BMC Genomics

RESEARCH ARTICLE Open Access

Comparative genomics of two inbred lines of the potato cyst nematode *Globodera rostochiensis* reveals disparate effector family-specific diversification patterns

Joris J.M. van Steenbrugge[1*], Sven van den Elsen[1], Martijn Holterman[1,2], Mark G. Sterken[1], Peter Thorpe[3], Aska Goverse[1], Geert Smant[1] and Johannes Helder[1]

Abstract

Background: Potato cyst nematodes belong to the most harmful pathogens in potato, and durable management of these parasites largely depends on host-plant resistances. These resistances are pathotype specific. The current *Globodera rostochiensis* pathotype scheme that defines five pathotypes (Ro1 - Ro5) is both fundamentally and practically of limited value. Hence, resistant potato varieties are used worldwide in a poorly informed manner.

Results: We generated two novel reference genomes of *G. rostochiensis* inbred lines derived from a Ro1 and a Ro5 population. These genome sequences comprise 173 and 189 scaffolds respectively, marking a ≈ 24-fold reduction in fragmentation as compared to the current reference genome. We provide copy number variations for 19 effector families. Four dorsal gland effector families were investigated in more detail. SPRYSECs, known to be implicated in plant defence suppression, constitute by far the most diversified family studied herein with 60 and 99 variants in Ro1 and Ro5 distributed over 18 and 26 scaffolds. In contrast, CLEs, effectors involved in feeding site induction, show strong physical clustering. The 10 and 16 variants cluster on respectively 2 and 1 scaffolds. Given that pathotypes are defined by their effectoromes, we pinpoint the disparate nature of the contributing effector families in terms of sequence diversification and loss and gain of variants.

Conclusions: Two novel reference genomes allow for nearly complete inventories of effector diversification and physical organisation within and between pathotypes. Combined with insights we provide on effector family-specific diversification patterns, this constitutes a basis for an effectorome-based virulence scheme for this notorious pathogen.

Keywords: Heterozygosity, Gland proteins, Innate immune system, SPRYSEC, CLE, Effectoromics

* Correspondence: joris.vansteenbrugge@wur.nl
[1] Laboratory of Nematology, Wageningen University & Research, Wageningen, The Netherlands
Full list of author information is available at the end of the article

Background

Plant-parasitic nematodes have a significant impact on food and feed production worldwide. Every cultivated crop can be parasitized by at least one nematode species, resulting in a net loss of over 70 billion US dollar annually [1]. From an economic point of view root-knot and cyst nematodes have the highest impact [2]. Whereas root-knot nematodes have a higher impact in warmer climate zones, cyst nematode problems mostly occur in the temperate regions. Unlike root-knot nematodes, most cyst nematodes have a defined center of origin. For example, soybean cyst nematodes originate from northeast Asia and have spread as a successful and highly harmful parasite to all major soybean-growing areas. Potato cyst nematodes diversified in the Andes in South America, and have now proliferated to all major potato production areas in the world (e.g. [3]). Outside of their centers of origin, cyst nematodes belong to the most harmful pathogens of the crops mentioned above.

One of the most widely applied control measures is the use of resistant host plants. Resistances against potato cyst nematodes tend to have a long agronomic life span due to cyst nematodes' unique biology. Potato cyst nematodes usually have only one generation per year through obligate sexual reproduction, go into diapause for months, and - once hatched - their motility is in the range of a few cm per day. Apart from this, remarkably low effective population sizes have been reported for multiple cyst nematode species [4, 5]. Together these characteristics drastically slow down the process of selection and proliferation of virulent individuals, a process that happens underground and therefore often goes unnoticed for years. Potato breeders have introgressed the resistance gene *H1* from *Solanum tuberosum* ssp. *andigena* CPC 1674 into numerous potato cultivars from the 1960's onwards [6]. The *H1* gene confers resistance against *G. rostochiensis* pathotypes Ro1 and Ro4 [7], and this resistance gene is still effective in virtually all major potato producing countries.

Based on a number of *Solanum* differentials, pathotypes have been defined within the two potato cyst nematode species *G. rostochiensis* and *G. pallida.* Five pathotypes named Ro1 - Ro5 have been proposed for *G. rostochiensis*, whereas three pathotypes (Pa1 - Pa3) were discriminated within *G. pallida* [8]. Apart from being laborious and time-consuming, the current pathotype scheme has limited value as it lacks a solid genetic basis. The distinction between for instance the *G. pallida* pathotypes Pa2 and Pa3 is elusive [9]. For *G. rostochiensis*, genome-wide allele frequencies correlate with the geographical distribution of populations, regardless of pathotype [10, 11]. This indicates that the genetic basis of the pre-defined pathotypes is small. A robust pathotyping scheme for potato cyst nematodes is highly desirable because it would lead to far more efficient and durable use of the limited number of host plant resistances currently available. The availability of high-quality reference genome sequences from individual pathotypes would be an ideal starting point for pathotypes' molecular characterization.

Resistant plant species deploy R proteins as surveillance molecules that recognize either directly or indirectly specific effector molecules - or their activities - secreted by nematodes. Nematodes use a protrusible stylet to inject effector proteins into plant cells. Effectors are diverse and fulfil functions ranging from plant cell wall degradation to the induction of a feeding site and suppressing the plant's innate immune system [12]. The nematode produces effectors mainly in the subventral and the dorsal esophageal glands. Effectors are usually members of diversified gene families, and potato cyst nematode typically produces multiple variants per effector. An example is the SPRYSEC gene family that codes for a highly expanded set of proteins that act as activators and suppressors of plant defence [13]. One variant of this family, RBP-1, was shown to trigger the activation of the potato resistance gene *Gpa2* [14] resulting in local hypersensitive response. Effector proteins secreted by the cyst nematode parasite are most likely responsible for the activation of plant resistance proteins. However, this was demonstrated for only a small number of resistance genes (*Gpa2;* [14], *Cf-2;* [15]).

Sequencing the genome of plant-parasitic nematodes is more challenging than for other, larger organisms. With the currently available methods, it is practically impossible to isolate and sequence DNA from an individual nematode to gain enough coverage to generate a high-quality reference genome sequence—especially when isolating high molecular weight DNA required for long-read sequencing technologies. Reference genomes of plant-parasitic nematodes are therefore often based on the genetic material from a population. Consequently, the reference genome includes a substantial heterozygosity level, as the starting material includes a high degree of allelic variation. The current reference genome sequences of potato cyst nematodes *Globodera rostochiensis* [16] and *G. pallida* [17] were each generated using heterozygous starting material (selected field populations), and are relatively fragmented (respectively 4,377 and 6,873 scaffolds). In *G. rostochiensis*, Eves-van den Akker et al. (2016) predicted 138 high confidence effector genes based on sequence similarity with previously described effector gene families. Furthermore, a third of these genes were identified to cluster on effector gene islands. Among these expanded gene families, sequence divergence between different pathotypes was estimated as well. While many single nucleotide polymorphisms and insertions/deletions were observed

[16], the highly fragmented reference genome sequence made it challenging to distinguish between sequence and copy number variation. Less fragmentation in the genome sequence would similarly make it possible to display the degree of clustering of effector genes more accurately.

We generated a new set of reference genome sequences to allow for the accurate organization of effector genes and to compare copy number variation and sequence variation between the Ro1 and Ro5 pathotypes. A precise representation of these two sources of genetic variation is essential for developing molecular pathotyping methods in the future. The current reference genome sequence of *G. rostochiensis* shows a haploid genome size of 95.9 Mb [16] and is expected to spread over eighteen diploid chromosomes [18]. For this, we used two *G. rostochiensis* lines, one fully avirulent and one fully virulent with regard to the *H1* gene. The starting materials for these lines were two distinct field populations sampled from The Netherlands (Ro1-Mierenbos) and Germany (Ro5-Hamerz) [19]. The selection process started with a single cross between an individual male and a female. After multiple generations, fully avirulent Ro1 (Gr-line19) and fully virulent Ro5 (Gr-line22) lines were generated regarding the *H1* resistance in potato [19]. As a result, both Gr-Line19 and Gr-Line22 harbour limited genetic variation, with a theoretical maximum of 4 alleles per locus. For diploid sexually reproducing species, this is the minimum level of heterozygosity that can be present in a population.

New genome assemblies were generated for each of the inbred lines based on PacBio long read-sequencing technology. Using these newly generated *G. rostochiensis* reference genome sequences with a substantially reduced number of scaffolds, we investigated the genomic organisation and the diversification of 19 effector families. A large number of differences in the number of paralogs and variation in sequence content were identified between the effector arsenals of the avirulent Gr-line19 and the virulent Gr-line22. These pathotype-specific effector variants form the basis for the generation of a virulence scheme for potato cyst nematodes.

Results

Genome Assemblies

Two inbred lines of the potato cyst nematode *G. rostochiensis* were initially derived from crossings between individuals from two populations, Ro1-Mierenbos and Ro5-Harmerz [19]. DNA from these lines, Gr-Line19 and Gr-line22, were sequenced using PacBio sequencing technology with respectively 119X and 132X coverage and assembled into two reference genome sequences (Table 1). Benefitting from this long read technology and the significantly smaller genetic background, the two newly generated *G. rostochiensis* genome assemblies are less fragmented than the first genome sequence that was published (nGr.v1.0) [16] while maintaining a comparable assembly size. The number of scaffolds in the new assemblies is about 24-fold lower than in the original *G. rostochiensis* reference genome sequence (Table 1). At the same time, the scaffold N50 increased about 20-fold from 0.085 to around 1.7 Mb. Regarding the assembly size and BUSCO score, the novel assemblies are comparable to the current reference. The assemblies of Gr-Line19 and Gr-Line22 harbor 2,733 and 6,572 gaps, respectively, covering in total 130 Kb and 150 Kb. As compared to the current *G. rostochiensis* reference [16], the number and lengths of gaps showed a 29-fold reduction.

The repeat content in both reference genome sequences is relatively low, 2.6 % for Gr-Line19 and 1.6 % for Gr-Line22. The GC content in repeat regions for Gr-Line19 (40.3 %) was comparable to this genotype's overall GC content (39.1 %). In Gr-Line22, the GC content in repeat regions (32.5 %) was lower than the overall GC content (38.3 %). In predicted protein-coding regions, the GC content is comparable between both reference genome sequences (Gr-Line19: 50.8 %, Gr-Line22: 50.9 %). Using Braker2 as a gene-prediction tool, 17,928 and 18,258 genes were predicted in the Gr-Line19 and Gr-Line22 genome assemblies, coded for 21,037 21,514 transcripts, respectively. The protein-coding regions take up approximately 33 % (Gr-Line19) and 30 % (Gr-Line22) of the genomes at an average density of 89.3 (Gr-Line19) and 86.6 (Gr-Line22) genes per Mb.

Table 1 Comparative genome statistics of three *G. rostochiensis* genome assemblies

***G. rostochiensis* population**	**Assembly ID**	**Size (Mb)**	**# scaffolds**	**Scaffold N50 (Mb)**	**BUSCO results (in %)**			
					Single	**duplicated**	**fragmented**	**missing**
Ro1 population from JHI PCN collection (JHI-Ro1)	nGr.v1.0 [16]	96	4377	0.085	82.8	1.0	8.3	7.9
Gr-Line19	Gr19v10	92	173	1.70	82.2	1.7	7.9	8.2
Gr-Line22	Gr22v10	101	189	1.80	81.5	1.3	8.3	8.9

BUSCO (Benchmarking Universal Single-Copy Orthologs) - eukaryota_odb10

Synteny between the newly generated genomes and the current reference genome [16] was evaluated using a progressive genome alignment. Homologous regions larger than 3 kb and their genomic organization are presented in Fig. 1. A broad span of regions in the nGr.v1.0 reference assembly shows homology to both new assemblies (respectively 67 %, 72 %, and 61 % of the total assembly sizes for JHI-Ro1, Gr-Line19 and Gr-Line22). While the total numbers of base pairs that are covered in a homologous region are roughly within a 10 % range of each other, both new assemblies show substantially larger continuous and so far uncovered regions.

Heterozygosity and structural variation between the two *Globodera rostochiensis* genomes

The inbred lines Gr-Line19 and Gr-Line22 originate from single crossings of individuals, and, as a result, the genetic variation is expected to be smaller than for field populations. To pinpoint the effect of this genetic bottleneck caused by a single crossing, a comparison was made between the proportion of heterozygous and homozygous single nucleotide variants between Gr-Line19 (Ro1) and JHI-Ro1, the selected field population used to generate the current *G. rostochiensis* reference genome [16] while using the Gr-Line22 (Ro5) genome as a reference. Among the called variants that passed the quality filter (JHI-Ro1 n = 584,145; Gr-Line19 n = 716,491), 37 % of the JHI-Ro1 loci were homozygous, as compared to 47 % of the variants in Gr-Line19 (Fig. 2 A). The increased level of homozygosity in Gr-Line19 reflects the relatively narrow genetic basis of this inbred line.

Secondly, structural variation (e.g. insertions, deletions, inversions) of approximately 1 kb or larger within the individual lines and between Gr-Line19 and Gr-Line22 was determined. The proportions of heterozygous and homozygous structural variants with fragment sizes > 1 kb were compared (Fig. 2B). The structural variation within Gr-Line19 and Gr-Line22 was minimal (Fig. 2B). This observation confirms the low level of structural intra-population heterozygosity. The proportion of

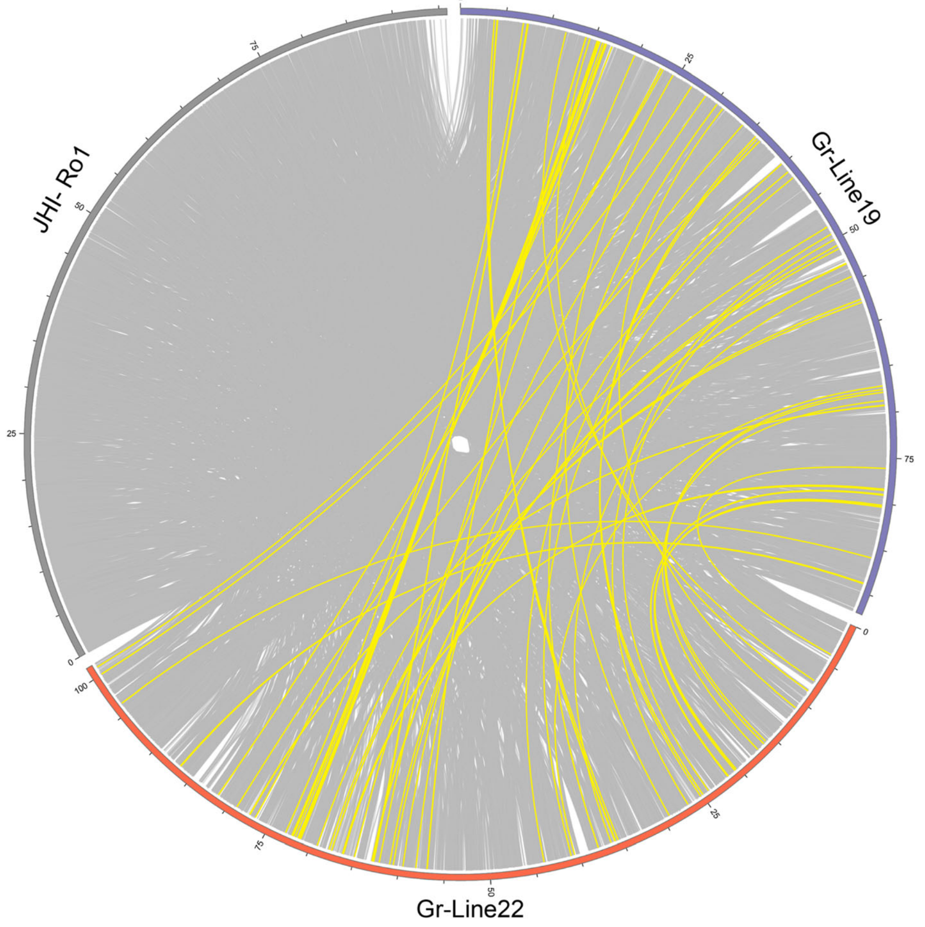

Fig. 1 Synteny between Gr-Line19, Gr-Line22, and JHI-Ro1 based on a progressive genome alignment in Mauve. Only syntenic regions larger than 3 kb are shown. Yellow lines represent regions that are exclusively syntenic between Gr-Line19 and Gr-Line22

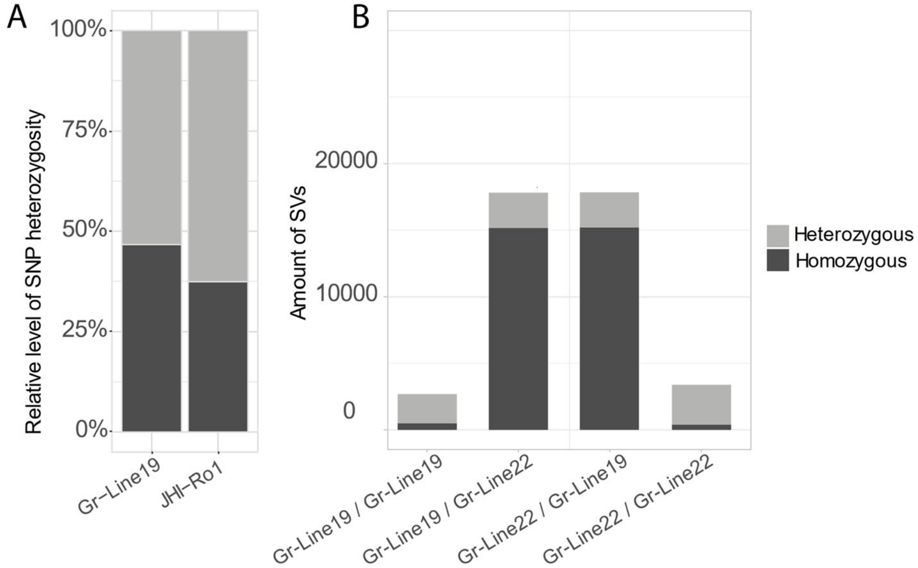

Fig. 2 (**A**) Relatively level of SNP heterozygosity of Gr-Line19 and JHI-Ro1 (JHI – James Hutton Institute, Scotland, UK). Both lines were compared to Gr-Line22. JHI Ro1 was used as the current *G. rostochiensis* reference genome. (**B**) Comparison overall genomic constitution of *G. rostochiensis* inbred lines 19 and 22. In this overview structural variants are shown. Structural variants are DNA region of approximately 1 kb and larger in size, and can include inversions, insertions, deletions

homozygous variants was nearly identical while comparing Gr-Line19 with Gr-Line22 and *vice versa* (Gr-Line22 versus Gr-Line19: 85.09 % & Gr-Line19 on Gr-Line22 85.06 %).

Expansion of Effector gene Families

We identified homologs of 19 known effector gene families from which at least one member was shown to be expressed in the subventral (6) or the dorsal (11) oesophageal gland cells, or in the amphids (1) (Table 2; Fig. 3 A).

For each of these gene families, the copy number differences between Gr-Line19 and Gr-Line22 were determined. The number of paralogs per effector families varied from 99 SPRYSEC variants in Gr-Line22 to a single Hg-GLAND14 gene with signal peptide in the same line. Among the 19 effector families, six have a lower number of paralogs in Gr-Line22, seven have an equal number of paralogs, whereas six have a higher number of variants in Gr-Line22 (Fig. 3 A).

Four effector families show a relatively large difference in the number of paralogs between Gr-Line19 and Gr-Line22. SPRYSEC is by far the most speciose effector family in both lines, but Gr-Line22 harbor 36 more paralogs with signal peptide than pathotype Ro1. Similarly, 11 Hg-GLAND5 homologs were present in Gr-Line19, while Gr-Line22 comprised 18 paralogs with a signal peptide. The reverse was also observed for the subventral gland effector family GH30. Whereas six variants with signal peptide were identified in Gr-Line19, only two were found in Gr-Line22. It is noted that the GH30 family harbors various glycoside hydrolases that were previously categorized as GH5.

Genomic organisation of effector genes

To characterise the genomic organisation of effector genes, the shortest distance between each gene and the closest adjacent gene was calculated (either at the 3’ or 5’-end of the full genomic sequence). This was done for effector genes, as well as for known non-effector genes (i.e. BUSCO gene set). The distances based on the full set of predicted genes ranged from extremely gene sparse to extremely gene dense regions (Fig. 3B). BUSCO genes are generally located in regions that are more gene dense than expected at random (Wilcoxon Rank Sum test $P < 0.0001$). Effector genes expressed in either the dorsal or the subventral esophageal gland cells are often located in more gene sparse regions both as compared to non-effector genes and to any random gene (Wilcoxon Rank Sum test, $P < 0.0001$).

Furthermore, the spatial organization and diversification between two pathotypes of *G. rostochiensis* lines is presented for four selected effector families that are expressed in the dorsal esophageal glands during parasitic life stages. Hg-GLAND5 effectors are known as plant triggered immunity suppressors [31]. Members of the effector family 1106 were demonstrated to suppress both plant triggered immunity and effector triggered immunity [29]. The highly speciose SPRYSEC family was shown to be involved in both the suppression and the activation of the plant immune system [12]. CLE-like

Table 2 Effector families mapped in genomes of *G. rostochiensis* lines Gr-line19 and Gr-line22

Expression	Effector family	Functionality / similarity		Reference
Subventral esophageal glands	GH5[a]	Beta 1,4 endoglucanase	CWDE	[20]
	GH30	xylanase, glucosylceramidase, etc.		[21]
	GH43	candidate arabinanase		[16]
	GH53	candidate arabinogalactanase		[16]
	PL3[b]	Pectate lyase		[22]
	Hg-GLAND 10	cellulose binding protein		[23, 24]
	VAL	Venom allergen-like protein	Immune	[25]
Dorsal esophageal gland	**SPRYSEC**	Suppression and activation of plant innate immunity	Immune	[26]
	GSS	glutathione synthetase-like effectors involved in redox regulation	Feeding site	[27]
	CLE	CLAVATA3/ESR-related peptides, mimic plant CLEs	Feeding site	[28]
	1106	PTI and ETI suppressor	Immune	[29]
	Hg16B09	Suppression plant innate immunity		[30]
	Hg-GLAND1	ETI suppressor		[24] [31]
	Hg-GLAND5	PTI suppressor		[24] [31]
	Hg-GLAND6 (4D06)	PTI suppressor		[24] [31]
	Hg-GLAND 12	Pioneer (function unknown)	Feeding site	[24]
	Hg-GLAND 13	Invertase (*Rhizobium*)		[24]
	Hg-GLAND 14	Endopeptidase (*Ascaris suum*)		[24]
Amphids	HYP	hyper-variable extracellular effector (function unknown)	?	[32]

[a] *GH* Glycoside Hydrolases; [b]*PL* Polysaccharide Lyases, Family numbering according to CAZy (http://www.cazy.org),
'Expression' (left column) refers to nematode organs in which at least one effector family member was shown to be expressed. Subventral esophageal glands of potato cyst nematode are mainly active during migration to the host plant, and host plant penetration. The dorsal esophageal gland shows highest activity during feeding site induction and maintenance. Amphids are chemosensory organs located at the head region of the nematode. For effector families in bold, diversification and physical distribution are investigated in detail. CWDE: cell wall-degrading enzymes, Immune: effector families for which at least one member is known to affect the plant innate immune system, Feeding site: effector families for which at least one member is known to be involved in feeding site induction or maintenance

effectors were demonstrated to be involved in feeding site induction by mimicking the functionality of endogenous host-plant CLE peptides [33]. Concentrating on the distribution of individual family members over the relevant scaffolds, large differences in the level of clustering per family are observed (Fig. 4). While the 60 and 99 SPRYSEC variants are distributed over respectively 18 and 26 scaffolds, the moderately diversified CLE family is concentrated on two scaffolds in case of Gr-Line19, and on a single scaffold for Gr-Line22.

A. Hg-GLAND5

Hg-GLAND5, also referred to as 'putative gland protein G11A06' [34], has first been discovered in soybean cyst nematode *Heterodera glycines*. This effector is expressed in the dorsal gland during a range of parasitic life stages, and it functions as a PTI suppressor [31]. In a transcriptional analysis of two *H. glycines* races, the expression level of Hg4J4-CT26, a GLAND5 family member, was shown to be highly race-dependent [35]. Searches in public genome database revealed that both PCN species harbour homologs of HG-GLAND5 [36].

In *G. rostochiensis*, the GLAND5 effector family comprises 13 and 23 members in Gr-Line19 and Gr-Line22, respectively. Among these variants, three and five are unlikely to be involved in parasitism as the corresponding protein sequences are not preceded by a signal peptide (SP) for secretion (Fig. 3 A). A phylogenetic analysis of the GLAND5 family based on the coding sequences using RaxML revealed an initial split between four effectors without an SP in one clade, and all functional GLAND5 effectors in the other (Fig. 5). Four clusters with mainly secreted GLAND5 variants could be discerned. Differences in numbers of effector paralogs were observed, and in three clusters more Gr-Line22 paralogs are present. Most notable is the diversification in Box I, where eight related GLAND5 representatives from Gr-Line22 surround a single Gr-Line19 variant. Box III and Box IV illustrate expansion in Gr-Line22 (or gene loss in Gr-Line19) as well, although less extreme. On the other hand, in Box II two paralogs of Gr-Line19 are present and a single variant of Gr-Line22.

B. 1106

The 1106 gene family encodes mainly secreted proteins, and members were demonstrated to suppress both PTI

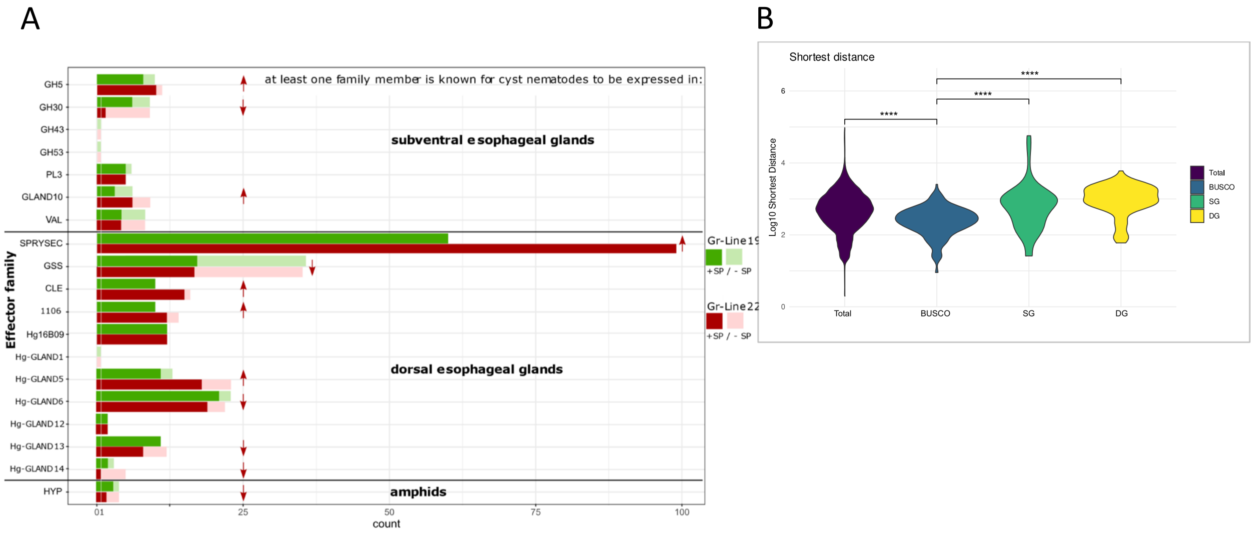

Fig. 3 (**A**) Copy numbers of effector gene families expressed in the secretory gland cells. A comparison was made between Gr-Line19 (Green) and Gr-Line22 (Red). Bright shades of green/red indicate copies that contain a signal peptide for secretion. Faded shades of green/red indicate the absence of a signal peptide for secretion. Upward and downward pointing arrows are used to indicate increase or decrease of the number of variants with a predicted signal peptide for secretion in Gr-Lin22 as compared to Gr-line19. (**B**) Gene density comparison of all genes (total), eukaryotic universal single copy genes (BUSCO), effector genes secreted from the subventral esophageal gland cells (SG), and effector genes secreted from the dorsal esophageal gland cells (DG). The shortest distance of each gene was based on the closest adjacent gene (3′ or 5′) and is measured as the log10 number of base pairs. Statistical significance for each group was determined by comparison with the BUSCO gene set using a Wilcoxon test ($P < 0.0001$)

and ETI responses [29]. In this previous study by Finkers-Tomczak et al. (2011), a conserved region of 1106 variants was shown to hybridize in the dorsal gland of infective juveniles of *G. rostochiensis*. Gr-Line19 contains ten paralogs, whereas 14 1106 paralogs were found in Gr-Line22. In terms of organization, the genes in Gr-Line22 and Gr-Line19 show a comparable degree of physical clustering (Fig. 4). To investigate the diversification of the effector family 1106, the phylogenetic relationship between the variants identified in Gr-Line 19 and Line 22 was examined (Fig. 6). In many cases, a 1106 variant in Gr-line 19 had a single, orthologous equivalent in Gr-line22 (see e.g. Gros19_g2102 and Gros22_g4744, and Gros19_g2104 and Gros22_g4746). Notably, the relationship between the small clusters of 1106 variants was largely unresolved. Within cluster I in Fig. 6, four Gr-Line22 variants were present, and only one representative from Gr-Line19. Two variants, Gros22_g4703 and Gros22_g4696, deviate substantially from the other 1106 family members. It is noted that these variants are not preceded by a signal peptide for secretion and thus are unlikely to act as effectors. Clusters II is highlighted as it represents a local expansion of this effector family in Gr-Line22.

C. SPRYSEC

The SPRYSEC gene family encodes for secreted proteins that contain an SPla and RYanodine receptor. SPRYSECs are produced in the dorsal esophageal glands, and this effector family is by far the most expanded one among the plant-parasitic nematodes [13]. Several SPRYSECs from *G. rostochiensis* were shown to be implicated in the suppression and the activation of defence-related cell death [37]. Suppression was demonstrated for the variants SPRYSEC-4, -5, -8, -15, -18, and – 19, whereas only SPRYSEC15 elicited a defence response in tobacco [12]. In the closely related cyst nematode species *G. pallida*, a single SPRYSEC variant - RBP-1 - was shown to be responsible for the evasion of the potato resistance gene Gpa2, thus preventing a local HR (Sacco et al. 2009). No direct ortholog of RBP-1 could be found among the *G. rostochiensis* SPRYSECs (identity lower than 50 %), with the used filtering criteria.

The diversification of the SPRYSEC-like variants in Gr-Line19 and Gr-Line22 was investigated by analysing the phylogenetic relationships. Although the number of SPRYSECs in Gr-Line19 (n = 60) was already higher than for any other effector family, Gr-Line22 was shown to harbour even more members of this effector family (*n* = 99) (Fig. 3 A). Maximum-likelihood-based inference revealed several SPRYSEC clusters (Fig. 7). Due to the poor backbone resolution, no statements can be made about the relationship between these clusters. It is noted that the support values for the more distal parts of the SPRYSEC tree are substantially higher than the support values for most of the more proximal bifurcations. Three

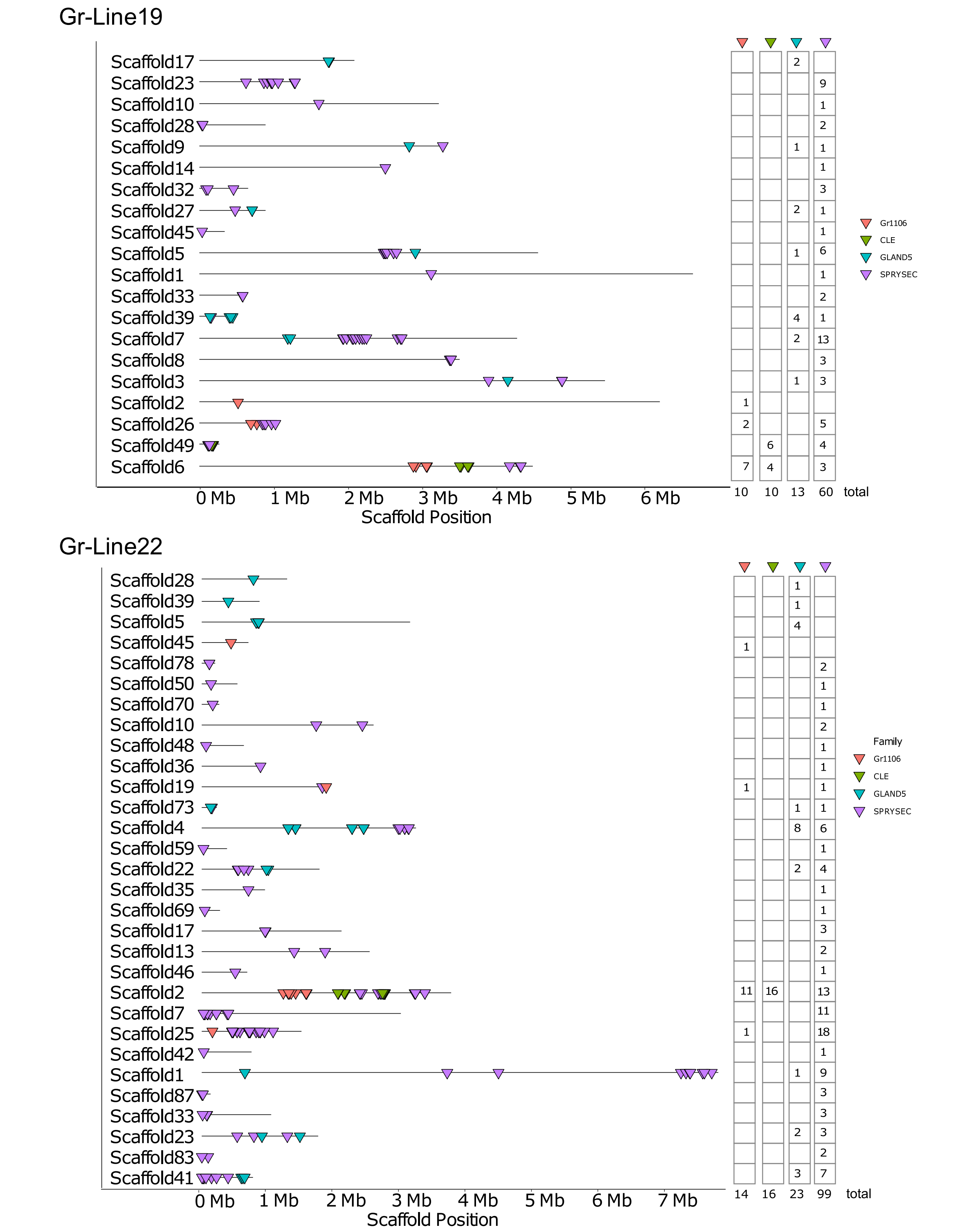

Fig. 4 Spatial distribution of genes belonging to the effector families Gr1106 (Red), CLE (Green), GLAND5 (blue), and SPRYSEC (purple). Each triangle indicates the genomic position of a single gene. At the right, the number of variants per effector family are given for each scaffold

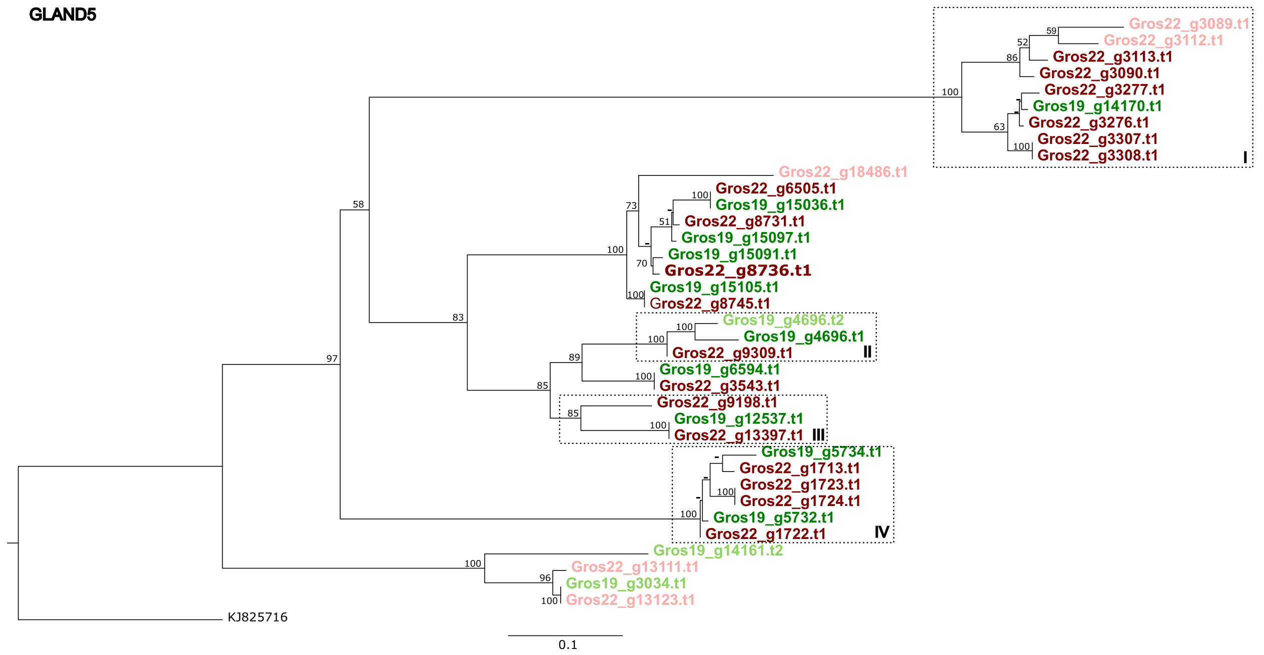

Fig. 5 Phylogeny of GLAND5 effector genes of both Gr-Line19 (Green) and Gr-Line22 (dark red). A multiple sequence alignment was made using MUSCLE on the coding sequence. A phylogenetic tree was made using RAxML using a GTRGAMMA model, validated by 100 bootstrap replicates. Bootstrap values < 50 % are indicated by a "-". Lighter shades of green or red indicate effector variants that lack a signal peptide for secretion. Boxed clusters (Roman I – IV) highlight asymmetric representations of variants among one of the two *G. rostochiensis* lines

large (A, B, and D) and two smaller (C, E) SPRYSEC clusters could be identified. The majority of Gr-Line19 gene family members have a single orthologous equivalent in Gr-Line22, while Gr-Line22 contains additional paralogs in each of the clusters.

As compared to B and D, cluster A shows the highest level of diversification. Both types of asymmetric SPRYSEC expansion were found in this cluster. Box I in Cluster A comprises a single Gr-Line22 and three Gr-Line-19 SPRYSECs. Box II exemplifies Gr-Line22 expansion, where four closely related Gr-Line22 SPRYSECs surround a single Gr-Line19 variant.

Cluster B harbours three of the SPRYSEC variants described in [12] (SPRYSEC-4, -5 and – 8), all of which seem to be represented by a single orthologous pair.

Cluster C is characterized by a set of genes homologous to SPRYSEC-15 that are considerably expanded in Gr-Line22. It is noted that Gros19_g2329.t1 also is the closest match of SPRYSEC-18, however only with 60 % identity.

Cluster D unites SPRYSEC variants with a low degree of diversification. Although most Gr-Line-19 variants have a single equivalent in Gr-Line22, there are a few examples of further diversification in Gr-Line22. Box III shows a notable example of a diversification event where a single Gr-Line19 variant has five closely related equivalents in Gr-Line22.

SPRYSEC-19, a variant that was demonstrated to suppresses programmed cell death mediated by several immune receptors [37], localized in cluster E. SPRYSEC-19 was first identified in a *G. rostochiensis* Ro1 Mierenbos population [26], which is the population Gr-Line19 was originally derived from. Cluster E shows the Gr-line 22 equivalent of SPRYSEC-19 (g7323).

D. CLE-like

The CLE-like gene family is an unusual effector family coding for prepropeptides that are delivered via the stylet of the infective J2 to the syncytial cell. For CLEs from the related cyst nematode species *Heterodera glycines* with domain structures similar to *G. rostochiensis*, it was shown that the mature propeptide comprised a nematode-specific translocation signal that facilitated the export from the developing syncytium to the apoplast [38]. Subsequently, the protein is cleaved outside the plant cell, and bioactive CLEs are released [39]. Two classes of CLE-like proteins were found to be expressed in the dorsal gland of *G. rostochiensis*. Members of the *Gr-CLE-1* class showed moderate (≈ 10 fold) upregulation in early parasitic life stages (peak in parasitic J-3), whereas *Gr-CLE-4* representatives showed an over 1,000 fold in later parasitic stages (at 21 dpi) (Lu et al., 2009).

The two *G. rostochiensis* lines 19 and 22 harbour 10 and 16 CLE variants, and both lines comprise members of the *Gr-CLE-1* and the *Gr-CLE-4* class. As shown in

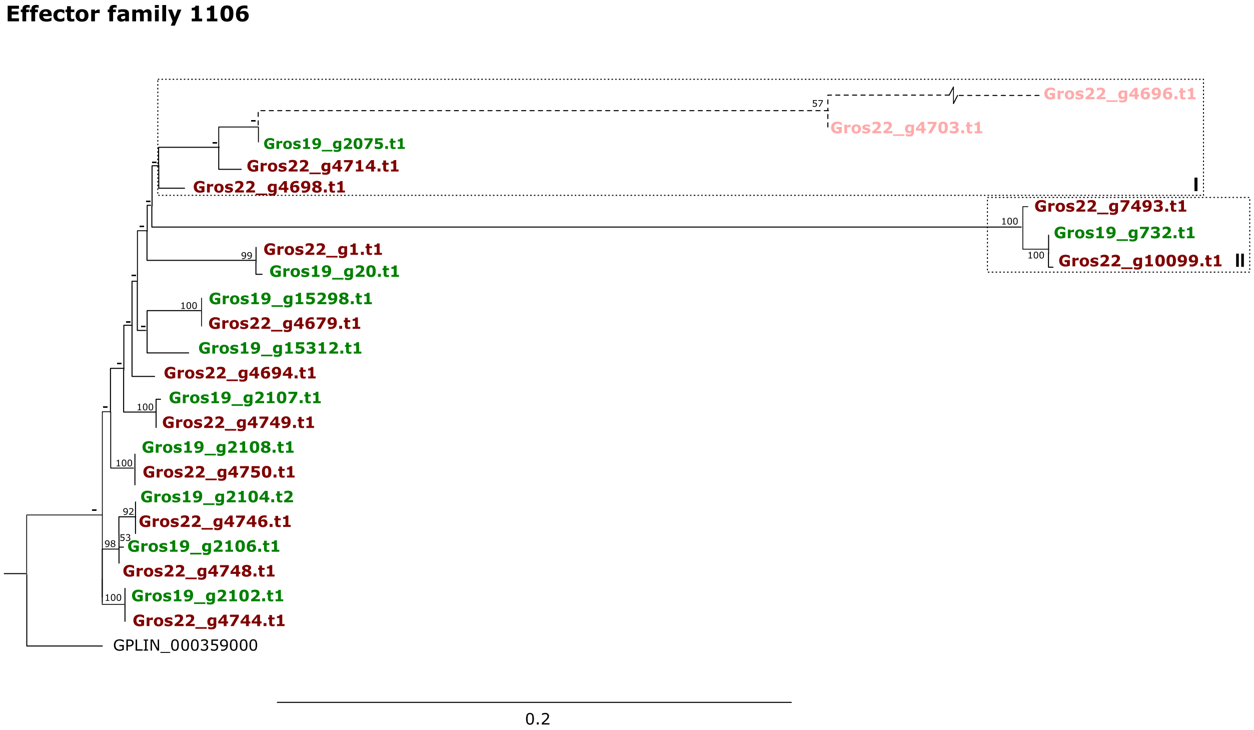

Fig. 6 Phylogeny of Gr1106 effector genes of both Gr-Line19 (green) and Gr-Line22 (dark red). A multiple sequence alignment was made using MUSCLE on the coding sequence. A phylogenetic tree was made using RAxML using a GTRGAMMA model, validated by 100 bootstrap replicates. Bootstrap values < 50 % are indicated by a "-". Lighter shades of green or red indicate effector variants that lack a signal peptide for secretion. Boxed clusters (Roman I - II) highlight asymmetric representations of variants among the two *G. rostochiensis* lines. Dashed lines are used in case there was uncertainty about the location of the 5′end of a given effector family variant

the phylogenetic analysis (Fig. 8), members of class *Gr-CLE-4* show little variation among each other, while *Gr-CLE-1* s show a higher level of diversification. As compared to Gr-Line19, the number of *Gr-CLE-4* variants had doubled from four to eight in Gr-Line22. On the contrary, each of the Gr-Line19 representatives of *Gr-CLE-1* had a single homolog in Gr-Line22. In Fig. 8, clusters A and B include four Gr-Line22 variants with no immediate ortholog in Gr-Line19. In cluster C, a Gr-Line19 variant is present that deviates substantially from the closest Gr-Line22 orthologous sequence. Cluster D contains an example of a homologous gene pair, with a tentative duplication in Gr-Line22.

In addition to sequence similarity within the Gr-CLE function classes, there is also a high degree of physical clustering (Fig. 9). The Gr-CLE-4 variants are all located adjacent to each other, and not interspersed by any other gene. Based on this remarkable physical organization, we hypothesize that one or more duplication events in this region underlies the copy number difference of Gr-CLE-4 effectors ($n = 4$ in Gr-Line19; $n = 8$ in Gr-Line22).

Discussion

Due to specific biological characteristics of plant-parasitic nematodes, host plant resistances tend to be a remarkably durable means to manage this category of soil-borne pathogens. The main challenge is the actual developing and breeding resistant host-plant varieties. As the genetic basis for virulence in plant-parasitic nematodes is unknown, breeding for resistance can only be done on a trial-and-error basis. The whole process is, therefore, inefficient, and thus time consuming and expensive. The availability of molecular-based pathotyping methods of plant-parasitic nematode populations would allow for the deployment of more targeted resistance. Here we concentrated on two *Globodera rostochiensis* inbred lines, Gr-Line19 and Gr-Line22, with distinct pathotypes [19]. Resulting from a single male-female crossing by Janssen et al. (1990), each of these lines' genomic background is small, with a maximum of 4 haplotypes per locus. These small genomic backgrounds signifcantly simplify the generation of high-quality reference genome sequences, which has been a challenge for sexually reproducing plant-parasitic nematodes in the past. Therefore, we expect that the reference genome sequences of Gr-Line19 and Gr-Line22 are a more accurate representation of the *G. rostochiensis* genome, making the process of molecular pathotyping a step closer. Furthermore, long-read sequencing technology allowed us to generate reference genomes about 24 fold

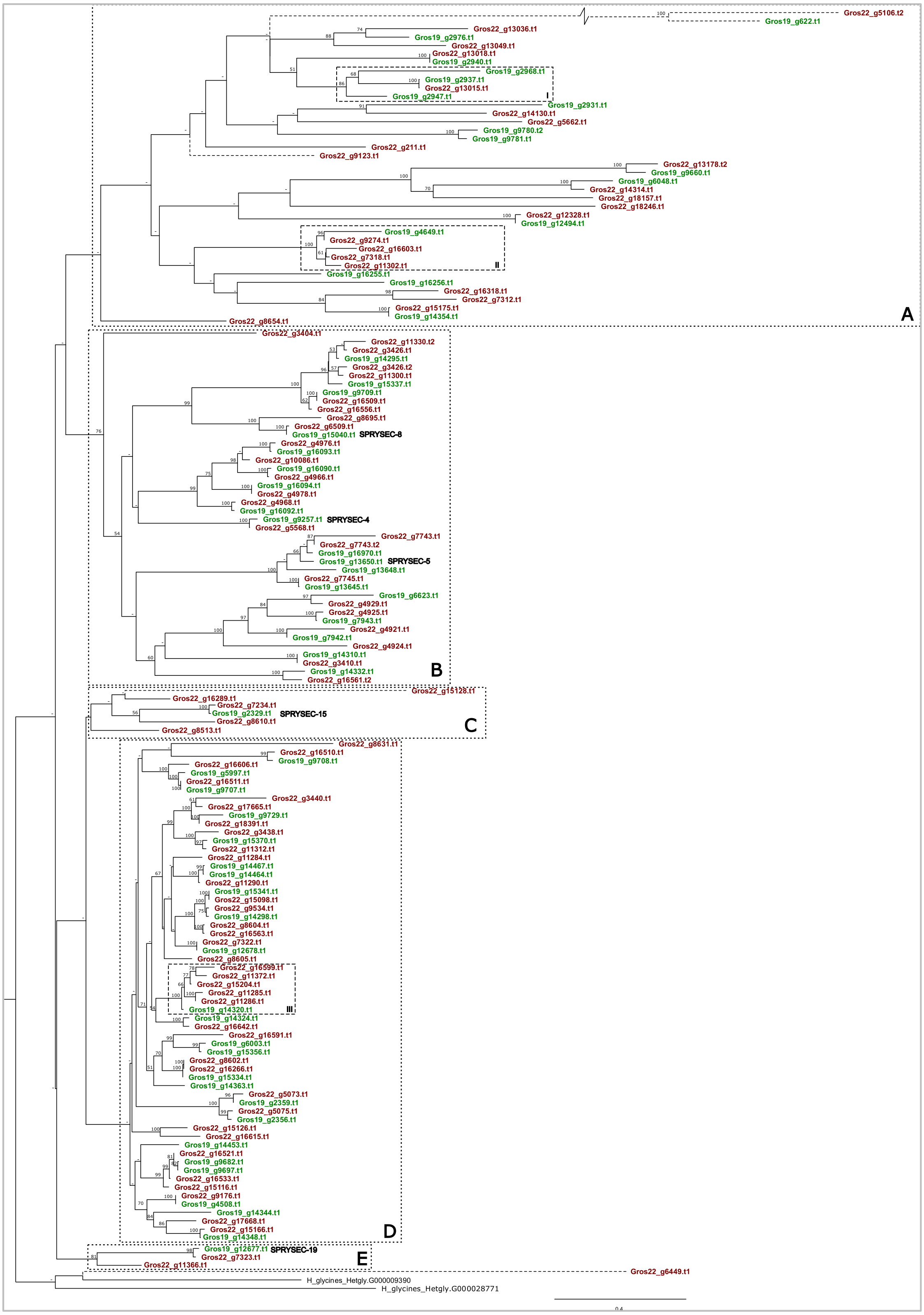

Fig. 7 (See legend on next page.)

(See figure on previous page.)

Fig. 7 Phylogeny of SPRYSEC effector genes of both Gr-Line19 (green) and Gr-Line22 (dark red). Only SPRY proteins with a signal peptide for secretion are included A multiple sequence alignment was made using MUSCLE on the coding sequence. A phylogenetic tree was made using RAxML using a GTRGAMMA model, validated by 100 bootstrap replicates. Bootstrap values < 50 % are indicated by a "-". Closest homologs to the functionally described SPRYSEC-4, SPRYSEC-5, SPRYSEC-8, SPRYSEC-15, and SPRYSEC-19 [10] are shown. Clusters of SPRYSEC variants are boxed (A-E). Boxed clusters (Roman I – III) highlight asymmetric representations of variants among the two *G. rostochiensis* lines. Dashed lines are used in case there was uncertainty about the location of the 5'end of a given effector family variant

less fragmented than the current reference genome [16]. This higher continuity made it possible to pinpoint the physical distribution and the diversification in a way that was not possible with the highly fragmented JHI-Ro1 reference genome sequence. Four effector families that, together with other effectors, define this potato cyst nematode's pathogenicity were explicitly studied in detail.

One of the main technical challenges we tried to overcome was generating high-quality reference genome

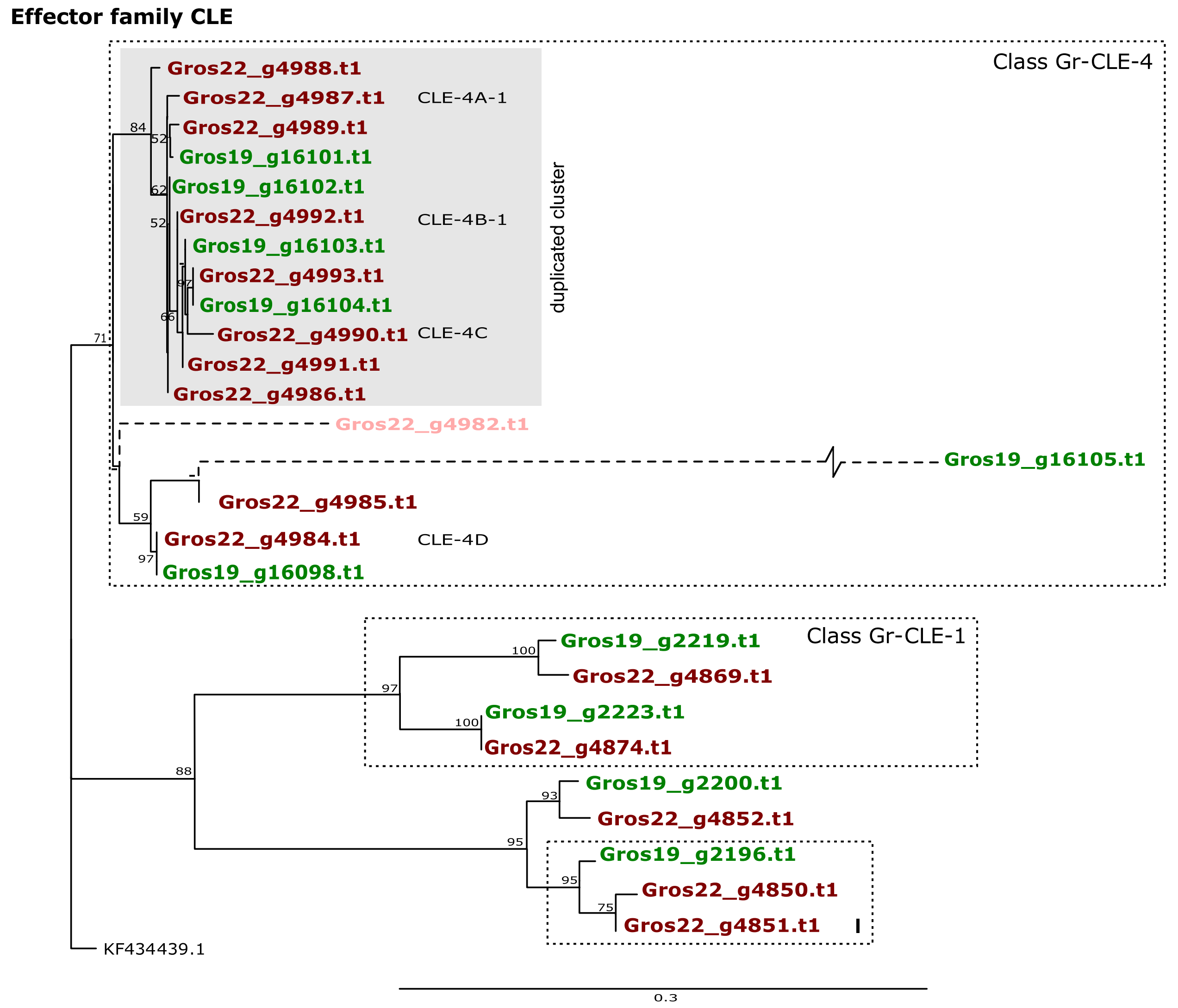

Fig. 8 Phylogeny of CLE effector genes of both Gr-Line19 (Green) and Gr-Line22 (dark red). A multiple sequence alignment was made using MUSCLE on the coding sequence. A phylogenetic tree was made using RAxML using a GTRGAMMA model, validated by 100 bootstrap replicates. Bootstrap values < 50 % are indicated by a "-". Lighter shades of green or red indicate effector variants that lack a signal peptide for secretion. Genes belonging to the functional classes Gr-CLE-1 and Gr-CLE-4 [28] are labeled with dashed boxes. A boxed cluster (Roman I) highlights an asymmetric representation of variants among of the two *G. rostochiensis* lines. Dashed lines are used in case there was uncertainty about the location of the 5'end of a given effector family variant

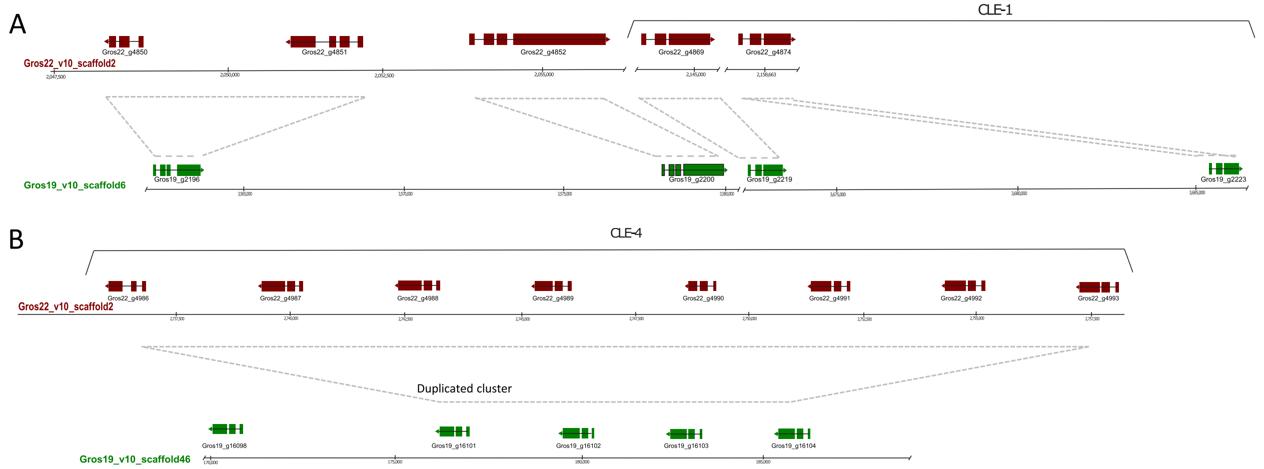

Fig. 9 Spatial organization of CLE functional classes CLE-1 (A) and CLE-4 (B) [28] for both Gr-Line19 (Green) and Gr-Line22 (Dark red). In panel A, the spatial organization of CLEs belonging to Class Gr-CLE-1 are shown for both lines, as well as the other CLEs residing in the same major CLE clade (Fig. 8). In panel B, the spatial organization of CLEs belonging to Class Gr-CLE-4 are shown (grey panel in Fig. 8), as well as CLE variant Gros19_16098 as this variant was located close to the Gr-Lin19 Class-4 CLEs

sequences of a highly heterozygous nematode species. In terms of assembly sizes, we see a comparable size to the *G. rostochiensis* JHI-Ro1 genome (Eves-van den Akker et al., 2016) as well as to the estimated *G. rostochiensis* genome size [40]. Therefore, it is likely that the high levels of heterozygosity did not negatively impact the assembly by the presence of haplotigs [41]. Another possibility is that the presence of many haplotypes negatively influenced the fragmentation of the assembly. Due to more variation in the bases, it might have been more challenging to combine more contigs into scaffolds. To further reduce the number of scaffolds, possibly to a chromosome level, it might be advantageous in the future to supplement long-read sequencing with other techniques such as optical mapping [42, 43].

We furthermore assessed the effect of generating a genome assembly of a highly inbred line instead of a regular population. A comparison was made between SNPs' zygosities called on short-read data and found that Gr-Line19 had a 10 % higher proportion of homozygous SNPs than the JHI-Ro1 population. This suggests that there is indeed a smaller number of haplotypes present in the inbred-lines than in the JHI-Ro1 population. Since more than 50 % of the called SNPs in Gr-Line19 are still heterozygous, it seems reasonable to assume that the measured heterozygosity levels provide a more realistic picture of the heterozygosity that is present in an individual. Which is in line with previous findings in the cyst nematode *Heterodera glycines* by Ste-Croix et al. (2021) [44], who described that individuals can have mixing levels of zygosity.

A more detailed analysis of four effector families for which at least a subset of members are known to be expressed in the dorsal gland of nematodes during feeding site induction or maintenance revealed dozens of novel potential virulence-associated variants.

To some extent, our starting point was comparable to the approach taken by Bekal et al. (2008) [45]. Within the soybean cyst nematode *Heterodera glycines* subsets of populations with comparable pathogenicity have been defined based on their multiplication characteristics on a set of seven soybean indicator lines. Populations that shared multiplication characteristics were coined 'HG types' [46]. Subsequently, Bekal and co-workers (2008) [45] used two inbred lines that were either avirulent ('TN10'; HG type 0)) or virulent ('TN20'; HG type 1, 2, 3, 4, 5, 6, 7). 454 micro-bead sequencing of these indicator lines resulted in the generation of tens of millions of short reads (110–120 bp), which allowed for whole-genome comparative analysis. These efforts resulted in 239 homozygous SNPs between TN20 and TN10 [45]. Although the relationship between these SNPs and pathogenicity is unclear, these SNPs could be considered one of the first molecular markers for pathogenicity in cyst nematodes. Here we took it one step further, by identifying copy number variation that might serve as potential pathotype specific molecular markers. Copy number variation is relevant, as it has been linked to virulence in various pathogens [47, 48] including plant parasitic nematodes [49].

For potato cyst nematodes, Folkertsma et al. (1996) [50] used AFLP assays [51] to characterize pathotypes of the potato cyst nematodes *G. rostochiensis* and *G. pallida*. Almost 1,000 marker loci were employed to genotype populations of both potato cyst nematode species. These analyses revealed genetic markers that can distinguish between the *G. rostochiensis* pathotypes Ro1, Ro3 and Ro4, while such loci appeared to be absent for the

G. pallida pathotypes Pa2 and Pa3. In a more extensive approach focussing on *G. rostochiensis* only, Mimee et al. (2015) [10] employed a restriction enzyme-based genotyping-by-sequencing approach. The genotypic characterization of 23 populations, covering all five pathotypes, revealed a clear distinction between pathotypes Ro1 and Ro2 on the one hand, and Ro,3, Ro4, and Ro5 on the other. Moreover, their analyses seemed to demonstrate intra-pathotype variation within Ro1. However, it is noted that with 14 populations from 9 different countries, Ro1 was overrepresented in this research.

The first reference genome for *G. rostochiensis* was published by Eves-van den Akker (2016) [16], and - in conjunction with this - the intra-species variation regarding members of known effector families was mapped. All five *G. rostochiensis* pathotypes were represented in this study, and whereas homozygous molecular markers to discriminate between Ro4 and Ro5 could be identified, this was not possible for the remaining three pathotypes. Moreover, this research confirmed the large genotypic diversity of populations that are all labeled Ro1, indicating that there are many possible genotypes that yield a similar Ro1-like virulence. Here it might be mentioned that Ro1 and Ro4 share the inability to parasitize potato genotypes that harbor the *H1* resistance gene from *Solanum tuberosum* ssp. *andigena* CPC 1673. Moreover, the H1 resistance genes have been introgressed in most commercial potato varieties, and potato cyst nematode populations worldwide have been exposed to these resistance genes likely including the ones characterized by [16]. So, although these pathotypes share their avirulence concerning the *H1* gene, they belong to another *G. rostochiensis* genotype and differ significantly in intra-pathotype variation.

Hence, as a starting point, we used pathotypically characterized inbred lines from which we generated new reference genome sequences. On this basis, complete effector families could be mapped and compared. In essence, the make-up of effector families in lines with distinct pathogenic characteristics could vary because of (1) non-synonymous variants in sequence in a given set of effector genes and/or (2) effector gene loss or gain (3) quantitative variation in expression levels due to SNPs in the promotor region (4) quantitative variation in expression levels due to copy number variation. The balance between these two (dependent) sources of variation varies in a pathogen-dependent manner. The genome-wide comparison of three *Microbotryum* species parasitizing distinct Caryophyllaceae allowed Beckerson et al. [52] to define the secretomes of the individual species. Their analyses revealed that host specificity was explained by rapid changes in effector genes rather than by variation in the effector copy numbers. With a similar underlying question, Qutob et al. (2019) [53] investigated two effector genes families of *Phytophthora sojae*, Avr1a and Avr3a in a range of races. The presence of multiple copies of nearly identical genes on the Avr1a and the Avr3a locus was suggested to contribute to the fitness of these races, and races with distinct pathogenicities were characterized by variations in effector gene numbers. These examples demonstrate that both sources of variation can generate differences in pathogenicity among plant pathogens. Here we specifically focussed on effector gene loss and gain effects, and observed that both events happen in the avirulent Gr-Line19 as well as the virulent Gr-Line22. Previous studies show that, at least in potato cyst nematodes, single nucleotide polymorphisms are also related to virulence (e.g. [14]), which indicates that both types of genomic variance are relatable with virulence.

In case of the tropical root-knot nematode *Meloidogyne incognita*, Castagnone-Sereno et al. (2019) [49] tried to pinpoint the genetic basis of avirulence and virulence with regard to the tomato resistance gene *Mi-1.2.* Genome-wide characterization of two pairs of avirulent and virulent lines revealed 20 gene families that all showed a lower number of copies in both virulent *M. incognita* lines. It is noted that the 20 families included pioneers and household genes, and not known effector families. Hence, although a lower copy number per gene family was associated with virulence, this research did not identify gene loss events that could be causally related to virulence.

We separately considered the dorsal esophageal gland-expressed effectors that are thought to be involved in immune response suppression and feeding site induction, and the subventral gland-expressed effectors that are active during plant penetration. Concerning dorsal gland-expressed effector families, *G. rostochiensis* Gr-Line19 harboured on average 14 genes per effector family, while on average, 19 members were identified per effector family in Gr-Line22, a homozygous virulent line regarding the H1 resistance gene. In our analysis, four effector families showed a higher number of variants in the avirulent Gr-Line 19, and four other families showed a reverse pattern (Fig. 3 A). The afore-mentioned difference in the average number of variants per family is explained by the differences in the extent to which the number of variants had changed in the two lines.

The effector families expressed in the subventral glands that are included in this study showed less expansion than the dorsal gland specific families, with only small differences in copy numbers between the two lines. Strikingly, a substantial number of genes belonging to this category lack a signal peptide presence. Since many of these genes (e.g., glycoside hydrolases, pectate lyases) code for cell wall-degrading or modifying enzymes, the proteins would have to be secreted to make physical contact with the plant in order to perform their function. One hypothesis could be that these genes are, in fact, pseudogenes. However, this seems unlikely as manual inspection showed that most of these SP lacking genes show a RNAseq signal (results not shown). Whereas ample RNAseq data allowed for an accurate prediction of the intron-exon structure, the transcription start site is more difficult to predict without additional experimental data. If transcription start sites were misplaced, we could have missed a preceding signal peptide. Alternatively, it could be that this cyst nematode genuinely harbours effector variants without apparent signal peptide similar to the invertases identified in *Meloidogyne incognita.* These effectors were suggested to be acquired at a late stage during cyst nematode evolution [54].

Phylogenetic analysis of effector families as presented here takes along both effector diversification and effector loss and gain. These data clearly demonstrate that the balance between both sources of variation differs per effector family. Whereas effector family 1106 showed overall little copy number variation, SPRYSEC genes were 65 % more abundant in Gr-Line22, and in case of the GLAND5 family significant diversification was accompanied by a large difference in copy numbers between both lines. Other population genetic studies on plant-pathogenic fungi and oomycetes showed exclusively low [55] or high [54] levels of diversification between effector genes. We are not aware of other plant pathogens for which such drastic contrasts in diversification pattern between effector families were described.

Because of its extreme level of physical clustering of CLE effectors in both *G. rostochiensis* inbred lines we investigated its genomic organisation in more detail. Potato cyst nematodes produce and secrete mimics of plant CLEs. Plant CLEs are signalling components that were shown to be conserved in both Arabidopsis and potato roots [28]. Among *Globodera* CLE genes two functional classes are distinguished, CLE1 and CLE4. The main difference between these classes is the composition of CLE peptides that are present as small cleavable units separated by small spacers at the protein's C terminus. [33] described a single CLE1 representative, and here a second potential CLE1 variant is identified in both *G. rostochiensis* lines (Fig. 8 & Supplemental Figure 3). This second variant has a domain structure similar to GrCLE1 (Supplemental Fig. 1), and the conservation of the domain structure makes it plausible this variant has a CLE1-like function. Notable is the putative duplication event of Gr-CLE4 genes in Gr-Line22. Gr-CLE-4 genes are highly conserved, even between pathotypes and we assume that this duplication event might result in a higher production of GrCLE4 peptides. A dose effect for a nematode effector was previously reported for the 32E03 effector of the beet cyst nematode *Heterodera schachtii* [56]. So our finding might suggest that the virulent *G. rostochiensis* line 22 might exert a stronger CLE4 peptide-based effects on its host.

Conclusions

Molecular pathotyping is an essential element in durable disease management. After all, this will allow breeders to use host plant resistances in a targeted way, and it allows farmers to make a more informed decision which potato variety to grow in the field. The existing pathotyping system for *G. rostochiensis* classifies populations into five pathotypes (Ro1-Ro5) on the basis of their relative multiplication rates on a number of *Solanum* differentials, and this systematic was used as starting point for the generation of new pathotyping platform. By generating high quality reference genomes from two pathotypically-distinct inbred lines, we were able to generate broad overviews of effector families including their diversification and spatial organisation. On the basis of a selection of four effector families, dozens of effector variants could be pinpointed that were unique for either of the two inbred lines Gr-Line19 (Ro1) and Gr-line22 (Ro5). Once these data are supplemented by re-sequencing data from well-characterized *G. rostochiensis* field populations, comparative effectoromics would be within reach. Comparative effectoromics will provide a foundation for our understanding of compatible and incompatible host-nematode interactions as well as for a new, biologically insightful pathotyping scheme as a basis for the durable use of host plant resistances.

Methods

DNA isolation and sequencing

Cysts from two *G. rostochiensis* lines that were previously selected by Janssen et al. (1990) [19] for being fully avirulent Ro1 (Gr-line19) or fully virulent Ro5 (Gr-line22) with regard to the *H1* gene were used as starting material for the collection of pre-parasitic second-stage juveniles (J2). J2 nematodes were concentrated, and sucrose centrifugation was used to purify the nematode suspension [57]. After multiple rounds of washing of the purified nematode suspension in 0.1 M NaCl, nematodes were resuspended in sterilized MQ water. Juveniles were lysed in a standard nematode lysis buffer with proteinase K and beta-mercaptoethanol at 60 °C for 1 h as described by Holterman et al. (2006) [58]. The lysate was mixed with an equal volume of phenol: chloroform: isoamyl alcohol (25:24:1) (pH 8.0) following a standard DNA purification procedure, and finally, DNA was precipitated with isopropanol. After washing the DNA pellet with 70 % ethanol for several times, it was resuspended in 10mM Tris-HCL (pH 8.0). DNAs of both inbred lines (each 10–20 μg) were sequenced using Pacific Biosciences SMRT sequencing technology at Bioscience

(Wageningen Research, Wageningen, The Netherlands) Gr-line19 was sequenced to a depth of approximately 119X with an average read length of 5,641 bp, whereas Gr-line22 was sequenced 132X with an average read length of 7,469 bp. Depth was calculated based on the assembly sizes. In parallel, a 2 × 250 bp Illumina NovaSeq run resulted in 188x coverage of paired-end reads per line used to polish the initial assemblies. The raw sequencing reads and the genome assemblies are available under NCBI accession PRJNA695196.

Genome assembly

Raw PacBio reads were first corrected by merging haplotypes with the correction mode of Canu v1.8 [59], allowing a corrected error rate of 15 % and a corrected coverage of 200. Using long-read assembler wtdgb2 v2.3 [60], approximately one hundred assemblies were generated per inbred line, optimizing the parameters minimal read length, k-mer size, and minimal read depth. The quality of the initial assemblies was assessed based on whether the assembly size was close to the genome size estimate [16]. Completeness of the genome was assessed using BUSCO v3 [61] using the standard library of eukaryotic single copy genes. Based on the criteria mentioned before, the most optimal assembly was then selected for each line and used for post-assembly processing. For Gr-Line19 a minimal read length of 6,000 was used, together with a k-mer size of 20 and a minimal read depth of 6. For Gr-Line22 a minimal read length of 5,000 was used, together with a k-mer size of 15 and a minimal read depth of 6.

After determining the most optimal assembly, remaining unmerged haplotigs were filtered from the assembly using Purge Haplotigs v1.0.4 [41]. The assembly was then tested for contamination using the blobtools pipeline v.1.0.1 [62] (Supplemental Fig. 1 & Suplemental Fig. 2). Contigs were scaffolded with PacBio reads using SSPACE-Longread with a minimum overlap length of 1000 bp and a minimum gap between two contigs of 500 bp [63]. The remaining gaps in the scaffolds were then filled using a consensus alignment approach with a minimum coverage per position of 10 reads [64]. NovaSeq data were used to polish the resulting assemblies using three iterations of Arrow v2.3.3 at default settings (https://github.com/ PacificBiosciences/GenomicConsensus) and five iterations of Pilon v1.23 [65] each. Repeat regions were soft masked using RepeatModeler v1.0.11 (https://github.com/Dfam-consortium/RepeatModeler) and RepeatMasker v4.0.9 (http://www.repeatmasker.org/ RepeatMasker/). Gene annotations in gff3 format were predicted for both assemblies using BRAKER v2.1.2 [66]. The prediction of gene models was aided by RNAseq datasets of different life stages of *G. rostochiensis* (NCBI BioProject accessions: PRJEB12075, PRJNA274143). While this data originates from a different *G. rostochiensis* population (JHI-Ro1), addition of this type of data greatly improved the quality of the gene predictions using Braker. Sequencing reads from these RNAseq datasets were mapped on both genomes using Hisat v2.2.0 [67]. All scripts used for the generation of the genome assemblies including all relevant details are available on Github (https://github.com/Jorisvansteenbrugge/GROS_genomes).

Genome Synteny

Genome synteny was determined between the genome assembly of Line19, Line22 and the previous Ro1 reference genome (NCBI BioProject PRJEB13504) through a progressive genome alignment using Mauve v2.4.0. The alignment was then visualized in Circos v0.69-9 [68], showing only syntenic regions of 3 kb and larger.

Estimating Heterozygosity levels and structural variation

Heterozygosity levels were estimated based on the frequency of heterozygous versus homozygous variants (SNPs and small indels). A comparison was made between Gr-Line19 and the JHI-Ro1 population, using the Gr-Line22 genome assembly as a reference. Illumina reads were mapped with Burrows-Wheeler Aligner [69] using default settings. For Gr-Line19, a library of Illumina NovaSeq reads (accessions: SRR13560389, SRR13560388) was used, and for JHI-Ro1, Illumina HiSeq reads (accessions: ERR114519) were mapped against the reference. Variants were called with bcftools v.1.9 [70] with multiallelic variant calling enabled, at a maximum depth of 1,000 reads.

The structural variation between the newly generated assemblies of Gr-Line19 and Gr-Line22 was estimated by the frequency of heterozygous *versus* homozygous structural variants (SVs) with fragment size > 1 kb. Raw Pacbio reads of Gr-Line19 were mapped against the Gr-Line22 (and *vice versa*) using NGMLR v0.2.7 [71] with default settings. SVs were then called using Sniffles v1.0.10 running the standard settings [71].

Identification of Effector Homologs

Effector genes were identified in both line Gr-line19 and Gr-line22 based on the proteomes predicted by BRAKER2 [66]. Phobius [72] was used to check for the presence of a signal peptide for secretion. Homologs for glycoside hydrolase (GH) families 5, 30, 43, 53, Pectate lyase 3, Glutathione Synthetase were identified with HMMER v3.2.1 [73] based on pre-calculated profile HMMs in the PFAM database [74] (entries PF00150, PF02055, PF04616 and PF07745, PF03211, PF03199 respectively). SPRYSEC homologs were identified by testing protein sequences for a SPRY domain (hmm profile PF00622). Arabinogalactan galactosidase homologs were identified with a custom profile HMM-based on UniProt sequences (entries O07012, Q65CX5, Q65CX4, D9SM34, P48841, O31529, Q8 × 168, Q5B153, O07013, P83692, P48842, P83691, Q4WJ80, B0XPR3, A1D3T4, Q2UN61, Q0CTQ7, A2RB93, Q9Y7F8, B8NNI2, Q76FP5). CLE-like homologs were identified with a custom profile HMM-based on UniProt sequences (D1FNJ7, D1FNK5, D1FNJ9, D1FNK2, D1FNK8, D1FNK3, D1FNK0, D1FNK4). GenBank peptide sequences JQ912480 to JQ912513 were used to generate a custom profile HMM for the effector family 1106. Based on GenBank entries KM206198 to KM206272, a custom profile HMM was made for the HYP effector family. Homologs of the *Hetereodera glycines* effector families Hg16B09 (GenBank: AAO85454) and GLAND1-18 (GenBank: KJ825712 to KJ825729) were identified with blastp, with the following cut-offs: an identity score higher than 35 %, a query coverage of at least 50 %, and an E-value lower than 0.0001.

Phylogeny

Multiple Sequence Alignments were generated based on the coding sequences of the orthologs per effector family, using Muscle v3.8.1551 [75] using standard options. To test for the best model of DNA substitution, ModelTest-NG [76] was used. Except for GLAND5, the best model for all effector families was GTRGAMMA. For GLAND5, GTRGAMMAI was marginally better. As the resulting phylogenetic tree was almost identical to the GTRGAMMA, we decided to stick to this model for sake of uniformity. Phylogenetic trees were then generated with RaxML v8.2.12 [77] running a GTRGAMMA model with 100 bootstrap replicates. The resulting trees were visualized and organized in Figtree (v. 1.4.4).

Abbreviations

CLE: CLAVATA3/Endosperm Surrounding Region-like proteins.; dpi: Days post inoculation.; ETI: Effector Triggered Immunity.; GH: Glycoside Hydrolase.; GSS: Glutathione synthetase.; Gr-Line19: Inbred line of *Globodera rostochiensis* derived from a pathotype Ro1 population.; Gr-Line22: Inbred line of *Globodera rostochiensis* derived from a pathotype Ro5 population.; HYP: Hyper-variable apoplastic effector family.; HR: Hypersensitive Response.; HMM: Hidden Markov Model.; J2: A second stage juvenile.; J3: A third stage juvenile.; Kb: Kilo bases (1,000 bases).; PTI: Pathogen Triggered Immunity.; Mb: Mega bases (1 milion bases).; PCN: Potato Cyst Nematode; RBP-1: Retinol binding protein; SNP: Single Nucelotide Polymorphism; SP: Signal Peptide for secretion; SPRYSEC: Secreted proteins containing a SP1a/RYanodine receptor-like domain; SV: Structural Variant; VAL: Venom Allergen-Like protein

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07914-6.

Additional file 1: Figure S1 - BlobTools-based interrogation of genome assembly of Gr-Line 19 to verify for single-taxon origin of the original sequences. Panel A: Each Gr-Line19 scaffold is represented by a single filled circle. Each scaffold is positioned in the main panel based on its GC proportion (x-axis) and coverage by reads from PacBio sequences (y-axis). On the top right the colours of the individual blobs are linked to their taxonomic origin. At the bottom of the main Blobtool figure, the size of the circles is linked to scaffold size. Panel B: on the left the % of unmapped *versus* mapped Gr-Line19 PacBio reads are presented, and the right the taxonomic origin of the reads. **Figure S2** - BlobTools-based interrogation of genome assembly of Gr-Line 22 to verify for single-taxon origin of the original sequences. Panel A: Each Gr-Line22 scaffold is represented by a single filled circle. Each scaffold is positioned in the main panel based on its GC proportion (x-axis) and coverage by reads from PacBio sequences (y-axis). On the top right the colours of the individual blobs are linked to their taxonomic origin. At the bottom of the main Blobtool figure, the size of the circles is linked to scaffold size. Panel B: on the left the % of unmapped *versus* mapped Gr-Line22 PacBio reads are presented, and the right the taxonomic origin of the reads. **Figure S3** - Multiple sequence alignment of Gr-CLE-1 protein sequences to verify the conservation of the CLE domain in putative CLE-1 genes. Each gene is represented by a gene identifier. Two genes are included for Gr-Line19 (Gros19_g2219.t1 and Gros19_g2223.t1) and two for Gr-Line22 (Gros22_g4869.t1 and Gros22_g4874.t1). The CLE1 sequence identified in *Heterodera glycines* (Q9BN21) is included as an outgroup.

Acknowledgements

We thank John T. Jones for sharing his expertise in the assembly of plant-parasitic nematode genome sequences.

Authors' contributions

SvdE performed the DNA extraction and library preparations. JvS, PT, and MH conceptualised the genome assembly pipeline. JvS generated the genome assemblies. JvS, MS and JH conceptualised the comparative genomics analyses. JvS performed the comparative genomics/effectoromics and phylogenetic analyses. JvS and JH wrote the manuscript. MS, AG and GS substantially revised/commented on the manuscript. The author(s) read and approved the final manuscript.

Funding

This work was funded as part of a grant by the Netherlands Organization for Scientific Research (NWO) as part of the Applied and Technical Science domain (TTW) under grant no. 14708. PT: bioinformatics and computational biology analyses were supported by the University of St Andrews Bioinformatics Unit (AMD3BIOINF), funded by Wellcome Trust ISSF award 105621/Z/14/Z.

Availability of data and materials

The Pacbio and Illumina Novaseq datasets for both lines that were used to generate the genome assemblies of Gr-Line19 and Gr-Line22 are available under NCBI BioProject accession PRJNA695196. Scripts used for the generation of genome assemblies, and resulting figures are available open-source under the MIT license on Github (https://github.com/jorisvansteenbrugge/GROS_genomes).

Reference *G. rostochiensis* transcriptome datasets that were analysed as part of this study are available under NCBI BioProject accessions PRJEB12075 and PRJNA274143. The genome assembly of the JHI-Ro1 G. rostochiensis population that was analysed as part of this study is available under NCBI BioProject accession PRJEB13504.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Laboratory of Nematology, Wageningen University & Research, Wageningen, The Netherlands. [2]Solynta, Dreijenlaan 2, 6703 HA Wageningen, The

Netherlands. [3]School of Medicine, Medical & Biological Sciences, University of St. Andrews, North Haugh, St Andrews, United Kingdom.

Received: 15 April 2021 Accepted: 27 July 2021
Published online: 11 August 2021

References

1. Nicol JM, Turner SJ, Coyne DL, den Nijs L, Hockland S, Maafi ZT: Current Nematode Threats to World Agriculture. In: Genomics and Molecular Genetics of Plant-Nematode Interactions. Edited by Jones J, Gheysen G, Fenoll C. Dordrecht (The Netherlands): Springer; 2011: 21–43.
2. Jones JT, Haegeman A, Danchin EGJ, Gaur HS, Helder J, Jones MGK, Kikuchi T, Manzanilla-L√≥pez R, Palomares-Rius JE, Wesemael WML, et al. Top 10 plant-parasitic nematodes in molecular plant pathology. Molecular Plant Pathology. 2013;14(9):946–61.
3. Plantard O, Picard D, Valette S, Scurrah M, Grenier E, Mugniéry D. Origin and genetic diversity of Western European populations of the potato cyst nematode (Globodera pallida) inferred from mitochondrial sequences and microsatellite loci. Mol Ecol. 2008;17(9):2208–18.
4. Montarry J, Bardou-Valette S, Mabon R, Jan PL, Fournet S, Grenier E, Petit EJ: Exploring the causes of small effective population sizes in cyst nematodes using artificial Globodera pallida populations. Proceedings of the Royal Society B: Biological Sciences 2019, 286(1894).
5. Jan PL, Gracianne C, Fournet S, Olivier E, Arnaud JF, Porte C, Bardou-Valette S, Denis MC, Petit EJ. Temporal sampling helps unravel the genetic structure of naturally occurring populations of a phytoparasitic nematode. 1. Insights from the estimation of effective population sizes. Evol Appl. 2016;9(3):489–501.
6. Fuller JM, Howard HW. Breeding for resistance to the white potato cyst-nematode, Heterodera pallida. Ann Appl Biol. 1974;77(2):121–112.
7. Toxopeus HJ, Huijsman CA. Genotypical background of resistance to Heterodera rostochiensis in Solanum tuberosum, var. andigenum [1]. Nature. 1952;170(4337):1016.
8. Kort J, Ross H, Rumpenhorst HJ, Stone AR. An International Scheme for Identifying and Classifying Pathotypes of Potato Cyst-Nematodes Globodera rostochiensis and G. pallida. Nematologica. 1977;23(3):333–9.
9. Phillips MS, Trudgill DL. Variations in the ability of Globodera pallida to produce females on potato clones bred from Solanum vernei or S. tuberosum ssp. andigena CPC 2802. Nematologica. 1983;29(2):217–26.
10. Mimee B, Duceppe MO, Véronneau PY, Lafond-Lapalme J, Jean M, Belzile F, Bélair G. A new method for studying population genetics of cyst nematodes based on Pool-Seq and genomewide allele frequency analysis. Molecular Ecology Resources. 2015;15(6):1356–65.
11. Thevenoux R, Folcher L, Esquibet M, Fouville D, Montarry J, Grenier E. The hidden diversity of the potato cyst nematode *Globodera pallida* in the south of Peru. Evol Appl. 2020;13(4):727–37.
12. Ali S, Magne M, Chen S, Obradovic N, Jamshaid L, Wang X, Bélair G, Moffett P. Analysis of Globodera rostochiensis effectors reveals conserved functions of SPRYSEC proteins in suppressing and eliciting plant immune responses. Front Plant Sci. 2015;6(AUG):623.
13. Diaz-Granados A, Petrescu AJ, Goverse A, Smant G. SPRYSEC Effectors: A Versatile Protein-Binding Platform to Disrupt Plant Innate Immunity. Front Plant Sci. 2016;7:1575.
14. Sacco MA, Koropacka K, Grenier E, Jaubert MJ, Blanchard A, Goverse A, Smant G, Moffett P. The cyst nematode SPRYSEC protein RBP-1 elicits Gpa2- and RanGAP2-dependent plant cell death. PLoS Pathog. 2009;5(8):e1000564.
15. Lozano-Torres JL, Wilbers RHP, Gawronski P, Boshoven JC, Finkers-Tomczak A, Cordewener JHG, America AHP, Overmars HA, Van 't Klooster JW, Baranowski L, et al. Dual disease resistance mediated by the immune receptor Cf-2 in tomato requires a common virulence target of a fungus and a nematode. Proc Natl Acad Sci USA. 2012;109(25):10119–24.
16. Eves-van den Akker S, Laetsch DR, Thorpe P, Lilley CJ, Danchin EGJ, Da Rocha M, Rancurel C, Holroyd NE, Cotton JA, Szitenberg A, et al. The genome of the yellow potato cyst nematode, *Globodera rostochiensis*, reveals insights into the basis of parasitism and virulence. Genome Biol. 2016;17(1):124.
17. Cotton JA, Lilley CJ, Jones LM, Kikuchi T, Reid AJ, Thorpe P, Tsai IJ, Beasley H, Blok V, Cock PJA, et al. The genome and life-stage specific transcriptomes of *Globodera pallida* elucidate key aspects of plant parasitism by a cyst nematode. Genome biology. 2014;15(3):R43.
18. Rouppe van der Voort JNAM, Van Enckevort ELJG, Pijnacker LP, Helder J, Gommers FJ, Bakker J. Chromosome number of the potato cyst nematode *Globodera rostochiensis*. Revue de Nématologie. 1996;19(4):369–74.
19. Janssen R, Bakker J, Gommers FJ. Selection of virulent and avirulent lines of *Globodera rostochiensis* for the H1 resistance gene in *Solanum tuberosum* ssp. *andigena* CPC 1673. Revue de Nématologie. 1990;13:265–8.
20. Smant G, Stokkermans J, Yan YT, de Boer JM, Baum TJ, Wang XH, Hussey RS, Gommers FJ, Henrissat B, Davis EL, et al. Endogenous cellulases in animals: Isolation of beta-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. Proc Nat Acad Sciences of the United States of America. 1998;95(9):4906–11.
21. Mitreva-Dautova M, Roze E, Overmars H, De Graaff L, Schots A, Helder J, Goverse A, Bakker J, Smant G. A symbiont-independent endo-1,4-beta-xylanase from the plant-parasitic nematode *Meloidogyne incognita*. Mol Plant Microbe Interact. 2006;19(5):521–9.
22. Popeijus H, Overmars H, Jones J, Blok V, Goverse A, Helder J, Schots A, Bakker J, Smant G. Enzymology: Degradation of plant cell walls by a nematode. Nature. 2000;406(6791):36–7.
23. Hewezi T, Howe P, Maier TR, Hussey RS, Mitchum MG, Davis EL, Baum TJ. Cellulose binding protein from the parasitic nematode *Heterodera schachtii* interacts with Arabidopsis pectin methylesterase: Cooperative cell wall modification during parasitism. Plant Cell. 2008;20(11):3080–93.
24. Noon JB, Hewezi T, Maier TR, Simmons C, Wei JZ, Wu G, Llaca V, Deschamps S, Davis EL, Mitchum MG, et al. Eighteen new candidate effectors of the phytonematode *Heterodera glycines* produced specifically in the secretory esophageal gland cells during parasitism. Phytopathology. 2015;105(10):1362–72.
25. Wilbers RHP, Schneiter R, Holterman MHM, Drurey C, Smant G, Asojo OA, Maizels RM, Lozano-Torres JL. Secreted venom allergen-like proteins of helminths: Conserved modulators of host responses in animals and plants. PLoS Pathog. 2018;14(10):e1007300.
26. Rehman S, Postma W, Tytgat T, Prins P, Qin L, Overmars H, Vossen J, Spiridon LN, Petrescu AJ, Goverse A, et al. A secreted SPRY domain-containing protein (SPRYSEC) from the plant-parasitic nematode *Globodera rostochiensis* interacts with a CC-NB-LRR protein from a susceptible tomato. Mol Plant Microbe Interact. 2009;22(3):330–40.
27. Lilley CJ, Maqbool A, Wu D, Yusup HB, Jones LM, Birch PRJ, Banfield MJ, Urwin PE. Eves-van den Akker S: Effector gene birth in plant parasitic nematodes: Neofunctionalization of a housekeeping glutathione synthetase gene. PLoS Genet. 2018;14(4):e1007310.
28. Lu SW, Chen S, Wang J, Yu H, Chronis D, Mitchum MG, Wang X. Structural and functional diversity of CLAVATA3/ESR (CLE)-like genes from the potato cyst nematode *Globodera rostochiensis*. Mol Plant Microbe Interact. 2009;22(9):1128–42.
29. Finkers-Tomczak A: Co-evolution between Globodera rostochiensis and potato driving sequence diversity of NB-LRR resistance loci and nematode suppressors of plant immunity. Wageningen University, PhD thesis; 2011.
30. Hu Y, You J, Li C, Pan F, Wang C. The *Heterodera glycines* effector Hg16B09 is required for nematode parasitism and suppresses plant defense response. *Plant Science* 2019, 289.
31. Pogorelko G, Wang J, Juvale PS, Mitchum MG, Baum TJ. Screening soybean cyst nematode effectors for their ability to suppress plant immunity. Molecular Plant Pathology. 2020;21(9):1240–7.
32. Eves-van den Akker S, Lilley CJ, Jones JT, Urwin PE. Identification and Characterisation of a Hyper-Variable Apoplastic Effector Gene Family of the Potato Cyst Nematodes. PLoS Pathog. 2014;10(9):e1004391.
33. Mitchum MG, Wang X, Wang J, Davis EL. Role of nematode peptides and other small molecules in plant parasitism. Annual Review of Phytopathology vol. 2012;50:175–95.
34. Gao B, Allen R, Maier T, Davis EL, Baum TJ, Hussey RS. The parasitome of the phytonematode *Heterodera glycines*. Mol Plant Microbe Interact. 2003;16(8):720–6.
35. Wang G, Peng D, Gao B, Huang W, Kong L, Long H, Peng H, Jian H. Comparative transcriptome analysis of two races of *Heterodera glycines* at different developmental stages. PLoS ONE. 2014;9(3):e91634.
36. Yang S, Pan L, Chen Y, Yang D, Liu Q, Jian H. *Heterodera avenae* GLAND5 Effector Interacts With Pyruvate Dehydrogenase Subunit of Plant to Promote Nematode Parasitism. Front Microbiol. 2019;10:1241.
37. Postma WJ, Slootweg EJ, Rehman S, Finkers-Tomczak A, Tytgat TOG, van Gelderen K, Lozano-Torres JL, Roosien J, Pomp R, van Schaik C, et al. The effector SPRYSEC-19 of *Globodera rostochiensis* suppresses CC-NB-LRR-mediated disease resistance in plants. Plant Physiol. 2012;160(2):944–54.

38. Wang J, Dhroso A, Liu X, Baum TJ, Hussey RS, Davis EL, Wang X, Korkin D, Mitchum MG. Phytonematode peptide effectors exploit a host post-translational trafficking mechanism to the ER using a novel translocation signal. New Phytol. 2021;229:563–74.
39. Gheysen G, Mitchum MG. Phytoparasitic nematode control of plant hormone pathways. Plant Physiol. 2019;179(4):1212–26.
40. Grisi E, Burrows PR, Perry RN, Hominick WM: The genome size and chromosome complement of the potato cyst nematode *Globodera pallida*. Fundam Appl Nematol 1995, 18:67–70.
41. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):460.
42. Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, May G, Lin H. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nat Commun. 2018;9(1):4844.
43. Field MA, Rosen BD, Dudchenko O, Chan EKF, Minoche AE, Edwards RJ, Barton K, Lyons RJ, Tuipulotu DE, Hayes VM *et al*: Canfam-GSD: De novo chromosome-length genome assembly of the German Shepherd Dog (Canis lupus familiaris) using a combination of long reads, optical mapping, and Hi-C. *GigaScience* 2020, 9(4):giaa027.
44. Ste-Croix DT, Gendron St-Marseille A-F, Lord E, Bélanger RR, Brodeur J, Mimee B. Genomic Profiling of Virulence in the Soybean Cyst Nematode Using Single-Nematode Sequencing. Phytopathology. 2021;111(1):137–48.
45. Bekal S, Craig JP, Hudson ME, Niblack TL, Domier LL, Lambert KN. Genomic DNA sequence comparison between two inbred soybean cyst nematode biotypes facilitated by massively parallel 454 micro-bead sequencing. Mol Genet Genomics. 2008;279(5):535–43.
46. Niblack TL, Arelli PR, Noel GR, Opperman CH, Orf JH, Schmitt DP, Shannon JG, Tylka GL. A revised classification scheme for genetically diverse populations of *Heterodera glycines*. Journal of Nematology. 2002;34(4):279–88.
47. Brynildsrud O, Gulla S, Feil EJ, Nørstebø SF, Rhodes LD. Identifying copy number variation of the dominant virulence factors msa and p22 within genomes of the fish pathogen *Renibacterium salmoninarum*. Microbial genomics. 2016;2(4):e000055.
48. Zhao S, Gibbons JG. A population genomic characterization of copy number variation in the opportunistic fungal pathogen *Aspergillus fumigatus*. PLoS ONE. 2018;13(8):e0201611.
49. Castagnone-Sereno P, Mulet K, Danchin EGJ, Koutsovoulos GD, Karaulic M, Da Rocha M, Bailly-Bechet M, Pratx L, Perfus-Barbeoch L, Abad P. Gene copy number variations as signatures of adaptive evolution in the parthenogenetic, plant-parasitic nematode *Meloidogyne incognita*. Mol Ecol. 2019;28(10):2559–72.
50. Folkertsma RT, Rouppe Van Der Voort JNAM, De Groot KE, Van Zandvoort PM, Schots A, Gommers FJ, Helder J, Bakker J. Gene pool similarities of potato cyst nematode populations assessed by AFLP analysis. Mol Plant Microbe Interact. 1996;9(1):47–54.
51. Vos P, Hogers R, Bleeker M, Reijans M, Vandelee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, et al. AFLP - a New Technique for DNA-Fingerprinting. Nucleic Acids Res. 1995;23(21):4407–14.
52. Beckerson WC, Rodríguez De La Vega RC, Hartmann FE, Duhamel M, Giraud T, Perlin MH. Cause and effectors: Whole-genome comparisons reveal shared but rapidly evolving effector sets among host-specific plant-castrating fungi. mBio. 2019;10(6):e02391.
53. Qutob D, Tedman-Jones J, Dong S, Kuflu K, Pham H, Wang Y, Dou D, Kale SD, Arredondo FD, Tyler BM, et al. Copy number variation and transcriptional polymorphisms of Phytophthora sojae RXLR effector genes Avr1a and Avr3a. PLoS ONE. 2009;4(4).
54. Flier WG, Grünwald NJ, Kroon LPNM, Sturbaum AK, Van Den Bosch TBM, Garay-Serrano E, Lozoya-Saldaña H, Fry WE, Turkensteen LJ. The population structure of *Phytophthora infestans* from the Toluca Valley of central Mexico suggests genetic differentiation between populations from cultivated potato and wild Solanum spp. Phytopathology. 2003;93(4):382–90.
55. Talas F, McDonald BA. Genome-wide analysis of *Fusarium graminearum* field populations reveals hotspots of recombination. BMC Genom. 2015;16(1):996.
56. Vijayapalani P, Hewezi T, Pontvianne F, Baum TJ. An effector from the cyst nematode *Heterodera schachtii* derepresses host rRNA genes by altering histone acetylation. Plant Cell. 2018;30(11):2795–812.
57. Jenkins WR. A rapid centrifugal-flotation technique for separating nematodes from soil. Plant Disease Reporter. 1964;48(9):48.
58. Holterman M, van der Wurff A, van den Elsen S, van Megen H, Bongers T, Holovachov O, Bakker J, Helder J. Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades. Mol Biol Evol. 2006;23(9):1792–800.
59. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.
60. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17(2):155–8.
61. Seppey M, Manni M, Zdobnov EM: BUSCO: Assessing genome assembly and annotation completeness. In: *Methods in Molecular Biology.* vol. 1962; 2019: 227–245.
62. Laetsch DR, Blaxter ML. BlobTools: Interrogation of genome assemblies. F1000Research. 2017;6:1287.
63. Boetzer M, Pirovano W. SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics. 2014;15(1):211.
64. van Steenbrugge JJM: Jorisvansteenbrugge/GapFiller: GROS Assembly version. Zenodo 2021.
65. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE. 2014;9(11):e112963.
66. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP + and AUGUSTUS supported by a protein database. NAR Genomics Bioinformatics. 2021;3(1):108.
67. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15.
68. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: An information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–45.
69. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
70. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.
71. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15(6):461–8.
72. Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 2004;338(5):1027–36.
73. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013;41(12):e121.
74. Mistry J, Finn R: Pfam: A domain-centric method for analyzing proteins and proteomes. In: Methods in Molecular Biology. vol. 396; 2007: 43–58.
75. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
76. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. Mol Biol Evol. 2020;37(1):291–4.
77. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.

Publisher's Note