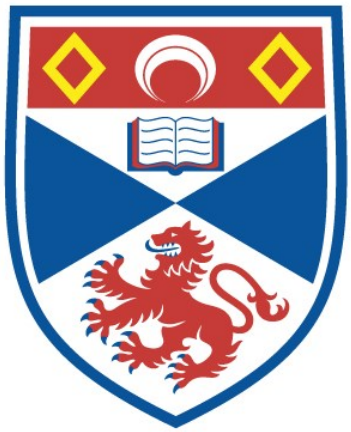
Applications of next-generation sequencing towards identifying novel disease resistance genes

Moray Smith

A thesis submitted for the degree of PhD
at the
University of St Andrews



2025

Full metadata for this thesis is available in
St Andrews Research Repository
at:
https://research-repository.st-andrews.ac.uk/

Identifier to use to cite or link to this thesis:
DOI: https://doi.org/10.17630/sta/1184

This item is protected by original copyright

This item is licensed under a
Creative Commons Licence
https://creativecommons.org/licenses/by-nc/4.0/

Declarations

Candidate's declaration

I, Moray Smith, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 50,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I was admitted as a research student at the University of St Andrews in September 2020. I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date 05/12/24 Signature of Candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date 05/12/24 Signature of Supervisor

Date 05/12/24 Signature of Supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Moray Smith, confirm that my thesis does not contain any third-party material that requires copyright clearance. The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 05/12/24 Signature of Candidate

Date 05/12/24 Signature of Supervisor

Date 05/12/24 Signature of Supervisor

Underpinning Research Data or Digital Outputs

Candidate's Declaration

I, Moray Smith, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 05/12/24 Signature of Candidate

I also called on Coates.
He was afraid he had mislaid her notes.
He took his article from a steel file:
“It’s accurate. I have not changed her style.
There’s one misprint - not that it matters much:
Mountain, not *fountain*. The majestic touch.”

- Vladimir Nabokov, *Pale Fire*

Acknowledgments

I write these acknowledgments on a Friday afternoon whilst I am feeling quite "done" with staring at my thesis. They are rushed, and can in no way fully capture the gratitude I feel to those who have supported me throughout this process.

Conducting this PhD has been a pleasure, largely due to the support of my supervisors Prof. Ingo Hein, for your limitless supply of ideas and useful suggestions, and Prof. John Jones, for your enthusiastic support and for introducing me to the eccentric nematology community.

I'd like to thank my colleagues who contributed directly to this thesis: Thomas Adams for his support in developing the `HISS` workflow, and Amanpreet Kaur for her excellent RNA-seq libraries for the *S. verrucosum* genome. Thank you also to all my colleagues and friends at the James Hutton Institute for making the last four years so enjoyable.

To Iona - my *Queen of Argyll* - words cannot express how you have supported me throughout this period of my life. We married at Fingask Castle on the 18th of March 2023, just a stone's throw from where I am sitting now. Thank you for enduring my rants about potatoes/sequencing/whatever, long may they continue.

To my father Callum, my mother Liz, and my brother Cameron - I dedicate this thesis to you all.

This work was supported through the East of Scotland Bioscience Doctoral Training Partnership (EASTBIO DTP), funded by the BBSRC award BB/T00875X/1.

Abstract

Improving disease resistance is a fundamental goal of plant breeding. The identification of novel resistance genes is a key step towards developing resistant potato varieties.

Nucleotide-binding site leucine-rich repeat (NLR) genes are a common class of resistance genes in plants. Their conserved and modular structure makes them ideal for automated identification in plant genomes through computational methods. Here, the novel NLR annotation program `Resistify` is presented, which is rapid, easy-to-use, and is the most sensitive NLR annotator to date. By applying `Resistify` to a Solanaceae pangenome, its performance is demonstrated, and a previously undescribed association between NLRs and *Helitron* transposable elements is revealed.

Wild potato genomes are a valuable source of novel resistance genes. The wild species *Solanum verrucosum* contains the *Rpi-ver1* gene which confers resistance to the late-blight pathogen *Phytophthora infestans*. Through a combination of HiFi and Nanopore sequencing, the genome of *S. verrucosum* is assembled, and the complete identity of the *Rpi-ver1* locus is resolved. Candidate genes within the locus are revealed including a severed NLR and a Jacalin-like lectin gene. *S. verrucosum* has the unusual and advantageous trait of being self-compatible - the *S-RNase* gene which imparts self-compatibility is identified within the genome. The genome of *S. verrucosum* also gives new insights into the centromeres of potato chromosomes. Some are formed of large tandem repeats which show evidence of being derived from transposable elements, whilst others are entirely repeatless and rich in transposable elements of the Tekay and CRM subfamilies.

In Scotland, the recent encroachment of the potato cyst nematode *Globodera pallida* has become an existential threat to the potato industry. The *G. pallida* resistance genes *H2* and *H3* are valuable sources of resistance, but their identity is unknown. Through a combination of RenSeq sequencing and association genetics, candidates of the *H2* and *H3* genes are identified, which are determined to be homologs of the *NRC3* and *R2* NLRs.

Contents

List of Figures

List of Tables

General introduction

Potato is the most important non-cereal food crop in the world. In 2022, 470.4 million tonnes of potatoes were produced globally, with China contributing 20.3%, India 11.9%, and Ukraine 4.4% - highlighting its ability to thrive in diverse regions and contribute to global food security (FAOSTAT 2023). They are a valuable source of essential vitamins and minerals including vitamin C, potassium, and vitamin B6, contributing to their importance as a staple food crop (Beals 2019). The cultural impact of potato must also not be understated, as they spawned a great number of traditional dishes in the countries of each port they landed in during their journey across the Atlantic, from *Pappas arrugadas* of the Canary Islands to the *Tattie scones* of Scotland. The increased agricultural productivity that occurred as a result of adoption of potatoes in Europe in the late 17th century led to an enduring reduction in civil conflict (Iyigun, Nunn, and Qian 2017). Consequently, the potato industry is valuable at a global and local level - production in Scotland contributes an estimated £507 million to the economy (Thomson 2024).

Pathogens and disease inflict substantial damage on global food systems and result in an estimated 17.2% yield loss for potato worldwide (Savary et al. 2019). Potatoes are vulnerable to a broad spectrum of pathogens including bacteria, fungi, viruses, insects, nematodes, and parasitic plants. The late blight oomycete *Phytophthora infestans* was a causal agent of a series of famines across Europe in the 1800s, most notably the Great Famine of Ireland which led to widespread death and population displacement. Although the availability of disease-resistant potato varieties has improved substantially from the then-popular but highly susceptible Lumper variety that drove such famines, pathogens such as *P. infestans* continue to be a significant threat to production worldwide. The burden of disease has had other impacts on potato - the value of seed potato production in Scotland is principally due to the climate which suppresses the spread of aphid, which act as vectors for viruses such as Potato Virus Y.

In Scotland, the potato cyst nematode (PCN) has emerged as an existential threat to the seed potato industry - without intervention, the industry will collapse by 2050 (Blok et al. 2020). This is an issue of phytosanitary health rather than yield loss, as seed potatoes grown in contaminated soil will carry and spread PCN to any field that they are subsequently grown in. There are strategies available to reduce PCN populations, but their effectiveness and long-term sustainability are in doubt. Fosthiazate is the only available nematicide in the United Kingdom following the consistent withdrawal of all other alternatives from the market, citing negative impacts on human and environmental health. Other strategies, such as crop rotations and trap cropping, are

available but present significant challenges. Crop rotations are relatively ineffective for controlling PCN, as the pest can persist in the soil for decades. Trap cropping requires specialist knowledge and management, and is difficult to implement effectively. Many of the species used for trap cropping do not thrive in Scottish conditions. Both methods also require avoiding the cultivation of seed potatoes, which is economically unacceptable.

The most effective option to control PCN - and other pathogens - in Scotland is to grow potatoes that possess disease resistance genes. Being a genetic trait, disease resistance does not require the expensive and potentially harmful application of pesticides, and does not require specialist per-field management to be utilised. Disease resistance does not reduce yield loss, but it can be used to directly manage pathogen populations. For PCN, juvenile nematodes will hatch and invade the crop but face an immune response that starves them, resulting in a failure to reproduce. Genetic disease resistance has a history of success in Scotland. The PCN resistance gene *H1* confers resistance to *G. rostochiensis* has seen widespread deployment for more than 50 years since its identification from a screen of more than 1,200 wild potato accessions [gartner_resisting_2021]. However, another PCN species, *G. pallida*, that is not affected by *H1* resistance has emerged since the widespread deployment of varieties containing *H1* and now threatens the Scottish seed potato industry (fig. 1).

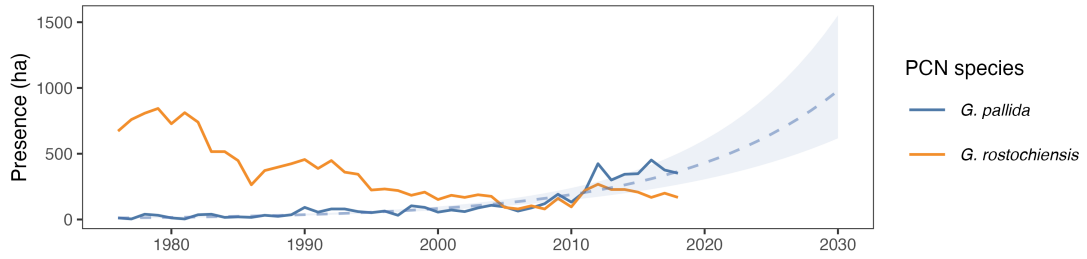


Figure 1: **Area testing positive for PCN in Angus.** Boundaries indicate the 95% confidence interval of a linear model. Data taken from Blok et al. (2020)

The development of new disease-resistant varieties is a key aim of international potato breeding programmes. In Scotland, identifying novel disease resistance genes effective against *G. pallida* and other pathogens is key to ensuring a sustainable future for the local potato industry. The advent of genomics has been critical in gaining an understanding of the form and function of the plant immune system, and has led to the development of tools which have enabled rapid characterising and breeding-led improvement of disease resistance.

The plant immune system

There are essentially three systems through which plant disease is manifested - immunity through physical barriers and pre-formed chemicals referred to as phytoanticipants, induced immunity through the recognition of molecular signatures of pathogen activity, and induced immunity through the recognition of pathogen effectors (fig. 2).

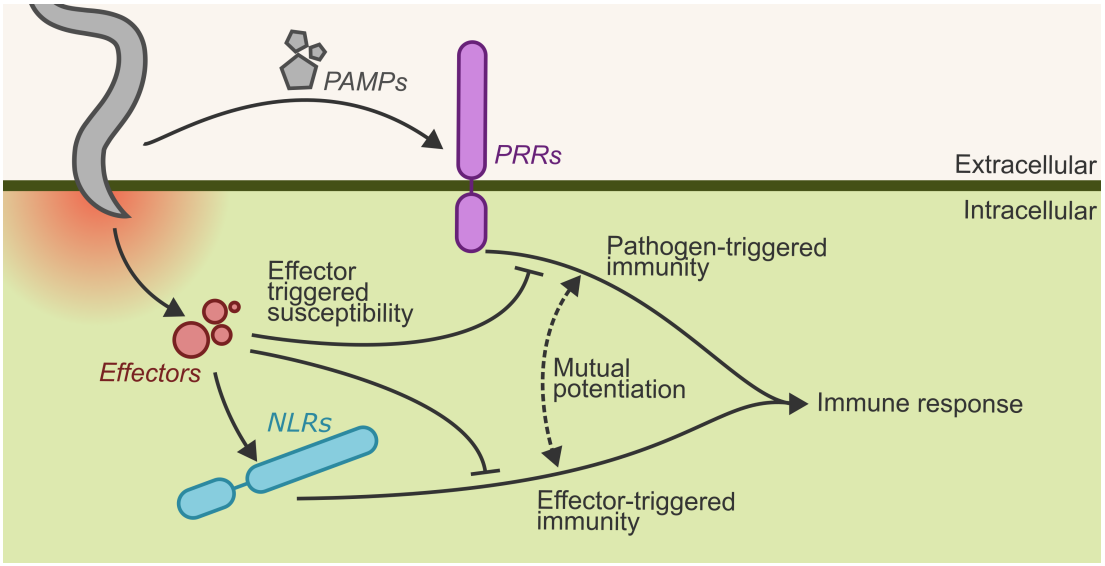


Figure 2: **The principal interactions of the plant immune system.** During pathogen invasion and infection, pathogens release pathogen-associated molecular patterns (PAMPs) that are detected by plant pattern-recognition receptors (PRRs). Pathogens release effectors into the host plant to suppress an immune response. Effectors can in turn be recognised by host nucleotide-binding leucine-rich repeat proteins (NLRs). Both PRRs and NLRs stimulate an immune response through different but complementary mechanisms.

Pathogen triggered immunity

In the process of establishing a site of infection, pathogens emit molecular signatures that give away their presence to the plant. One of the earliest and most well-characterised of these is flg22, an N-terminal fragment of the bacteria flagellin protein which is essential for movement (Sun et al. 2013). Treatment of plants with flg22 causes strong growth impairment of the plant due to activation of an immune response (Gómez-Gómez, Felix, and Boller 1999). The flg22 peptide is recognised directly by the LRR Receptor-like kinase (RLK) FLS2 of *A. thaliana* which activates the immune response through brassinosteroid signalling, a characteristic of which is impaired growth [Chinchilla et al. (2007);chinchilla_arabidopsis_2006]. Accordingly, flg22 is a pathogen-associated molecular pattern (PAMP) that initiates pathogen triggered immunity (PTI) through the pattern-recognition receptor (PRR) FLS2.

Additional PAMPs have been characterised including the wound-response hormone Systemin of Solanaceae which stimulates PTI through the PRR systemin receptor 160 (SR160). Raising a PTI response against PAMPs which are conserved and essential to the pathogen, such as flagellum proteins, is important for establishing a durable and broad-spectrum immune system. However, the ability to raise a PTI response to a PAMP can show variation within a genus - the csp22 peptide of bacterial cold-shock proteins is recognised by the CORE PRR in *Solanum lycopersicum* but not the wild tomato species *S. penellii* (L. Wang et al. 2016). This variation in PTI activity was leveraged to identify the causative PRR and serves to highlight that disease immunity is a genetic heritable system. That the transformation and expression of the CORE

PRR into the genetically distant *Arabidopsis* demonstrates that PRRs act as sensors that can "plug in" to a conserved immune network in the plant.

The functional origin of molecules that can be sensed as PAMPs in the plant immune system is diverse. Ascaroside pheromones are an essential component of nematode communication, and the most abundant ascaroside Ascr18 is detectable by the *Arabidopsis* RLK NEMATODE-INDUCED LRR-RLK1 (NILR1) (Huang et al. 2023). Pretreatment of *Arabidopsis* with Ascr18 primes the plant for resistance to *Pseudomonas* bacterium, and this effect is lost in mutants with defective NILR1 receptors (Mendy et al. 2017). It is not yet clear whether this is due to an ascaroside-like molecular being present in *Pseudomonas* or if NILR1 can function as a co-receptor for another PRR. Interestingly, the immune response to Ascr18 is multifaceted - *Arabidopsis* possesses enzymes that metabolise Ascr18 into a derivative pheromone that functions as a nematode repellent (Manohar et al. 2020).

PRRs may be either RLKs or RLPs which both have an extracellular domain that acts as the PAMP sensor, but differ in their intracellular domain which is a kinase in RLKs but not for RLPs. Hundreds of RLKs or RLPs may be present in a single genome - potato contains approximately 268 in total (Ngou, Ding, and Jones 2022).

Effectors drive susceptibility

The PTI response is complex and includes bursts of Ca^{2+} ion and reactive oxygen species, and the recruitment of protein kinases and other signal transducers to deliver a coordinated immune response (Bigeard, Colcombet, and Hirt 2015). Pathogens can suppress PTI through the delivery of effector proteins to the plant in a system called effector triggered susceptibility (ETS). For example, the *Pseudomonas syringae* effector HopM1 disrupts the function of the plant protein GRF8/AtMIN10, a signal transducer of PTI which in turn inhibits the reactive oxidative burst and physiological immune responses of the stomata (Lozano-Durán et al. 2014). Effectors have also been identified which can act upstream of PTI by soaking up PAMPs and preventing PRR activation. The *Cladosporium fulvum* effector Ecp6 is one example of this, which sequesters fungal chitin PAMPs away from the chitin-sensitive PRRs to prevent a PTI response (Jonge et al. 2010). The role of effectors is not exclusively to repress the immune response - they have diverse roles in establishing infection or maintaining parasitism. For example, plant parasitic nematodes deliver a cocktail of effectors to the plant which degrade cell walls, suppress the immune system, and reprogram the plant tissue into complex feeding sites such as the giant cells of the root knot nematode or the syncytia of the cyst nematodes (Khan and Khan 2021).

NLRs sense effectors

Recognising the presence or activity of pathogen effectors is thus critical towards maintaining an immune response following ETS. Plant nucleotide binding leucine rich repeat proteins (NLRs) represent the vast majority of a plant's effector sensing capability. There are several mechanisms through which NLRs can sense effectors. The most simple is to directly bind to the effector and elicit an immune response. The

S. chacoense NLR Rpi-chc1 and allelic variants of it directly recognise members of the *P. infestans* PexRD12/31 effector superfamily (Monino-Lopez et al. 2021). Directly targeting effectors is not a common mechanism for eliciting an immune response, as sensing can be compromised through mutations in the effector which maintain function but evade NLR binding. In addition, given the huge number of pathogens that infect plants and the many hundreds of effectors than can be present in any given pathogen species, direct recognition of effectors would require a huge number of NLRs.

A more effective mechanism is for NLRs to sense the activity of effectors indirectly. A classical example of this is the Pto/Prf complex in tomato (Ntoukakis et al. 2014). The NLR Prf forms a complex with the protein kinase Pto which mimics kinases that function downstream of PRR. When Pto is inadvertently modified by pathogen effectors seeking to disrupt PRR signal transduction, this change is sensed by Prf which subsequently elicits an immune response. It should also be noted that NLRs are not restricted to sensing effectors - the *S. lycopersicum* NLR Sw-5b binds directly to the tospovirus movement protein NSm - a PAMP - to deliver broad spectrum resistance (Zhu et al. 2017).

Beyond mimicry, NLRs can guard host proteins that are manipulated by effectors. Ran GTPase Activating Protein 2 (RanGAP2) is a regulator of nucleocytoplasmic trafficking and interacts directly with the NLR *Gpa2* which confers resistance to *G. pallida* (Sacco et al. 2009). Resistance is imparted through *Gpa2* by the sensing of the *G. pallida* effector RBP-1 which seeks to alter RanGAP2 kinetics to manipulate trafficking of cell components (Putker et al. 2024). Interestingly, RanGAP2 also interacts with the closely related NLR *Rx* which imparts resistance to potato virus X (PVX) through the recognition of a viral coat protein (Tameling and Baulcombe 2007).

NLRs have a conserved, modular protein domain structure. All canonical NLRs share a central and highly conserved NB-ARC domain and a downstream hypervariable LRR domain. The NB-ARC domain is an ADP/ATP exchange site and plays an essential role in activation and signal transduction (Maruta et al. 2022). The LRR domain maintains the NB-ARC domain in an inactivate ADP-bound state by wrapping across the nucleotide binding site. Direct or indirect effector recognition by the LRR domain releases it from this conformation leading to NLR activation (Maruta et al. 2022).

NLRs are further sorted into subclasses based on a variable N-terminal domain, being either coiled-coil (CC) NLRs (CNLs), Toll/ Interleukin-1 receptor/Resistance (TIR) NLRs (TNLs), or Resistance to Powdery Mildew 8 (RPW8) NLRs (RNLs). The role of these conserved domains varies - CC domains likely play a role in the formation of calcium efflux channels; the TIR domain has NADase activity that plays a key role in signal transduction; the less common RPW8 domain is associated with NLRs involved in signal transduction rather than pathogen sensing (J. Wang et al. 2019; Bernoux et al. 2011; Feehan et al. 2020).

NLRs are held in an inactive ADP-bound state until pathogen sensing triggers a conformational change and an ADP-ATP exchange. Once activated, NLRs form oligomers called 'resistosomes' that function as Ca^{2+} cell membrane channels that drive downstream defence responses. In Solanaceae, the NLR REQUIRED FOR CELL DEATH

(NRC) family of NLRs play an essential role in NLR signalling whereby activated sensor NLRs interact and activate the NRC, promoting it to oligomerize and accumulate at the cell membrane (Mauricio P. Contreras et al. 2022). In contrast to the sensor NLRs, NRCs exhibit limited diversity and play redundant roles in sensor NLR signal transduction.

NLRs are abundant and diverse

A single NLR can provide resistance against a single or multiple pathogens. To provide resistance against a diverse range of pathogens, plants have evolved large inventories of NLRs and can contain tens to thousands of NLRs within their genomes - potato contains around 700 (Barragan and Weigel 2021; Smith, Jones, and Hein 2024). NLR inventories are diverse, containing core families that are conserved across species and families that exhibit presence absence variation (Van de Weyer et al. 2019). The number of each subclass of NLR also varies between plant species - Solanaceae have an expanded inventory of CNLs, some members of the Magnoliid genus have completely lost their inventories of TNLs (Seo et al. 2016; Wu, Xue, and Van de Peer 2021).

The spatial distribution of NLRs within the genome also varies. NLRs show - at odds with other gene families - a tendency to form large gene clusters which diverge substantially between species, driven by gene duplication, unequal cross over, and the activity of transposable elements (Wersch and Li 2019). Although an exact biological explanation for this is not available, the divergence of clusters between species suggests that NLR clusters could act as evolutionary engines that facilitate the rapid turnover of novel and surplus NLRs. Helper NLRs are known to be embedded in NLR clusters with sensor NLR neighbours - the close genetic proximity could allow sharing of regulatory elements and potentially enhanced oligomerization.

Families of NLRs linked with direct recognition have been identified that are undergoing rapid diversification and interestingly are also highly expressed, lowly methylated, and are often in close proximity with transposable elements (Sutherland et al. 2024). High expression levels appears to be a key feature of pathogen-sensing NLRs - a feature exploited by the NLRseek platform which filters by expression to rapidly identify novel NLRs (Brabham et al. 2024).

Unifying plant immunity

An early model to describe the sequential layers of plant immunity was the zigzag model, which described the resistance mediated through PRRs, the susceptibility mediated through pathogen effectors, and the immunity restored by NLRs as three distinct phases of the immune response (J. D. G. Jones and Dangl 2006). Since the models inception it has become clear that the systems of PRRs and NLRs are not distinct and in fact function cooperatively to deliver immunity. PRRs are essential for NLR-mediated immunity, and NLRs potentiate PTI through upregulation of signalling components (Ngou et al. 2021; Yuan et al. 2021).

Solanum genomes

The potato genome was first resolved in 2011 by the Potato Genome Sequencing Consortium following the sequencing of the doubled monoploid *S. tuberosum* group Phureja DM1-3 516 R44 - hereafter referred to as DM (The Potato Genome Sequencing Consortium 2011). In total, 86% of the 844 Mbp genome was assembled into 12 chromosomes which contained 90.3% of the predicted genes in the DM assembly. Improved genetic mapping led the release of an improved DM assembly in 2013 with 93% of the assembly being placed into chromosomes which also increased the gene content to 96% (Sharma et al. 2013). The development of high accuracy long read sequencing and chromatin confirmation capture technologies contributed to the release of a further improved DM assembly in 2020 which led to a 595-fold increase in the size of contigs (assembled fragments without gaps) (Pham et al. 2020). Two years later saw the release of the first gapless assembly of DM which used ultra-long nanopore sequencing and gap-closing strategies to create a "telomere-to-telomere" assembly whereby each chromosome is represented by a single contig, providing full coverage across centromeres and other complex repetitive regions (X. Yang et al. 2023).

In parallel to recent improvements to the DM reference genome through long read sequencing and efficient assembly algorithms has been the release of high quality cultivated, landrace, and wild potato genomes. Hundreds of short-read sequenced potato genomes are now available which have facilitated broad studies on the genetic diversity of potato, and recent long-read sequencing potato pangenomes have enabled the study of large structural variations that influence tuberisation (Bozan et al. 2023; Tang et al. 2022). The accumulation of a diverse set of potato genomes is a key step towards the development of pangenomes which can be used to study valuable traits whilst accounting for the full genetic diversity of potato, rather than just a single reference genome (Shi et al. 2023).

A current challenge not exclusive to potato is to produce haplotype-phased assemblies of tetraploid genomes (Yibin Wang et al. 2023). Attempts to assembly tetraploid genomes are thwarted by problems such as multiple haplotypes being collapsed into a single assembly, high levels of heterozygosity leading to fragmented assemblies, and gaps leading to assemblies that represent a mosaic of different haplotypes. Recent successful strategies have included single-cell pollen sequencing and progeny sequencing of a selfing population - both methods rely on the segregating haplotypes in the gametes and progeny respectively as evidence to phase the assembly into four subgenomes (Bao et al. 2022; Serra Mari et al. 2024). Low-depth sequencing of the progeny from potato cultivar crossbreeding, which are routinely developed in potato breeding, has also been demonstrated to be an effective source of evidence for haplotype-phasing (Serra Mari et al. 2024). State-of-the-art nanopore sequencing strategies have proven effective at producing gap-free, haplotype-phased assemblies of human and plant genomes (Stanojević et al. 2024). As sequencing and assembly methods continue to improve assembly quality (fig. 3), it is likely that polyploid genome assemblies will also soon be achievable from sampling single genomes.

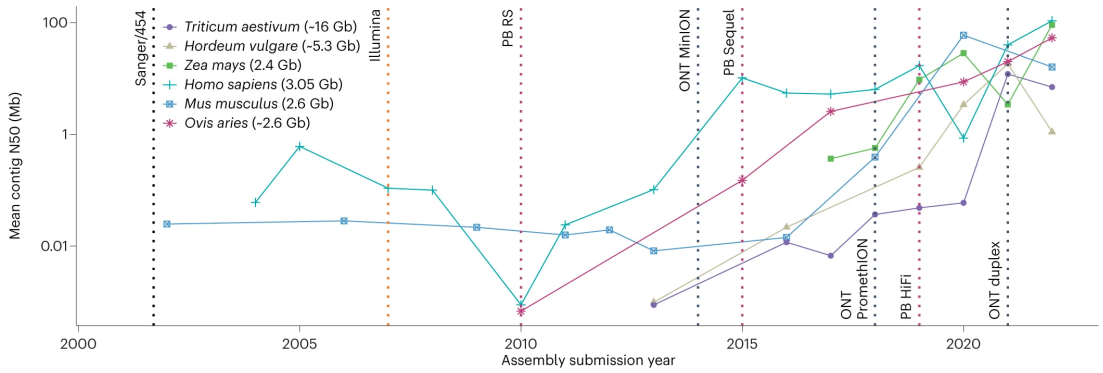


Figure 3: **Advancements in genomics through next-generation sequencing.** Contig N50 - that is, 50% of the genome being covered by contigs larger than the specified length - has improved as new sequencing technologies have been developed. Taken from Kovaka et al. (2023)

Disease resistance discovery through genomics

A conventional approach to identifying disease resistance genes involves using genetic markers on the progeny of a cross between susceptible and resistant cultivars to map genetic loci associated with resistance. In some cases, this has led to the identification of the exact gene imparting resistance, such as the potato cyst nematode resistance gene *Hero* (Ernst et al. 2002). The success of this approach generally depends on resistance being imparted by a single gene or genes in close genetic proximity and the extent to which fine mapping was conducted.

The release of the potato genome has enabled high-throughput, marker-agnostic genotyping approaches, such as QTL-mapping, which have been used to identify resistance loci effective against diseases such as early blight and late blight (Odilbekov et al. 2020; Lindqvist-Kreuze et al. 2014). While resolving the full genome is valuable for identifying resistance genes, the high cost of genome sequencing has proved to be a barrier towards capturing the full diversity of resistance genes in potato and other species.

To overcome this, one approach has been enrichment sequencing of NLRs using a technique called RenSeq (Jupe et al. 2013). Essentially, RenSeq and its derivate approaches use a bait library of oligos that share sequence similarity with NLRs to extract NLR-like sequences from genomic DNA, which can then be sequenced in isolation. RenSeq is a cost-effective strategy for sequencing NLRs and has seen success as a platform for conducting bulk segregant analysis and genome wide-association studies. These approaches have been used to map the *H2* potato cyst nematode resistance locus, to identify multiple novel late blight resistance genes in *Solanum americanum*, and also to identify novel PRRs using a RenSeq variant that targets PRR rather than NLR sequences (Lin et al. 2023, 2020; S. M. Strachan et al. 2019).

Diagnostic RenSeq (dRenSeq) is a derivative approach which allows users to identify known NLRs in any given genome, greatly accelerating breeding for disease resistance which previously relied on phenotype data or genetic markers (Armstrong et al. 2019).

In another approach called AgRenSeq, RenSeq can be combined with phenotype data to directly identify novel NLRs associated with an unknown source of disease resistance, an approach that has seen success in wheat and potato (Arora et al. 2019; Adams et al. 2023; Yuhan Wang et al. 2023). RenSeq has also seen improvements with the uptake of long-read sequencing which has enabled the *de novo* assembly of NLRs without the need for a reference genome, further accelerating resistance gene discovery (Witek et al. 2016). HiFi-RenSeq has been used to capture the species-wide diversity of NLRs in *Arabidopsis*, providing valuable insight into the diversification and evolutionary trajectory of NLR subfamilies - a study that would be prohibitively expensive through conventional genome assembly (Van de Weyer et al. 2019).

Disease resistance engineering

Identifying disease resistance loci and genes has been instrumental in the development of genetic markers that are used to efficiently introgress resistance into commercial cultivars. Introducing stacks of multiple resistance genes effective against multiple pathogens is an important step towards delivering disease resistance that is broad-spectrum and durable (Mundt 2018). Mobilising multiple genes from different sources into a single cultivar is laborious and time-consuming - high-quality markers accelerate this process by avoiding the need to phenotype breeding intermediates (Hafeez et al. 2021). Breeding to incorporate multiple resistance genes against a single pathogen is challenging due to each obscuring the others phenotype. Markers resolve this by confirming their presence without the reliance on phenotype data. Facilitating a more dynamic system of resistance breeding reduces the potential for resistance genes to be squandered by overuse leading to pathogen selection and "loss" of resistance genes.

Direct genetic modification is an attractive option for improving disease resistance through NLRs. A field trial of transgenic tomato expressing the pepper NLR *Bs2* provided consistent improvements in marketable yield in the absence of pesticides (Horvath et al. 2012). Recent field trials in the Netherlands and Ireland with transgenic potatoes encoding NLR stacks effective against late blight have also proven effective at reducing the need for expensive and harmful fungicides (Ortiz, Phelan, and Mullins 2016; Kessel et al. 2018). To date, NLR transgenesis has mostly been used in experimental approaches and has yet to see broad commercial application in Europe, largely due to restrictions and legislation. A change in attitude towards genetically modified crops will be necessary for the value of NLR engineering to be fully realised. The in-field performance of transgenic late blight resistant potato is striking, and has facilitated a direct change in consumer attitudes towards genetically modified crops through field visits (Bubolz et al. 2022).

As demonstrated with *Bs2*, transgenic NLRs may be effective in a broad range of genetic backgrounds, although this is reliant on the conservation of downstream signalling elements which the NLR interacts with, such as the helper NLR network of Solanaceae. In an alternative approach, pre-existing but non-functional NLRs may be "resurrected" to reestablish disease resistance, for example by reversing mutations in the helper NLR NRC2 that have led to pathogen insensitivity (Mauricio P. Contreras et al. 2023).

Recently, synthetic NLRs have been demonstrated as a sustainable and adaptable

alternative source of disease resistance. The CNLs Pik-1 and Pik-2 function as a pair whereby an integrated domain on Pik-1 interacts with a corresponding effector which subsequently leads to Pik-2 mediated immune signalling. Swapping out the integrated domain for a nanobody that binds to a different peptide can still result in an immune response (Kourelis et al. 2023). Given that large libraries of nanobodies can be rapidly generated against a specific source, this approach may enable tailor-made NLR resistance for a broad range of pathogens. Nanobodies are not strictly required in this approach - swapping the integrated domain of Pik-1 for a host protein targeted by *Magnaporthe oryzae* provided broad and durable resistance (Zdrzałek et al. 2024). The applicability of these approaches has yet to be fully demonstrated, however it is clear that identifying and characterising novel NLR-sensing mechanisms plays a key role in developing synthetic resistance systems.

Thesis aims

The identification of novel NLRs is critical to establishing effective and durable disease resistance in potato. They can be employed directly in cultivars to provide resistance, and can be valuable tools in furthering our understanding of NLR-mediated resistance and how pathogens trigger a response. As pathogen populations continue to adapt and new strains emerge, developing tools that enable the rapid identification of novel NLRs is key to ensuring that the arsenal of available NLRs continues to grow.

This thesis aims to develop tools that take advantage of recent advances in next-generation sequencing and genomics to rapidly and accurately identify NLRs from potato genomes, which should be easy-to-use and broad in their applicability so that they may be used by researchers with other crops. The genome of the wild potato *Solanum verrucosum* will be resolved with the aim of identifying a previously uncharacterised late blight resistance gene, and also to explore how next-generation sequencing and *Solanum* genomics can enhance our understanding of NLR diversity as well as other genetic features such as centromeres. Finally, this thesis aims to apply both next-generation sequencing and the development of improved bioinformatic workflows to a broad panel of wild and cultivated potato genomes with the aim of identifying novel potato cyst nematode resistance genes.

Automated NLR discovery

Introduction

As aforementioned, the structure of NLRs is conserved and modular, comprised of a central NB-ARC and downstream LRR domain, and an upstream domain which can be CC, TIR, or RPW8 fig. 4. As such, NLRs can be broken into three categories - CNL, TNL, and RNL respectively. Given that NLR domains are generally well-conserved across the plant kingdom, each can be identified computationally.

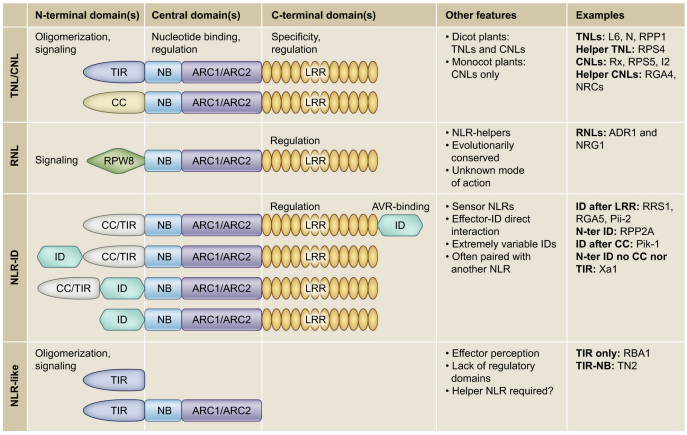


Figure 4: **The diversity of plant NLR domains.** The N-terminal of NLRs is diverse and broadly recognised into three categories - containing a coiled-coil (CC) domain (CNL), a toll/interleukin-1 receptor (TIR) domain (TNL), or a resistance to powdery mildew 8 (RPW8) domain (RNL). Downstream of the NB-ARC domain is the diverse leucine-rich repeat (LRR) domain. Integrated domains (ID) may also be present. Figure taken from Cesari (2018)

The NB-ARC domain of NLRs is subdivided into three conserved subdomains - the nucleotide-binding (NB) fold, and the downstream ARC1 and ARC2 subdomains which form four-helix and winged-helix folds (Ooijen et al. 2008). There are nine motifs associated with the NB-ARC domain - VG, P-loop (Walker A), RNBS-A, and Walker B are in the NBD subdomain; RNBS-B and RNBS-C are in the ARC1 subdomain; GLPL, RNBS-D, and MHD are in the ARC2 subdomain (Eliza C. Martin et al. 2022). Given the conservation of the NB-ARC domain, one common strategy to identify putative NLRs is to search for proteins that contain NB-ARC sequences. For this purpose, hidden Markov model (HMM) profiles are a popular method (Eddy 2009). HMM profiles

can be built from multiple sequence alignments of a given domain, allowing them to account for sequence diversity in a single search, unlike other 1:1 search strategies such as BLAST.

The LRR domain is defined as a hypervariable region interspersed with LxxLxL motifs (where "x" can be any amino acid) that are generally ~30aa apart (Kobe and Kajava 2001). The repeating motifs give rise to a solenoid protein structure - a distinctive feature of NLRs and RLK/RLPs - the concave surface of which tends to be involved in ligand binding (Padmanabhan, Cournoyer, and Dinesh-Kumar 2009). Annotating the LRR domain remains a challenge - HMM profile performance is poor given their diversity and fail to capture the individual tandem repeat subunits, and more sensitive machine-learning or structural methods are computationally taxing (Xu et al. 2023).

Both the TIR and RPW8 domain are relatively conserved, making them both suitable for representation with HMM profiles. Within the TIR domain are six conserved motifs - βA, αA, βC, αC, βD-αD1, and αD3 which are associated with the α-helix and β-strand folds of the domain (Toshchakov and Neuwald 2020). The CC domain is highly variable and can evade strategies that rely primarily upon HMM profiles. Instead, coiled-coil structures can be predicted through machine-learning approaches such as `DeepCoil`, and for plant NLRs through the presence of a conserved EDVID motif (Ludwiczak et al. 2019; Eliza C. Martin et al. 2022).

Beyond the core domains, additional sequence elements are emerging as indicators of NLR subclass. The MADA motif is a short N-terminal sequence that is a key indicator of the NRC helper-NLR family and is present in ~20% of CNLs (Adachi et al. 2019). The C-terminal jelly roll/Ig-like (C-JID) domain is a post-LRR structure common to dicotyledonous plant TNLs (R. Martin et al. 2020). Given their conservation, both domains can be represented by HMM profiles and serve as useful additional evidence for an NLRs classification.

Identifying and classifying NLRs according to their structure is a key step in identifying novel resistance genes and understanding their diversity in new genomes. To date, multiple tools have been developed to achieve this including `DRAGO`, `NLGenomeSweeper`, `NLR-Annotator`, `RGAugury`, `RRGPredictor`, and `NLRtracker` (tbl. 1).

Table 1: **Available NLR annotation software.** A summary of currently available NLR annotation tools. Adapted from Kourelis et al. (2021)

Tool	Input data	Dependencies	Distribution
`DRAGO2`	Protein, Transcript	None	Online only, API available
`NLGenomeSweeper`	Transcript, Genomic	InterProScan, MUSCLE, TransDecoder, samtools, bedtools, BLAST, HMMER, Python	GitHub, Manual dependency installation

Tool	Input data	Dependencies	Distribution
`NLR-Annotator`	Transcript, Genomic	Java, MEME	GitHub
`NLRexpress`	Protein	HMMER, Python	GitHub, Online or local, Conda environment provided
`NLR-Parser`	Genomic	Java, MEME	GitHub
`NLRtracker`	Protein, Transcript	InterProScan, HMMER, MEME, R	GitHub, Manual dependency installation
`RGAugury`	Protein	InterProScan, BLAST, HMMER, Java, PfamScan, Phobious, ncoils	Bitbucket, Online or local webservice, Docker container
`RRGPredictor`	Protein, Transcript	InterProScan, Perl	GitHub, Manual dependency installation

The earliest available tool was `NLR-Parser` which uses motif alignment and search tool (MAST) searches for 20 conserved protein motifs that were initially identified in potato but are broadly applicable to plant genomes (Steuernagel et al. 2015; Jupe et al. 2012). Based on the presence and absence of the NLR-associated motifs, `NLR-Parser` can determine which subfamily an NLR belongs to and can also make an assessment on whether the sequence represents a complete or a partial NLR. An extension of `NLR-Parser` is the tool `NLR-Annotator` which can identify NLR boundaries in genomic sequences by searching for the motifs in a 6-frame translation of the genome, followed by postprocessing to stitch open reading frames together into putative NLR loci (Steuernagel et al. 2020).

Beyond NLRs, other gene families that are also associated with disease resistance such as RLKs and RLPs can also be identified through searches for their associated protein signatures. To achieve this, `RGAugury` uses several programs - `InterProScan` to identify resistance-associated protein domains, `nCoils` to identify coiled-coil domains, `pfam_scan` to specifically identify the NB-ARC domain, and `Phobius` to identify transmembrane domains that are a key feature of RLKs and RLPs (P. Li et al. 2016). Programs were selected for their performance, for example `pfam_scan` was determined to outperform `InterProScan` in identifying the NB-ARC domain, and `nCoils` is specifically designed to identify the coiled-coil domain which as a diverse can evade sequence-based approaches. As both `InterProScan` and `nCoils` are computationally intensive, `RGAugury` initially performs a `BLASTp` search to remove sequences that are unlikely to be disease resistance genes.

`DRAGO` uses a similar approach, although instead of `InterProScan` it uses a custom set HMM models based on sequences in the complementary Plant Resistance Genes database (PRGdb), and uses `COILS` and `TMHMM` to identify coiled-coil and transmembrane

domains (Osuna-Cruz et al. 2018). Since its development, DRAGO has received a number of updates as the PRGdb has continued to expand which have improved its accuracy (Calle García et al. 2022). RRGPredictor is a simple set of Perl scripts that parse the output of InterProScan to identify NLRs (Santana Silva and Micheli 2020). NLGenomeSweeper also uses InterProScan, but searches through open reading frames identified in the genomic sequence, akin to NLR-Annotator (Toda et al. 2020).

The most recently developed tool is NLRtracker, which combines InterProScan and motif searches to identify and classify NLRs (Kourelis et al. 2021). Crucially, NLRtracker was developed in parallel with the RefPlantNLR database - a collection of experimentally validated NLRs from a diversity of genetic backgrounds. Importantly, the development of RefPlantNLR allowed a comprehensive benchmarking of the available tools for their performance in identifying and correctly classifying NLRs. Performance was varied - RRGPredictor had the lowest sensitivity, whilst DRAGO and RGAugury were high accuracy but suffered from a higher false positive rate. NLRtracker was the most performant tool according to benchmarking.

Recently, NLRexpress was released which identifies NLR-associated motifs but with a marked increase in accuracy over previous approaches, particularly in identifying LRR and coiled-coil motifs (Eliza C. Martin et al. 2022). The LRR domain is highly variable and differs in structure between RLK/RLPs and NLRs - identifying the LRR domain by analysing the distribution of motifs is preferable to other domain sequence similarity approaches. The machine-learning approach that NLRexpress takes is a good balance between performance and accuracy, and is relatively agnostic as to whether the repeats are from RLK/RLP or NLRs. Despite the good performance of NLRexpress, it has yet to be included in any tool for identifies and performs classification of NLRs.

Software accessibility

The availability and ease-of-installation of bioinformatics software is a key issue in providing accessible and reproducible science. A recent survey of computational tools revealed that only half of them are "easy to install" and a quarter are completely uninstallable (Mangul et al. 2019). Concerningly, 57% of software failed to be installed when the recommended installation instructions - when available - were followed explicitly. Key points that improved the usability of software included hosting software on repositories such as GitHub and GitLab, providing a proper installation interface such as Conda that automatically handle dependencies, and other sensible choices such as not hardcoding input/output file paths and designing the software to be platform agnostic.

One drawback with the currently available NLR annotators is the reliance on InterProScan as the backend for domain annotation. InterProScan is designed as a comprehensive and generalised domain annotation tool (P. Jones et al. 2014). As a result, each input sequence is scanned against several databases each containing in total more than 180,000 protein signatures, the vast majority of which are not NLR associated. Compounding this, NLRs only represent a fraction of a plant proteome resulting in unnecessary searches against non-NLR sequences. In addition, domains common to NLRs are present in InterProScan databases with differing

levels of curation which often results in overlapping or fragmented annotations. These must be parsed, particularly the LRR domain which is represented by multiple `InterProScan` signatures. The initial filtering that `RGAugury` takes to remove obvious non-candidates is a good example of how the search space can be easily reduced to improve performance. Other dependencies are also known to suffer from poor performance - `nCoils` and `Phobius` used by `RGAugury` do not support multithreading, and whilst `jackhmmer` that is used by `NLRexpress` does support multithreading, on modern disks it is generally limited to around 2-4 threads as I/O is a major bottleneck.

Another requirement for software is for it to be freely available and easy to install. The package manager conda is a gold standard for distributing bioinformatics software due to its widespread support on HPCs, lack of requirements for root access, and crucially its dependency manager which ensures that any dependencies that are also available through conda can be installed automatically. Containerisation is also become increasingly sought approach as it allows software to be run in a consistent computational environment, and ensures that differences in the underlying environment do not result in irreproducible errors (Di Tommaso et al. 2017).

Unfortunately, none of the programs are available on Conda, and only `RGAugury` is available through a docker container, although the large `InterProScan` databases must be downloaded manually. `DRAGO` is available only through an online portal and the source code is not available, making its usage dependent on a stable internet connection and continued support of the webtool. `NLRexpress` comes with a recommended Conda environment for running the software, but the code itself must be independently downloaded and databases set up.

Software that is readily available with minimum manual configuration is critical for projects that make use of workflow managers such as Snakemake or Nextflow (Di Tommaso et al. 2017; Mölder et al. 2021). Workflow managers are particularly valuable when multiple samples, as they support parallelisation, can be deployed on diverse HPC environments, can automatically download required software, and can robustly handle errors that might otherwise propagate through the analysis. For example, an ideal Nextflow process that identifies NLRs in a given genome using the tool `examplenlrtool` might be:

```
process IdentifyNlrs {
    container 'docker://quay.io/biocontainers/examplenlrtool:1.0'
    conda 'bioconda::examplenlrtool==1.0'
    input:
    path fasta
    output:
    path "results"
    script:
    """
    examplenlrtool --input ${fasta} --output results
    """
}
```

Here, `examplenlrtool v1.0` is strictly defined and is available as a Docker container

and through Conda. Accordingly, Nextflow will automatically download and deploy either the docker image or a conda environment containing the required program. If this process was part of a study involving many genomes for the purpose of analysing NLR diversity, this would ensure that each sample is processed identically and could be fully replicated by any reviewer or other third party. If the pipeline is to be shared with other users for their own analysis, this also ensures that the analysis is easy and replicable. If these practices are not followed, then setting up the process becomes more involved and inaccessible to inexperienced users.

Chapter aims

As genome sequencing continues to become more affordable and accessible, the increasing availability of genomes will be a valuable asset for identifying novel NLRs and characterising their diversity. Developing NLR annotation software that is easy-to-use, rapid, and accurate is a key step towards achieving these goals. The primary aims of this chapter are to:

- Develop a new NLR annotator that takes advantage of the latest approaches, is user friendly, and is faster than the currently available alternatives
- Apply the annotator to a pangenome of Solanaceae genomes to obtain a brief overview of the diversity of NLRs

Methods

The development of Resistify

`Resistify` was implemented in python as a command line executable program. It takes a user-provided `.fasta` file of protein sequences as input. These are pre-processed with `Biopython` to remove stop-codons at the end of sequences, and warn and exit if the user provides a file with internal stop-codons. Descriptions in entry headers are also removed. The cleaned `.fasta` file is moved to a temporary directory created by the program to reduce I/O pressure in certain HPC environments.

Its operation is separated into two stages. First, `Resistify` searches through the sequences with `hmmsearch` for a curated set Pfam entries (tbl. 2). The results table produced by `--domtblout` is parsed with `the Bio.SearchIO` module and by default E-value of 0.00001, with the exception of the RPW8 domain which is hard-filtered by a score of 20 based on previous studies. Domain annotations of the same type are merged if they overlap or are within 100 amino acids of each other. Preliminary testing in development showed that this is necessary to overcome NB-ARC domain annotations which can become split. Proteins are then initially classified as belonging to either a CN, RN, TN, or N classification. Proteins which do not have any evidence of an NB-ARC domain are discarded at this stage. A new `.fasta` file containing this subset of sequences is moved to the temporary directory.

Table 2: **HMM models included in the initial hmmsearch stage of Resistify.**

Domain	Source
NB-ARC	https://www.ebi.ac.uk/interpro/entry/pfam/PF00931/
TIR	https://www.ebi.ac.uk/interpro/entry/pfam/PF01582/
TIR	https://www.ebi.ac.uk/interpro/entry/pfam/PF13676/
RPW8	https://www.ebi.ac.uk/interpro/entry/pfam/PF05659/
CC	https://www.ebi.ac.uk/interpro/entry/pfam/PF18052/
C-JID	https://www.ebi.ac.uk/interpro/entry/pfam/PF20160/
MADA	https://cdn.elifesciences.org/articles/49956/elife-49956-supp2-v2.hmm

Following this, the filtered set of NB-ARC containing proteins is queried against a database of NLR sequences provided by `NLRexpress` using `jackhmmer` with the options `--noali -N 2 -E 1e-5 --domE 1e-5 --chkhmm`, as specified in the original `NLRexpress` software. As `jackhmmer` is severely I/O limited - the option is also provided of splitting the input `.fasta` into chunks of N sequences which are then queried with `jackhmmer` in parallel. The resulting files are then merged and parsed to create matrices of HMM emission probabilities. The resulting matrices are passed as input to each of the 17 multilayer perceptron network models provided by `NLRexpress`, which predict a probability for each NLR-associated motif at each valid position in the sequence.

Sequences are then reclassified using the following logic. If a protein belongs to class N (i.e., does not have any evidence of an upstream CC, RPW8, or TIR domain) it is scanned for upstream TIR or CC motifs. As motif searches can be more promiscuous, restricting motif searches to this condition prevents them from interfering with unambiguous NLR classifications. The sequence is then screened for LxLxxL motifs to define the LRR domain. Following a previous definition, an LRR domain is annotated if four or more LxLxxL motifs are identified with inter-motif gaps of less than 75 amino acids (Eliza C. Martin et al. 2023). Gaps larger than 75 amino acids are predicted to be a break in the LRR domain and so the LRR annotation process is restarted from that position onwards. If less than four motifs exist across the whole sequence this process is skipped.

This combined evidence is then integrated into the domain annotation data. The domains are sorted by start position and a "domain string" is formed. For example, if the sorted domains took the order TIR, NB-ARC, LRR, then the domain string would be TNL. Alternatively, a canonical CNL would take the form CNL. The domain strings are searched for substrings CNL, RNL, TNL, or NL and classified accordingly. A motif string is also formed representing the distribution of motifs in the sequence. The presence of MADA, MADA-like, and C-JID domains is also determined for each sequence.

The primary output of `Resistify` is a table detailing the NLRs identified and the specific classification for each sequence. The complete motif string, domain string, classification, NB-ARC motif count, and MADA, MADA-like, and CJID status are listed. A complete list of all annotations and NLRexpress motifs are given as separate

tables, with motif sequences provided. Additionally, all NB-ARC domain sequences are extracted and provided as a FASTA file for downstream phylogenetic analysis.

This strategy has several key advantages over previous methods. Firstly, the databases required for `hmmsearch` and `NLRexpress` are relatively small allowing them to be distributed alongside the code, enabling greater portability. As `hmmsearch` is the only external dependency - and is simple to install - this also improves portability. Secondly, `Resistify` benefits from an improved accuracy over previous tools due to the inclusion of `NLRexpress`, which has been previously demonstrated to be performant over alternative motif identifiers. Finally, the results of `Resistify` are generated to be easy to parse and interpret by inexperienced users, and outputs useful files for downstream analysis such as the NB-ARC domains.

Distribution of Resistify

`Resistify` was developed and is distributed on GitHub at https://github.com/SwiftSeal/resistify. `Resistify` is also available on the PyPi database at http://pypi.org/project/resistify/, which is used in a Bioconda distribution. All databases and models are distributed with the executable so that manual setup is not required.

RefPlantNLR benchmarking

Protein sequences of the RefPlantNLR database members were retrieved and used as input for `Resistify` with default settings (Kourelis et al. 2021) `Resistify` classifications were compared directly in `R v4.3.2` with `tidyverse v2.0.0`. Any sequence where the `Resistify` classification did not exactly match with the provided RefPlantNLR structure were considered as a potential misclassification and taken forward for manual inspection.

Araport11 benchmarking

The latest release of Araport11 representative gene model protein sequences were downloaded from TAIR and used as input for `Resistify` with default settings (Kourelis et al. 2021). A phylogenetic tree was built from the Resistify-extracted NB-ARC domain sequences with `mafft v7.52.0` and `fasttree v2.1.11` with default settings. The phylogenetic tree was visualised in `R v4.3.2` with `ggtree v3.19`.

Pangenome pipeline

A Snakemake workflow was developed to predict genes and NLRs in a *Solanum* pangenome comprised of chromosome-scale genomes. Genomes Snakemake was executed in a `mamba v1.4.2` environment with `snakemake v7.32.3` and `cookiecutter v1.7.3` (Mölder et al. 2021). Genes were predicted de-novo from the genome sequence alone using `Helixer v0.3.2` with the `land_plant_v0.3_a_0080.h5` model (Holst et al. 2023). Protein sequences were extracted using `AGAT v1.2.0` and used as input for `Resistify` with default settings (Dainat et al. 2023). Homolog analysis was carried out using `OrthoFinder v2.5.5` using all predicted protein sequences (Emms and

Kelly 2019). Transposable elements were annotated with the latest GitHub release of EDTA (Ou et al. 2019).

Gene models were translated to bed format using `AGAT v1.2.0` and overlaps with intact transposable elements were identified using `bedtools v2.31.1` with the command `bedtools intersect` using the options `-f 0.9 -wo` (Quinlan and Hall 2010). To identify previously characterised NLRs, protein sequences of a subset of *Solanum*-originating NLRs in the RefPlantNLR database were queried against each genome with `blastp v2.15.0` (Camacho et al. 2009). The full pipeline and all post-hoc analysis are available at https://github.com/SwiftSeal/pangenomics/.

Results

Evaluating available NLR annotators

Available NLR annotation tools have been evaluated extensively in the previous RefPlantNLR study. As an additional evaluation metric, tools were evaluated on their accessibility and general usability. To assess their performance, tools were applied to the RefPlantNLR database represents the best currently available curated source of annotated NLRs, and the *Arabidopsis* Araport11 protein assembly. The evaluation was focused on tools that identify NLRs in pre-defined gene sequences - `NLGenomeSweeper` and `NLR-Annotator` were therefore excluded as they identify NLR loci in genomic sequence.

Downloading `DRAGO` is simple as it only requires a single bash script available on GitHub that makes an API request. It is rapid when executed against the RefPlantNLR database, returning results in under one minute. The output of `DRAGO` is a tab-delimited table with columns containing the sequence ID, sequence classification, domain classification, and start and end coordinates of the domain. Row represents each identified domain per sequence. As a result, parsing the output is therefore straightforward, although additional statistics such as domain E-values would be useful in assessing the quality and confidence of domain annotations. Unfortunately, `DRAGO` failed on multiple attempts against the Araport11 annotation, producing a 404 "Not found" error. It is unclear if this is due to a limitation on input size, or if the service was experiencing internal errors. Integrating `DRAGO` into a pipeline that handles a large number of genomes would be possible, but would likely require the development of a more rigorous script for handling API requests to ensure that erroneous results are correctly handled. From the documentation it is also unclear if there are any request limits for the provided service.

`RGAugury` was evaluated. It was noted that the web version was no longer functional and pointed to a Bitbucket "Resource not available" web-page. Following the provided installation instructions, use of a provided Docker container is advised to avoid having to manually install all the required dependencies. On the Crop Diversity HPC - and many other HPCs - Docker is unavailable due to its required elevated privileges. Apptainer is a commonly available alternative for running containers on HPCs, and can run Docker containers. Unfortunately, the provided container was not functional with

Apptainer. The internal `$PATH` and other environment variables were not configured correctly and so the script continued to fail. Attempts to fix the container were made, but without a provided Dockerfile, it was eventually decided to be infeasible to achieve in a reasonable timeframe. Additionally, although the container in theory could reduce the overhead of manual installation, the Pfam database and `InterProScan` needed to be installed manually.

`RRGPredictor` is available on GitHub and was downloaded. The Perl and `InterProScan` dependencies were manually installed after which the software could be run. As the pipeline is only comprised of two short Perl scripts, execution was relatively straightforward. The output of `RRGPredictor` is a series of text files encompassing the output of `InterProScan` split into a file for each NLR domain, and a second series of files for each NLR classification containing a list of sequence IDs. Parsing of these would be relatively simple, although splitting of the classifications into individual files is an unusual approach. Beyond sequence ID, no additional information is provided. Additionally, the `RRGPredictor` scripts must be in the same directory as the `InterProScan` output files that it uses as input, and the output can only be written to the current working directory. Integrating this into a workflow manager like Nextflow would require additional configuration to handle this non-standard approach, likely softlinking of the `RRGPredictor` scripts.

`NLRtracker` was downloaded from GitHub. Dependencies were manually installed in a conda environment - it was noted that the required `MEME v5.2.0` is unavailable through conda, but `MEME v5.5.6` is available and is a functional alternative. A manual installation of `InterProScan` also needed to be added to path due to it being incompatible with the conda environment. Execution of `NLRtracker` was relatively straightforward with proper argument and input/output handling. An output directory is created with a comprehensive set of annotation results and extracted domain sequences.

`NLRexpress` is also available on GitHub and comes with a conda environment to automate the installation of dependencies. Machine learning predictor weights must be downloaded from an external link prior to running the tool. Execution of the `NLRexpress` python script is straightforward and uses sensible argument and input/output handling. The result file is a table containing the sequence ID, residue, motif ID, probability, and motif sequence of each motif identified in the input. The tabular format is non-standard and requires a degree of manual configuration to be correctly parsed. As `NLRexpress` is principally a motif-finder, no NLR classifications are determined.

To summarise, of the available tools that identify NLRs from protein sequences, only `RRGPredictor`, `NLRtracker`, and `NLRexpress` could be executed, of which `NLRtracker` is the more useable tool for the purpose of classifying NLRs and produces the most comprehensive set of results. Despite this, all tools still require manual installation and dependency resolving.

An overview of Resistify

`Resistify` was designed to make use of two sources of evidence for identifying NLRs - a streamlined database of high-quality HMM models derived from Pfam representing common NLR domains, and the machine-learning predictors of `NLRexpress` to accurately identify NLR motifs. First, the input protein sequences are searched for CC, TIR, RPW8, and NB-ARC domains via `hmmsearch`, but not the LRR domain which is often poorly represented by HMM models. In the standard mode, sequences that do not have evidence of an NB-ARC domain are discarded as they are unlikely to be canonical NLRs. This reduces the search space for the subsequent `NLRexpress` analysis which is high accuracy but more computationally intensive. A key change introduced in `Resistify` since `v0.1.1` has been the introduction of multithreading to the `jackhmmer` search that is used as extrinsic evidence by `NLRexpress`. This process normally maxes out at 2-4 threads due to I/O limitations - `Resistify` resolves this by internally splitting the input sequences and running `jackhmmer` in parallel. This in effect leads to a linear improvement in `NLRexpress` performance as thread count increases.

`Resistify` then uses both the HMM domain hits and NLR motifs to classify NLRs according to their architecture. Crucially, `Resistify` takes advantage of the high-quality LRR motifs provided by `NLRexpress` to provide consistent classification of LRR domains across diverse sequences. CC domains are also frequently missed by HMM searches alone - `Resistify` conservatively searches for CC domains in any sequences that do not have a clear upstream domain.

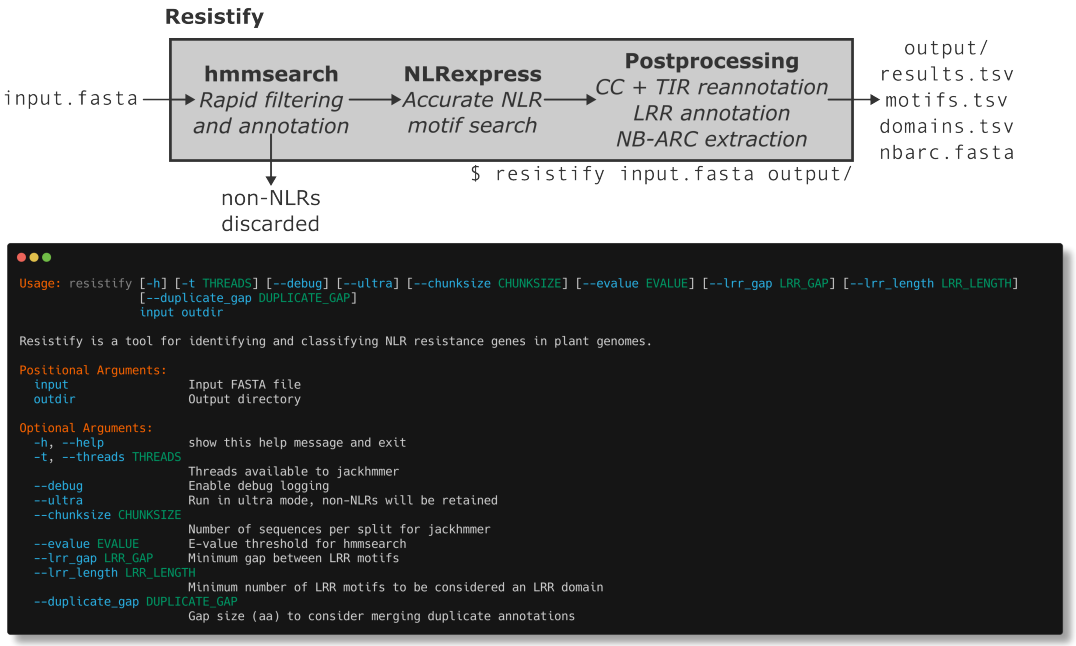


Figure 5: **The `Resistify` program.** Top: an overview of the internal processing and logical flow of `Resistify`. Bottom: A screen capture of the command-line output of `Resistify` detailing its arguments and available options.

To reduce the burden on users when installing `Resistify`, the python code was

uploaded to PyPi from which a Conda package was created through the Bioconda channel which handles all dependencies. As `Resistify` is made available on Bioconda, a functional Docker container is automatically made available through https://quay.io/repository/biocontainers/resistify. The HMM database and `NLRexpress` predictors are also included with the program so that they do not require manual installation.

Performance against RefPlantNLR

To evaluate the performance of `Resistify` (fig. 5), it was applied against the RefPlantNLR database - a curated set of 415 previously cloned NLRs from a diverse range of species. `NLRtracker` was also applied to the RefPlantNLR database for comparison.

In the default mode, only three RefPlantNLR entries were not identified by `Resistify` - *AtNRG1.3* which carries only an LRR domain and *Pb1* and *RXL* which have significantly truncated NB-ARC domain that are not listed in the RefPlantNLR database. However, if required, these genes can be identified with `Resistify` using `--ultra` mode which skips the initial filtering stage and reports and sequence which contains at least one NLRexpress motif. Sequences are reported as an unmerged string of NLR motifs. Consequently, *AtNRG1.3*, *Pb1*, and *RXL* are reported as NNLLLLLLL, CNNNLLLLLLLLLL, and CNNNLLLLLLLL respectively.

The largest source of variation between RefPlantNLR and `Resistify` classification was 29 NLRs which had an NL structure according to RefPlantNLR but CNL according to `Resistify`. All of these belonged to CNL-associated subclasses which are known to have CC domains that are challenging to identify. This included 15 CC_{G10}-NLRs including *Pvr4*, *Tsw*, *RPS2*, *RPS5*, *SUT1* and *SUMM2* which have previously been noted for their lack of the CC-associated EDVID motif (H. Lee et al. 2021). Others included six members of the *Pm5* locus which despite not containing a CC domain in RefPlantNLR, have previously been identified to contain CC-like domains [xie_2020]. This analysis demonstrates that `Resistify` is highly sensitive at retrieving canonical NLRs and accurately describing their structure.

As `NLRtracker` does not exclude NLRs without NB-ARC domains, *AtNRG1.3*, *Pb1* and *RXL* were identified and classified as "CC-NLR or CCR-NLR or CC_{G10}-NLR", "CC-NLR", and "CC-NLR" respectively. `NLRtracker` relies on the NB-ARC associated RNBS-D motif to classify NLRs as "CC-NLR or CCR-NLR or CC_{G10}-NLR". As a result, 26 of the conflicting CNLs identified by `Resistify` were classified as "CC-NLR or CCR-NLR or CC_{G10}-NLR" despite failing to identify a CC domain. The exception of this was for *SpNBS-LRR*, *Rpp1-R1*, and *Rpp4C4* which were classified as "UNDETERMINED". Reliance on RNBS-D also led to classification of *NtTPN1* as "CC-NLR or CCR-NLR or CC_{G10}-NLR". Whilst *NtTPN1* is a CNL homolog, it lacks any upstream domain and is structurally an NL. Unexpectedly, the TNL *DSC1* was misclassified as "CC-NLR" despite having a domain structure of "(TIR)(NBARC)(LRR)(CJID)" according to `NLRtracker`.

In summary, `Resistify` performs well at identifying canonical NLRs from a diverse range of species. In default mode, it does not assign genes as NLRs with extremely truncated or entirely absent NB-ARC domains, unlike `NLRtracker` which reports any sequence with NLR-associated domains. However, this can be replicated in `Resistify`

with the `--ultra` mode. Thus, `Resistify`'s structure-based classification method is well suited for correctly classifying NLRs, including members of challenging CNL subclasses.

Performance against the Araport11 proteome

To assess the performance of `Resistify` across a well characterised and annotated genome, the representative gene models of Araport11 were analysed (C.-Y. Cheng et al. 2017). In total, `Resistify` identified 166 NLRs - the majority of which were TNLs and CNLs (fig. 6 a). Of the CNLs, 25% had a MADA motif, and 44.4% and 41.2% of NLs and TNLs had C-JID domain respectively. Partial NLRs either without an N-terminal or LRR domain were also identified.

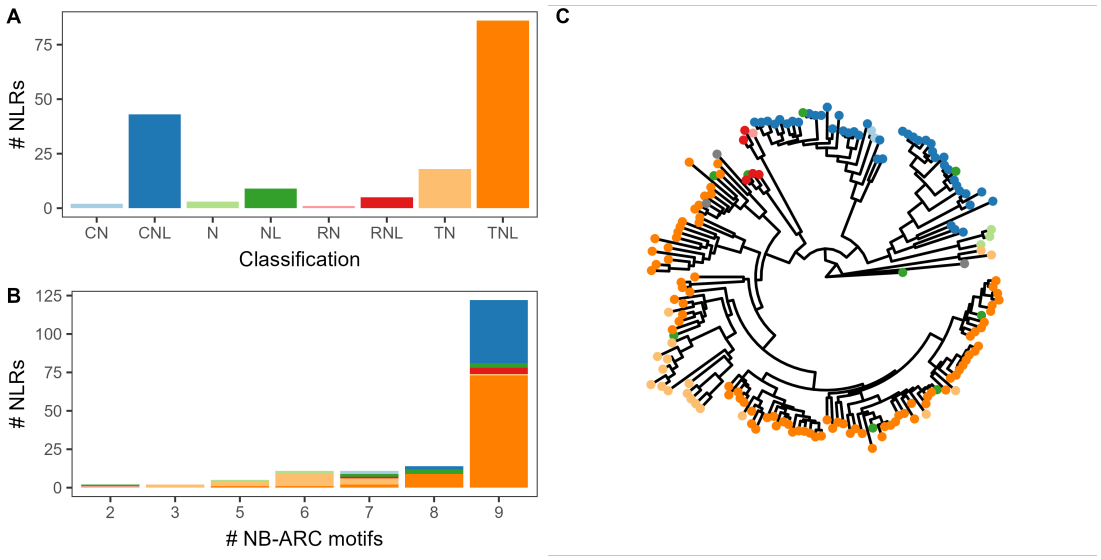


Figure 6: **`Resistify` applied to Araport11.** a) The number of NLRs belonging to each identified classification. b) The number of NLRs grouped by number of identified unique NB-ARC associated motifs. c) A phylogenetic tree of NLRs based on the `Resistify` extracted NB-ARC domain sequences. The NB-ARC domain of CED-4 has been included as an outgroup.

Manual inspection of the motifs within these sequences confirmed that they were not due to a failure to identify these elements. The majority of NLRs carried all nine conserved NB-ARC motifs - those with fewer increasingly belonged to partial NLR classifications (fig. 6 b). NLRs with as few as two of the conserved NB-ARC motifs were successfully identified. A phylogenetic tree of the `Resistify`-identified NB-ARC domains independently validated the classifications and allowed the placement of ambiguous NL or N classified sequences into subclasses (fig. 6 c).

`NLRtracker` identified an additional 48 sequences, however these sequences did not contain any NB-ARC domain annotation according to either tool. Three sequences - *AT4G19060.1*, *AT4G19060.1*, and *AT5G45440.1* - were not identified by `NLRtracker`. According to `Resistify` these contained a single NB-ARC domain each, with 5, 6, and 6 NB-ARC motifs respectively. Overall, `Resistify` performs well at identifying and

classifying NLRs from whole proteomes and successfully retrieves NB-ARC domain annotations for phylogenetic analyses.

Application against an example workflow

To demonstrate how `Resistify` might be implemented to identify novel resistance genes, a pangenome experiment was performed. Eighteen contiguous *Solanum* genomes were downloaded in addition to the *S. verrucosum* genome presented in Chapter 2, (Tang et al. 2022; N. Li et al. 2023). The *Capsicum annuum* genome was also included as an outgroup (F. Liu et al. 2023). Genomes were subsequently processed with a simple workflow which predicts genes *de novo*, identifies homologues, and classifies NLRs with `Resistify`. For gene annotations, the recently developed tool `Helixer` was selected for its near reference quality predictions and lack of requirement for repeat-masking which is known to result in false-negative NLR annotations (Bayer, Edwards, and Batley 2018; Holst et al. 2023). The workflow also predicts transposable elements with `EDTA`. The *C. annuum* genome was included as an outgroup, but also because recent analysis has suggested a substantial expansion of NLRs in this species associated with LTR transposable elements (Kim et al. 2017).

Predicted gene content varied between genomes from 29,223 in *S. habrochaites* to 61,015 in the larger *C. annuum* genome (tbl. 3). Transposable element content ranged from 55.6% in *S. chmielewskii* to 76.8% in *C. annuum*. Whilst there was no significant difference in total TE content between tuber-bearing and non-tuber-bearing genomes (p=0.644), there was a significant increase in the number of intact TEs reported by EDTA in the genomes of tuber-bearing species (p=0.008).

Table 3: **Annotation statistics.** Transposable element and gene annotation statistics for each genome used in this study. The tuberising status is also listed according to their classification in source publications. Genomes originated from (Tang et al. 2022; N. Li et al. 2023; F. Liu et al. 2023).

Genome	Intact TEs	% TE	Genes	Genome size	Tuberising
C. annuum	15108	76.78	61,015	3.02 Gbp	False
S. candolleanum	10311	59.18	38,121	714.80 Mbp	True
S. chilense	8186	57.47	33,285	807.50 Mbp	False
S. chmielewskii	5339	55.65	29,270	734.85 Mbp	False
S. corneliomulleri	7926	59.43	29,793	776.61 Mbp	False
S. etuberosum	5499	60.81	32,082	684.12 Mbp	False
S. galapagense	5753	56.72	32,244	800.98 Mbp	False
S. habrochaites	3606	55.63	29,223	825.88 Mbp	False
S. lycopersicoides	9417	67.24	40,831	1.11 Gbp	False
S. lycopersicum	4858	55.68	28,927	736.82 Mbp	False
S. neorickii	5329	55.1	29,308	732.11 Mbp	False
S. peruvianum	8733	60.44	29,764	796.99 Mbp	False
S. pimpinellifolium	5432	57.64	29,243	760.60 Mbp	False
S. phureja (E4-63)	10639	58.49	38,037	741.82 Mbp	True

Genome	Intact TEs	% TE	Genes	Genome size	Tuberising
S. phureja (E86-69)	10334	56.81	37,325	705.19 Mbp	True
S. stenotomum (A6-26)	10614	59.49	37,494	729.41 Mbp	True
S. stenotomum (PG6359)	10935	59.77	38,174	745.62 Mbp	True
S. tuberosum_RH	11045	59.31	39,079	750.89 Mbp	True
S. tuberosum (RH10-15)	10842	59.11	38,243	753.92 Mbp	True
S. verrucosum	8744	58.37	36,002	699.05 Mbp	True

In total, 8144 NLRs were identified across all genomes (fig. 7). CNLs were the most abundant classification of NLR identified, ranging from 84 in *S. habrochaites* to 422 in *S. tuberosum* (group tuberosum RH10-15). There was a notable expansion of NLRs in tuber-bearing *Solanum* species. This is in agreement with the previous observation that potato-bearing *Solanum* species have an expansion of tuber expressed NLRs (Tang et al. 2022). This effect does not correlate with the proportion of genome occupied by transposable elements of any classification.

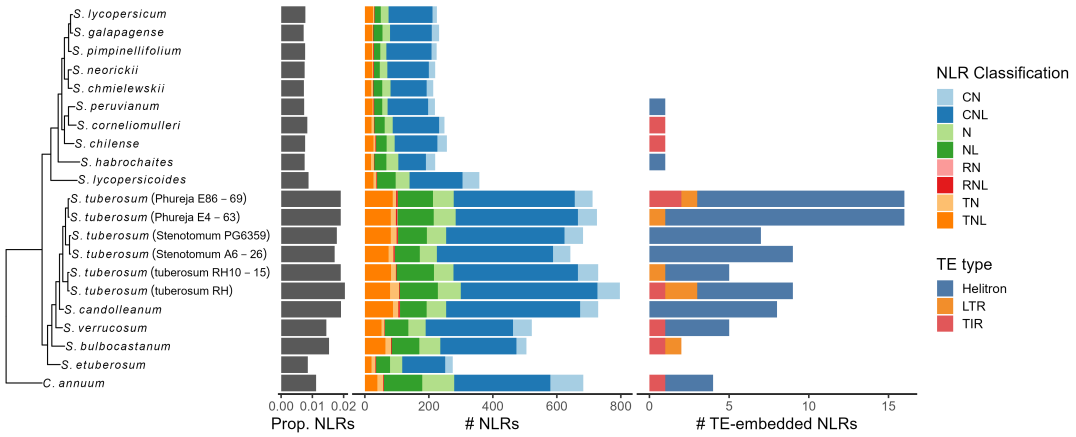


Figure 7: **NLRs identified across the *Solanum* pangenome.** The proportion of NLRs identified within the predicted genes for each genome, the number of NLRs of each classification, and the number of NLRs identified to be embedded within transposable elements, is listed for each genome.

In total, 38,590 orthogroups were identified, of which 687 (1.8%) contained NLRs. The distribution of orthogroups showed a clear divide between core and shell/cloud orthogroups within the pangenome with the majority of NLR orthogroups existing within the shell/cloud (fig. 8 a). This reflects the relatively large width of the pangenome, which captures NLR variation over a Genus level. Species specific NLRomes also exhibit an abundance of cloud orthographs, reflecting the high variability of NLRs in genomes (Van de Weyer et al. 2019). Many (48.1%, n=371) orthogroups were classified as containing N or NL NLRs (fig. 8 b). Whilst these classifications are less abundant (24.6%, n=2081), they were often associated with smaller orthogroups contributing to this inflation.

The distribution of previously identified Solanaceae NLR homologs across the

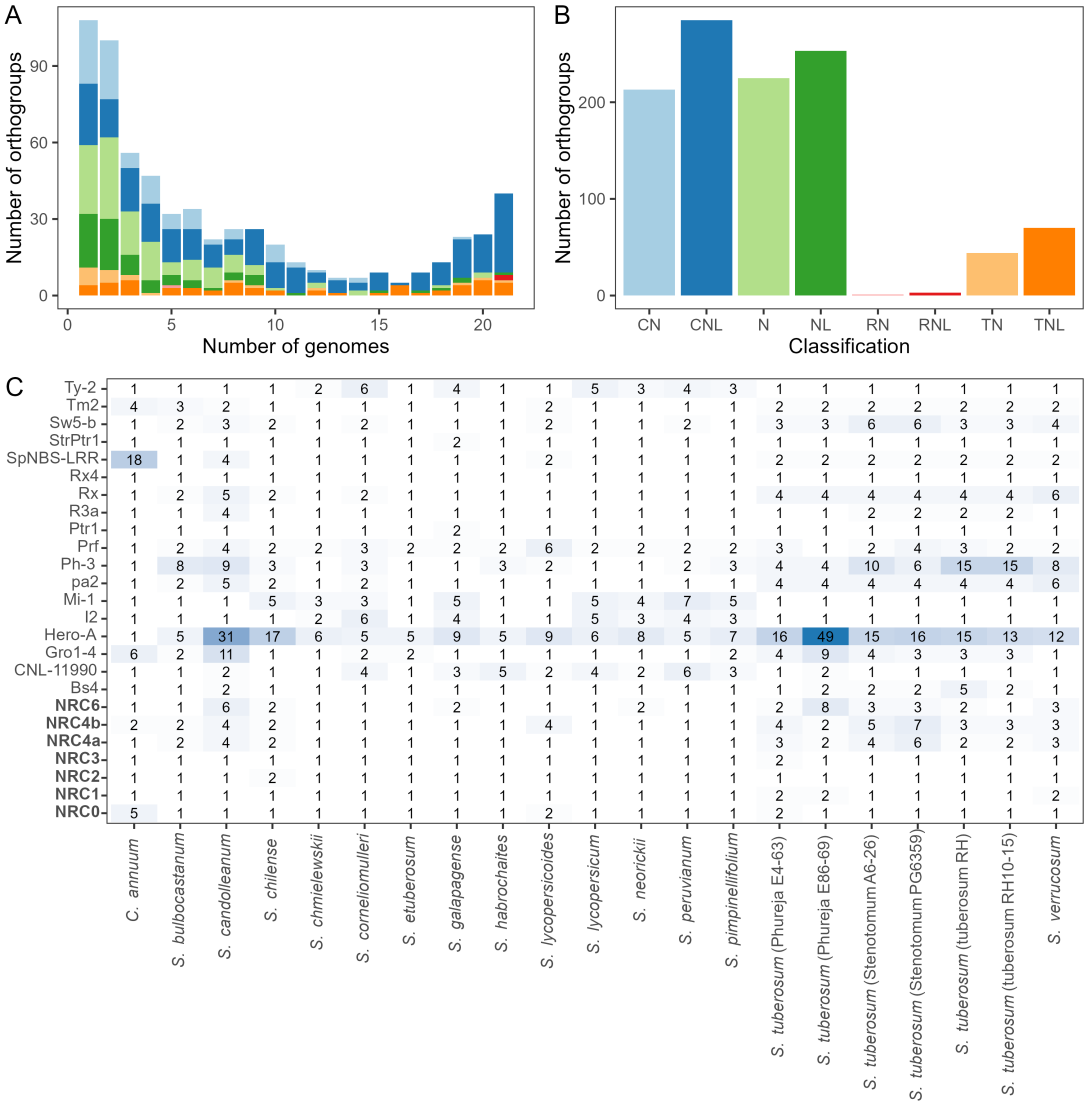


Figure 8: **NLR-containing orthogroups.** a) The number of orthogroups shared between genomes. b) The number of orthogroups according to NLR classification. c) The number of homologs of known *Solanaceae* NLRs identified in each genome

pangenome was examined (fig. 8 c). In agreement with previous findings, members of the NRC group remained relatively stable across the pangenome except for *NRC4a*, *NRC4b*, and *NRC6* which were expanded in the tuber-bearing genomes. By contrast, an expansion of *Mi-1* and *Hero* was evident in the non-tuber bearing and tuber bearing genomes respectively. At least one homologs was identified in each genome for all genomes.

It has previously been reported that there is a vast expansion of NLRs in the *C. annum* genome due to retroduplication; NLRs nested within LTRs represented a large proportion (~13%) of NLRs within the genome, and this effect is also seen in tomato (8%) and potato (18%) (Kim et al. 2017). To explore whether this effect could be linked to the expansion of NLRs within the tuberising members of *Solanum*, a similar analysis was repeated. Intact TEs were identified and considered to interact with NLRs if they covered >90% of the gene annotation.

Unexpectedly, the effect could not be replicated and across all genomes only five intact LTRs were identified as containing NLRs (fig. 7). None were identified in *C. annuum* but instead only in the *S. tuberosum* group Phureja and clone RH, and *S. bulbocastanum*. The putative retrotransposed NLRs within these all belonged to the same orthogroup (OG0000639) which was expanded in both Phureja and RH, but not *S. stenotomum*.

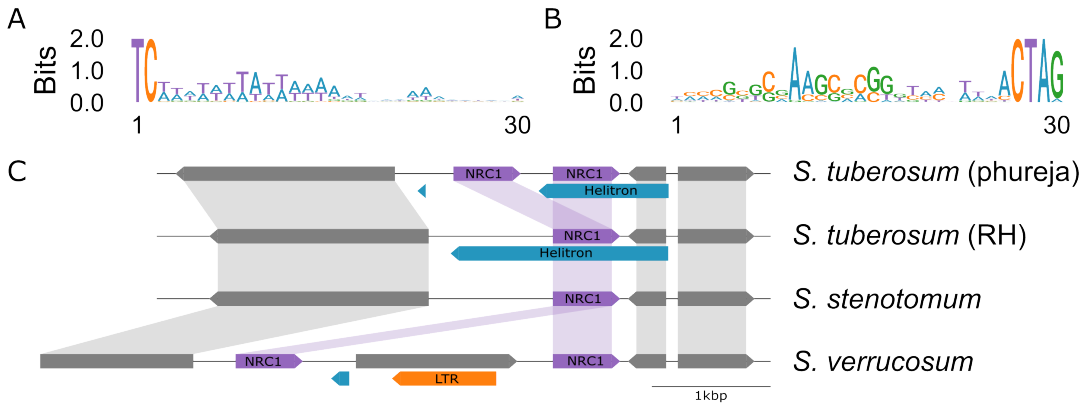


Figure 9: ***Helitrons* are associated with NLRs.** a) The 5' motif of NLR overlapping *Helitrons*. b) The 3' motif of NLR overlapping *Helitrons*. c) The *NRC1* locus and its association with *Helitrons* across four genomes. Helitron (blue) and LTR (orange) transposable elements are highlighted.

By contrast, a proliferation of *Helitron*-associated NLRs was identified in the tuber-bearing genomes. All predicted *Helitrons* carried the expected 5'-TC...CTRR-3' signature as well as a GC-rich region in the 3' terminal (fig. 9 a,b). An example of an NLR which may have undergone *Helitron*-associated duplication is *NRC1* (fig. 9 c) which is present as a single copy except in both *S. tuberosum* group Phureja genomes and *S. verrucosum*. Close examination of the *NRC1* locus reveals a duplication event in S. tuberosum group Phureja where one *NRC1* gene is nested entirely within a predicted *Helitron*. In *S. tuberosum* clone RH which only has one copy of *NRC1*, *NRC1* is similarly nested within a *Helitron* although this appears to be extended to a distant *Helitron* terminator past the site of duplication in group Phureja. In *S. stenotomum*,

only one *NRC1* is present and no *Helitron* predicted. Interestingly, *S. verrucosum* has two copies of *NRC1* and whilst it does not have a nested *Helitron* copy, the leftmost *NRC1* copy does have a short predicted *Helitron* upstream of its start position in the same orientation as the other *Helitron* elements. In this case, the two copies of *NRC1* appear to have been further separated by an intact LTR insertion which has resulted in an additional gene prediction.

Discussion

`Resistify` is presented as a highly accurate and easy-to-use tool to aid the identification of novel resistance genes. Applying `Resistify` against the RefPlantNLR database demonstrates that it is effective at identifying NLRs from a diverse range of plant genomes. It is highly sensitive at classifying challenging CNL families which often evade other NLR classification tools due to the inclusion of the performant `NLRexpress` motif models. Members of the CC_{G10}-NLR clade have previously been described as lacking the CC-associated EDVID motif and are often predicted as having an NL structure in the RefPlantNLR database (H. Lee et al. 2021). By utilising `NLRexpress`' highly sensitive extended EDVID motif model, `Resistify` successfully classifies these elusive NLR clades, highlighting it as the most sensitive NLR classifier to date.

As demonstrated, `Resistify` can be easily integrated into workflows and is scalable to large pangenome experiments. As high-quality genome assemblies become more common, providing accessible tools for the study of NLRs will be crucial for fully appreciating their diversity and identifying novel sources of disease resistance. The recent releases of chromosome level assemblies for potato and tomato pangenomes are a valuable resource for understanding the diversity of NLRs within *Solanum* (Tang et al. 2022; N. Li et al. 2023). An expansion of NLRs including essential families such as the NRCs is apparent in tuber-bearing species, as well as other functional NLR families.

As `Resistify` relies on an initial NB-ARC domain search to reduce the search space prior to NLRexpress motif identification, NLRs with very truncated or entirely absent NB-ARC domains can be missed. To resolve this, an additional `--ultra` mode is provided which does not perform initial NB-ARC domain filtering which identifies sequences with any NLR motifs, which proves to be more sensitive than `NLRtracker` for detecting highly fragmented NLRs.

An unexpected finding of this study was the failure to replicate the previous observation of an abundance of LTR associated NLRs across Solanaceae (Kim et al. 2017). Differences in gene and TE annotation and NLR classification likely contributed to this. Although `Helixer` predicted ~70% more genes in *C. annuum* than were identified in the previous study, this did not translate to an increase in NLRs identified. Instead, only 673 NLRs were identified here in comparison to the previous estimate of 835. Previously, a `tBLASTn` search of the NB-ARC domain followed by ORF identification and `BLASTp` searches against GenBank to classify NLRs was used to identify NLRs in the genome (Seo et al. 2016). This method may result in more false positives in comparison to identifying NLRs directly from gene annotations. Here, `EDTA` was

selected for identifying TEs which has an improved sensitivity and selectivity for identifying intact LTRs in comparison to `LTRHarvest` which was used previously. Although `Helixer` does not require repeat masking which can result in false negative NLR annotations, annotations used in training the model (which may have relied upon repeat masking) might introduce TE-avoidant behaviour (Bayer, Edwards, and Batley 2018). However, the identification of several LTR-associated NLRs indicates that this is unlikely to be the source of the difference in the result.

The identification of *Helitron*-associated NLR expansion has not been previously reported. *Helitrons* are challenging to identify due to their lack of structural elements and as a result the pipeline used in this study is known to overestimate *Helitron* density (Baril, Galbraith, and Hayward 2024). However, a large part of this overestimation is due to `EDTA` reannotating the genome with `Helitronscanner` predictions, which itself suffers from a high false positive rate, leading to a proliferation of fragmented *Helitron* annotations across the genome. For this study, only intact *Helitrons* which passed `EDTA`s stringent filter were considered. As a result, all predicted *Helitrons* had the required structural elements for activity. Further validation would be required to determine if *Helitron* association is a valid mechanism of NLR expansion. The *Helitron*/*NRC1* relationship highlighted here would be a good starting point for unpicking this mechanism.

Improving Resistify

`Resistify` was principally designed as a tool to provide "good enough" annotations of canonical NLRs in genomes, as they are the most frequent and desirable source of disease resistance in potato - and other crop - genomes. Canonical NLRs are represent the vast majority of identified functional NLRs to date, and studies on NLR diversity tend to focus predominantly on those with canonical architecture. In situations where non-canonical genes may be of interest to researchers, for example in screening a mapped locus, `--ultra` mode is provided which can indicate highly fragmented NLRs to researchers if necessary. However, there are several areas where `Resistify` could be improved with minimal effort.

Programs such as `RGAugury` and to a lesser extent `NLRtracker` identify in addition to NLRs, classes of other resistance-related gene families such as RLK/RLPs, Jacalin-lectin domains, and transmembrane-CC proteins. Both `RGAugury` and `NLRtracker` achieve this principally through `InterProScan` hits, and `RGAugury` additionally uses transmembrane domain prediction software. Integrating similar annotations in `Resistify` should be relatively simple. High-quality Pfam domain annotations exist for the majority of these families which could be included in the initial HMM search with minimal impact on performance. The LRR domain strategy of `Resistify` is also well suited for identifying LRR domains from both NLRs and RLK/RLPs, and would be more performant that relying on `InterProscan` LRR annotations. Since the development of `RGAugury`, more performant transmembrane domain annotators are also available such as `DeepTMHMM` and `TOPCONS`, however to date no annotators are distributed in a way that permits easy installation or allows predictions to be run locally (Hallgren et al. 2022; Tsirigos et al. 2015). Caution should be taken when integrating software to

ensure that accessibility is not compromised.

Unlike `NLRtracker`, `Resistify` does not search for noncanonical integrated domains (Kroj et al. 2016). As integrated domains are widespread and often critical to NLR function, their identification and analysis are an important factor in studies of NLR diversity (Barragan and Weigel 2021). Pairing `Resistify` with a secondary sweep for these domains with `NLRtracker` or `InterProScan` would permit this analysis whilst also benefiting from a reduced search space and high-quality NLR classifications from `Resistify`. Integrating this directly into `Resistify` would be unwise, as it would greatly increase the dependency and database requirement, and would be better suited for a workflow manager such as Nextflow.

In addition to classifying NLRs, `NLRtracker` also provides ready-to-use GFF annotations and formatted data for producing phylogenetic visualisation with iTOL. These functions are useful for users with less bioinformatics experience, who may be less comfortable in parsing and creating similar files. The tabular data of `Resistify` has been designed to be easy to parse, but further efforts in this area would be valuable.

Assembly and analysis of the *Solanum verrucosum* genome

Introduction

Solanum verrucosum is a wild potato species native to Mexico. It has been widely studied as it contains valuable sources of disease resistance and has the ability to act as a bridge species, enabling crosses between some species with different Endosperm Balance Number (EBN) values. In addition, it is self-compatible, unlike most tuber bearing species of *Solanum*, and has unusual centromeres (Eijlander et al. 2000; Gong et al. 2012). These properties have led to *S. verrucosum* being a focus for genomic studies (Hosaka, Sanetomo, and Hosaka 2022; Paajanen et al. 2019).

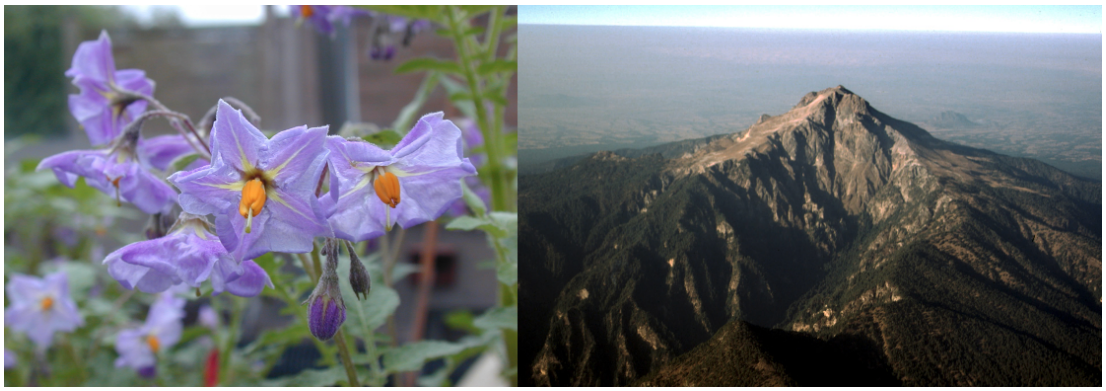


Figure 10: **Solanum verrucosum.** Left: A photograph of the flowers of *S. verrucosum* clone 54 (source: https://ics.hutton.ac.uk/germinate-cpc/; ID: 10.18730/5ANEF). Right: A photograph of La Malinche in Mexico, the original sample site of *S. verrucosum* clone 54 (source: https://volcano.si.edu/volcano.cfm?vn=341091)

S. verrucosum is resistant to late blight and other potato diseases

Late blight, or potato blight, is a serious disease that has impacted potato cultivation and breeding for over 150 years (Ivanov, Ukladov, and Golubeva 2021). The oomycete pathogen *Phytophthora infestans* is the causal agent of late blight. *S. verrucosum* has been long noted for resistance against this pathogen, going as far back as 1852 in The Florist and Horticultural Journal, when it was noted that:

> "...the plants of *Solanum verrucosum* were quite free from disease, although common Potatoes planted by their side were attacked as early as the end of July; and we have the evidence of Prof. Decaisne, that in the Garden of Plants, at Paris, the same power of resisting disease, was remarked."

Functional orthologues of the *RB* gene have been identified in this species while more recent studies have identified that a novel broad spectrum dominant late blight resistance gene, *Rpi-ver1*, is present on chromosome 9 of this species (Z. Liu and Halterman 2006; X. Chen et al. 2018). Resistance to plant viruses and tomato psyllid have also been reported in *S. verrucosum*, while one accession also showed resistance to a broad range of *G. pallida* populations (Carrasco et al. 2000; Cooper and Bamberg 2016; Castelli et al. 2005). A more detailed analysis of the genome of *S. verrucosum* will therefore improve our understanding of disease resistance in *Solanum* as well as the processes underpinning genome evolution and structure. This is particularly true for the *Rpi-ver1* resistance, which although has been localised to chromosome 9, has yet to be identified at the gene level.

Plant genomes are shaped by transposable elements

Transposable elements occupy a large fraction of plant genomes. Generally, the frequency of transposable elements is positively associated with genome size - *Arabidopsis* for example has a genome of ~125 Mbp of which ~10% is transposable elements (The Arabidopsis Genome Initiative 2000), whereas larger genomes such as the ~14Gbp wheat genome is composed of ~85% transposable elements (Aury et al. 2022). The ~800 Mbp tomato and potato genomes, which lie close to the 700 Mbp median size of plant genomes, are composed of ~65% and 75% transposable elements respectively (Bozan et al. 2023; Su et al. 2021; Akakpo et al. 2020).

Representing such a significant proportion of the genome, it is not surprising that transposable element activity can be linked to many agronomic traits as well as driving a wide range of evolutionary processes. Transposable elements in plants are broadly separated into two classification groups - Class 1 retroelements and class 2 DNA elements. Class 1 elements make use of an RNA intermediate during transposition and as a result do not require excision from their locus prior to transposition. Class 2 elements do not have an RNA intermediate, but are instead excised from the host genome via a transposase followed by re-integration into the genome elsewhere. Within each classification, further subfamilies exist which exhibit considerable variation in their structure, "behaviour", and effect on plant genomes.

The family of class 1 elements that generally occupies the largest fraction of plant genomes are the long terminal repeat (LTR) retrotransposons. Autonomous LTRs encode all the necessary components for retrotransposition - *gag*, a caspid protein, protease, integrase, reverse transcriptase, and a ribonuclease. The Ty3-*gypsy* and Ty1-*copia* LTR superfamilies are abundant in plant genomes where they are respectively associated with non-genic and genic regions of the chromosomes (Galindo-González et al. 2017). Non-LTR class 1 elements include the long interspersed nuclear elements (LINEs), less prevalent in plants but that curiously have invaded the centromeres of *Poplar* trees; and the short interspersed nuclear elements (SINEs), non-autonomous ele-

ments that hijack the transposition machinery of LINEs and are abundant in Solanaceae species where they contribute to gene evolution (Xin et al. 2024; Seibt et al. 2016).

Class 2 elements include the *hAT*, CACTA, *Mariner*, *Mutator*-like transposable elements (MULEs), and *Helitron* elements which have distinct structural and mechanistic features. For example, MULEs are flanked by terminal inverted repeats (TIRs) and contain a transposase domain which serves to bind the TIRs together, cleave them from the genome, and inserts them into a new site (Dupeyron et al. 2019). *Helitron* elements are flanked by small 5' TY and 3'CTRR motifs alongside a 3' palindromic sequence, and autonomous *Helitrons* encode a *Helicase* enzyme that facilitates transposition through a rolling-circle mechanism (K. Hu et al. 2019). Derivatives of class 2 elements also exist such as miniature inverted-repeat transposable elements (MITEs) which frequently insert into gene-rich regions and can influence local gene expression (Jiongjiong Chen et al. 2014).

Transposable elements drive evolution in plants. Highlighted below are some hallmark features of *Solanum* genomes that are influenced by transposable element activity.

Polyploidisation

Polyploidisation is often associated with an expansion of transposable elements in the genome. One reason that increased transposable element activity might be more tolerated in this instance is that the newly duplicated copy of genes essentially acts as a buffer against deleterious transposable element insertions (Akakpo et al. 2020). Transposable elements insertions near essential genes drives the evolution of these genes, potentially to the advantage of the plant in adapting to new stressors. Thus, polyploidisation can act as a trigger to mediate transposable element led adaptation.

The transposable element burst associated with polyploidisation is transient, and generally TE activity is suppressed in the subsequent generations, although the new inventory of TEs can continue to exhibit a degree of activity. Transposable element activation is often driven by widespread DNA demethylation which is observed in newly formed polyploids and which gradually transitions back to hypermethylation of transposable elements in the subsequent generations (Kraitshtein et al. 2010). Given the frequent variations in ploidy within *Solanum* species, it can be expected that transposable elements are a key driver of potato adaptation and diversification.

Transposable elements shape centromeres

Centromeres are a specialised chromosomal domain which act as the assembly site of the kinetichore - a protein complex which acts as a physical tether and mediates the interaction between the chromosome and spindle microtubules involved in chromosome translocation (Talbert and Henikoff 2020).

A hallmark of the centromere in eukaryotic genomes is the presence of the centromeric histone H3 variant CENH3, also referred to as CENP-A in mammals. The mechanisms which dictate which region of the chromosome will be occupied by CENH3 varies across the tree of life. In budding yeast, the site of CenH3 inclusion is dictated by a ~120bp DNA sequence, often referred to as a "Point centromere" (Henikoff and

Henikoff 2012). However, this is atypical for eukaryotes where it is determined epigenetically (Cuacos, H. Franklin, and Heckmann 2015). The sequence of the human centromere is composed primarily of megabase arrays of ~171bp satellite DNA repeats, which are further arranged into higher order repeat units on the order of kilobases (Altemose et al. 2022). Human centromeres are further defined by DNA methylation, exhibiting satellite repeats with dense CG hypermethylation as well as the well-defined "centromere dip region" - a depression of hypomethylation where CENP-A binding is highest. Despite their essential and conserved function, centromeres are the sites of some of the fastest evolution in human genomes and exhibit significant sequence diversity between individuals (Logsdon et al. 2024).

Similar centromere compositions are seen in plant genomes. The centromeres of *Arabidopsis* are composed of megabase satellite arrays with ~178bp subunits which are similarly organised into higher order repeats (Naish et al. 2021). The centromeres are densely methylated in the CG, CHG, and CHH contexts although satellite arrays depleted in the CHG context are also present, likely due to H3 substitution with CENH3. *Arabidopsis* centromeres similarly substantial diversity between individuals. They are also the site of frequent incursions of the *ATHILA* LTR retrotransposons which may have specialised to invade centromeric DNA, possibly by adapting to recognise the chromatin state of centromeres. *ATHILA* LTRs are in turn purged from the centromere through an unknown homologous recombination system. This leads to a cyclic expansion and contraction of *ATHILA* elements in the centromere, driving evolution and speciation within *Arabidopsis* (Wlodzimierz et al. 2023).

Many plant genomes have centromeres that do not fit this general pattern. Coimmunoprecipitation of CENH3-bound DNA in *S. tuberosum* revealed uneven and fragmented coverage across the potato genome (Gong et al. 2012). Large peaks of ChIP-seq signal were detected on five of the twelve potato chromosomes. Clustering of CENH3-ChIP reads identified kilobase monomers which were unique to chromosomes and when visualised through FISH, were found to be arranged in megabase centromeric-arrays. These large repeat monomers shared sequence similarity with LTR retrotransposons, indicating their role in centromere evolution in *Solanum* (Gong et al. 2012). Repeats showed varying conservation across six wild *Solanum* species, including *S. verrucosum*, and were often entirely absent. The presence of both repeatless and repeat-based centromeres with homology to LTR retrotransposons suggests a mechanism of centromere formation from repeatless "neocentromeres" by expansion of satellite repeats. Under this model, the *S. tuberosum* centromeres can be considered as being in a mixed state of well-established centromeres, and neocentromeres evolving into a repeat-based structure. A follow up study on the sequence of identity of *S. verrucosum* centromeres revealed that only a single centromere was present that shared repeats with *S. tuberosum*, and that the remaining centromeres contained unique repeats or were in a repeatless state (H. Zhang et al. 2014).

Transposable element annotation strategies

There are multiple strategies available for identifying transposable elements in a plant genome. `RepeatModeler2` is a popular tool which identifies transposable elements by constructing seed alignments of known transposable element families, followed by

successive rounds of searching and consensus refinement of these families (Flynn et al. 2020). One recent advancement of `RepeatModeler2` was to carry out LTR detection in a separate module which takes advantage of their distinct structural elements. In a final step, annotations are merged into a non-redundant library of transposable elements which is classified by comparing sequences to a database of previously identified transposable elements. Despite automated curation, a substantial amount of fragmentation and redundancy often remains in transposable element libraries. To resolve these, recent tools such as `Earl Grey` have been developed which are pitched to resolve this through additional post-processing of the `RepeatModeler2` transposable element library (Baril, Galbraith, and Hayward 2024).

The transposable element annotation `EDTA` uses an alternative strategy whereby a series of tools are executed which identify transposable elements according to their structural elements (Ou et al. 2019). For example, *Helitron* elements are identified through their 5' and 3' signatures, rather than by homology to previously identified *Helitrons*. A benefit of this strategy is that a bias is not given towards identifying "more of the same", and can be used to curate novel transposable elements.

The performance of transposable element annotation software is a hotly debated topic and has resulted in a series of studies lambasting each other's weaknesses, mostly through a series of benchmarks against model organism genomes (Ou et al. 2019; Gozashti and Hoekstra 2024; Baril, Galbraith, and Hayward 2024). This is probably in part due to the genomes in which the tools were developed - `EDTA` was developed with the rice genome and places an emphasis on LTR, MITE and *Helitron* annotation, whereas tools like `RepeatModeler2` are more aligned with non-plant genomes, which are less dominated by LTRs and other plant-prevalent elements. Plant transposable elements are very poorly represented in databases such as `Dfam` from which homology searches are often conducted.

The role of DNA methylation in plant genomes

In plants, DNA methylation of cytosine nucleotides can occur in three contexts - CG, CHG, or CHH, where H is a C, T, or A nucleotide. Transposable elements tend to be densely hypermethylated across their sequence and as a result represent a significant proportion of the total DNA methylation in plant genomes (Baduel and Colot 2021). Levels of CG methylation are generally highest whereas CHH methylation is more sparse. This is a result of the separate mechanisms which perpetuate these types of DNA methylation - CG methylation is copied directly as new DNA is being replicated; CHG methylation is copied through the recognition of H3K9me2 histone modifications which mark CHG methylation sites; CHH methylation must be repeatedly re-established during replication owing to its asymmetric structure [baduel_epiallelic_2021]. DNA methylation works in concert with other epigenetic marks to suppress transposable elements by preventing the expression of transposition machinery (W. Zhou et al. 2020).

The function of methylation in the context of genes is rather opaque. A common trend amongst plant genomes where methylation data is available is that gene body methylation (gbM) in the CG context only is positively associated with gene expres-

sion and expression stability across different tissues (Muyle et al. 2022). The trend is subtle, contradictory, and in the case of *Eutrema salsugineum* which lacks gbM entirely, non-existent (Bewick et al. 2016). Methylation of exons in the CHG context is mostly associated with a decrease in gene expression and in Maize has been used as a strategy for flagging genes as pseudogenes or misannotations derived from transposable elements.

In *Solanum*, the status and function of DNA methylation has been studied in several contexts. A study on *S. lycopersicum* and *S pimpinellifolium* and their reciprocal hybrids demonstrated that hybrids had lower levels of DNA methylation in LTR retrotransposons and genes with a variety of functions (Raza et al. 2017). In *S. tuberosum*, the application of DNA methylation inhibitors could promote tuberisation in genotypes exhibiting photoperiod-sensitive tuberisation (Ai et al. 2021).

S. verrucosum as a bridge species

Although the majority of cultivated potatoes are tetraploid, the vast majority of wild *Solanum* species are diploid (Hijmans et al. 2007). Whilst wild varieties often exhibit traits that are desirable for introgression into cultivated potato, traits often cannot be directly integrated into cultivated potato. Any systems which can overcome this barrier are of interest to breeders.

The endosperm is a tissue that surrounds the embryo of angiosperms, providing a vital food source for the developing embryo which absorbs the endosperm whilst developing into a mature seed (Carputo et al. 1999). In angiosperms - including *Solanum* - the endosperm is formed by a sperm cell fusing with the polar nuclei of the central cell in conjunction with the fertilisation of the egg by a second sperm cell. The central cell, being diploid, leads to the development of a triploid endosperm upon fusion with the haploid sperm cell. This 2:1 Maternal:Paternal ratio is an important factor in proper endosperm development and forms a core component of the Endosperm Balance Number (EBN) hypothesis (Carputo et al. 1999). In crosses between parents of different ploidy this ratio is broken, leading to maternal or paternal excess. Under the EBN hypothesis, only *Solanum* species with matching EBN will produce viable seeds. Although first considered to be linked directly with ploidy, the EBN hypothesis has since been adapted as it has emerged that factors other than ploidy, such as the relative expression of genes involved in endosperm development, are important in determining the outcome of crossing.

Cases also exist where EBN values do not need to match for successful endosperm development. For example, *S. verrucosum* has an EBN value of two, but can readily cross with species with EBN values of one with high efficiency, including *S. lignicaule* and *S. dolichocremastrum* (W. Behling et al. 2024). Given that cultivated *Solanum* species also have an EBN value of two or four, *S. verrucosum* has the potential to act as a bridge species to introgress traits from EBN 1 species into agriculturally important cultivars. However, variation in the ability of *S. verrucosum* to act as a bridge to EBN 1 species does exist, depending on the accession of *S. verrucosum* being used and the species that the cross is being performed with (W. Behling et al. 2024). As highlighted in W. Behling et al. (2024), commitment to the EBN hypothesis as a set

rule for breeding compatibility may have led breeders away from useful and viable crosses.

S. verrucosum also has the trait - rare among Solanaceae species - of being self-compatible (W. L. Behling and Douches 2023). Self-incompatibility is a genetic mechanism preventing self-fertilisation and is the norm in many diploid potato species. Self-compatibility is advantageous in terms of plant breeding as it allows the development of inbred lines which can be used to steadily select for advantageous alleles through backcrosses and inbreeding selection. Underpinning self-incompatibility is the *S*-locus - a genetic locus that leads to the inhibition of the growing pollen tube if the S-locus haplotype of the pollen and pistil are the same (McClure, Cruz-García, and Romero 2011). Within the *S*-locus is an *S-RNase* gene which exerts a cytotoxic effect from the female tissue onto the male pollen tube, regardless of the self- or non-self identity of the pollen tube. Degradation of the pollen tube in non-self conditions is mediated by the *S*-locus F box (SLF) genes, which are expressed by the pollen tube and inhibit non-self S-RNases.

Given that S-RNase functions to maintain the barrier of self-pollen tube growth, one factor that can lead to self-compatibility is mutations impacting the function or expression of S-RNase. The tomato species *S. lycopersicum*, *S. pimpinellifolium*, *S. galapagense*, *S.cheesemaniae*, *S. neorickii*, and *S. chmielewskii* are all self-compatible, and even in other species that exhibit self-incompatibility, populations that are self-compatible do still occur (Broz et al. 2021). Underpinning these self-compatible species are deletions, frame-shifts, nonsense mutations, or transcriptional silencing of the *S-RNase* gene that prevent its normal function. In potato, disrupting the *S-RNase* gene with a CRISPR-Cas9 system proved effective at inducing self-compatibility that is stable across generations (Enciso-Rodriguez et al. 2019). The function of S-RNase in potato can also be impacted by distant genetic elements. For example, it was demonstrated that the dominant S-locus inhibitor (*Sli*) gene of self-compatible *S. chacoense* is the result of a MITE transposable element insertion providing a promoter to an F-box protein on chromosome 12, leading to its expression and subsequent repression of S-RNase (Eggers et al. 2021). Simultaneously, a similar study also identified *Sli* in the self-compatible *S. tuberosum* line RH, but attributed its activity to mutations in its protein sequence. Subsequent to these observations, a survey of *Sli* diversity in wild potato species, including *S. verrucosum*, demonstrated that the MITE insertion upstream of *Sli* is occasionally present in several *Solanum* species, but that it is not a reliable predictor of self-compatibility (Ames et al. 2024). The true impact of MITE elements on *Sli* promoter regions remains unclear.

Interestingly, MITE elements are also implicated in modifying the function or activity of S-RNase. In the self-compatible citrus species *Fortunella hindsii*, a MITE insertion into the promoter region of S-RNase inhibits expression, and self-incompatibility is restored when this is deleted (J. Hu et al. 2024). MITE insertions in the S-RNase promoter region were identified in other self-compatible citrus species, although this was not a perfect predictor, and the position of the MITE insertion may be a factor in determining the repression of expression. Certain self-compatible tomato species also possess an upstream MITE insertion, but again this does not appear to be directly linked to the transition to self-compatibility (Broz et al. 2021). That MITE insertions

are present in both citrus and tomato suggests that this relationship might be common in eudicot species.

It appears that the promoter region of S-RNase is a very attractive target of transposable element insertions in general. The self-incompatible *Nicotiana alata* has a Ty3 LTR insertion upstream of its S-RNase and whilst it does not impact expression directly, it may be the causal element behind enrichment of CHH methylation near to the locus in the pistil, where it is expressed in abundance. Recently, a novel mechanism of self-compatibility has been observed in *Poncirus trifoliata* whereby recombination has resulted in the formation of a "super *S*-haplotype" containing two copies of the *S*-locus from independent haplotypes (J. Hu et al. 2024). The two *S*-locus copies cause self-recognition in the pollen, breaking self-incompatibility and leading to self-compatibility. This rare recombination event may have been driven in part by MITE insertions owing to their presence at the recombination breakpoint.

The causal agent of *S. verrucosum* self-compatibility is not known. It has been attributed to a lack of S-RNase protein, although whether this is due to a loss of S-RNase gene function or another mechanism independent of the *S*-locus, is not known. When compared with other species carrying knockouts of S-RNase, *S. verrucosum* still exhibited a greater degree of interspecific compatibility, suggesting that mechanisms other than S-locus regulation are contributing to its phenotype. Being not only self-compatible, which permits inbreeding, but also having a lack of interspecific reproductive barriers, makes *S. verrucosum* a key species for breeding and mobilising important traits from wild germplasm collections, including EBN1 species.

An unexplored avenue in *Solanum* interspecific reproductive barriers is the interplay between DNA methylation and transposable elements (Bozan et al. 2023). DNA methylation is a critical modular of gene silencing in the endosperm, which is necessary to prevent high gene doses leading to incorrect development in the endosperm of wild *Solanum* species (D. Lu, Zhai, and Xi 2022; Roth et al. 2018). Given that DNA methylation also regulates transposable element activity, and transposable element content varies significantly between incompatible potato species, coordination between these factors may be acting as a reproductive barrier in *Solanum* (Bozan et al. 2023). Given the consistency of transposable element insertions close to essential components of the S-locus system, such a hypothesis might not be too much of a reach.

Previous assemblies of *S. verrucosum*

The first available genome for *S. verrucosum* was produced in 2019 following the sequencing of the ver54 line. Whilst the final outcome of this research was a scaffolded genome of *S. verrucosum*, the research itself was largely focused on assessing the performance of different sequencing strategies to resolve plant genomes. Three different sequencing methods were used to provide the initial contigs - a 500bp insert, 250bp paired-end HiSeq 2500 library; a 650bp insert, 100/150bp HiSeq 2500 library; and a 13.5kbp PacBio RSII P6C4 library. The Illumina libraries were highly fragmented, producing contig N50s of 77kbp and 75kbp respectively. These were improved through the use of a 10kbp insert, 500bp Miseq mate-pair libraries, which produced scaffolds of 858kbp and 331kbp respectively. The PacBio library consistently outperformed the

Illumina libraries, with a best N50 of 712kbp.

A variety of scaffolding approaches were then assessed including Dovetail "Chicago" libraries which are analogous to modern Hi-C approaches but create contact matrices for fragments rather than whole genomes, Bionano physical mapping, and a 10X Genomics approach which created barcoded read libraries for individual large DNA fragments. Following comprehensive assessment of these approaches, the method which produced the largest contig N50 of 2.868 Mbp was by initially scaffolding the PacBio library with the Dovetail library, followed by Bionano scaffolding. Despite producing a relatively high-quality assembly for *S. verrucosum*, the genome itself saw little study of its content following its publication.

The second *S. verrucosum* genome was produced in 2022 following the sequencing of a monohaploid clone of *S. verrucosum* 11H23. The approach used was essentially the status quo of genome assembly during this time period - PacBio HiFi sequencing, assembly with `hifiasm`, and scaffolding with Hi-C sequencing. From 46.5Gbp of PacBio Hifi reads and 101 million Hi-C reads, a final assembly of 684 Mbp with a contig N50 of 21.0 Mbp was produced, representing a vast improvement of the previously available assembly. The high contiguity of the assemblies permitted the mapping of previously available *CENH3* ChIP reads to the genome which identified putative centromeres that were noted to have little conservation with the centromeres of *S. phureja*. The genome was further annotated for genes, although notably only using evidence from the *S. phureja* DM genome and *ab initio* gene prediction models pre-trained on tomato. Transposable elements were also annotated using EDTA, and the genome was used to study the synteny and orthologous relationships of the genome to other *Solanum* genomes. One interesting observation from this was that *S. verrucosum* was most closely related to *S. chacoense* despite both having significantly different geographical distributions. Intrachromosomal rearrangements were noted between the two species.

Aims

The primary aims of this chapter are to:

- Produce a high-quality assembly of the *S. verrucosum* genome
- Identify and characterise the *Rpi-ver1* locus providing resistance to *P. infestans*
- Explore the inventory of NLR genes in the context of sequence, expression, and epigenetics
- Determine the presence and status of *S*-RNase
- Characterise the centromeres of *S. verrucosum*

Methods

High-molecular-weight DNA extraction

For all protocols, young flash frozen leaf tissue was prepared by grinding in a mortar and pestle for 10 minutes. Long grinding protocols over more than 20 minutes led to a

noticeable improvement on yield and quality of DNA extracted, with no impact on high molecular weight (HMW) DNA fragment size.

The Promega Wizard® HMW DNA Extraction Kit was used to extract HMW DNA for nanopore sequencing following manufacturer's guidelines. Up to 80mg of tissue could be used in this protocol without any negative affect on final DNA quality.

Circulomics Nanobind Plant Nuclei Big DNA extractions were carried out using the following nuclei isolation protocol. Briefly, 5g of leaf tissue was added to 50mL ice-cold nuclei isolation buffer (1% PVP360 (W/V), 0.25% Triton X-100 (V/V), 0.035M 2-mercaptoethanol, 0.01M Trizma, 0.08M KCl, 0.01M EDTA, 1.7mM Spermidine, 1.7mM Spermine, 0.5M sucrose) and end-over-end for 15 minutes at room temperature. The lysate was gravity-filtered through a 20µm mesh and nuclei were pelleted from the collected liquid phase (7000g, 20m, 4°C). Filtering was repeated until the liquid appeared clear. Nuclei DNA extraction was carried out following Circulomics guidelines.

For all protocols, protein LoBind® tubes were used for all intermediate steps to reduce protein contaminant carryover before final elution into DNA LoBind® to minimise yield loss. If eluted DNA formed a precipitate, it was diluted to <200ngμL^{-1} and gently agitated with a standard P200 tip until clear.

Short read elimination

20µg of HMW DNA was loaded onto a BluePippin 0.75% High Pass cassette with a >15kbp size selection cutoff. Samples were collected and purified with a 1:1 AMPure XP cleanup and a 20 minute incubation at 37°C into 60µL nuclease-free water. 10µg of DNA was thoroughly mixed 1:1 with a size selection solution (3% PVP 360000, 1.2M NaCl, 20mM Tris-HCl, pH 8) and centrifuged (10000g, 30m, RT). The supernatant was removed without disrupting the pellet. The pellet was washed with two 70% ethanol washes and resuspended in 50µL nuclease-free water.

Oxford Nanopore sequencing

Input DNA purity was measured via nanodrop for an optimum OD 260/280 of 1.9, and OD260/230 of 2.0-2.2. Approximate size distribution was measured by running 500ng of DNA on a 0.4% agarose gel, looking for the absence of a smear in the <10kbp region. DNA concentration was measured by Qubit™ BR Assay Kit measurements of the top, middle, and bottom of the sample. If measurements deviated by more than 10%, the DNA sample was mixed eight times with a wide-bore 200µL pipette tip. The absence of RNA was determined via the Qubit™ RNA BR Assay kit.

Nanopore libraries were prepared using the SQK-LSK110 Ligation Sequencing Kit. Three micrograms of high-quality HMW DNA, representing approximately 200fmol assuming 30kbp fragments. DNA was end-repaired by addition of 1µL DNA CS, 2µL NEBNext® FFPE RNA Repair Mix, 3µL Ultra II End-prep enzyme mix, and 3.5µL NEBNext® FFPE DNA Repair Buffer and Ultra II End-prep reaction buffer to 47µL DNA. The reaction was incubated at 20°C for five minutes and 65°C for five minutes. End-repaired DNA was purified with a 1:1 AMPure XP cleanup, with a ten minute

incubation in 61µL nuclease-free water as opposed to the recommended two minute incubation to promote HMW DNA solubilisation.

25µL Ligation Buffer, 10µL NEBNext® Quick T4 DNA Ligase, and 5µL Adapter Mix-F were added to 60µL end-repaired DNA, followed by a ten minute incubation at RT. Adapter-ligated DNA was purified with a 1:0.4 AMPure XP cleanup using Long Fragment Buffer to deplete short fragments, and a 20 minute incubation at 37°C in 15µL Elution Buffer to promote HMW DNA solubilisation.

Library preparation efficiency was determined by Qubit™ HS Assay kit measurements after each cleanup and libraries were immediately taken forward for sequencing. Preferably, 50-75fmol library DNA was loaded onto R9.4.1 minION flow cells or 20-30fmol onto flongle flow cells as moderate overloading appeared to improve pore occupancy with no negative effects. MinIONs were loaded following manufacturer recommendations, and when necessary, cells were washed with Flow Cell Wash Kit EXP-WSH003 and reloaded with a new library.

The recommended protocol for loading flongle libraries directly into the loading port often resulted in an immediate and severe (>90%) loss of pores available for sequencing. An alternative protocol was followed whereby one waste port was sealed with tape, the priming solution dropped *onto* the loading port, and pulled into the flow cell by applying negative pressure with a pipette inserted into the other, unsealed waste port. The tape was then removed, and the remaining priming solution was loaded dropwise onto the loading port, which would now drain into port directly. The library was then loaded dropwise.

PacBio sequencing

20µg HMW DNA was extracted using the Nucleobond protocol and sequenced by the Norwegian Sequencing Centre. Their protocol included the following steps: DNA was fragmented to 15-20kbp using Megaruptor3 and size selected with Bluepippin using a 10kbp cutoff. The library was prepared using SMRTbell® ExpressTemplate Prep Kit 3.0. The library was sequenced on one 8M SMRT cell on a Sequel II using Sequel II Binding kit 2.2 and Sequencing chemistry v2.0. Loading was performed by adaptive loading with a movie time of 30 hours and pre-extension of 2 hours.

Hi-C sequencing

Preparation of material for Hi-C sequencing was carried out using the Dovetail® Omni-C® kit. Approximately 300mg of fresh frozen leaf tissue was ground for more than 10 minutes in a mortar and pestle in liquid nitrogen. Powdered tissue was suspended in 4mL 1x PBS and 50µL 0.3M DSG and rotated for 10 minutes at room temperature. Then, 108µL 37% formaldehyde was added to give a final concentration of 1% and rotated for 10 minutes at room temperature. The sample was centrifuged (5000g, 5m, RT), the supernatant removed, and pellet washed twice with 1X wash buffer. The pellet was resuspended in 1mL 1x wash buffer and successively filtered through 200µm and 50µm filters before being separated into three aliquots. Each aliquot was pelleted (2000g, 5m, RT) and resuspended in 1x nuclease buffer.

Nuclei aliquots were heated to 30°C and dilutions of nuclease enzyme mix were added corresponding to 0.01x, 0.001x, and 0.0005x of nuclease to each aliquot. Aliquots were incubated for 30 minutes at 30°C in a thermal mixer at 1250rpm. The reaction was stopped by addition of 5µL 0.5M EDTA. Cells were permeabilised by addition of 3µL 20% SDS and incubation for 5 minutes at 30°C at 1250rpm.

A 2.5µL aliquot of each digestion was quantified by proteinase K digestion and clean-up with a Zymo DNA Clean & Concentrator™-5 kit. DNA yield was determined using a Qubit™ HS Assay kit and size distribution with an Agilent Bioanalyzer DNA 7500 kit. Digestions were taken forward if >50% of the DNA was between 100-2500bp in length and yield of the total lysate was >1000ng.

For proximity ligation, 500ng of lysate was added to 100µL of chromatin capture beads and incubated for 10 minutes at room temperature. Beads were magnetically separated and washed twice with 150µL 1X wash buffer. While remaining bound to the beads, DNA was end-polished with the addition of 50µL end polishing buffer and 3.5µL end polishing enzyme mix. The bead mix was incubated for 30 minutes at 22°C followed by 30 minutes at 65°C in a thermal mixer at 1250rpm. Beads were washed with 1X wash buffer and resuspended in 50µL bridge ligation mix and 1µL T4 DNA ligase. The ligation mix was incubated for 30 minutes at 22°C at 1250rpm. Beads were washed with 1X wash buffer and resuspended in 50µL intra-aggregate ligation buffer and 2µL intra-aggregate ligation enzyme mix. The beads were incubated overnight at 22°C with shaking at 1250rpm.

To reverse DNA crosslinking, beads were washed with 1X wash buffer and resuspended in 50µL 1X crosslink reversal buffer and 1.5µL proteinase K. Beads were incubated for 15 minutes at 55°C followed by 45 minutes at 68°C at 1250rpm. Beads were magnetically separated, and the supernatant was retained. DNA was purified with a 0.7X AMPure XP bead cleanup and quantified via a Qubit™ HS Assay kit.

To prepare the library for ligation capture, 7µL end repair buffer, 3µL end repair enzyme mix, and 0.5µL 350mM DTT were added to 150ng of DNA in 50µL. The library was incubated for 30 minutes at 20°C followed by 30 minutes at 65°C. Then, 2.5µL Illumina adaptor, 1µL ligation enhancer, and 30µL ligation enzyme mix was added and incubated for 15 minutes at 20°C. Following this, 3µL USER enzyme mix was added and incubated for 15 minutes at 37°C. The library was purified with a 0.8X AMPure XP bead cleanup.

For ligation capture, 95µL of purified library was added to a resuspension of streptavidin beads in 100µL of 2X NTB. The mix was incubated for 30 minutes at 35°C at 1250rpm. The beads were magnetically separated and washed with 200µL LWB, NWB, and 1X wash buffer sequentially. The beads were resuspended in 25µL HotStart PCR Ready Mix, 5µL UDI primer pair, and 20µL nuclease-free water. The mix was incubated in a PCR machine with parameters:

Temperature	Time	Cycle
98°C	3 min	1x
98°C	20 sec	12x

Temperature	Time	Cycle
65°C	30 sec	↓
72°C	30 sec	↓
72°C	1 min	1x
12°C	hold	↓

For size selection, the mix was magnetically separated and 47µL of the supernatant adjusted to 100µL with TE buffer. To this, 50µL AMPure XP beads were added and 145µL of supernatant was removed, to which an additional 30µL of beads were added. The beads were then washed and eluted in 30µL TE buffer and quantified via a Qubit™ HS Assay kit.

Two independent Hi-C libraries were prepared and sequenced with a NextSeq P2 kit with 2x150bp reads.

RNA sequencing

RNA extractions and sequencing were conducted by Amanpreet Kaur. Tissue culture plants of *S. verrucosum* were maintained on MS20 medium and kept in a growth room at a light intensity of 110µmol$m^{-2}s^{-1}$, a temperature of 18±2°C, and a photoperiod of 16/8h light/dark.

Healthy three-week-old plantlets with fully expanded leaves were selected. *In vitro* shoots with roots were gently removed from the media and dipped for one minute in a zoospore suspension of *P. infestans* isolate W9928C adjusted to 4x10^6spores/mL. Dip-inoculated microshoots were gently blotted on sterile paper towels and again planted in fresh MS20 media in vented tissue culture grade glass containers (Generon, UK). The infected plants were kept in darkness for 16 hours and then incubated under the growth conditions mentioned above. The disease severity was recorded by counting the number of leaves showing disease symptoms in 24 hour intervals. The leaf samples from three independent replicates were collected after 0 and 24 hours post infection and immediately immersed in liquid nitrogen before storing at -70°C for further processing.

From each replicate, leaf samples were crushed to a fine powder and 400mg of ground sample was resuspended in 2mL of TRI reagent and vortexed after addition of 10µL ß-mercaptoethanol. The slurry was left to stand at room temperature for five minutes before centrifugation (10,000g for 10min @ 4°C). To the supernatant, 0.2mL chloroform was added (per 1mL) and incubated at room temperature for five minutes before centrifugation (10,000g for 10min @ 4°C). The aqueous layer was transferred, 0.5mL isopropanol added, and transferred to a QIA RNAeasy spin column for washing in RPE buffer twice. RNA was eluted in 50µL RNase free water and the integrity checked with a Bioanalyzer 2100 (Agilent).

For RNAseq, samples were checked for a RIN value ≥8 and were processed at the James Hutton Institute's Genomics facility for generating RNA sequencing libraries using the standard Illumina mRNA Prep kit and Integrated DNA technology (IDT) RNA

unique dual UD Indices as recommended, with 100ng total RNA per sample. Libraries were checked on a Qubit fluorimeter and Bioanalyzer 2100 prior to pooling equimolar amounts before sequencing. Sequencing was conduced on a NextSeq 2000 Sequences at a loading concentration of 750pM using a P3 200 kit, generating paired-end 100bp reads.

Primary genome assembly

The nanopore Canu assembly was produced with `canu v2.2` (Koren et al. 2017) using the options `genomeSize=750m -nanopore`. The HiCanu assembly was produced using `canu v2.2` (Nurk et al. 2020) using the options `genomeSize=750m -pacbio-hifi`. The HiFi Hifiasm assembly was produced using `hifiasm v0.16.1-r375` (H. Cheng et al. 2021) using the options `--primary -t 64`. The HiFi La Jolla Assembler (LJA) assembly was produced using `LJA v0.2` (Bankevich et al. 2022) using the default options.

Genome assembly benchmarking

To benchmark the various assemblies produced, a Snakemake workflow was developed to produce summary statistics including size distribution, BUSCO scores, and merfin k* completeness and QV histograms. The workflow is available at https://github.com/SwiftSeal/assembly_olympics.

Briefly, k-mers were counted from HiFi reads using `kmc v3.2.1` with options `-k21 -t10 -m64 -ci1 -cs10000` and transformed to a histogram using `kmc_tools transform histogram` with option `-cx10000` (Kokot, Długosz, and Deorowicz 2017). The k-mer histogram was used to produce a genomescope profile using `genomescope2 v2.0` with default options (Ranallo-Benavidez, Jaron, and Schatz 2020).

To calculate merfin statistics, HiFi read k-mers were counted using `meryl v2013` using options `-k21` (Rhie et al. 2020). K* completeness was calculated using `merfin v1.0` with `-completeness` mode using the meryl k-mers and the lookup table and kcov value extracted from genomescope (Formenti et al. 2022). QV histograms were produced using the `-hist` mode and plotted using `seaborn v0.11.2` with `kdeplot` (Waskom 2021; Hunter 2007).

BUSCO statistics were calculated using `BUSCO 5.4.3` with lineage `solanales_odb10 2020-08-05` (Manni et al. 2021).

Genome scaffolding

Hi-C reads were aligned to the primary genome assembly contigs using `BWA-MEM` with options `-5SP -T0` to skip pairing and mate rescue and record all alignments (H. Li 2013). Valid ligation events were recorded using `pairtools parse` with options `--min-mapq 40 --walks-policy 5unique --max-inter-align-gap 3` and sorted with `pairtools sort` (Open2C et al. 2023). PCR duplicates were removed using `pairtools dedup`. The final alignment and pair files were produced using `pairtools split`.

The Hi-C alignment data was used to scaffold the genome using `yahs` with option `--no-contig-ec` to prevent contig splitting (C. Zhou, McCarthy, and Durbin 2023). The scaffolded genome was prepared for manual curation the Juicer GUI with the `yahs juicer pre` module and `juicertools` (Durand et al. 2016). Minor corrections for chromosome fusions were made and a final assembly was generated.

To label chromosomes, the final assembly was aligned to the DM1–3 516 R44 (v6.1) (Pham et al. 2020) genome using `mashmap v3.0.6` (Kille et al. 2023) with default settings. The alignment was visualised with the D-GENIES web tool and chromosome identities were inferred.

Repetitive element annotation

`EDTA` transposable elements annotations were generated with `EDTA v2.2.0` with options `--anno 1 --sensitive 1`. EarlGrey annotations were generated with `EarlGrey v4.2.4` using the default settings. To classify LTR elements into clades, libraries were provided to `TEsorter v1.4.6` using the `rexdb-plant` database. To compare the `EDTA` and `EarlGrey` libraries, mean lengths, family counts, and genomic coverage were calculated with `polars` and visualised with `seaborn`.

Tandem repeats were identified with `TRASH` with the options `--win 10000 --m 9000`. Repeats were classified by their homology to the EarlGrey TE library - a fasta of repeats was extracted from the `TRASH .bed` file, a non-redundant library was created with `seqkit rmdup`, and this was used as query against the TE library. Subject hits with the highest bit score were used to classify repeats according to their LTR clade, and manual validation was carried out to ensure no misclassification occurred. Repeat classifications were merged with the original `TRASH .bed` file to produce an informative `.bed` file for use in subsequent analysis.

Gene annotation

Gene models were predicted using `BRAKER3` using RNAseq and protein sequence evidence. All RNA-seq data was aligned to the unmasked genome via `STAR v2.7.10` (Dobin et al. 2013). Individual samples were aligned with the following command:

```
STAR \
--genomeDir $starIndex \
--readFilesIn $infileR1 $infileR2 \
--runThreadN $SLURM_CPUS_PER_TASK \
--outBAMsortingThreadN $SLURM_CPUS_PER_TASK \
--outFileNamePrefix $output_dir/$sampleName"_pass1_" \
--outBAMcompression 10 \
--outSAMattrRGline ID:$sampleName SM:$sampleName PL:Illumina \
--twopassMode Basic \
--alignIntronMin 60 \
--alignIntronMax 15000 \
--alignMatesGapMax 2000 \
--alignEndsType Local \
```

```
--alignSoftClipAtReferenceEnds No \
--outSAMprimaryFlag AllBestScore \
--outFilterMismatchNoverLmax 0.02 \
--outFilterMismatchNmax 999 \
--outFilterMismatchNoverReadLmax 1 \
--outFilterMatchNmin 0 \
--outFilterMatchNminOverLread 0 \
--outFilterMultimapNmax 15 \
--outSAMstrandField intronMotif \
--outSAMtype BAM SortedByCoordinate \
--alignTranscriptsPerReadNmax 30000 \
--readFilesCommand zcat \
--outReadsUnmapped Fastx \
--alignSJoverhangMin 7 \
--alignSJDBoverhangMin 7 \
--alignSJstitchMismatchNmax 0 1 0 0
```

The resulting BAM files were merged, indexed, and sorted with `samtools v1.16.1` (H. Li et al. 2009). Viridiplantae OrthoDB v.11 protein sequences were retrieved from G. Both sets of data were provided as input to `BRAKER v3.0.7` using the `EDTA` softmasked genome.

Helixer gene annotations were generated with `Helixer v0.3.0` using the model `land_plant_v0.3_a_0080.h5`.

Stringtie gene annotations were generated with `Stringtie v2.2.3` using the default settings with the individual `STAR` alignments. Individual predictions were merged with Stringtie's merge function which was used as input for `TransDecoder v5.7.1` using the recommended pipeline of `gtf_genome_to_cdna_fasta.pl`, `gtf_to_alignment_gff3.pl`, `TransDecoder.LongOrfs`, `TransDecoder.Predict`, `cdna_alignment_orf_to_genome_orf.pl` to generate an annotation.

To measure the completeness of gene annotations, `BUSCO v5.7.0` was applied to peptide sequences of each gene prediction.

To generate the final gene annotation, `Helixer` predictions that did not overlap with `BRAKER3` annotations and were predicted to be NLRs by `Resistify` were merged into the `BRAKER3` annotation.

Gene ontology predictions were generated with `EggNOG-mapper v2.1.12` with the default settings.

Deepsignal-plant methylation analysis

Nanopore DNA methylation analysis was carried out using `deepsignal-plant v0.1.6` (Ni et al. 2021). Basecalled fast5 files were re-squiggled with `tombo v1.5.1` (Stoiber et al. 2017) and extracted with `deepsignal_plant extract`. Modifications were called in GPU mode using `model.dp2.CNN.arabnrice2-1_120m_R9.4plus_tem.bn13_sn16.both_bilstm.epoch6.ckpt` as the trained

model.

For the majority of methylation analysis, the resulting methylation frequency data was parsed into a bedgraph format and, when necessary, compressed into the BigWig format using `bedGraphToBigWig v377` (Kent et al. 2010).

Data exploration was carried out with `deeptools v3.5.1` (Ramírez et al. 2016). Binned and scaled data for methylation plots was calculated using `computematrix scale-region` with arguments `-m 3000 -b 3000 -a 3000`. Per-feature mean methylation levels were calculated with `bedtools v2.31.1 map`.

RNA-seq analysis

RNAseq mapping and read count estimation was carried out using the NF-core RNAeq analysis pipeline, using the default settings.

The `salmon` quantified read counts were imported into R with the `tximport v3.19` `tximport()` function and all subsequent analysis was conducted with `deseq2 v3.19`. Infection, tissue specific, and temperature response assays were analysed in separate RNA-seq experiments. Differential expression analysis was conduction with `DESeq()` using the default settings, followed by `lfcShrink()` with the `apeglm` shrinkage estimator. Genes were considered as differentially expressed if $p_{adj.} < 0.01$ and $|\log2(FC)| > 1$.

NLR analysis

NLRs were identified and classified with `Resistify v0.2.2` with the option `--ultra` enabled to identify partial NLRs. NLR phylogenetic trees were built by aligning the NBARC domains extracted by `Resistify` with `MAFFT v7.525` using the default settings. The best model - JTT+G4 - was selected with `ModelTest-NG v0.1.7` and built with `RAxML-NG v1.2.2` providing the multiple sequence alignment as input.

Homologs of known NLRs were identified by a `BLASTp` search against the RefPlantNLR database. Genes were considered as homologs if they exceeded a percentage identity of more than 85%.

Differential expression and $\log_{10}$(TPM) values were taken from the analysis in sec. . Statistical analysis of expression was conducted with `statsmodel v0.14.1` in a single OLS model of the formula `lcpm ~ condition + helixer`.

To identify the *Rpi-ver1* locus, previously generated KASP markers were mapped to the genome with `BLASTn` and filtered for hits with 100% identity and query length (X. Chen et al. 2018). The *Rpi-ver1* locus was determined to be the locus delimited by the high-confidence KASP markers. To verify that the locus fully represented *Rpi-ver1*, bulk segregant RenSeq reads from the original study were mapped to the *S. verrucosum* genome and filtered with the expected homozygosity for the parent datasets, and heterozygosity for the F_1 progeny bulks.

To screen for candidate *Rpi-ver1* genes, `Resistify --ultra` motif annotations were used to identify potential NLRs. Eggnog GO terms were also scanned for GO terms

associated with disease resistance or known *P. infestans* resistance pathways. To filter out low priority candidates, genes with a mean TPM < 1 were discarded.

Hi-C analysis

Reads were aligned to the scaffolded genome and processed as described in sec. . The Hi-C contact matrix was built using `cooler v0.9.2` (Abdennur and Mirny 2019) with the command `cooler cload pairs` and the final `.mcool` was built using `cooler zoomify`.

For the final assembly, Hi-C analysis was conducted with `nf-core/hic v2.1.0` using the default settings (Servant et al. 2023, 2015; P. A. Ewels et al. 2020; P. Ewels et al. 2016). Visualisations were produced with `cooltools v0.7.0` (Open2C et al. 2022).

Centromere analysis

Centromeres were identified by aligning CENH3 ChIP reads against the genome using `bowtie v2.5.3` (Langmead and Salzberg 2012) with default settings. Reads were obtained from a previous project (H. Zhang et al. 2014). To define centromere boundaries, CENH3 mapping was visualised in `igv` and centromeric regions were selected manually.

To identify transposable elements significantly associated with centromeres, contingency tables of counts of LTR families inside and outside were calculated and used as input for fisher exact tests with `scipy v1.13.1` using the function `fisher_exact`. Families with a p-value < 0.05 and odds ratio > 1 were considered as being potentially centromere-biased.

Organelle assembly

Organelles were assembled from HiFi reads using `oatk v1.0` with the setting `-c 100` to exclude syncmers below 100x coverage. The plastid genome was annotated with the `GeSeq` webtool.

Characterisation of the S-locus

To identify the homolog of *S-RNase* in *S. verrucosum*, known *S-RNase* sequences from *S. neorickii* and *S. chilense* (`BAC00940.1` and `BAC00934.1`) were queried against the final annotation protein sequences of *S. verrucosum* with `BLASTp v2.9.0`. The *S-RNase* homologs were determined by its higher percentage identity as compared to non- S-locus RNases, and the genes location on chromosome 1 which was the anticipated position of the S-locus. As no RNAseq reads were available as evidence for its annotation, the gene model produced by `BRAKER3` was compared to the `Helixer` annotation and verified as identical.

Similarly, SLF homologs were identified by `BLASTp` searches with *Petunia integrifolia* SLF `AAS79485.1` as a query. Homologs were filtered by their presence on chromosome 1.

Results

The *S. verrucosum* genome is highly contiguous

The genome of *S. verrucosum* clone 54 was highly homozygous with an estimated homozygosity of 88.5% (fig. 11). To obtain the highest quality primary assembly of *Solanum verrucosum*, multiple sequencing technologies and assembly methods were compared. Of the multiple assemblies produced, a `hifiasm` assembly using solely HiFi reads resulted in the highest-quality primary assembly with a contig N50 of 46.3 MB (# 1667) and k* completeness of 0.938 tbl. 5.

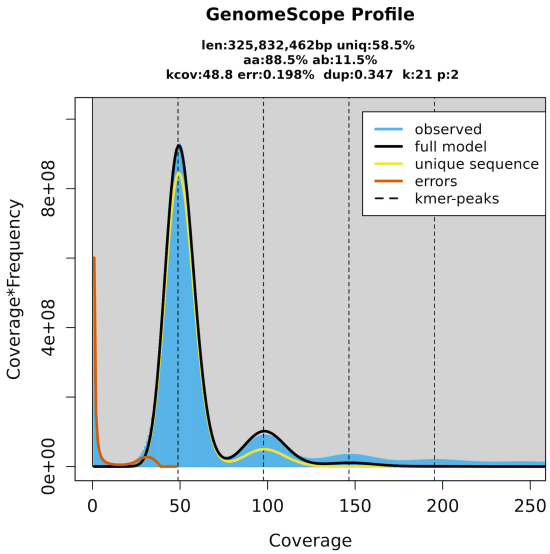


Figure 11: **GenomeScope profile**. Transformed linear model of HiFi *21*-mer coverage across the assembly with key calculated statistics highlighted.

The BUSCO score was also high, capturing 98.5% complete BUSCOs (single: 95.8%, duplicated: 2.7%). The density of K* values was also tightest for the hifiasm assembly, indicative of a low rate of collapsed and expanded k-mers. That several of the higher quality assemblies did not exceed a BUSCO completeness of 98.5% suggests this may be the upper limit for *S. verrucosum*.

From communications with other users of `hifiasm`, it was noted that breakpoints in the assembly were not necessarily due to overly complex regions that could not be confidently assembled. To verify this and further increase the contiguity of the assembly, `quickmerge` was used to merge the HiFi-based `hifiasm` and ONT-based `flye` assemblies. This resulted in a moderate increase in contiguity which was manually validated to be the result of valid merges.

Table 5: **Primary assembly statistics.** Statistics of assembly programs tested. The `hifiasm`, `LJA`, and `HiCanu` assemblies were produced using the HiFi reads. The `Canu` assembly was produced with Nanopore reads.

Assembly	Completeness	# Contigs	Contig N50	Total Length	BUSCO (%)
`hifiasm`	0.9383	1667	46.3 Mbp	753.3 Mbp	98.5
`LJA`	0.8863	991	22.1 Mbp	713.2 Mbp	98.5
`HiCanu`	0.9037	4043	18.8 Mbp	803.0 Mbp	98.5
`Canu`	0.7163	1490	2.8 Mbp	728.3 Mbp	94.2

Two circular mitochondrial genomes were assembled which were 417.9kbp and 49.3kbp in size respectively. A single circular chloroplast genome of 155.5kbp was also assembled (fig. 12). This was in line with a recent 155.5kbp assembly of the *S. verrucousm* chloroplast which curiously was lacking in a *ycf1* annotation at the IRB-SSC boundary, (L. Zhang et al. 2024).

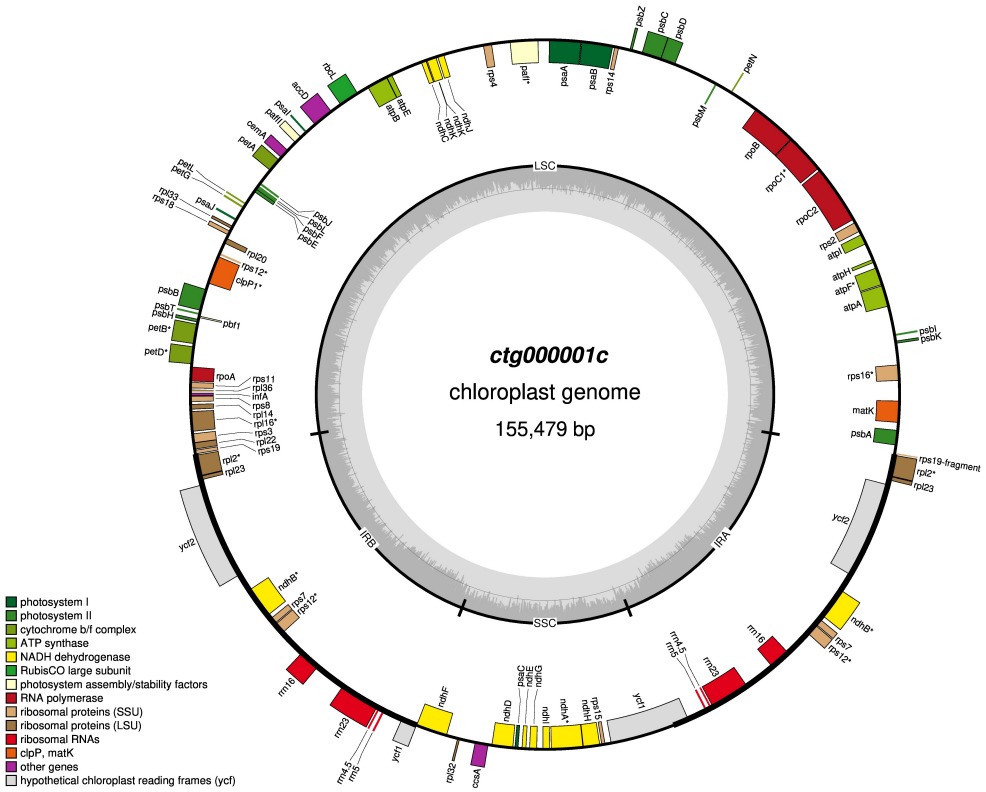


Figure 12: **The complete chloroplast of *S. verrucosum*.** Annotations and figure produced by the `GeSeq` webtool. GC content is indicated by the inner grey ring. Annotated genes are displayed on the outer ring and coloured by group according to the key presented.

Multiple genomic features including DNA methylation, genes, and transposable elements were mapped across the genome (fig. 13). These showed distinct correlations

across the genome - DNA methylation and Ty3 LTR content increased towards the centromeric regions, whilst gene content decreased.

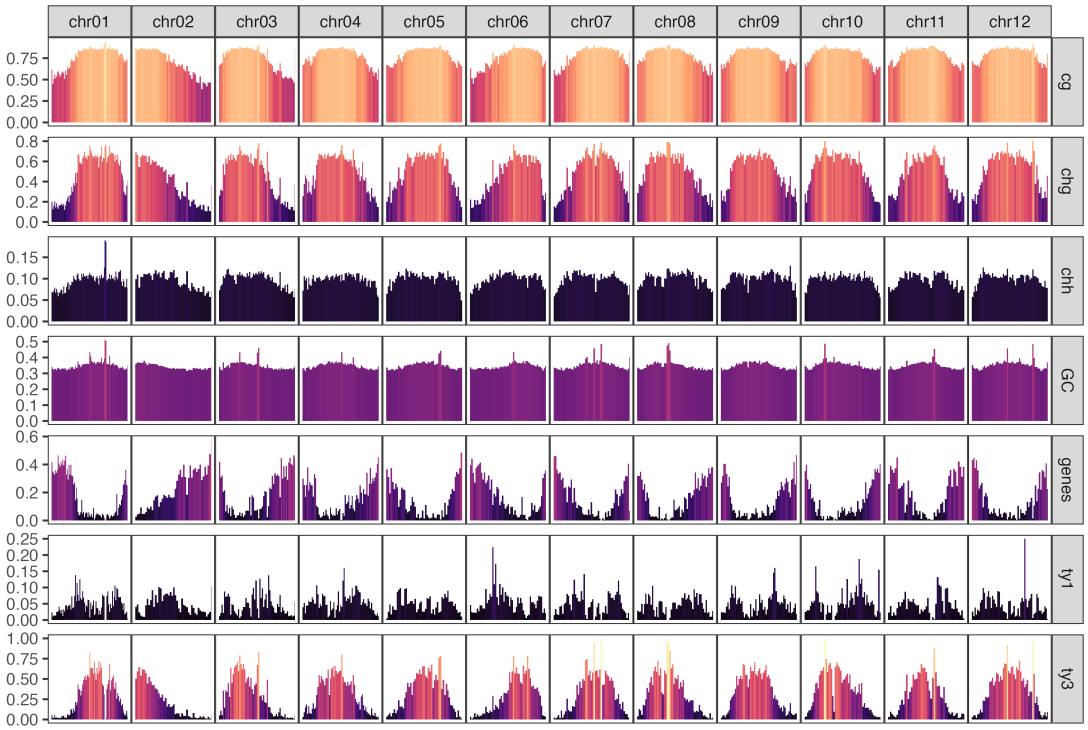


Figure 13: **Landscape of the *S. verrucosum* genome.** From left to right: chromosome 1 to chromosome 12. From top to bottom: CG, CHG, and CHH methylation, GC content, genes, Ty1 and Ty3 LTRs. Y-axis are variable per-feature, colour gradient is proportional from 0-1 for all features. Features counter per 100Kbp windows.

Multiple gene annotation strategies were tested. Of these, a combination of BRAKER3 and `Helixer` produced the most comprehensive annotation, producing 38,710 genes with a mean gene length of 3043bp, 1.2 transcripts per gene, 4.3 exons per transcript, and a multi:single exon ratio of 2.75.

Recently, it has been highlighted that genes in plant genomes can be classified according to the levels of CG and CHG methylation in their exons. Genes unmethylated in either context are unmethylated (UM). Genes exhibiting predominantly CG methylation in their exons can be classified as gene body methylation (gbM). Genes with high levels of both CG and CHG methylation are referred to as TE-like methylation (teM).

The distribution of genes according to their exon methylation in the CG and CHG contexts showed a similar distribution to what has been observed previously in Maize, demonstrating that the nanopore-derived methylation signals are applicable to *Solanum*, and that *Solanum* also exhibits a similar trend (fig. 14 b). In the Maize analysis, genes were categorised into gbM, UM, and teM categories based on the following rules:

- UM = CG <= 0.05, CHG <= 0.05
- gbM = CG > 0.2, CHG <=0.05

- teM = CG > 0.4, CHG > 0.4

Given that the distribution of methylation in *S. verruocusm* was similar to that of maize, the same criteria were selected here. Accordingly, 5940 genes were classified as UM, 17151 as gbM, 8076 as teM, and 7957 as ambiguous. The proportion of UM genes was substantially lower than in the Maize dataset, whilst the proportion of gbM and teM genes was higher. The higher proportion of teM genes could be explained at least in part by the lack of curation in the gene annotation presented here - the number of teM genes was reduced significantly in the Maize dataset when considering only core genes, suggesting that teM genes might be pseudogenes or misannotations. The increased proportion of gbM is however unexplained.

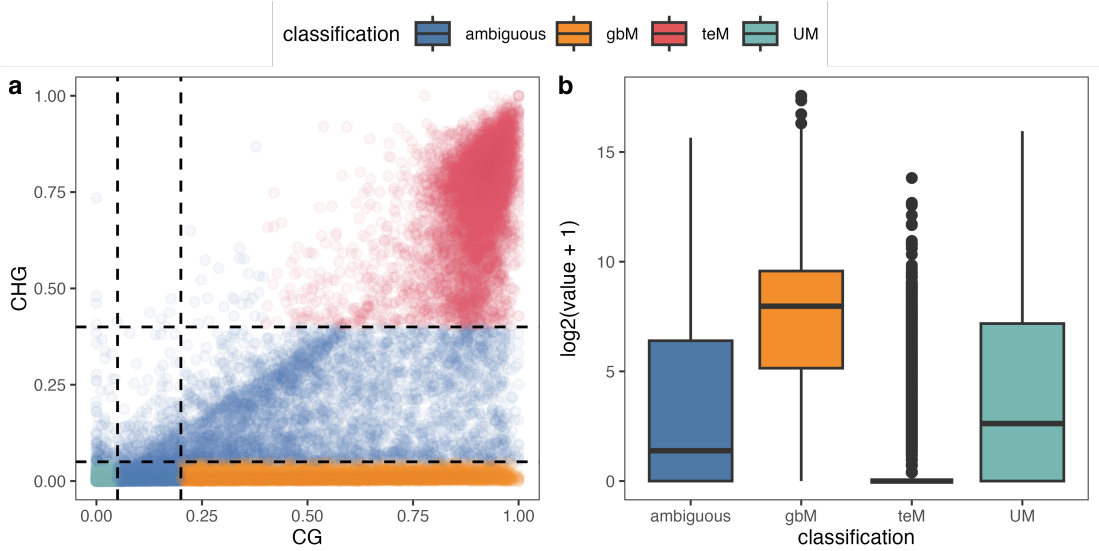


Figure 14: **Gene body methylation**. a) Scatterplot of genes according to the proportion of CG and CHG methylation in their exons. Lines representing the categorisation thresholds have been drawn. b) Boxplot of gene leaf expression for each methylation state.

The average level of methylation in the genome was 70%, 46.2%, and 9.4% in the CG, CHG, and CHH contexts, respectively. This is in line with previous estimates for *S. lycopersicum* and *S. melogena* (Cui et al. 2021; Y. Lu et al. 2021). The maize genome is 86.4%, 70.9%, and 1.2% methylated in the CG, CHG, and CHH contexts, which perhaps indicates that overall levels of methylation in the genome do not strictly correlate with increased gene methylation (West et al. 2014). Given that Maize has a similar number of genes (42,580) but a larger genome (2.42Gb), the higher methylation proportion is likely to have been driven by a larger repeat content (88.37%) (Jian Chen et al. 2023).

In leaf tissue, the expression of gbM genes was significantly higher than UM (ß=3.29 $\log_2$(TPM+1), SE=0.05, $p<2e^{-16}$), and the expression of tEM was significantly lower (ß=-3.38 $\log_2$(TPM+1), SE=0.06, $p<2e^{-16}$) (fig. 14 b). Previously, unmethylated genes have shown to mostly exhibit tissue-specific expression (Zeng, Dawe, and Gent 2023). A similar significant trend was not seen in *S. verrucosum* when expression levels between root, shoot, and leaf tissue were examined where 22127, 3872, and 9787

constitutively expressed, tissue specific, and silent genes were identified. The probable cause of this is that fewer tissue-specific RNAseq conditions are available in this study in comparison to the previous (leaf tip, leaf middle, leaf base, root, shoot, ear, anther, tassel, endosperm, and embryo), leading to fewer genes being identified as being "tissue specific".

The repeat landscape of *S. verrucosum*

To characterise the repetitive fraction of the *S. verrucosum* genome, the performance of the `EarlGrey` and `EDTA` TE annotation pipelines were first compared. Each pipeline uses a slightly different nomenclature for TE classifications, owing to their internal architecture. To aide in comparison, all classifications were reduced to being either an LTR, LINE, SINE, DNA, *Helitron*, or 'Other' type of TE classification (fig. 15).

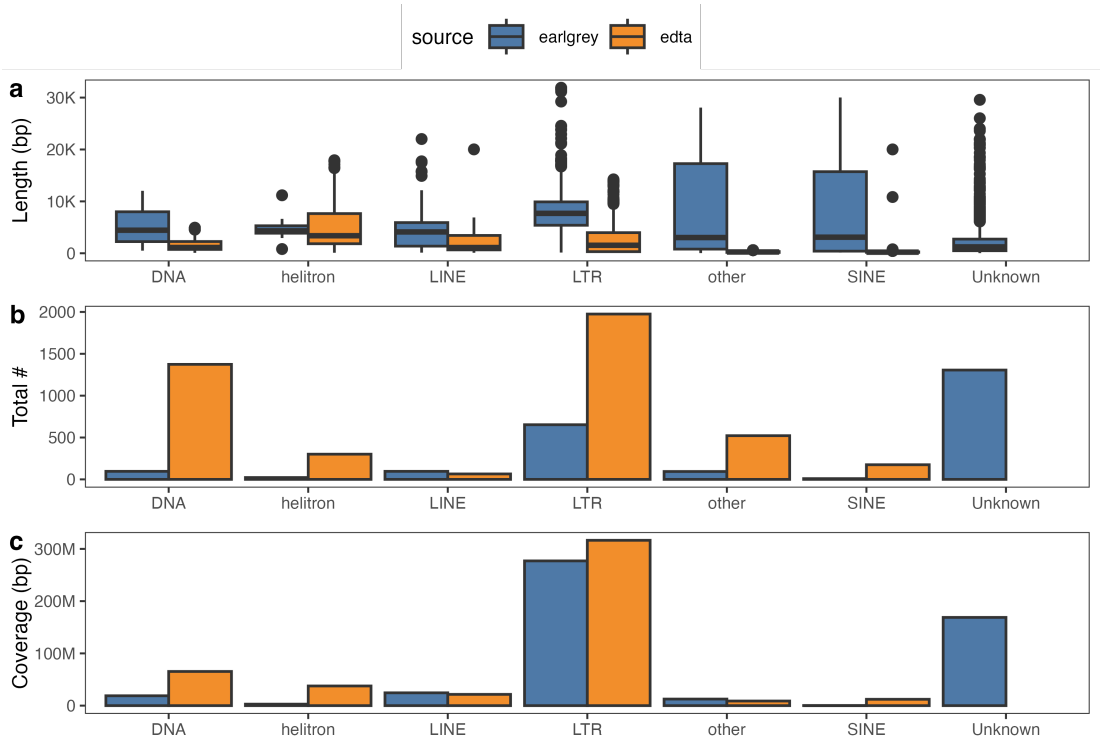


Figure 15: **Transposable element annotation comparison.** Summary statistics for `EDTA` (blue) and `EarlGrey` (orange) are presented across different families of transposable elements.

EarlGrey classified 67.6% of the genome as being repetitive, whereas `EDTA` classified 56.7%. As expected, the largest proportion of repeats were classified as being LTR-derived with both pipelines identifying similar proportions of 39.7% and 36.6%. The fraction of LINE, SINE, and *Helitron* elements deviated between pipelines, notably an additional 4.2% of the genome being identified as *Helitron*-derived by `EDTA`. The largest source of variation in the fraction of genome as being repetitive was due to the 22.3% of genome classified as "Unclassified" by EarlGrey.

The number of unique families identified varied considerably for each TE classification.

EarlGrey identified fewer LTR families (n = 632) than EDTA (n = 1975) and they were of significantly greater mean length. A similar trend was also seen for DNA, *Helitron*, and SINE elements, but not LINEs, which EarlGrey identified a larger number of families for.

To further assess the completeness of the TE libraries assembled by both tools, TEsorter was applied to classify TEs by their domains. The fraction of LTRs that failed to be classified by TEsorter was 63.1% and 9.8% for EDTA and EarlGrey respectively. TEsorter further classified 38.6% (n=252) of LTRs as being complete in the EarlGrey library, but only 15.5% (n=306) in the EDTA library. This indicates that the TE library produced by EarlGrey is substantially more complete than the library generated by EDTA. Three *Helitron* families identified by EarlGrey contained both HEL1 and HEL2 domains, indicating the presence of functional *Helitron* elements in the genome and their successful identification. In the EDTA library, give families were identified as containing HEL1/HEL2 domains, and a further five were identified which contained either a HEL1 or HEL2 domain.

The smaller number of TE families, greater mean length of LTR families, and higher rate of successful LTR classification by TEsorter were taken as evidence of EarlGrey outperforming EDTA at producing a high-quality TE library. As a result, the EarlGrey library was taken forward for further examination.

In the genetic history of *S. verrucosum*, two potential bursts of LTR insertion are noted, indicated by the two peaks in the repeat landscape (fig. 16). The most recent burst also appears to have been associated with DNA type transposable elements. Interestingly, upon closer examination of the LTR history it appears that the recent activity was associated with Tekay, Athila, and CRM Ty3 LTRs, whereas the Ty1 Clades TAR, Ikeros, and Bianca contributed to a peak of activity approximately between the two transposable element bursts.

As methylation data was derived from long nanopore reads, the methylation profile of transposable elements could be accurately profiled (fig. 17). In general, transposable elements elevated levels of DNA methylation in all three contexts, although profiles varied substantially between transposable element families. Interestingly, hypermethylation in the CHH methylation context were more apparent for some, but not all, DNA type transposable elements, as well as members of the SINE/tRNA-Deu-RTE family. The boundaries of transposable elements also varied in their definition - high copy number classifications such as the Ty3 and Ty1 LTRs had clear methylation boundaries, whereas lower copy number classifications in the DNA and LINE groups often had ill-defined boundaries or messy profiles.

Given the variability of DNA methylation amongst transposable elements, it was reasoned that they could be classified according to their CG and CHG proportions as was the case for the genes. Accordingly, a clear majority of transposable elements families exhibited elevated levels of CG and CHG methylation as is generally expected for transposable elements (fig. 18). Many families of the LINE and *Helitron* classifications were not methylated in the CHG context, potentially indicating that they are not true positive annotations.

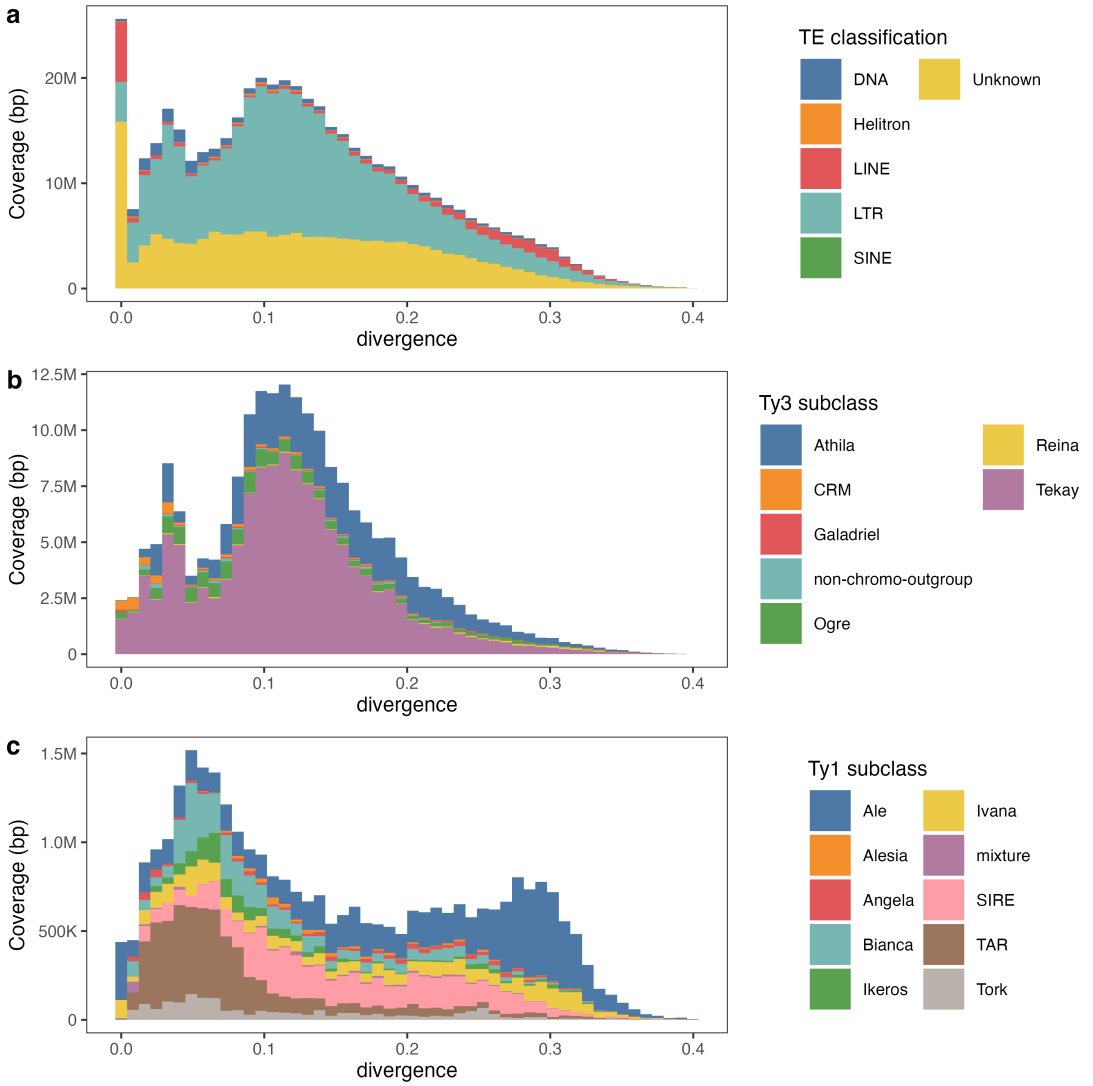


Figure 16: **Repeat divergence in *S. verrucosum*.** a) The density of transposable element classifications by their divergence from family consensus. b) Divergence of Ty3 LTR clades. c) Divergence of Ty1 LTR clades. Note that the y-scale has changed by a factor.

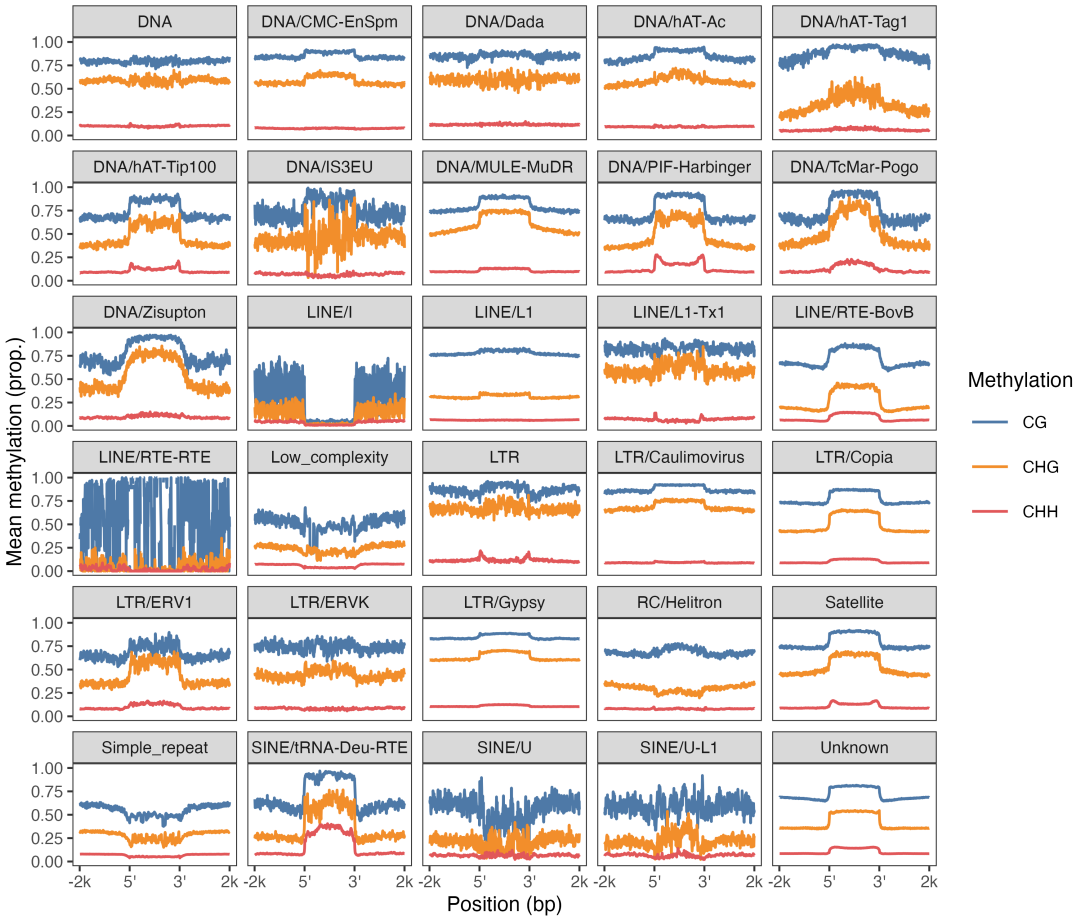


Figure 17: **Transposable element methylation profiles.** The mean proportion of DNA methylation in CG, CHG, and CHH contexts across transposable element classifications derived from EarlGrey. Profiles encompass +/-2kbp around the annotated element.

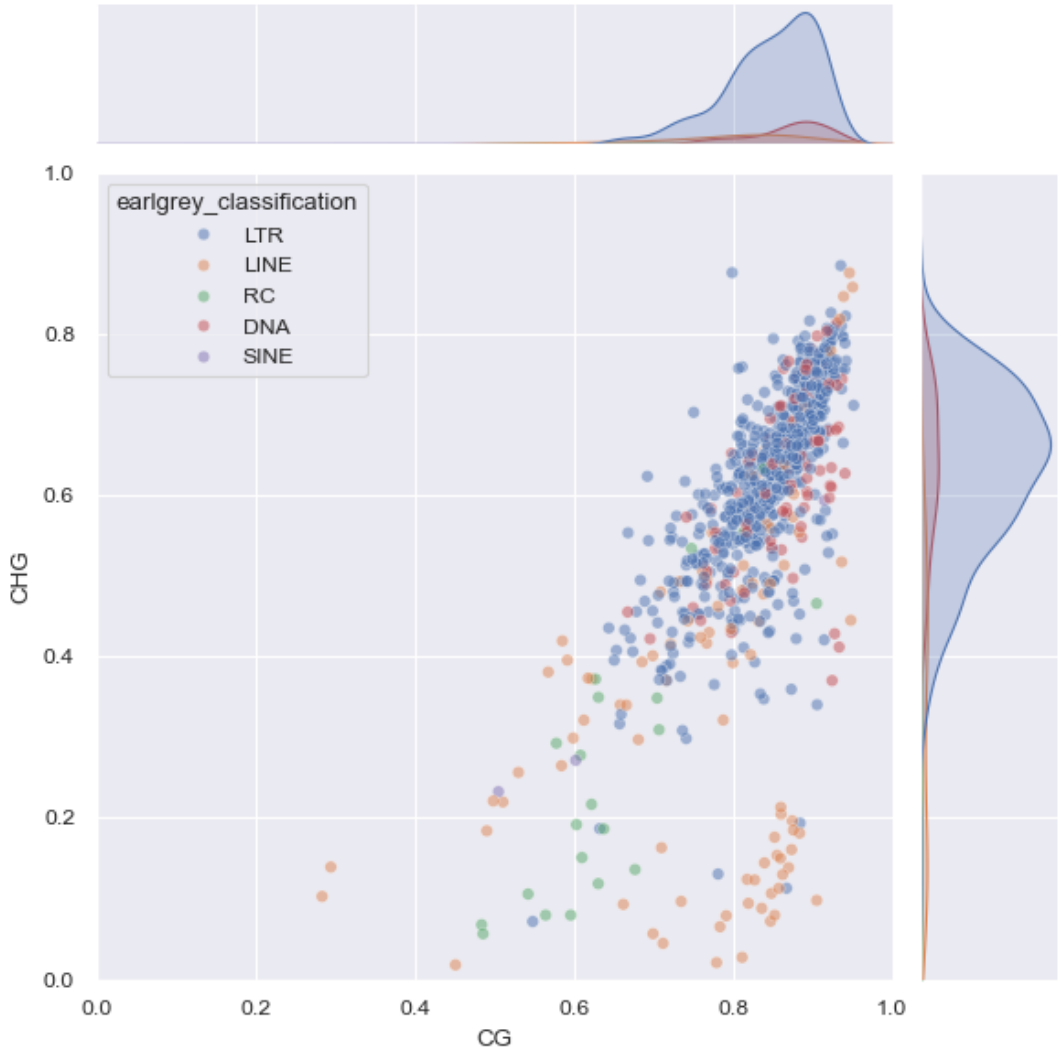


Figure 18: **Transposable element methylation ratio.** EarlGrey derived transposable element families distributed by their mean CG and CHG methylation levels.

RNAseq analysis

High depth RNAseq libraries with repeats were developed for root, shoot, and leaf tissue, infected and uninfected *P. infestans* conditions, and at 4°C, 25°C, and 35°C (tbl. 6). Whole plantlets were used for infection libraries, whereas leaf tissue was used for temperature conditions. Repeats within libraries clustered tightly and distinct clusters were evident for root, shoot, and infected samples (fig. 19 a). To identify differentially expressed genes, contrast groups were established comparing tissue levels, infection status, and temperature differences. Within contrast groups, the number of differentially expressed genes varied between contrast groups (fig. 19 b). A large number of genes were differentially expressed following incompatible infection, and differential expression was also noted for the various tissue levels. The smallest differential expression result was within the cold stress group.

Table 6: **RNAseq libraries prepared for *Solanum verrucosum*.** For infection conditions, *in vitro* roots and shoots were treated with *P. infestans* isolate W9928C, followed by RNA extraction at the specified timepoint from whole plantlet tissue. For temperature conditions, RNA was extracted from leaf tissue.

Condition	Rep	Total reads	Duplicates	Uniquely mapping
Infection 0hpi	1	53,925,964	36%	90%
Infection 0hpi	2	62,270,652	38%	89%
Infection 0hpi	3	68,241,131	35%	90%
Infection 24hpi	1	59,017,589	41%	86%
Infection 24hpi	2	56,916,433	39%	88%
Infection 24hpi	3	65,742,064	40%	89%
Leaf	1	55,457,446	35%	86%
Leaf	2	55,866,776	36%	88%
Leaf	3	56,785,668	37%	90%
Root	1	54,483,222	35%	90%
Root	2	61,491,956	38%	88%
Root	3	61,258,248	35%	92%
Shoot	1	53,692,166	34%	90%
Shoot	2	60,784,503	36%	89%
Shoot	3	55,940,978	35%	92%
25°C	1	61,537,559	39%	87%
25°C	2	58,842,187	37%	87%
25°C	3	55,652,255	35%	92%
35°C	1	56,643,303	35%	86%
35°C	2	59,834,457	32%	88%
35°C	3	62,788,765	35%	89%
4°C	1	58,432,736	38%	88%
4°C	2	61,348,502	39%	88%
4°C	3	57,388,920	36%	90%

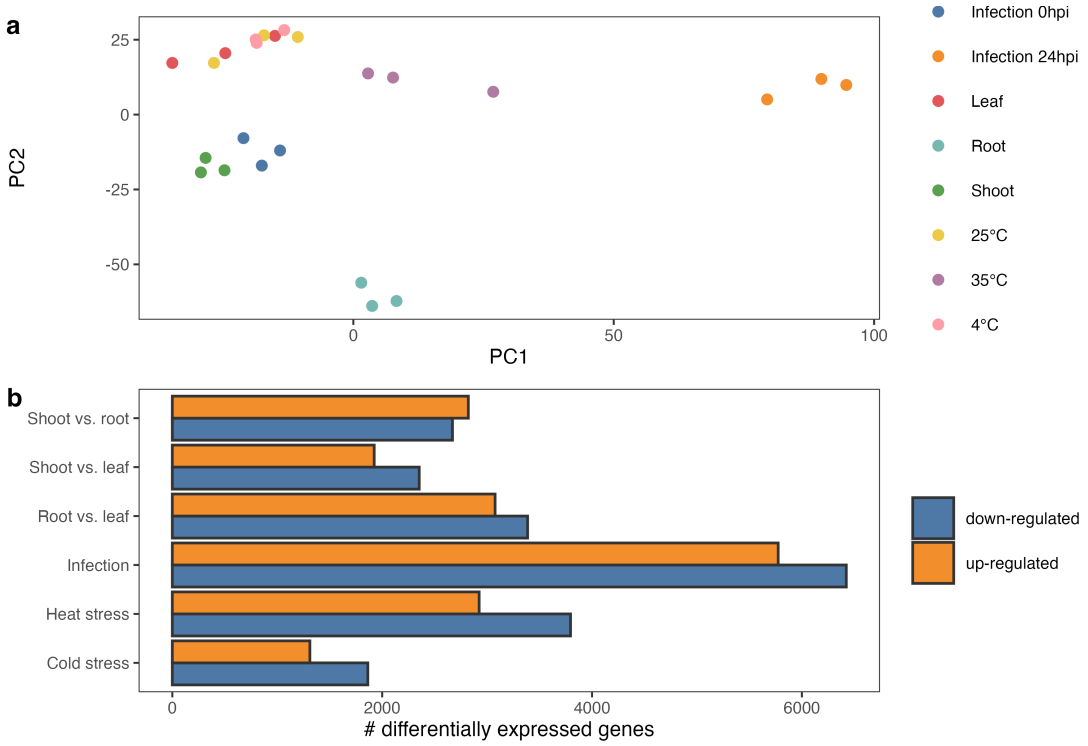


Figure 19: **RNAseq analysis.** a) Sample distribution in a PCA of the blind dispersion estimate from DESeq2. b) Differentially expressed genes in each contrast group. Genes are considered differentially expressed when $p_{adj.}$<0.01 and $|\log_2(FC)| > 1$.

Taking inventory of resistance genes

In total, 502 NLRs were identified fig. 20. Of these, 248 were identified as canonical CNLs, 45 as TNLs, and two as RNLs. MADA and C-JID domains were also identified - 11% of CNLs contained a MADA domain whilst 52.6% of TNLs had a C-JID domain. At a maximum gap width of 30kbp, 223 NLRs were identified in clusters of more than two NLRs, whilst 283 NLRs were identified as singletons or pairs.

Of the identified NLRs, 23.1% exhibited low to no expression across all RNAseq conditions explored in this study. A linear model indicated that NLRs identified exclusively by Helixer exhibited significantly lower expression (ß = -2.76 $\log_2$(TPM), SE=0.11, p=2.8e-137), potentially indicating an expression-led bias in misidentification by BRAKER3. Overall NLR expression showed little variation between biotic and abiotic conditions except for 24 hours following *P. infestans* infection, where NLR expression was significantly reduced (ß = 01.44 $\log_2$(TPM), SE=0.22, p=2.03e-11).

This trend was reflected in the number of differentially expressed NLRs identified. Conditions which induced the largest shift in NLR expression were root versus leaf tissue, root versus shoot, and 24 hours following *P. infestans* infection. Expression changes between tissue levels were mostly balanced, with a similar proportion of upregulated and downregulated NLRs identified. This is at odds with a recent report of root-biased NLR expression in *S. lycopersicum* (Lüdke et al. 2023). Post infection, the majority of differentially expressed NLRs were downregulated. Clustering of

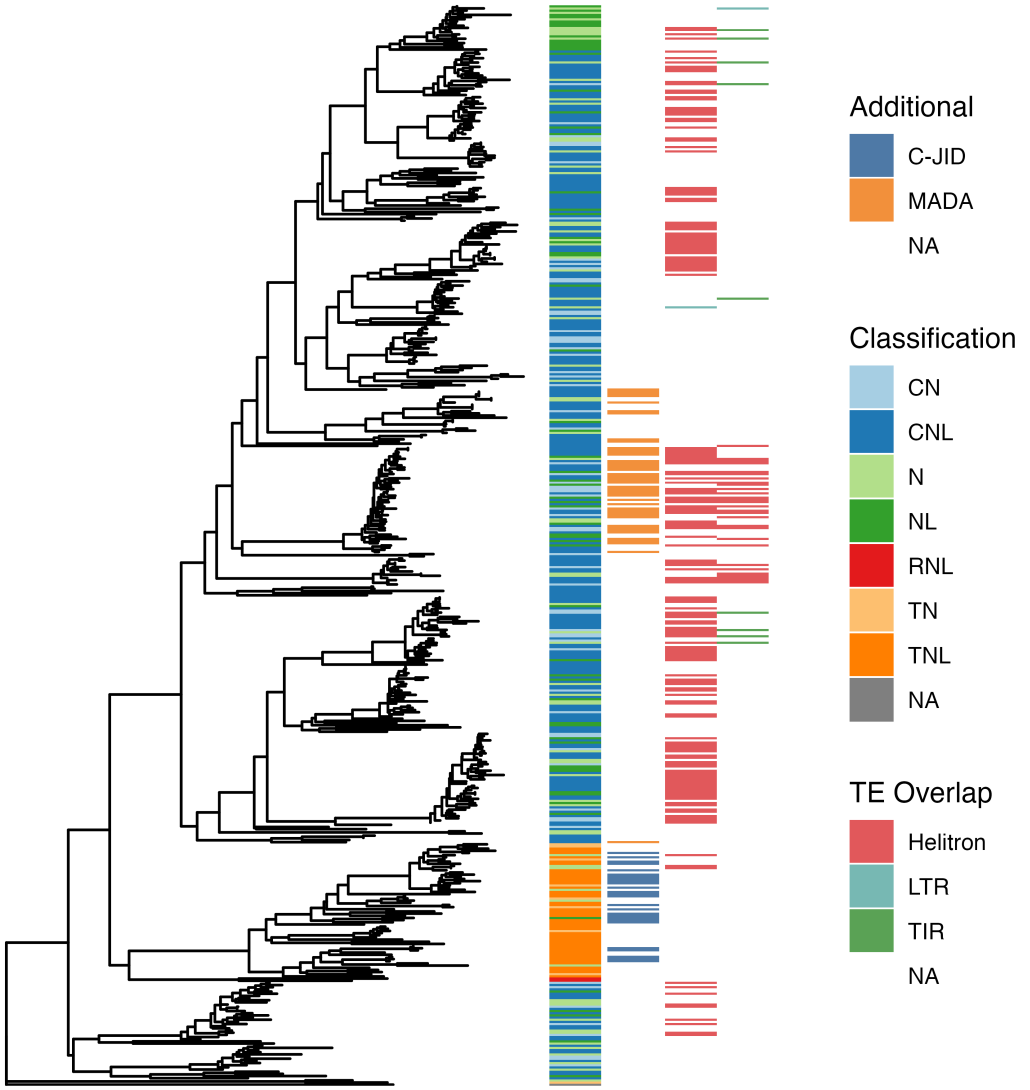


Figure 20: **The NLRs of *S. verrucosum*.** Phylogenetic tree rooted on *C. elegans CED-4*. Annotations, from left to right - NLR classification according to `Resistify`, Presence of C-JID domain or MADA motif, `EarlGrey` TE overlaps, and `EDTA` TE overlaps.

NLRs by their expression across conditions identified several distinct profiles. Most clusters exhibited uniform expression of NLRs across conditions with varying levels of expression. Root expression was somewhat of an outlier across multiple clusters, exhibiting either reduced or enhanced expression.

When clustered by their expression profiles across conditions, distinct clusters of NLR expression emerged (fig. 21). The largest cluster 1 was comprised of NLRs that exhibited little (LCPM < 1) to no expression across all conditions which included homologs of functional NLRs. Clusters 11, 12, 13, and 14 were comprised of NLRs exhibited very high expression and included homologs of *NRC1*, *NRC2*, *NRC3*, and *NRC4a*, and the *Pseudomonas syringae* resistance gene *Prf*. Recently, root-specific expression of a cluster of NLRs has been identified in *S. lycopersicum*, including *Hero* and *NRC6* (Lüdke et al. 2023). A similar pattern was not seen here - whilst three *NRC6* homologs exist in *S. verrucosum*, they do not exhibit root-specific expression and instead span three separate clusters of varying NLR expression. Whilst several clusters to exhibit root specific over- or underexpression, none of them contained NLR homologs that indicated a helper/sensor co-expression network. Factoring into this is the lack of a skew of NLR expression towards roots highlighted in the article that is not present here.

In Chapter 1 I reported a previously undescribed association between NLRs undergoing expansion in *Solanacaeae* and *Helitron* transposable elements: this was followed up using both the `EDTA` and `EarlGrey` annotations produced in this study fig. 20. For `EDTA`, a similar trend was seen – 29 NLRs were enclosed by *Helitrons*, seven by TIR transposons, and one by a Ty1 LTR retrotransposon. Of the overlapping *Helitron* annotation, four of these were complete elements with intact *Helitron* motifs, the remaining 25 were annotated by homology. For `EarlGrey`, 167 NLRs were identified to be enclosed by *Helitrons*, and one by a Ty1 LTR retrotransposon. Of the four intact *Helitrons* identified by `EDTA` that overlapped with NLRs, only one was also identified by `EarlGrey`.

To remap the *Rpi-ver1* locus in the new assembly, pre-existing KASP markers linked with the resistant phenotype were mapped to the assembly. As expected, all KASP markers mapped at least once to the genome with 100% identity, and all to the distal arm of chromosome 9. The marker `DMG400017237` mapped with equal E-values to multiple loci in close proximity on chromosome 9.

In total, the KASP markers defined a locus 11.5 Mbp in size, from 47.1 Mbp to 58.6 Mbp. The two KASP markers that defined the highest confidence locus in the original publication, `DMG400017237` and `DMG400017146`, defined a locus 1.3 Mbp in size, from 54.6 Mbp to 56.0 Mbp. This was a significant reduction in locus size in comparison to the 4.3 Mbp locus reported in the original publication that was based on the DM reference genome.

Table 7: **BLASTn result of KASP markers against the *S. verrucosum* genome.**

KASP marker	Chrom.	ID (%)	Length	Start	End	E-value
DMG400012878	chr09	99.01	101	46494373	46494273	2.42e-45

KASP marker	Chrom.	ID (%)	Length	Start	End	E-value
DMG400019345	chr09	100	101	47056573	47056473	1.87e-46
NLR0215	chr09	100	101	47796311	47796211	1.87e-46
DMG400031427	chr09	100	101	48156538	48156438	1.87e-46
DMG400010295	chr09	100	101	48630799	48630699	1.87e-46
DMG400003805	chr09	100	101	49747933	49747833	1.87e-46
DMG400016850	chr09	100	101	50196342	50196242	1.87e-46
DMG400011361	chr09	100	101	50704065	50703965	1.87e-46
DMG400011401	chr09	100	101	51196801	51196701	1.87e-46
DMG400011401	chr03	90.164	61	59315195	59315136	1.17e-13
DMG400017237	chr09	100	101	54606140	54606040	1.87e-46
DMG400017237	chr09	100	101	54656340	54656240	1.87e-46
DMG400017237	chr09	100	101	54725451	54725351	1.87e-46
DMG400017237	chr09	88.889	99	54554664	54554566	5.35e-27
DMG400017237	chr09	88.889	99	54564259	54564161	5.35e-27
DMG400017237	chr09	88.889	99	54614461	54614363	5.35e-27
DMG400017237	chr09	88.889	99	54683570	54683472	5.35e-27
DMG400017237	chr09	87.879	99	54561459	54561361	2.49e-25
DMG400017237	chr09	87.879	99	54611661	54611563	2.49e-25
DMG400017237	chr09	87.879	99	54754218	54754120	2.49e-25
DMG400017237	chr09	86.869	99	54761977	54761879	1.16e-23
DMG400017146	chr09	100	101	55951434	55951334	1.87e-46
NLR0226	chr09	100	101	58124547	58124447	1.87e-46
NLR0226	chr09	97.143	35	57906076	57906110	4.25e-08
DMG400031521	chr09	100	101	58562723	58562623	1.87e-46
DMG400031521	chr09	97.143	35	59135746	59135780	4.25e-08
DMG400031521	chr11	97.143	35	31215073	31215039	4.25e-08
DMG400031521	chr11	97.143	35	31259721	31259687	4.25e-08

To identify candidates for the *Rpi-ver1* gene, it was first decided to scan the small, high confidence locus for NLRs. In total, 108 genes were identified, and after filtering for genes with an expression greater than 1 TPM, 72 remained. Of the genes identified, only a single NLR was identified. Unusually, the NLR - which lies directly on top of the `NLR0226` marker that defines the border of this locus - has an upstream CC and partial NB-ARC domain that has been broken by a frameshift mutation in the NB-ARC locus. No evidence was found for an intron at the site of this frameshift in the RNAseq data, however all tested gene annotation methods would either introduce an intron or break the gene into two separate models to resolve this break. Mapping of raw HiFi reads to the gene locus validated that this was not due to a misassembly error. To validate this candidate as *Rpi-ver1*, two copies of the gene were cloned, one with the internal stop codon removed by the addition of an additional two nucleotides to prevent the frameshift. Candidates were externally validated by Jamie Orr - neither copy successfully induced an HR response when transiently expressed with *Phytohptora infestans*.

To identify additional candidates of *rpi-ver1*, a search for non-canonical NLRs was

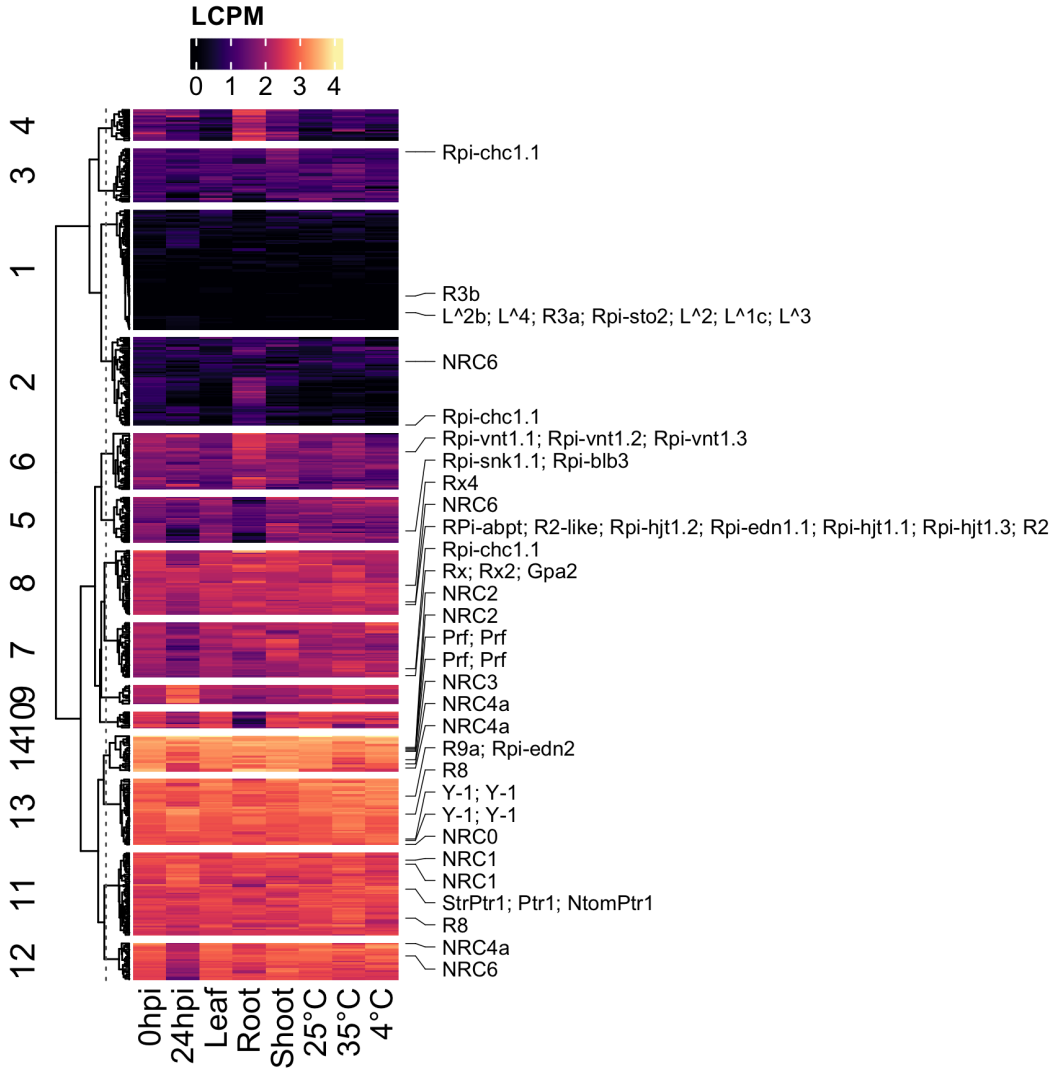


Figure 21: **NLRs clustered by expression.** NLRs separated into 14 clusters based on LCPM values. Identifiable homologs are highlighted.

conducted with Resistify `--ultra`. A second sequence was identified with an Rx-CC domain and a "NCLL" motif structure. This sequence was determined to be an Rx-CC Jacalin-like lectin domain protein, which have been previously implicated in fungal pathogen resistance in plants (Esch and Schaffrath 2017). A BLAST search against the NCBI nt database revealed a sequence with 99% identity (`KAH0640457.1`) but with an additional 340 amino acid upstream domain that contained a second Jacalin domain in the potato cultivar 'Otava'. A manual search upstream of the candidate did not indicate the presence of this domain, and so it was considered as an additional candidate. Transient expression assays by Jamie Orr revealed a strong hypersensitive response in tissue even in unchallenged conditions.

Aside from NLR and NLR-like genes, some genes are present that might be implicated in *P. infestans* disease resistance. There are two receptor-like kinases `g27418` and `g27419`, although `g27418` has a 100% identical, 99% coverage match to *S. tuberosum* gene `XP_006354477.1`, and `g27419` has a 98.89% identical, 99% coverage match with 'Otava' gene `KAH0640440.1`.

Table 8: **Annotated genes identified in the *Rpi-ver1* locus.** Annotations are derived from mapping against the eggNOG database. NLR motifs refers to motifs identified by `Resistify --ultra`.

Gene	Annotation	LCPM	NLR motifs
g27413	lactate/malate dehydrogenase, alpha/beta C-terminal domain	8.97003	None
g27461	RNA recognition motif	8.21552	None
g27475	Prp19/Pso4-like	7.98544	None
g27415	Zinc finger, C3HC4 type (RING finger)	7.84136	None
g27417	Methyltransferase	7.80978	C
g27423	-	7.79277	None
g27418	Serine threonine-protein kinase	7.77435	None
g27439	Catalyzes the reaction of cyanate with bicarbonate to produce ammonia and carbon dioxide	7.75106	None
g27467	Initiation factor	7.63276	None
g27438	RNA recognition motif	7.42566	None
g27409	FAD binding domain of DNA photolyase	7.41532	None
g27421	belongs to the protein kinase superfamily	7.3802	None
g27407	Inorganic phosphate transporter	7.32103	CC
g27470	60S ribosomal protein	6.96758	None
g27429	PAP_fibrillin	6.95747	None
g27460	Flavin containing amine oxidoreductase	6.91384	N
g27473	Domain of unknown function (DUF3527)	6.89458	None
g27419	Protein kinase domain	6.88528	None

Gene	Annotation	LCPM	NLR motifs
g27403	Cupin	6.4419	None
g27431	Long-chain fatty alcohol oxidase involved in the omega- oxidation pathway of lipid degradation	6.32319	None
g27389	Electron transfer flavoprotein	6.31753	None
g27437	Ribosomal L28 family	6.31	None
g27466	proton gradient regulation 5	6.26113	None
g27406	serine threonine-protein phosphatase	6.18445	None
g27405	ASCH	6.12274	None
g27388	helix loop helix domain	6.05149	None
g27410	Belongs to the actin-binding proteins ADF family	6.04789	None
g27393	DNA-binding domain in plant proteins such as APETALA2 and EREBPs	5.88064	None
g27443	Nucleolar GTP-binding protein	5.86158	None
g27444	POT family	5.82176	None
g27433	bromo domain	5.744	NT
g27457	Flavin containing amine oxidoreductase	5.70502	None
g27412	Myb-like DNA-binding domain	5.65496	T
g27436	Translation initiation factor	5.59655	None
g27401	Cupin	5.37784	None
g27478	Belongs to the disease resistance NB-LRR family	5.33557	NNNNNNLL...
g27463	-	5.22085	None
g27432	EXS family	5.20558	None
g27411	Histone acetyltransferase subunit NuA4	5.20072	None
g27430	Transducin WD40 repeat-like superfamily protein	5.16312	None
g27402	SET (Su(var)3-9, Enhancer-of-zeste, Trithorax) domain	5.05784	None
g27469	Acid phosphatase	4.78262	None
g27476	domain in TBC and LysM domain containing proteins	4.72666	None
g27386	Belongs to the Casparian strip membrane proteins (CASP) family	4.68651	None
g27465	Family of unknown function (DUF716)	4.58402	C
g27420	Pentatricopeptide repeat-containing protein	4.48252	None
g27434	-	4.43007	None
g27426	MazG-like family	4.42901	None
g27440	PPR repeat	4.36934	None
g27425	RNA recognition motif	4.35413	T
g27416	Protein of unknown function (DUF640)	4.0657	None

Gene	Annotation	LCPM	NLR motifs
g27458	Flavin containing amine oxidoreductase	3.90286	None
g27428	Belongs to the TRAFAC class myosin-kinesin ATPase superfamily. Kinesin family	3.6543	None
g27474	CCT motif	3.42983	None
g27477	Cysteine-rich repeat secretory protein 3-like	3.32828	None
g27462	-	3.16011	None
g27464	Functions as actin-binding component of the Arp2 3 complex which is involved in regulation of actin polymerization and together with an activating nucleation- promoting factor (NPF) mediates the formation of branched actin networks	3.091	None
g27390	finger protein	3.03577	None
g27408	phosphate transporter	1.88097	None
g27385	non-haem dioxygenase in morphine synthesis N-terminal	1.82132	None
g27398	Cupin domain	1.78344	None
g27452	Copper amine oxidase, enzyme domain	1.71443	None
g27455	Flavin containing amine oxidoreductase	1.6868	None
g27391	ethylene-responsive transcription factor	1.64642	None
g27394	BEST Arabidopsis thaliana protein match is Uncharacterised conserved protein UCP015417	1.63235	None
g27472	Ubiquitin-binding WIYLD domain	1.46496	None
g27383	Belongs to the iron ascorbate-dependent oxidoreductase family	1.36236	None
g27400	Germin-like protein subfamily 1 member 17	1.34638	None
g27392	Ethylene-responsive transcription factor	1.28423	None
g27471	Belongs to the disease resistance NB-LRR family	1.19566	C
g27387	Auxin-induced protein 5NG4	1.19075	None
g27384	Belongs to the iron ascorbate-dependent oxidoreductase family	1.16013	None
g27441	ADP binding	1.05941	NCLL

To identify candidates beyond the high confidence locus, the search for NLRs was expanded to the markers `DMG00011401` and `NLR0226`, encompassing a much wider region of 6.9 Mbp. Within this locus, 11 NLRs were identified of which four were canonical TNLs and three were canonical CNLs. All canonical NLRs were highly expressed apart from the CNL `solanum_verrucosum_chr09_001508` which had no expression, a teM methylation profile, and overlapped with *Helitron* annotations in both the `EarlGrey` and `EDTA` datasets. Given that this wider locus is defined by markers that were not fully associated with resistance in the original screen, it is difficult to establish whether these could be *Rpi-ver1* without further mapping experiments.

To summarise, re-mapping of existing data to the *S. verrucosum* genome revealed and reduced the known size of the *Rpi-ver1* locus. The absence of canonical NLRs suggests that resistance is imparted by a non-canonical or non-NLR gene. Future validation of *Rpi-ver1* resistance would benefit from the development of additional markers using this new locus to determine the causative gene.

Mixed state centromeres

To identify putative centromeric regions, previously generated CENH3 ChIP reads were realigned to the *S. verrucosum* genome. In total, 98% of the reads aligned to the chromosomal assembly, indicating that the vast majority of centromere sequence had been successfully assembled. This was an improvement on the previous *S. verrucosum* assembly, of which 17.2% of the reads aligned to unanchored contigs, indicating a failure to fully resolve the centromeres (Hosaka, Sanetomo, and Hosaka 2022).

The centromeres varied in size and composition (fig. 22). Centromere one exhibited highly variable CENH3 ChIP read mapping, reflecting the underlying sequence which was partially repetitive. The repetitive region was locally depleted in CG but enriched for CHG methylation. The majority of the centromere was composed of LTR-derived sequences, including the repetitive region. Chromosomes two and three exhibited a similar distribution of CENH3 ChIP reads, but did not contain evidence of a repetitive region. Their sequence was similarly composed of LTR-derived sequence. Chromosome four exhibited a well-defined repetitive region which covered almost all mapped reads. The repetitive region was composed of identifiable LTRs, although invasion of LTRs was apparent.

Similar variation in architecture was apparent in the remaining centromeres - chromosomes five, six, nine, and twelve had no repetitive regions and five showed large regions depleted in read mapping; chromosomes seven, ten, and eleven were highly repetitive but with evidence of LTR invasion; chromosome eight had a repetitive region that did not correspond with the largest peak of read mapping.

To characterise the repetitive landscape of the *S. verrucosum* centromeres, a search for tandem repeats with a maximum size of 9kbp was conducted across the genome, based on the observation of *S. tuberosum* containing repeat monomers on the order of kilobases. Repeat monomer size varied across the genome with the majority of repeats being <2kbp, although a cluster of repeats >6kbp was also noted. Repeats that lay inside the centromeric regions exhibited a different distribution - although a peak

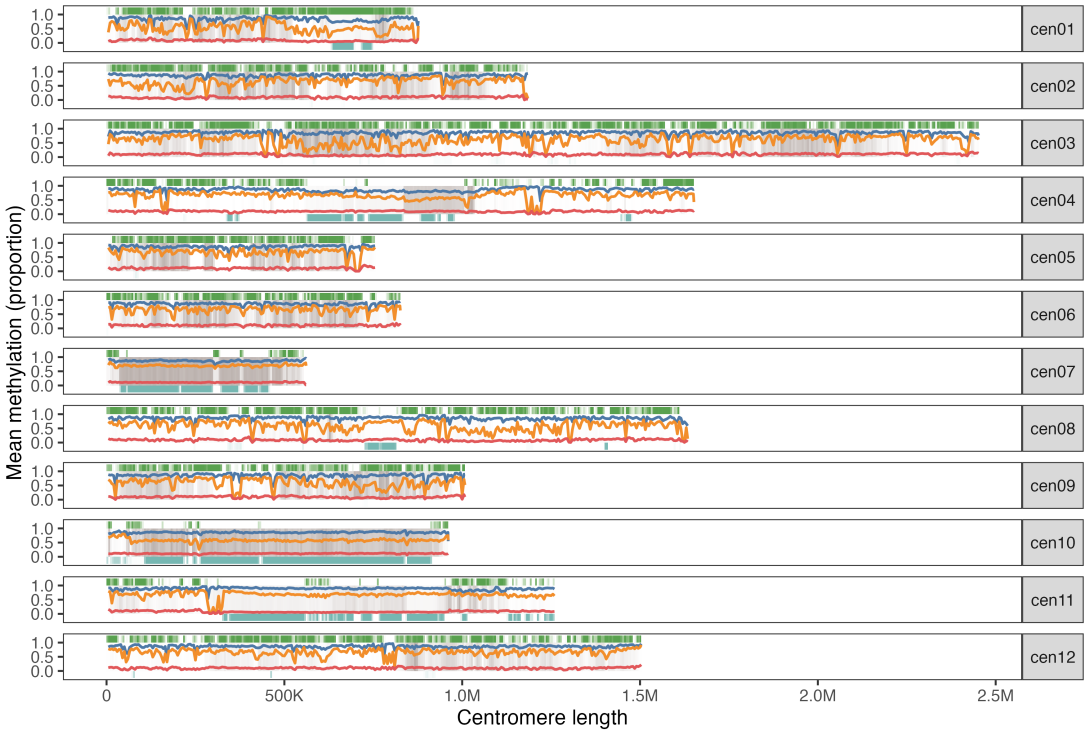


Figure 22: **The centromeres of *S. verrucosum*.** Features spanning the centromeric regions of each chromosome. Grey background corresponds to the per-centromere min-max normalisation of mapped CENH3 reads. Upper green heatmap corresponds to density of EarlGrey Ty3 LTR annotations. Lower aqua heatmap corresponds to density of Tandem Repeat Annotation and Structural Hierarchy (TRASH) annotations. Inner lines correspond to proportion of CG (blue), CHG (orange), and CHH (red) methylation respectively.

of monomers <2kbp was identified, a large complement of repeats lay in the 2-4kbp range (fig. 23 a).

To identify repeats enriched for CENH3, and therefore represent centromere repeat subunits, the CENH3 ChIP read depth of centromeric repeats was examined (fig. 23 b). As anticipated, repeats that lay within the bounds of the centromeres were enriched for CENH3. Clusters of monomers with similar repeat size and read depth were evident. To further characterise these clusters, their distribution amongst the chromosomes was assessed (fig. 23 c). In line with the observation that *Solanum* centromeres exhibit - when present - unique repeats, each cluster was linked to a distinct centromere. Clusters enriched for CENH3 reads were evident for chromosomes four, seven, ten, and eleven. No evidence of clusters being shared by chromosomes was evident. Whilst satellite repeats in "normal" centromeres are generally stable in terms of their size, clusters exhibited a degree of freedom in terms of their monomer size.

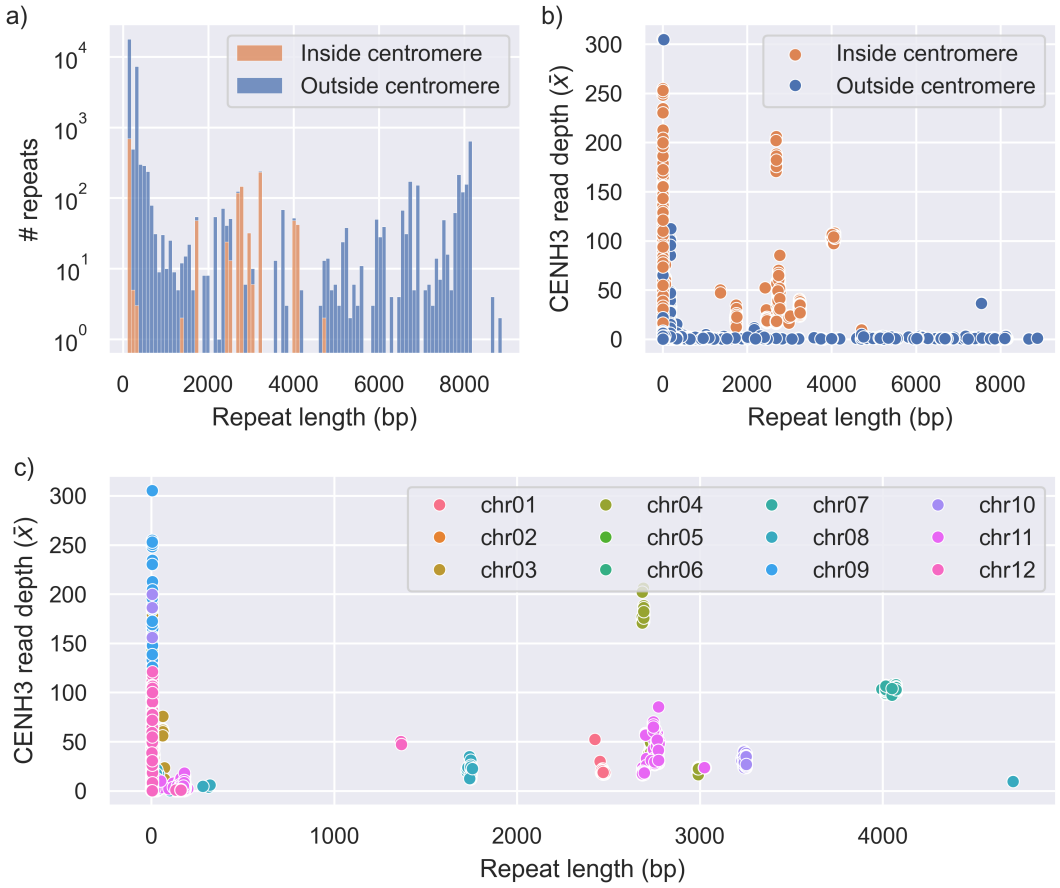


Figure 23: **Centromere-associated repeats.** a) The copy number and distribution of monomer sizes of repeats identified by TRASH. Repeats present inside of the *S. verrucosum* centromeres are highlighted. b) The mean depth of CENH3 ChIP read mapping against the monomer size of TRASH repeats. Repeats smaller than 100bp or with a mean read depth of < 10 have been pre-filtered. c) The distribution of monomers across chromosomes.

As a large proportion of the centromeres were comprised of dense LTR annotations, and evidence was seen for recent invasions, the transposable element library was searched for transposable element families which exhibited a bias towards the centromeric region. Odds ratios for centromere:non-centromere bias were calculated, and significantly biased families were identified. The distribution of odds ratios of significantly varying families indicated a small group of transposable elements with a bias greater than 10, which was hence selected as an arbitrary cutoff (fig. 24).

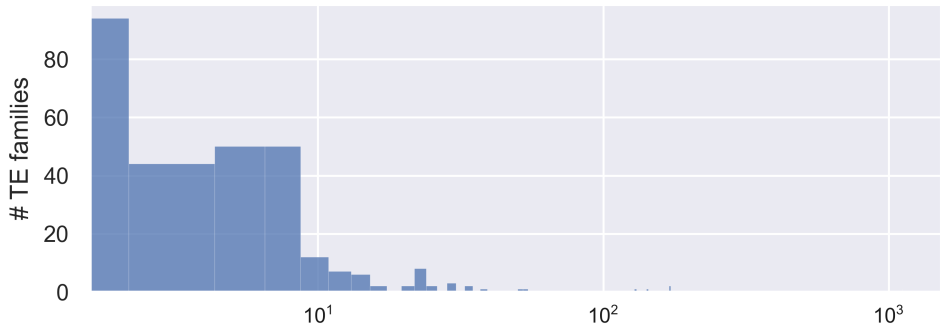


Figure 24: **Distribution of centromere biased TEs.** The distribution of centromere:non-centromere odds ratios for significantly varying ($p<0.05$) transposable element families.

A total of 46 families were identified above this threshold. Of these, 25 were classified by `EarlGrey` as Ty3 LTRs, one as a Caulimovirus, and the remaining as satellite repeats or unknown. To examine these families closer, each family was further classified by `TEsorter` (tbl. 9). Encouragingly, five of the families classified by `EarlGrey` as Satellites were classified as belonging to the CRM clade of LTRs including one family with an extreme bias towards the centromeres. The remaining identifiable clades associated with the centromeres were Tekay and Athila elements, of which Tekay elements represented the majority. In total, 13 CRM, 17 Tekay, and 3 Athila families were identified to be invading the centromeres. The estimated completeness of elements varied, likely due to potential degradation of these elements following centromere invasion.

Table 9: **Centromere-biased transposable elements.** Transposable element families identified by `EarlGrey` and their classification by `TEsorter`. Odds ratio and p-value from fisher exact test. Order, Clade, and Complete columns are via `TEsorter`.

EarlGrey ID	Odds ratio	p-value	Order	Clade	Complete
rnd-4_family-935#Satellite	1027.353622	2.5295e-24	LTR	CRM	no
rnd-4_family-936#Satellite	157.873412	0.000466	LTR	CRM	no
rnd-1_family-236#LTR/Ty3	50.125391	7.1504e-52	LTR	CRM	yes

EarlGrey ID	Odds ratio	p-value	Order	Clade	Complete
rnd-5_family-12753#Satellite	46.471079	6.5051e-13	LTR	CRM	no
rnd-4_family-648#Satellite	34.763622	5.2730e-13	LTR	CRM	no
rnd-1_family-528#LTR/Ty3	31.890542	4.9883e-29	LTR	Tekay	no
rnd-5_family-24514#Satellite	31.574516	0.003153	LTR	CRM	no
rnd-1_family-513#LTR/Ty3	30.396724	5.1540e-131	LTR	CRM	yes
rnd-1_family-329#Unknown	26.806389	1.1364e-20	LTR	CRM	no
rnd-1_family-627#LTR/Ty3	26.729238	1.9598e-25	LTR	Tekay	yes
rnd-1_family-529#LTR/Ty3	26.322606	4.2495e-7	LTR	CRM	no
rnd-1_family-84#LTR/Ty3	22.557655	0.000065	LTR	Athila	no
rnd-1_family-714#LTR/Ty3	21.575135	1.2532e-22	LTR	Tekay	yes
rnd-5_family-896#LTR/Ty3	21.292146	6.8659e-20	LTR	CRM	no
rnd-4_family-647#LTR/Ty3	21.15486	1.3842e-21	LTR	CRM	yes
rnd-1_family-724#LTR/Ty3	21.053783	0.000082	LTR	Tekay	yes
rnd-1_family-111#Satellite	20.799432	1.6863e-14	LTR	Tekay	no
rnd-1_family-295#LTR/Ty3	20.733289	2.3713e-15	LTR	Tekay	yes
rnd-5_family-3233#LTR/Ty3	20.47124	2.5421e-27	LTR	Tekay	yes
rnd-5_family-5977#LTR/Ty3	19.796611	2.4688e-30	LTR	CRM	no
rnd-1_family-473#LTR/Ty3	19.733994	0.00659	LTR	Tekay	no
rnd-1_family-500#LTR/Ty3	19.733994	0.00659	LTR	Tekay	no
rnd-4_family-2245#LTR/Ty3	15.735114	5.9992e-26	LTR	Tekay	yes
rnd-5_family-501#LTR/Ty3	14.109814	3.2580e-61	LTR	Tekay	no
rnd-5_family-317#LTR/Ty3	14.023845	1.0247e-21	LTR	Tekay	yes

EarlGrey ID	Odds ratio	p-value	Order	Clade	Complete
rnd-4_family-2247#LTR/Ty3	13.352526	1.6539e-20	LTR	Tekay	yes
rnd-5_family-9169#LTR/Caulimovirus	13.171131	2.7883e-11	pararetrovirus	unknown	unknown
rnd-4_family-3092#LTR/Ty3	12.31341	1.4245e-9	LTR	Tekay	yes
rnd-1_family-265#LTR/Ty3	11.841442	0.002876	LTR	unknown	none
rnd-1_family-62#Satellite	11.279689	0.00014	LTR	Tekay	no
rnd-5_family-8459#LTR/Ty3	11.279689	0.00014	LTR	Athila	yes
rnd-1_family-439#LTR/Ty3	10.577162	1.7224e-14	LTR	CRM	no
rnd-1_family-282#LTR/Ty3	10.481859	1.7552e-9	LTR	Athila	no
rnd-1_family-61#LTR/Ty3	10.065481	6.8202e-19	LTR	Tekay	no

As multiple Tekay families exhibited a strong bias towards the centromeres, their phylogenetic relationship was examined (fig. 25). The Tekay family which exhibited the greatest bias towards the centromere, rnd-1_family-52, was closely related to a clade of Tekay elements which all exhibited a strong bias to the centromeres. This was at odds with their neighbours which, whilst exhibited some degree of bias towards the centromere, this bias was much reduced. This possibly demonstrates that this subfamily of Tekay elements has specialised in invading the centromeres or is being selected for in *S. verrucosum*.

Whilst, a centromere-bias likely indicates a role in maintaining or disrupting the centromeres, it may not be directly linked to CENH3 enrichment. To compare how the centromere-bias related to CENH3 read mapping, transposable element families were explored for CENH3 read mapping. As with the calculated centromere-bias, CENH3 read mapping showed a clear subset of LTRs that were enriched (fig. 26). All LTRs enriched for CENH3 ChIP reads were of the Ty3 clade. When centromere-bias was compared against CENH3 ChIP enrichment, a positive association was noted (fig. 26). LTRs with the greatest bias and CENH3 ChIP enrichment all belonged to the CRM clade, followed by Tekay elements. The association of centromere-bias and CENH3 read mapping suggests that Tekay and CRM elements are both accepted by *S. verrucosum* to be present in the centromere - they are not being treated as invaders. No elements were identified that significantly deviated from this trend.

As stated, several of the centromeres showed evidence of large satellite repeats. Whilst EarlGrey could in some circumstances identify mosaics of repeats overlapping with these regions, many of these repetitive sections of the centromeres remained unannotated. To fully annotate the tandem repeats present in *S. verrucosum* centromeres,

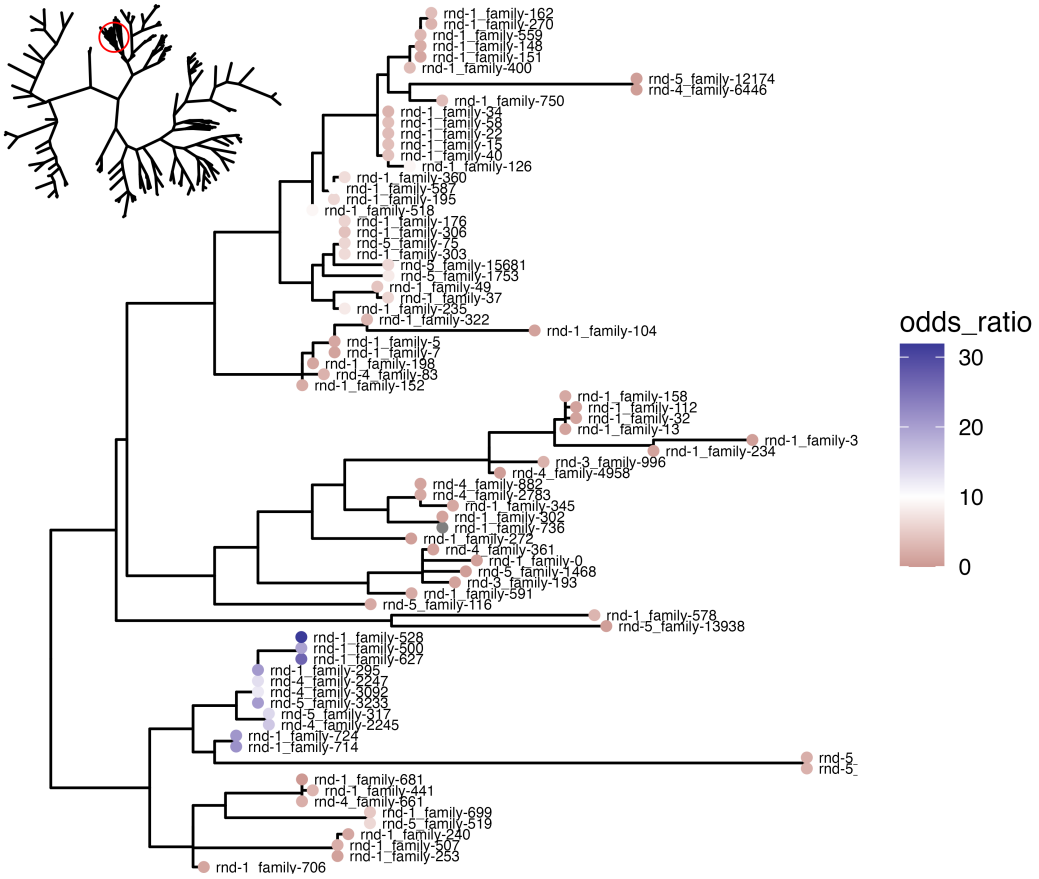


Figure 25: **Phylogenetic relationship of centromere-biased LTRs.** Phylogenetic tree of LTRs based on `TEsorter` extracted domains. Subtree is centred on the highest scoring Tekay family rnd-1_family-52 with nine steps taken back in the tree. A midpoint odds ratio of 10 has been selected. Position of the subtree in the wider phylogeny is circled in the upper left tree.

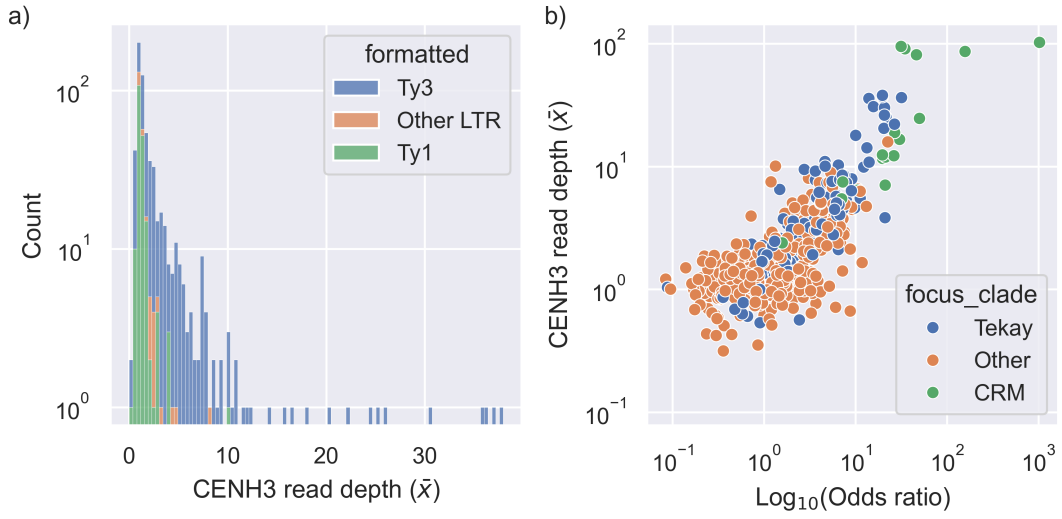


Figure 26: **LTRs are CENH3 ChIP enriched**. a) Histogram of the number of LTRs by their mean CENH3 ChIP read depth. b) Mean CENH3 ChiP read depth by centomere-bias odds ratio. Tekay and CRM elements have been highlighted.

Tandem Repeat Annotation and Structural Hierarchy (TRASH) was selected as an alternative. Although TRASH has been designed to identify typical satellite repeats on the order of 100s of basepairs, it successfully identified repeats on the order of kilobasepairs in the *S verrucosum* genome.

Repeats that were associated with the centromeres ranged from 2000 to 6000bp (fig. 27 a). This is in line with a previous observation that potato has kilobasepair centromeric repeats and is the first absolute measurement of centromere repeat size in *Solanum*. Centromeric repeats were also enriched for ChIP enrichment, supporting their role as centromeric subunits (fig. 27 b). Interestingly, centromeric repeats overwhelmingly formed chromosome-specific clusters in the repetitive centromeres of chromosomes four, seven, ten, and eleven (fig. 27). The size of repeat in each cluster was distinct between chromosomes and showed a degree of variation within clusters. This suggested that repetitive centromeres in *S. verrucosum* are formed from distinct repeat units of different sequence origin.

Given previous contradicting reports of centromere-invading *ATHILA* LTRs in *Arabidopsis* exhibiting elevated CHG methylation, but centromere-invading *CRM* LTRs in *C. annuum* being hypomethylated, the levels of CHG methylation here were examined. In general, the families identified to be centromere-biased in the *S. verrucosum* genome were found to be slightly yet significantly demethylated when comparing insertions inside and outside of the centromere (ß = -0.0290 CHG (prop.), SE = 0.006, p = 0.000...).

Given the density of transposable elements within the *S. verrucosum* centromeres, I hypothesised that these repeats were originally seeded by LTR retrotransposons, likely of the CRM or Tekay clades. To verify this, repeats were aligned to the TE library to identify their LTRs of origin. Accordingly, the LTR origins of repetitive centromeres could be resolved: chromosome one and seven centromeres are of CRM

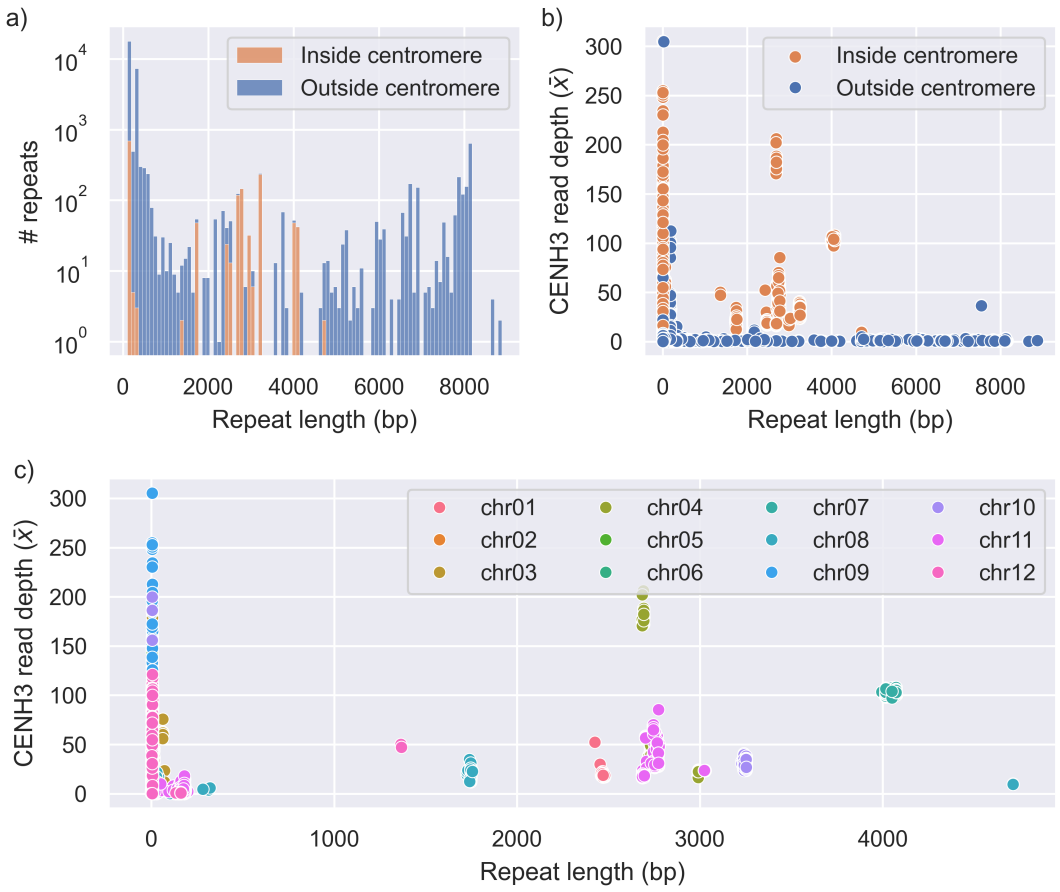


Figure 27: **Tandem repeats in *S. verrucosum* centromeres.** a) Distribution of repeats identified by TRASH. Repeats identified inside of centromeres are highlighted. b) Distribution of repeats and mean CENH3 ChIP read depth. Centromeric repeats are highlighted. c) The same distribution but with chromosomes highlighted.

origin, chromosome 10 is of Tekay origins, chromosome four is a hybrid of CRM and Tekay repeats with the Tekay elements being *more* enriched for CENH3, and chromosome 11 has repeats that are not of any clear transposable element origin. There was evidence of higher order repeats within the centromeres, indicating that they are not formed via unequal crossover. As previously suggested, they shared little homology with the centromeres of *S. tuberosum*, taking chromosome seven as an example (fig. 28).

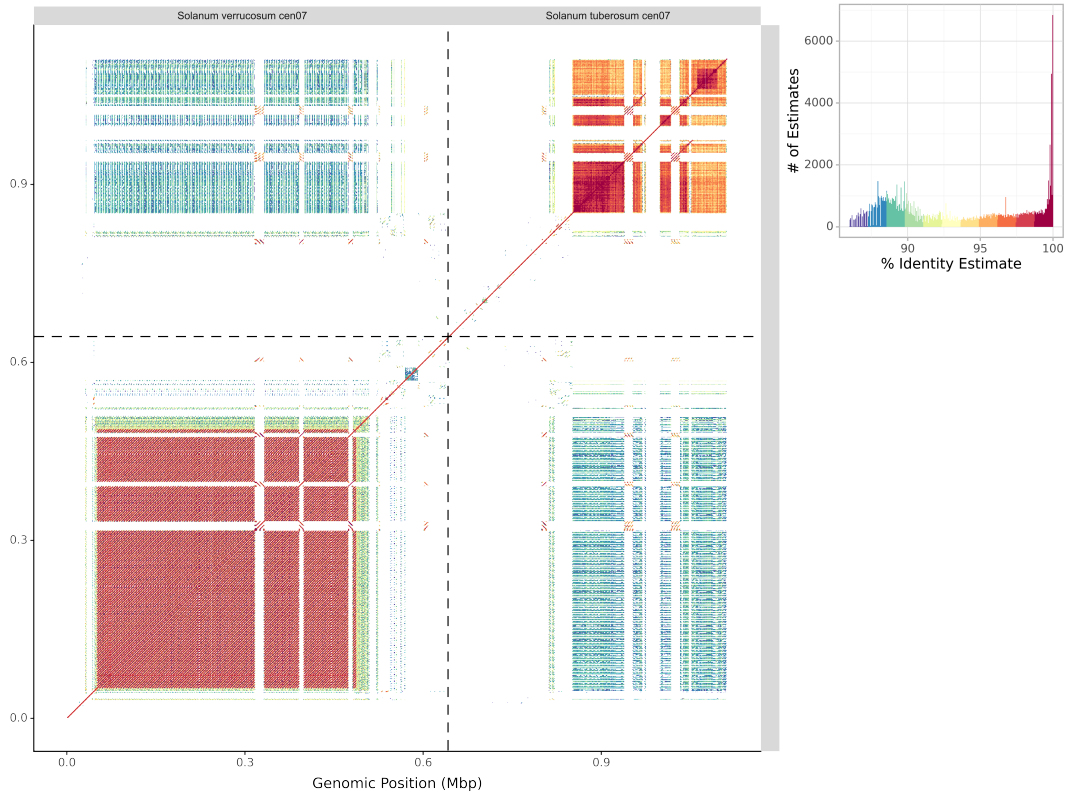


Figure 28: **The structure of centromere 7.** A graphical representation of the structural similarity within and between centromere 7 of *S. verrucosum* and *S. tuberosum*. The large blocks of homologous sequence do not indicate a higher order repeat structure. Recent LTR insertions are evident in both centromeres by gaps in the homology blocks.

One surprising observation is that CRM and Tekay derived repeat clusters are not unique to the centromeres, and in fact appear to be a common feature of the *S. verrucosum* genome. Indeed, each chromosome had at least one Tekay-derived repeat cluster which was not centromeric. This is surprising - given that Tekay and CRM elements both seem to be acting as centromere repeat arrays, the function of these non-centromeric arrays is not clear.

Large gaps of CENH3 read coverage were noted in both the repetitive and repeatless centromeres. Given the interspecies diversity of centromeres in *Arabidopsis* and even between the tissue of distal branches of trees, it can be expected that the *S. verrucosum*

individual sequenced in this project has diverged from the individual that CENH3 reads were generated for. Repeating CENH3-ChIP (and ideally simultaneously resequencing the centromeres) would be necessary to determine whether these are true regions depleted in CENH3, or are the result of recent insertions that have been established as centromeres.

To summarise, the centromeres of *S. verrucosum* are in a mixed state of being composed of tandem repeat arrays with kilobase repeat subunits and no evidence of higher order repeat topology, and being composed of diverse LTR sequences with no tandem repeat arrays present. Some centromeres such as 7 are formed almost exclusively by tandem repeats, others such as centromere 4 have a partial tandem repeat array, but also contain a significant fraction of LTRs enriched for CENH3 reads. Other centromeres such as 6 are repeatless and formed exclusively of LTRs. Some tandem array subunits have identifiable LTR domains, others do not show evidence of being derived from LTRs. LTRs present within centromeres are of the Tekay and CRM families, and distinct subfamilies show a strong bias towards being present in the centromere. DNA methylation varies across centromeres but no dip region of association with tandem arrays is evident.

S. verrucosum has S-RNase alleles

The self-compatibility of *S. verrucosum* has been attributed to a lack of a functional *S-RNase* in the genome, either through mutation, a lack of expression, or is completely absent. To identify *S-RNase* homologs in the genome, a BLASTp search was conducted using the complete sequence of an *S. tuberosum S-RNase* (`Q01796.1`). A single high-scoring hit (Identity: 53.5% E-value: $1.84e^{-77}$) was identified on chromosome 1, the expected location of the S-locus. As expected for *S-RNase* which is expressed exclusively in the stylet, no RNA reads from the available RNAseq conditions, and so the gene model was produced entirely *ab initio* - the annotations produced by `BRAKER3` and `Helixer` were identical.

From the Solanaceae pangenome produced in Chapter 1, the *S. verrucosum S-RNase* belongs to an orthogroup that contains all other identifiable *S-RNase* homologs from the other genomes (fig. 29). The *S. verrucosum* homolog is closely related to homologs from *S. neorickii* and *S. chmielewskii*. It is interesting to note that *S. neorickii* and *S. peruvianum* contain multiple *S-RNases*, but that it was seemingly completely absent from the tuber-bearing cultivars *S. candolleanum*, *S. tuberosum* Group Stenotomum (PG6359), *S. tuberosum* Group Phureja (E4-63 and E86-69). Their absence was not due to a failure of Helixer to annotate the gene models - a tBLASTn search for `Q01796.1` did not identify any significant hits on chromosome 1 for any of the genomes. PG6359 is naturally self-compatible, a trait which has been previously attributed to the high expression of the S_{s12} *S-RNase* allele but low expression of the S_{s11} allele, which leads to S_{s12} but not S_{s11} pollen being rejected (C. Zhang et al. 2019). The sequences of each allele were determined by de novo assembly of stylet RNA sequencing which are unfortunately unavailable, so it is challenging to determine the exact status of *S-RNase* in this line. *S. tuberosum* Group Phureja E86-69 and its inbred descendent E4-63 are self-compatible through the introgression of *Sli* from a *S. chacoense* breeding line (C.

Zhang et al. 2021). Why *S-RNase* is absent from these is also unclear.

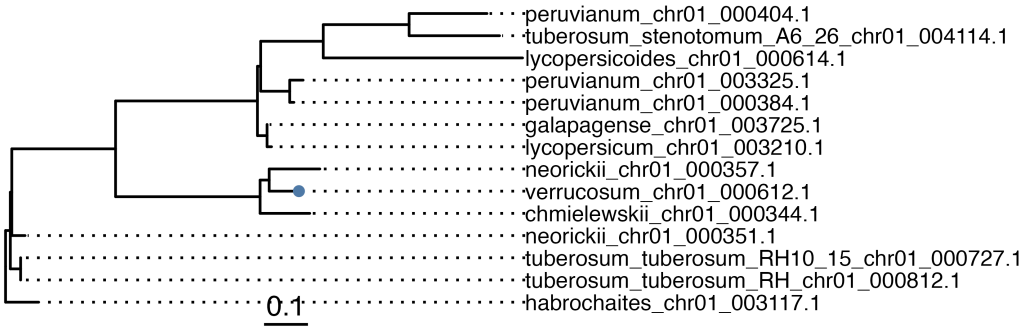


Figure 29: **S-RNase homologs in the Solanaceae pangenome.** Members of the single orthogroup determined to contain *S-RNase* in the pangenome produced in Chapter 1.

When searched against RefSeq for closest homologs, the *S-RNase* was found to be identical with a previously identified *S. verrucosum S-RNase* and was clustered with other Solanaceae *S-RNases*. Surprisingly, a search against the nr database revealed a 97.8% shared identity with *S-RNase* 7 from *S. tuberosum* (`QYF06681.1`). It is not exactly clear what the source of this *S-RNase* is - it is cited as being cloned from "diploid potatoes" and is potentially *S. stenotomum*, *S. goniocalys*, or *S. tuberosum* Group phureja based on the material with an *S-RNase* 7 haplotype in the study.

The sequence of *S. verrucosum S-RNase* was examined for any mutations that could result in non-functionality. The sequence is complete - there is no evidence of any truncations or insertions that might lead to a loss of function. Previously identified mutations that have led to reduced or a loss of *S-RNase* function including lost N-glycosylation site and histidine active sites are also not present (Broz et al. 2021).

The *S-RNase* gene lies in a pericentromeric region dense exceptionally dense in LTRs. In the promoter region 4kbp upstream of the *S-RNase* is a complex of repeats that are likely of transposable element origin (fig. 30). `EarlGrey` and `EDTA` produced conflicting annotations - `EDTA` identified two TIR fragments of the Mariner and CACTA family in a repetitive region that `EarlGrey` did not fully classify, but which instead identified two large LINE/L1 fragments and an intermediate MULE-MuDR element. Of the annotation, only the consensus sequence of the LINE/L1 family had a structural element that supported its classification - a LINE family RNase H domain identified by `TEsorter`. It is not possible to determine whether this, or another transposable element, could be an insertion that has inhibited *S-RNase* expression. The promoter regions of the other Solanaceae *S-RNase* examined here also showed an abundance of transposable elements of various origin.

To summarise, *S. verrucosum* has an intact *S-RNase* protein that is closely related to tomato *S-RNases*. The promoter region likely contains a genuine LINE/L1 insertion, and may be the site of other transposable elements. Whether or not the stylet expression of the *S-RNase* is impacted by these is not currently known.

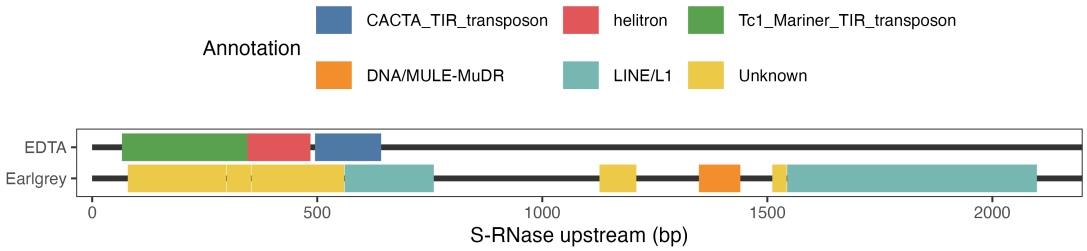


Figure 30: **S-RNase upstream region.** Transposable element annotations from `EDTA` and `EarlGrey` in the 2 kbp promoter region of *S-RNase*.

Discussion

Assembly and annotation performance

Of the various assembly strategies assessed in this study, `hifiasm` was a clear outlier in terms of final contiguity. This is reflected in the recent abundance of genome assembly projects, of which the HiFi + `hifiasm` combination has formed the core of most assemblies. Still, it should be noted that inclusion of nanopore-derived `flye` assemblies did moderately increase the contiguity further, and so a hybrid approach might be suitable where data is available.

Through the combination of HiFi, Nanopore, and Hi-C scaffolding, the most contiguous assembly of *S. verrucosum* was produced to date. The high contiguity of assembly was likely due in part to the unusual centromeres of *S. verrucosum*. Typical satellite repeat centromeres are typically challenging to fully resolve with HiFi reads resulting in breaks in the assembly that require ultra-long nanopore reads to fully resolve.

Even in the relatively short period since this projects inception, new methods and strategies have been developed that would yield more performative results when applied to *S. verrucosum* or genomes of similar structure. In particular, the emergence of ultra-long nanopore sequencing and next generation assemblers such as the hybrid HiFi/ONT `verkko` has allowed the routine construction of telomere-to-telomere plant genomes (Rautiainen et al. 2023). The read quality achieved by duplex basecalling of the recent R10.4.1 nanopore chemistry permits genome assembly exclusively from nanopore reads, further increasing the accessibility of assembling plant genomes (Koren et al. 2024). Higher accuracy basecalling models will improve the accessibility and quality of assemblies further, as demonstrated in the recent release of the HERRO model that can achieve a 100 fold improvement in read quality of nanopore reads (Stanojević et al. 2024).

A shift towards nanopore-exclusive assemblies could in theory also enable improved genome annotation of genes, transposable elements, or other features marked by DNA modifications. It has recently been demonstrated that methylation data can be used to aide in haplotype phasing of human genomes (Fu et al. 2023). As demonstrated in this study, transposable elements exhibit distinctive DNA methylation signatures and genes exhibit differential exon-methylation which can be used as an indicator of silent or transposable element derived genes. DNA methylation could be utilised by

annotation software as an additional layer of evidence to identify for example introns or transposable element boundaries.

S. verrucosum has dynamic centromeres

Previously, it has been demonstrated that the centromeres of potato exist in an unusual mixed state of repetitive and repeatless sequence structure, and this is also true of *S. verrucosum* (Gong et al. 2012; H. Zhang et al. 2014). Here, the full sequence of the *S. verrucosum* centromeres is recovered, revealing a complex of repeatless and repetitive sequences that appears to be actively shaped by transposable elements.

The repeatless centromeres of *S. verrucosum* are similar in structure to two recent assemblies of the *C. annuum* and *C. rhomboideum* genomes (Jian Chen et al. 2023). In both cases, the centromeres are formed of a rich landscape of Ty3 LTRs with members of the Athila, Tekay, and CRM clades being present. In *Capsicum*, CRM LTRs are particularly dominant in the repeatless centromeres. One explanation for this enrichment is that CRM clades have specialised chromodomain and CR motifs which enable centromere targeting through an unknown mechanism (Neumann et al. 2011). The absence of these structural elements in families such as the centromere-biased Tekay LTRs observed in *S. verrucosum* indicates that other mechanisms exist which drive centromere insertion or retention. Recently, a mechanism that drives centromere-bias of LTRs has been elucidated for the *ATHILA* family in *Arabidopsis*, whereby centromeres become "addicted" to *ATHILA* insertions due to them silencing transcription whilst simultaneously providing small RNAs that restore normal centromere function (Shimada et al. 2023).

The repetitive and partially repetitive centromeres of *S. verrucosum* are also very unusual. The combination of Tekay-derived, CRM-derived, and hybrid centromeres makes it challenging to suggest a single mechanism that results in the formation of these centromeres.

A potential conclusion from the *S. verrucosum* genome is that repeat arrays are the "end-point" of centromere development, and that the repeatless centromeres merely exist as a transitory step in centromere formation. However, the persistence of similar repeatless centromeres in *C. annuum* since their divergence indicates that these are stably functional and that there is no drive towards repeat formation.

The epigenetic status of *S. verrucosum* centromeres remains to be fully resolved here. It is clear from this analysis that differential DNA methylation is not a primary feature of the repeatless or repetitive centromeres as centromere dip regions or any other distinctive DNA methylation pattern are not present. In the *C. annuum* repeatless centromeres, the levels of DNA methylation are also not distinctive, although invading CRM elements do have a slightly reduced level of CHG methylation which is also observed here. Contrarily, in *Arabidopsis* the centromeres are depleted in the CHG context, attributed to the co-observed depletion of H3K9me2 modifications which are associated with the maintenance of non-CG methylation maintenance. Yet *ATHILA* LTRs invading the centromeres of *Arabidopsis* were observed to have an *increased* level of CHG methylation.

One surprisingly observation in *S. verrucosum* is that while CRM and Tekay repeat arrays appear to function as centromeres, similar arrays are also present outside of the centromeres. Tandem arrays of LTR elements have been infrequently observed outside of *S. verrucosum.* Arrays of *Cassandra* LTRs have been observed in several plant genomes with varying degrees of copy number (Kalendar et al. 2020). The function of these elements is not clear - they contain a 5S rRNA domain and their repeat structure is reminiscent of the more well-characterised arrays of 5s rRNAs. The method by which these repeat arrays initially form is also not clear - arrays might form from interchromosomal recombination leading to expansion, or template switching during reverse transcription might lead to multi-copy cDNA. Once a tandem array of LTRs has been seeded, arrays might further expand through unequal crossovers akin to 5s RNA array expansion, although this would result in variations in the sequence of the *Cassandra* monomers, which is not observed. Interestingly, species in hot and dry climates appear to have a higher *Cassandra* copy number but lower tandem array size.

In yeast, the Tf1 LTRs retain insertion activity in the absence of integrase and readily form repeat arrays through insertion into related transposable element sequences (F. Li et al. 2022). Strikingly, insertion of Tf1 is mediated by an integrase independent system, whereby transposition occurs through homologous recombination promoted by the DNA repair mediator Rad52, rather than through the typical Rad51 system, biasing it towards insertion into the related LTR Tf2 and driving repeat array formation.

Given the observation of both centromeric and non-centromeric *Tekay* repeat arrays in *S. verrucosum*, there is a possibility that a similar mechanism is shared in the formation of nascent repetitive centromeres derived from LTRs. The potential link to abiotic stress responses is also conceivable, as if such a mechanism is linked to centromere development, then it could be reasoned to influence evolution or even speciation in response to stress. Partnered with this would be a stress-associated burst in transposable element activity, which might act as further feedstock for tandem arrays, or increase the frequency of LTR-dimers which act as seeds for future arrays.

Thus, there are two reasons for investigating the LTR tandem arrays of *S. verrucosum* further. The first is to establish *how* and *why* they function as centromeres for some, but not all, chromosomes in the genome. Examining their intergenerational stability would be valuable in determining the frequency of LTR insertions and expansion/contractions of these arrays. The self-compatibility of *S. verrucosum* and the relative ease to which these centromeres were resolved make it an excellent platform the observe these dynamics further. Further characterisation of the epigenetic profile of *S. verrucosum* centromeres, such as the distribution of histone methylations will also be useful.

The second reason is to characterise the mechanism which enables these LTR arrays to form. It is clear that they are not the result of unequal cross over given their lack of higher order repeat structure. Determining how these loci respond in, for example, a genetic background deficient in DNA methylation could be informative as it has been in other organisms for understanding array formation.

Canonical NLRs are missing in gene annotations

One striking finding in this analysis is the discrepancy between the number of NLRs identified in the gene annotations produced by BRAKER3 and `Helixer`. An assumption might be made that identifying NLRs would be fairly trivial in a high-quality genome assembly, as their high-copy number and conservation should lend itself towards *ab initio* predictions, and their conserved protein domains should be readily identifiable. However, the strategies used to identify NLRs in the recent literature suggests that this is more challenging than anticipated. A recent potato pangenome resolved to reannotate all NLRs in the genome using a combination of `NLR-annotator`, a set of 7007 NLR protein sequences retrieved from a variety of databases, and a combination of SNAP, AUGUSTUS, and MAKER2 prediction pipelines to generate a set of NLR gene models (Tang et al. 2022). Although not explicitly stated in the article, this rather involved strategy for NLR gene annotation - which was entirely independent of the gene annotation strategy used for the remainder of the genome - perhaps indicates that the researchers recognised that a large fraction of NLRs were being missed through conventional gene annotation. In another recent pangenome of *Solanum americanum*, loci putatively containing NLRs were identified with `NLR-annotator`, which were then extracted alongside their flanking loci, from which NLR gene annotations were predicted using the tomato model of AUGUSTUS followed by manual curation (Lin et al. 2023). Again, this strategy for annotating NLRs was at odds with the remainder of the genome, which used a more standard strategy for gene annotation.

This issue does not appear to be exclusive to *Solanum* - as discussed in Chapter 1, the *C annuum* genome which led to the suggestion of NLR retroduplication used a very complex strategy. Briefly, this involved a search in open-reading frames for NB-ARC domains, the building of a species-specific NB-ARC HMM profile, application of this against the genome to identify NB-ARC domains, extraction of the NB-ARC domains, re-searching against the genome with tBLASTn, followed by extraction of a 10kbp loci around positive hits. Genes were annotated in the loci with a gene annotation method (the exact method is not described), followed by a second search using the NB-ARC domain, followed by a bespoke search for TIR, CC, and LRR domains to classify the NLRs.

I propose that Helixer is a well performing and simple alternative for identifying the full complement of NLRs in the genome. A similar conclusion is reached in the original Helixer publication, which demonstrates that plausible genes may be missed even in very high-quality annotations such as the *Arabidopsis* TAIR10 dataset (Holst et al. 2023) The strategy of merging only non-overlapping Helixer NLR annotations into the final annotation means that any NLR annotation that benefited from the higher-quality BRAKER3 annotation are not impacted by replacement with an *ab initio* model. It is surprising that the missing fraction of NLRs identified by Helixer exhibited a substantially lower level of expression. Whilst RNAseq evidence is used in the BRAKER3 annotations, the missing NLRs are not diverged from other NLRs to the extent where they would likely be missed by the protein homolog searches and *ab initio* modelling that are also a part of BRAKER3. This discrepancy might be linked to the observation that repeat-masking can impact resistance gene prediction due to NLR fragments contaminating repeat databases (Bayer, Edwards, and Batley 2018). A

surprising conclusion from this communication is that NLRs putatively embedded in transposable elements represent false positives. I suggest that any sequence that can be reasonably classified as an NLR, should be considered as one when studying NLR evolution or searching for resistance gene candidates.

NLR-Helitron intersections are consistently predicted

In Chapter 2, a previously unidentified association between *Helitrons* and NLRs was established which appeared to be linked to the expansion of NLRs in tuber-bearing *Solanum* genomes. Helitron elements are notoriously difficult to identify and annotations are often impacted by false positives. To reduce the false positive rate, only *Helitron* elements that were identified by `EDTA` as intact were analysed, which in theory have all the sequence elements necessary to undergo *Helitron*-mediated transposition. Here, this threshold was relaxed slightly to include `EDTA` *Helitrons* that were identified by homology to the intact elements, which will doubtlessly increase any associated false positive rate but will indicate clades that might be undergoing *Helitron*-based expansion. The NRC clade, for example, showed exclusive associated with *Helitron* elements.

As an alternative to `EDTA`'s structure-based annotations of *Helitrons*, the annotations produced by `EarlGrey` were also examined given its overall improved performance when compared in this Chapter. With `EarlGrey`, *Helitrons* are classified from repeats through homology searches against the Dfam transposable element database. Despite this different approach, *Helitron*-embedded NLRs were still identified and at a much greater abundance than via `EDTA`. *Helitron*-embedded NLRs were identified exclusively for CNLs, with the exception of three TNL-related sequences that were annotated only as "N" by `Resistify`.

Whilst a similar association was observed with both `EDTA` and `EarlGrey`, this should not be taken as direct evidence of *Helitron*-embedded NLRs. The false positive rate of `EDTA`-derived classifications remains high, and there is the possibility of NLRs contaminating the Dfam database leading to spurious annotations with `EarlGrey`. However, the presence of fully intact *Helitron* motifs encapsulating NLRs does suggest that *Helitron*-mediated transposition might be possible. The successful revival of the autonomous *Helitron Helraiser* is a good example of how *Helitrons* can show activity from their simple terminal motifs, and whether these non-autonomous NLR elements can also function is a topic worth exploring. The identification of autonomous *Helitrons* in *S. verrocusum* also provides candidates for *Helitron* elements which the non-autonomous NLR *Helitrons* may be exploiting to undergo transposition.

It is interesting to note that despite *Helitrons* having the capacity to in theory relocate to anywhere in the genome, the NRC duplications all occurred in local clusters. This restriction may be due in part to the *Helitrons* bias towards survival in gene rich regions and in close proximity to other *Helitrons*, potentially due to an underlying chromatin signature which marks *Helitrons* (L. Yang and Bennetzen 2009). A bias towards insertion near the donor locus termed "local hopping" is observed for some transposable element families, although this effect was not seen for the revived *Helraiser Helitron* (Muñoz-López and García-Pérez 2010; Grabundzija et al. 2016). The

lack of components necessary for autonomous transposition would also impact their rate of expansion, particularly if autonomous *Helitrons* are being actively repressed by the host genome.

Rpi-ver1 remains elusive

Assembly of the *S. verrucosum* genome revealed the complete content of the *Rpi-ver1* locus, which is present on chromosome 9 between 54.6 Mbp and 56.0 Mbp. Surprisingly, no canonical NLRs were identified within this locus, and the two non-canonical NLRs - a Rx-CC Jacalin-related lectin protein and a frameshifted CNL - did not provide resistance when their functionality was assessed independently. It is conceivable that *rpi-ver1* is not an NLR but instead a gene that provides resistance through some other mechanism. Non-NLR genes that provide resistance have been previously identified, for example *Rhg4* which is a serine hydroxymethyltransferase that provides resistance against the soybean cyst nematode (S. Liu et al. 2012). No genes were identified that both contained a domain associated with plant disease resistance and were unique to *S. verrucosum*. Although re-analysis of the *Rpi-ver1* KASP markers with the genome did reduce the loci from 4.3 Mbp to 1.4 Mbp owing to the difference between the DM and *S. verrucosum* genome, it is likely that further mapping will be required to identify the causative gene to avoid costly screening of the 74 genes high-confidence genes identified. By identifying the *S. verrucosum Rpi-ver1* locus, markers can now be designed to provide fine-mapping of *Rpi-ver1*.

Beyond *P. infestans*, *S. verrucosum* has been previously highlighted for other disease resistances including viruses, insects, and nematodes (Carrasco et al. 2000; Cooper and Bamberg 2016; Castelli et al. 2005). Exploring the genetic basis of these will be empowered by the *S. verrucosum* genome.

On the topic of disease resistance, it is interesting to note that *S. verrucosum* does not exhibit the *NRC6* root-exclusive expression observed in *S. lycopersicum* (Lüdke et al. 2023). Given that *NRC6* is an ancient and highly conserved network, it is suggested that root-specific expression is a strategy for nematode-specific resistance. As *NRC6* is known to act as a node for mediating the response of dozens of sensor NLRs, the expression difference between *S. verrucosum* and *S. lycopersicum* is likely to have a significant impact on NLR-mediated resistance as a whole. Additional root RNA sequencing across *Solanum* would be valuable in determining where this root-specific expression emerges.

It should also be noted that in the *S. lycopersicum* study, *S. verrucosum* is listed as having only one copy of *NRC6*. In the original study, homologs were identified from the result of an `NLRtracker` run against the NCBI hosted *S. verrucosum* gene annotation. Indeed, the second *NRC6* homolog identified here, which is of lower expression, was identified by *Helixer* alone. Revisiting the distribution of NRCs across Solanaceae with the inclusion of `Helixer` annotations would be valuable in fully capturing the diversity of NRCs across the genus. Nonetheless, the presence of multiple NRC6 copies in *S. verrucosum* is an additional indicator that it has deviated from *S. lycopersicum*, which has only one.

The role that DNA methylation and gene body methylation in the CG and CHG contexts plays on NLR function is yet to be fully realised. In *Arabidopsis*, NLRs with high intraspecific diversity (termed hvNLRs; highly variable NLRs) are more expressed, less CG methylated, and are more frequently overlapped with transposable element annotations (Sutherland et al. 2024). A recent survey of hvNLRs in Maize did not indicate any definitive pattern of CG and CHG methylation distribution - association with transposable elements was not assessed (Prigozhin et al. 2024).

Given the clear association between gene body methylation and expression, characterising NLR methylation across plant species will be invaluable in understanding how this impacts disease resistance. Methylation levels could also be used as an additional layer of evidence for identifying high-confidence candidate NLRs in the absence of expression data. The main barrier towards this analysis is the relatively slow uptake of nanopore sequencing, and the current restrictions on depositing raw nanopore data on sequence databases. However, given the clear advantage of ultra-long nanopore reads in assembling plant genomes, and the development of more lightweight raw data formats such as POD5, these barriers may soon be overcome.

The molecular basis of self-compatibility

A key aim in resolving the *S. verrucosum* genome is to understand the mechanisms that underpin its exceptional self-compatibility and interspecific breeding. Here, the presence of *S-RNase* was determined, which does not exhibit any structural deformities that would indicate non-functionality. It is apparent that the S-RNase of *S. verrucosum* is densely methylated and has transposable element insertions in its promoter region - how this correlates to expression in the pistil is unknown, and warrants further investigation. Beyond *S-RNase*, the role of other self-compatibility determining factors such as *SLF* continues to emerge. The *S. verrucosum* genome is a useful resource for this purpose (W. L. Behling and Douches 2023).

PCN resistance gene discovery

Introduction

Nematodes are the most abundant animal life form on earth. In the soil, they function at all major trophic levels and are crucial to nutrient turnover, carbon sequestration, and can even protect plants from root-feeding pests (Hoogen et al. 2019; J. G. Ali et al. 2012). However, nematodes can also be remarkable plant parasites. Plant parasitism has evolved independently in four clades of the Phylum Nematoda (fig. 31). These include migratory ectoparasites such as members of the *Trichodorus* and *Xiphinema* Genera, in Clades 1 and 2 respectively, which remain in the soil to graze from the root surface, often acting as vectors for plant viruses in the process (Raski et al. 1983). A small number of Bursaphelenchus species from Clade 10 are insect vectored pathogens of trees.

However, the largest number of plant-parasitic nematodes are found in Clade 12. Some such as *Pratylenchus* are migratory endoparasites which enter the plant root and migrate inter- or intracellularly, causing significant damage to the root structure in the process. However, the most complex and highly evolved interactions are seen in the biotrophic sedentary endoparasitic nematodes, including the cyst and root-knot nematodes. These nematodes form feeding structures in the roots of their host plants (syncytia in the case of cyst nematodes, giant cells in root-knot nematodes) which the nematode solely relies upon for food throughout maturation. Development of these structures is underpinned by extensive reprogramming of host gene expression as well as the evasion of host resistance mechanisms (M. A. Ali et al. 2017). In crops, these nematodes act as a 'nutrient sink' that reduces biomass and yield.

It has been estimated that plant parasitic nematodes (PPN) cause yield losses valued $78-125 billion annually (Abd-Elgawad and Askary 2015). However, the true extent of damage caused is difficult to establish as determining levels of field infestations is challenging, and growers are often unaware of PPNs as they are microscopically small, soil dwelling, and cause non-specific symptoms (J. T. Jones et al. 2013). The most damaging nematode pest in the UK is the potato cyst nematode (PCN).

In Scotland, PCN caused estimated losses of £25 million in 2019, potentially rising to £125 million by 2040 (Blok et al. 2020). For PCN the cost impact is not just associated with yield loss and costs of control measures, but is also due to restrictions against exporting seed potatoes from infested land to curb the spread of PCN. Despite current quarantine controls, PCN has spread globally, with severe infestations identified in

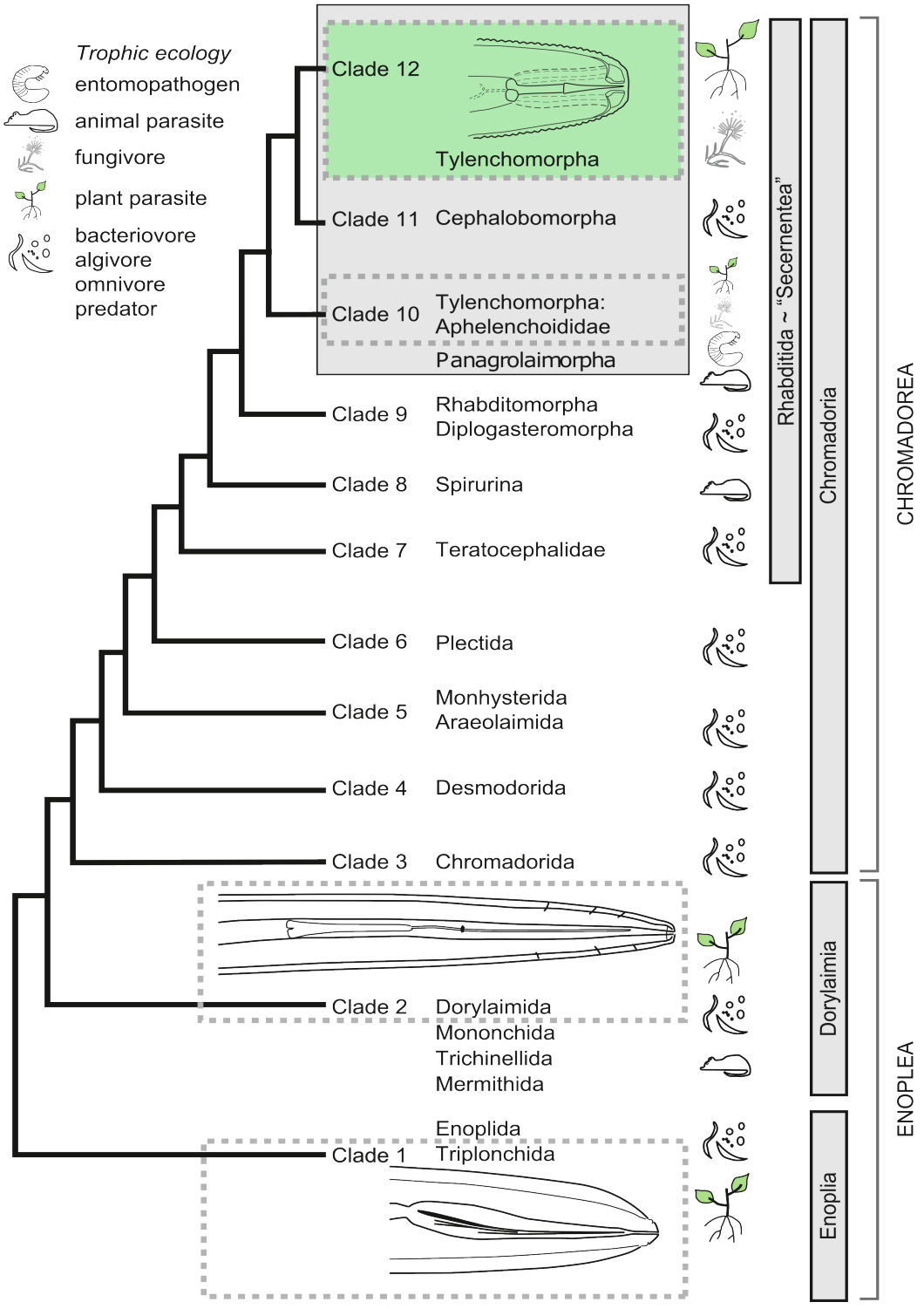


Figure 31: **Overview of the Nematoda phylum.** Adapted from Bert, Karssen, and Helder (2011). Major plant parasitic clades are indicated by dotted boxes, clade 12 (highlighted in green) contains the majority of the most damaging PPN species including PCN.

East Africa since its discovery there in 2015 (Mburu et al. 2020).

The Potato Cyst Nematode

Cyst nematodes are one of the most economically important group of PPN and infect many staple crops including soybean, cereals, and potatoes (J. T. Jones et al. 2013). In total, there are eight genera containing cyst nematodes within the Heteroderinae subfamily, the most economically important being *Heterodera* and *Globodera* (fig. 31) (Moens, Perry, and Jones 2018). PCN are within the genus *Globodera* with the most important species being *Globodera rostochiensis* and *G. pallida*.

PCN are sedentary endoparasites - juveniles embed themselves within the plant root and form a permanent feeding structure called a syncytium which acts as their sole source of nutrients through all sedentary stages of development. Their name refers to the cysts produced by the hardening of the mature female's body to create a protective egg case that can persist in soil for multiple seasons until favourable conditions arise.

This creates two major difficulties in the control of PCN. As it can persist in the field for up to 30 years PCN requires long crop rotations to effectively reduce populations, and its relative inaccessibility as a soil-bound parasite means that alternative treatment options such as fumigation are costly and damaging to the environment (Back et al. 2018). In the UK, *G. rostochiensis* was originally the predominant species of PCN present, but the incidence of *G. pallida* has increased substantially over the past 30 years due to over-reliance upon cultivars containing the *H1* gene which is effective only against *G. rostochiensis* (Blok et al. 2020).

Life cycle

Hatching of the dormant second stage juvenile (J2) from eggs within cysts in the soil is initiated under favourable environmental conditions and importantly, in the presence of host root diffusates. The sensitivity of PCN to root diffusate relates to its narrow host range as it ensures that the life cycle is only restarted in the presence of a host. By contrast, cyst nematodes that have broader host ranges are less dependent on diffusates for hatch (Moens, Perry, and Jones 2018). When the preferred conditions are met, the J2 emerges from the egg casing using its stylet.

Once hatched, the J2 rapidly migrates towards the host root and penetrates close to the root tip in the elongation zone (Perry and Moens 2011). The J2 then migrates intracellularly through the root towards the vascular cylinder causing considerable damage in the process. Here, the J2 induces a permanent feeding site called a syncytium (fig. 32). The syncytium is formed by fusion of protoplasts of adjacent root cells that are incorporated sequentially into the developing syncytium following controlled breakdown of the cell wall. After establishing a syncytium, the J2 becomes sedentary, feeding off plant resources as it undergoes three moults until developing into a male or female adult - lower nutrient availability leads to male development.

By this stage the female has become large and swollen and protrudes from the root. By comparison, the male is significantly smaller and worm-like and exits the root to locate and mate with the protruding female. Following successful mating, eggs develop within

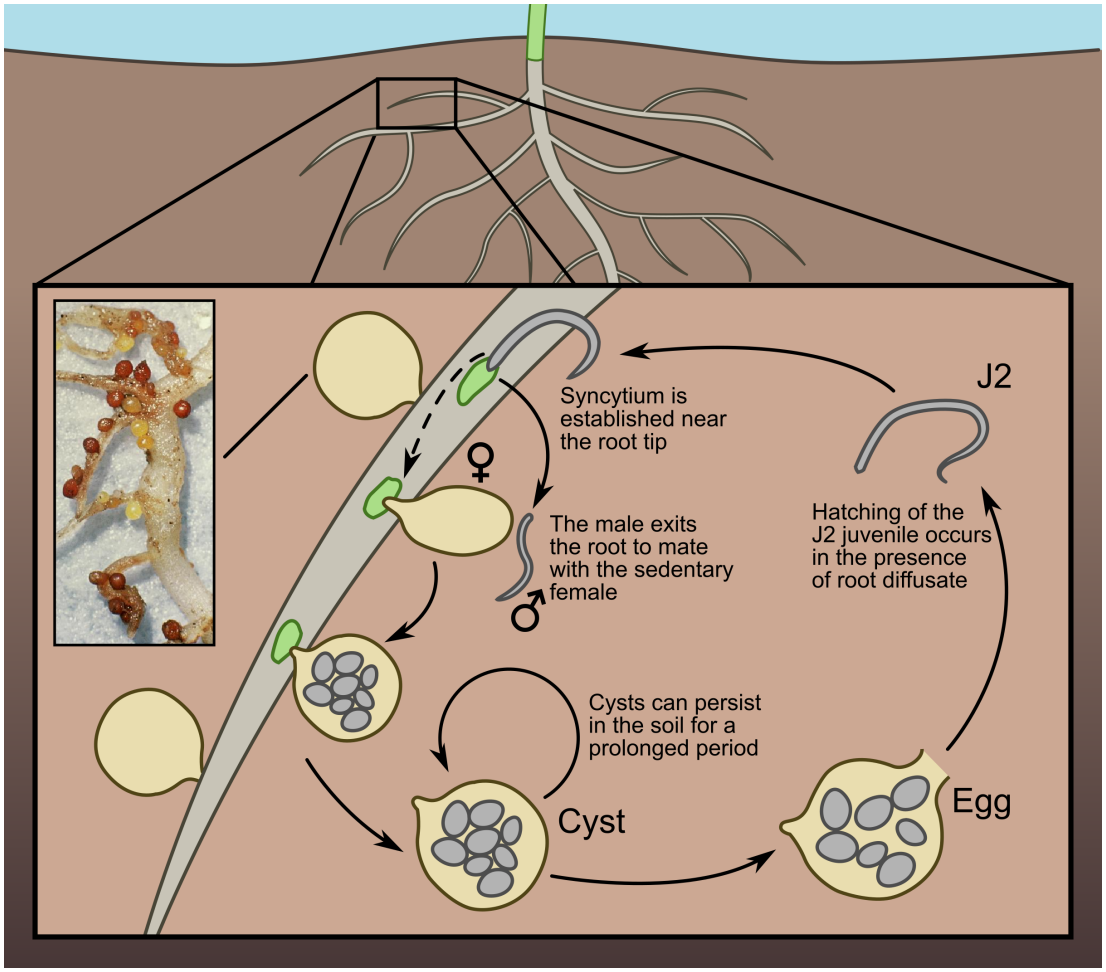


Figure 32: **An outline of the lifecycle of Globodera.** Soil bound cysts hatch in response to root diffusate and the developed J2 juvenile migrates to the root. Following the formation of a syncytium by the J2, the nematode undergoes further development and sexual differentiation. Eggs develop within the fertilised female which subsequently detaches from the root and re-enters the soil phase. Pictured are mature cysts formed on potato root (source: https://www.agric.wa.gov.au/potatoes/potato-cyst-nematode-western-australia)

the female body which subsequently dies as the cuticle hardens to form a protective case around the eggs. Once detached from the host root, the unhatched nematodes within the eggs in the cyst enter a period of overwintering dormancy, preventing hatch even in the presence of root exudates, which normally ends following a rise in soil temperature in the late spring, coinciding with the arrival of the next potato crop (Perry and Moens 2011).

Nematode effectors manipulate plant immunity

All biotrophic plant pathogens secrete effectors into their host plants which manipulate host metabolism to the benefit of the pathogen. For PPNs, effectors are secreted into the host in order to degrade physical barriers, overcome plant immunity, and induce the developmental changes in plant tissue required for establishment and maintenance of the feeding structure. In PPNs, the majority of effectors are produced in two sets of glands (the subventral and dorsal gland cells) and delivered to the host cells through the stylet (J. T. Jones and Mitchum 2018).

As effectors manipulate such a broad range of host functions and are subject to intense selection pressure, they exhibit remarkable diversity and abundance. In *G. pallida* alone there are several hundred putative effectors which exhibit distinct patterns of temporal expression throughout the nematode's lifecycle, indicating that they function in different phases of parasitism (Thorpe et al. 2014). Effectors from diverse plant pathogens, including PPN, migrate to distinct subcellular locations within the host cells, again suggesting that they manipulate distinct host functions (Thorpe et al. 2014; S. Wang et al. 2019).

The first barrier to infection encountered by the migrating cyst nematode is the physical barrier presented by the cell wall. Cyst nematodes secrete cell wall degrading enzymes during migration to soften and aid the mechanical breakdown of plant cell walls during migration (Smant et al. 1998; Bohlmann and Sobczak 2014). Expansins are also secreted by the nematode (Qin et al. 2004; Wieczorek et al. 2006) which do not have any enzymatic activity, but instead weaken the cell wall by disrupting internal hydrogen bonds within its structure (McQueen-Mason and Cosgrove 1994).

Cyst nematode effectors have been demonstrated to undermine PTI and induce ETS. For example, transient expression of the 12-amino acid *G. rostochiensis* effector GrCEP12 suppressed the PTI responses of plants challenged with flg22 (S. Chen, Chronis, and Wang 2013). The rapid production of ROS was reduced, as was the expression of two PTI marker genes *NbPti5* and *NbAcre31*.

Transgenic potato lines expressing GrCEP12 are susceptible to infection by both *G. rostochiensis* and the unrelated scab causing *Streptomyces scabies* (Chronis et al. 2013), demonstrating that effectors compromise plant immunity as well as the broad-spectrum resistance mediated by PTI. Effectors also suppress components of plant immunity that regulate ETI. For example, the *G. rostochiensis* effector SPRYSEC-19 can suppress the signalling response of multiple plant resistance proteins following effector recognition (Postma et al. 2012).

Cyst nematodes can also directly target NLRs to suppress the immune system. The *G.*

rostochiensis effector SPRYSEC-15 binds to the NB-ARC domain of the helper NLRs NRC2 and NRC3 (Derevnina et al. 2021). By binding to a "hinge" in the NB-ARC domain, the NRC is immobilised and cannot form the resistosome structure necessary for proper resistance function. Given that the NRCs act as central nodes in NLR mediated immunity, this serves to suppress the activity of a significant proportion of the NLR inventory of the host genome. Interestingly, oomycete effectors have also been identified that interact with the same NRC surface but have evolved independently of *G. rostochiensis*, highlighting that pathogens will converge on pivotal systems to suppress immunity.

Resistance gene discovery and its challenges

Natural disease resistance is the most cost efficient, environmentally friendly, and effective way of controlling PCN. Resistance genes against PCN do not prevent infection by the J2 but suppress reproduction of the nematode, reducing population levels in the soil.

Resistance can operate in several distinct ways. The hypersensitive response (HR) may, in some cases, be targeted against the developing feeding site at an early stage of infection, as seen most clearly for the root-knot nematode resistance gene *Mi-1.2* (Milligan et al. 1998). However, for most disease resistance genes effective against PCN, the resistance response is initiated later during infection, often targeting cells surrounding the developing syncytium and restricting development of this structure (Sobczak et al. 2005). This may cause a failure of parasitism or, in some cases, a shift in the sex ratio towards males, reflecting the fact that sex in PCN is determined by environment (food abundance) rather than genetics (Trudgill 1967). In either case, the impact is a reduction in the number of cysts formed.

In order to maintain and improve crop productivity, new genetic sources of disease resistance must be continuously delivered to commercial crop lines. Germplasm collections such as the Commonwealth Potato Collection (CPC) offer a significant untapped source of potential resistance traits from wild relatives of current crop cultivars. Resistances encoded by a single genetic locus are particularly desirable due to their ease of discovery and introgression into commercial lines.

The *H1* gene was the first source of resistance against PCN to be successfully utilised. Effective against the Ro1 pathotype of *G. rostochiensis*, *H1* was sourced from the CPC accession *S. tuberosum* ssp. *andigena* (CPC 1673) in 1952 and introduced into cultivars following three backcrosses (Bradshaw and Ramsay 2005). Interestingly, despite long term efforts to identify *H1*, the gene itself has evaded molecular cloning and characterisation. *H1* is still in use today nearly 70 years after its discovery, although over-reliance on this source has led to selection for *G. pallida*, against which *H1* is not effective, requiring new resistance sources to be utilised. Since this time screens of a variety of germplasm source have led to the identification of a large number of resistance genes against PCN (tbl. 10).

Table 10: **Known sources of PCN resistance.** PCN resistance genes and their chromosome location (chrom.), type (gene, variant of a gene, or locus), population they are effective against (Ro - *G. rostochiensis*; Pa - *G. pallida*), and their species of origin/identification. Adapted from (Gartner et al. 2021).

Gene/Locus	Chrom.	Type	Pathotype	Source
Gro1-4	7	Gene	Ro1	*S. spegazzinii*
Gpa2	12	Gene	Pa2/3	*S. tuberosum* ssp. *andigena*
Hero	4	Gene	Ro1, Pa2/3	*S. pimpinellifolium*
$NRC3^{K316}$	5	Variant	Ro*	Complex
H2	5	Locus	Pa1, Pa2/3	*S. multidissectum*
$Gpa^{V}spl$	5	Locus	Pa2/3	*S. sparsipilum*
$Gpa^{XI}spl$	11	Locus	Pa2/3	*S. sparsipilum*
Gpa	5	Locus	Pa2/3	*S. spegazzinii*
GpaM1	5	Locus	Pa2/3	*S. spegazzinii*
GpaM2	6	Locus	Pa2/3	*S. spegazzinii*
GpaM3	12	Locus	Pa2/3	*S. spegazzinii*
Gro1.2	10	Locus	Ro1	*S. spegazzinii*
Gro1.3	11	Locus	Ro1	*S. spegazzinii*
Gro1.4	3	Locus	Ro1	*S. spegazzinii*
$Gpa^{XII}tar$	11	Locus	Pa3	*S. tarijense*
H1	5	Locus	Ro1, Ro4	*S. tuberosum* ssp. *andigena*
H3	4	Locus	Pa2/3	*S. tuberosum* ssp. *andigena*
GpaV	5	Locus	Pa2/3	*S. vernei*
GpaVI	9	Locus	Pa2/3	*S. vernei*
GroVI	5	Locus	Ro1, Ro4	*S. vernei*
Grp1	5	Locus	Ro5, Pa2/3	*S. tuberosum*, *S. oplocense*, *S. vernei*
Ro2_A	5	Locus	Ro2	*S. tuberosum* ssp. *andigena*, *S. vernei*
Ro2_B	5	Locus	Ro2	*S. tuberosum* ssp. *andigena*, *S. vernei*
Pa2/3_A	5	Locus	Pa2/3	*S. tuberosum* ssp. *andigena*, *S. vernei*
Pa2/3_B	10	Locus	Pa2/3	*S. tuberosum* ssp. *andigena*, *S. vernei*
GpaIV	4	Locus	Pa2/3	*S. tuberosum* ssp. *andigena*, *S. vernei*

Recent advances in sequencing technologies have enhanced the rate of resistance gene discovery and characterisation. Resistance gene enrichment sequencing (RenSeq) allows the rapid annotation and mapping of NLRs in plants (Jupe et al. 2013). RenSeq utilises a library of biotinylated DNA probes designed from the annotated potato genome to enrich samples of plant DNA for NLR-like sequences which can be subsequently sequenced and mapped to the potato genome. This strategy focuses sequencing

power on the part of the genome of interest, in this case the *R* genes. Jupe et al. (2013) validated this method against the sequenced 'DM' reference potato clone and identified a further 317 novel NLRs that had been missed in the initial automated annotation. A variation of this is diagnostic RenSeq (dRenSeq) which can be applied in crop breeding programmes to validate the presence of desired resistance genes in breeding lines, as well as detect polymorphisms of these (Armstrong et al. 2019).

Recently, a combination of RenSeq and the derivative GenSeq that enriches for low copy number genes across the potato genome (X. Chen et al. 2018) were used in mapping the *H2* resistance gene (S. M. Strachan et al. 2019), which confers resistance to the *G. pallida* Pa1 pathotype (Blok and Phillips 2012). Crosses of *H2* carrying clone P55/7 and susceptible 'Picasso' were screened for resistance to *G. pallida* and subsequently subject to RenSeq and GenSeq analysis. No previously identified nematode resistance genes could account for the resistant phenotype of P55/7, and both enrichment sequencing approaches identified SNPs closely associated with resistance at the distal end of chromosome 5. This region was further narrowed down using an allele specific PCR marker assay to a size of 4.7Mb. The *H2* locus was further reduced to a 0.8Mbp locus using an extended segregating host population (S. Strachan 2018). Other approaches have combined RenSeq with long read sequence technology to capture complete NLR genes, better represent genes in homologous NLR clusters, and capture adjacent regulatory elements (Witek et al. 2016; Belinchon-Moreno et al. 2023).

Whilst sequencing technologies have improved our tracking and identification of NLR genes, a major bottleneck that remains in PCN resistance gene discovery is verifying that a candidate gene is indeed creating resistance. Cyst nematodes are difficult to culture and resistant phenotypes can be challenging to assess. Possessing a complementary effector for the candidate can improve this somewhat as candidate host genes can be tested via co-expression with the cognate effector to induce HR. Unfortunately, effectors recognised by resistance genes are rarely identified prior to genetic studies.

Another challenge is that NLR clusters are laborious to navigate experimentally where a single resistance gene locus may contain dozens of highly similar candidate genes. This complexity is increased substantially when resistance is controlled by more than one resistance gene, or by a signalling partner that is not present in the susceptible background when producing transgenic plants. For example, the tomato PCN resistance gene *Hero* could confer resistance when expressed in a susceptible tomato background but not in potato, perhaps due to the lack of signalling partners that are present in its native background (Sobczak et al. 2005). Alternatively, this could be due to divergence in a guarded protein between potato and tomato, rendering *Hero* non-functional.

Our understanding of the genetic and molecular basis of plant immunity is rapidly expanding, with advances in technologies such as RenSeq aiding resistance gene discovery and the recent discovery of the structural basis to NLR activity. Despite this, details of the mechanisms of resistance against economically important pathogens such as the cyst nematodes remain unclear and there are still relatively few resistance genes against nematodes identified. Investigating these mechanisms not only contributes

to our understanding of plant immunity as a whole, but will also have a significant translational impact on our approach to crop protection.

Chapter aims

The aims of this chapter are to:

- Develop high-throughput methods for resistance gene identification through RenSeq-based sequencing
- Identify and characterise the *H2* resistance gene
- Identify and characterise the *H3* resistance gene

Methods

HiFi-RenSeq assembly

The initial RenSeq assembly of P55/7 using CCS sequencing and analysis leading to the identification of 14 candidates was conducted prior to this study.

For the repeat of the RenSeq of P55/7 using HiFi sequencing, assemblies were produced using `hifiasm v0.19.9`, `flye v2.9.4`, and `canu v2.2`. To check for potential haplotype collapsing, `coverm v0.7.0` was used to calculate the mean read coverage of each contig for each assembly. To predict the number of NLRs identified in each assembly, open reading frames were identified and search for with the NB-ARC HMM profile (Pfam: `PF00931`) with `hmmsearch v3.4` using the default settings. The output multiple sequence alignment was reformatted to individual fasta entries with the `esl-reformat` tool packaged with `hmmsearch`. To identify homologs of known NLRs, this fasta was searched against the RefPlantNLR database.

H2 RenSeq variant analysis

Previously generated RenSeq data of a bulk segregant analysis of a P55/7 x Picasso cross was acquired (S. M. Strachan et al. 2019). Reads were trimmed with `fastp v0.23.4` and aligned to the P55/7 contigs with `bowtie2 v2.5.1` using the options `--score-min L,-0.18,-0.18 --phred33 --fr --maxins 1000 --very-sensitive --no-unal --no-discordant`. Alignments were sorted and pileups built with `samtools v1.17`.

Variants were called with `VarScan v2.4.4` using the option `--strand-filter 0` and filtered with a custom python script. Variants were retained if they had a frequency between 20-30% in P55/7 and 0-5% in Picasso. For the bulk progeny samples, variants were retained in linkage if they had a frequency of 20-30% in the resistant bulk and 95-100% in the susceptible bulk. The reciprocal condition for genes in repulsion was also calculated. Common variants between the parent and bulk samples were merged using a custom python script.

The full pipeline is available at https://github.com/swiftseal/renseq_mapping.

To compare the haplotypes of *NRC3* in P55/7, the candidate identified in the variant analysis was searched against the remaining contigs with `blastn v2.16.0` using the default settings. The full sequence of *NRC3* was predicted with `AUGUSTUS v3.50` using the default settings with the *S. lycopersicum* model. The *NRC3* haplotypes were aligned with the *N. benthamiana NRC3* sequence from the RefPlantNLR database with `mafft v7.525` using the default settings, and visualised with the Bioconductor `msa v1.36.0` package.

KASP marker design

KASP markers were designed using a python script. Briefly, `.vcf` files of variants identified in the parental and bulk population intersections were used to obtain variant positions and identities. The `.fasta` assembly was then used to capture the 50bp upstream and downstream sequence of variants identified in the bulk populations intersect. Within this sequence, variants present in the parental intersect were added and represented with IUPAC codes. The target variant was encapsulated in square brackets to comply with the LGC input format.

KASP marker sequences were filtered by `blastn v2.9.0` searching against the complete set of contigs. Marker sequences that reported significant non-self results (>95% identity, >28bp) were discarded. When multiple valid sequences were available, sequences with the fewest parental variants in the ±50bp region were selected (tbl. 11). Marker sequences were submitted to LGC for primer design.

Table 11: **Final KASP markers selected for further study.**

Contig	KASP marker sequence
utg000005l	CTCAAGTATCCGCTCTTTACTATTCAATGCCACCAGTGRTGATCAGTATA[M]TACAATGGCGCGTGATATCTCMTTCATCCTTAAYAGCTTCAAACTTGTTA
utg000010l	ATACTGAATTGCCTCAATGCCCAAACAAAAGAGCTAATAGTCATWTCACT[K]GTTGGTATGGRCGGTATAGGTAAGACAACTCTTGCTAGAAAAGTTTTTGR
utg000184l	CRTTATCAACGACCTCAAGTGTTGGTGAAAAGGCWCCCGTGTTTTGGTGC[Y]CRGAGATACAAACTTGGAAAGCTCATTACTGAAACTAACGCTCACTCGAC
utg000186l	GTCTTGGATGATATGTGGGATTGTAYGGTGTGGGATGACTTAAGGCTTTG[Y]TTTCCAGATGTTGGAAATAGAAGCAGAATAGTAATAACAACTCGACTTGA
utg000489l	TCCTAAAGATTACGAAATTCCAGTGTCTGATCTACTCAAGTGGTGGATAG[M]TGAGGAGTTTGTGCAGAACATTGACACGTTGAAGCTAGAAGAATTATCAG
utg002172l	ACTTTATCAAAGTTYCGTCTACCTTGGACCCAAATTTCGATCATTGCAGA[R]CTGCCCAACTTGGTGATTCTTAAGTTATTGCTCAGAGCCTTTGAAGGGGA
utg002291l	TATTCATTCTGGCTAARGACTTGGAGACAATCACATCCTGCAATCTCTTC[Y]ACTTCTTTTTTAATATCGAGAGCTGAAAGCTTAAGATACTTRTTGCAGCG
utg002533l	TCATTGCTCTCATTTCTGGAAGTTCTGGCMATAGAGRCAGCAATTGCAAT[M]TACTCCTTGTGACATGGATTTGGAGAACAGTATTGCTGAAGTAGACCATA
utg003022l	CCTTGTTCTGGTATAAATTCCACTTTTTCTCGTTCCAGACCTGTGTGATC[W]TGCTCTCCGATAAATCAAGAACTGCAATCTTTCGGGACAACAACTCTGGA
utg003165l	TTGAAACTTTAATAGTTAAAGGACTTGGAGGACGAGTAACTTTACCAGAY[R]CCATCTGGAAGATGGTCAAGTTGCGCCATTTGCACGTATACAACCRCGCT
utg004020l	GTTCACTATTGTTATATMTACTGCTCCTGATCCCAATGTAACGTTCTGGT[R]GTTCATCGTTGTYTTCCCTCATCTGTAACAAGGTCTCAACAAGCTCAGGC
utg004097l	AAGGTTGAGTATGAGTTCTTCACCCGCGTTTGGGTATATGTCTCTCAAAC[R]TTCAAGAGAAGGGAAATATTTCTCAACATTATCAGCAAGTTYACTCGAAA
utg004897l	TCCGTGGACGGGTCCAGCACTCAGCAAATGAGGCTTCCATTATCTGATCT[K]CTACAAGAGATTGAGACTGYCAAGGTAGAGTTCAGAAAAGTATTCTTTCA
utg004988l	ATGCTTTTCCTCGCCTTGAACACTTGGTACTGAGAAGATGTCGATATCTC[R]AGGCAATCCCTTCTCGCTTTGGAGACATCACATCTCTAATATCCATTGAG

KASP genotyping was carried out on the Applied Biosystems™ StepOne™ system. Plant DNA was isolated from 100mg leaf tissue via the Qiagen DNeasy® Plant Mini Kit according to the manufacturer's guidelines. Frozen tissue was disrupted using a liquid-nitrogen-cooled TissueLyser II with a metal bead (2x 1min, 30Hz). DNA was eluted in 100µL Buffer AE or nuclease-free water, although this could be reduced to 50µL when higher concentrations were required. DNA concentration was quantified by Qubit™ dsDNA HS assay kit.
The KASP PCR was conducted in a reaction containing 5µL DNA (20ngμL^{-1}), 5µL 2X KASP-TF Master Mix, and 0.14µL KASP Assay Mix. Samples underwent a PCR cycle for amplification prior to genotyping (tbl. 12).

Table 12: **PCR conditions used for KASP assay.**

Temperature	Time	Cycle
94°C	15 min	x1
94°C	20 sec	x10
65°C (Δ-0.8°C)	60 sec	↓
94°C	20 sec	x26
57°C	60 sec	↓
94°C	20 sec	x3
57°C	60 sec	↓

Sample genotypes were determined by plotting HEX vs FAM fluorescence normalised to ROX reference dye fluorescence. Genotype was assigned when three sample replicates clustered with a minimum of five parental samples.

When KASP assays were not successful for determining the genotype of progeny, sanger sequencing was instead conducted. PCR was conducted with primers (tbl. 13). PCR products were sequenced at the James Hutton Institute and genotypes determined by aligning sanger sequencing traces to contigs and manually inspecting SNPs with Geneious Prime.

Table 13: **PCR primers used to determining contig genotype of P55/7 x Picasso progeny.**

Contig	Left primer	Right primer
utg000184l	GGAAGTCCTGCAATGACGGA	GTCTATGGATGGCGGAAGGG
utg000186l	AGGCCCTACTCTGTGGTTGA	TGCACGAATCCTTCTGCGAT
utg000489l	AAGTCATGTCCTTCGCAGCT	TTCGAGTCATGTCAGCCACG
utg003165l	GGTTGTGGACAGCAGAAGGT	GCAAATTCAGCTCAGTGGCC
utg004020l	AGCTGGATTTGCTGCCAAGA	TCCGCTATGTGATGATGGGC
utg004097l	ATGAGTTCTTCACCCGCGTT	TGCAACATGTCATGAACGCG

RNA-RenSeq

RNA-RenSeq of P55/7 was carried out using a myBaits hybridization capture kit for NLR sequence enrichment. RNA was extracted from root and leaf tissue under PCN infected and control conditions with the Qiagen RNeasy Plant Mini Kit The cDNA was fragmented, end repaired, and dA-tailed by combining 7µL NEBNext® Ultra™ II FS Reaction Buffer with 2µL NEBNext® Ultra™ II FS Enzyme Mix to 26µL DNA (500ng; 19.23ngµL^{-1}). A fragmentation size of 500-1000bp was achieved by incubating at 37°C for two minutes followed by 30 minutes at 65°C. To the FS reaction mix, 30µL NEBNext® Ultra™ II Ligation Master Mix, 1µL NEBNext® Ligation Enhancer, and 2.5µL NEBNext® Adaptor were added. The adaptor ligation mix was incubated at 20°C for 15 minutes, 3µL USER™ enzyme added, then incubated at 37°C for 15 minutes.

Libraries were purified with a 1:1 AMPure XP bead cleanup into 15µL 0.1X TE. PCR enrichment for adapter-ligated sequence was achieved by addition of 25µL NEBNext® Ultra™ II Q5 Master Mix and a 10µL mix of the index and universal primer, followed by thermocycling:

Temperature	Time	Cycle
98°C	30 sec	1x
98°C	10 sec	5x
65°C	75 sec	↓
65°C	5 min	1x

Adapter-ligated DNA was purified with a 1:1 AMPure XP bead cleanup into 30µL buffer AE. Target yield for RenSeq was ~750ng - the number of PCR cycles could be adjusted accordingly. DNA size distribution was determined on a 0.8% agarose gel.

For hybridisation, 500ng of the indexed DNA libraries was evaporation-concentrated to a volume of 5.9µL. To this, 2.5µL human Cot-1 DNA (1µgµL^{-1}), 2.5µL salmon sperm DNA (1µgµL^{-1}), and 0.6µL myBaits blocking agent was added and incubated for five minutes at 95°C. This was transferred to a capture bait mix of 5µL capture probe and 1µL RNase block which had been pre-warmed to 65°C. To this, 10.5µL of a 65°C pre-warmed hybridisation master mix (7.1µL 20X SSPE, 2.84µL 50X Denhardt's solution, 0.28µL 0.5M EDTA, 0.28µL 10% SDS) was added immediately. Hybridisation proceeded for 24 hours at 65°C.

To recover the captured targets, 50µL Dynabeads® MyOne™ Streptavidin C1 magnetic beads were washed and resuspended in 20µL myBaits Binding Buffer and added to the hybridisation reaction mix. The mix was incubated at 65°C for 15 minutes. Beads were pelleted and washed three times in myBaits Wash Buffer 2 with complete resuspension and incubation at 65°C for five minutes. Captured sequencing library was resuspended in 30µL nuclease-free water.

Post-hybridisation PCR amplification was carried out using 15µL of the adapter-ligated DNA bead suspension. To it, 25µL 2X KAPA HiFi HS RM, 2.5µL of each primer, and 5µL nuclease-free water was added. This was thermocycled as followed:

Temperature	Time	Cycle
98°C	45 sec	1x
98°C	15 sec	10x
60°C	30 sec	↓
72°C	60 sec	↓
72°C	5 min	1x

The reaction mix was cleared of residual Dyanbeads. Libraries were purified with a 1.8X AMPure XP bead cleanup and eluted into 20µL nuclease-free water. Libraries were sequenced on a MiSeq sequencing platform at an equimolar ratio.

Iso-RenSeq of 12601ab1 leaf tissue was carried out independently of this thesis by Arbor Biosciences.

`HISS` AgRenSeq workflow

An automated AgRenSeq workflow was developed with `snakemake v7.20.0`. For compatibility with the SLURM workload manager, `cookiecutter v2.1.1` was used to develop a SLURM compatible workflow from the profile available at https://github.com/Snakemake-Profiles/slurm.

Diversity panel RenSeq reads were pre-processed with `fastp v0.23.2` using the default options. From these, *k*-mers were counted with `jellyfish v2.2.10` with the options `count -C -m 51 -s 1G -t 4`. Tab-delimited dump files were then created with the options `dump -L 10 -ct` and added to a tab-delimited configuration dataframe as a prerequisite for the AgRenSeq *k*-mer presence/absence matrix.

The AgRenSeq matrix was created using `AgRenSeq_CreatePresenceMatrix.jar` with the default settings.

Association scoring was conducted with `AgRenSeq_RunAssociation.jar` using the presence/absence matrix and the phenotype scores of the cultivars.

Results of the AgRenSeq run were parsed with `R v4.2.2` with the libraries `dplyr v1.0.9` and `ggplot2 v3.3.6`. The approximate location of contigs in the reference potato genome were predicted using `BLASTn v2.13.0` with the default options.

The AgRenSeq workflow was integrated into the `HISS` package (available at https://github.com/SwiftSeal/HISS).

H3 candidate identification

An assembly of 12601ab1 HiFi-RenSeq sequencing was produced via the `HISS v2.1.1 smrtrenseq_assembly` workflow with the default settings. Contigs containing NLRs identified by the workflow were used as input for the `HISS v2.1.1 agrenseq` workflow using a panel of known *H3* positive and negative cultivars. Contigs containing *k*-mers with an association score equal or greater than 27 were selected as *H3* candidates.

Iso-RenSeq reads of 12601ab1 leaf tissue were trimmed using `cutadapt v4.7` with the options `-a A(x100) -g T(x100) -O 12 -j 4`, following by a second trim with the options `-a CCCATGT$ -g ^ACATGGG -j 4`. Trimmed reads were aligned to the contigs using `minimap2 v2.28 -ax splice:hq -ub -G 5000 --secondary=no`, which were sorted and index using `samtools v1.20` with the default settings. The per-base depth of Iso-RenSeq reads using `bedtools v2.31.1 coverage` with the NLR coordinates determined by NLR Annotator as part of the `smrtrenseq_assembly` workflow. The mean depth of NLRs was determined with `polars v1.2.1` and any NLRs with zero coverage were removed.

To further reduce the candidate list, candidates were screened against a dRenSeq panel of 1577 cultivars, breeding clones, and wild species using a reimplementation of `HISS` in Nextflow (available at https://www.github.com/SwiftSeal/nfHISS). Briefly, reads were filtered with `fastp v0.23.4` using the default settings and aligned to the candidate contigs with `bowtie2 v2.5.3` using the options `--score-min L,0,-0.24 --phred33 --fr --maxins 1000 --very-sensitive --no-unal --no-discordant -k 10`. Alignments were sorted with `samtools v1.19.2` using the default settings, reads with no mismatches were filtered for with `sambamba v1.0` using the option `--filter='[NM] == 0'`, and NLR read coverage calculated with `bedtools v2.31.1 coverage` using the default settings. NLR coverage values for each cultivar were merged into a single matrix with `r-tidyverse v1.2.1`. The dRenSeq matrix was visualised with `ComplexHeatmap v3.19` and manually inspected to remove NLRs that exhibited false positives and false negatives.

Simultaneously, unfiltered RenSeq read alignments were used for variant identification for the purpose of KASP marker design. Variants were identified with `freebayes v1.3.6` using the options `--min-alternate-count 2 --min-alternate-fraction 0.05 --ploidy 4 -m 0 --legacy-gls`. Variants were sorted with `bcftools v1.19 sort`, compressed with `bgzip v1.19.1`, indexed with `tabix v1.19.1`, and merged with `bcftools v1.19 merge`.

To identify homologs of candidate NLRs, the NLR Annotator sequences were searched for open reading frames with `getorf v6.5.7` using the default settings, and searched for NB-ARC domains with `hmmsearch v3.4` with the PF00931 model. The multiple sequence alignment file produced by option `-A` was reformatted as a fasta with `esl-reformat v3.4`. To identify homologs of known NLRs, the NB-ARC domain sequences were searched against the RefPlantNLR database with `blastp v1.15.0` using the default settings. To identify homologs in the *S. stenotomum* genome, NB-ARC domains were searched against the gene predictions produced in Chapter 1 using the same strategy.

Transgenics

H2 candidate genes with their 2kbp upstream region were synthesised by GenScript using the P55/7 RenSeq assembly and cloned into a pCR8/GW/TOPO vector. The exCNL candidate was unnstable and prepared in the GenetoxicV2 strain. Candidates were cloned into pK7WG destination vectors using 150ng of entry and destination vector in 8µL TE to which 2µL Gateway™ LR Clonase™ II was added. The reaction

was incubated for one hour at 25°C before addition of 1µL proteinase K and incubation for 10 minutes at 37°C. Destination vectors were transformed into LBA4404 via electroporation using 10µL cell culture, 10µL 10% glycerol, and 1µL plasmid. Transformed cells recovered in 1mL YEB for four hours at 28°C without selection before plating on YM plates (100µgmL^{-1} kanamycin, 100µgmL^{-1} streptomycin) at a 1:10 dilution.

Agrobacterium cultures were initiated from 1mL cryopreserved glycerol stock in 5mL LB medium with 50mgL^{-1} kanamycin monosulfate at 28°C with agitation for 24 hours. Then, the 24 hour *Agrobacterium* culture was diluted into 100mL LB medium with 50 mgL^{-1} kanamycin monosulfate and incubated at 28°C with agitation for 6 hours prior to transformations.

25mL of *Agrobacterium* culture was sedimented at 1500g for 10 minutes and resuspended in 5mL MS20. Internode and petiole explants from 6-8 week old *in vitro* cultured plantlets were submerged in 15mL MS20 supplemented with 1mL of resuspended *Agrobacterium* for 5-10 minutes. Explants were blotted on sterile filter paper and placed onto HB1 medium for two days under low light conditions at 18-22°C. Explants were sub-cultured onto HB1 regeneration media (MS20 with 8gL^{-1} agar, 0.2mgL^{-1} NAA, 0.02mgL^{-1} GA3, 2.5mgL^{-1} Zeatin riboside) supplemented with 500 mgL^{-1} cefotaxime for two weeks and then transferred to HB2 regeneration media (MS20 with 8gL^{-1} agar, 0.02mgL^{-1} NAA, 0.02mgL^{-1} GA3, 2mgL^{-1} Zeatin riboside) with cefotaxime and kanamycin for two weeks. Explants were transferred to fresh HB2 regeneration media with cefotaxime and kanamycin every two weeks. Regenerated shoots were removed from explants and rooted on MS20 with 500mgL^{-1} cefotaxime and 50mgL^{-1} kanamycin. Explants were transferred to glasshouse and maintained through cuttings.

Results

Identifying *H2* candidates

Previously, the *H2* gene has been mapped to a 4.7Mbp interval on chromosome 5 of the DM v4.03 reference potato genome. A previously generated assembly of RenSeq-enriched PacBio CCS reads for P55/7 which contains *H2* was obtained. From the assembly, 14 contigs had been identified which contained SNPs associated with *H2* resistance based on mapping of the previously RenSeq mapping data (S. M. Strachan et al. 2019).

To further reduce the candidate list, a combination of KASP markers and PCR assays were carried out to identify those that were highly linked with resistance. Of the 14 KASP markers designed, 9 were deemed sufficient to identify the resistant and susceptible haplotypes. Where KASP markers could not readily differentiate between the resistant and susceptible genotypes, PCR amplification and sanger sequencing was conducted to manually validate the SNP identity. Rather than screening the full progeny panel with markers, a subset of seven resistant and seven susceptible progeny that showed recombination close to the *H2* locus were selected. From this panel, only two contigs contigs - `utg0040971` and `utg0048971` - showed complete association

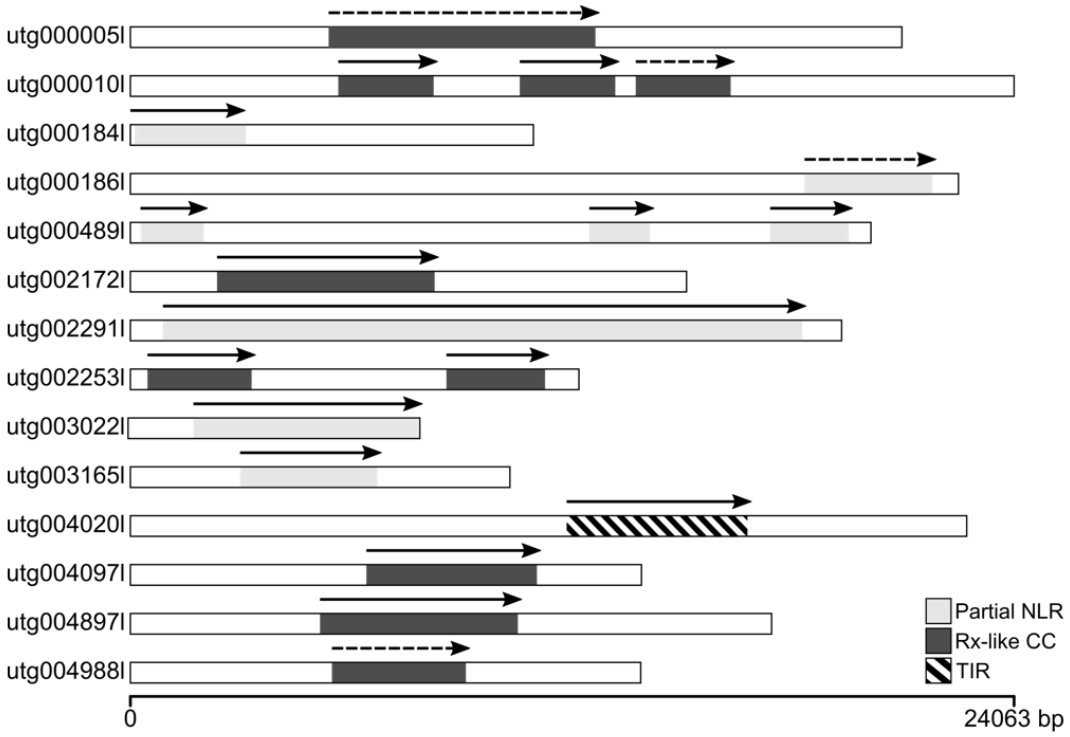


Figure 33: **H2 candidates identified by RenSeq.** Fourteen contigs were identified that were linked with H2 resistance and contained NLRs. Both RxCC and TIR NLRs were identified by NLR Annotator, along with NLRs that could not be classified. Some contigs contained multiple NLRs.

with the resistant phenotype (tbl. 16).

Table 16: **Genotypes of the segregating progeny of the P55/7 x Picasso cross.** R: Resistant, S: Susceptible, -: Ambiguous. Resistant progeny are in the order 61, 113, 133, 137, 278, 374, 604. Susceptible progeny are in the order 41, 64, 104, 175, 331, 481, 584. Numbers are linked to progeny IDs used in S. Strachan (2018).

Marker	Resistant progeny	Susceptible progeny
utg000005l	R,S,R,S,S,S,S	S,S,S,S,R,R,R
utg000010l	R,R,R,S,S,S,S	S,S,S,S,R,R,R
utg000184l	R,S,R,R,S,R,R	R,S,R,S,R,S,S
utg000186l	R,S,R,S,S,S,S	S,S,S,S,R,R,R
utg000489l	-,-,-,-,-,-,R	S,S,S,-,-,R,R
utg002172l	R,S,R,S,S,S,S	S,S,S,S,R,R,R
utg002291l	S,R,R,R,S,S,R	S,S,S,R,S,S,R
utg002533l	R,R,S,R,S,S,S	S,S,R,S,S,R,S
utg003022l	R,S,R,S,S,S,S	S,S,S,S,R,R,R
utg003165l	R,S,R,S,S,S,S	S,S,S,S,R,R,R
utg004020l	S,R,S,R,S,R,R	R,R,R,S,R,R,S
utg004097l	R,R,R,R,R,R,R	S,S,S,S,S,S,S
utg004897l	R,R,R,R,R,R,R	S,S,S,S,S,S,S
utg004988l	R,S,R,S,S,S,S	S,S,S,S,R,R,R

Each contig contained a single NLR, both of which had a canonical CNL structure. The utg0040971 NLR is a close homolog of *NRC3* (pident: 96.86%, length: 891, evalue: 0.0, bitscore: 1784), whereas utg0048971 is most similar to the *N. benthamiana* NLR *NbPRFa* (pident: 67.31%, length: 1043, evalue: 0.0, bitscore: 1404) and contains a Solanaceae domain (Pfam signature PF12061), suggesting that it is a member of the extended CNLS (exCNLs) family of *Solanaceae* (Seong et al. 2020).

To validate the predicted NLRs and verify their expression, cDNA-RenSeq was carried out for leaf and root tissue across infected and uninfected conditions. No substantial difference in expression across the conditions was observed for either gene. For both candidates, the cDNA-RenSeq data validated the predicted gene structures and confirmed that they were unique to P55/7 and not present in Picasso (fig. 34).

To identify whether each gene was the true *H2* gene, both candidates were transformed into a susceptible Desiree background. The exCNL candidate was unstable in plasmid and required additional transgenic plants due to it exerting a strong stress response in the plant tissue. In total, 30 transgenic lines were established for the exCNL candidate and 28 for the *NRC3* candidate. Unfortunately, all transgenic material failed in the glasshouse due to extraneous variables and was not recoverable before phenotyping could be conducted.

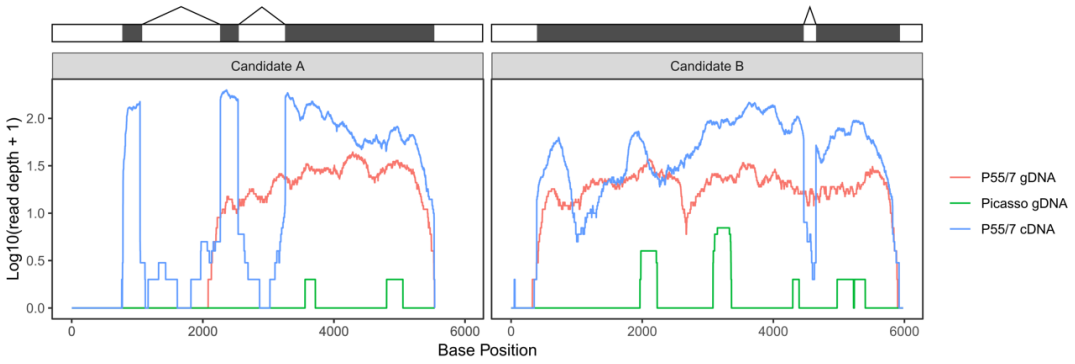


Figure 34: **H2 candidate expression.** The log read depth of gDNA and cDNA RenSeq libraries from P55/7. gDNA RenSeq data for Picasso is also plotted. For brevity, all RNA sequencing conditions have been merged. Left: the NRC3 homolog identified in `utg0040971`. Right: the exCNL candidate identified in `utg0048971`.

Revisiting *H2*

To obtain a more complete assembly of *H2* candidates, HiFi-RenSeq of P55/7 was conducted to replace the previous CCS reads which have a higher error rate. In total, 787,980 reads of an average length of 4.5kbp and quality of 28.95 were obtained, representing in total 3.57Gb of sequence data. Multiple assembly strategies were selected which produced varying numbers of contig count, contig size, and total coverage (tbl. 17). The assembly produced by `flye` was an outlier with a significantly lower number of contigs and the highest contiguity. To determine how well each assembly strategy predicted the NLR inventory of P55/7, the assemblies were searched for NB-ARC domains indicative of NLRs. The `canu` assembly produced the largest number of identifiable NB-ARC domains whilst `hifiasm` and `flye` produced similar counts, despite `flye` assembly having fewer contigs in total.

Table 17: **Assembly statistics for P55/7 HiFi-RenSeq.** The total number and N50 values of contigs produced by Hifiasm, Canu, and Flye. The total size of the final assembly is also present.

Assembler	# Contigs	Contig N50	Total size	NBARCs
Hifiasm	6,987	11.5 Mbp	81.8 Mbp	2,995
Canu	7,936	10.2 Mbp	80.5 Mbp	3,444
Flye	4,278	13.2 Mbp	52.3 Mbp	2,963

It was hypothesised that the lower contig count in the `flye` assembly was due to haplotypes being collapsed into single contigs. To visualise this, the read coverage of contigs was calculated, with the assumption that collapsed contigs would be identifiable as a multiple of the mean coverage. Across all three assemblies, no evidence of significant levels of haplotype collapsing was observed although a minor "shoulder" was noted at 50x coverage in the assembly produced by `flye`. (fig. 35).

The assemblies were searched for homologs of known NLRs to visualise how different

assembly strategies impact NLR recovery (tbl. 18). An identical or very similar number of copies was identified for the majority of NLRs including the essential NRC family. Interestingly `hifiasm` failed to capture homologs of the NLRs *Rpi-chc1.1* and variants of *Tm2*. The number of copies identified by `canu` was elevated for some NLRs.

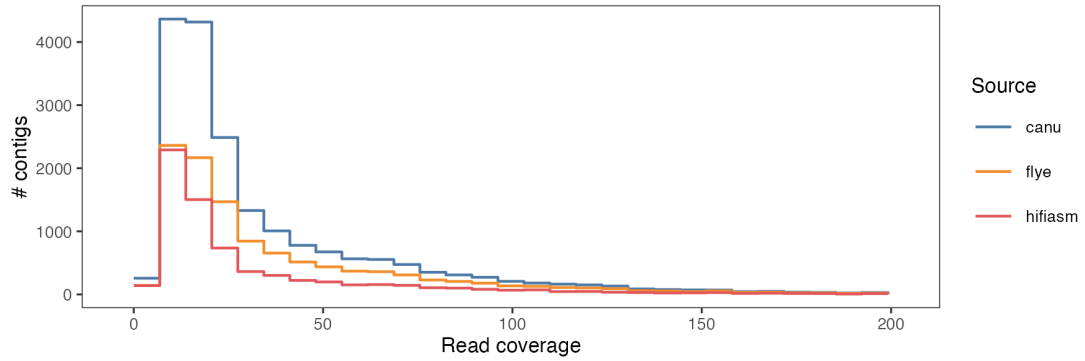


Figure 35: **HiFi-RenSeq assembly read coverage.** The mean read coverage of contigs of the `flye`, `hifiasm`, and `canu` assemblies was calculated with `coverm`. Significant levels of haplotype collapsing would be identifiable as a multimodal distribution.

Table 18: **NLR Homologs identified in P55/7 assemblies.**

Homolog	Hifiasm	Flye	Canu
Rpi-sto1	2	2	1
R9a	9	10	10
Rx	4	5	5
R8	8	5	11
Rpi-chc1.1	0	1	1
StrPtr1	2	3	3
Sw5-b	5	4	9
Rpi-pta1	2	2	1
R1	1	1	1
Rpi-mcq1.1	1	1	1
NRC2	2	2	2
NbADR1	3	3	3
Ry_sto	6	6	7
Rpi-blb1	2	2	1
Gro1-4	3	3	4
NRC0	5	5	5
Y-1	1	1	1
NRC3	3	3	3
Rpi-hjt1.3	3	4	4
Rpi-hjt1.2	1	1	1
RPi-abpt	1	1	1
Tm2^2	0	1	1
Rpi-bt1	5	4	5

Homolog	Hifiasm	Flye	Canu
Prf	2	2	2
Rpi-hjt1.1	3	4	4
R3b	1	2	1
NRC6	15	13	15
Rpi-hcb1.1	5	5	4
Rpi-edn2	9	10	10
Rpi-mcq1.2	1	1	2
Ptr1	2	3	3
R2-like	1	1	1
Tm2	0	1	1
tm2_sus	0	1	1

It was concluded that whilst all assembly strategies perform well in terms of NLR count and haplotype collapsing, canu would be the most appropriate assembly strategy for identifying unknown NLRs, as it identified the largest (albeit potentially inflated) complement of NLRs for P55/7.

The *H2* RenSeq variant analysis was repeated using the canu assembly and a variant filtration protocol that searched for variants represented in both the bulk and parental samples and had an allele frequency expected of the 1:4 resistant:susceptible ratio. Two contigs were identified through this approach - `tig00000244` and `tig00000556` - both of which contained a single SNP with a shared allele frequency indicative of being resistant haplotype.

The contig `tig00000244` contained two expressed NLRs, both with homology to the *R8* resistance gene of *S. demissum* (Identity: 39.07%, E-value: $7x10^{-134}$; Identity 40.19%, E-value: $3.3x10^{-132}$). Although *R8* is mapped to chromosome 9 of *S. demissum*, both NLRs share a closer homology with uncharacterised NLRs in *S. pinnatisectum* (`KAK4715368.1`, 87.48% identity, 0.0 E-value) and *S. tuberosum* (`KAH0668551.1`, 93.94% identity, 0.0 E-value), both of which are located on chromosome 5 of their respective genomes as expected of *H2*. Remapping of the KASP markers indicated that `tig00000244` was equivalent to `utg0004891` in the original assembly which had failed KASP marker and sanger sequencing analysis, although large gaps and polymorphisms in `utg0004891` were noted.

The contig `tig00000556` contained a single NLR which was determined to be *NRC3*. Recently, it has been observed that variants of *NRC3* can evade *G. rostochiensis* through disruption of the binding surface of nematode effector SS15. Interestingly, although the *NRC3* homolog of `tig00000556` did not have the K316 polymorphism that provides this resistance, another homolog on contig `tig00000987` was identified that did. Mutations that were exclusive to `tig00000556` were also present.

AgRenSeq identifies *H3* candidates

To identify candidates for the *H3* resistance gene, HiFi-RenSeq reads of the breeding clone 12601ab1 which is known to be duplex for *H3* were assembled using the same

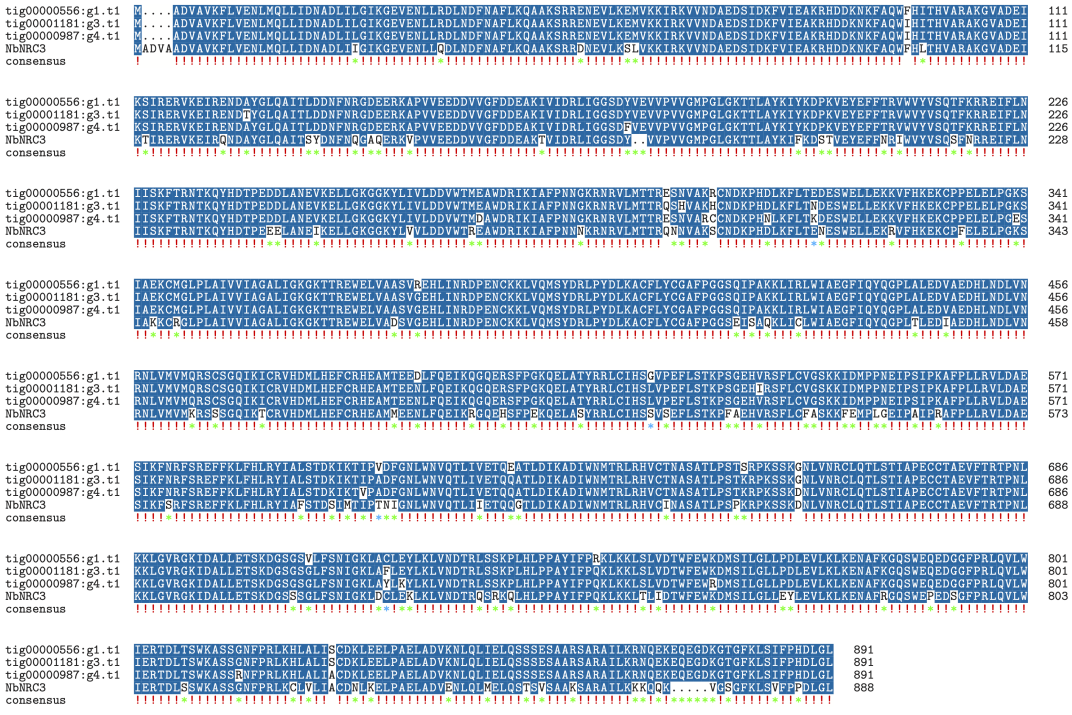


Figure 36: **NRC3 haplotypes of P55/7**. Multiple sequence alignment of the *NRC3* haplotypes and the *NRC3* homolog of *N. benthamiana*.

canu assembly strategy as used for *H2*. In total, 2790 contigs were assembled with an N50 of 10.7kbp. Of these, 1673 were identified to contain NLRs, in total containing an estimated 2364 NLR sequences. To reduce this set of candidates, an AgRenSeq approach was performed using a panel of 48 *H3* negative and 29 *H3* positive cultivars (tbl. 19). It should be noted that of the *H3* positive cultivars, 19 were selected based on their positive status for two H3 markers rather than phenotype data.

Table 19: **Table of accessions, scores, and reason for selection of the initial AgRenSeq analysis conducted for H3.** Phenotype scores are used in the AgRenSeq analysis to identify phenotype-linked *k*-mers. Status is based on internal known breeding data or marker data.

Accession	Score	Reason
Picasso	-1	Known negative
Athlete	-1	Known negative
Jelly_tp50_v5	-1	Known negative
Alouette	-1	Known negative
CPC_3534_gandarillasii	-1	Known negative
Maris_Peer_tp20_v3	-1	Known negative
Saturna_tp50_v5	-1	Known negative
Saturna_HR02_v5	-1	Known negative

Accession	Score	Reason
Rooster_tp20_v3	-1	Known negative
Nectar_tp50_v5	-1	Known negative
Cammeo	-1	Known negative
Lady_Balfour_v3	-1	Known negative
Maris_Peer_HR02_v5	-1	Known negative
Estima_tp20_v3	-1	Known negative
LaStrada	-1	Known negative
Osprey_HR02_v5	-1	Known negative
Rooster_HR02_v5	-1	Known negative
P55	-1	Known negative
Emma	-1	Known negative
Pentland_Ivory	-1	Known negative
sandra	-1	Known negative
Compass	-1	Known negative
FINGAL_12290_af_20	-1	Known negative
Inca_rosa	-1	Known negative
Jester	-1	Known negative
Juliette	-1	Known negative
Orla	-1	Known negative
Record	-1	Known negative
AUSONIA	-1	Known negative
Navan	-1	Known negative
Pixie	-1	Known negative
18_WC_1_a_23	-1	Known negative
18_WC_1_a_56	-1	Known negative
Arizona	-1	Known negative
Amora	-1	Known negative
Annabelle	-1	Known negative
Arran_Victory	-1	Known negative
Asparges	-1	Known negative
Balmoral	-1	Known negative
Ambo	-1	Known negative
Asterix	-1	Known negative
Avalanche	-1	Known negative
Bambino	-1	Known negative
Barna	-1	Known negative
Belle_de_Fontenay	-1	Known negative
Bonnie	-1	Known negative
Bounty	-1	Known negative
Almera	-1	Known negative
Buster	1	Known positive
Olympus_HR03b	1	Known positive
Midas_HR03b	1	Known positive
12601_ab1	1	Known positive
Lorimer_HR03b	1	Known positive

Accession	Score	Reason
Rocket_HR03b	1	Known positive
Paladin_HR03b	1	Known positive
Strachan_HR03b	1	Known positive
BENOL	1	Known positive
JUBILEE	1	Known positive
67_VALES_EVEREST	1	Known positive
4_18_WC_6_A_36	1	Known positive
32_12601_ab_1	1	Known positive
34_15_JHL_125_a_2	1	Dual marker positive
35_15_JHL_125_a_3	1	Dual marker positive
13_P_7_A_4	1	Dual marker positive
15_JHL_125_A_4	1	Dual marker positive
15_JHL_126_A_14	1	Dual marker positive
15_JHL_126_A_22	1	Dual marker positive
15_JHL_127_A_21	1	Dual marker positive
15_JHL_127_A_8	1	Dual marker positive
15_JHL_128_A_1	1	Dual marker positive
15_JHL_137_A_3	1	Dual marker positive
15_JHL_140_A_1	1	Dual marker positive
92_PD_20_b_1	1	Dual marker positive
92_PD_20_b_16	1	Dual marker positive
92_PD_27_C_17	1	Dual marker positive
92_PD_27_C_24	1	Dual marker positive
ROYAL	1	Known positive
CHICAGO	1	Known positive

The AgRenSeq approach reduced the number of contigs from 1673 to 23, which contained in total 32 predicted NLRs. As high levels of constitutive expression is an expected pre-requisite of canonical NLRs, the expression of each candidate was examined. Of the NLRs, 16 had no mapped RNAseq reads and five had low expression, defined here as an average read depth of less than ten. Contigs containing multiple NLRs with mixed levels of expression were noted, for example `tig00000004` contains 9 predicted NLRs but only two were supported with RNA reads. Accordingly, the list of candidates was reduced to 12 NLRs.

It was reasoned that other H3-containing cultivars likely exist, and that dRenSeq could be used to further reduce the candidate list by examining the distribution of NLRs across cultivars. As a result, a dRenSeq panel containing 1577 samples was applied to the candidates. The *H3* candidates showed variable presence across the dRenSeq panel (fig. 37). Manual inspection of the dRenSeq matrix indicated that six candidates - `tig00000601_nlr_1`, `tig00000883_nlr_1`, `tig00000004_nlr_5`, `tig00000004_nlr_4`, `tig00001378_nlr_1`, and `tig00000178_nlr_1` - could be discarded based on their presence in varieties that do not contain *H3* resistance. Other NLRs could be discarded based on their absence in varieties known to contain *H3*, including `tig00001101_nlr_1` `tig00001101_nlr_2`, and `tig00000402_nlr_1`. As a

result of this, the number of high-confidence *H3* candidates could be reduced down to nine.

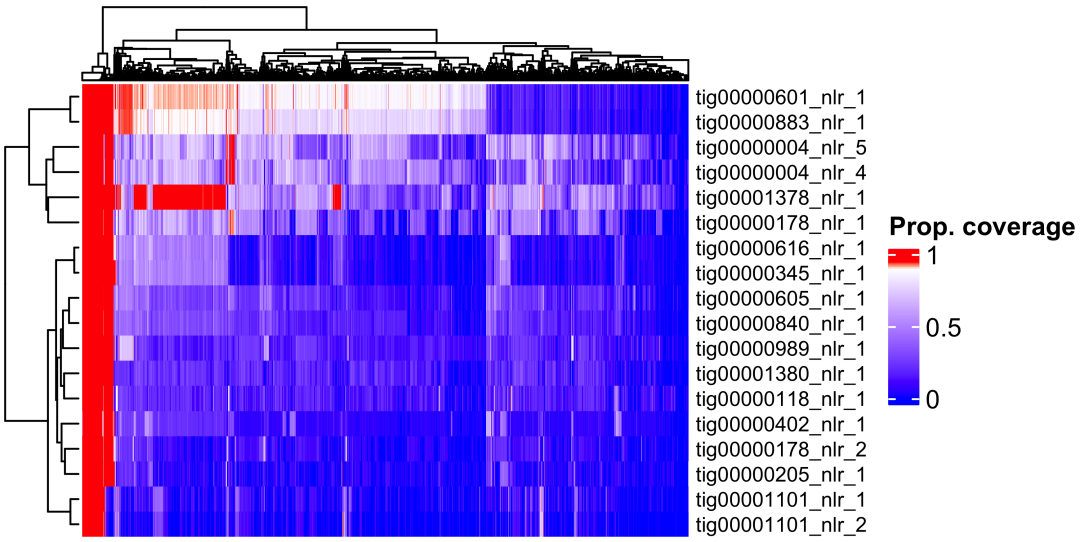


Figure 37: **dRenSeq of H3 candidates.** The dRenSeq coverage of *H3* candidates has been clustered by their distribution amongst 1577 cultivars. The colour scale for proportion of NLR covered by reads has been centred on 0.95, the threshold at which an NLR is considered to be present.

One interesting observation is that *S. stenotomum* (CIP 704369) contained complete coverage of all the *H3* candidates. Currently, the earliest known introgression of *H3* is from *S. tuberosum* ssp. *andigena*, and so this may represent an origin species for *H3* resistance. A search for candidates in the *S. stenotomum* A6-2 genome identified several genes on chromosome 4, the expected location of *H3*. Within the *Solanaceae* pangenome all candidate hits belonged to a single orthogroup. The orthogroup had undergone expansion in the tuberising *Solanaceae* genomes and all members were present on chromosome 4. This suggests that *H3* is the result of an expanding NLR locus. All *H3* candidates were closely related to the late-blight resistance NLR *R2*, a CNL (fig. 38).

Development of HISS

To enable rapid HiFi-RenSeq, AgRenSeq, and dRenSeq analysis, the High-throughput SMRT-AgRenSeq-d Snakemake (HISS) pipeline was developed (Adams et al. 2023). HISS is implemented with the workflow management system Snakemake and allows users to execute automated RenSeq analysis workflows (fig. 39). Users provide tabular data containing sample IDs, read paths, and phenotype scores for AgRenSeq as input. Software dependencies are downloaded and installed automatically to ensure reproducible analysis between users.

Subsequent to its development, the release of Snakemake version 8 broke support for the Slurm HPC workflow manager which HISS was developed on. As a result, a second version of the HISS pipeline was developed which used Nextflow rather than

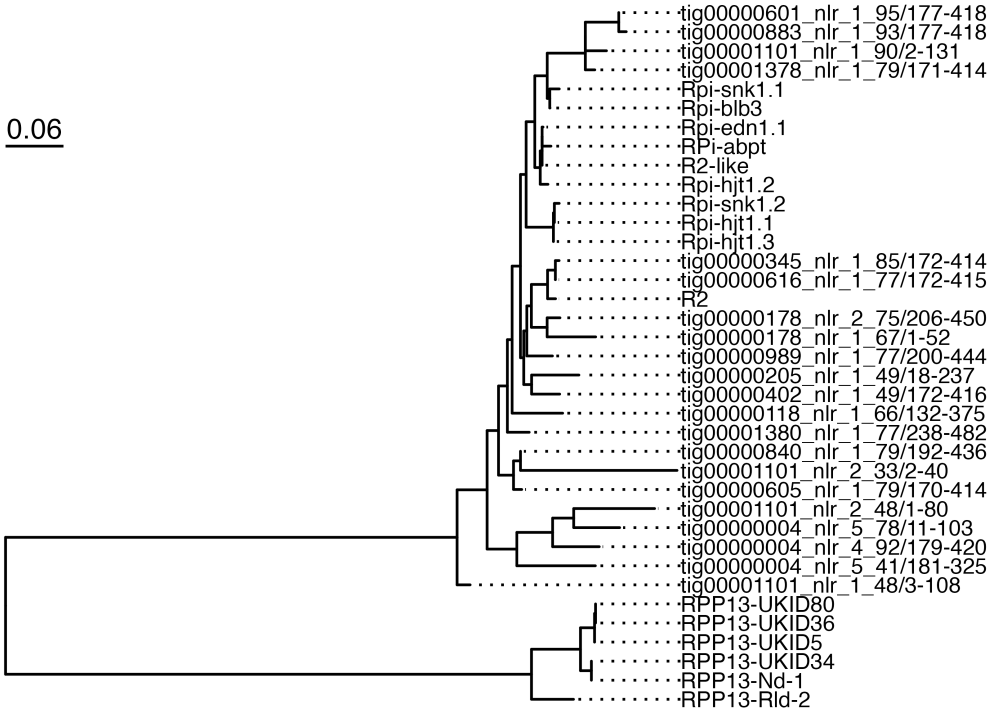


Figure 38: **Phylogeny of *H3* candidates**. Phylogenetic tree of all curated *H3* candidate NB-ARC domains identified in this study aligned to the RefPlantNLR database. Candidates are indicated by their "tig" nomenclature.

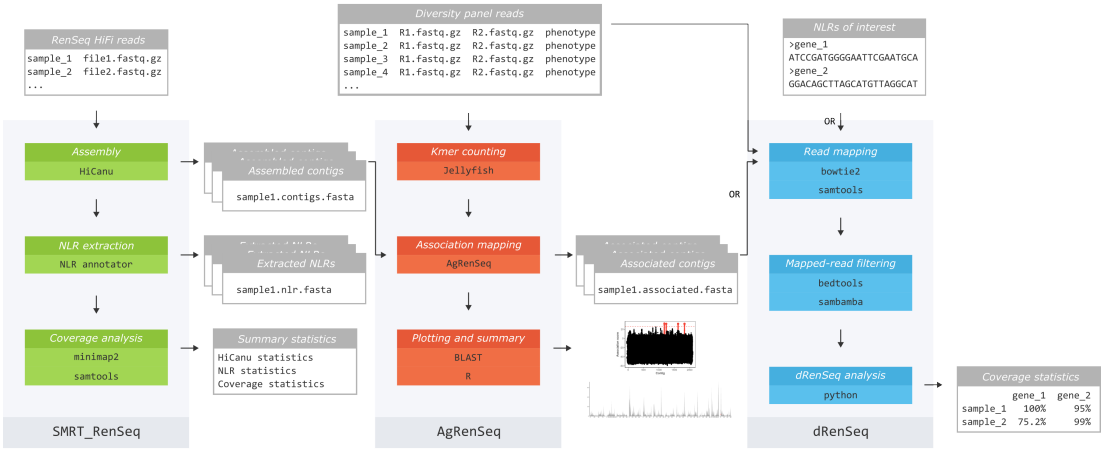


Figure 39: **An overview of the HISS pipeline.** SMRT-RenSeq (green) assembles RenSeq HiFi reads and produces assembly and NLR summary statistics. AgRenSeq (red) takes a metadata file of diversity panel reads and can use the output of SMRT-RenSeq as a reference for k-mer mapping. It outputs highly associated contigs and NLR loci as well as *k*-mer scoring plots and mapping of contigs to a reference genome. dRenSeq (blue) can use a list of NLRs of interest or the output from AgRenSeq to calculate read coverage.

Snakemake (available at https://github.com/swiftseal/nfhiss). This reimplementation offers several advantages over the original HISS workflow - software dependencies are handled by a single docker container rather than individual conda environments improving performance and reliability; the dRenSeq workflow carries out variant calling and filtering for the purpose of KASP marker design; the pipeline can be executed with a single Nextflow command which automatically downloads the latest version of HISS.

Discussion

Strong candidates for *H2* and *H3* resistance are identified

The data presented here suggests that *H2* resistance is the result of interactions with *NRC3* or is linked with another NLR in close proximity to *NRC3*. It is interesting to note that although the *NRC3* homolog linked with resistance identified here did not contain the known variant which conveys resistance to *G. rostochiensis*, another homolog which does contain this variant also exists in P55/7. Further examinations of the distribution of this variant across potato varieties would be valuable in understanding the presence of this variant throughout *Solanum* - HiFi-RenSeq is a suitable platform for this.

The status of the other *H2* candidates identified in this chapter is unclear. Both the exCNL candidate identified in the original screen and the *R8* homologs identified in the updated analysis are diverged from known NLRs. That different candidates were identified between the two screens is an indicator that further investigation would be beneficial for identifying *H2*. Obtaining the complete sequence of the *H2* locus would reveal how these candidates are physically linked to *NRC3*, would yield possible non-NLR candidates, would yield candidates that do not support variants of the expected 1:4 ratio (e.g., genes undergoing presence-absence variation), and would permit the design of additional markers for fine mapping. Unfortunately the genetic material from the previous *H2* crosses is no longer available, so a repeat of the P55/7 x Picasso cross and phenotyping would be required.

The identity of *H3* resistance remains more complex. This chapter has narrowed down *H3* to nine candidate NLRs. All candidates were homologs of the *R2* NLR and belonged to a single orthogroup in the *Solanaceae* pangenome. It is interesting to note that through the dRenSeq analysis, *S. stenotomum* CIP 704369 was identified to have the full complement of *H3* candidates. The most distant known origin of *H3* is *S. tuberosum* ssp. *andigena* (tbl. 10), and it is possible that this observation points closer to the true origin of *H3* resistance. *S. stenotomum* is believed to be a primitive origin of modern-day cultivated *S. tuberosum* ssp. *andigena* (Yan et al. 2021). Assembling the genome of *S. stenotomum* CIP 704369 or another *H3* positive cultivar would allow full resolution of the *H3* locus for the purpose of fine mapping of the true *H3* gene.

NLR discovery through RenSeq

As demonstrated in this chapter, RenSeq and its derivates are an effective means for identifying and characterising NLR-based disease resistance in plant genomes. Through HiFi-RenSeq, the inventory of NLRs within a genome can be captured; through AgRenSeq, NLRs that are strongly associated with a particular resistance phenotype can be identified; through dRenSeq, the presence and absence of known NLRs in a cultivar can be revealed. The development of high-throughput workflows such as the `HISS` platform developed here enable rapid and reproducible analysis that is accessible to users with less bioinformatics experience. Subsequent to its development, `HISS` has been used to identify candidates for the *G. pallida* resistance gene *Gpa5* and is currently being used to identify *H1*, *Sen1*, *Rpi-smira1*, *Rpi-R4* (Yuhan Wang et al. 2023).

Whilst this study demonstrates that enrichment sequencing methods are a valid strategy for identifying resistance genes, it also highlights some of their drawbacks. As a targeted sequence approach, the success of the identification of *H2* and *H3* relies on resistance being directly linked to the sequence of NLRs. Resistance that is mediated through *cis*-regulatory elements that affect NLR activity, through multiple interacting NLRs, or through non-NLR mechanisms are very likely to be missed by RenSeq alone. Indeed, the genes controlling two different resistances against *Heterodera glycines* - *rhg-1* and *rhg-4* - have been shown to be based on non-NLR mechanisms (Cook et al. 2012; Bayless et al. 2016). In addition, it must also be considered that HiFi-RenSeq may not resolve the full coding sequence on an NLR - the *P. infestans* resistance gene *Rpi-amr1* has short 3' exons with >2kbp introns separating them from the remaining sequence (Witek et al. 2021). Care must also be taken to avoid potential haplotype collapsed assemblies which can result in candidates that represent a mosaic of haplotypes.

It has recently been demonstrated that Nanopore adaptive sampling can be used to selectively sequence NLRs without the need for RenSeq enrichment (Belinchon-Moreno et al. 2023). Although its performance in capturing the full inventory of NLRs is variable, it could be reasoned that adaptive sampling of a known resistance locus could be used to cost-effectively assemble the full locus in resistance gene containing cultivars. Such a strategy might be suitable for *H3*, where the approximate locus of *H3* in previously assembled *S. stenotomum* genomes could be used as the target sequence for adaptive sampling. However, challenges in assembling tetraploid genomes may render this ineffective.

Candidate validation is a bottleneck

Probably the largest bottleneck in the resistant phenotype to cloned resistance gene pipeline is validating candidate genes. In other disease systems such as *P. infestans*, validating a single resistance gene is a relatively simple process of transiently expressing the candidate in leaf tissue - normally *N. benthamiana* - and inoculating the tissue with the appropriate strain of pathogen or co-infiltration with the cognate effector, if known. Multiple candidates can be screened simultaneously through this method, allowing high-throughput validation of candidates. For more complex traits, falling back on knockdown or knockout methodologies such as VIGS, RNAi, or Crispr-Cas9

might be necessary.

For PCN, this process is more laborious. The soil bound nature of PCN and their complex parasitism necessitates validation in root tissue and in a species that is both attractive to the nematode and that can support their infection. As a result, transformation of candidate genes into susceptible cultivars or knockout of the candidate gene in its cultivar of origin is necessary. Both methods have their drawbacks - transgenesis through *Agrobacterium* often result in off-target modifications to the genome; RNAi can be hard to deliver stably to root tissue, and failing to correctly phenotype a true candidate is in some ways worse than a false positive. Both methods require a significant time investment and rely on skilled labour. Future developments that offer rapid and high-throughput introduction of candidate NLRs into susceptible potato tissue or deletion of candidate NLRs in resistant lines will enable a greater rate of NLR identification. Recent advances in CRISPR loss-of-function experiments in *Solanum* has enabled rapid and efficient gene characterisation in a more diverse set of *Solanum* genetic backgrounds which may prove useful here (Satterlee et al. 2024).

Future strategies

Given the value of assembling the source genome for the purpose of identifying resistance genes, and the tetraploid status of potato, developing strategies that permit tetraploid genome assembly is an important step that requires further investigation. Many challenges exist in assembling polyploid genomes. Haplotypes are frequently collapsed even in diploid assemblies, and in tetraploids assigning a genetic locus to four independent haplotypes is challenging without additional evidence. Compounding this, homologous locus frequently result in switching errors, where contigs become mosaics of different haplotypes, as the intermediate homologous loci cannot be fully resolved. Given that SMRT-RenSeq cannot take advantage of these strategies, the user is limited to a decision of whether to be more permissive of mismatches during contig assembly, leading to collapses, or implementing strict parameters, leading to over-splitting of haplotypes and the potential inclusion of sequencing errors.

An effective strategy of resolving tetraploid genomes has been to utilise sequence data from the offspring of the cultivar of interest. A hallmark demonstration of this is the assembly of the autotetraploid potato cultivar *Altus* (Serra Mari et al. 2024). In this approach, a `hifiasm` assembly *Altus* was generated, and *k*-mers from assembled unitigs. Given that unitgs represent uncollapsed (but also, higher error rate) haplotypes, the unique *k*-mers could be used to estimate the *dosage* (i.e., coverage) of each unitig in the genome, thus calculating the number of haplotypes it represents. This offers a distinct advantage over other approaches that work from *contigs* - which will be collapsed in a tetraploid - and unzipping them into distinct haplotypes *post hoc*. The distribution of unique *k*-mers was examined in short-read sequencing data of a panel of 193 offspring, which could then be used to infer the inheritance of haplotypes in the offspring. From this, unitigs could be clustered into distinct haplotype clusters, permitting haplotype-resolved chromosome assembly.

The primary advantage of this approach is that it is relatively simple - low-depth sequencing data of progeny is needed, but this is often already routinely generated

during breeding programmes. A typical strategy towards identifying a novel PCN resistance gene is to cross a resistant cultivar with a susceptible cultivar, determine the distribution of the phenotype in the offspring, then sequence the offspring to identify informative variants for the purpose of mapping the resistance trait. Here, sequencing of the progeny could be extended towards producing a haplotype-phased assembly of the resistant parent. This approach would exhibit several advantages over the currently used methods.

First, it would provide a haplotype phased assembly of the resistant parent. This would be an invaluable resource in mapping, particularly for complex NLR loci which exhibit considerable haplotype diversity and copy number variation. As demonstrated in the *S. verrucosum* assembly chapter, remapping of RenSeq data to the genome of origin rather than a reference produced allowed the physical distance between candidate genes to be determined, which HiFi-RenSeq cannot do. The haplotype-phased assembly would also be a useful resource beyond just resistance gene identification. Additionally, a haplotype-phased assembly of the susceptible parent could be generated with additional HiFi-sequencing.

Secondly, if the offspring were to be simultaneously phenotyped for the segregating resistant phenotype, this could be combined with the offspring to be used for trait mapping directly. Given that hundreds of progeny can be phenotyped in a PCN resistance screen, this would provide a very large sample size for trait mapping which should offer a massive increase in resolution over current approaches that rely on sequencing of pools of resistant and susceptible samples. As this would provide whole genome sequence data rather than RenSeq data, state-of-the-art ploidy-aware variant callers could be used to provide more reliable variants. By implementing a phenotype first approach, a library of phenotyped samples could be accessed if and when greater mapping resolution or further haplotype-phasing is required.

Given the consistent decrease in sequencing costs, such an approach should be suitable for implementation in the near future. Whilst it is likely that polyploid assembly will be feasible without the prerequisite of progeny sequencing or trio-binning, the need for phenotype data will remain. This will be particularly true of traits, such as *H2*, which have poor representation in the available genetic diversity of potato, and so will require crosses and phenotype segregation to be identified.

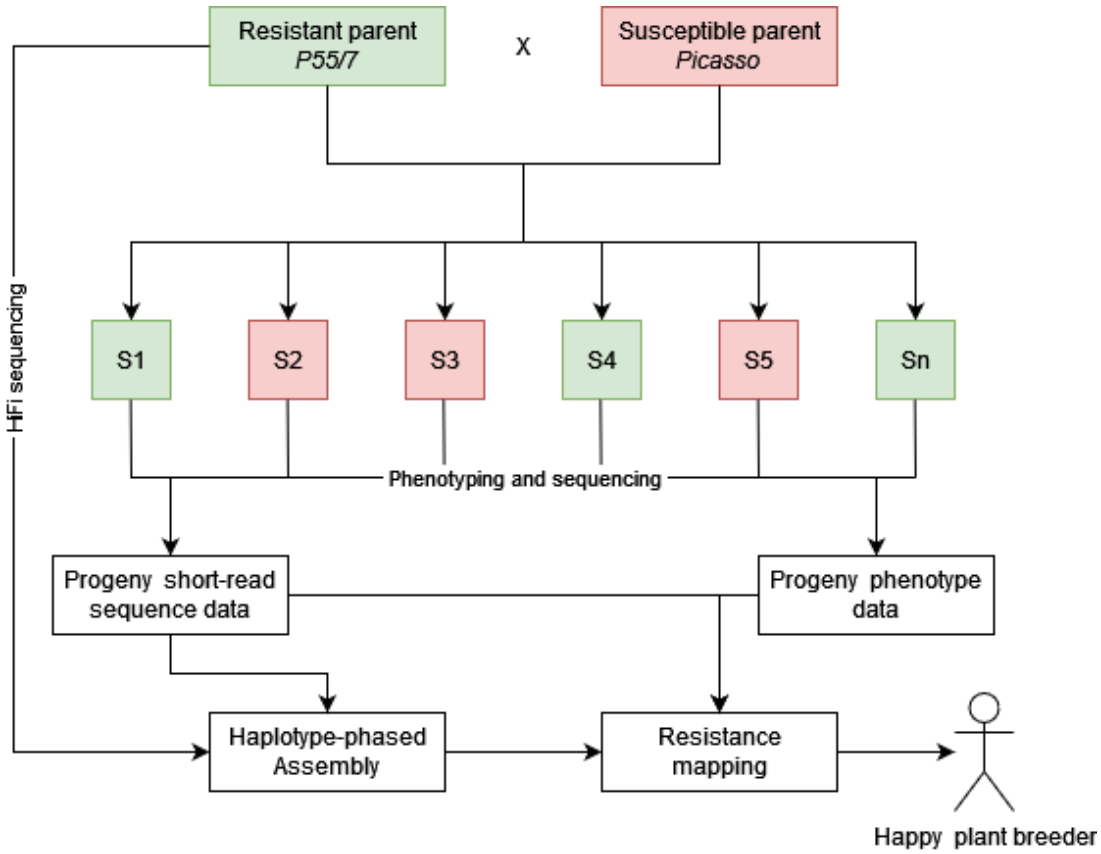


Figure 40: **Haplotype-resolved resistance mapping.** Sequence data of progeny used for producing a haplotype-resolved parental genome can be simultaneously used to map disease resistance genes based on the progeny phenotype.

Discussion

The primary aim of this thesis was to explore the applications of next-generation sequencing, principally towards the discovery of novel disease resistance genes. To achieve this, three experimental aims were explored:

- Develop tools that take advantage of recent advances in next-generation sequencing and genomics to rapidly and accurately identify NLRs from potato genomes, which should be easy-to-use and broad in their applicability so that they may be used by researchers with other crops.
- Resolve the genome of the wild potato *Solanum verrucosum* accession 54 through next-generation sequencing with the aim of identifying a previously uncharacterised late blight resistance gene, and also to explore how next-generation sequencing and *Solanum* genomics can enhance our understanding of NLR diversity as well as other genetic features such as centromeres.
- Apply both next-generation sequencing and the development of novel tools to a broad panel of wild and cultivated potato genomes with the aim of identifying novel potato cyst nematode resistance genes.

Annotating resistance genes

The development and release of `Resistify` succeeded in producing an NLR annotation tool that is rapid, accurate, and easy-to-use. `Resistify` is more performant than other available tools in identifying canonical NLRs in plant genomes, and also provides the option for very sensitive motif searches where the identification of fragmented NLRs may be necessary. Since it's release, `Resistify` has had good engagement with users at the James Hutton Institute, as well as users in Europe and China via communications on GitHub. At the time of writing, `Resistify` has been downloaded over 1,600 times through Bioconda, and the Biocontainers release has also begun to see good engagement. It will be interesting to see where `Resistify` emerges as a helpful tool for other researchers, and I intend to support or even extend `Resistify` with additional functionality while it continues to prove useful to users.

`Resistify`'s easy-of-use and strong performance in screening whole genomes should make it well suited for future projects that seek to establish 'pan-NLRomes' that fully capture the diversity of NLRs in a given species (Barragan and Weigel 2021). Such strategies have already been shown to be complementary to projects seeking to

identify novel NLRs, as demonstrated in a recent *S. americanum* pan-NLRome (Lin et al. 2023). As genomes that have been historically less represented in sequencing efforts continue to emerge, observations of NLR diversity have led to surprising discoveries such as non-canonical CNL families in wild tomato, massive losses of NLR subfamilies in Magnoliids, and large concentrations of NLRs on single chromosomes in Conifers (Wu, Xue, and Van de Peer 2021; Woudstra et al. 2024; Seong et al. 2020). Such observations are both interesting in their own right and are informative of the evolutionary mechanisms that underpin the diversification of NLRs - `Resistify` should provide an accessible interface for future discoveries in these areas.

Other gene families are also associated with disease resistance, and in future it would be valuable to include these into `Resistify`. In particular, identifying and characterising RLK/RLPs should be a particular strength of `Resistify` given its use of the `NLRexpress` LRR motif predictors which can be applied to both NLR and non-NLR LRR domains. Inclusion of these gene families will not cause a substantial decrease in speed, and would be valuable in characterising resistance loci which may not necessarily be the result of NLRs. One necessary layer of evidence to include would be transmembrane domain annotations. Since the development of `RGAugury`, improved prediction programs have been released such as `DeepTMHMM`, but to-date none provide an option to be installed and run locally (Hallgren et al. 2022).

Currently, `Resistify` cannot be directly applied to genomic sequences to identify NLRs in the absence of gene annotations. It was decided that this would be advantageous, as gene annotation is a task better suited to tools specialised for this function. Nonetheless, for the purpose of identifying genes in situations where gene annotations are absent, developing an application for `Resistify` might be valuable. For whole genomes, simply pairing `Resistify` with *ab initio* predictors such as `Helixer` would likely be sufficient. For more fragmented assemblies as is the case for HiFi-RenSeq (`Helixer` performs poorly here), devising a strategy that can take advantage of the accuracy of `Resistify` could prove useful, possibly through interpreting ORFs to predict NLR loci.

A second development of the first aim has been the release of the `HISS` package, which provides a suite of workflows that handle large RenSeq sequencing datasets for the purpose of identifying novel resistance genes. Currently, RenSeq is the most cost-effective approach for accessing the NLR inventory of a given sample. At the James Hutton Institute, the `HISS` workflows have been helpful to this thesis and other users within the potato group. `HISS` provides an accessible and robust interface with RenSeq data, enabling users who are less comfortable with bioinformatics to engage with their data in a reproducible manner. The current development of a Nextflow reimplementation of `HISS` will improve the longevity of the platform and its accessibility.

Isolating novel NLRs

Another key aim of this thesis was to identify the causative genes of late-blight *Rpi-ver1* resistance, and the PCN resistance genes *H2* and *H3*. Two different strategies

were used to achieve this. To identify *Rpi-ver1*, the complete genome of *S. verrucosum* was assembled for the purpose of re-analysing bulk-segregant variants derived from RenSeq sequencing (X. Chen et al. 2018). In doing so, the associated locus was reduced from a 4.3 Mbp region on chromosome 9 to a 1.3 Mbp region. To identify *H2* and *H3*, HiFi-RenSeq assemblies of resistant cultivars were taken and candidates identified through bulk-segregant analysis and AgRenSeq respectively. In addition, transcriptome sequencing was conducted to remove candidates that had little-to-no expression.

A key advantage of HiFi-RenSeq is that it offers a cost-effective means of identifying the full complement of NLRs within a given sample. Since its inception, bait libraries have been optimised and long-read sequencing has seen continuous cost-improvements, making it an accessible option for identifying candidate resistance genes. Additionally, the high-accuracy of HiFi reads makes it suitable for assembling the different haplotypes of a resistance gene, an advantage for potato where the tetraploid genome is a barrier to whole-genome sequencing.

As demonstrated by the analysis of *H2* and *H3*, HiFi-RenSeq assemblies are suitable for association genetic approaches, with the key assumption that the candidate gene has a suitable variant on a single resistance gene. This does make identifying candidate genes challenging in situations where NLR functionality is more complex, such as resistance that is imparted by paired NLRs or epigenetic regulation (Xi, Cesari, and Kroj 2022; Tsuchiya and Eulgem 2013). As non-NLR genes are not represented in HiFi-RenSeq datasets, assessing their association with a resistance phenotype is not possible. As NLRs are largely represented on individual contigs in HiFi-RenSeq data, it is also not possible to determine the physical association between candidates. Given that *H3* is a homolog of *R2* and is thus likely to be part of a dense NLR cluster and that *H2* is linked to *NRC3* and possibly an additional NLR, physical mapping information would be invaluable in determining the true causative gene.

For the purpose of identifying a causative resistance gene, obtaining a full assembly of the locus is likely the gold-standard approach going forward. In the case of *Rpi-ver1*, it seems unlikely that resistance is imparted by a canonical NLR and the assembled locus will be a key asset in future fine-mapping or candidate selection experiments. Producing similar genomes for P55/7 and 12601ab1 would allow similar analysis to be conducted for *H2* and *H3* respectively. A key challenge to this is that both are tetraploids making producing a genome assembly far more complex. However, a series of recent haplotype-phased tetraploid potato genome assemblies do indicate that this is feasible and should be considered (Bao et al. 2022; Hoopes et al. 2022; F. Wang et al. 2022). Previously, such assemblies have relied upon selfing population, single-cell pollen, or parental sequencing data to correctly phase haplotypes, but more recently it has been demonstrated that similar results can be achieved through low-depth offspring sequencing (Serra Mari et al. 2024). Given that offspring sequencing is a requirement for resistance gene mapping, extending this to simultaneously provide a haplotype-phased genome should be feasible.

An alternative strategy for NLR identification that warrants further investigation in potato is whole-transcriptome sequencing which, in combination with the assumption

that functional NLRs are highly expressed, is used to identify candidates in the recent NLRseek project (Brabham et al. 2024). This would not be dissimilar to HiFi-RenSeq, but would have the benefit of providing more reliable gene annotations and would be more NLR-agnostic in its approach. Across diverse germplasm, association genetics techniques such as those employed by AgRenSeq may also be suitable here (Arora et al. 2019). Long-read transcriptome sequencing remains quite expensive for high-throughput applications, but recent advances such as the release of PacBio Revio and Kinnex sequencing should make this more accessible. Of course, as the whole transcriptome is being sequenced, such data would be invaluable for studies on potato diversity beyond just resistance genes.

Even as next-generation sequencing becomes increasingly accessible and expedited NLR discovery becomes the norm, the challenges associated with PCN resistance phenotyping will continue to be a bottleneck in efforts to reduce its impact in Scotland. Currently, transforming candidate genes into a susceptible background is the most common method for assessing their function - recent advances in *Solanum* CRISPR methods may make validating resistance genes in their genome of origin a more efficient option (Satterlee et al. 2024).

How are NLRs and transposable elements linked?

It is becoming increasingly apparent that NLR diversity is directly influenced by the activity of transposable elements. NLRs exhibiting high allelic diversity in *Arabidopsis* are in close proximity with transposable elements, an LTR insertion in the NLR *RPP7* was domesticated to function as a regulator of expression, and the NLR inventory of *Capsicum* appears to have been greatly expanded by retroduplication (Sutherland et al. 2024; Tsuchiya and Eulgem 2013; Kim et al. 2017).

An unexpected finding that emerged during the development of `Resistify` is a possible relationship between NLRs and *Helitron* transposable elements in tuber-bearing *Solanum* genomes. Given that *Helitrons* have very few structural motifs, they are challenging to determine from sequence alone. Further validation would be necessary to determine if this is a genuine mechanism for NLR diversification. That autonomous *Helitron* elements were readily identifiable within the *S. verrucosum* genome, and that only the strictest definition of *Helitrons* were considered for those that were NLR-associated, does make it feasible that they at least express a degree of non-autonomous activity.

Transposable element activity results in genetic instability which can result in the production of new resistance genes, but their insertion can also impact the transcription of nearby NLRs as evidenced for *RPP7* and in *Arabidopsis* (Sutherland et al. 2024; Tsuchiya and Eulgem 2013). A recent survey of NLR clusters in *Arabidopsis* also demonstrated that they are the sites of frequent - and young - transposable element insertion events (Teasdale et al. 2024). The impact of insertions on the methylation, and subsequent transcription of NLRs is yet to be fully evaluated. In the NLR cluster analysis in *Arabidopsis*, no difference was noted between NLRs inside and outside of NLR clusters, but methylation was only evaluated in the CG context and attention was

not paid to the variability of NLRs within individual clusters. From a recent analysis in Maize and in this thesis, it is apparent that the combination of methylation in the CG and CHG contexts is a more reliable indicator of gene expression, and studies in soybean demonstrate that NLR expression within clusters is more of a mosaic than a broad on/off system (Prigozhin et al. 2024; W. Wang et al. 2021).

Given that NLR clusters are the site of many functional NLRs, a comprehensive assessment of clusters in potato would be valuable. Such an analysis was beyond the scope of this thesis, but as demonstrated in Chapter 1 there are a sufficient number of publicly available chromosome-scale *Solanum* genomes to begin this, and the assembly of additional genomes will improve the quality of this analysis further. Also demonstrated in Chapter 1 is that particular attention should be paid to the quality of NLR gene annotations, which are frequently missed even by performant tools such as BRAKER3. The *S. verrucosum* genome demonstrates that ONT sequencing is a suitable method for obtaining high-quality and informative methylation data 'for free', and its inclusion into future studies will be valuable for assessing the epigenetic diversity of NLR clusters, which is yet to be fully understood.

Regardless of their association with NLRs, curation of transposable elements within *Solanum* would be of value. As demonstrated in Chapter 2, tools such as `Earlgrey` are invaluable in providing high-quality libraries of transposable elements, and secondary analysis with tools such as `TEsorter` are useful for further classification (Baril, Galbraith, and Hayward 2024; R.-G. Zhang et al. 2022). Without curation, annotation relies on *de novo* methods or homology comparisons from distantly related plant species which makes comparisons between *Solanum* genomes challenging. It is possible that ONT-derived methylation data could also be used as an additional layer of evidence for filtering out low-confidence annotations and determining transposable element boundaries.

S. verrucosum as a bridge species

S. verrucosum is useful in breeding programs for its function as a bridge species, and the genetic basis for this condition is yet to be resolved. Previously, this has been attributed to a lack of a functional *S-RNase* gene - a key determinant of self-compatibility (W. L. Behling and Douches 2023). Here, it appears that a functional S-RNase is present in *S. verrucosum*, which may rule out at least one hypothesis for the cause of this trait. The expression of this gene deserves investigation, which may possibly be impacted by transposable element activity in its promoter that is common in *Solanum*. It is clear that other genetic systems such as *HT* also serve a role in *Solanum* self-compatibility, evaluating the status of these will also be important in determining the basis of this important trait (S. Lee et al. 2023.)

Solanum centromeres

A fortuitous insight from the *S. verrucosum* genome is the unexpected nature of its centromeres. Some centromeres are repeatless and the site of frequent transposable

element activity, whereas others are formed of large repeat arrays with massive subunits that appear to *sometimes* be derived from transposable elements, although through what mechanism is not clear. The form of the centromeres in *S. verrucosum* is somewhat at odds with our understanding of the character and mechanism of centromere formation, and highlights the diversity of these essential structures across the plant kingdom. Analysis on the transposable elements and methylation status of *S. verrucosum* offers an early but incomplete insight into their function.

Much of our recent understanding on the organisation of plant centromeres has been established in *Arabidopsis*, where they are composed of short tandem repeats arranged into higher-order structures, are genetically diverse between individuals, and experience rapid cycles of transposon invasion and purging that result in diversification (Naish et al. 2021; Wlodzimierz et al. 2023). Long-read sequencing has proved particularly useful in these investigations, as reads are now of sufficient length and quality to bridge complex repeats which were otherwise collapsed in previous short read approaches. As our understanding of centromere structure and function advances through studies in *Arabidopsis*, further investigations on the organisation of *Solanum* centromeres would be beneficial to broaden our understanding. Of greatest interest would be an assessment on how dynamic the repetitive and repeatless centromeres are with respect to transposable element activity. Sequencing additional *S. verrucosum* accessions or conducting a transgenerational survey of the centromeres would allow this - fortunately the centromeres are relatively straightforward to assemble due to their lack of highly repetitive arrays. It is apparent that DNA methylation is not strongly altered in the repeatless or repetitive centromeric elements, but the status of other epigenetic marks is unclear. Further exploration of the distribution of histone modifications would prove valuable here.

Although we have only a preliminary understanding of centromere function in plants, centromere engineering offers exciting new avenues for plant breeding. Synthetic chromosomes established through centromere manipulation could be potent vectors of agronomic traits, and centromere repositioning could fundamentally alter the recombination strategies of breeding programs (Naish and Henderson 2024). The *S. verrucosum* genome provides the most recent insight into these systems in potato.

References

Abd-Elgawad, M. M. M., and T. H. Askary. 2015. "Impact of Phytonematodes on Agriculture Economy." *Biocontrol Agents of Phytonematodes*, CABI Books, January, 3–49. https://doi.org/10.1079/9781780643755.0003.

Abdennur, Nezar, and Leonid A Mirny. 2019. "Cooler: Scalable Storage for Hi-C Data and Other Genomically Labeled Arrays." *Bioinformatics* 36 (1): 311–16. https://doi.org/10.1093/bioinformatics/btz540.

Adachi, Hiroaki, Mauricio P Contreras, Adeline Harant, Chih-hang Wu, Lida Derevnina, Toshiyuki Sakai, Cian Duggan, et al. 2019. "An N-Terminal Motif in NLR Immune Receptors Is Functionally Conserved Across Distantly Related Plant Species." Edited by Jian-Min Zhou, Detlef Weigel, and Jian-Min Zhou. *eLife* 8 (November): e49956. https://doi.org/10.7554/eLife.49956.

Adams, Thomas M., Moray Smith, Yuhan Wang, Lynn H. Brown, Micha M. Bayer, and Ingo Hein. 2023. "HISS: Snakemake-Based Workflows for Performing SMRT-RenSeq Assembly, AgRenSeq and dRenSeq for the Discovery of Novel Plant Disease Resistance Genes." *BMC Bioinformatics* 24 (1): 204. https://doi.org/10.1186/s12859-023-05335-8.

Ai, Yanjun, Shenglin Jing, Zhengnan Cheng, Botao Song, Conghua Xie, Jun Liu, and Jun Zhou. 2021. "DNA Methylation Affects Photoperiodic Tuberization in Potato (Solanum Tuberosum L.) by Mediating the Expression of Genes Related to the Photoperiod and GA Pathways." *Horticulture Research* 8 (January): 181. https://doi.org/10.1038/s41438-021-00619-7.

Akakpo, Roland, Marie-Christine Carpentier, Yue Ie Hsing, and Olivier Panaud. 2020. "The Impact of Transposable Elements on the Structure, Evolution and Function of the Rice Genome." *New Phytologist* 226 (1): 44–49. https://doi.org/10.1111/nph.16356.

Ali, Jared G., Hans T. Alborn, Raquel Campos-Herrera, Fatma Kaplan, Larry W. Duncan, Cesar Rodriguez-Saona, Albrecht M. Koppenhöfer, and Lukasz L. Stelinski. 2012. "Subterranean, Herbivore-Induced Plant Volatile Increases Biological Control Activity of Multiple Beneficial Nematode Species in Distinct Habitats." *PLOS ONE* 7 (6): e38146. https://doi.org/10.1371/journal.pone.0038146.

Ali, Muhammad A., Farrukh Azeem, Hongjie Li, and Holger Bohlmann. 2017. "Smart Parasitic Nematodes Use Multifaceted Strategies to Parasitize Plants." *Frontiers in Plant Science* 8 (October). https://doi.org/10.3389/fpls.2017.01699.

Altemose, Nicolas, Glennis A. Logsdon, Andrey V. Bzikadze, Pragya Sidhwani, Sasha A. Langley, Gina V. Caldas, Savannah J. Hoyt, et al. 2022. "Complete Genomic and Epigenetic Maps of Human Centromeres." *Science* 376 (6588): eabl4178. https:

//doi.org/10.1126/science.abl4178.

Ames, Mercedes, Andy Hamernik, William Behling, David S. Douches, Dennis A. Halterman, and Paul C. Bethke. 2024. "A Survey of the Sli Gene in Wild and Cultivated Potato." *Plant Direct* 8 (5): e589. https://doi.org/10.1002/pld3.589.

Armstrong, Miles R., Jack Vossen, Tze Yin Lim, Ronald C. B. Hutten, Jianfei Xu, Shona M. Strachan, Brian Harrower, Nicolas Champouret, Eleanor M. Gilroy, and Ingo Hein. 2019. "Tracking Disease Resistance Deployment in Potato Breeding by Enrichment Sequencing." *Plant Biotechnology Journal* 17 (2): 540–49. https://doi.org/10.1111/pbi.12997.

Arora, Sanu, Burkhard Steuernagel, Kumar Gaurav, Sutha Chandramohan, Yunming Long, Oadi Matny, Ryan Johnson, et al. 2019. "Resistance Gene Cloning from a Wild Crop Relative by Sequence Capture and Association Genetics." *Nature Biotechnology* 37 (2): 139–43. https://doi.org/10.1038/s41587-018-0007-9.

Aury, Jean-Marc, Stefan Engelen, Benjamin Istace, Cécile Monat, Pauline Lasserre-Zuber, Caroline Belser, Corinne Cruaud, et al. 2022. "Long-Read and Chromosome-Scale Assembly of the Hexaploid Wheat Genome Achieves High Resolution for Research and Breeding." *GigaScience* 11 (January): giac034. https://doi.org/10.1093/gigascience/giac034.

Back, M. A., L. Cortada, I. G. Grove, and V. Taylor. 2018. "Field Management and Control Strategies." *Cyst Nematodes*, CABI Books, January, 305–36. https://doi.org/10.1079/9781786390837.0305.

Baduel, Pierre, and Vincent Colot. 2021. "The Epiallelic Potential of Transposable Elements and Its Evolutionary Significance in Plants." *Philosophical Transactions of the Royal Society B: Biological Sciences* 376 (1826): 20200123. https://doi.org/10.1098/rstb.2020.0123.

Bankevich, Anton, Andrey V. Bzikadze, Mikhail Kolmogorov, Dmitry Antipov, and Pavel A. Pevzner. 2022. "Multiplex de Bruijn Graphs Enable Genome Assembly from Long, High-Fidelity Reads." *Nature Biotechnology* 40 (7): 1075–81. https://doi.org/10.1038/s41587-022-01220-6.

Bao, Zhigui, Canhui Li, Guangcun Li, Pei Wang, Zhen Peng, Lin Cheng, Hongbo Li, et al. 2022. "Genome Architecture and Tetrasomic Inheritance of Autotetraploid Potato." *Molecular Plant* 15 (7): 1211–26. https://doi.org/10.1016/j.molp.2022.06.009.

Baril, Tobias, James Galbraith, and Alex Hayward. 2024. "Earl Grey: A Fully Automated User-Friendly Transposable Element Annotation and Analysis Pipeline." *Molecular Biology and Evolution* 41 (4): msae068. https://doi.org/10.1093/molbev/msae068.

Barragan, A Cristina, and Detlef Weigel. 2021. "Plant NLR Diversity: The Known Unknowns of Pan-NLRomes." *The Plant Cell* 33 (4): 814–31. https://doi.org/10.1093/plcell/koaa002.

Bayer, Philipp E., David Edwards, and Jacqueline Batley. 2018. "Bias in Resistance Gene Prediction Due to Repeat Masking." *Nature Plants* 4 (10): 762–65. https://doi.org/10.1038/s41477-018-0264-0.

Bayless, Adam M., John M. Smith, Junqi Song, Patrick H. McMinn, Alice Teillet, Benjamin K. August, and Andrew F. Bent. 2016. "Disease Resistance Through Impairment of α-SNAP–NSF Interaction and Vesicular Trafficking by Soybean Rhg1." *Proceedings of the National Academy of Sciences* 113 (47): E7375–82. https://doi.org/10.1073/pnas.1610150113.

Beals, Katherine A. 2019. "Potatoes, Nutrition and Health." *American Journal of Potato Research* 96 (2): 102–10. https://doi.org/10.1007/s12230-018-09705-4.

Behling, William L., and David S. Douches. 2023. "The Effect of Self-Compatibility Factors on Interspecific Compatibility in Solanum Section Petota." *Plants* 12 (8): 1709. https://doi.org/10.3390/plants12081709.

Behling, William, Joseph Coombs, Paul Collins, and David Douches. 2024. "An Analysis of Inter-Endosperm Balance Number Crosses with the Wild Potato Solanum Verrucosum." *American Journal of Potato Research* 101 (1): 34–44. https://doi.org/10.1007/s12230-023-09937-z.

Belinchon-Moreno, Javier, Aurelie Berard, Aurelie Canaguier, Véronique Chovelon, Corinne Cruaud, Stéfan Engelen, Rafael Feriche-Linares, et al. 2023. "Nanopore Adaptive Sampling to Identify the NLR-Gene Family in Melon (Cucumis Melo L.)." bioRxiv. https://doi.org/10.1101/2023.12.20.572599.

Bernoux, Maud, Thomas Ve, Simon Williams, Christopher Warren, Danny Hatters, Eugene Valkov, Xiaoxiao Zhang, Jeffrey G. Ellis, Bostjan Kobe, and Peter N. Dodds. 2011. "Structural and Functional Analysis of a Plant Resistance Protein TIR Domain Reveals Interfaces for Self-Association, Signaling, and Autoregulation." *Cell Host & Microbe* 9 (3): 200–211. https://doi.org/10.1016/j.chom.2011.02.009.

Bert, Wim, Gerrit Karssen, and Johannes Helder. 2011. "Phylogeny and Evolution of Nematodes." In *Genomics and Molecular Genetics of Plant-Nematode Interactions*, edited by John T. Jones, Godelieve Gheysen, and Carmen Fenoll, 45–59. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0434-3_3.

Bewick, Adam J., Lexiang Ji, Chad E. Niederhuth, Eva-Maria Willing, Brigitte T. Hofmeister, Xiuling Shi, Li Wang, et al. 2016. "On the Origin and Evolutionary Consequences of Gene Body DNA Methylation." *Proceedings of the National Academy of Sciences* 113 (32): 9111–16. https://doi.org/10.1073/pnas.1604666113.

Bigeard, Jean, Jean Colcombet, and Heribert Hirt. 2015. "Signaling Mechanisms in Pattern-Triggered Immunity (PTI)." *Molecular Plant* 8 (4): 521–39. https://doi.org/10.1016/j.molp.2014.12.022.

Blok, Vivian C., and Mark S. Phillips. 2012. "Biological Characterisation of Globodera Pallida from Idaho." *Nematology* 14 (7): 817–26. https://doi.org/10.1163/156854112X627336.

Blok, Vivian C., Jon Pickup, Kim Davie, Helen Kettle, David Ewing, Adrian Roberts, Laure Kuhfuss, Adam Kleczkowski, and Beth McDougall. 2020. "The Future Threat of PCN in Scotland." PHC2018/16. Scotland's Centre of Expertise for Plant Health. 10.5281/zenodo.3889965.

Bohlmann, Holger, and Miroslaw Sobczak. 2014. "The Plant Cell Wall in the Feeding Sites of Cyst Nematodes." *Frontiers in Plant Science* 5 (March). https://doi.org/10.3389/fpls.2014.00089.

Bozan, Ilayda, Sai Reddy Achakkagari, Noelle L. Anglin, David Ellis, Helen H. Tai, and Martina V. Strömvik. 2023. "Pangenome Analyses Reveal Impact of Transposable Elements and Ploidy on the Evolution of Potato Species." *Proceedings of the National Academy of Sciences* 120 (31): e2211117120. https://doi.org/10.1073/pnas.2211117120.

Brabham, Helen J., Inmaculada Hernández-Pinzón, Chizu Yanagihara, Noriko Ishikawa, Toshiyuki Komori, Oadi N. Matny, Amelia Hubbard, et al. 2024. "Discovery of

Functional NLRs Using Expression Level, High-Throughput Transformation, and Large-Scale Phenotyping." bioRxiv. https://doi.org/10.1101/2024.06.25.599845.

Bradshaw, John E., and Gavin Ramsay. 2005. "Utilisation of the Commonwealth Potato Collection in Potato Breeding." *Euphytica* 146 (1): 9–19. https://doi.org/10.1007/s10681-005-3881-4.

Broz, Amanda K., Christopher M. Miller, You Soon Baek, Alejandro Tovar-Méndez, Pablo Geovanny Acosta-Quezada, Tanya Elizabet Riofrío-Cuenca, Douglas B. Rusch, and Patricia A. Bedinger. 2021. "S-RNase Alleles Associated With Self-Compatibility in the Tomato Clade: Structure, Origins, and Expression Plasticity." *Frontiers in Genetics* 12 (December). https://doi.org/10.3389/fgene.2021.780793.

Bubolz, Jéssica, Patrycja Sleboda, Anna Lehrman, Sven-Ove Hansson, Carl Johan Lagerkvist, Björn Andersson, Marit Lenman, et al. 2022. "Genetically Modified (GM) Late Blight-Resistant Potato and Consumer Attitudes Before and After a Field Visit." *GM Crops & Food* 13 (1): 290–98. https://doi.org/10.1080/21645698.2022.2133396.

Calle García, Joan, Anna Guadagno, Andreu Paytuvi-Gallart, Alfonso Saera-Vila, Ciro Gianmaria Amoroso, Daniela D'Esposito, Giuseppe Andolfo, Riccardo Aiese Cigliano, Walter Sanseverino, and Maria Raffaella Ercolano. 2022. "PRGdb 4.0: An Updated Database Dedicated to Genes Involved in Plant Disease Resistance Process." *Nucleic Acids Research* 50 (D1): D1483–90. https://doi.org/10.1093/nar/gkab1087.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.

Carputo, D., L. Monti, J. E. Werner, and L. Frusciante. 1999. "Uses and Usefulness of Endosperm Balance Number." *Theoretical and Applied Genetics* 98 (3): 478–84. https://doi.org/10.1007/s001220051095.

Carrasco, A., J. I. Ruiz de Galarreta, A. Rico, and E. Ritter. 2000. "Transfer of PLRV Resistance fromSolanum Verrucosum Schlechdt to Potato (S. Tuberosum L.) by Protoplast Electrofusion." *Potato Research* 43 (1): 31–42. https://doi.org/10.1007/BF02358511.

Castelli, Lydia, Glenn Bryan, Vivian C. Blok, Gavin Ramsay, and Mark S. Phillips. 2005. "Life Stage Responses Observed Amongst Fifteen Wild Solanum Species Resistant to Globodera Pallida." *Nematology* 7 (5): 701–11. https://doi.org/10.1163/156854105775142955.

Cesari, Stella. 2018. "Multiple Strategies for Pathogen Perception by Plant Immune Receptors." *New Phytologist* 219 (1): 17–24. https://doi.org/10.1111/nph.14877.

Chen, Jian, Zijian Wang, Kaiwen Tan, Wei Huang, Junpeng Shi, Tong Li, Jiang Hu, et al. 2023. "A Complete Telomere-to-Telomere Assembly of the Maize Genome." *Nature Genetics* 55 (7): 1221–31. https://doi.org/10.1038/s41588-023-01419-6.

Chen, Jiongjiong, Qun Hu, Yu Zhang, Chen Lu, and Hanhui Kuang. 2014. "P-MITE: A Database for Plant Miniature Inverted-Repeat Transposable Elements." *Nucleic Acids Research* 42 (D1): D1176–81. https://doi.org/10.1093/nar/gkt1000.

Chen, Shiyan, Demosthenis Chronis, and Xiaohong Wang. 2013. "The Novel GrCEP12 Peptide from the Plant-Parasitic Nematode Globodera Rostochiensis Suppresses Flg22-Mediated PTI." *Plant Signaling & Behavior* 8 (9): e25359. https://doi.org/10.4

161/psb.25359.

Chen, Xinwei, Dominika Lewandowska, Miles R. Armstrong, Katie Baker, Tze-Yin Lim, Micha Bayer, Brian Harrower, et al. 2018. "Identification and Rapid Mapping of a Gene Conferring Broad-Spectrum Late Blight Resistance in the Diploid Potato Species Solanum Verrucosum Through DNA Capture Technologies." *Theoretical and Applied Genetics* 131 (6): 1287–97. https://doi.org/10.1007/s00122-018-3078-6.

Cheng, Chia-Yi, Vivek Krishnakumar, Agnes P. Chan, Françoise Thibaud-Nissen, Seth Schobel, and Christopher D. Town. 2017. "Araport11: A Complete Reannotation of the Arabidopsis Thaliana Reference Genome." *The Plant Journal* 89 (4): 789–804. https://doi.org/10.1111/tpj.13415.

Cheng, Haoyu, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. 2021. "Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm." *Nature Methods* 18 (2): 170–75. https://doi.org/10.1038/s41592-020-01056-5.

Chinchilla, Delphine, Cyril Zipfel, Silke Robatzek, Birgit Kemmerling, Thorsten Nürnberger, Jonathan D. G. Jones, Georg Felix, and Thomas Boller. 2007. "A Flagellin-Induced Complex of the Receptor FLS2 and BAK1 Initiates Plant Defence." *Nature* 448 (7152): 497–500. https://doi.org/10.1038/nature05999.

Chronis, Demosthenis, Shiyan Chen, Shunwen Lu, Tarek Hewezi, Sara C. D. Carpenter, Rosemary Loria, Thomas J. Baum, and Xiaohong Wang. 2013. "A Ubiquitin Carboxyl Extension Protein Secreted from a Plant-Parasitic Nematode Globodera Rostochiensis Is Cleaved in Planta to Promote Plant Parasitism." *The Plant Journal* 74 (2): 185–96. https://doi.org/10.1111/tpj.12125.

Contreras, Mauricio P., Hsuan Pai, Muniyandi Selvaraj, AmirAli Toghani, David M. Lawson, Yasin Tumtas, Cian Duggan, et al. 2023. "Resurrection of Plant Disease Resistance Proteins via Helper NLR Bioengineering." *Science Advances* 9 (18): eadg3861. https://doi.org/10.1126/sciadv.adg3861.

Contreras, Mauricio P, Hsuan Pai, Yasin Tumtas, Cian Duggan, Enoch Lok Him Yuen, Angel Vergara Cruces, Jiorgos Kourelis, et al. 2022. "Sensor NLR Immune Proteins Activate Oligomerization of Their NRC Helpers in Response to Plant Pathogens." *The EMBO Journal* 42 (5): e111519. https://doi.org/10.15252/embj.2022111519.

Cook, David E., Tong Geon Lee, Xiaoli Guo, Sara Melito, Kai Wang, Adam M. Bayless, Jianping Wang, et al. 2012. "Copy Number Variation of Multiple Genes at Rhg1 Mediates Nematode Resistance in Soybean." *Science* 338 (6111): 1206–9. https://doi.org/10.1126/science.1228746.

Cooper, W. Rodney, and John B. Bamberg. 2016. "Variation in Susceptibility to Potato Psyllid, Bactericera Cockerelli (Hemiptera: Triozidae), Among Solanum Verrucosum Germplasm Accessions." *American Journal of Potato Research* 93 (4): 386–91. https://doi.org/10.1007/s12230-016-9512-x.

Cuacos, Maria, F. Chris H. Franklin, and Stefan Heckmann. 2015. "Atypical Centromeres in Plants—What They Can Tell Us." *Frontiers in Plant Science* 6 (October). https://doi.org/10.3389/fpls.2015.00913.

Cui, Weina, Peidong Tai, Xiaojun Li, Chunyun Jia, Honghong Yuan, Lei He, and Lizong Sun. 2021. "A Reduction in Cadmium Accumulation and Sulphur Containing Compounds Resulting from Grafting in Eggplants (Solanum Melogena) Is Associated with DNA Methylation." *Plant and Soil* 468 (1): 183–96. https:

//doi.org/10.1007/s11104-021-05122-5.

Dainat, Jacques, Darío Hereñú, Dr K. D. Murray, Ed Davis, Kathryn Crouch, LucileSol, Nuno Agostinho, pascal-git, Zachary Zollman, and tayyrov. 2023. "NBISweden/AGAT: AGAT-V1.2.0." Zenodo. https://doi.org/10.5281/zenodo.8178877.

Derevnina, Lida, Mauricio P. Contreras, Hiroaki Adachi, Jessica Upson, Angel Vergara Cruces, Rongrong Xie, Jan Skłenar, et al. 2021. "Plant Pathogens Convergently Evolved to Counteract Redundant Nodes of an NLR Immune Receptor Network." *PLOS Biology* 19 (8): e3001136. https://doi.org/10.1371/journal.pbio.3001136.

Di Tommaso, Paolo, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. "Nextflow Enables Reproducible Computational Workflows." *Nature Biotechnology* 35 (4): 316–19. https://doi.org/10.1038/nbt.3820.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics (Oxford, England)* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Dupeyron, Mathilde, Kumar S. Singh, Chris Bass, and Alexander Hayward. 2019. "Evolution of Mutator Transposable Elements Across Eukaryotic Diversity." *Mobile DNA* 10 (1): 12. https://doi.org/10.1186/s13100-019-0153-8.

Durand, Neva C., Muhammad S. Shamim, Ido Machol, Suhas S. P. Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. 2016. "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments." *Cell Systems* 3 (1): 95–98. https://doi.org/10.1016/j.cels.2016.07.002.

Eddy, Sean R. 2009. "A New Generation of Homology Search Tools Based on Probabilistic Inference." In *Genome Informatics 2009*, 205–11. PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO. https://doi.org/10.1142/9781848165632_0019.

Eggers, Ernst-Jan, Ate van der Burgt, Sjaak A. W. van Heusden, Michiel E. de Vries, Richard G. F. Visser, Christian W. B. Bachem, and Pim Lindhout. 2021. "Neofunctionalisation of the Sli Gene Leads to Self-Compatibility and Facilitates Precision Breeding in Potato." *Nature Communications* 12 (1): 4141. https://doi.org/10.1038/s41467-021-24267-6.

Eijlander, Ronald, Wendy ter Laak, Jan G.Th. Hermsen, M. S. Ramanna, and Evert Jacobsen. 2000. "Occurrence of Self-Compatibility, Self-Incompatibility and Unilateral Incompatibility After Crossing Diploid S. Tuberosum (SI) with S. Verrucosum (SC): I. Expression and Inheritance of Self-Compatibility." *Euphytica* 115 (2): 127–39. https://doi.org/10.1023/A:1003902907599.

Emms, David M., and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 1–14. https://doi.org/10.1186/s13059-019-1832-y.

Enciso-Rodriguez, Felix, Norma C. Manrique-Carpintero, Satya Swathi Nadakuduti, C. Robin Buell, Daniel Zarka, and David Douches. 2019. "Overcoming Self-Incompatibility in Diploid Potato Using CRISPR-Cas9." *Frontiers in Plant Science* 10 (April). https://doi.org/10.3389/fpls.2019.00376.

Ernst, Karin, Amar Kumar, Doris Kriseleit, Dorothee-U. Kloos, Mark S. Phillips, and Martin W. Ganal. 2002. "The Broad-Spectrum Potato Cyst Nematode Resistance

Gene (Hero) from Tomato Is the Only Member of a Large Gene Family of NBS-LRR Genes with an Unusual Amino Acid Repeat in the LRR Region." *The Plant Journal* 31 (2): 127–36. https://doi.org/10.1046/j.1365-313X.2002.01341.x.

Esch, Lara, and Ulrich Schaffrath. 2017. "An Update on Jacalin-Like Lectins and Their Role in Plant Defense." *International Journal of Molecular Sciences* 18 (7): 1592. https://doi.org/10.3390/ijms18071592.

Ewels, Philip A., Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2020. "The Nf-Core Framework for Community-Curated Bioinformatics Pipelines." *Nature Biotechnology* 38 (3): 276–78. https://doi.org/10.1038/s41587-020-0439-x.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report." *Bioinformatics (Oxford, England)* 32 (19): 3047–48. https://doi.org/10.1093/bioinformatics/btw354.

FAOSTAT. 2023. "Crops and Livestock Products." https://www.fao.org/faostat/en/#data/QCL/metadata.

Feehan, Joanna M, Baptiste Castel, Adam R Bentham, and Jonathan DG Jones. 2020. "Plant NLRs Get by with a Little Help from Their Friends." *Current Opinion in Plant Biology*, Biotic interactions ● AGRI 2019, 56 (August): 99–108. https://doi.org/10.1016/j.pbi.2020.04.006.

Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. 2020. "RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families." *Proceedings of the National Academy of Sciences* 117 (17): 9451–57. https://doi.org/10.1073/pnas.1921046117.

Formenti, Giulio, Arang Rhie, Brian P. Walenz, Françoise Thibaud-Nissen, Kishwar Shafin, Sergey Koren, Eugene W. Myers, Erich D. Jarvis, and Adam M. Phillippy. 2022. "Merfin: Improved Variant Filtering, Assembly Evaluation and Polishing via k-Mer Validation." *Nature Methods* 19 (6): 696–704. https://doi.org/10.1038/s41592-022-01445-y.

Fu, Yilei, Sergey Aganezov, Medhat Mahmoud, John Beaulaurier, Sissel Juul, Todd J. Treangen, and Fritz J. Sedlazeck. 2023. "MethPhaser: Methylation-Based Haplotype Phasing of Human Genomes." bioRxiv. https://doi.org/10.1101/2023.05.12.540573.

Galindo-González, Leonardo, Corinne Mhiri, Michael K. Deyholos, and Marie-Angèle Grandbastien. 2017. "LTR-Retrotransposons in Plants: Engines of Evolution." *Gene* 626 (August): 14–25. https://doi.org/10.1016/j.gene.2017.04.051.

Gartner, Ulrike, Ingo Hein, Lynn H. Brown, Xinwei Chen, Sophie Mantelin, Sanjeev K. Sharma, Louise-Marie Dandurand, et al. 2021. "Resisting Potato Cyst Nematodes With Resistance." *Frontiers in Plant Science* 12 (March). https://doi.org/10.3389/fpls.2021.661194.

Gómez-Gómez, Lourdes, Georg Felix, and Thomas Boller. 1999. "A Single Locus Determines Sensitivity to Bacterial Flagellin in Arabidopsis Thaliana." *The Plant Journal* 18 (3): 277–84. https://doi.org/10.1046/j.1365-313X.1999.00451.x.

Gong, Zhiyun, Yufeng Wu, Andrea Koblížková, Giovana A. Torres, Kai Wang, Marina Iovene, Pavel Neumann, et al. 2012. "Repeatless and Repeat-Based Centromeres in Potato: Implications for Centromere Evolution." *The Plant Cell* 24 (9): 3559–74. https://doi.org/10.1105/tpc.112.100511.

Gozashti, Landen, and Hopi E. Hoekstra. 2024. "Accounting for Diverse Transposable Element Landscapes Is Key to Developing and Evaluating Accurate de Novo Annotation Strategies." *Genome Biology* 25 (1): 4. https://doi.org/10.1186/s13059-023-03118-1.

Grabundzija, Ivana, Simon A. Messing, Jainy Thomas, Rachel L. Cosby, Ilija Bilic, Csaba Miskey, Andreas Gogol-Döring, et al. 2016. "A Helitron Transposon Reconstructed from Bats Reveals a Novel Mechanism of Genome Shuffling in Eukaryotes." *Nature Communications* 7 (1): 10716. https://doi.org/10.1038/ncomms10716.

Hafeez, Amber N., Sanu Arora, Sreya Ghosh, David Gilbert, Robert L. Bowden, and Brande B. H. Wulff. 2021. "Creation and Judicious Application of a Wheat Resistance Gene Atlas." *Molecular Plant* 14 (7): 1053–70. https://doi.org/10.1016/j.molp.2021.05.014.

Hallgren, Jeppe, Konstantinos D. Tsirigos, Mads Damgaard Pedersen, José Juan Almagro Armenteros, Paolo Marcatili, Henrik Nielsen, Anders Krogh, and Ole Winther. 2022. "DeepTMHMM Predicts Alpha and Beta Transmembrane Proteins Using Deep Neural Networks." bioRxiv. https://doi.org/10.1101/2022.04.08.487609.

Henikoff, Steven, and Jorja G. Henikoff. 2012. "'Point' Centromeres of Saccharomyces Harbor Single Centromere-Specific Nucleosomes." *Genetics* 190 (4): 1575. https://doi.org/10.1534/genetics.111.137711.

Hijmans, Robert J., Tatjana Gavrilenko, Sarah Stephenson, John Bamberg, Alberto Salas, and David M. Spooner. 2007. "Geographical and Environmental Range Expansion Through Polyploidy in Wild Potatoes (Solanum Section Petota)." *Global Ecology and Biogeography* 16 (4): 485–95. https://doi.org/10.1111/j.1466-8238.2007.00308.x.

Holst, Felix, Anthony Bolger, Christopher Günther, Janina Maß, Sebastian Triesch, Felicitas Kindel, Niklas Kiel, et al. 2023. "Helixer–de Novo Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model." bioRxiv. https://doi.org/10.1101/2023.02.06.527280.

Hoogen, Johan van den, Stefan Geisen, Devin Routh, Howard Ferris, Walter Traunspurger, David A. Wardle, Ron G. M. de Goede, et al. 2019. "Soil Nematode Abundance and Functional Group Composition at a Global Scale." *Nature* 572 (7768): 194–98. https://doi.org/10.1038/s41586-019-1418-6.

Hoopes, Genevieve, Xiaoxi Meng, John P. Hamilton, Sai Reddy Achakkagari, Fernanda de Alves Freitas Guesdes, Marie E. Bolger, Joseph J. Coombs, et al. 2022. "Phased, Chromosome-Scale Genome Assemblies of Tetraploid Potato Reveal a Complex Genome, Transcriptome, and Predicted Proteome Landscape Underpinning Genetic Diversity." *Molecular Plant* 15 (3): 520–36. https://doi.org/10.1016/j.molp.2022.01.003.

Horvath, Diana M., Robert E. Stall, Jeffrey B. Jones, Michael H. Pauly, Gary E. Vallad, Doug Dahlbeck, Brian J. Staskawicz, and John W. Scott. 2012. "Transgenic Resistance Confers Effective Field Level Control of Bacterial Spot Disease in Tomato." *PLOS ONE* 7 (8): e42036. https://doi.org/10.1371/journal.pone.0042036.

Hosaka, Awie J, Rena Sanetomo, and Kazuyoshi Hosaka. 2022. "A de Novo Genome Assembly of Solanum Verrucosum Schlechtendal, a Mexican Diploid Species Geographically Isolated from Other Diploid A-Genome Species of Potato Relatives." *G3 Genes|Genomes|Genetics* 12 (8): jkac166. https://doi.org/10.1093/g3journal/jkac166.

Hu, Jianbing, Chenchen Liu, Zezhen Du, Furong Guo, Dan Song, Nan Wang, Zhuang-

min Wei, et al. 2024. "Transposable Elements Cause the Loss of Self-Incompatibility in Citrus." *Plant Biotechnology Journal* 22 (5): 1113–31. https://doi.org/10.1111/pbi.14250.

Hu, Kaining, Kai Xu, Jing Wen, Bin Yi, Jinxiong Shen, Chaozhi Ma, Tingdong Fu, Yidan Ouyang, and Jinxing Tu. 2019. "Helitron Distribution in Brassicaceae and Whole Genome Helitron Density as a Character for Distinguishing Plant Species." *BMC Bioinformatics* 20 (1): 354. https://doi.org/10.1186/s12859-019-2945-8.

Huang, Li, Yulin Yuan, Chloe Lewis, Joanna Kud, Joseph C. Kuhl, Allan Caplan, Louise-Marie Dandurand, Inga Zasada, and Fangming Xiao. 2023. "NILR1 Perceives a Nematode Ascaroside Triggering Immune Signaling and Resistance." *Current Biology* 33 (18): 3992–3997.e3. https://doi.org/10.1016/j.cub.2023.08.017.

Hunter, John D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

Ivanov, Artemii A., Egor O. Ukladov, and Tatiana S. Golubeva. 2021. "Phytophthora Infestans: An Overview of Methods and Attempts to Combat Late Blight." *Journal of Fungi* 7 (12): 1071. https://doi.org/10.3390/jof7121071.

Iyigun, Murat, Nathan Nunn, and Nancy Qian. 2017. "The Long-Run Effects of Agricultural Productivity on Conflict, 1400-1900." Working {Paper}. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w24066.

Jones, John T., Annelies Haegeman, Etienne G. J. Danchin, Hari S. Gaur, Johannes Helder, Michael G. K. Jones, Taisei Kikuchi, et al. 2013. "Top 10 Plant-Parasitic Nematodes in Molecular Plant Pathology." *Molecular Plant Pathology* 14 (9): 946–61. https://doi.org/10.1111/mpp.12057.

Jones, John T., and Melissa G. Mitchum. 2018. "Biology of Effectors." In *Cyst Nematodes*, 74–88. CABI.

Jones, Jonathan D. G., and Jeffery L. Dangl. 2006. "The Plant Immune System." *Nature* 444 (7117): 323–29. https://doi.org/10.1038/nature05286.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." *Bioinformatics* 30 (9): 1236–40. https://doi.org/10.1093/bioinformatics/btu031.

Jonge, Ronnie de, H. Peter van Esse, Anja Kombrink, Tomonori Shinya, Yoshitake Desaki, Ralph Bours, Sander van der Krol, Naoto Shibuya, Matthieu H. A. J. Joosten, and Bart P. H. J. Thomma. 2010. "Conserved Fungal LysM Effector Ecp6 Prevents Chitin-Triggered Immunity in Plants." *Science (New York, N.Y.)* 329 (5994): 953–55. https://doi.org/10.1126/science.1190859.

Jupe, Florian, Leighton Pritchard, Graham J. Etherington, Katrin MacKenzie, Peter JA Cock, Frank Wright, Sanjeev Kumar Sharma, et al. 2012. "Identification and Localisation of the NB-LRR Gene Family Within the Potato Genome." *BMC Genomics* 13 (1): 75. https://doi.org/10.1186/1471-2164-13-75.

Jupe, Florian, Kamil Witek, Walter Verweij, Jadwiga Śliwka, Leighton Pritchard, Graham J. Etherington, Dan Maclean, et al. 2013. "Resistance Gene Enrichment Sequencing (RenSeq) Enables Reannotation of the NB-LRR Gene Family from Sequenced Plant Genomes and Rapid Mapping of Resistance Loci in Segregating Populations." *The Plant Journal* 76 (3): 530–44. https://doi.org/10.1111/tpj.12307.

Kalendar, Ruslan, Olga Raskina, Alexander Belyayev, and Alan H. Schulman. 2020.

"Long Tandem Arrays of Cassandra Retroelements and Their Role in Genome Dynamics in Plants." *International Journal of Molecular Sciences* 21 (8): 2931. https://doi.org/10.3390/ijms21082931.

Kent, W. J., A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik. 2010. "BigWig and BigBed: Enabling Browsing of Large Distributed Datasets." *Bioinformatics* 26 (17): 2204–7. https://doi.org/10.1093/bioinformatics/btq351.

Kessel, Geert J. T., Ewen Mullins, Albartus Evenhuis, Jeroen Stellingwerf, Vilma Ortiz Cortes, Sinead Phelan, Trudy van den Bosch, et al. 2018. "Development and Validation of IPM Strategies for the Cultivation of Cisgenically Modified Late Blight Resistant Potato." *European Journal of Agronomy* 96 (May): 146–55. https://doi.org/10.1016/j.eja.2018.01.012.

Khan, Masudulla, and Azhar U. Khan. 2021. "Plant Parasitic Nematodes Effectors and Their Crosstalk with Defense Response of Host Plants: A Battle Underground." *Rhizosphere* 17 (March): 100288. https://doi.org/10.1016/j.rhisph.2020.100288.

Kille, Bryce, Erik Garrison, Todd J. Treangen, and Adam M. Phillippy. 2023. "Minmers Are a Generalization of Minimizers That Enable Unbiased Local Jaccard Estimation." bioRxiv. https://doi.org/10.1101/2023.05.16.540882.

Kim, Seungill, Jieun Park, Seon-In Yeom, Yong-Min Kim, Eunyoung Seo, Ki-Tae Kim, Myung-Shin Kim, et al. 2017. "New Reference Genome Sequences of Hot Pepper Reveal the Massive Evolution of Plant Disease-Resistance Genes by Retroduplication." *Genome Biology* 18 (1): 210. https://doi.org/10.1186/s13059-017-1341-9.

Kobe, Bostjan, and Andrey V Kajava. 2001. "The Leucine-Rich Repeat as a Protein Recognition Motif." *Current Opinion in Structural Biology* 11 (6): 725–32. https://doi.org/10.1016/S0959-440X(01)00266-4.

Kokot, Marek, Maciej Długosz, and Sebastian Deorowicz. 2017. "KMC 3: Counting and Manipulating k-Mer Statistics." *Bioinformatics* 33 (17): 2759–61. https://doi.org/10.1093/bioinformatics/btx304.

Koren, Sergey, Zhigui Bao, Andrea Guarracino, Shujun Ou, Sara Goodwin, Katharine M. Jenike, Julian Lucas, et al. 2024. "Gapless Assembly of Complete Human and Plant Chromosomes Using Only Nanopore Sequencing." *bioRxiv*, March, 2024.03.15.585294. https://doi.org/10.1101/2024.03.15.585294.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. "Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation." *Genome Research* 27 (5): 722–36. https://doi.org/10.1101/gr.215087.116.

Kourelis, Jiorgos, Clemence Marchal, Andres Posbeyikian, Adeline Harant, and Sophien Kamoun. 2023. "NLR Immune Receptor–Nanobody Fusions Confer Plant Disease Resistance." *Science* 379 (6635): 934–39. https://doi.org/10.1126/science.abn4116.

Kourelis, Jiorgos, Toshiyuki Sakai, Hiroaki Adachi, and Sophien Kamoun. 2021. "RefPlantNLR Is a Comprehensive Collection of Experimentally Validated Plant Disease Resistance Proteins from the NLR Family." *PLOS Biology* 19 (10): e3001124. https://doi.org/10.1371/journal.pbio.3001124.

Kovaka, Sam, Shujun Ou, Katharine M. Jenike, and Michael C. Schatz. 2023. "Approaching Complete Genomes, Transcriptomes and Epi-Omes with Accurate Long-Read Sequencing." *Nature Methods* 20 (1): 12–16. https://doi.org/10.1038/s41592-022-01716-8.

Kraitshtein, Zina, Beery Yaakov, Vadim Khasdan, and Khalil Kashkush. 2010. "Genetic and Epigenetic Dynamics of a Retrotransposon After Allopolyploidization of Wheat." *Genetics* 186 (3): 801–12. https://doi.org/10.1534/genetics.110.120790.

Kroj, Thomas, Emilie Chanclud, Corinne Michel-Romiti, Xavier Grand, and Jean-Benoit Morel. 2016. "Integration of Decoy Domains Derived from Protein Targets of Pathogen Effectors into Plant Immune Receptors Is Widespread." *New Phytologist* 210 (2): 618–26. https://doi.org/10.1111/nph.13869.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lee, Hye-Young, Hyunggon Mang, Eunhye Choi, Ye-Eun Seo, Myung-Shin Kim, Soohyun Oh, Saet-Byul Kim, and Doil Choi. 2021. "Genome-wide Functional Analysis of Hot Pepper Immune Receptors Reveals an Autonomous NLR Clade in Seed Plants." *The New Phytologist* 229 (1): 532–47. https://doi.org/10.1111/nph.16878.

Lee, Sarah, Felix E. Enciso-Rodriguez, William Behling, Thilani Jayakody, Kaela Panicucci, Daniel Zarka, Satya Swathi Nadakuduti, C. Robin Buell, Norma C. Manrique-Carpintero, and David S. Douches. 2023. "HT-B and S-RNase CRISPR-Cas9 Double Knockouts Show Enhanced Self-Fertility in Diploid Solanum Tuberosum." *Frontiers in Plant Science* 14 (May). https://doi.org/10.3389/fpls.2023.1151347.

Li, Feng, Michael Lee, Caroline Esnault, Katie Wendover, Yabin Guo, Paul Atkins, Mikel Zaratiegui, and Henry L. Levin. 2022. "Identification of an Integrase-Independent Pathway of Retrotransposition." *Science Advances* 8 (26): eabm9390. https://doi.org/10.1126/sciadv.abm9390.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." arXiv. https://doi.org/10.48550/arXiv.1303.3997.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Ning, Qiang He, Juan Wang, Baike Wang, Jiantao Zhao, Shaoyong Huang, Tao Yang, et al. 2023. "Super-Pangenome Analyses Highlight Genomic Diversity and Structural Variation Across Wild and Cultivated Tomato Species." *Nature Genetics* 55 (5): 852–60. https://doi.org/10.1038/s41588-023-01340-y.

Li, Pingchuan, Xiande Quan, Gaofeng Jia, Jin Xiao, Sylvie Cloutier, and Frank M. You. 2016. "RGAugury: A Pipeline for Genome-Wide Prediction of Resistance Gene Analogs (RGAs) in Plants." *BMC Genomics* 17 (1): 852. https://doi.org/10.1186/s12864-016-3197-x.

Lin, Xiao, Miles Armstrong, Katie Baker, Doret Wouters, Richard G. F. Visser, Pieter J. Wolters, Ingo Hein, and Vivianne G. A. A. Vleeshouwers. 2020. "RLP/K Enrichment Sequencing; a Novel Method to Identify Receptor-Like Protein (RLP) and Receptor-Like Kinase (RLK) Genes." *New Phytologist* 227 (4): 1264–76. https://doi.org/10.1111/nph.16608.

Lin, Xiao, Yuxin Jia, Robert Heal, Maxim Prokchorchik, Maria Sindalovskaya, Andrea Olave-Achury, Moffat Makechemu, et al. 2023. "Solanum Americanum Genome-Assisted Discovery of Immune Receptors That Detect Potato Late Blight Pathogen Effectors." *Nature Genetics* 55 (9): 1579–88. https://doi.org/10.1038/s41588-023-01486-9.

Lindqvist-Kreuze, Hannele, Manuel Gastelo, Willmer Perez, Gregory A. Forbes, David de Koeyer, and Merideth Bonierbale. 2014. "Phenotypic Stability and Genome-Wide Association Study of Late Blight Resistance in Potato Genotypes Adapted to the Tropical Highlands." *Phytopathology®* 104 (6): 624–33. https://doi.org/10.1094/PHYTO-10-13-0270-R.

Liu, Feng, Jiantao Zhao, Honghe Sun, Cheng Xiong, Xuepeng Sun, Xin Wang, Zhongyi Wang, et al. 2023. "Genomes of Cultivated and Wild Capsicum Species Provide Insights into Pepper Domestication and Population Differentiation." *Nature Communications* 14 (1): 5487. https://doi.org/10.1038/s41467-023-41251-4.

Liu, Shiming, Pramod K. Kandoth, Samantha D. Warren, Greg Yeckel, Robert Heinz, John Alden, Chunling Yang, et al. 2012. "A Soybean Cyst Nematode Resistance Gene Points to a New Mechanism of Plant Resistance to Pathogens." *Nature* 492 (7428): 256–60. https://doi.org/10.1038/nature11651.

Liu, Zhenyu, and Dennis Halterman. 2006. "Identification and Characterization of *RB*-Orthologous Genes from the Late Blight Resistant Wild Potato Species *Solanum Verrucosum*." *Physiological and Molecular Plant Pathology* 69 (4): 230–39. https://doi.org/10.1016/j.pmpp.2007.05.002.

Logsdon, Glennis A., Allison N. Rozanski, Fedor Ryabov, Tamara Potapova, Valery A. Shepelev, Claudia R. Catacchio, David Porubsky, et al. 2024. "The Variation and Evolution of Complete Human Centromeres." *Nature* 629 (8010): 136–45. https://doi.org/10.1038/s41586-024-07278-3.

Lozano-Durán, Rosa, Gildas Bourdais, Sheng Yang He, and Silke Robatzek. 2014. "The Bacterial Effector HopM1 Suppresses PAMP-Triggered Oxidative Burst and Stomatal Immunity." *New Phytologist* 202 (1): 259–69. https://doi.org/10.1111/nph.12651.

Lu, Dongdong, Jixian Zhai, and Mengli Xi. 2022. "Regulation of DNA Methylation During Plant Endosperm Development." *Frontiers in Genetics* 13 (February). https://doi.org/10.3389/fgene.2022.760690.

Lu, Yunlong, Yunyun Song, Lingtong Liu, and Tai Wang. 2021. "DNA Methylation Dynamics of Sperm Cell Lineage Development in Tomato." *The Plant Journal* 105 (3): 565–79. https://doi.org/10.1111/tpj.15098.

Lüdke, Daniel, Toshiyuki Sakai, Jiorgos Kourelis, AmirAli Toghani, Hiroaki Adachi, Andrés Posbeyikian, Raoul Frijters, et al. 2023. "A Root-Specific NLR Network Confers Resistance to Plant Parasitic Nematodes." bioRxiv. https://doi.org/10.1101/2023.12.14.571630.

Ludwiczak, Jan, Aleksander Winski, Krzysztof Szczepaniak, Vikram Alva, and Stanislaw Dunin-Horkawicz. 2019. "DeepCoil—a Fast and Accurate Prediction of Coiled-Coil Domains in Protein Sequences." *Bioinformatics* 35 (16): 2790–95. https://doi.org/10.1093/bioinformatics/bty1062.

Mangul, Serghei, Thiago Mosqueiro, Richard J. Abdill, Dat Duong, Keith Mitchell, Varuni Sarwal, Brian Hill, et al. 2019. "Challenges and Recommendations to Improve the Installability and Archival Stability of Omics Computational Tools." *PLOS Biology* 17 (6): e3000333. https://doi.org/10.1371/journal.pbio.3000333.

Manni, Mosè, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, and Evgeny M Zdobnov. 2021. "BUSCO Update: Novel and Streamlined Workflows Along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic,

and Viral Genomes." *Molecular Biology and Evolution* 38 (10): 4647–54. https://doi.org/10.1093/molbev/msab199.

Manohar, Murli, Francisco Tenjo-Castano, Shiyan Chen, Ying K. Zhang, Anshu Kumari, Valerie M. Williamson, Xiaohong Wang, Daniel F. Klessig, and Frank C. Schroeder. 2020. "Plant Metabolism of Nematode Pheromones Mediates Plant-Nematode Interactions." *Nature Communications* 11 (1): 208. https://doi.org/10.1038/s41467-019-14104-2.

Martin, Eliza C, Catalin F Ion, Florin Ifrimescu, Laurentiu Spiridon, Jaap Bakker, Aska Goverse, and Andrei-J Petrescu. 2023. "NLRscape: An Atlas of Plant NLR Proteins." *Nucleic Acids Research* 51 (D1): D1470–82. https://doi.org/10.1093/nar/gkac1014.

Martin, Eliza C., Laurentiu Spiridon, Aska Goverse, and Andrei-José Petrescu. 2022. "NLRexpress—A Bundle of Machine Learning Motif Predictors—Reveals Motif Stability Underlying Plant Nod-Like Receptors Diversity." *Frontiers in Plant Science* 13. https://www.frontiersin.org/articles/10.3389/fpls.2022.975888.

Martin, Raoul, Tiancong Qi, Haibo Zhang, Furong Liu, Miles King, Claire Toth, Eva Nogales, and Brian J. Staskawicz. 2020. "Structure of the Activated ROQ1 Resistosome Directly Recognizing the Pathogen Effector XopQ." *Science* 370 (6521): eabd9993. https://doi.org/10.1126/science.abd9993.

Maruta, Natsumi, Hayden Burdett, Bryan Y. J. Lim, Xiahao Hu, Sneha Desa, Mohammad Kawsar Manik, and Bostjan Kobe. 2022. "Structural Basis of NLR Activation and Innate Immune Signalling in Plants." *Immunogenetics* 74 (1): 5–26. https://doi.org/10.1007/s00251-021-01242-5.

Mburu, Harrison, Laura Cortada, Solveig Haukeland, Wilson Ronno, Moses Nyongesa, Zachary Kinyua, Joel L. Bargul, and Danny Coyne. 2020. "Potato Cyst Nematodes: A New Threat to Potato Production in East Africa." *Frontiers in Plant Science* 11 (May). https://doi.org/10.3389/fpls.2020.00670.

McClure, Bruce, Felipe Cruz-García, and Carlos Romero. 2011. "Compatibility and Incompatibility in S-RNase-Based Systems." *Annals of Botany* 108 (4): 647–58. https://doi.org/10.1093/aob/mcr179.

McQueen-Mason, S, and D J Cosgrove. 1994. "Disruption of Hydrogen Bonding Between Plant Cell Wall Polymers by Proteins That Induce Wall Extension." *Proceedings of the National Academy of Sciences* 91 (14): 6574–78. https://doi.org/10.1073/pnas.91.14.6574.

Mendy, Badou, Mary Wanjiku Wang'ombe, Zoran S. Radakovic, Julia Holbein, Muhammad Ilyas, Divykriti Chopra, Nick Holton, Cyril Zipfel, Florian M. W. Grundler, and Shahid Siddique. 2017. "Arabidopsis Leucine-Rich Repeat Receptor–Like Kinase NILR1 Is Required for Induction of Innate Immunity to Parasitic Nematodes." *PLOS Pathogens* 13 (4): e1006284. https://doi.org/10.1371/journal.ppat.1006284.

Milligan, Stephen B., John Bodeau, Jafar Yaghoobi, Isgouhi Kaloshian, Pim Zabel, and Valerie M. Williamson. 1998. "The Root Knot Nematode Resistance Gene Mi from Tomato Is a Member of the Leucine Zipper, Nucleotide Binding, Leucine-Rich Repeat Family of Plant Genes." *The Plant Cell* 10 (8): 1307–19. https://doi.org/10.1105/tpc.10.8.1307.

Moens, M., R. N. Perry, and J. T. Jones. 2018. "Cyst Nematodes - Life Cycle and Economic Importance." *Cyst Nematodes*, CABI Books, January, 1–26. https://doi.org/10.1079/9781786390837.0001.

Mölder, Felix, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, et al. 2021. "Sustainable Data Analysis with Snakemake." 10:33. F1000Research. https://doi.org/10.12688/f1000research.29032.2.

Monino-Lopez, Daniel, Maarten Nijenhuis, Linda Kodde, Sophien Kamoun, Hamed Salehian, Kyrylo Schentsnyi, Remco Stam, et al. 2021. "Allelic Variants of the NLR Protein Rpi-Chc1 Differentially Recognize Members of the Phytophthora Infestans PexRD12/31 Effector Superfamily Through the Leucine-Rich Repeat Domain." *The Plant Journal* 107 (1): 182–97. https://doi.org/10.1111/tpj.15284.

Mundt, Christopher C. 2018. "Pyramiding for Resistance Durability: Theory and Practice." *Phytopathology®* 108 (7): 792–802. https://doi.org/10.1094/PHYTO-12-17-0426-RVW.

Muñoz-López, Martín, and José L. García-Pérez. 2010. "DNA Transposons: Nature and Applications in Genomics." *Current Genomics* 11 (2): 115. https://doi.org/10.2174/138920210790886871.

Muyle, Aline M, Danelle K Seymour, Yuanda Lv, Bruno Huettel, and Brandon S Gaut. 2022. "Gene Body Methylation in Plants: Mechanisms, Functions, and Important Implications for Understanding Evolutionary Processes." Edited by Tanja Slotte. *Genome Biology and Evolution* 14 (4): evac038. https://doi.org/10.1093/gbe/evac038.

Naish, Matthew, Michael Alonge, Piotr Wlodzimierz, Andrew J. Tock, Bradley W. Abramson, Anna Schmücker, Terezie Mandáková, et al. 2021. "The Genetic and Epigenetic Landscape of the Arabidopsis Centromeres." *Science* 374 (6569): eabi7489. https://doi.org/10.1126/science.abi7489.

Naish, Matthew, and Ian R. Henderson. 2024. "The Structure, Function, and Evolution of Plant Centromeres." *Genome Research* 34 (2): 161–78. https://doi.org/10.1101/gr.278409.123.

Neumann, Pavel, Alice Navrátilová, Andrea Koblížková, Eduard Kejnovský, Eva Hřibová, Roman Hobza, Alex Widmer, Jaroslav Doležel, and Jiří Macas. 2011. "Plant Centromeric Retrotransposons: A Structural and Cytogenetic Perspective." *Mobile DNA* 2 (1): 4. https://doi.org/10.1186/1759-8753-2-4.

Ngou, Bruno Pok Man, Hee-Kyung Ahn, Pingtao Ding, and Jonathan D. G. Jones. 2021. "Mutual Potentiation of Plant Immunity by Cell-Surface and Intracellular Receptors." *Nature* 592 (7852): 110–15. https://doi.org/10.1038/s41586-021-03315-7.

Ngou, Bruno Pok Man, Pingtao Ding, and Jonathan D G Jones. 2022. "Thirty Years of Resistance: Zig-Zag Through the Plant Immune System." *The Plant Cell* 34 (5): 1447–78. https://doi.org/10.1093/plcell/koac041.

Ni, Peng, Neng Huang, Fan Nie, Jun Zhang, Zhi Zhang, Bo Wu, Lu Bai, et al. 2021. "Genome-Wide Detection of Cytosine Methylations in Plant from Nanopore Data Using Deep Learning." *Nature Communications* 12 (1): 5976. https://doi.org/10.1038/s41467-021-26278-9.

Ntoukakis, Vardis, Isabel ML Saur, Brendon Conlan, and John P Rathjen. 2014. "The Changing of the Guard: The Pto/Prf Receptor Complex of Tomato and Pathogen Recognition." *Current Opinion in Plant Biology*, SI: Biotic interactions, 20 (August): 69–74. https://doi.org/10.1016/j.pbi.2014.04.002.

Nurk, Sergey, Brian P. Walenz, Arang Rhie, Mitchell R. Vollger, Glennis A. Logsdon, Robert Grothe, Karen H. Miga, Evan E. Eichler, Adam M. Phillippy, and Sergey

Koren. 2020. "HiCanu: Accurate Assembly of Segmental Duplications, Satellites, and Allelic Variants from High-Fidelity Long Reads." *Genome Research* 30 (9): 1291–1305. https://doi.org/10.1101/gr.263566.120.

Odilbekov, Firuz, Catja Selga, Rodomiro Ortiz, Aakash Chawade, and Erland Liljeroth. 2020. "QTL Mapping for Resistance to Early Blight in a Tetraploid Potato Population." *Agronomy* 10 (5): 728. https://doi.org/10.3390/agronomy10050728.

Ooijen, Gerben van, Gabriele Mayr, Mobien M. A. Kasiem, Mario Albrecht, Ben J. C. Cornelissen, and Frank L. W. Takken. 2008. "Structure–Function Analysis of the NB-ARC Domain of Plant Disease Resistance Proteins." *Journal of Experimental Botany* 59 (6): 1383–97. https://doi.org/10.1093/jxb/ern045.

Open2C, Nezar Abdennur, Sameer Abraham, Geoffrey Fudenberg, Ilya M. Flyamer, Aleksandra A. Galitsyna, Anton Goloborodko, Maxim Imakaev, Betul A. Oksuz, and Sergey V. Venev. 2022. "Cooltools: Enabling High-Resolution Hi-C Analysis in Python." bioRxiv. https://doi.org/10.1101/2022.10.31.514564.

Open2C, Nezar Abdennur, Geoffrey Fudenberg, Ilya M. Flyamer, Aleksandra A. Galitsyna, Anton Goloborodko, Maxim Imakaev, and Sergey V. Venev. 2023. "Pairtools: From Sequencing Data to Chromosome Contacts." bioRxiv. https://doi.org/10.1101/2023.02.13.528389.

Ortiz, Vilma, Sinead Phelan, and Ewen Mullins. 2016. "A Temporal Assessment of Nematode Community Structure and Diversity in the Rhizosphere of Cisgenic Phytophthora Infestans-Resistant Potatoes." *BMC Ecology* 16 (1): 55. https://doi.org/10.1186/s12898-016-0109-5.

Osuna-Cruz, Cristina M, Andreu Paytuvi-Gallart, Antimo Di Donato, Vicky Sundesha, Giuseppe Andolfo, Riccardo Aiese Cigliano, Walter Sanseverino, and Maria R Ercolano. 2018. "PRGdb 3.0: A Comprehensive Platform for Prediction and Analysis of Plant Disease Resistance Genes." *Nucleic Acids Research* 46 (D1): D1197–1201. https://doi.org/10.1093/nar/gkx1119.

Ou, Shujun, Weija Su, Yi Liao, Kapeel Chougule, Jireh R. A. Agda, Adam J. Hellinga, Carlos Santiago Blanco Lugo, et al. 2019. "Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline." *Genome Biology* 20 (1): 275. https://doi.org/10.1186/s13059-019-1905-y.

Paajanen, Pirita, George Kettleborough, Elena López-Girona, Michael Giolai, Darren Heavens, David Baker, Ashleigh Lister, et al. 2019. "A Critical Comparison of Technologies for a Plant Genome Sequencing Project." *GigaScience* 8 (3): giy163. https://doi.org/10.1093/gigascience/giy163.

Padmanabhan, Meenu, Patrick Cournoyer, and S. P. Dinesh-Kumar. 2009. "The Leucine-Rich Repeat Domain in Plant Innate Immunity: A Wealth of Possibilities." *Cellular Microbiology* 11 (2): 191–98. https://doi.org/10.1111/j.1462-5822.2008.01260.x.

Perry, Roland N., and Maurice Moens. 2011. "Introduction to Plant-Parasitic Nematodes; Modes of Parasitism." In *Genomics and Molecular Genetics of Plant-Nematode Interactions*, edited by John Jones, Godelieve Gheysen, and Carmen Fenoll, 3–20. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0434-3_1.

Pham, Gina M, John P Hamilton, Joshua C Wood, Joseph T Burke, Hainan Zhao, Brieanne Vaillancourt, Shujun Ou, Jiming Jiang, and C Robin Buell. 2020. "Construction of a Chromosome-Scale Long-Read Reference Genome Assembly for Potato." *GigaScience* 9 (9): giaa100. https://doi.org/10.1093/gigascience/giaa100.

Postma, Wiebe J., Erik J. Slootweg, Sajid Rehman, Anna Finkers-Tomczak, Tom O. G. Tytgat, Kasper van Gelderen, Jose L. Lozano-Torres, et al. 2012. "The Effector SPRYSEC-19 of Globodera Rostochiensis Suppresses CC-NB-LRR-Mediated Disease Resistance in Plants." *Plant Physiology* 160 (2): 944–54. https://doi.org/10.1104/pp.112.200188.

Prigozhin, Daniil M., Chandler A. Sutherland, Sanjay Rangavajjhala, and Ksenia V. Krasileva. 2024. "Majority of the Highly Variable NLRs in Maize Share Genomic Location and Contain Additional Target-Binding Domains." bioRxiv. https://doi.org/10.1101/2022.10.05.510735.

Putker, V., Q. Zheng, Ana Catarina Silva, M. J. Zerdoner, A. Zivkovic, O. C. A. Sukarta, and A. Goverse. 2024. "Cellular Dynamics Underlying Globodera Pallida Effector RBP-1 Recognition and Function." In. https://research.wur.nl/en/publications/cellular-dynamics-underlying-globodera-pallida-effector-rbp-1-rec.

Qin, Ling, Urszula Kudla, Erwin H. A. Roze, Aska Goverse, Herman Popeijus, Jeroen Nieuwland, Hein Overmars, et al. 2004. "A Nematode Expansin Acting on Plants." *Nature* 427 (6969): 30–30. https://doi.org/10.1038/427030a.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Ramírez, Fidel, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. 2016. "deepTools2: A Next Generation Web Server for Deep-Sequencing Data Analysis." *Nucleic Acids Research* 44 (W1): W160–65. https://doi.org/10.1093/nar/gkw257.

Ranallo-Benavidez, T. Rhyker, Kamil S. Jaron, and Michael C. Schatz. 2020. "GenomeScope 2.0 and Smudgeplot for Reference-Free Profiling of Polyploid Genomes." *Nature Communications* 11 (1): 1432. https://doi.org/10.1038/s41467-020-14998-3.

Raski, D. J., A. C. Goheen, L. A. Lider, and C. P. Meredith. 1983. "Strategies Against Grapevine Fanleaf Virus and Its Nematode Vector." *Plant Disease* 67 (3): 335–39. https://doi.org/10.1094/PD-67-335.

Rautiainen, Mikko, Sergey Nurk, Brian P. Walenz, Glennis A. Logsdon, David Porubsky, Arang Rhie, Evan E. Eichler, Adam M. Phillippy, and Sergey Koren. 2023. "Telomere-to-Telomere Assembly of Diploid Chromosomes with Verkko." *Nature Biotechnology* 41 (10): 1474–82. https://doi.org/10.1038/s41587-023-01662-6.

Raza, Muhammad Ammar, Ningning Yu, Dan Wang, Liwen Cao, Susheng Gan, and Liping Chen. 2017. "Differential DNA Methylation and Gene Expression in Reciprocal Hybrids Between Solanum Lycopersicum and S. Pimpinellifolium." *DNA Research* 24 (6): 597–607. https://doi.org/10.1093/dnares/dsx028.

Rhie, Arang, Brian P. Walenz, Sergey Koren, and Adam M. Phillippy. 2020. "Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies." *Genome Biology* 21 (1): 245. https://doi.org/10.1186/s13059-020-02134-9.

Roth, Morgane, Ana M. Florez-Rueda, Margot Paris, and Thomas Städler. 2018. "Wild Tomato Endosperm Transcriptomes Reveal Common Roles of Genomic Imprinting in Both Nuclear and Cellular Endosperm." *The Plant Journal* 95 (6): 1084–1101. https://doi.org/10.1111/tpj.14012.

Sacco, Melanie Ann, Kamila Koropacka, Eric Grenier, Marianne J. Jaubert, Alexandra

Blanchard, Aska Goverse, Geert Smant, and Peter Moffett. 2009. "The Cyst Nematode SPRYSEC Protein RBP-1 Elicits Gpa2- and RanGAP2-Dependent Plant Cell Death." *PLoS Pathogens* 5 (8): e1000564. https://doi.org/10.1371/journal.ppat.1000564.

Santana Silva, Raner José, and Fabienne Micheli. 2020. "RRGPredictor, a Set-Theory-Based Tool for Predicting Pathogen-Associated Molecular Pattern Receptors (PRRs) and Resistance (R) Proteins from Plants." *Genomics* 112 (3): 2666–76. https://doi.org/10.1016/j.ygeno.2020.03.001.

Satterlee, James W., David Alonso, Pietro Gramazio, Katharine M. Jenike, Jia He, Andrea Arrones, Gloria Villanueva, et al. 2024. "Convergent Evolution of Plant Prickles by Repeated Gene Co-Option over Deep Time." *Science* 385 (6708): eado1663. https://doi.org/10.1126/science.ado1663.

Savary, Serge, Laetitia Willocquet, Sarah Jane Pethybridge, Paul Esker, Neil McRoberts, and Andy Nelson. 2019. "The Global Burden of Pathogens and Pests on Major Food Crops." *Nature Ecology & Evolution* 3 (3): 430–39. https://doi.org/10.1038/s41559-018-0793-y.

Seibt, Kathrin M., Torsten Wenke, Katja Muders, Bernd Truberg, and Thomas Schmidt. 2016. "Short Interspersed Nuclear Elements (SINEs) Are Abundant in Solanaceae and Have a Family-Specific Impact on Gene Structure and Genome Organization." *The Plant Journal* 86 (3): 268–85. https://doi.org/10.1111/tpj.13170.

Seo, Eunyoung, Seungill Kim, Seon-In Yeom, and Doil Choi. 2016. "Genome-Wide Comparative Analyses Reveal the Dynamic Evolution of Nucleotide-Binding Leucine-Rich Repeat Gene Family Among Solanaceae Plants." *Frontiers in Plant Science* 7. https://www.frontiersin.org/articles/10.3389/fpls.2016.01205.

Seong, Kyungyong, Eunyoung Seo, Kamil Witek, Meng Li, and Brian Staskawicz. 2020. "Evolution of NLR Resistance Genes with Noncanonical N-Terminal Domains in Wild Tomato Species." *New Phytologist* 227 (5): 1530–43. https://doi.org/10.1111/nph.16628.

Serra Mari, Rebecca, Sven Schrinner, Richard Finkers, Freya Maria Rosemarie Ziegler, Paul Arens, Maximilian H.-W. Schmidt, Björn Usadel, Gunnar W. Klau, and Tobias Marschall. 2024. "Haplotype-Resolved Assembly of a Tetraploid Potato Genome Using Long Reads and Low-Depth Offspring Data." *Genome Biology* 25 (1): 26. https://doi.org/10.1186/s13059-023-03160-z.

Servant, Nicolas, nf-core bot, Phil Ewels, Maxime U. Garcia, Adam Talbot, Alexander Peltzer, Edmund Miller, et al. 2023. "Nf-Core/Hic: Nf-Core/Hic V2.1.0." Zenodo. https://doi.org/10.5281/zenodo.7994878.

Servant, Nicolas, Nelle Varoquaux, Bryan R. Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. 2015. "HiC-Pro: An Optimized and Flexible Pipeline for Hi-C Data Processing." *Genome Biology* 16 (1): 259. https://doi.org/10.1186/s13059-015-0831-x.

Sharma, Sanjeev Kumar, Daniel Bolser, Jan de Boer, Mads Sønderkær, Walter Amoros, Martin Federico Carboni, Juan Martín D'Ambrosio, et al. 2013. "Construction of Reference Chromosome-Scale Pseudomolecules for Potato: Integrating the Potato Genome with Genetic and Physical Maps." *G3 Genes|Genomes|Genetics* 3 (11): 2031–47. https://doi.org/10.1534/g3.113.007153.

Shi, Junpeng, Zhixi Tian, Jinsheng Lai, and Xuehui Huang. 2023. "Plant Pan-Genomics

and Its Applications." *Molecular Plant* 16 (1): 168–86. https://doi.org/10.1016/j.molp.2022.12.009.

Shimada, Atsushi, Jonathan Cahn, Evan Ernst, Jason Lynn, Daniel Grimanelli, Ian Henderson, Tetsuji Kakutani, and Robert A. Martienssen. 2023. "Retrotransposon Addiction Promotes Centromere Function via Epigenetically Activated Small RNAs." bioRxiv. https://doi.org/10.1101/2023.08.02.551486.

Smant, Geert, Jack P. W. G. Stokkermans, Yitang Yan, Jan M. de Boer, Thomas J. Baum, Xiaohong Wang, Richard S. Hussey, et al. 1998. "Endogenous Cellulases in Animals: Isolation of β-1,4-Endoglucanase Genes from Two Species of Plant-Parasitic Cyst Nematodes." *Proceedings of the National Academy of Sciences* 95 (9): 4906–11. https://doi.org/10.1073/pnas.95.9.4906.

Smith, Moray, John T. Jones, and Ingo Hein. 2024. "Resistify - A Rapid and Accurate Annotation Tool to Identify NLRs and Study Their Genomic Organisation." bioRxiv. https://doi.org/10.1101/2024.02.14.580321.

Sobczak, Miroslaw, Anna Avrova, Justyna Jupowicz, Mark S. Phillips, Karin Ernst, and Amar Kumar. 2005. "Characterization of Susceptibility and Resistance Responses to Potato Cyst Nematode (Globodera Spp.) Infection of Tomato Lines in the Absence and Presence of the Broad-Spectrum Nematode Resistance Hero Gene." *Molecular Plant-Microbe Interactions®* 18 (2): 158–68. https://doi.org/10.1094/MPMI-18-0158.

Stanojević, Dominik, Dehui Lin, Paola Florez de Sessions, and Mile Šikić. 2024. "Telomere-to-Telomere Phased Genome Assembly Using Error-Corrected Simplex Nanopore Reads." bioRxiv. https://doi.org/10.1101/2024.05.18.594796.

Steuernagel, Burkhard, Florian Jupe, Kamil Witek, Jonathan D. G. Jones, and Brande B. H. Wulff. 2015. "NLR-Parser: Rapid Annotation of Plant NLR Complements." *Bioinformatics (Oxford, England)* 31 (10): 1665–67. https://doi.org/10.1093/bioinformatics/btv005.

Steuernagel, Burkhard, Kamil Witek, Simon G. Krattinger, Ricardo H. Ramirez-Gonzalez, Henk-jan Schoonbeek, Guotai Yu, Erin Baggs, et al. 2020. "The NLR-Annotator Tool Enables Annotation of the Intracellular Immune Receptor Repertoire1 [OPEN]." *Plant Physiology* 183 (2): 468–82. https://doi.org/10.1104/pp.19.01273.

Stoiber, Marcus, Joshua Quick, Rob Egan, Ji Eun Lee, Susan Celniker, Robert K. Neely, Nicholas Loman, Len A. Pennacchio, and James Brown. 2017. "De Novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing." bioRxiv. https://doi.org/10.1101/094672.

Strachan, Shona. 2018. "Characterisation of the Potato H2 Resistance Gene Against Globodera Pallida." Thesis, University of St Andrews. https://doi.org/10.17630/10023-17228.

Strachan, Shona M., Miles R. Armstrong, Amanpreet Kaur, Kathryn M. Wright, Tze Yin Lim, Katie Baker, John Jones, Glenn Bryan, Vivian Blok, and Ingo Hein. 2019. "Mapping the H2 Resistance Effective Against Globodera Pallida Pathotype Pa1 in Tetraploid Potato." *Theoretical and Applied Genetics* 132 (4): 1283–94. https://doi.org/10.1007/s00122-019-03278-4.

Su, Xiao, Baoan Wang, Xiaolin Geng, Yuefan Du, Qinqin Yang, Bin Liang, Ge Meng, et al. 2021. "A High-Continuity and Annotated Tomato Reference Genome." *BMC Genomics* 22 (1): 898. https://doi.org/10.1186/s12864-021-08212-x.

Sun, Yadong, Lei Li, Alberto P. Macho, Zhifu Han, Zehan Hu, Cyril Zipfel, Jian-Min Zhou, and Jijie Chai. 2013. "Structural Basis for Flg22-Induced Activation of the Arabidopsis FLS2-BAK1 Immune Complex." *Science* 342 (6158): 624–28. https://doi.org/10.1126/science.1243825.

Sutherland, Chandler A, Daniil M Prigozhin, J Grey Monroe, and Ksenia V Krasileva. 2024. "High Allelic Diversity in Arabidopsis NLRs Is Associated with Distinct Genomic Features." *EMBO Reports* 25 (5): 2306–22. https://doi.org/10.1038/s44319-024-00122-9.

Talbert, Paul B., and Steven Henikoff. 2020. "What Makes a Centromere?" *Experimental Cell Research* 389 (2): 111895. https://doi.org/10.1016/j.yexcr.2020.111895.

Tameling, Wladimir I. L., and David C. Baulcombe. 2007. "Physical Association of the NB-LRR Resistance Protein Rx with a Ran GTPase–Activating Protein Is Required for Extreme Resistance to Potato Virus X." *The Plant Cell* 19 (5): 1682–94. https://doi.org/10.1105/tpc.107.050880.

Tang, Dié, Yuxin Jia, Jinzhe Zhang, Hongbo Li, Lin Cheng, Pei Wang, Zhigui Bao, et al. 2022. "Genome Evolution and Diversity of Wild and Cultivated Potatoes." *Nature* 606 (7914): 535–41. https://doi.org/10.1038/s41586-022-04822-x.

Teasdale, Luisa C., Kevin D. Murray, Max Collenberg, Adrian Contreras-Garrido, Theresa Schlegel, Leon van Ess, Justina Jüttner, et al. 2024. "Pangenomic Context Reveals the Extent of Intraspecific Plant NLR Evolution." bioRxiv. https://doi.org/10.1101/2024.09.02.610789.

The Arabidopsis Genome Initiative. 2000. "Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana." *Nature* 408 (6814): 796–815. https://doi.org/10.1038/35048692.

The Potato Genome Sequencing Consortium. 2011. "Genome Sequence and Analysis of the Tuber Crop Potato." *Nature* 475 (7355): 189–95. https://doi.org/10.1038/nature10158.

Thomson, Steven. 2024. "The Estimated Economic Contribution of Scotland's Seed and Ware Potato Sectors. An Output of WP1: Economic Assessment of the Scottish Government Funded Project:"Delivering a Sustainable Potato Industry for Scotland Through Management of Potato Cyst Nematode (PCN)." Scotland's Rural College. https://doi.org/10.58073/SRUC.25244626.v1.

Thorpe, Peter, Sophie Mantelin, Peter JA Cock, Vivian C. Blok, Mirela C. Coke, Sebastian Eves-van den Akker, Elena Guzeeva, et al. 2014. "Genomic Characterisation of the Effector Complement of the Potato Cyst Nematode Globodera Pallida." *BMC Genomics* 15 (1): 923. https://doi.org/10.1186/1471-2164-15-923.

Toda, Nicholas, Camille Rustenholz, Agnès Baud, Marie-Christine Le Paslier, Joelle Amselem, Didier Merdinoglu, and Patricia Faivre-Rampant. 2020. "NLGenomeSweeper: A Tool for Genome-Wide NBS-LRR Resistance Gene Identification." *Genes* 11 (3): 333. https://doi.org/10.3390/genes11030333.

Toshchakov, Vladimir Y., and Andrew F. Neuwald. 2020. "A Survey of TIR Domain Sequence and Structure Divergence." *Immunogenetics* 72 (3): 181–203. https://doi.org/10.1007/s00251-020-01157-7.

Trudgill, D. L. 1967. "The Effect of Environment on Sex Determination in Heterodera Rostochiensis." *Nematologica* 13 (2): 263–72. https://doi.org/10.1163/187529267X00120.

Tsirigos, Konstantinos D., Christoph Peters, Nanjiang Shu, Lukas Käll, and Arne Elofsson. 2015. "The TOPCONS Web Server for Consensus Prediction of Membrane Protein Topology and Signal Peptides." *Nucleic Acids Research* 43 (Web Server issue): W401–7. https://doi.org/10.1093/nar/gkv485.

Tsuchiya, Tokuji, and Thomas Eulgem. 2013. "An Alternative Polyadenylation Mechanism Coopted to the Arabidopsis RPP7 Gene Through Intronic Retrotransposon Domestication." *Proceedings of the National Academy of Sciences* 110 (37): E3535–43. https://doi.org/10.1073/pnas.1312545110.

Van de Weyer, Anna-Lena, Freddy Monteiro, Oliver J. Furzer, Marc T. Nishimura, Volkan Cevik, Kamil Witek, Jonathan D. G. Jones, Jeffery L. Dangl, Detlef Weigel, and Felix Bemm. 2019. "A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis Thaliana." *Cell* 178 (5): 1260–1272.e14. https://doi.org/10.1016/j.cell.2019.07.038.

Wang, Fang, Zhiqiang Xia, Meiling Zou, Long Zhao, Sirong Jiang, Yun Zhou, Chenji Zhang, et al. 2022. "The Autotetraploid Potato Genome Provides Insights into Highly Heterozygous Species." *Plant Biotechnology Journal* 20 (10): 1996–2005. https://doi.org/10.1111/pbi.13883.

Wang, Jizong, Meijuan Hu, Jia Wang, Jinfeng Qi, Zhifu Han, Guoxun Wang, Yijun Qi, Hong-Wei Wang, Jian-Min Zhou, and Jijie Chai. 2019. "Reconstitution and Structure of a Plant NLR Resistosome Conferring Immunity." *Science (New York, N.Y.)* 364 (6435): eaav5870. https://doi.org/10.1126/science.aav5870.

Wang, Lei, Markus Albert, Elias Einig, Ursula Fürst, Damaris Krust, and Georg Felix. 2016. "The Pattern-Recognition Receptor CORE of Solanaceae Detects Bacterial Cold-Shock Protein." *Nature Plants* 2 (12): 1–9. https://doi.org/10.1038/nplants.2016.185.

Wang, Shumei, Hazel McLellan, Tatyana Bukharova, Qin He, Fraser Murphy, Jiayang Shi, Shaohui Sun, et al. 2019. "Phytophthora Infestans RXLR Effectors Act in Concert at Diverse Subcellular Locations to Enhance Host Colonization." *Journal of Experimental Botany* 70 (1): 343–56. https://doi.org/10.1093/jxb/ery360.

Wang, Weidong, Liyang Chen, Kevin Fengler, Joy Bolar, Victor Llaca, Xutong Wang, Chancelor B. Clark, et al. 2021. "A Giant NLR Gene Confers Broad-Spectrum Resistance to Phytophthora Sojae in Soybean." *Nature Communications* 12 (1): 6263. https://doi.org/10.1038/s41467-021-26554-8.

Wang, Yibin, Jiaxin Yu, Mengwei Jiang, Wenlong Lei, Xingtan Zhang, and Haibao Tang. 2023. "Sequencing and Assembly of Polyploid Genomes." In *Polyploidy: Methods and Protocols*, edited by Yves Van de Peer, 429–58. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-2561-3_23.

Wang, Yuhan, Lynn H Brown, Thomas M Adams, Yuk Woon Cheung, Jie Li, Vanessa Young, Drummond T Todd, et al. 2023. "SMRT–AgRenSeq-d in Potato (Solanum Tuberosum) as a Method to Identify Candidates for the Nematode Resistance Gpa5." *Horticulture Research* 10 (11): uhad211. https://doi.org/10.1093/hr/uhad211.

Waskom, Michael L. 2021. "Seaborn: Statistical Data Visualization." *Journal of Open Source Software* 6 (60): 3021. https://doi.org/10.21105/joss.03021.

Wersch, Solveig van, and Xin Li. 2019. "Stronger When Together: Clustering of Plant NLR Disease Resistance Genes." *Trends in Plant Science* 24 (8): 688–99. https://doi.org/10.1016/j.tplants.2019.05.005.

West, Patrick T., Qing Li, Lexiang Ji, Steven R. Eichten, Jawon Song, Matthew W. Vaughn, Robert J. Schmitz, and Nathan M. Springer. 2014. "Genomic Distribution of H3K9me2 and DNA Methylation in a Maize Genome." *PLOS ONE* 9 (8): e105267. https://doi.org/10.1371/journal.pone.0105267.

Wieczorek, Krzysztof, Bettina Golecki, Lars Gerdes, Petra Heinen, Dagmar Szakasits, Daniel M. Durachko, Daniel J. Cosgrove, et al. 2006. "Expansins Are Involved in the Formation of Nematode-Induced Syncytia in Roots of Arabidopsis Thaliana." *The Plant Journal* 48 (1): 98–112. https://doi.org/10.1111/j.1365-313X.2006.02856.x.

Witek, Kamil, Florian Jupe, Agnieszka I. Witek, David Baker, Matthew D. Clark, and Jonathan D. G. Jones. 2016. "Accelerated Cloning of a Potato Late Blight–Resistance Gene Using RenSeq and SMRT Sequencing." *Nature Biotechnology* 34 (6): 656–60. https://doi.org/10.1038/nbt.3540.

Witek, Kamil, Xiao Lin, Hari S. Karki, Florian Jupe, Agnieszka I. Witek, Burkhard Steuernagel, Remco Stam, et al. 2021. "A Complex Resistance Locus in Solanum Americanum Recognizes a Conserved Phytophthora Effector." *Nature Plants* 7 (2): 198–208. https://doi.org/10.1038/s41477-021-00854-9.

Wlodzimierz, Piotr, Fernando A. Rabanal, Robin Burns, Matthew Naish, Elias Primetis, Alison Scott, Terezie Mandáková, et al. 2023. "Cycles of Satellite and Transposon Evolution in Arabidopsis Centromeres." *Nature* 618 (7965): 557–65. https://doi.org/10.1038/s41586-023-06062-z.

Woudstra, Yannick, Hayley Tumas, Cyril van Ghelder, Tin Hang Hung, Joana J Ilska, Sebastien Girardi, Stuart A'Hara, et al. 2024. "Conifers Concentrate Large Numbers of NLR Immune Receptor Genes on One Chromosome." *Genome Biology and Evolution* 16 (6): evae113. https://doi.org/10.1093/gbe/evae113.

Wu, Jia-Yi, Jia-Yu Xue, and Yves Van de Peer. 2021. "Evolution of NLR Resistance Genes in Magnoliids: Dramatic Expansions of CNLs and Multiple Losses of TNLs." *Frontiers in Plant Science* 12 (December). https://doi.org/10.3389/fpls.2021.777157.

Xi, Yuxuan, Stella Cesari, and Thomas Kroj. 2022. "Insight into the Structure and Molecular Mode of Action of Plant Paired NLR Immune Receptors." *Essays in Biochemistry* 66 (5): 513–26. https://doi.org/10.1042/EBC20210079.

Xin, Haoyang, Yiduo Wang, Wenli Zhang, Yu Bao, Pavel Neumann, Yihang Ning, Tao Zhang, et al. 2024. "Celine, a Long Interspersed Nuclear Element Retrotransposon, Colonizes in the Centromeres of Poplar Chromosomes." *Plant Physiology*, April, kiae214. https://doi.org/10.1093/plphys/kiae214.

Xu, Boyan, Alois Cerbu, Daven Lim, Christopher J Tralie, and Ksenia Krasileva. 2023. "Structure-Aware Annotation of Leucine-Rich Repeat Domains." *bioRxiv*, November, 2023.10.27.562987. https://doi.org/10.1101/2023.10.27.562987.

Yan, Lang, Yizheng Zhang, Guangze Cai, Yuan Qing, Jiling Song, Haiyan Wang, Xuemei Tan, et al. 2021. "Genome Assembly of Primitive Cultivated Potato Solanum Stenotomum Provides Insights into Potato Evolution." *G3 Genes|Genomes|Genetics* 11 (10): jkab262. https://doi.org/10.1093/g3journal/jkab262.

Yang, Lixing, and Jeffrey L. Bennetzen. 2009. "Distribution, Diversity, Evolution, and Survival of Helitrons in the Maize Genome." *Proceedings of the National Academy of Sciences* 106 (47): 19922–27. https://doi.org/10.1073/pnas.0908008106.

Yang, Xiaohui, Lingkui Zhang, Xiao Guo, Jianfei Xu, Kang Zhang, Yinqing Yang, Yu Yang, et al. 2023. "The Gap-Free Potato Genome Assembly Reveals Large Tandem

Gene Clusters of Agronomical Importance in Highly Repeated Genomic Regions." *Molecular Plant* 16 (2): 314–17. https://doi.org/10.1016/j.molp.2022.12.010.

Yuan, Minhang, Zeyu Jiang, Guozhi Bi, Kinya Nomura, Menghui Liu, Yiping Wang, Boying Cai, Jian-Min Zhou, Sheng Yang He, and Xiu-Fang Xin. 2021. "Pattern-Recognition Receptors Are Required for NLR-Mediated Plant Immunity." *Nature* 592 (7852): 105–9. https://doi.org/10.1038/s41586-021-03316-6.

Zdrzałek, Rafał, Yuxuan Xi, Thorsten Langner, Adam R. Bentham, Yohann Petit-Houdenot, Juan Carlos De la Concepcion, Adeline Harant, et al. 2024. "Bioengineering a Plant NLR Immune Receptor with a Robust Binding Interface Towards a Conserved Fungal Pathogen Effector." bioRxiv. https://doi.org/10.1101/2024.01.20.576400.

Zeng, Yibing, R Kelly Dawe, and Jonathan I Gent. 2023. "Natural Methylation Epialleles Correlate with Gene Expression in Maize." *Genetics* 225 (2): iyad146. https://doi.org/10.1093/genetics/iyad146.

Zhang, Chunzhi, Pei Wang, Die Tang, Zhongmin Yang, Fei Lu, Jianjian Qi, Nilesh R. Tawari, Yi Shang, Canhui Li, and Sanwen Huang. 2019. "The Genetic Basis of Inbreeding Depression in Potato." *Nature Genetics* 51 (3): 374–78. https://doi.org/10.1038/s41588-018-0319-1.

Zhang, Chunzhi, Zhongmin Yang, Dié Tang, Yanhui Zhu, Pei Wang, Dawei Li, Guangtao Zhu, et al. 2021. "Genome Design of Hybrid Potato." *Cell* 184 (15): 3873–3883.e12. https://doi.org/10.1016/j.cell.2021.06.006.

Zhang, Haiqin, Andrea Koblížková, Kai Wang, Zhiyun Gong, Ludmila Oliveira, Giovana A. Torres, Yufeng Wu, et al. 2014. "Boom-Bust Turnovers of Megabase-Sized Centromeric DNA in Solanum Species: Rapid Evolution of DNA Sequences Associated with Centromeres." *The Plant Cell* 26 (4): 1436–47. https://doi.org/10.1105/tpc.114.123877.

Zhang, Longhao, Chengqi Yi, Xin Xia, Zheng Jiang, Lihui Du, Shixin Yang, and Xu Yang. 2024. "Solanum Aculeatissimum and Solanum Torvum Chloroplast Genome Sequences: A Comparative Analysis with Other Solanum Chloroplast Genomes." *BMC Genomics* 25 (1): 412. https://doi.org/10.1186/s12864-024-10190-9.

Zhang, Ren-Gang, Guang-Yuan Li, Xiao-Ling Wang, Jacques Dainat, Zhao-Xuan Wang, Shujun Ou, and Yongpeng Ma. 2022. "TEsorter: An Accurate and Fast Method to Classify LTR-Retrotransposons in Plant Genomes." *Horticulture Research* 9 (January): uhac017. https://doi.org/10.1093/hr/uhac017.

Zhou, Chenxi, Shane A McCarthy, and Richard Durbin. 2023. "YaHS: Yet Another Hi-C Scaffolding Tool." *Bioinformatics* 39 (1): btac808. https://doi.org/10.1093/bioinformatics/btac808.

Zhou, Wanding, Gangning Liang, Peter L. Molloy, and Peter A. Jones. 2020. "DNA Methylation Enables Transposable Element-Driven Genome Expansion." *Proceedings of the National Academy of Sciences* 117 (32): 19359–66. https://doi.org/10.1073/pnas.1921719117.

Zhu, Min, Lei Jiang, Baohui Bai, Wenyang Zhao, Xiaojiao Chen, Jia Li, Yong Liu, et al. 2017. "The Intracellular Immune Receptor Sw-5b Confers Broad-Spectrum Resistance to Tospoviruses Through Recognition of a Conserved 21-Amino Acid Viral Effector Epitope." *The Plant Cell* 29 (9): 2214–32. https://doi.org/10.1105/tpc.17.00180.