two sequences. (**A**) NAS filtered assembly of Chang-Bougi (y-axis) against the 15 target regions from Anso77 (x-axis). (**B**) NAS filtered assembly of Chang-Bougi (y-axis) against the 18 NLR clusters identified in the Chang-Bougi assembly published by Shin et al. [33]

the percentage of masking and the sequence depth, as expected (Additional file 1: Figure S4B).

We proposed two enrichment measurements adapted from previous studies in which NAS was tested on metagenomics samples or panels of many small genomic regions [16, 57]: enrichment by yield, a simple widespread metric; and enrichment by selection, a metric that is not biased by the sequencing behavior of each target region. These enrichment measures made sense in our study because our goal was to increase the coverage in complex ROIs to generate more accurate assemblies.

We achieved up to 3.7-fold average enrichment by yield and up to 69-fold average enrichment by selection, even when the reference was genetically distant from the sequenced accession. These findings were comparable to the best results obtained in previous studies involving the enrichment of individuals in metagenomics samples [16, 23], and they were better than the enrichment values previously obtained with loci panels [57]. Our successful enrichment with NAS could have been linked to the percentage of the genome targeted here (~1.41%), the target sizes or the DNA fragment sizes, which have
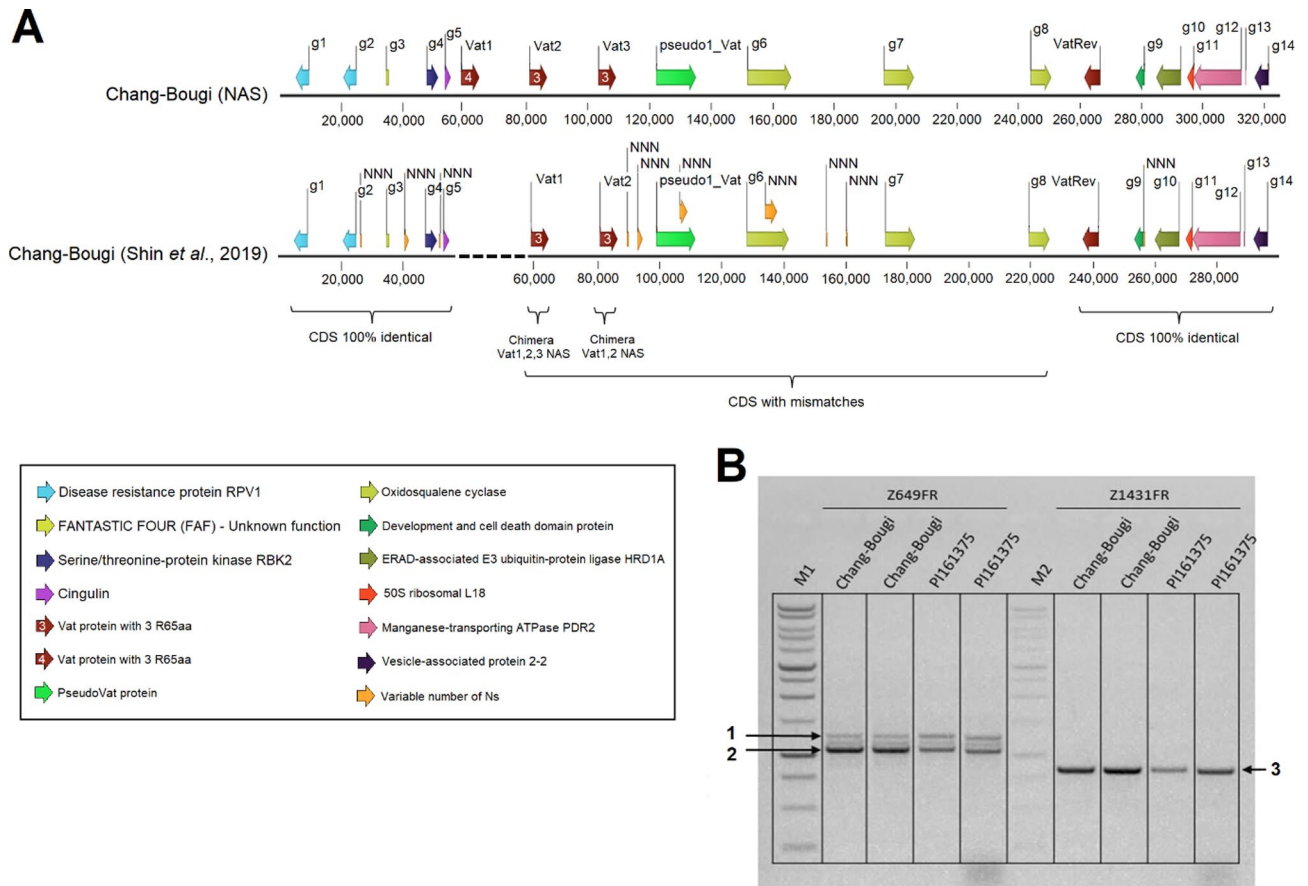
Fig. 9 Manual annotation and validation of the *Vat* region of Chang-Bougi. (**A**) Genes identified after manual annotation within the *Vat* regions of Chang-Bougi. The sequence above was obtained from the NAS assembly, while the sequence below was recovered from the publicly available genome assembly [33]. (**B**) Agarose gel electrophoresis of PCR products obtained using primers Z649FR and Z1431FR. Lanes M1 and M2 are the two 1 kb DNA ladders (Promega, Madison, WI, USA). PI161375 was used as a control having a *Vat1* with four R65aa motifs and a *Vat2* with three R65aa motifs. Bands pointed with arrows represent an amplicon of four R65aa motifs (1), an amplicon of three R65aa motifs (2), and a specific amplicon of four R65aa motifs (3)
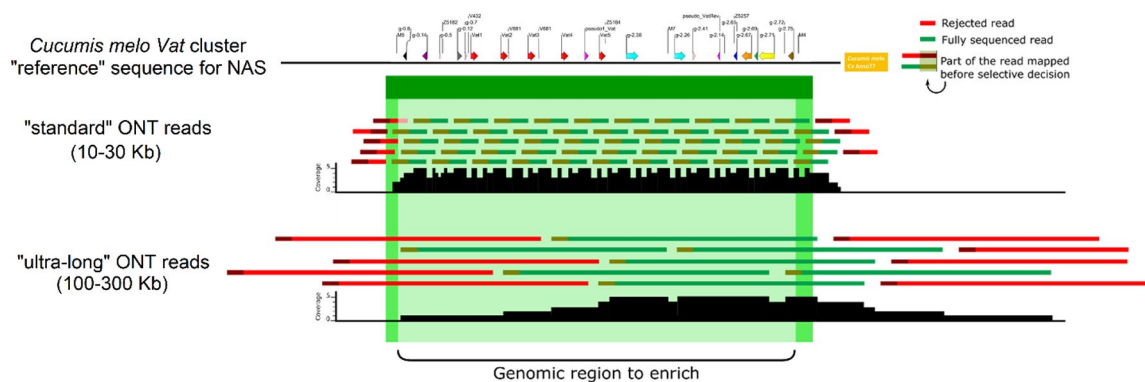
Fig. 10 Comparison of the use of standard and ultra-long reads in the enrichment of the *Vat* cluster. The diagram illustrates the difference in coverage and extent of the area outside the region to be enriched for standard size fragments (10-30 kb) and for ultra-high molecular weight fragments (100-300 kb) on the *Vat* (melon) cluster. For the same yield (illustrated here by an arbitrary overall depth of 5X), standard fragments make it possible to achieve a depth more concentrated on the area to be enriched and to sequence less outside this area. For convenience, only reads oriented from 5' to 3' are represented

been demonstrated to be key factors in determining the enrichment rate [16, 30]. In addition, the fact that we performed a late nuclease flush when the percentage of sequencing pores was around 10% rather than at a fixed time might have contributed to the good performance [16, 58, 59]. The temporal enrichment patterns of the different target regions were higher and more variable at the beginning of the run but then generally stabilized after 70 h (Fig. 5). Actually, channel inactivation occurred faster on the NAS half-flowcell, which could have been due to the repetitive potential flipping to reject off-target sequences or simply because the likelihood of channel clogging is statistically related to the number of sequenced molecules [13, 16].

In previous studies, NAS has been used to enrich specific species in metagenomics samples [16, 18, 60] and relatively small sequences within an organism, e.g. panels of exons or of key variant loci [57, 59, 61]. Here we demonstrated the power of NAS as a tool for enriching ROIs representing isolated NLR genes or complex NLR gene clusters in a plant crop species. The correct assembly of these complex regions typically requires long reads, such as those generated by ONT sequencing, or dedicated laborious approaches such as the resistance gene enrichment sequencing (RenSeq) method [6, 62–65]. In fact, NLR genes may sometimes be miss-predicted, especially when short-read sequencing technologies are used. Two to four functional NLR genes and some pseudo NRLs were found when deciphering the *Vat* region in the DHL92 genome [31, 66]. When screening for these NLRs in the early released DHL92 genome, they were found to be misassembled. It was only when a high-quality genome (with long reads, optical maps or HiC) was released [55], that the *Vat* gene assemblies were finally in line with those obtained via Sanger sequencing using long-range PCR [31]. Moreover, we compared the *Vat* cluster in the Chang-Bougi cultivar derived from short WGS [33] and long NAS reads. We showed that the WGS assembly was erroneous in terms of the homologous gene numbers and sequences (Fig. 9) and that the NAS assembly accurately reconstructed the region.

NLR genes are located in the dispensable portion of the genome [26, 67], and therefore the NLR reference used for NAS should be carefully selected when targeting the NLRome of a species. Our predictions of the number of NBS domains in Anso77 and Doublon, alongside all previously published melon genome assemblies, consistently yielded similar values (Additional file 2: Table S1). In all cases, we did not detect more than 15 groups of NLR genes regardless of the subspecies assessed, i.e. *melo* or *agrestis*. These findings indicated a well-conserved number and location of NLR genes in melon. We chose Anso77—a Spanish cultivar belonging to ssp. *melo* and the *inodorus* botanical group—as reference for the

NAS approach because it contained the highest number of *Vat* homologs within the *Vat* region used for benchmarking [31]. Our results obtained with Doublon and Chang-Bougi suggested that our strategy was suitable. Doublon is a French melon line belonging to ssp. *melo*, and the *cantalupensis* botanical group. When NAS was used without any short-read polishing, we obtained a *Vat* cluster with CDSs identical to those derived from an assembly using HW-DNA, PacBio and ONT long sequences, Illumina short sequences and optical maps. Chang-Bougi is a Korean melon line belonging to ssp. *agrestis* and the *makuwa* botanical group. The *Vat* cluster we obtained using NAS was highly consistent with that of PI 161,375 [31], a Korean line belonging to ssp. *agrestis*. However, a limitation was noted regarding very large SVs not present in the reference, i.e. that identified in Chang-Bougi chromosome 11 which turned out to belong to the oriental melon clade. Using different high-quality reference genomes or even combining them in an "artificial" reference genome could address this shortcoming. Existing software packages such as BOSS-RUNS [15] already enable dynamic updating of decision strategies during a run, thereby enhancing the good balance of the target regions or multiplexed sample sequencing depth. Yet it is unclear whether NAS would be able to discover extra NLR gene clusters if they were to exist. This should be possible if the additional NBS domains are sufficiently conserved to match the provided reference.

Overall, the findings of our study provide a blueprint for the selective capture of NLRome in melon and it could be extended to other key crop species. The nucleotide variation observed in the NLR regions is biologically significant as it directly impacts the diversity of resistance genes involved in plant immunity. Variations in these regions, such as mutations, duplications, and copy number variations, contribute to the functional diversity of NLR receptors, enabling plants to recognize a wide array of pathogen effectors [26]. Understanding this variation is crucial for identifying novel resistance genes, which could be used in breeding programs to enhance disease and pest resistance [68]. Furthermore, exploring the biological significance of this variation provides insights into the evolutionary mechanisms that shape plant immune responses [69].

NLR gene numbers are generally low in the Cucurbitaceae family [26, 70], and they represent an ideal percentage of the genome to be targeted using NAS. However, this ideal situation does not correspond to reality in other species, as the number of NLR genes is highly variable between plant species independently of their genome size [26]. To adapt the NAS procedure implemented here for NLR-rich plant species, certain adjustments should be made to align with the ideal targeted percentage of the genome. First, a reduction in the length of flanking

regions surrounding ROIs is recommended. Subsequently, a more rigorous definition of NLR clusters would help reduce the targeted percentage of the genome. Finally, strict NAS targeting of clustered NLRs could be done, recovering the easiest-to-assemble isolated NLRs with the low-pass reads rejected by NAS.

Conclusion

NAS offered flexible real-time enrichment of selected NLR gene clusters while reducing costs as compared to the WGS approach. This target enrichment did not require any laborious or expensive library preparation, nor probe design and synthesis, unlike previously developed target sequencing methods. This is particularly advantageous for researchers who may not have access to special molecular biology techniques or who seek to conduct in-field experiments. NAS only requires an ONT sequencing device (e.g. the low-cost MinION device), a reference genome, and one or several ROIs. In addition, the fast enrichment obtained here may be of marked interest when time is a critical factor. Moreover, we highlighted the ability of NAS to reduce the high off-target data volume generated by WGS, addressing the growing challenges of data management and storage in the field of bioinformatics and genomics. This method, which we validated here on three melon cultivars, shows promise for application when dealing with a large number of accessions. This is particularly relevant for breeding and pre-breeding applications, including the identification of novel resistance genes and the creation of varieties including these genes. Pre-breeding programs aim to enhance the diversity and spectrum of resistance to pests and pathogens within the genetic pools used for breeding. A thorough understanding of NLRs can facilitate the creation of core collections of reduced size for each species, while maximizing the diversity of NLRs. These reduced collections can then be biotested against a broader range of pathogens and pests than current methods allow, which is especially critical given the current challenges of high-speed phenotyping for resistance. Furthermore, knowledge of NLRs and their flanking genetic context simplifies the development of specific markers for each effective NLR homolog. Tapping into the NLRome diversity opens opportunities for discovering previously uncharacterized resistance genes. Altogether, this approach could pave the way for developing multiresistant crop varieties, ultimately leading to a significant reduction in pesticide use.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11295-5.

Additional file 1

Additional file 2

Additional file 3

Additional file 4

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

References

1. Lee RRQ, Chae E. Variation patterns of NLR clusters in Arabidopsis thaliana genomes. Plant Commun. 2020;1(4):100089.
2. Mohamed M, Dang NTM, Ogyama Y, Burlet N, Mugat B, Boulesteix M, et al. A transposon story: from TE content to TE dynamic invasion of Drosophila genomes using the single-molecule sequencing technology from Oxford Nanopore. Cells. 2020;9(8):1776.
3. Lieberman NAP, Armstrong TD, Chung B, Pfalmer D, Hennelly CM, Haynes A, et al. High-throughput nanopore sequencing of Treponema pallidum tandem repeat genes arp and tp0470 reveals clade-specific patterns and recapitulates global whole genome phylogeny. Front Microbiol. 2022;13:1007056.
4. Warburton PE, Sebra RP. Long-read DNA sequencing: recent advances and remaining challenges. Annu Rev Genomics Hum Genet. 2023;24(1):109–32.
5. Hook PW, Timp W. Beyond assembly: the increasing flexibility of single-molecule sequencing technology. Nat Rev Genet. 2023;24(9):627–41.
6. Witek K, Jupe F, Witek AI, Baker D, Clark MD, Jones JD. Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing. Nat Biotechnol. 2016;34(6):656–60.
7. Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W. Nanopore sequencing detects structural variants in cancer. Cancer Biol Ther. 2016;17(3):246–53.
8. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat Biotechnol. 2020;38(4):433–38.
9. Madsen EB, Höijer I, Kvist T, Ameur A, Mikkelsen MJ, Xdrop. Targeted sequencing of long DNA molecules from low input samples using droplet sorting. Hum Mutat. 2020;41(9):1671–9.
10. Gabrieli T, Sharim H, Michaeli Y, Ebenstein Y. Cas9-Assisted Targeting of CHromosome segments (CATCH) for targeted nanopore sequencing and optical genome mapping. Preprint at https://www.biorxiv.org/content/https://doi.org/10.1101/110163v3 (2017).
11. Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. Nat Methods. 2016;13(9):751–54.
12. Edwards HS, Krishnakumar R, Sinha A, Bird SW, Patel KD, Bartsch MS. Real-time selective sequencing with RUBRIC: read until with basecall and reference-informed Criteria. Sci Rep. 2019;9(1):11475.
13. Kovaka S, Fan Y, Ni B, Timp W, Schatz MC. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. Nat Biotechnol. 2021;39(4):431–41.
14. Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. Nat Biotechnol. 2021;39(4):442–50.
15. Weilguny L, De Maio N, Munro R, Manser C, Birney E, Loose M, et al. Dynamic, adaptive sampling during nanopore sequencing using bayesian experimental design. Nat Biotechnol. 2023;41:1018–25.
16. Martin S, Heavens D, Lan Y, Horsfield S, Clark MD, Leggett RM. Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. Genome Biol. 2022;23(1):11.
17. Miyatake S, Koshimizu E, Fujita A, Doi H, Okubo M, Wada T, et al. Rapid and comprehensive diagnostic method for repeat expansion diseases using nanopore sequencing. Npj Genomic Med. 2022;7(1):62.
18. Kipp EJ, Armstrong T, Faulk C, Oliver J, Larsen P, et al. Metagenomic surveillance for bacterial tick-borne pathogens using nanopore adaptive sampling. Sci Rep. 2023;13(1):10991.
19. Greer SU, Botello J, Hongo D, Levy B, Shah P, Rabinowitz M, et al. Implementation of Nanopore sequencing as a pragmatic workflow for copy number variant confirmation in the clinic. J Transl Med. 2023;21(1):378.
20. Hewel C, Schmidt H, Runkel S, Kohnen W, Schweiger-Seemann S, Michel A et al. Nanopore adaptive sampling of a metagenomic sample derived from a human monkeypox case. J Med Virol. 2024;96(5).
21. Su J, Lui WW, Lee Y, Zheng Z, Siu GK, Ng TT, et al. Evaluation of Mycobacterium tuberculosis enrichment in metagenomic samples using ONT adaptive sequencing and amplicon sequencing for identification and variant calling. Sci Rep. 2023;13(1):5237.
22. Wrenn DC, Drown DM. Nanopore adaptive sampling enriches for antimicrobial resistance genes in microbial communities. GigaByte. 2023. https://doi.org/10.1101/2023.06.27.546783.
23. De Meulenaere K, Cuypers WL, Gauglitz JM, Guetens P, Rosanas-Urgell A, Laukens K, Cuypers B. Selective whole-genome sequencing of Plasmodium parasites directly from blood samples by nanopore adaptive sampling. mBio. 2024. https://doi.org/10.1128/mbio.01967-23.
24. Stevanovski I, Chintalaphani SR, Gamaarachchi H, Ferguson JM, Pineda SS, Scriba CK, et al. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. Sci Adv. 2022;8(9):eabm5386.
25. Liu J, Liu X, Dai L, Wang G. Recent progress in elucidating the structure, function and evolution of disease resistance genes in plants. J Genet Genomics. 2007;34(9):765–76.
26. Barragan AC, Weigel D. Plant NLR diversity: the known unknowns of pan-NLRomes. Plant Cell. 2021;33(4):814–31. https://doi.org/10.1093/plcell/koaa002.
27. Zhang W, Yuan Q, Wu Y, Zhang J, Nie J. Genome-wide identification and characterization of the CC-NBS-LRR gene family in cucumber (Cucumis sativus L). Int J Mol Sci. 2022;23(9):5048.
28. Van Wersch S, Li X. Stronger when together: clustering of plant NLR disease resistance genes. Trends Plant Sci. 2019;24(8):688–99.
29. González VM, Aventín N, Centeno E, Puigdomènech P. High presence/absence gene variability in defense-related gene clusters of Cucumis melo. BMC Genomics. 2013;14(1):782.
30. Nanopore Community. Adaptive sampling methodology best practices. 2020. https://community.nanoporetech.com/docs/plan/best_practice/adaptive-sampling/v/ads_s1016_v1_revi_12nov2020. Accessed 20 December 2023.
31. Chovelon V, Feriche-Linares R, Barreau G, Chadoeuf J, Callot C, Gautier V, et al. Building a cluster of NLR genes conferring resistance to pests and pathogens: the story of the vat gene cluster in cucurbits. Hortic Res. 2021;8:72.
32. Boissot N, Chovelon V, Rittener-Ruff V, Giovinazzo N, Mistral P, Pitrat M, et al. A highly diversified NLR cluster in melon contains homologs that confer powdery mildew and aphid resistance. Hortic Res. 2023;11(1):uhad256.
33. Shin AY, Koo N, Kim S, Sim YM, Choi D, Kim YM, et al. Draft genome sequences of two oriental melons, Cucumis melo L. var. Makuwa Sci Data. 2019;6(1):220.
34. Salinier J, Lefebvre V, Besombes D, Burck H, Causse MC, Daunay M-C, et al. The INRAE Centre for Vegetable Germplasm: geographically and phenotypically diverse collections and their use in Genetics and Plant breeding. Plants. 2022;11(3):347.
35. Sallet E, Gouzy J, Schiex T. EuGene: an automated integrative gene finder for eukaryotes and prokaryotes. In: Gene prediction: Methods and protocols. 2019;97–120.
36. Holst F, Bolger A, Günther C, Maß J, Triesch S, Kindel F et al. Helixer–de novo Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model. Preprint at https://www.biorxiv.org/content/https://doi.org/10.1101/2023.02.06.527280v2.abstract (2023).
37. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021;38(12):5825–9.
38. Toda N, Rustenholz C, Baud A, Le Paslier MC, Amselem J, Merdinoglu D, et al. NLGenomeSweeper: a tool for genome-wide NBS-LRR resistance gene identification. Genes. 2020;11(3):333.
39. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in repbase: RepbaseSubmitter and Censor. BMC Bioinformatics. 2006;7(1):474.
40. Shen W, Le S, Li Y, Hu F, SeqKit. A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11(10):e0163962.
41. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
42. Pedersen BS, Quinlan AR, Mosdepth. Quick coverage calculation for genomes and exomes. Bioinformatics. 2018;34(5):867–8.
43. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM, Canu. Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.
44. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–46.
45. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol. 2020;38(9):1044–53.
46. Chen Y, Nie F, Xie S-Q, Zheng Y-F, Dai Q, Bray T, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat Commun. 2021;12(1):60.

47. Vaser R, Šikić M. Time- and memory-efficient genome assembly with Raven. Nat Comput Sci. 2021;1(5):332–36.

48. Liu H, Wu S, Li A, Ruan J, SMARTdenovo. A de novo assembler using long noisy reads. GigaByte. 2021. https://doi.org/10.46471/gigabyte.15.

49. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol. 2018;14(1):e1005944.

50. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.

51. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. 2021. Available from: https://www.R-project.org/

52. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37 Suppl 1.

53. Oren E, Dafna A, Tzuri G, Halperin I, Isaacson T, Elkabetz M, et al. Pan-genome and multi-parental framework for high-resolution trait dissection in melon (*Cucumis melo*). Plant J. 2022;112(6):1525–42.

54. Wei M, Huang Y, Mo C, Wang H, Zeng Q, Yang W, et al. Telomere-to-telomere genome assembly of melon (Cucumis melo L. var. Inodorus) provides a high-quality reference for meta-QTL analysis of important traits. Horticult Res. 2023;10(10):uhad189.

55. Castanera R, Ruggieri V, Pujol M, Garcia-Mas J, Casacuberta JM. An improved melon reference genome with single-molecule sequencing uncovers a recent burst of transposable elements with potential impact on genes. Front Plant Sci. 2020;10.

56. Zhang H, Li H, Jain C, Cheng H, Au KF, Li H, et al. Real-time mapping of nanopore raw signals. Bioinformatics. 2021;37(Suppl 1):i483.

57. Hogers R, Wittenberg A, Roelofs D. Adaptive sequencing in crop species. 2020. https://www.keygene.com/wp-content/uploads/2020/06/white-paper-read-until-plants-at-keygene.pdf

58. Payne A, Munro R, Holmes N, Moore C, Carlile M, Loose M. Barcode aware adaptive sampling for GridION and PromethION Oxford Nanopore sequencers. Preprint at https://www.biorxiv.org/content/https://doi.org/10.1101/2021.12.01.470722v2.abstract (2021).

59. Nakamura W, Hirata M, Oda S, Chiba K, Okada A, Mateos RN et al. A comprehensive workflow for target adaptive sampling long-read sequencing applied to hereditary cancer patient genomes. Preprint at https://www.medrxiv.org/content/https://doi.org/10.1101/2023.05.30.23289318v1 (2023).

60. Ulrich JU, Lutfi A, Rutzen K, Renard BY. ReadBouncer: precise and scalable adaptive sampling for nanopore sequencing. Bioinformatics. 2022;38(Suppl 1):i160.

61. Filser M, Schwartz M, Merchadou K, Hamza A, Villy M-C, Decees A, et al. Adaptive nanopore sequencing to determine pathogenicity of BRCA1 exonic duplication. J Med Genet. 2023;60(12):1206–9.

62. Van de Weyer AL, Monteiro F, Furzer OJ, Nishimura MT, Cevik V, Witek K, et al. A species-wide inventory of NLR genes and alleles in Arabidopsis thaliana. Cell. 2019;178(5):1260–72.

63. Huang Z, Qiao F, Yang B, Liu J, Liu Y, Wulff BBH, et al. Genome-wide identification of the NLR gene family in Haynaldia villosa by SMRT-RenSeq. BMC Genomics. 2022;23(1):118.

64. Vendelbo NM, Mahmood K, Steuernagel B, Wulff BB, Sarup P, Hovmøller MS, et al. Discovery of resistance genes in rye by targeted long-read sequencing and association genetics. Cells. 2022;11(8):1273.

65. Adams TM, Smith M, Wang Y, Brown LH, Bayer MM, Hein I. HISS: Snakemake-based workflows for performing SMRT-RenSeq assembly, AgRenSeq and dRenSeq for the discovery of novel plant disease resistance genes. BMC Bioinformatics. 2023;24(1):204.

66. Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, González VM, et al. The genome of melon (Cucumis melo L). Proc Natl Acad Sci U S A. 2012;109(29):11872–7.

67. Shang L, Li X, He H, Yuan Q, Song Y, Wei Z, et al. A super pan-genomic landscape of rice. Cell Res. 2022;32(10):878–96.

68. Boissot N. NLRs are highly relevant resistance genes for aphid pests. Curr Opin Insect Sci. 2023;56:101008.

69. Liu Y, Zhang YM, Tang Y, Chen JQ, Shao ZQ. The evolution of plant NLR immune receptors and downstream signal components. Curr Opin Plant Biol. 2023;73:102363.

70. Baggs E, Dagdas G, Krasileva K. NLR diversity, helpers and integrated domains: making sense of the NLR IDentity. Curr Opin Plant Biol. 2017;38:59–67. https://doi.org/10.1016/j.pbi.2017.04.012.

Publisher's note