

NLR STATION

MAR 8, 2021

Introduction

NLR STATION

MAR 8, 2021

Warm-up questions

What are NLRs and why should I care?

Mystery plant NLR annotation

NLR-omes, Pan-genomes, and Microbiomes*, oh my!

Introduction

Take a minute to make sure your preferred name is displayed in Zoom

Introduction

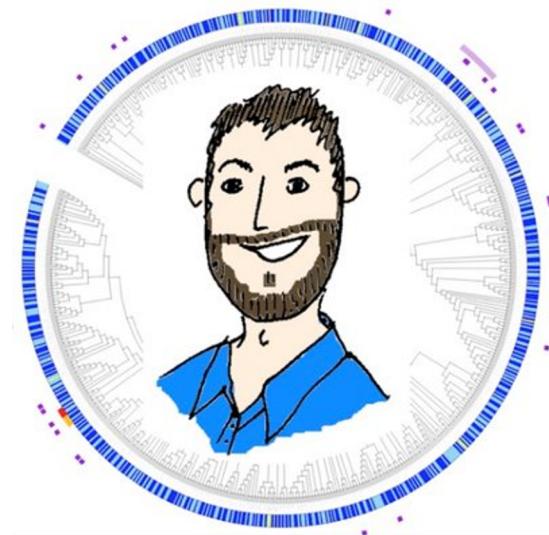
B.A. in Ecology, Evolution, Conservation Bio from
University of Hawai'i at Hilo

- Research in a plant virology lab

Research Associate with Monsanto in Cambridge, MA
-Insect-control proteins from bacteria

PhD in Plant-Microbe Biology from Cornell U.
-Molecular warfare during bacterial infection of rice

Postdoc with Nathan Springer
-Epigenetics, what even is it?



@hashtag_read
read0094@umn.edu

Warm-up Question 1:

“NLR” is an acronym for:

-  Nucleotide-binding and Leucine-rich Repeat
-  NOD-Like Receptor
-  Nuclear Localized Resistance gene
-  I thought these things were called NB-LRRs or R-genes

Warm-up Question 1:

“NLR” is an acronym for:



Nucleotide-binding and Leucine-rich Repeat



NOD-Like Receptor



Nuclear Localized Resistance gene



I thought these things were called NB-LRRs or R-genes

Warm-up Question 2:

According to the reading, which of these plants has the largest NLR repertoire?

 Giant Sequoia

 Apple

 Wheat

 Cucumber

Warm-up Question 2:

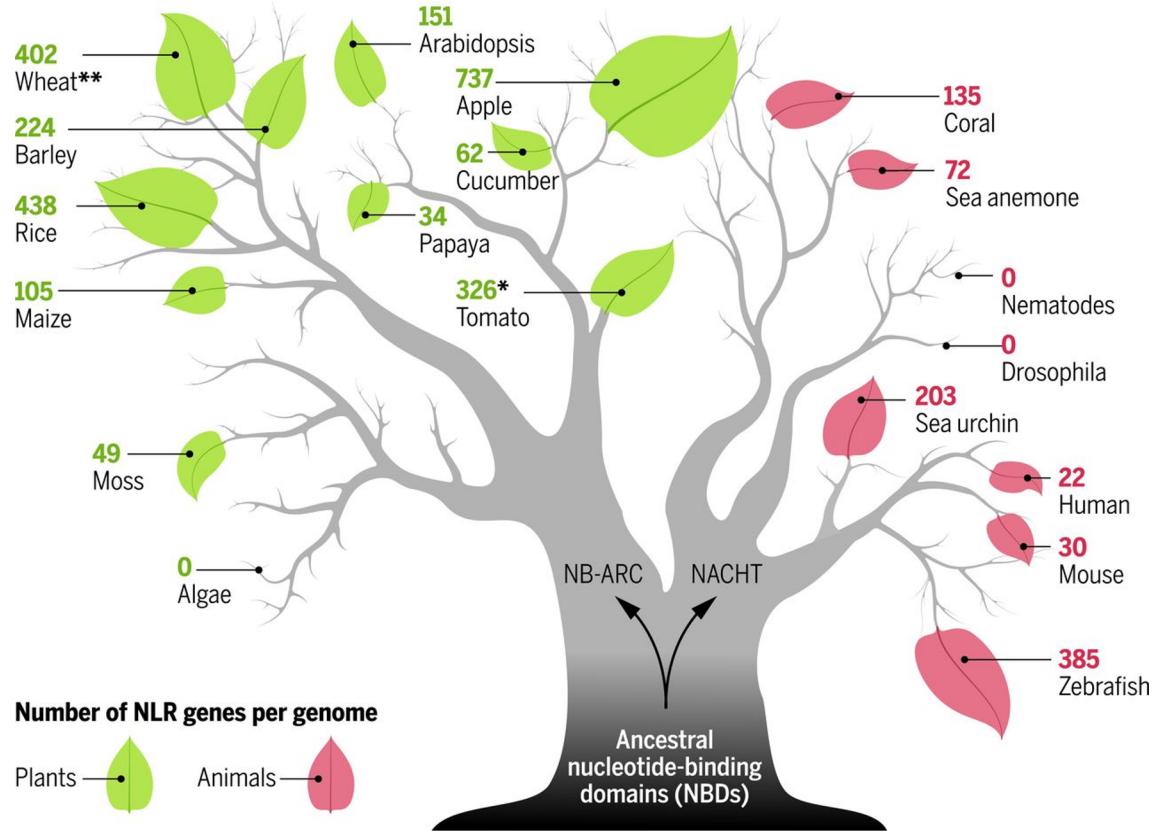
According to the reading, which of these plants has the largest NLR repertoire?

 Giant Sequoia

 Apple

 Wheat

 Cucumber



Intracellular innate immune surveillance devices in plants and animals

BY JONATHAN D. G. JONES, RUSSELL E. VANCE, JEFFERY L. DANGL
SCIENCE 02 DEC 2016

Warm-up Question 3:

According to the reading, what is the largest NLR cluster?



RGC2 in lettuce



Xo1 in rice



RPPS in Arabidopsis



NLR2000 in baobab tree

Warm-up Question 3:

According to the reading, what is the largest NLR cluster?

-  **RGC2 in lettuce 24 cluster members - ~3.5 Megabases!**
-  Xo1 in rice
-  RPPS in Arabidopsis
-  NLR2000 in baobab tree

Life is hard for plants....

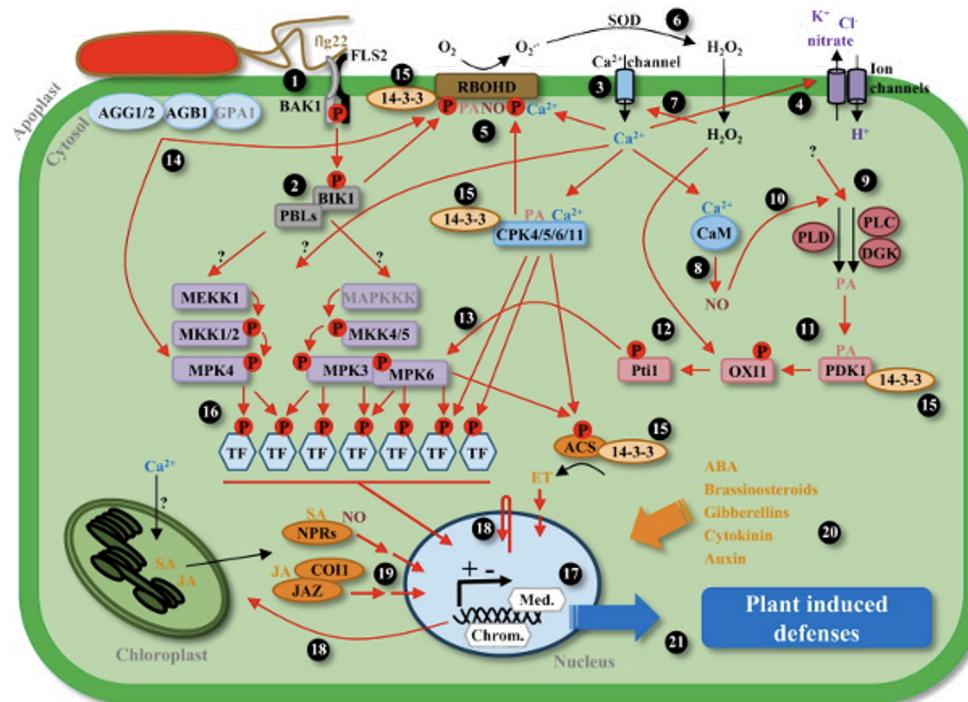
DALE WALTERS

FORTRESS PLANT

*How to survive
when everything
wants to eat you*

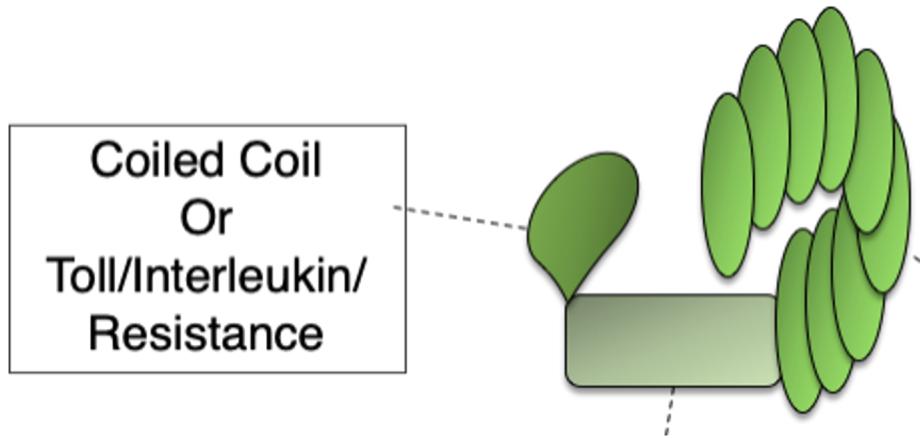


Molecular plant defense is complicated

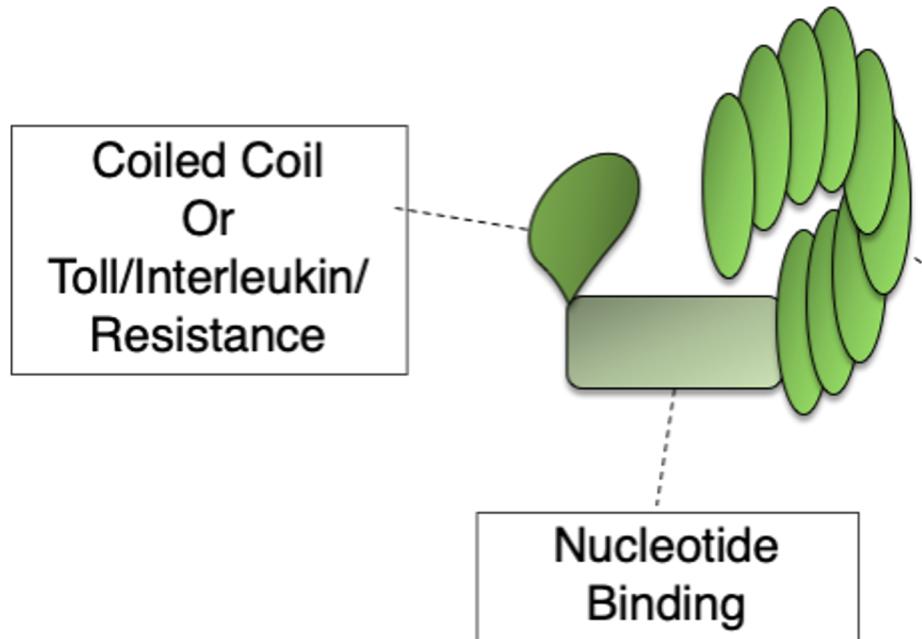


Bigeard, Colcombet, Hirt - Molecular Plant, 2015

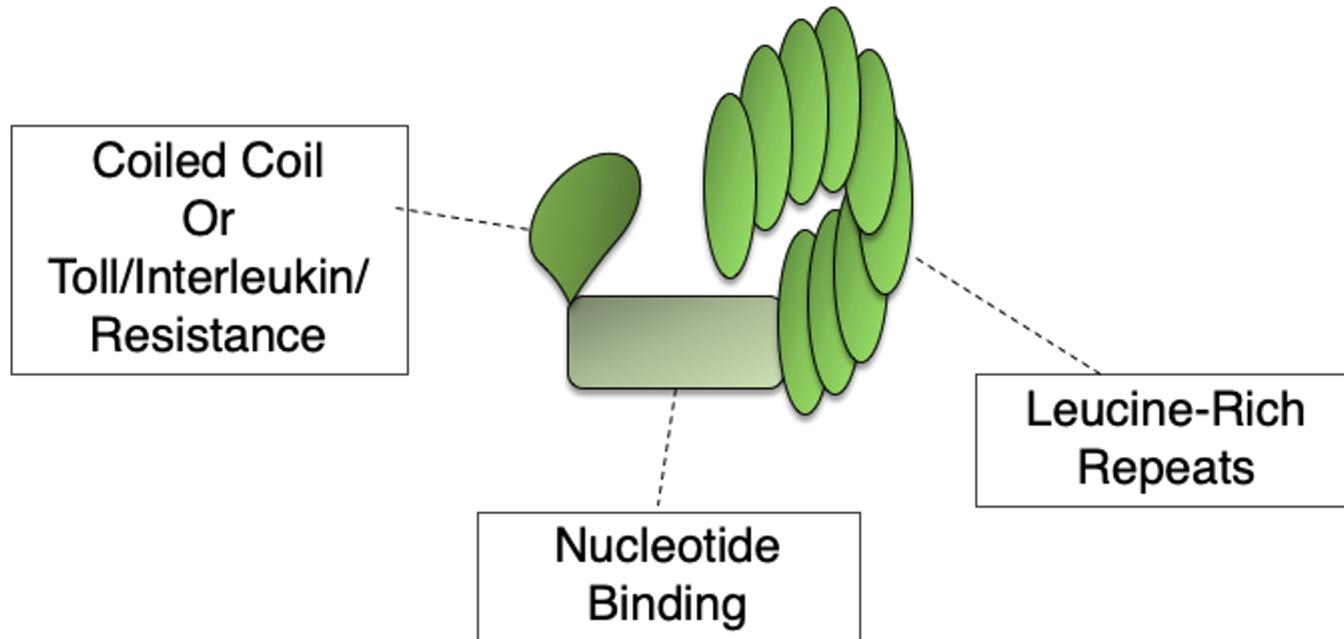
Introducing the NLR gene family



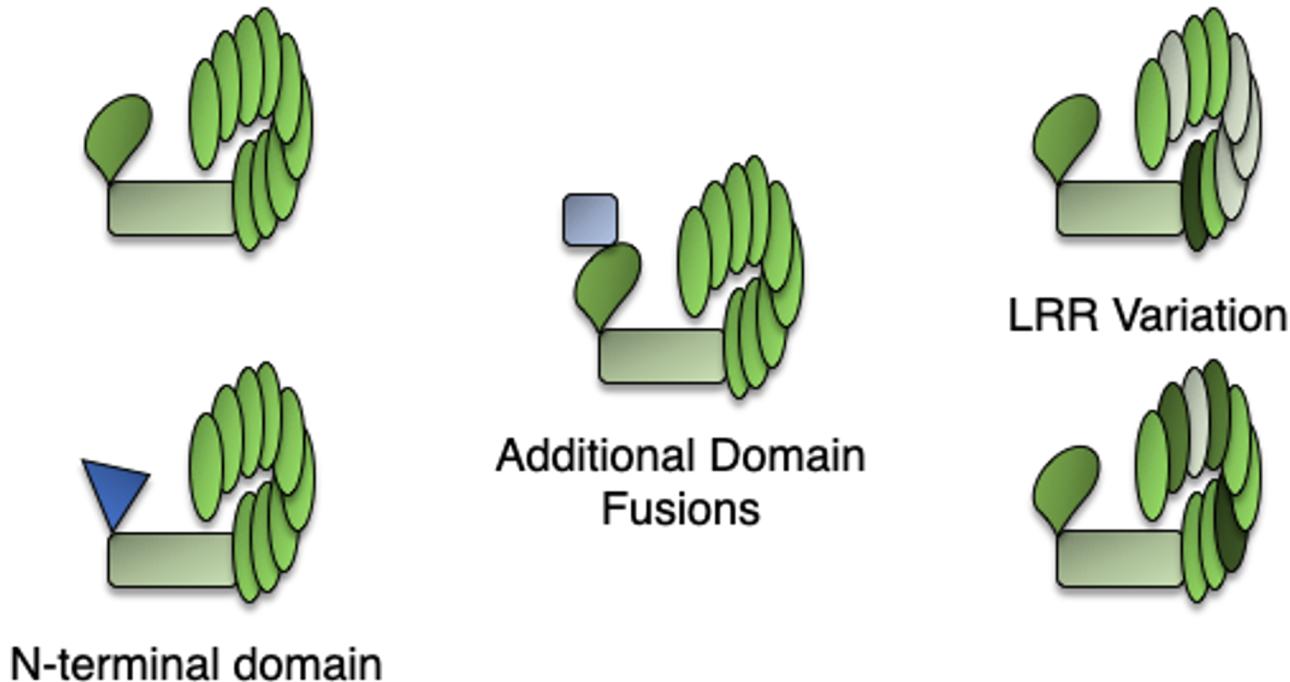
Introducing the NLR gene family



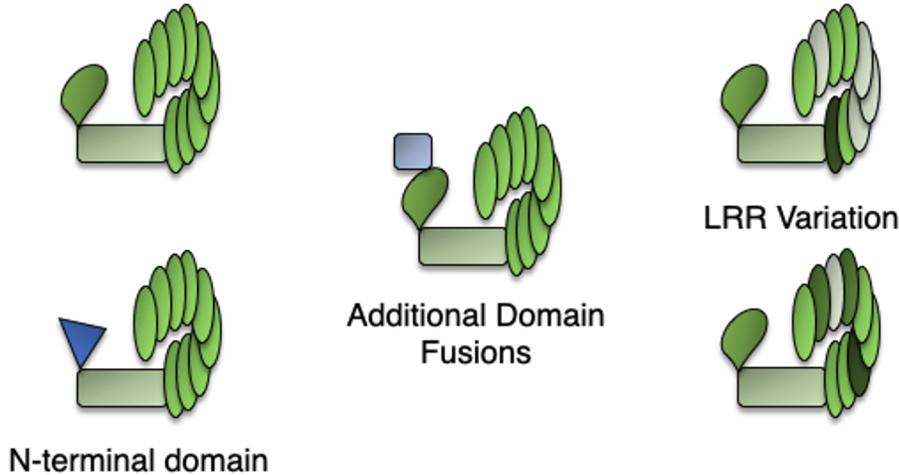
Introducing the NLR gene family



NLRs are among the most diverse plant gene families -
Both within individuals and at the population level

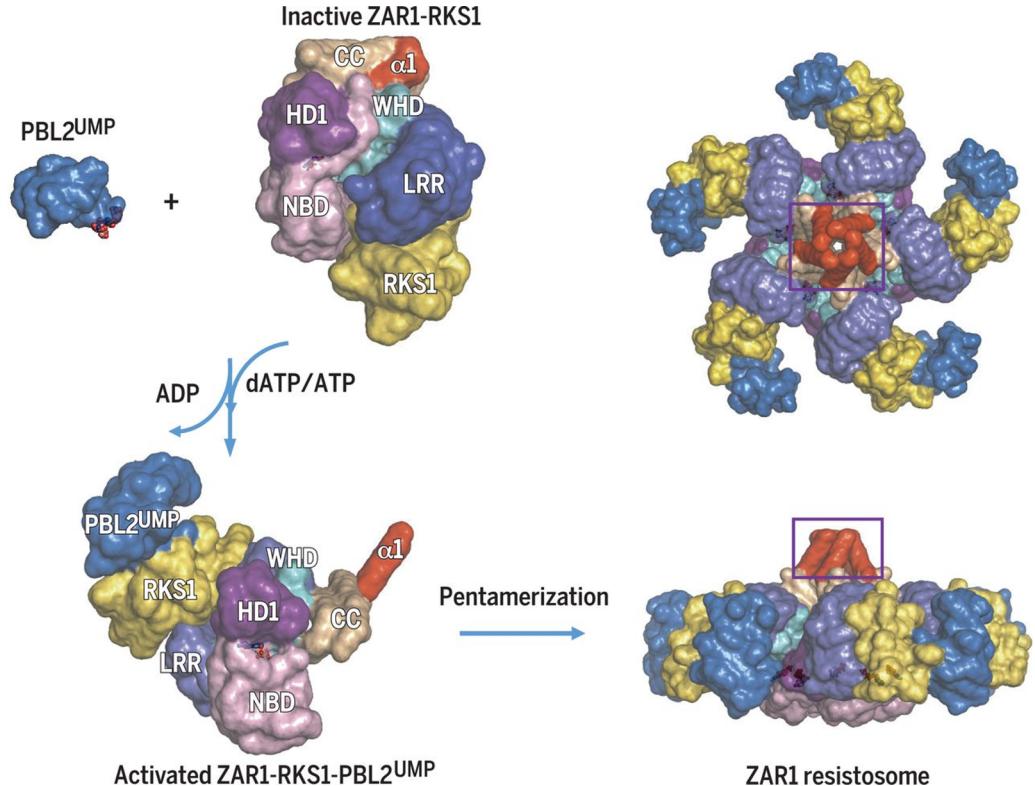


NLRs are among the most diverse plant gene families -
Both within individuals and at the population level



Plants lack an adaptive immune system (antibodies) - the large repertoire of NLRs act as an alternative mechanism to recognize 'non-self'

How do NLRs ‘work’?



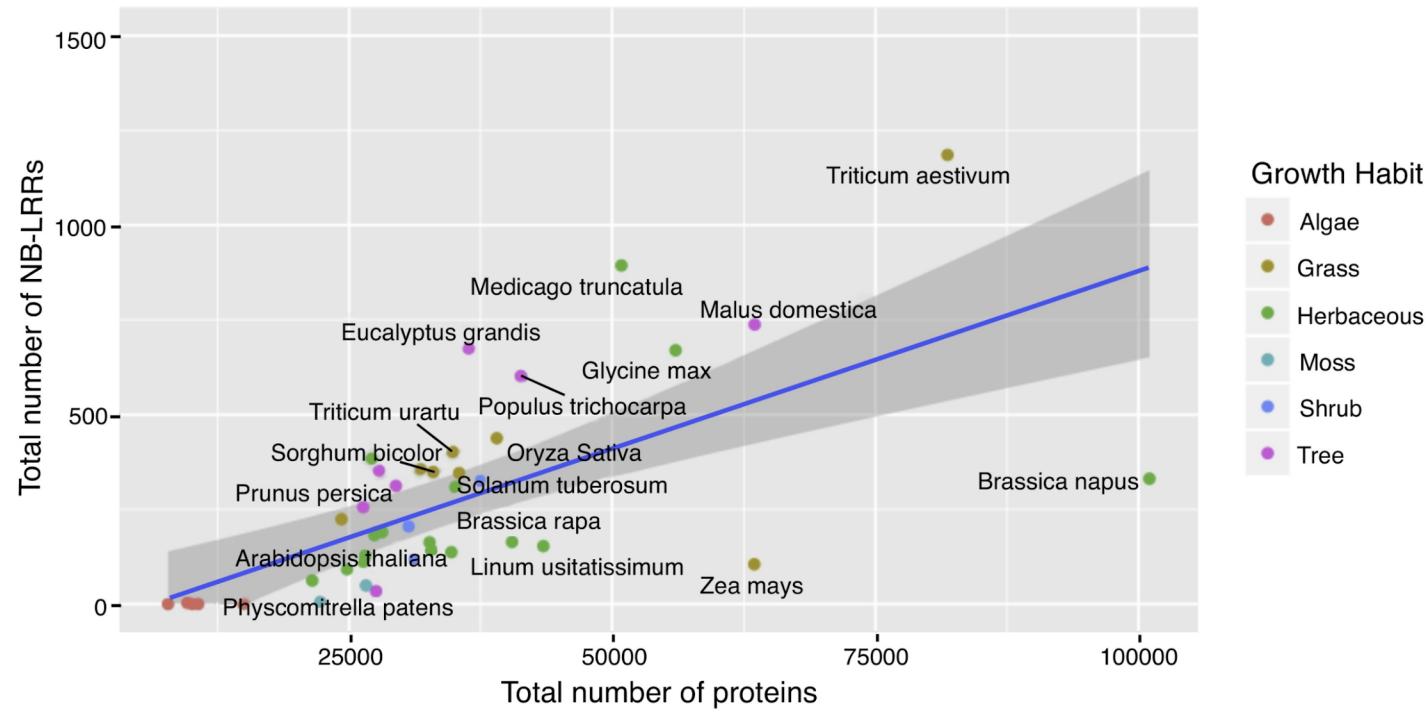
Reconstitution and structure of a plant NLR resistosome conferring immunity

BY JIZONG WANG, MEIJUAN HU, JIA WANG, JINFENG QI, ZHIFU HAN, GUOXUN WANG, YIJUN QI, HONG-WEI WANG, JIAN-MIN

JIJIE CHAI

SCIENCE 05 APR 2019

NLRs are diverse within and between individual genomes



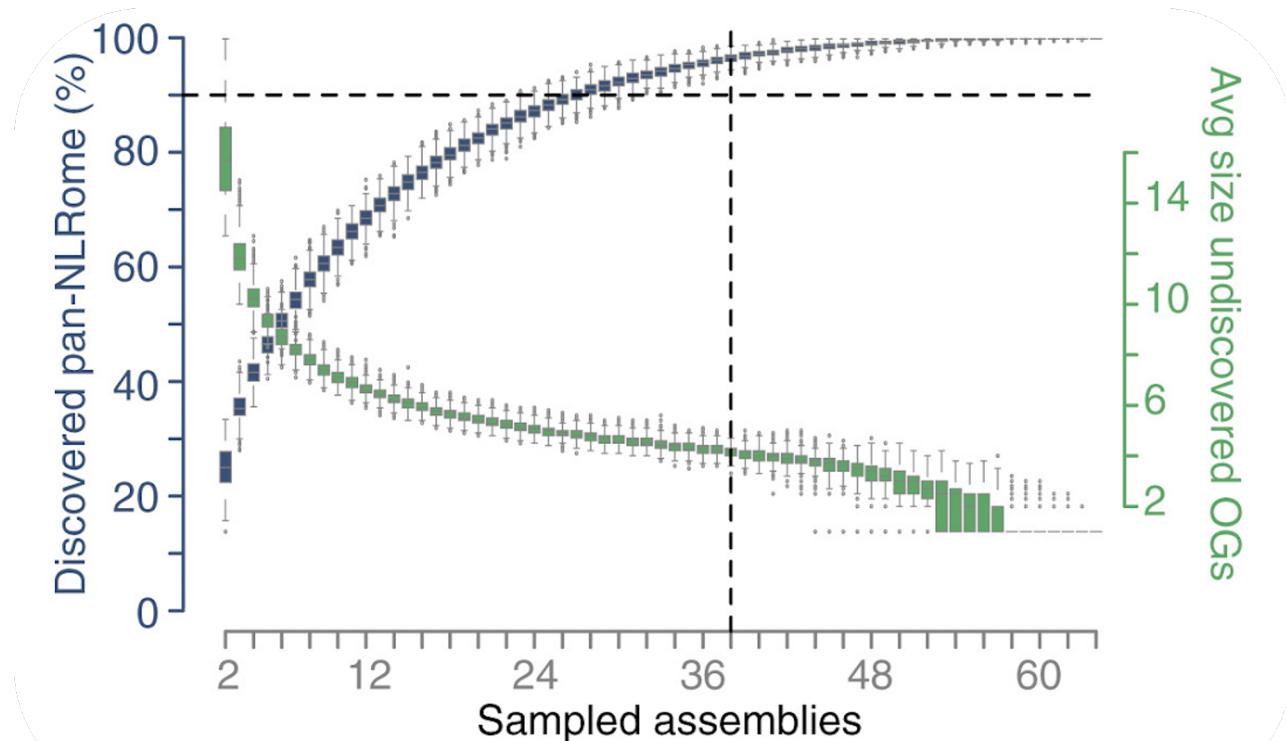
That's a lot of diversity within individuals, What about at the species level?

Table 1

Examples of plant pan-genome or pan-NRLome studies and key observations on NLRs or *R* genes

Plant pan-genome or pan-NRLome studies					
Organism (s)	<i>N</i>	Sequencing technology	Assembly strategy	Reference	Key observations on NLR or <i>R</i> genes
<i>Brassica rapa</i>	3	Short reads	De novo assemblies	Lin et al. (2014)	- NLRs enriched in dispensable genome.
<i>Glycine soja</i>	7	Short reads	De novo assemblies	Li et al. (2014)	- CNVs common in NLRs: candidates of resistance differences between wild and cultivated accessions. - NLRs numbers and domain architectures varying between species. - NLRs enriched in dispensable genome.
<i>Zea mays</i>	503	Short reads	De novo transcriptome	Hirsch et al. (2014)	
<i>Oryza sativa</i>	3	Short reads	De novo assemblies	Schatz et al. (2014)	- NLRs enriched in dispensable genome, e.g. 12% shell versus 0.35% core genes are NLRs.
<i>O. sativa</i>	1483	Short reads	Iterative assembly	Yao et al. (2015)	- NLRs enriched in dispensable genome.
<i>Populus</i> clade	7	Short reads	Map to reference	Pinosio et al. (2016)	- CNVs and SVs enriched for NLRs.
<i>Brassica oleracea</i>	10	Short reads	Iterative assembly	Golicz et al. (2016)	- NLRs enriched in genes showing PAV. - 43% of NLRs dispensable, 45% in clusters, and 60% absent from reference.

How to capture NLR diversity? Enter the panNLR-ome



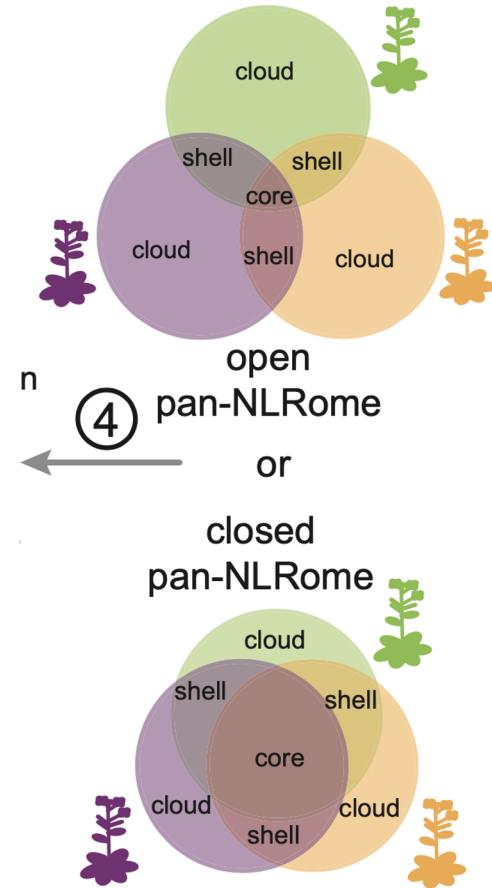
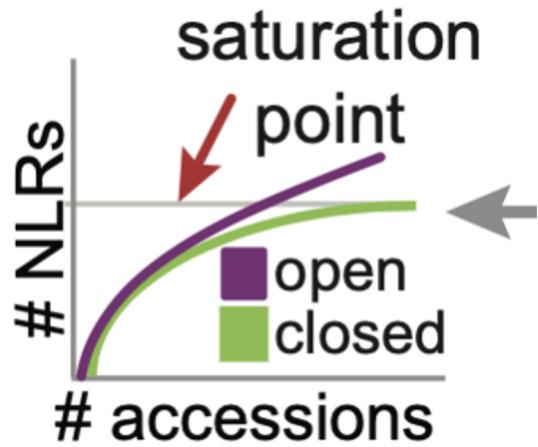
CELL VOLUME 178, ISSUE 5, P1260-1272.E14, AUGUST 22, 2019

A Species-Wide Inventory of NLR Genes and Alleles in *Arabidopsis thaliana*

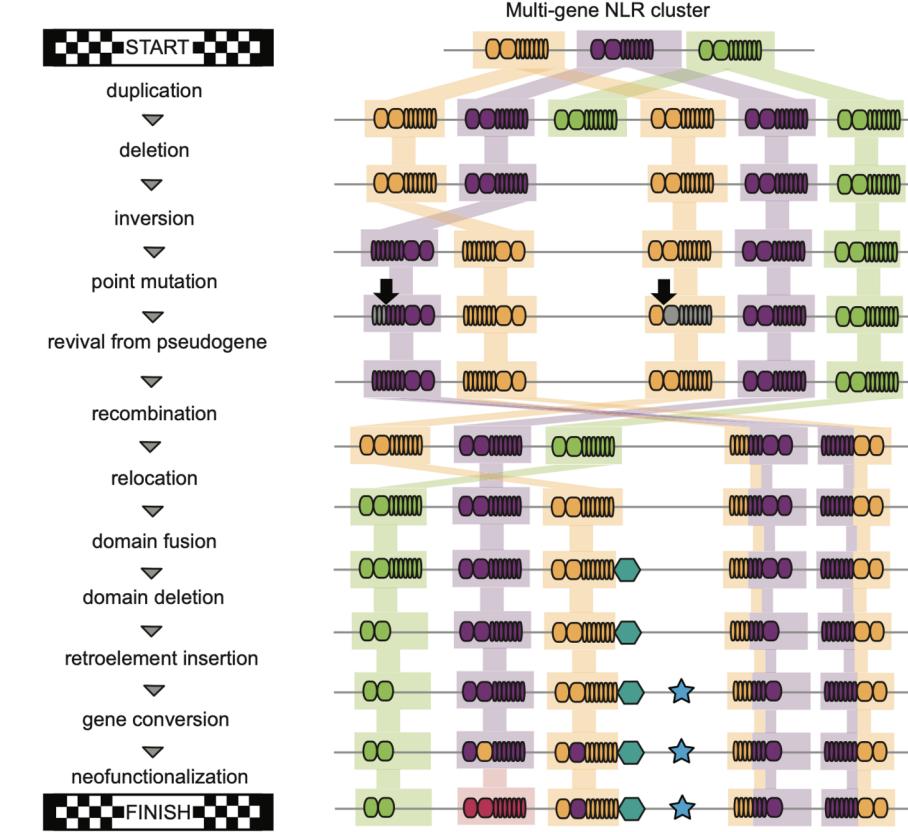
Anna-Lena Van de Weyer 8Freddy Monteiro 8Oliver J. Furzer 8Marc T. Nishimura 8Volkan Cevik 8Kamil Witek 8Jonathan D.G. Jones 8Jeffery L. Dangl 8Detlef Weigel 9Felix Bemm 8

Will all species have similar curves?

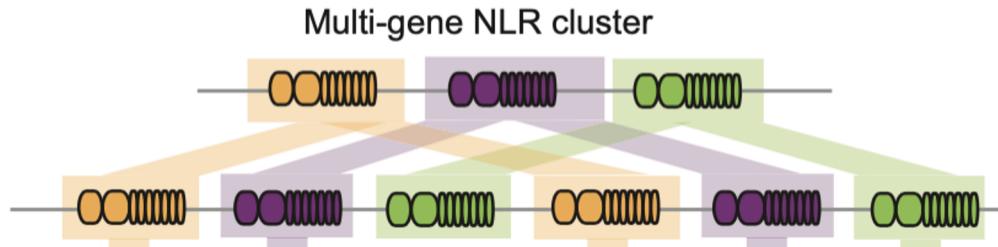
Will all species have similar curves?



But, NLRs are challenging to sequence, assemble, and annotate!



But, NLRs are challenging to sequence, assemble, and annotate!



Gedankenexperiment time!

You are given enough money to sequence your favorite plant with one sequencing technology, and your goal is to identify NLR diversity - what technology do you choose? Why? Are there any down-sides to this choice?

Gedankenexperiment time!

👍 Illumina

😂 PacBio

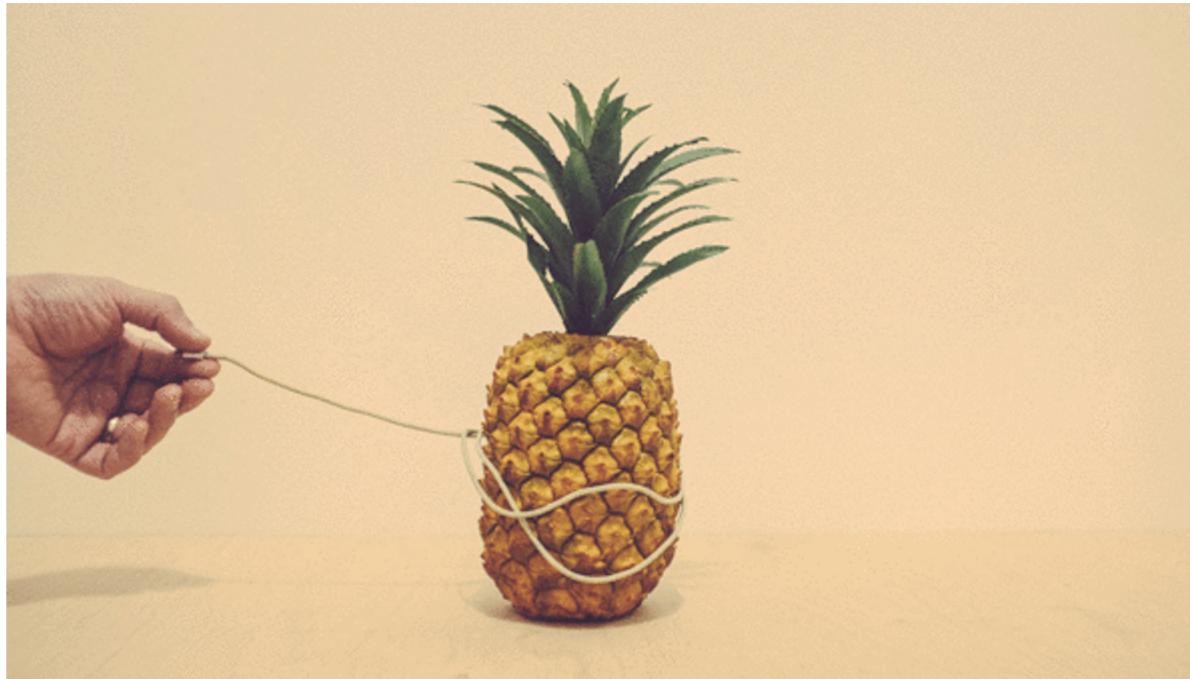
😯 Oxford Nanopore

👏 Other

Four candidate genomes to hunt for NLRs

And the winner is.....

And the winner is..... PINEAPPLE!!!!



Pineapple facts:

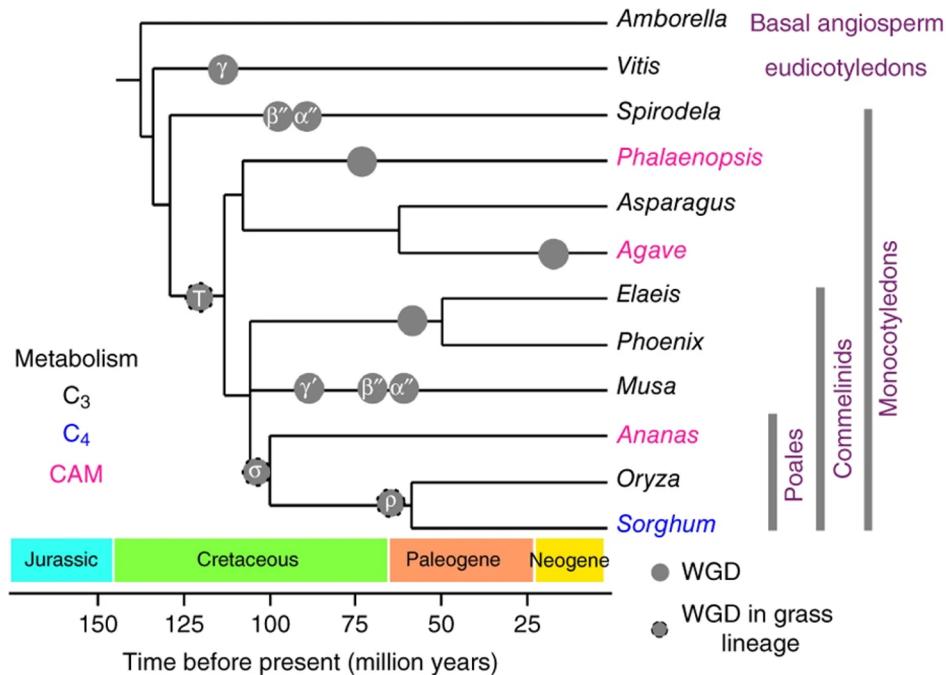
25 chromosomes

Genome size ~526 megabases

Assembly size 382 megabases

MAKER used for gene annotation -
27,024 gene models

TEs make up 44% of the assembly



Pineapple facts:

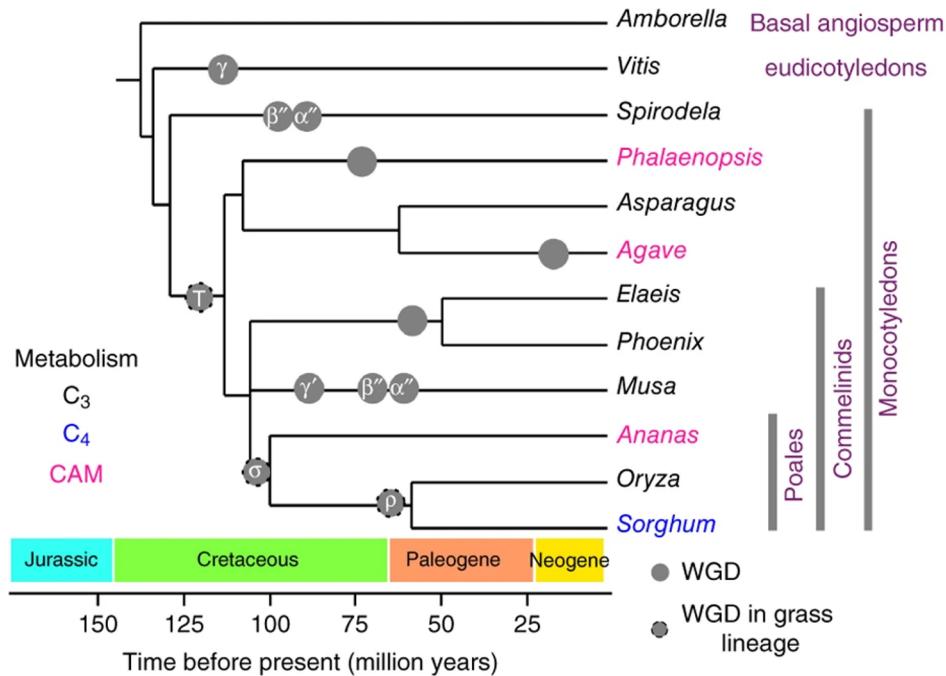
25 chromosomes

Genome size ~526 megabases

Assembly size 382 megabases

MAKER used for gene annotation -
27,024 gene models

TEs make up 44% of the assembly



Any NLR number guesses?

We have a genome - time to find some NLRs

Evidence-based

Our basic pipeline

0. Obtain protein sequences of species of interest and organise them into a directory.

We follow the Phytozome organisation of `master_dir/species/annotation/species_version_proteins.fa` where each species is denoted by the first letter of the genus name and all letters in the species names, for example `Athaliana`

1. Pfam-based annotation of domains

usage: `bash run_pfam_scan.sh dir`

Dependencies:

- HMMER software (<http://hmmer.janelia.org/>) including pfam_scan.pl (part of HMMER) Move in same directory as this script or set path at command string.
- Pfam database (<http://pfam.xfam.org/>)
- File names should be consistent with Phytozome and include Species_*_protein.fa
- perl modules specified in the scripts (best to install with cpan: <http://www.cpan.org/modules/>)

2. Parsing the pfamscan output with K-parse_Pfam_domains_v3.1.pl

- The script parses the output of pfam_scan.pl
- The script extracts all domains for each proteins and removes redundant nested hits with larger e-values.
- Domains are printed out in the order of appearance in the query.
- By default, Pfam_B domains are skipped.

How many NLRs - Evidence-based?

Open Access | Published: 02 November 2015

The pineapple genome and the evolution of CAM photosynthesis

Ray Ming , Robert VanBuren, [...] Qingyi Yu 

Nature Genetics 47, 1435–1442(2015) | [Cite this article](#)

11k Accesses | 220 Citations | 137 Altmetric | [Metrics](#)

From the publication - MAKER annotation and PFAM based determination of protein domains
237 “Disease Resistance Proteins”
136 “NBS-LRRs”

Ab initio **NLR-Annotator**

github.com/steuernb/NLR-Annotator

Ab initio **NLR-Annotator**

github.com/steuernb/NLR-Annotator

3 steps:

1. ChopSequence.jar -- chop the genome into 20kb fragments with 5kb overlaps
2. NLR-Parser.jar -- 6 frame translation of each genomic fragment
 - Translations scanned for NLR domains
3. NLR-Annotator.jar -- 20kb fragments are merged
 - Potential NLR-containing regions are identified
 - NLR domains are listed and NLRs are classified (Complete/Partial, Pseudo)
 - Optional NB domain alignment for building trees

NLR-Annotator

```
[(base) read0094@ln0004 [~/Scriptz/NLR-Annotator] % ls -lh
total 3.6M
-rw-----. 1 read0094 springer 140K Mar  6 16:54 ChopSequence.jar
-rwx-----. 1 read0094 springer 455K Mar  6 16:54 mast
-rw-----. 1 read0094 springer 1.7M Mar  6 16:54 meme.xml
-rw-----. 1 read0094 springer 140K Mar  6 16:54 NLR-Annotator.jar
-rw-----. 1 read0094 springer  50K Mar  6 16:54 NLR-Parser3.jar
-rw-----. 1 read0094 springer 140K Mar  6 16:54 NLR-Parser.jar
-rw-----. 1 read0094 springer 4.7K Sep 28 17:00 readme.md
```

NLR-Annotator

```
[(base) read0094@ln0004 [~/Scriptz/NLR-Annotator] % ls -lh
total 3.6M
-rw-----. 1 read0094 springer 140K Mar  6 16:54 ChopSequence.jar
-rwx-----. 1 read0094 springer 455K Mar  6 16:54 mast
-rw-----. 1 read0094 springer 1.7M Mar  6 16:54 meme.xml
-rw-----. 1 read0094 springer 140K Mar  6 16:54 NLR-Annotator.jar
-rw-----. 1 read0094 springer  50K Mar  6 16:54 NLR-Parser3.jar
-rw-----. 1 read0094 springer 140K Mar  6 16:54 NLR-Parser.jar
-rw-----. 1 read0094 springer 4.7K Sep 28 17:00 readme.md
```

```
[(base) read0094@ln0004 [~/Pineapple] % ls -lh
total 1011M
-rw-----. 1 read0094 springer  23K Mar  4 13:51 Acomosus_321_v3.ANNOTATED.txt
-rw-----. 1 read0094 springer 482M Mar  2 16:26 Acomosus_321_v3.CHOPPED.fa
-rw-----. 1 read0094 springer 371M Mar  2 16:22 Acomosus_321_v3.fa
-rw-----. 1 read0094 springer  39K Mar  4 13:51 Acomosus_321_v3.NB_Align.fasta
-rw-----. 1 read0094 springer  32M Mar  6 16:57 Acomosus_321_v3.PARSED.xml
-rw-----. 1 read0094 springer 13M Mar  2 16:22 Acomosus_321_v3.protein_primaryTranscriptOnly.fa
```

How many NLRs? - NLR-Annotator-based

From our *ab initio* analysis - NLR-Annotator
235 total NLRs

Row Labels	Count
complete	166
complete (pseudogene)	47
partial	15
partial (pseudogene)	7

NLR-Annotator output

Chromosome	NLR Number	Status	Start	End	Orientation	Domains Detected
LG14	LG14_nlr_4	complete	10797640	10799456	-	1,4,5,10,3,12,2,7,9,9,11
LG14	LG14_nlr_5	complete (pseudogene)	10786515	10790841	-	17,1,4,5,10,3,12,2,7,9,9,11,11,11,20
LG14	LG14_nlr_6	complete	10774651	10776772	-	1,4,5,10,3,12,2,7,9,9,11,11,11,11,11
LG14	LG14_nlr_7	complete	6147706	6149928	-	16,1,6,4,5,12,2,8,7,9
LG12	LG12_nlr_1	partial	10806563	10807847	-	2,8,7,11,11,19,11
LG12	LG12_nlr_2	complete	10740682	10743802	-	17,1,6,4,5,10,3,12,2,8,9,9,11,11,11,11,11,11
LG12	LG12_nlr_3	complete	10717114	10719598	-	1,6,4,10,3,2,8,9,9,11,11,11,11,11,11,11,11
LG12	LG12_nlr_4	complete	10697160	10701330	-	1,6,4,10,3,12,2,8,11,9,9,11,11,11,11,11,11,11,11,19
LG11	LG11_nlr_1	complete	269416	273145	+	16,1,6,4,5,10,3,12,2,8,7,19,11,11,11,11,11,11,11

NLR-Annotator output

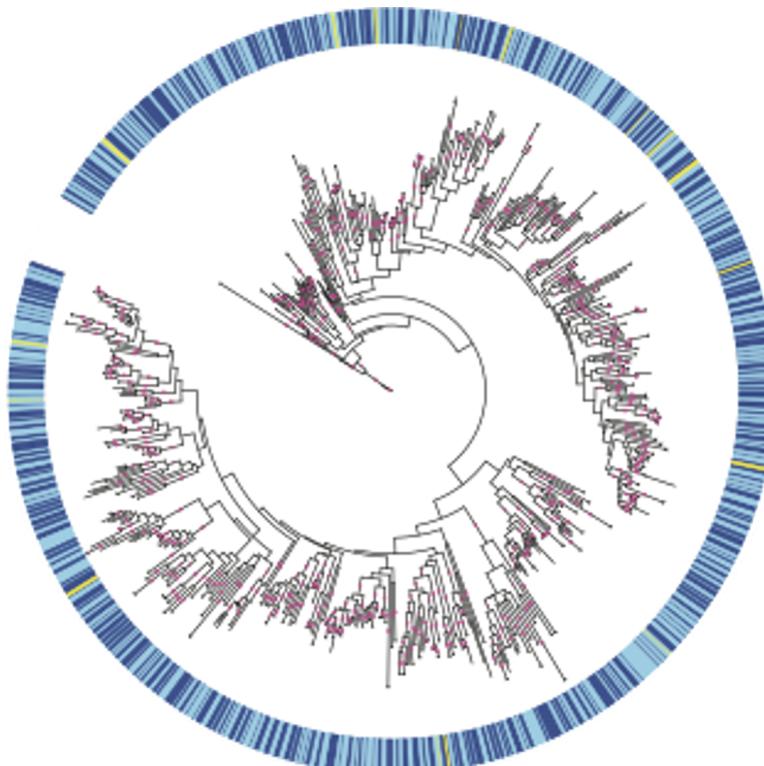
Chromosome	NLR Number	Status	Start	End	Orientation	Domains Detected
LG14	LG14_nlr_4	complete	10797640	10799456	-	1,4,5,10,3,12,2,7,9,9,11
LG14	LG14_nlr_5	complete (pseudogene)	10786515	10790841	-	17,1,4,5,10,3,12,2,7,9,9,11,11,11,20
LG14	LG14_nlr_6	complete	10774651	10776772	-	1,4,5,10,3,12,2,7,9,9,11,11,11,11,11
LG14	LG14_nlr_7	complete	6147706	6149928	-	16,1,6,4,5,12,2,8,7,9
LG12	LG12_nlr_1	partial	10806563	10807847	-	2,8,7,11,11,19,11
LG12	LG12_nlr_2	complete	10740682	10743802	-	17,1,6,4,5,10,3,12,2,8,9,9,11,11,11,11,11,11
LG12	LG12_nlr_3	complete	10717114	10719598	-	1,6,4,10,3,2,8,9,9,11,11,11,11,11,11,11,11
LG12	LG12_nlr_4	complete	10697160	10701330	-	1,6,4,10,3,12,2,8,11,9,9,11,11,11,11,11,11,11,11,19
LG11	LG11_nlr_1	complete	269416	273145	+	16,1,6,4,5,10,3,12,2,8,7,19,11,11,11,11,11,11,11

NLR-Annotator output - sequence alignment of NB-domain

CLUSTAL O(1.2.4) multiple sequence alignment

scaffold_2252_nlr_1	SIVGTAGIGKTAQIYHNHEFPMRIWVYVSEKFNVRKLTVKMIQRALDFMLVLVDDVWE	60
scaffold_1587_nlr_1	AITGMMGGAGKTTLAAYVYKLFHFVDCAWVPVHQDPDITELVIMIMIDQLGLYLIVLDDIWR	60
scaffold_1429_nlr_1	SVLGMGGGLGKTTLVTHVYNIIISFDACAWVAQSQSYETHDLLRQILNEFCRYVLLILDDVWS	60
scaffold_1429_nlr_2	SVLGMGGGLGKTTLVTHVYNIIISFDACAWVAQSQSYETDPLLRLQILKELYRYVLLILDDVWS	60
scaffold_763_nlr_1	PVVGIGGVGKTTLVQLVYSKDVFHLLRMWICVSNAFDARRITKEVVESASFFLLVLLDDVWS	60
scaffold_1186_nlr_1	TIVGIGGLGKTTLAALKVYSDHEFDSIMMWCVSLDFDAAAIIRKILNYSHSFLLILDDVWN	60
	: * . * ***:*. : *. * *: * : : : : : : : : : : : : : : : : :	
scaffold_2252_nlr_1	EIYSEKRSVVLLTSQIERSVVKQIETLDTDDYWRLFKHCAFLLGSVLRNND	120
scaffold_1587_nlr_1	KRVWDNGSRIIVTTRFKDVAIYHDPLNEEDSWRLFSTKVFCAGLPLALVVLLGLVSIKLK	120
scaffold_1429_nlr_1	ADVMNSKSRSIVLTTTIRDVALLANELEADHSWDLFCKNAFGSGGLPLAIVSIGRLLSFREQ	120
scaffold_1429_nlr_2	TDVLDCKSRISIVLTTTIRDVALLASELEADHSWDLFCKNAFCGGGLPLAIVSIARLLSFREK	120
scaffold_763_nlr_1	DDQVQQGSKIVVTTRNSSVARIKGMLSEEDSLSLFTHHAFNLNGSPLAAKTLGQLCNSKLG	120
scaffold_1186_nlr_1	DDKTMRGSKILLTTRMDSVADMAAKLEESDCMSLNFNNYAFLGGSPLAIEVIGGMLNYCME	120
	* : : * : : * . * . * . . ** * * * : * : : :	
scaffold_2252_nlr_1	-----ELKHCFVYCAVFPQGYILEKDKVVQMWWAQ	150
scaffold_1587_nlr_1	HCLEILALSYYDLPYNMKFCFLYLGAFPQGSEISASKLTKLWIGD	165
scaffold_1429_nlr_1	TVHDILKLSSLDDLPHYLQNCFLYCSSFPEDYRIOQSDMLIWLWVAE	165
scaffold_1429_nlr_2	TVHDILKLSSLDDLPHYLQNCFLYCSSFPEDYRFQSDRLIRLWVAE	165
scaffold_763_nlr_1	KIMPILKLTYNHLPAAHLKQCFACCSVFPKEYDFDKDDLVRMWMGQ	165
scaffold_1186_nlr_1	YIMAVLRLSYDHLPDTDLQLCFRYCSLFPQDYTFKRKKLVLNMWIGS	165
	: : * * : : * * : : . : : : * : :	

NLR-Annotator output - sequence alignment of NB-domain



Maximum Likelihood tree
Built with RaxML

Advantages of the different NLR annotation methods

Evidence-based

Sequence-based

Advantages of the different NLR annotation methods

Evidence-based

More likely to get the ‘true’ or ‘active’ NLRs - this represents the plant’s potential defense gene repertoire (important for breeders/pathologists)

Sequence-based

No genome masking required
Can identify pseudogenes - not important for current resistance but may be building blocks for future genes or teach us about evolutionary history

But - identifies regions, not genes - is not splice-aware

Misses certain classes of NLRs

Advantages of the different NLR annotation methods

Evidence-based

More likely to get the ‘true’ or ‘active’ NLRs - this represents the plant’s potential defense gene repertoire (important for breeders/pathologists)

Sequence-based

No genome masking required
Can identify pseudogenes - not important for current resistance but may be building blocks for future genes or teach us about evolutionary history

But - identifies regions, not genes - is not splice-aware

Misses certain classes of NLRs

For pan-genome comparisons - probably fine as long as you are consistent in your method across samples

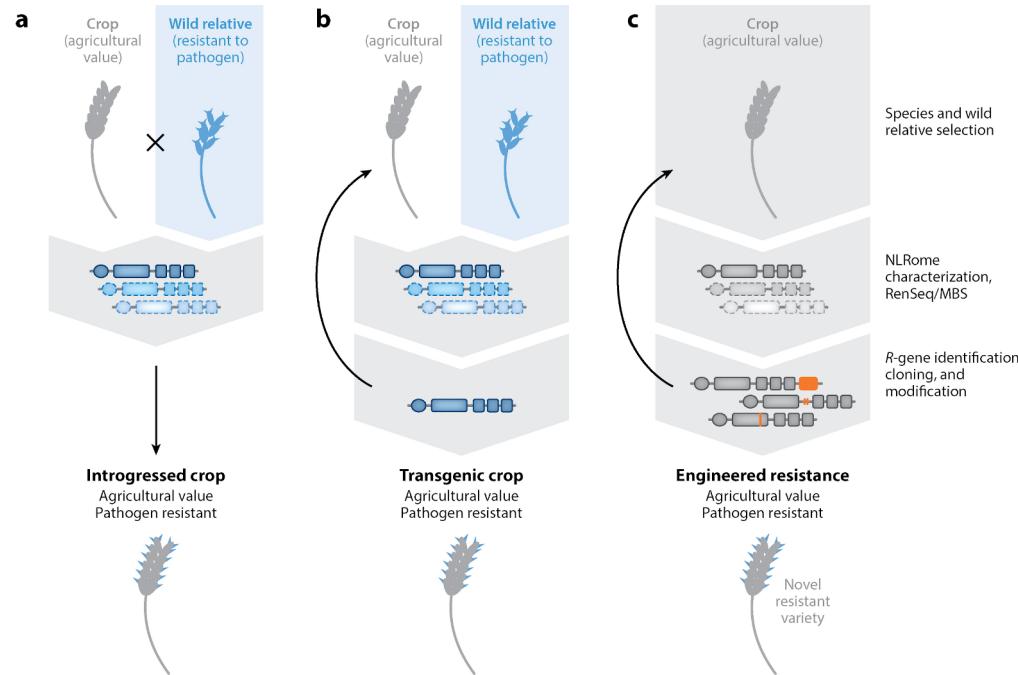
How to resolve the two predictions?

Repeat NLR Annotation (whichever method) on all cultivars
in your pangenome

But how many genomes do you need to capture the pan-
NL Rome?

As sequencing gets cheaper, and annotations get better, pangenomes will be more readily available... How will you leverage the data?

Pan-genomes can inform disease resistance breeding

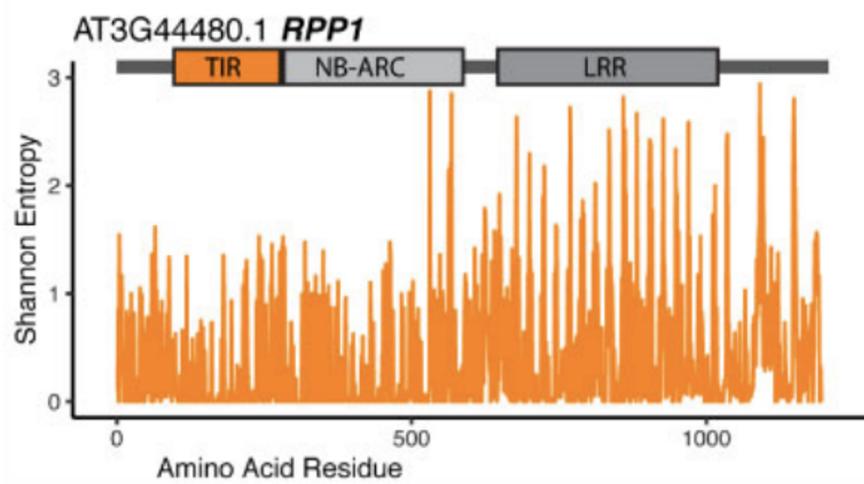
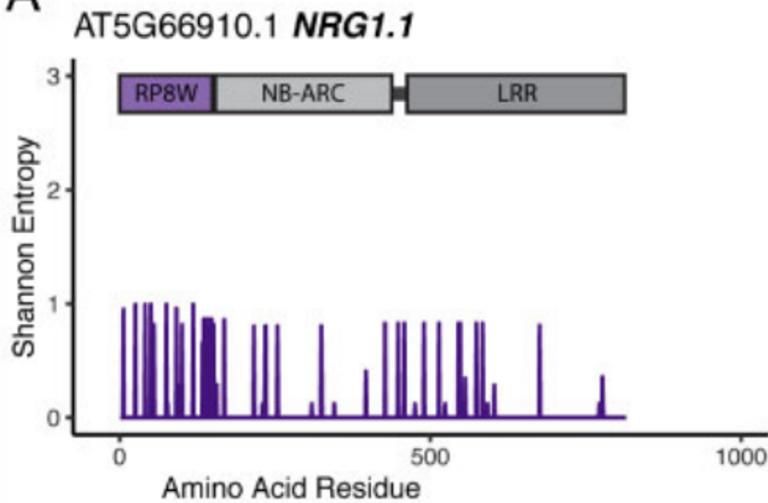


Monteiro F, Nishimura MT. 2018.
Annu. Rev. Phytopathol. 56:243–67

Structural, Functional, and Genomic Diversity of Plant NLR Proteins: An Evolved Resource for Rational Engineering of Plant Immunity Annual Review of Phytopathology Vol. 56:243-267 (Volume publication date August 2018)

Engineered recognition - a disease resistance ‘grail’

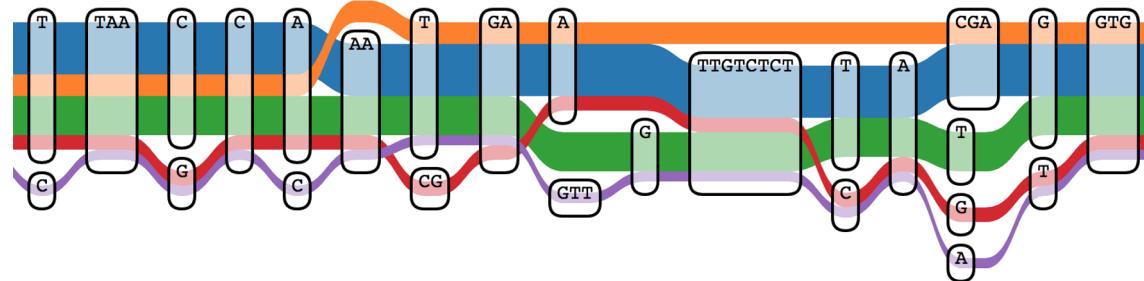
A



High entropy values correlate strongly with surface exposure and hydrophilic character (Liao et al., 2005) and can be used to predict rapidly evolving ligand-binding sites

Thanks for riding, and one last thing....

We've talked about NLR structure, diversity, and identification as well as the value of pangenomes -- **take five minutes or so to write an exam question that requires synthesis of some of this information to answer, if your question is selected for the exam you will receive extra credit :)**



@hashtag_read
read0094@umn.edu