Title: Conifers concentrate large numbers of NLR immune receptor genes on one chromosome

Authors: Yannick Woudstra[1], Hayley Tumas[1], Cyril van Ghelder[2], Tin Hang Hung[1], Joana J. Ilska[3], Sebastien Girardi[4,5], Stuart A'Hara[6], Paul McLean[6], Steve J. Lee[6], Joan Cottrell[6], Joerg Bohlmann[7], Jean Bousquet[4], Inanc Birol[8], John A. Woolliams[3], John J. MacKay[1,*]

Affiliations:
[1] Department of Biology, University of Oxford, OX1 3RB, Oxford, UK
[2] INRAE, Université Côte d'Azur, CNRS, ISA, 06903 Sophia Antipolis, France
[3] The Roslin Institute, Royal (Dick) School of Veterinary Science, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, Scotland, UK
[4] Canada Research Chair in Forest Genomics, Forest Research Centre, Université Laval, Québec, QC, G1V 0A6, Canada
[5] Institute for Systems and Integrative Biology, Université Laval, Québec, QC, GIV 0A6, Canada
[6] Forest Research, Northern Research Station, Roslin, Midlothian, EH25 9SY, UK
[7] Michael Smith Laboratories, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada
[8] Canada's Michael Smith Genome Sciences Centre, Vancouver, BC, V5Z 4S6, Canada
[*] Author for correspondence: john.mackay@biology.ox.ac.uk, +44 (0) 1865 275088

Word counts (including (sub)headings):
Summary: 198
Introduction: 1,083
Materials & methods: 1,015
Results: 979
Discussion: 2,046
Total: 5,124

Summary

Research conducted

Nucleotide-binding domain and Leucine-rich Repeat (NLR) immune receptor genes form a major line of defence in plants, acting in both pathogen recognition and resistance machinery activation. NLRs are reported to form large gene clusters in pine but it is unknown how widespread this genomic architecture may be among the extant species of conifers (Pinophyta). We used comparative genomic analyses to assess patterns in the abundance, diversity and genomic distribution of NLR genes.

Methods

Chromosome-level whole genome assemblies and high-density linkage maps in the Pinaceae, Cupressaceae, Taxaceae and other gymnosperms were scanned for NLR genes using existing and customised pipelines. Discovered genes were mapped across chromosomes and linkage groups, and analysed phylogenetically for evolutionary history.

Key results

Conifer genomes are characterised by dense clusters of NLR genes, highly localised on one chromosome. These clusters are rich in TNL-encoding genes, which seem to have formed through multiple tandem duplication events.

Main conclusion

In contrast to angiosperms and non-coniferous gymnosperms, genomic clustering of NLR genes is ubiquitous in conifers. NLR-dense genomic regions are likely to influence a large part of the plant's resistance, informing our understanding of adaptation to biotic stress and the development of genetic resources through breeding.

Key words

Resistance genes, NBS-LRR, genomic architecture, comparative genomics, gene family evolution, gene clusters, gymnosperms, conservation biology

Plain language summary:

NLR immune receptor genes are important in pest, disease and drought resistance of plants. In the giga-genomes of conifers, they concentrate on very small chromosomal regions. These regions act as important reservoirs for NLR diversity and can be used in breeding to improve the resilience of conifer trees.

Introduction

Disease resistance is one of the key aspects of plant genetics and evolution studies with implications for conservation and ecosystem health, as well as breeding. Decades of research have improved our understanding of the identity and interplay of gene families involved in disease resistance have led to a decent understanding of the involved gene families and their interplay (Ngou, Ding, and Jones 2022). One of the first events in plant defence mechanisms is pathogen recognition, in which the Nucleotide-binding domain and Leucine-Rich Repeat (NBS-LRR or NLR) immune receptor gene family plays a central role (Duxbury, Wu, and Ding 2021). The products of NLR genes occur intracellularly and can bind directly to specific pathogenic effectors (pathogen-encoded proteins) or detect modifications of plant proteins induced by such effectors, thus activating a cascade of defence mechanisms upon perception (Ngou, Ding, and Jones 2022). NLRs provide a typical example of an evolutionary arms race in which pathogenic effectors evolve to evade detection by host NLRs which, in turn, evolve to recognise the new variants (e.g., the *Capsicum chinense* Jacq. NLR *Tsw* versus the pathogen ‘Tomato spotted wilt virus’ (Chen et al. 2023)). Unsurprisingly therefore, NLRs are diverse and abundant in many plant species with several hundred different NLR genes found in a range of land plant lineages (Barragan and Weigel 2021). While NLRs have been studied extensively in angiosperms (i.e, Cucurbitaceae (Lin et al. 2013), Rosaceae (Jia et al. 2015) and Solanaceae (Seo et al. 2016); Angiosperm NLR Atlas (Y. Liu et al. 2021)), studies in conifers are rare (Van Ghelder et al. 2019; J.-J. Liu et al. 2019; Ence et al. 2022).

NLRs exhibit a conserved tripartite structure consisting of 1) a non-conserved N-terminal domain; 2) a conserved central nucleotide-binding domain (NB-ARC, defined as “a nucleotide-binding adaptor shared by APAF-1, certain *R* gene products and CED-4” (van der Biezen and Jones 1998)); and 3) a C-terminal leucine-rich repeat (LRR) domain which can vary in length. These resistance genes seem to have originated before the rise of land plants (Shao et al. 2019) and have since diversified into three main classes based on the character of the N-terminal domain: CNL, RNL and TNL. CNLs (N-terminal ‘coiled-coil’ domain) and

RNLs (N-terminal 'resistance to powdery mildew 8 (RPW8) domain') are closely related classes, whereas TNLs (N-terminal 'Toll interleukin-1 receptor' domain) form a distinct class. All three classes are unusually abundant and diverse in conifers, with an RNL diversity that is distinctly higher than in any other group of land plants (Van Ghelder et al. 2019). Furthermore, a genomic distribution analysis in limber pine (*Pinus flexilis* E.James) revealed an unbalanced intragenomic and intrachromosomal distribution of NLRs (J.-J. Liu et al. 2019). In this species, one chromosome contained a dense cluster of NLRs comprising mainly TNLs, indicating a high rate of tandem duplications. Besides disease resistance, NLRs have been shown to be responsive to drought stress when investigated in the conifer white spruce (*Picea glauca* (Moench) Voss) (Van Ghelder et al. 2019). Controlling both disease and drought resistance, NLR-rich genomic regions will be particularly interesting candidates for genomic breeding purposes and genetic resource management in conifers, especially under climate change and the spread of tree pests and diseases.

Conifers are found in diverse ecosystems and several species are used in productive forestry across the globe, some of which involves breeding programs (Mullin et al. 2011). Breeding in conifers traditionally relies on pedigree analysis and phenotypic evaluations such as growth rates, wood yield and properties, and susceptibility to biotic and abiotic threats (White, Adams, and Neale 2007). Emerging and intensifying threats to the health of conifers in natural populations and managed forests involve a range of biotic stressors including oomycetes, fungi, herbivorous insects, and nematodes (Mitton and Ferrenberg 2012; Jakoby, Lischke, and Wermelinger 2019; Brar et al. 2018; Mota et al. 1999). In response to these challenges, molecular tools and genomic resources are being developed to support both fundamental research and diverse applications such as an acceleration of breeding outputs (Neale and Kremer 2011; Stocks et al. 2019; Bousquet et al. 2021). For example, investigations have linked genetic resistance to fusiform rusts in *Pinus taeda* L. to TNL-encoding sequences and to genomic clustering of resistance genes (Wilcox et al. 1996; Quesada et al. 2014; Ence et al. 2022). Genomic selection methods may have the potential to be developed to enhance resistance, such as rust resistance in *Pinus taeda* L. (Ence et al. 2022) and insect resistance in Norway spruce (*Picea abies* (L.) H.Karst.) (Lenz et al. 2020). However, the success of this relies on a more complete understanding of the genomic architecture of conifer NLRs (Ence et al. 2022).

The large size (often ≥ 10 Gbp) and relative complexity of conifer nuclear genomes has challenged whole genome sequence assembly (e.g., Nystedt et al. 2013; Birol et al. 2013; Zimin et al. 2014) but new methods have greatly improved the contiguity of genome assemblies as seen in *Sequoiadendron giganteum* (Lindl.) J.Buchholz (Scott et al. 2020), *Taxus chinensis* (Pilg.) Rehder (Xiong et al. 2021), and *Pinus tabuliformis* Carrière (Niu et al. 2022). In some species that lack such assemblies, high-density genetic maps are available to probe genome architecture (J.-J. Liu et al. 2019; Bernhardsson et al. 2019; Gagalova et al. 2022). Together, these assembled genome sequences and genetic maps, along with diverse transcriptomes (e.g., Van Ghelder et al. 2019) open the doors to more comprehensive analyses of resistance genes and their genomic architecture in conifer trees. Considering the relatively high level of genome conservation across conifers, we may predict that genomic clustering as observed in *P. flexilis* (J.-J. Liu et al. 2019) will occur across conifer taxa, that is to the extent they result from shared ancestral evolutionary events.

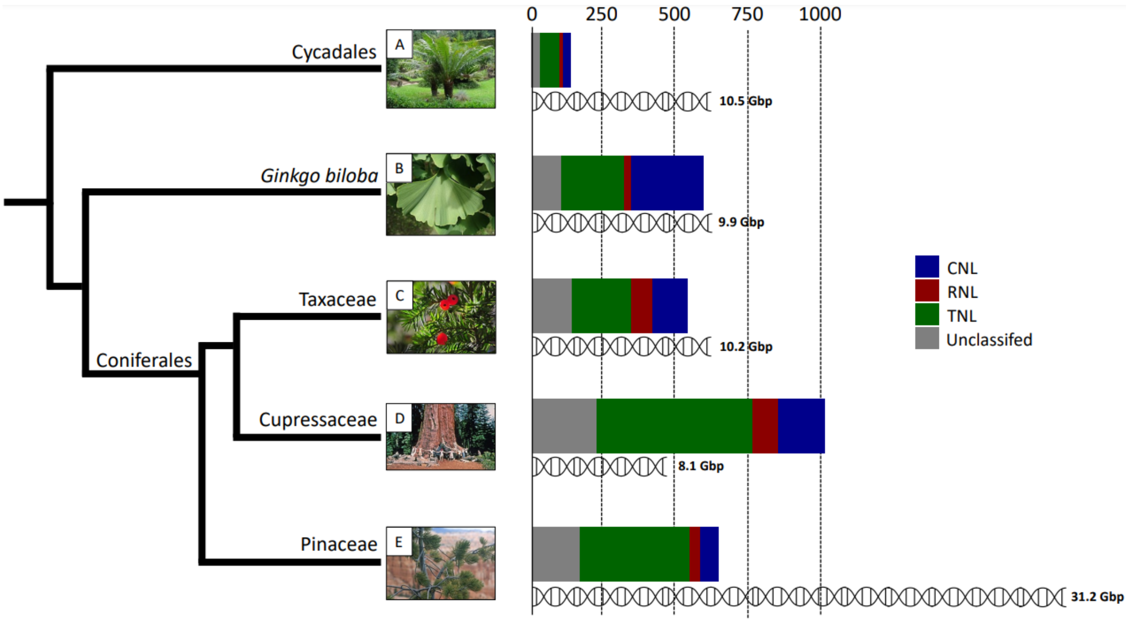

Figure 1: Overview of main gymnosperm clades, with total number of NLR genes and their corresponding categories indicated as bar charts, as found in this study (Table 1). Cladogram is based on the phylogeny presented in Leslie et al. (2018). Genome sizes are indicated beneath each bar chart and are based on the chromosome-level assemblies used in this study (Materials & methods section 1 for details) or, in the case of Pinaceae (*Pinus flexilis*), obtained from the Kew Plant DNA C-values database (release 7.1, Pellicer and Leitch 2020). Gymnosperms invariably have large genomes (~10 Gbp), but display large variations in NLR gene numbers. Despite the three-fold increase in genome size observed in Pinaceae, the number of discovered NLR genes remained within the average range of conifers. Pictures were obtained from the Wikimedia Commons repository (https://commons.wikimedia.org) and correspond to the broader taxonomic clades: A – *Cycas rumphii* Miq., andy_king50; B – *Ginkgo biloba* L., Susanna Giaccai; C – *Taxus baccata* L., Mykola Swarnyk; D – *Sequoiadendron giganteum*, W. Bulach; E – *Pinus flexilis*, Greg Woodhouse.

Considering the potential benefits of genomic breeding with NLR dense genome segments for conservation and industry, we investigated NLR gene clustering patterns across conifers. We leveraged recently published diploid high-density linkage maps and chromosome-level whole genome assemblies for genomic mapping of NLR genes. Results for conifers were contrasted with non-conifer gymnosperms from the Ginkgoales and Cycadales (Figure 1). To elucidate the evolutionary trajectories towards the observed clustering patterns, NLR genes were analysed in a phylogenetic framework. Our results indicate consistently uneven genomic distribution patterns of NLR genes across all conifers, with large and heavily concentrated reservoirs of NLR genes located on specific chromosomes. This knowledge on resistance genes will be informative for both breeding and conservation of these economically and ecologically important trees.

Materials & methods

1. Genomic distribution of NLR genes in Pinaceae and other conifer families

Following the evidence for dense genomic clusters of NLR immune receptor genes in *Pinus flexilis* (J.-J. Liu et al. 2019), we tested whether this is a Pinaceae family-wide phenomenon by examining publicly available genomic resources in other members of the family. To characterise the distribution of NLR genes in the speciose *Picea* genus, we deployed a high-density linkage map for *P. abies* (Bernhardsson et al. 2019) and high-quality genome assemblies of *P. glauca* and *P. sitchensis* (Gagalova et al. 2022). Although these assemblies do not have chromosome-level contiguity, they are suitably scaffolded into linkage groups corresponding to an updated version of the original *P. glauca* high-density linkage map (Pavy et al. 2017).

NLR genes were identified using the NLR Annotator pipeline v2.1 (Steuernagel et al. 2020) and linkage map positions were recorded to map NLR gene distribution on the twelve different linkage groups. We utilised the recent linkage map comparison work by Tumas et al. (2023) to determine the syntenic linkage groups. NLR distribution data for *P. flexilis* were taken from the original high-density linkage map publication (J.-J. Liu et al. 2019).

To compare distribution patterns across main lineages of conifers and gymnosperms, we applied the same pipeline to recently published chromosome-level genome assemblies: *Sequoiadendron giganteum* (Cupressaceae) (Scott et al. 2020), *Taxus chinensis* (Taxaceae) (Xiong et al. 2021), *Ginkgo biloba* (Ginkgoales) (H. Liu et al. 2021), *Pinus tabuliformis* (Niu et al. 2022), and *Cycas panzihuaensis* L.Zhou & S.Y.Yang (Cycadales) (Y. Liu et al. 2022). Nucleotide positions of NLR genes were recorded for each chromosome.

To visualise the distribution of NLR genes, histograms were produced in R (R Core Development Team 2010) with ggplot2 v3.4.2 (Wickham 2016) where genes were plotted along the length of chromosomes (linkage groups for linkage maps) using the starting position in Mbp (cM for linkage maps). The bin width was chosen to correspond roughly to 1% of the largest chromosome (or linkage group), i.e.: 13 Mbp for *C. panzihuaensis*, 12 Mbp for *G. biloba*, 10 Mbp for *S. giganteum* and *T. chinensis*, 4 Mbp for *Picea glauca*, 2 Mbp for *P. sitchensis* (Bong.) Carrière, 4 cM for *P. abies* (L.) H.Karst., and 2 cM for *Pinus flexilis*. We calculated the expected number of NLR genes for each chromosome (or linkage group) to quantify abnormal distribution patterns:

$$\#\text{[illegible]}_{\text{[illegible]}} = \frac{\text{[illegible]}}{\text{[illegible]}\ (\text{[illegible]})\ \text{[illegible]}\ (\text{[illegible]})} \times \text{[illegible]}\ (\text{[illegible]})\ \text{[illegible]}$$

The anomalies were visualised by plotting the observed values against the expected values in a scatter plot. To determine whether these anomalies were statistically significant we performed two-character Fisher's exact tests in R (R Core Team 2021) and computed the p-values.

2. NLR classification

The NLR Annotator pipeline describes the motifs discovered in each NLR based on a curated list of NLR motifs (Jupe et al. 2012). We utilised a custom python script (available on https://github.com/hung-th/NLRmeta), adapted from open-source code by Philipp Bayer (https://gist.github.com/philippbayer/0052f5ad56121cd2252a1c5b90154ed1) and based on the motif table in Jupe et al. (2012), to extract the motif output from NLR Annotator and convert it into CNL or TNL subfamily classification. The third subfamily, RNL, is characterised by the variable N-terminal RPW8 domain but is not annotated by NLR Annotator. In a previous study on conifer NLR genes, Van Ghelder et al. (2019) discovered two RNL-characteristic motifs located in the RNBS-D domain. We searched for these motifs in the generated NLR datasets by deploying the MAST software from the MEME suite (Bailey et al. 2015) and classified NLR genes containing these as RNLs. A further search for RPW8 domains was performed with tblastn (eV <0.05) using the conifer RPW8 sequences characterised by Van Ghelder et al. (2019) against the genome assemblies and linkage map loci. We only retained RPW8 hits with ≥1/3 amino acid identity in the reference sequences. NLRs with an RPW8 domain fused to the N-terminal side were thereby classified as RNLs regardless of their RNBS-D motif composition. Separate RPW8 sequences (e.g., not fused to an NLR gene) were also characterised as RNL genes. A fourth category of "unclassified" NLRs encompasses NLR genes that could not be classified into one of the three subfamilies due to a lack of characteristic domains and motifs. NLR genes containing motifs characteristic of CNL as well as TNL were also labelled as "unclassified". For *P. flexilis*, we utilised the annotation information from the original linkage map publication (J.-J. Liu et al. 2019) to divide NLRs into classes. RNLs in *P. flexilis* were classified in the same way as for the other species. NLR class information was used to further annotate the intrachromosomal NLR distribution histograms (Materials & methods section 1) to visualise patterns of class distribution.

NLR Annotator further determines whether detected NLR genes are complete, partial (missing domains) or pseudogenes (unexpected stop codon in sequence), which we recorded for each discovered gene, except for the separate RPW8 sequences.

3. NLR phylogenies

To determine the evolutionary history of discovered NLR genes, maximum likelihood phylogenies were generated from alignments of the central (conserved) NB-ARC domain. For *Pinus flexilis*, NB-ARC sequences were identified through a BlastP search (eV <0.05) of the reference NB-ARC sequence used by NLR Annotator against the translated NLR sequences. All hits with ≥60 amino acid residues were extracted and aligned using MAFFT v7 (Katoh and Standley 2013) using default settings. For the other species, we utilised the '-a' flag in NLR Annotator to obtain NB-ARC domains of all complete NLR genes. Maximum likelihood phylogenetics was performed with IQTree v1.6.12 (Nguyen et al. 2015) using the 'GTR20' model for protein evolution and 1000 ultra-fast bootstrap replicates to calculate node support values.

Phylogenies were visualised in the online Interactive Tree of Life (iTOL) tool (Letunic and Bork 2021) and rerooted at the node separating RNL/CNL and TNL clades. NLR class and chromosome were mapped onto the topologies using the iTOL annotation editor.

Results

NLR immune receptor gene diversity and abundance were annotated in the genomes of six conifer species (members of the Pinaceae, Cupressaceae and Taxaceae) and two other gymnosperms (*Ginkgo biloba* and a member of the Cycadales), with varying levels of contiguity. We deployed an automated NLR annotation pipeline and additional manual BLAST procedures on recently published high-density linkage maps and chromosome-level whole genome assemblies to physically map the genomic distribution of NLR genes in gymnosperms. We discovered consistent patterns of genomic clustering of NLR genes among conifers, but not in other gymnosperms. In each of the analysed conifers, a particularly dense cluster of NLR genes occurred on a single chromosome which contained between 18% and 34% of the total number of NLR genes within only a short segment of a few Mbp (or cM in the case of linkage maps).

We restricted our analysis to diploid genomes to avoid potential issues of false discovery for duplicated genes, due to incomplete phasing of the haplotypes in polyploid genome assemblies. We therefore omitted the genome assembly of the hexaploid *Sequoia sempervirens* Endl. (Neale et al. 2022) from our analysis.

Our scan of genome assemblies and high-density linkage maps for conifers and other gymnosperms did include the recently published mega-genome assembly (25.4 Gbp) of the diploid *Pinus tabuliformis* (Niu et al. 2022), where some chromosomes surpass 2 Gbp in length. Unfortunately, the NLR Annotator pipeline (Steuernagel et al. 2020) is currently insufficiently programmed for such large contigs, limiting the output of our NLR analysis on this genome assembly. Furthermore, our preliminary results indicated extremely high NLR numbers (>4000) in this assembly, which we hypothesise to be an overestimation, considering the large disjunction with other conifers (1,002 in *S. giganteum*) and members of the *Pinus* genus (639 in *P. flexilis*). We therefore omitted *P. tabuliformis* from further analysis.

1. <u>Gymnosperm NLR abundance and diversity</u>

The number of NLR genes discovered in conifer genomic datasets varied nearly two-fold (Figure 1), between 533 (*Taxus chinensis*) and 1002 (*Sequoiadendron giganteum*). Of the non-conifers, *Ginkgo biloba* contained 585 NLR genes, which is similar to the average found in conifers (Table 1). In contrast, there were only 136 NLR genes found in the Cycad genome. Of the conifers, the *Sequoiadendron giganteum* genome contained the highest number of NLR genes (1,002) followed by (*Pinus flexilis*) which had considerably less (639). In all conifers, the TNL class was the most abundant, contributing between 33% (81/245, *Picea sitchensis*) to 59% (376/639, *Pinus flexilis*) to the total number of NLR genes (Table 1). In contrast, the CNL class was most abundant in *Ginkgo biloba*, in which they contributed 43% to the total number of NLR genes. RNLs were more abundant in conifers than in other gymnosperms. This class was proportionally most abundant in *Picea* species (16-20%), with half of the RNLs in *P. glauca* and *P. sitchensis* consisted only of RPW8 domains. The number of unclassified NLR genes (missing or ambiguous C-terminal domain) grew proportionally with the total number of NLR genes, representing 16-25% of the genes in complete genomic datasets.

Table 1: Number of NLR genes discovered in each taxon of this study, separated by class. As some NLR genes were located on unassigned scaffolds in *C. panzhihuaensis*, *G. biloba* and *T. chinensis*, the numbers of NLR genes mapped onto chromosomes are recorded separately in brackets alongside the total number of NLR genes discovered for these species. Some RNL comprised genes only in the RPW8 domain, and these are recorded separately in brackets for each taxon. **P. flexilis* values are based on the results from J.-J. Liu et al (2019). **Values for *Picea* taxa are based on NLR annotation of high-density linkage maps or incomplete genome assemblies that were built using high-density linkage maps (more details are provided in Materials & methods sections section 1 & 2). ^Listed only for the chromosome with the highest NLR-density. The method for calculating the ratio of observed /predicted NLR clustering on a single chromosome is explained in Materials & methods section 1.

Species	*Total #NLR (on chr.)*	*#CNL*	*#RNL (RPW8 only)*	*#TNL*	*Uncl.*	*Densest chr. (% of total)*	*#Observed /#Predicted^*	*p-value (distrib.)*
Cycas panzhihuaensis	136 (127)	29	13 (8)	65	29	24 (19%)	2.02	0.02958
Ginkgo Biloba	585 (570)	249	25 (5)	213	98	83 (14%)	2.04	6.70e-08
Taxaceae								
Taxus chinensis	533 (496)	124	73 (29)	201	135	146 (29%)	4.30	< 2.2e-16
Cupressaceae								
Sequoiadendron giganteum	1,002	159	89 (17)	533	221	312 (31%)	2.47	< 2.2e-16
Pinaceae								
*Pinus flexilis**	639	64	36 (-)	376	163	266 (42%)	3.55	< 2.2e-16
*Picea abies***	173	37	30 (5)	85	21	68 (39%)	3.37	4.78e-05
*Picea glauca***	273	74	43 (21)	105	51	89 (33%)	3.26	1.07e-07
*Picea sitchensis***	245	73	50 (26)	81	41	57 (23%)	2.61	0.003825

2. Chromosomal and Intrachromosomal NLR distributions

In each of the conifer species we analysed, one chromosome displayed a disproportionately high NLR content (Table 1 and Figure 2A), a phenomenon that was absent in non-coniferous gymnosperms. In conifers, the chromosome which displayed the highest clustering of NLR genes contained between 29% (*T. chinensis*) and 42% (*P. flexilis*) of the total number of NLR genes found in the respective genome (Table 1). Even when there was incomplete genome coverage, the results for *Picea* species were within this range, the only outlier being *P. sitchensis* at 23%. Large numbers of NLR genes form dense clusters on these NLR-rich chromosomes or linkage groups (Figure 2B-E). These clusters contain high proportions of the total amount of NLR genes found in the genomic dataset, with up to 34% (*Pinus flexilis*, Figure 2C) of NLR genes concentrated in the space of 21% (44 cM) of a chromosome (linkage group).

Although the distribution of NLR genes was non-random in all gymnosperms analysed, it was considerably less clustered in *C. panzhihuaensis* and *G. biloba* compared to the full genomic datasets of the conifers we analysed (Table 1). In the non-conifer species, only one small cluster of 13 genes (10% of total) was found on chromosome #10 in *C. panzhihuaensis* (Figure 3) and a small cluster of 45 genes (7.9% of total) occurred on chromosome #3 in *G. biloba*, both proportionally smaller than the clusters in conifer species (on average 25% of total). Non-uniform distribution of NLR genes in conifers did not only occur between but also within chromosomes (Figure 3) with large regions of NLR-rich chromosomes being devoid of NLR genes, particularly in *Taxus chinensis*.

3. Intragenomic diversification and evolution of NLR genes

Phylogenetic relationships between intragenomic NLR genes followed expected class distinctions in all conifers and *Cycas panzhihuaensis* but not in *Ginkgo biloba* (Figure 4). Although conifer TNLs show a strong overall monophyletic correlation, *Taxus chinensis* (Fig. 4D) is the only gymnosperm where all TNLs share one most recent common ancestor (MRCA). *Sequoiadendron giganteum* (Fig. 4C) even has a small monophyletic clade of TNLs nested within the larger CNL/RNL clade. In all non-cycad gymnosperms, there is at least one monophyletic RNL clade with a different MRCA than the main CNL clade. All conifer CNL clades contained a few derived RNL sequences restricted to one phylogenetic subclade of the CNL clade. Both the *T. chinensis* and *S. giganteum* genomes contain a smaller monophyletic clade with a unique MRCA that contains almost exclusively unclassified NLR genes.

The densest clusters in conifers are mainly composed of TNLs (Fig. 2B-E), which correlates with their intrachromosomal diversification (Fig. 4G). Most TNLs that occur on the same chromosome are also phylogenetically correlated. In the Pinaceae this phylogenetically correlated diversification even occurred on the same syntenic linkage group (Supporting information).

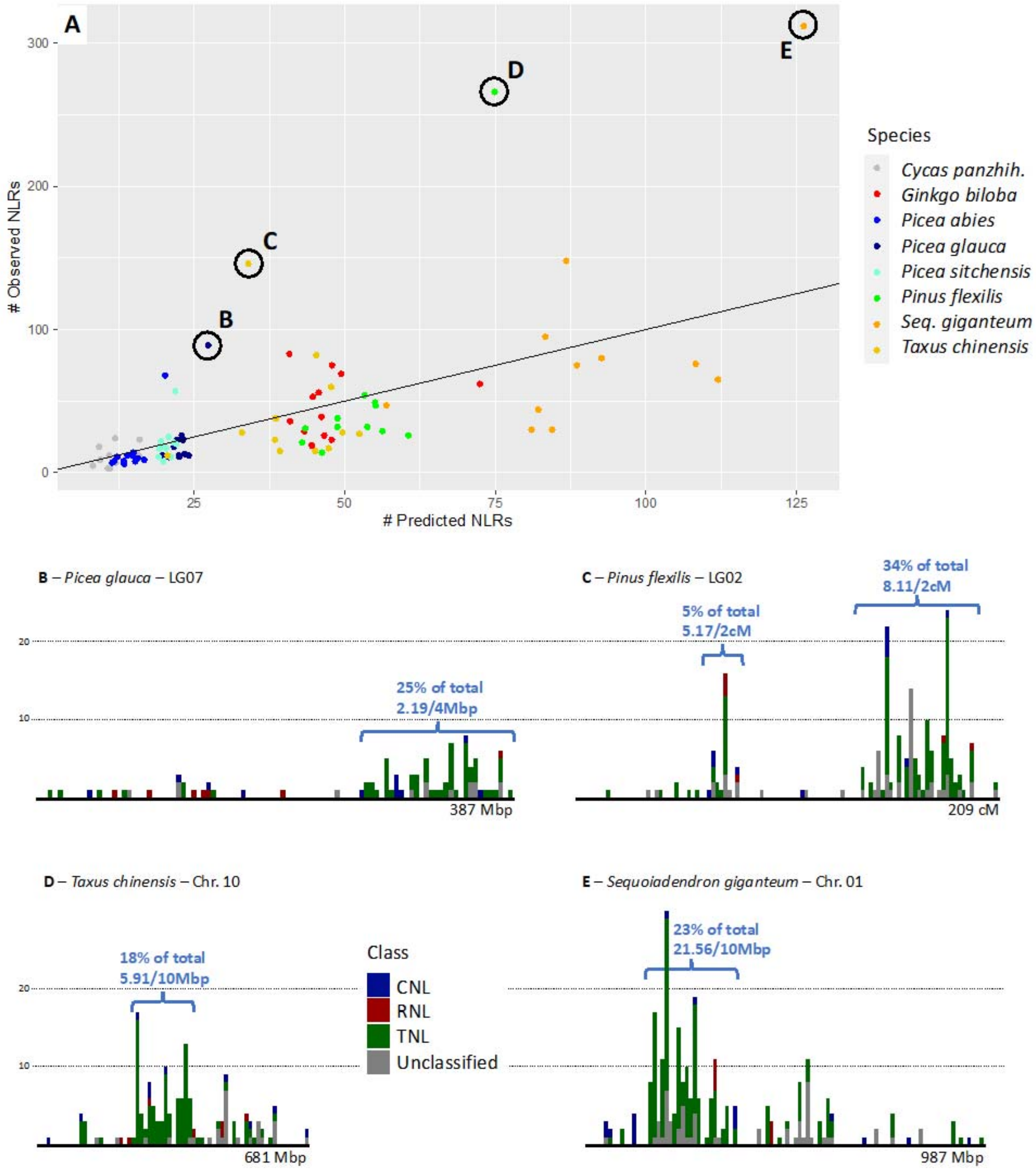

Figure 2: Chromosomal distribution of NLR genes across gymnosperms. A – Scatter plot indicating the observed number of NLR genes versus the expected number of NLR genes (calculated based on the length of the chromosome and the total number of NLR genes discovered in the genome, Materials & methods section 1) for each chromosome in each taxon analysed in this study. Black line follows the function y=x, indicating a perfectly homogeneous distribution of NLR genes over the chromosomes. Deviations from this line therefore indicate a non-homogeneous distribution. Highly deviant chromosomes of four taxa are highlighted and have their intrachromosomal NLR distribution displayed in histograms (B-E). Bin width equals ±1% of the length of the largest chromosome in the genome of the respective taxon (Materials & methods section 1). Colours indicate NLR class as determined with NLR Annotator (Steuernagel et al. 2020) and manual BLASTs (Materials & methods section 2). Ultra-dense NLR clusters are indicated for each taxon.

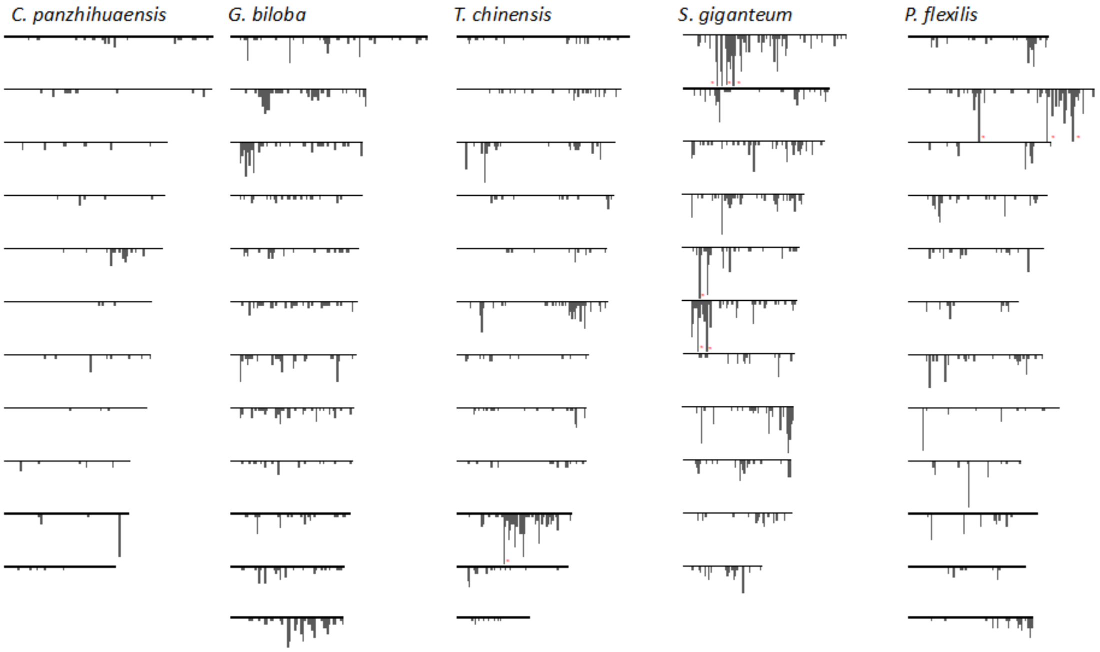

Figure 3: Histogram plots of NLR genes on each chromosome within the genome of five gymnosperm lineages. Chromosomes are ordered based on the ordering in the respective assembly and do not reflect synteny. Red ***** indicates that a particularly dense bin has been cropped manually in order to fit all the histograms into a single image.

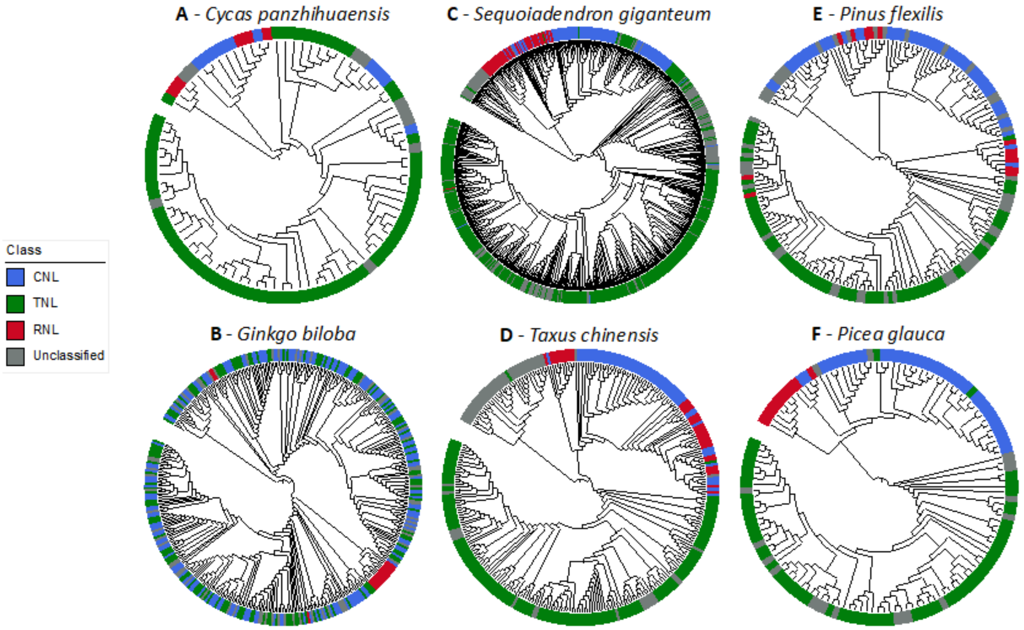

Figure 4: Intragenomic phylogenetic relationships of NLR genes based on the conserved central NB-ARC domain, calculated with maximum likelihood algorithms using IQTree v1.6.12 (Nguyen et al. 2015) (Materials & methods section 3) and annotated using the iTOL web server (Letunic and Bork 2021). Colour strips around the circular trees indicate NLR class as determined with NLR Annotator (Steuernagel et al. 2020) and manual BLASTs (Materials & methods section 2). Main gymnosperm lineages represented by Cycadales (A), Ginkgoales (B), Cuppressaceae (C), Taxaceae (D), and Pinaceae: *Pinus* (E), and *Picea* (F).

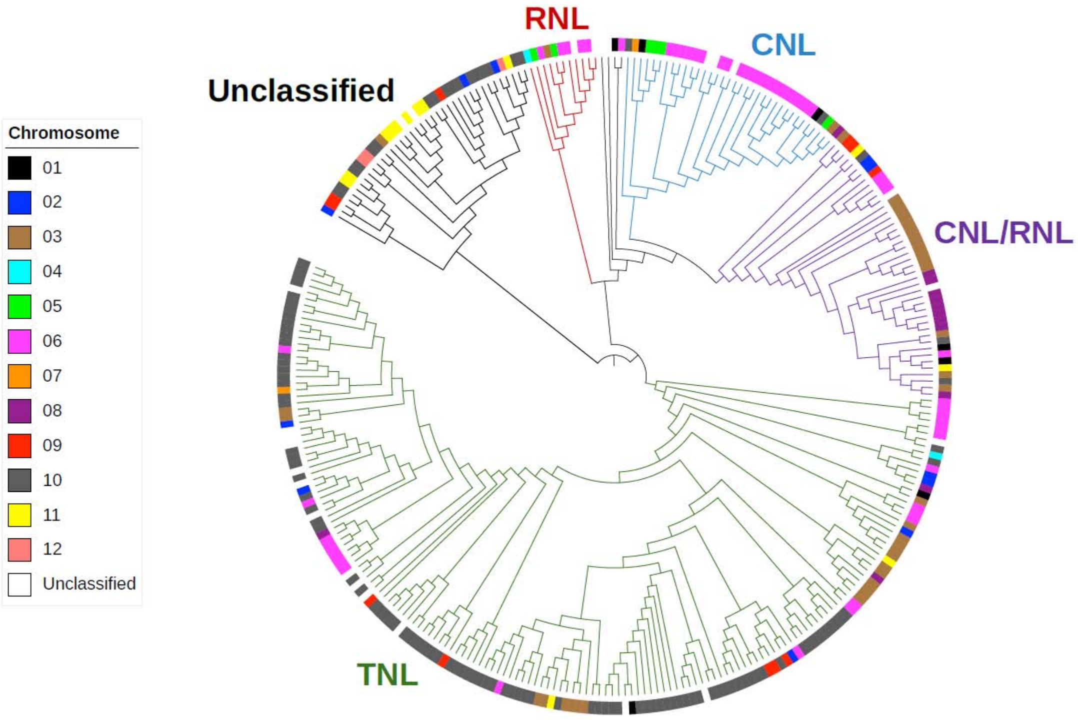

Figure 5: Chromosomal structuring in phylogenetic relationships within a conifer species (*Taxus chinensis*) displayed in an intragenomic framework. Branch colours correspond to the NLR gene subfamily, as indicated in the same colours next to the phylogeny. Coloured squares at the tips of branches represent the chromosome on which the respective NLR genes are located. NLR genes found on scaffolds that were not assembled into chromosomes are indicated with empty colour squares ("Unclassified").

Discussion

The study of NLR immune receptor evolution and genomic architecture in conifers is motivated in part by the increased threats from diverse biotic aggressors. The 615 species of extant conifers span all continents except Antarctica and are classified in eight families with the largest being Pinaceae, Cupressaceae, and Podocarpaceae (Farjon and Filer 2013). Forest trees, including conifers, have several co-evolved biotic aggressors ranging from rust diseases to herbivorous insects which may damage or even kill trees. However, more severe attacks, range expansions or species introductions, and infection of previously unknown hosts have become increasingly prevalent and linked to climate change (Teshome, Zharare, and Naidoo 2020). Major herbivorous insects are expanding their range and intensifying damage levels in conjunction with climate change, such as mountain pine beetles (e.g. Mitton and Ferrenberg 2012). New diseases or outbreaks in several conifers are being linked to Oomycetes, namely species of *Phytophthora* de Barys (e.g., Brasier and Webber 2010, Brar et al. 2018). There is growing evidence of the involvement of TNLs in resistance particularly to rust diseases affecting pines (Quesada et al. 2014; Amerson et al. 2015; Ence et al. 2022), but much less is known when it comes to insects and newly emerging biotic threats. Understanding the genomic architecture of NLR genes will therefore be critical for informing breeding programs and other conservation practices seeking to mitigate the effects of climate change.

Ubiquitous genomic clustering of NLR genes in conifers

Through comparative genome-wide analysis of NLR immune receptor genes in conifers, we elucidated a conserved non-random chromosomal distribution of NLR genes. In all conifer genomes analysed here we found one chromosome containing a disproportionately large number of NLR genes (Figure 2). Without exception, these chromosomes contained dense clusters of NLR genes (Figure 3), comprising on average a quarter of the total number of NLR genes in the genome (Figure 2B-E). These clusters predominantly contain TNL genes, which share an ancestral origin (Figure 4 & 5). High-density gene regions are a hallmark of conifer genomes (Pavy et al. 2017), hinting at tandem duplications, as detected frequently for NLR genes (Pavy et al. 2017) and for other conifer gene families (e.g. Guillet-Claude et al. 2004). We therefore conclude that the unbalanced genomic clustering pattern of NLR genes found previously in *Pinus flexilis* (J.-J. Liu et al. 2019) occurs in a taxonomically broad range of other conifer species. The absence of this pattern in non-coniferous gymnosperms (Figure 3 & Figure 4A-B) indicates that NLR gene clusters arose in the ancestor of conifers. Diversification within these large groups of sequences appears to be variable and largely lineage-specific (such as in the Pinaceae). Partial functional redundancy is likely, in response to similar environmental cues and selection pressures among conifer taxa, thus contributing to the high abundance of NLR genes in conifers. These ancestral clusters may have enabled lineage-specific tandem duplications leading to the high (and variable) abundance of NLR genes observed in conifers (Van Ghelder et al. 2019; this study), as compared to non-coniferous plants such as cycads (Table 1, Figure 4A) and many angiosperms. A high diversity of NLR genes may lead to a high versatility in drought and disease resistance for these woody perennials characterised by delayed sexual maturity for reproduction. These genes are likely to contribute to the ecological dominance of conifers in many boreal and temperate forests (Bonello et al. 2006), including in highly inhospitable habitats (Laberge et al. 2000).

Genomic clusters of NLR genes have previously been found in a variety of angiosperm lineages (Wersch and Li 2019), such as *Arabidopsis* Heinh. (Brassicaceae) (Meyers et al.

2003), lettuce (Asteraceae) (Christopoulou et al. 2015), peach (Rosaceae) (Verde et al. 2013), potatoes (Solanaceae) (Seo et al. 2016), and wheat (Poaceae) (Smith et al. 2007). In contrast to the studied conifers, NLR clustering is frequent but not as ubiquitous in angiosperms. What further sets conifers apart from angiosperms in regards to NLRs, is the ubiquitous presence and abundance of all three NLR subfamilies (CNLs, RNLs and TNLs) (Van Ghelder et al. 2019, Table 1 of this study). RNL abundance is rare in angiosperms and TNLs are absent in monocots (Van Ghelder et al. 2019). We found RNLs comprised 4-14% of the total NLR diversity in all gymnosperms, indicating that RNL abundance is an ancestral trait to the gymnosperms. RNLs were consistently divided over two phylogenetic clades, one of which comprised mainly CNLs (Figure 4). This is consistent with an evolutionary divergence between TNLs and CNL/RNLs predating that between CNLs and RNLs (Shao et al. 2019). Interestingly, CNLs and TNLs shared ancestral origins in *Ginkgo biloba* (Figure 4B). This potentially indicates frequent domain swapping between NLR genes, highlighting the dynamic nature of these resistance genes.

Evolution of genomic architecture in conifer giga-genomes

Conifers have very large genomes (18-34 Gb) and harbour large amounts of repetitive DNA sequence (Mackay et al. 2012; Nystedt et al. 2013; Birol et al. 2013; Zimin et al. 2014; De La Torre et al. 2014). Genome evolution is considered to be less dynamic in conifers and other gymnosperms compared to flowering plants (Leitch and Leitch. 2012), which may suggest lower rates of gene diversification. On the other hand, conifers have suites of rapidly evolving genes (Gagalova et al. 2022) and highly diversified gene families or sub-families (e.g., Bedon et al. 2010; Stival Sena et al. 2018; Van Ghelder et al. 2019) both related to stimuli and stress response. Several comparative studies in conifers have shown high levels of intergeneric macro-synteny and macro-collinearity among *Pinaceae* taxa (Pavy et al. 2012) (Pelgas et al. 2006; Ritland et al. 2011; Westbrook et al. 2015) and clear chromosomal rearrangements when comparing Pinaceae and Cupressaceae (Moriguchi et al. 2012, de Miguel et al. 2015). These observations are consistent with a small number of whole genome duplications early in conifer evolution (Li et al. 2015). In contrast, our study has focused on the genomic architecture of a targeted gene family, showing conserved localised clustering and shedding insights into the evolutionary trajectory of NLR genes. Our investigation was possible due to two types of relatively recent genomic resources: 1) highly contiguous genome assemblies, e.g., *Sequoiadendron giganteum* (Lindl.) J.Buchholz (Scott et al. 2020), *Taxus chinensis* (Pilg.) Rehder (Xiong et al. 2021), among others, which are developed using proximity ligation (e.g., HiC (Belton et al. 2012)) and long-read sequencing (e.g., PacBio SMRT); 2) high density genetic maps which are available for *Pinus* L. *spp.* (e.g., J.-J. Liu et al. 2019), and *Picea* A.Dietr. *spp.*(Bernhardsson et al. 2019; Gagalova et al. 2022; Tumas et al. 2023).

A specific feature of conifer genomes probably enabled the accumulation of NLR genes and facilitated the formation of very large gene clusters. Conifers are inefficient at removing extra copies of DNA sequence through proof-editing, hence their propensity to accumulate gigabases of repetitive sequences such as the type I transposable elements (e.g. copia and gypsy sequences) (Nystedt et al. 2013; Zimin et al. 2014). Conifers also retain a high proportion of pseudogenes (Warren et al. 2015) and single copy sequences that are similar to protein coding genes (Pellicer et al. 2018). Therefore, we could expect a significant proportion of genomic NLR sequences in conifers to represent pseudogenes. However, RNA sequencing has identified between 271 and 725 NLR genes expressed across a suite of Pinaceae and Cupressaceae species (Van Ghelder et al. 2019; Liu et al., 2021; Ence et al., 2022). Expression of selected NLRs was shown to be responsive to infection by *Phytophthora ramorum* in *Larix spp.* (Dun et al. 2022) or drought in *Picea glauca* (Van Ghelder et al. 2019) and to be variable across different seed families in *Pinus flexilis* (J.-J. Liu et al. 2021). The genomic sequences identified here may prompt further work to determine which of these are expressed and under what circumstances. Studies of functional divergence among sequences and of evolutionary rates, aiming at identifying footprints of natural selection (e.g., Guillet-Claude et al. 2004) should also be considered (e.g., Chia and Carella 2023).

Opportunities for genomic breeding using NLR genes for pest, disease and drought resistance

Resistance to viruses, bacteria, oomycetes, fungi and some insects has been linked to NLR genes in a range of flowering plants (Kourelis and van der Hoorn 2018) but our understanding of their contribution to resistance in conifers is very rudimentary. Dissection of the genetic resistance to fusiform rusts (*Cronartium quercuum* (Berk.) Miyabe ex Shirai f.sp. *fusiforme*) in *Pinus taeda* L. breeding populations has provided evidence for genetic resistance (Wilcox et al. 1996) and implicated NLR-encoding genes as likely candidates (Quesada et al. 2014; Ence et al. 2022). In *P. taeda*, nine different fusiform rust resistance loci were identified across three linkage groups (Amerson et al. 2015). These are hypothesised to contain NLR sequences, which were found to vary in numbers of genomic sequences when comparing populations from different geographic areas (Ence et al. 2022). Similarly, in *Pinus flexilis*, fine genetic dissection, evolutionary analysis, and expression profiling have identified two NLR genes as candidates for resistance linked to the Cr4 locus (J.-J. Liu et al. 2021) among 155 NLR genes mapped across 12 linkage groups to date. Interestingly, the major clusters found in *P. flexilis* were on linkage groups distinct to the Cr4 locus. Taken together, these results indicate that resistance-associated NLR genes may be distributed across the genome. To our knowledge, resistance phenotypes have been identified in only a few studies in conifers and these have involved fungal rusts infecting pines, although expression profiling indicated responsiveness to *Phytophthora ramorum* (Dun et al. 2022) and drought in other species (Van Ghelder et al. 2019).

NLR-dense genomic regions could act as potential reservoirs for NLR diversity. The NLR clusters found in gymnosperm genomes probably arose by tandem duplications of TNL genes, as indicated by their consistent composition in conifers (Figure 2B-D) and the close phylogenetic relationships of the TNLs on the same chromosome (Figure 5). Although tandem duplication inevitably leads to identical gene copies initially (paralogs), it can eventually lead to gene diversification through differential mutation trajectories and domain swapping (Ostermeier and Benkovic 2001). There are documented examples of such neofunctionalisation for resistance genes in different plant lineages (Kong and Ranganathan 2008; Wei et al. 2023) and for genes encoding transcription factors in the Pinaceae (Guillet-Claude et al. 2004). Given the long evolutionary history of conifer lineages (Leslie et al. 2018) and their considerable NLR diversity, a high degree of non-canonical NLR genes is expected, especially in and around these dense NLR gene clusters. Considering whole-genome datasets, about one in four NLR genes is unclassified in conifers (Table 1), meaning that a distinctive N-terminal domain is missing or ambiguous (e.g., fusion of TIR and CC domains). Closer inspection of NLR-dense regions could therefore reveal interesting non-canonical domains in conifer NLR genes. Together with the overall diversity of canonical NLR genes in these clusters, this could further emphasise their potential in genomics-assisted breeding to improve disease and drought resistance.

The NLR gene family is among the most studied in plants due to its agronomic importance (Kourelis and van der Hoorn, 2018) with many linked breeding and genetic engineering applications proposed to improve resistance in crops (van Wersch et al. 2020) and to a lesser extent in forest trees (Ence et al. 2022). We have shown how improved genome sequences along with transcriptome data may enhance our understanding of NLR genomic architecture in this understudied group. In order to develop genetic resources that will help to respond to emerging threats in conifers, three other components are needed: 1) populations of phenotypically diverse individuals in which to study resistance traits (Ence et al. 2022; Liu et al., 2019); 2) efficient and accurate assessment of the susceptibility and resistance phenotypes to relevant pests and diseases, which is difficult to accomplish and is therefore often either lacking or sub-optimal in conifers; 3) fast and accurate genome scanning methods that are suitable for differentiating among genes and alleles within and among populations. The large size and the variability of the NLR gene family adds to the challenge of linking genes to resistance phenotypes; however, knowledge of the position of clusters of these immune receptor genes paves the way to more focused investigations in conjunction with genome selection and other genome-wide analyses.

Acknowledgements

We are grateful for feedback to this study and its results from the Sitka Spruced (https://sitkaspruced.web.ox.ac.uk/) and Spruce-Up communities (https://spruce-up.ca/en/). As part of the Sitka Spruced project (BB/P020488/1), this research was funded by BBSRC and a group of forest and wood processing industries.

Competing interests

None declared.

Author contributions

YW & JJM designed the study. YW conducted the bioinformatic analyses with input from HT, CVG & THH. YW & JJM wrote the manuscript with input from all authors.

Data availability

Genomic mapping data of NLR gene distributions generated for this study is available in the Figshare digital repository under the following DOI: 10.6084/m9.figshare.24412579

References

Amerson, Henry V., C. Dana Nelson, Thomas L. Kubisiak, E. George Kuhlman, and Saul A. Garcia. 2015. 'Identification of Nine Pathotype-Specific Genes Conferring Resistance to Fusiform Rust in Loblolly Pine (*Pinus taeda* L.)'. *Forests* 6 (8): 2739–61. https://doi.org/10.3390/f6082739.

Bailey, Timothy L., James Johnson, Charles E. Grant, and William S. Noble. 2015. 'The MEME Suite'. *Nucleic Acids Research* 43 (W1): W39–49. https://doi.org/10.1093/nar/gkv416.

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.

Belton, Jon-Matthew, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. 2012. 'Hi–C: A Comprehensive Technique to Capture the Conformation of Genomes'. *Methods*, 3D chromatin architecture, 58 (3): 268–76. https://doi.org/10.1016/j.ymeth.2012.05.001.

Bernhardsson, Carolina, Amaryllis Vidalis, Xi Wang, Douglas G Scofield, Bastian Schiffthaler, John Baison, Nathaniel R Street, M Rosario García-Gil, and Pär K Ingvarsson. 2019. 'An Ultra-Dense Haploid Genetic Map for Evaluating the Highly Fragmented Genome Assembly of Norway Spruce (*Picea abies*)'. *G3 Genes|Genomes|Genetics* 9 (5): 1623–32. https://doi.org/10.1534/g3.118.200840.

Biezen, Erik A. van der, and Jonathan D. G. Jones. 1998. 'The NB-ARC Domain: A Novel Signalling Motif Shared by Plant Resistance Gene Products and Regulators of Cell Death in Animals'. *Current Biology* 8 (7): R226–28. https://doi.org/10.1016/S0960-9822(98)70145-9.

Birol, Inanc, Anthony Raymond, Shaun D. Jackman, Stephen Pleasance, Robin Coope, Greg A. Taylor, Macaire Man Saint Yuen, et al. 2013. 'Assembling the 20 Gb White Spruce (Picea Glauca) Genome from Whole-Genome Shotgun Sequencing Data'. *Bioinformatics* 29 (12): 1492–97. https://doi.org/10.1093/bioinformatics/btt178.

Bonello, Pierluigi, Thomas R. Gordon, Daniel A. Herms, David L. Wood, and Nadir Erbilgin. 2006. 'Nature and Ecological Implications of Pathogen-Induced Systemic Resistance in Conifers: A Novel Hypothesis'. *Physiological and Molecular Plant Pathology* 68 (4): 95–104. https://doi.org/10.1016/j.pmpp.2006.12.002.

Brar, S., J. F. Tabima, R. L. McDougal, P.-Y. Dupont, N. Feau, R. C. Hamelin, P. Panda, et al. 2018. 'Genetic Diversity of Phytophthora Pluvialis, a Pathogen of Conifers, in

New Zealand and the West Coast of the United States of America’. *Plant Pathology* 67 (5): 1131–39. https://doi.org/10.1111/ppa.12812.

Brasier, Clive, and Joan Webber. 2010. ‘Sudden Larch Death’. *Nature* 466 (7308): 824–25. https://doi.org/10.1038/466824a.

Chen, Jing, Yanxiao Zhao, Xuanjie Luo, Hao Hong, Tongqing Yang, Shen Huang, Chunli Wang, et al. 2023. ‘NLR Surveillance of Pathogen Interference with Hormone Receptors Induces Immunity’. *Nature* 613 (7942): 145–52. https://doi.org/10.1038/s41586-022-05529-9.

Chia, Khong-Sam, and Philip Carella. 2023. ‘Taking the Lead: NLR Immune Receptor N-Terminal Domains Execute Plant Immune Responses’. *New Phytologist* n/a (n/a). https://doi.org/10.1111/nph.19170.

Christopoulou, Marilena, Sebastian Reyes-Chin Wo, Alex Kozik, Leah K McHale, Maria-Jose Truco, Tadeusz Wroblewski, and Richard W Michelmore. 2015. ‘Genome-Wide Architecture of Disease Resistance Genes in Lettuce’. *G3 Genes|Genomes|Genetics* 5 (12): 2655–69. https://doi.org/10.1534/g3.115.020818.

De La Torre, Amanda R., Inanc Birol, Jean Bousquet, Pär K. Ingvarsson, Stefan Jansson, Steven J.M. Jones, Christopher I. Keeling, et al. 2014. ‘Insights into Conifer Giga-Genomes’. *Plant Physiology* 166 (4): 1724–32. https://doi.org/10.1104/pp.114.248708.

Dun, Heather F., Tin Hang Hung, Sarah Green, and John J. MacKay. 2022. ‘Comparative Transcriptomic Responses of European and Japanese Larches to Infection by Phytophthora Ramorum’. *BMC Plant Biology* 22 (1): 480. https://doi.org/10.1186/s12870-022-03806-3.

Duxbury, Zane, Chih-hang Wu, and Pingtao Ding. 2021. ‘A Comparative Overview of the Intracellular Guardians of Plants and Animals: NLRs in Innate Immunity and Beyond’. *Annual Review of Plant Biology* 72 (1): 155–84. https://doi.org/10.1146/annurev-arplant-080620-104948.

Ence, Daniel, Katherine E Smith, Shenghua Fan, Leandro Gomide Neves, Robin Paul, Jill Wegrzyn, Gary F Peter, et al. 2022. ‘NLR Diversity and Candidate Fusiform Rust Resistance Genes in Loblolly Pine’. *G3 Genes|Genomes|Genetics* 12 (2): jkab421. https://doi.org/10.1093/g3journal/jkab421.

Farjon, Aljos, and Denis Filer. 2013. ‘An Atlas of the World’s Conifers: An Analysis of Their Distribution, Biogeography, Diversity and Conservation Status’. In *An Atlas of the World’s Conifers*. Brill. https://brill.com/display/title/20587.

Gagalova, Kristina K., René L. Warren, Lauren Coombe, Johnathan Wong, Ka Ming Nip, Macaire Man Saint Yuen, Justin G. A. Whitehill, et al. 2022. ‘Spruce Giga-Genomes: Structurally Similar yet Distinctive with Differentially Expanding Gene Families and Rapidly Evolving Genes’. *The Plant Journal* 111 (5): 1469–85. https://doi.org/10.1111/tpj.15889.

Jakoby, Oliver, Heike Lischke, and Beat Wermelinger. 2019. ‘Climate Change Alters Elevational Phenology Patterns of the European Spruce Bark Beetle (Ips Typographus)’. *Global Change Biology* 25 (12): 4048–63. https://doi.org/10.1111/gcb.14766.

Jia, YanXiao, Yang Yuan, Yanchun Zhang, Sihai Yang, and Xiaohui Zhang. 2015. ‘Extreme Expansion of NBS-Encoding Genes in Rosaceae’. *BMC Genetics* 16 (1): 48. https://doi.org/10.1186/s12863-015-0208-x.

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.

Katoh, Kazutaka, and Daron M. Standley. 2013. ‘MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability’. *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Kong, Lesheng, and Shoba Ranganathan. 2008. ‘Tandem Duplication, Circular Permutation, Molecular Adaptation: How Solanaceae Resist Pests via Inhibitors’. *BMC Bioinformatics* 9 (1): S22. https://doi.org/10.1186/1471-2105-9-S1-S22.

Kourelis, Jiorgos, and Renier A.L. van der Hoorn. 2018. ‘Defended to the Nines: 25 Years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function’. *The Plant Cell* 30 (2): 285–99. https://doi.org/10.1105/tpc.17.00579.

Lenz, Patrick R. N., Simon Nadeau, Marie-Josée Mottet, Martin Perron, Nathalie Isabel, Jean Beaulieu, and Jean Bousquet. 2020. ‘Multi-Trait Genomic Selection for Weevil Resistance, Growth, and Wood Quality in Norway Spruce’. *Evolutionary Applications* 13 (1): 76–94. https://doi.org/10.1111/eva.12823.

Leslie, Andrew B., Jeremy Beaulieu, Garth Holman, Christopher S. Campbell, Wenbin Mei, Linda R. Raubeson, and Sarah Mathews. 2018. ‘An Overview of Extant Conifer Evolution from the Perspective of the Fossil Record’. *American Journal of Botany* 105 (9): 1531–44. https://doi.org/10.1002/ajb2.1143.

Letunic, Ivica, and Peer Bork. 2021. ‘Interactive Tree Of Life (iTOL) v5: An Online Tool for

Phylogenetic Tree Display and Annotation'. *Nucleic Acids Research* 49 (W1): W293–96. https://doi.org/10.1093/nar/gkab301.

Li, Zheng, Anthony E. Baniaga, Emily B. Sessa, Moira Scascitelli, Sean W. Graham, Loren H. Rieseberg, and Michael S. Barker. 2015. 'Early Genome Duplications in Conifers and Other Seed Plants'. *Science Advances* 1 (10): e1501084. https://doi.org/10.1126/sciadv.1501084.

Lin, Xiao, Yu Zhang, Hanhui Kuang, and Jiongjiong Chen. 2013. 'Frequent Loss of Lineages and Deficient Duplications Accounted for Low Copy Number of Disease Resistance Genes in Cucurbitaceae'. *BMC Genomics* 14 (1): 335. https://doi.org/10.1186/1471-2164-14-335.

Liu, Hailin, Xiaobo Wang, Guibin Wang, Peng Cui, Shigang Wu, Cheng Ai, Nan Hu, et al. 2021. 'The Nearly Complete Genome of Ginkgo Biloba Illuminates Gymnosperm Evolution'. *Nature Plants* 7 (6): 748–56. https://doi.org/10.1038/s41477-021-00933-x.

Liu, Jun-Jun, Anna W. Schoettle, Richard A. Sniezko, Holly Williams, Arezoo Zamany, and Benjamin Rancourt. 2021. 'Fine Dissection of Limber Pine Resistance to Cronartium Ribicola Using Targeted Sequencing of the NLR Family'. *BMC Genomics* 22 (1): 567. https://doi.org/10.1186/s12864-021-07885-8.

Liu, Jun-Jun, Anna W. Schoettle, Richard A. Sniezko, Fupan Yao, Arezoo Zamany, Holly Williams, and Benjamin Rancourt. 2019. 'Limber Pine (Pinus Flexilis James) Genetic Map Constructed by Exome-Seq Provides Insight into the Evolution of Disease Resistance and a Genomic Resource for Genomics-Based Breeding'. *The Plant Journal* 98 (4): 745–58. https://doi.org/10.1111/tpj.14270.

Liu, Yang, Sibo Wang, Linzhou Li, Ting Yang, Shanshan Dong, Tong Wei, Shengdan Wu, et al. 2022. 'The Cycas Genome and the Early Evolution of Seed Plants'. *Nature Plants* 8 (4): 389–401. https://doi.org/10.1038/s41477-022-01129-7.

Liu, Yang, Zhen Zeng, Yan-Mei Zhang, Qian Li, Xing-Mei Jiang, Zhen Jiang, Ji-Hong Tang, et al. 2021. 'An Angiosperm NLR Atlas Reveals That NLR Gene Reduction Is Associated with Ecological Specialization and Signal Transduction Component Deletion'. *Molecular Plant* 14 (12): 2015–31. https://doi.org/10.1016/j.molp.2021.08.001.

Mackay, John, Jeffrey F. D. Dean, Christophe Plomion, Daniel G. Peterson, Francisco M. Cánovas, Nathalie Pavy, Pär K. Ingvarsson, et al. 2012. 'Towards Decoding the Conifer Giga-Genome'. *Plant Molecular Biology* 80 (6): 555–69.

https://doi.org/10.1007/s11103-012-9961-7.

Meyers, Blake C., Alexander Kozik, Alyssa Griego, Hanhui Kuang, and Richard W. Michelmore. 2003. ‘Genome-Wide Analysis of NBS-LRR–Encoding Genes in Arabidopsis[W]’. *The Plant Cell* 15 (4): 809–34. https://doi.org/10.1105/tpc.009308.

Mitton, Jeffry B., and Scott M. Ferrenberg. 2012. ‘Mountain Pine Beetle Develops an Unprecedented Summer Generation in Response to Climate Warming.’ *The American Naturalist* 179 (5): E163–71. https://doi.org/10.1086/665007.

Mota, Manuel M., Helen Braasch, Maria Antonia Bravo, Ana Catarina Penas, Wolfgang Burgermeister, Kai Metge, and Edmundo Sousa. 1999. ‘First Report of Bursaphelenchus Xylophilus in Portugal and in Europe’. *Nematology* 1 (7): 727–34. https://doi.org/10.1163/156854199508757.

Mullin, Tim J., B. Andersson, J. C. Bastien, J. Beaulieu, R. D. Burdon, W. S. Dvorak, C. Plomion, J. Bousquet, and C. Kole. 2011. *Genetics, Genomics and Breeding of Conifers*. Enfield, NH, USA: Science Publishers.

Neale, David B, Aleksey V Zimin, Sumaira Zaman, Alison D Scott, Bikash Shrestha, Rachael E Workman, Daniela Puiu, et al. 2022. ‘Assembled and Annotated 26.5 Gbp Coast Redwood Genome: A Resource for Estimating Evolutionary Adaptive Potential and Investigating Hexaploid Origin’. *G3 Genes|Genomes|Genetics* 12 (1): jkab380. https://doi.org/10.1093/g3journal/jkab380.

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.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. ‘IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies’. *Molecular Biology and Evolution* 32 (1): 268–74. https://doi.org/10.1093/molbev/msu300.

Niu, Shihui, Jiang Li, Wenhao Bo, Weifei Yang, Andrea Zuccolo, Stefania Giacomello, Xi Chen, et al. 2022. ‘The Chinese Pine Genome and Methylome Unveil Key Features of Conifer Evolution’. *Cell* 185 (1): 204-217.e14. https://doi.org/10.1016/j.cell.2021.12.006.

Nystedt, Björn, Nathaniel R. Street, Anna Wetterbom, Andrea Zuccolo, Yao-Cheng Lin, Douglas G. Scofield, Francesco Vezzi, et al. 2013. ‘The Norway Spruce Genome Sequence and Conifer Genome Evolution’. *Nature* 497 (7451): 579–84. https://doi.org/10.1038/nature12211.

Ostermeier, Marc, and Stephen J. Benkovic. 2001. 'Evolution of Protein Function by Domain Swapping'. In *Advances in Protein Chemistry*, 55:29–77. Evolutionary Protein Design. Academic Press. https://doi.org/10.1016/S0065-3233(01)55002-0.

Pavy, Nathalie, Manuel Lamothe, Betty Pelgas, France Gagnon, Inanç Birol, Joerg Bohlmann, John Mackay, Nathalie Isabel, and Jean Bousquet. 2017. 'A High-Resolution Reference Genetic Map Positioning 8.8 K Genes for the Conifer White Spruce: Structural Genomics Implications and Correspondence with Physical Distance'. *The Plant Journal* 90 (1): 189–203. https://doi.org/10.1111/tpj.13478.

Pavy, Nathalie, Betty Pelgas, Jérôme Laroche, Philippe Rigault, Nathalie Isabel, and Jean Bousquet. 2012. 'A Spruce Gene Map Infers Ancient Plant Genome Reshuffling and Subsequent Slow Evolution in the Gymnosperm Lineage Leading to Extant Conifers'. *BMC Biology* 10 (1): 84. https://doi.org/10.1186/1741-7007-10-84.

Pellicer, Jaume, Oriane Hidalgo, Steven Dodsworth, and Ilia J. Leitch. 2018. 'Genome Size Diversity and Its Impact on the Evolution of Land Plants'. *Genes* 9 (2): 88. https://doi.org/10.3390/genes9020088.

Pellicer, Jaume, and Ilia J. Leitch. 2020. 'The Plant DNA C-Values Database (Release 7.1): An Updated Online Repository of Plant Genome Size Data for Comparative Studies'. *New Phytologist* 226 (2): 301–5. https://doi.org/10.1111/nph.16261.

Quesada, Tania, Marcio F. R. Resende Jr., Patricio Muñoz, Jill L. Wegrzyn, David B. Neale, Matias Kirst, Gary F. Peter, Salvador A. Gezan, C. Dana Nelson, and John M. Davis. 2014. 'Mapping Fusiform Rust Resistance Genes within a Complex Mating Design of Loblolly Pine'. *Forests* 5 (2): 347–62. https://doi.org/10.3390/f5020347.

R Core Team. 2021. 'R: A Language and Environment for Statistical Computing.' Vienna, Austria: R Foundation for Statistical Computing.

Scott, Alison D, Aleksey V Zimin, Daniela Puiu, Rachael Workman, Monica Britton, Sumaira Zaman, Madison Caballero, et al. 2020. 'A Reference Genome Sequence for Giant Sequoia'. *G3 Genes|Genomes|Genetics* 10 (11): 3907–19. https://doi.org/10.1534/g3.120.401612.

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.

Shao, Zhu-Qing, Jia-Yu Xue, Qiang Wang, Bin Wang, and Jian-Qun Chen. 2019. 'Revisiting the Origin of Plant NBS-LRR Genes'. *Trends in Plant Science* 24 (1): 9–12.

https://doi.org/10.1016/j.tplants.2018.10.015.

Smith, P.H., J. Hadfield, N.J. Hart, R.M.D. Koebner, and L.A. Boyd. 2007. ‘STS Markers for the Wheat Yellow Rust Resistance Gene Yr5 Suggest a NBS–LRR-Type Resistance Gene Cluster’. *Genome* 50 (3): 259–65. https://doi.org/10.1139/G07-004.

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.

Teshome, Demissew Tesfaye, Godfrey Elijah Zharare, and Sanushka Naidoo. 2020. ‘The Threat of the Combined Effect of Biotic and Abiotic Stress Factors in Forestry Under a Changing Climate’. *Frontiers in Plant Science* 11. https://www.frontiersin.org/articles/10.3389/fpls.2020.601009.

Tumas, Hayley, Joana J. Ilska, Sebastien Girardi, Jerome Laroche, Stuart A’Hara, Brian Boyle, Mateja Janes, et al. 2023. ‘High-Density Genetic Linkage Mapping in Sitka Spruce Advances the Integration of Genomic Resources in Conifers’. bioRxiv. https://doi.org/10.1101/2023.08.21.554184.

Van Ghelder, Cyril, Geneviève J. Parent, Philippe Rigault, Julien Prunier, Isabelle Giguère, Sébastien Caron, Juliana Stival Sena, et al. 2019. ‘The Large Repertoire of Conifer NLR Resistance Genes Includes Drought Responsive and Highly Diversified RNLs’. *Scientific Reports* 9 (1): 11614. https://doi.org/10.1038/s41598-019-47950-7.

Verde, Ignazio, Albert G. Abbott, Simone Scalabrin, Sook Jung, Shengqiang Shu, Fabio Marroni, Tatyana Zhebentyayeva, et al. 2013. ‘The High-Quality Draft Genome of Peach (Prunus Persica) Identifies Unique Patterns of Genetic Diversity, Domestication and Genome Evolution’. *Nature Genetics* 45 (5): 487–94. https://doi.org/10.1038/ng.2586.

Wei, Wei, Xing Wu, Alexandre Garcia, Nancy McCoppin, João Paulo Gomes Viana, Praerona S. Murad, David R. Walker, et al. 2023. ‘An NBS-LRR Protein in the Rpp1 Locus Negates the Dominance of Rpp1-Mediated Resistance against Phakopsora Pachyrhizi in Soybean’. *The Plant Journal* 113 (5): 915–33. https://doi.org/10.1111/tpj.16038.

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.

Wersch, Solveig van, Lei Tian, Ryan Hoy, and Xin Li. 2020. ‘Plant NLRs: The

Whistleblowers of Plant Immunity'. *Plant Communications* 1 (1): 100016. https://doi.org/10.1016/j.xplc.2019.100016.

White, Timothy L., Wesley Thomas Adams, and David B. Neale (eds.). 2007. *Forest Genetics*. Cabi.

Wickham, H. 2016. 'Ggplot2: Elegant Graphics for Data Analysis.' R. New York: Springer-Verlag.

Wilcox, P L, H V Amerson, E G Kuhlman, B H Liu, D M O'Malley, and R R Sederoff. 1996. 'Detection of a Major Gene for Resistance to Fusiform Rust Disease in Loblolly Pine by Genomic Mapping.' *Proceedings of the National Academy of Sciences* 93 (9): 3859–64. https://doi.org/10.1073/pnas.93.9.3859.

Xiong, Xingyao, Junbo Gou, Qinggang Liao, Yanlin Li, Qian Zhou, Guiqi Bi, Chong Li, et al. 2021. 'The Taxus Genome Provides Insights into Paclitaxel Biosynthesis'. *Nature Plants* 7 (8): 1026–36. https://doi.org/10.1038/s41477-021-00963-5.

Zimin, Aleksey, Kristian A Stevens, Marc W Crepeau, Ann Holtz-Morris, Maxim Koriabine, Guillaume Marçais, Daniela Puiu, et al. 2014. 'Sequencing and Assembly of the 22-Gb Loblolly Pine Genome'. *Genetics* 196 (3): 875–90. https://doi.org/10.1534/genetics.113.159715.

Supporting information

A detailed description of the phylogenetic analysis of NLR genes in Pinaceae is available in the Supporting Information. The following figures are available in this document:

Fig. S1: NLR phylogeny for Pinus flexilis.

Fig. S2: NLR phylogeny for Picea abies.

Fig. S3: NLR phylogeny for Picea glauca.

Fig. S4: NLR phylogeny for Picea sitchensis.

Figure legends

Figure 1: Overview of main gymnosperm clades, with total number of NLR genes and their corresponding categories indicated as bar charts, as found in this study (Table 1). Cladogram is based on the phylogeny presented in Leslie et al. (2018). Genome sizes are indicated beneath each bar chart and are based on the chromosome-level assemblies used in this study (Materials & methods section 1 for details) or, in the case of Pinaceae (Pinus flexilis), obtained from the Kew Plant DNA C-values database (release 7.1, Pellicer and Leitch 2020). Gymnosperms invariably have large genomes (~10 Gbp), but display large variations in NLR gene numbers. Despite the three-fold increase in genome size observed in Pinaceae, the number of discovered NLR genes remained within the average range of conifers. Pictures were obtained from the Wikimedia Commons repository (https://commons.wikimedia.org) and correspond to the broader taxonomic clades: A – Cycas rumphii Miq., andy_king50; B – Ginkgo biloba L., Susanna Giaccai; C – Taxus baccata L., Mykola Swarnyk; D – Sequoiadendron giganteum, W. Bulach; E – Pinus flexilis, Greg Woodhouse.

Figure 2: Chromosomal distribution of NLR genes across gymnosperms. A – Scatter plot indicating the observed number of NLR genes versus the expected number of NLR genes (calculated based on the length of the chromosome and the total number of NLR genes discovered in the genome, Materials & methods section 1) for each chromosome in each taxon analysed in this study. Black line follows the function y=x, indicating a perfectly homogeneous distribution of NLR genes over the chromosomes. Deviations from this line therefore indicate a non-homogeneous distribution. Highly deviant chromosomes of four taxa are highlighted and have their intrachromosomal NLR distribution displayed in histograms (B-E). Bin width equals ±1% of the length of the largest chromosome in the genome of the respective taxon (Materials & methods section 1). Colours indicate NLR class as determined with NLR Annotator (Steuernagel et al. 2020) and manual BLASTs (Materials & methods section 2). Ultra-dense NLR clusters are indicated for each taxon.

Figure 3: Histogram plots of NLR genes on each chromosome within the genome of five gymnosperm lineages. Chromosomes are ordered based on the ordering in the respective assembly and do not reflect synteny. Red * indicates that a particularly dense bin has been cropped manually in order to fit all the histograms into a single image.

Figure 4: Intragenomic phylogenetic relationships of NLR genes based on the conserved central NB-ARC domain, calculated with maximum likelihood algorithms using IQTree v1.6.12 (Nguyen et al. 2015) (Materials & methods section 3) and annotated using the iTOL web server (Letunic and Bork 2021). Colour strips around the circular trees indicate NLR class as determined with NLR Annotator (Steuernagel et al. 2020) and manual BLASTs (Materials & methods section 2). Main gymnosperm lineages represented by Cycadales (A), Ginkgoales (B), Cuppressaceae (C), Taxaceae (D), and Pinaceae: Pinus (E), and Picea (F).

Figure 5: Chromosomal structuring in phylogenetic relationships within a conifer species (Taxus chinensis) displayed in an intragenomic framework. Branch colours correspond to the NLR gene subfamily, as indicated in the same colours next to the phylogeny. Coloured squares at the tips of branches represent the chromosome on which the respective NLR genes are located. NLR genes found on scaffolds that were not assembled into chromosomes are indicated with empty colour squares (“Unclassified”).