Using a FAIR database and bioinformatics analyses to improve plant, human, animal and ecosystem health

Martin Urban and Alayne Cuzick
Kim Hammond-Kosack

UK node

The Pathogen-Host Interactions database: PHI-base

A multispecies database used to understand the mechanisms underlying pathogenesis (disease formation) and explore new options to protect crop, human, animal and ecosystem health

What is PHI-base?

- Web accessible relational **phenotype** database that **catalogues genes whose contribution to the pathogenic process has been tested experimentally**

  - **Manually curated** by domain / species experts
  - **Entries supported by peer reviewed literature** references
    (*pathogen gene deletion / gene modification vs WT strain*)
  - Broad scope of pathogen and host species
  - Positive and **negative data** included

- Freely accessible at **www.phi-base.org** since **2005**
- **Regular users based in 130 countries**
- **> 12,000 users / annum**

Winnenburg et al. (2006 and 2008) Nucleic Acids Research (Database issues)
Urban et al. (2015 and **2017**) **Nucleic Acids Research (Database issues)**
Urban et al. (2015) Frontiers in Plant Sciences, doi: 10.3389/fpls.2015.00605
Brown et al. (2016) FEMS Microbiological Reviews 40, 19-40

Motivation for Pathogen-Host interactions database

- Major crop disease outbreaks, human and animal epidemics
- Emerging fungicide / anti-infective resistance

PHI-base is a multi-species database for fungi, protists and bacteria pathogens (Version 4.8- Sept 2019)

60% plant pathogens, 35% human_animal_fish pathogens, 5% insect_others

Total ~ 250 pathogenic species (fixed)

Top 15 pathogen species in PHI-base

Host Plant Focus
- Agricultural crops
- Horticultural crops
- Commercial trees
  (timber, fruit)
- Model species
- Emerging crop plant
  threats (pandemics)

Host Animal Focus
- Key human pathogens
- 3Rs species alternative
- Plant attacking nematodes
- Plant attacking insects
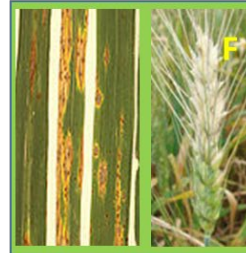
Fusarium graminearum (F)
Magnaporthe oryzae (F)
Ralstonia solanacearum (B)
Salmonella enterica (B)
Xanthomonas oryzae (B)
Erwinia amylovora (B)
Candida albicans (F)
Pseudomonas aeruginosa (B)
Botrytis cinerea (F)
Ustilago maydis (F)
Aspergillus fumigatus (F)
Cryptococcus neoformans (F)
Pseudomonas syringae (B)
Escherichia coli (B)
Staphylococcus aureus (B)
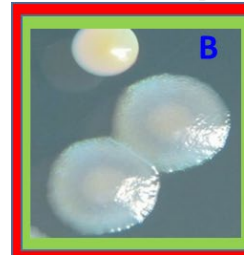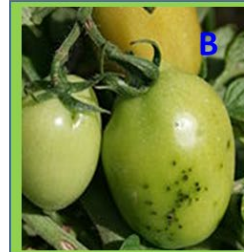
Plant pathogen   Animal pathogen   Plant and animal pathogen

Current PHI-base version 4.8 release (Sept 2019)

	www.phibase.org
References	**3454**
Genes	**8351**
Interactions	**13801**
Pathogens	268
Hosts	210
Diseases	502
Chemistries	34 fungicides 149 anti-infectives

- Pathogens: 75% eukaryotes (fungi_protists), 25% bacteria
- Hosts: 60% plant (50:50 cereal:non-cereal), 40% animal_human
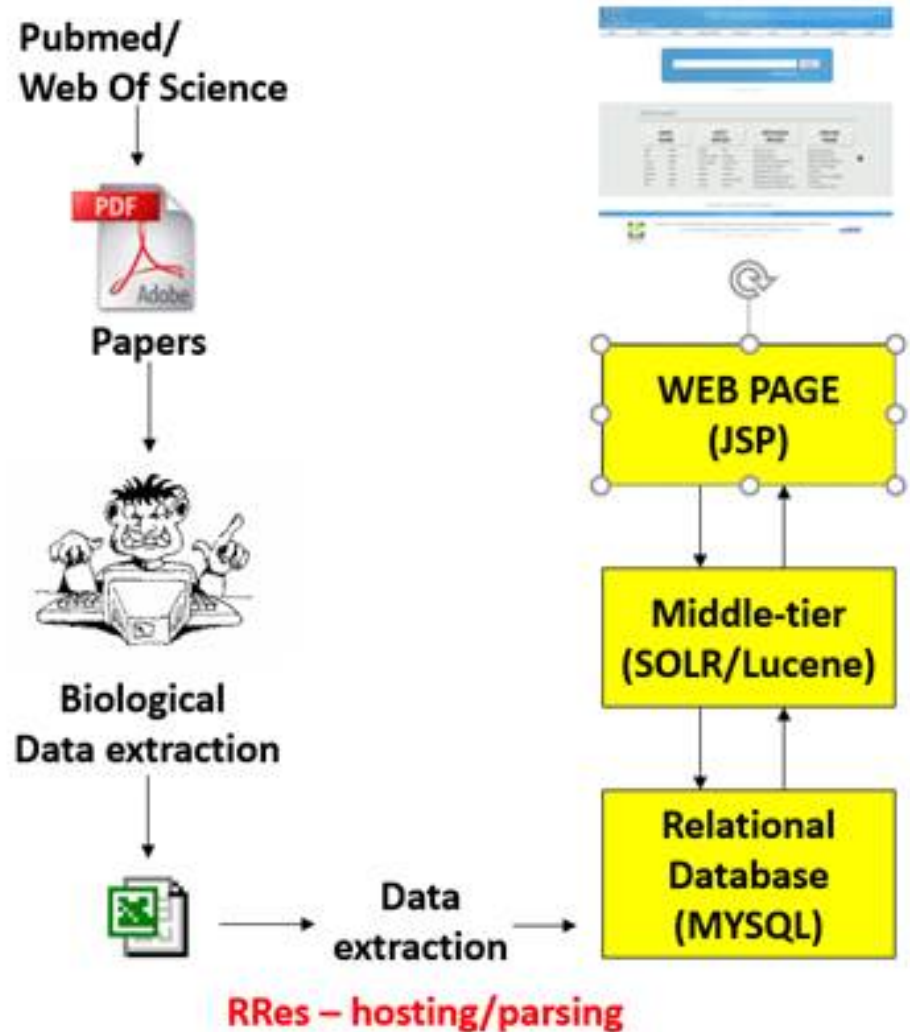
Data and knowledge explosion over the last 40 years

Increasing number of publications on virulence associated genes and their phenotypes in microbial pathogens (Web of Sciences/PubMed)
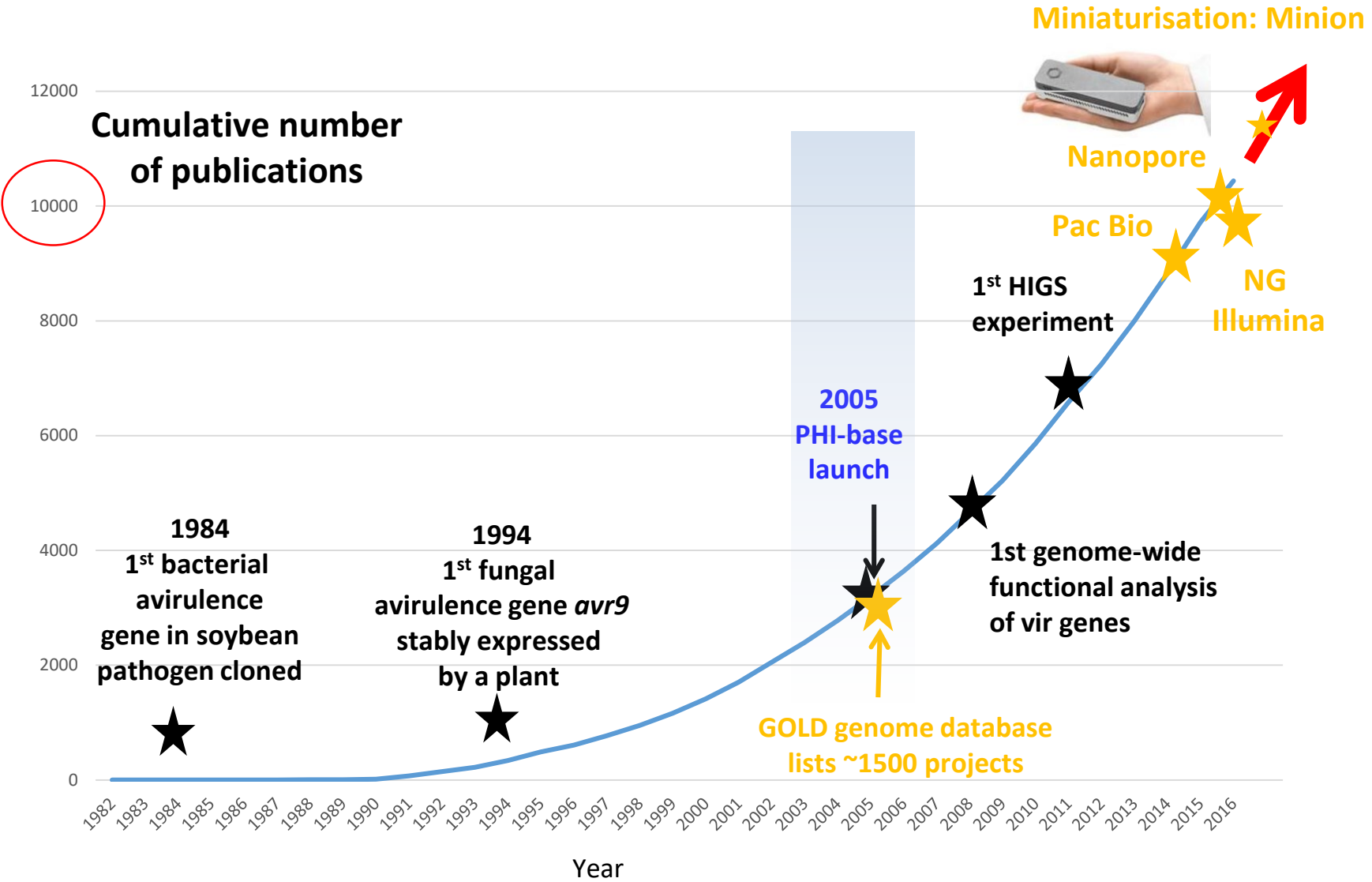
Database curation workflow
at RRes and Molecular Connections (MC) India

- **MC: keyword searches for new papers to generate candidate list (trialling machine learning)**

- **RRes: review of list and add in submissions from species experts/users, collect PDFs, send to MC**

- **MC: curation of 10 papers/week; forward data in monthly batches**
  Up to 81 data entries / gene

- **RRes: review of data 2 database releases/year**

Pubmed/
Web Of Science

Papers

Biological
Data extraction

Data extraction

WEB PAGE (JSP)

Middle-tier (SOLR/Lucene)

Relational Database (MYSQL)

RRes – hosting/parsing

The PHI-base interface has a faceted views to allow complete visualisation of all curated PHI-base data!

PHI-base protein BLASTP search functionality

Ref: http://www.sequenceserver.com/

Main uses of PHI-base

> 360 publications citing PHI-base
All papers are cited in the about section of the database

- Quick lookup of gene mutant_ phenotype relationships
  (human readable knowledge database/computer readable)
  key word search, BLAST tool and a full download function

- **Provide phenotype annotations for genome browsers
  and knowledge networks**

- **Annotate novel datasets for candidate gene selection:**
  Newly sequenced pathogen genomes or exploring variomes (variant genomes)
  RNA-seq and microarray data sets (enrichment of virulence genes)
  Other candidate gene lists (i.e. from forward genetics screens, GWAS)

- Comparative genomics/phenomics
  – Conserved themes vs species-specific differences
  – *In silico* predictions of candidate virulence using protein-protein interaction networks

Nine high level PHI-base phenotype outcomes

Generic (species neutral) terms to permit wide inter-species comparisons

Virulence terms
1. Loss of pathogenicity
2. Reduced virulence
3. Increased virulence
4. Unaffected pathogenicity
5. Effector (transferred 'entity' in an interaction)
6. Enhanced antagonism (endophytes)
7. Essential
8. Resistance to chemistry
9. Sensitive to chemistry

FRAC FUNGICIDE RESISTANCE ACTION COMMITTEE

ChEMBL

35 fungicides

Terms are routinely published in research articles but mapping to GO terms is not supported due to their high-level nature.

Urban et al. (2015), Frontiers Plant Science, doi: 10.1093/nar/gku1165

High level phenotypes are assigned to each host-pathogen interaction

An **interaction** is defined as the function of one gene, on one host and one tissue type from one publication. **One high-level phenotype term is assigned.**

	Host/tissue	Interaction	Phenotype
	Wheat ear	1	Reduced virulence
	Wheat seedling	2	Reduced virulence
	Maize cob	3	Unaffected pathogenicity

Tri5 mutant

Pathogen species:
Fusarium graminearum

This permits narrow and wide intra- and inter-species comparisons

Summary of phenotypic outcome content in PHI-base (ver 4.8, Sept 2019)

PHENOTYPIC OUTCOME	%
Loss of pathogenicity	6.6%
Reduced virulence	44.4%
Increased virulence	4.6%
Unaffected pathogenicity	26.4%
Effector (transferred 'entity')	16.4%
Enhanced antagonism	0.1%
Essential	1.3%
Resistance to chemistry	0.3%
Sensitive to chemistry	0.1%

Integration of PHI-base data with Ensembl Genomes
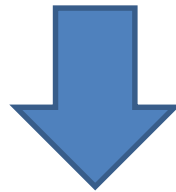

since 2010

Data:
Phenotype
Experimental information

Data:
Reference genomes for pathogens
Variation
Comparative analysis (**Compara Tool to explore closely related sub-taxa**)

Phenotype information mapped onto Ensembl Genomes
(direct hits (**UniProt Ids**) and sequence similarity (**100% identity BlastP within sub-taxa**)

> - Ensembl Fungi: 166 genomes
> - Ensembl Protists: 44 genomes
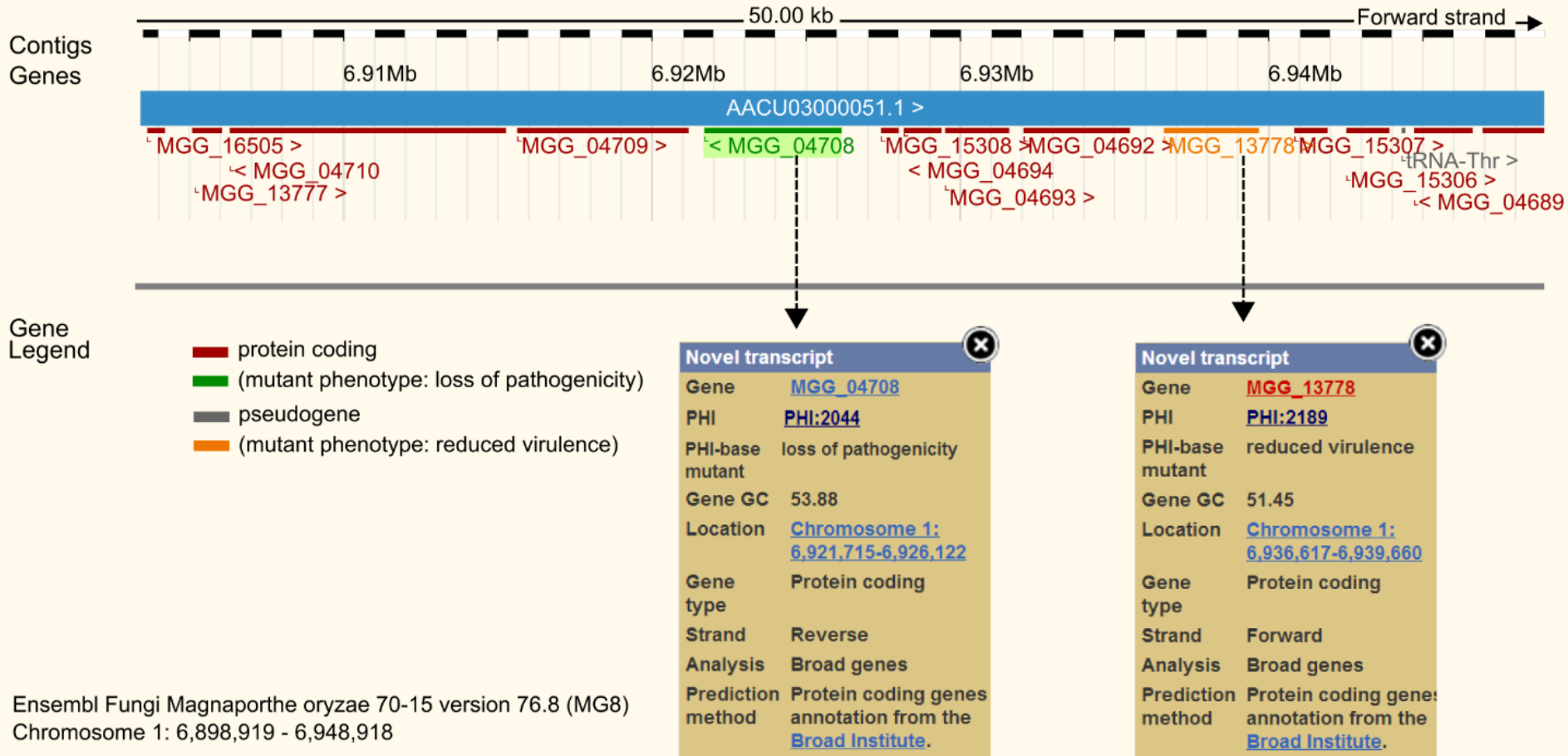> - Ensembl Bacteria: 820 genomes

BioMART search for PHI-base annotated pathogen genes

Display phenotypes directly in Ensembl Genomes browser

Ensembl Fungi

Magnaporthe oryzae
Magnaporthe oryzae 70-15

Ensembl Fungi Magnaporthe oryzae 70-15 version 76.8 (MG8)
Chromosome 1: 6,898,919 - 6,948,918

fungi.ensembl.org

Search *Zymoseptoria tritici* genes by phenotype using the BioMart tool

fungi.ensembl.org

Full BioMart results output and export options

Using the COMPARA tool to explore homologues in other taxa

Name	Transcript ID	bp	Protein	Biotype	UniProt	Flags
Novel	Mycgr3T98311	1514	460aa	Protein coding	F9WW98	

Gene Tree ENSGT00930000004479
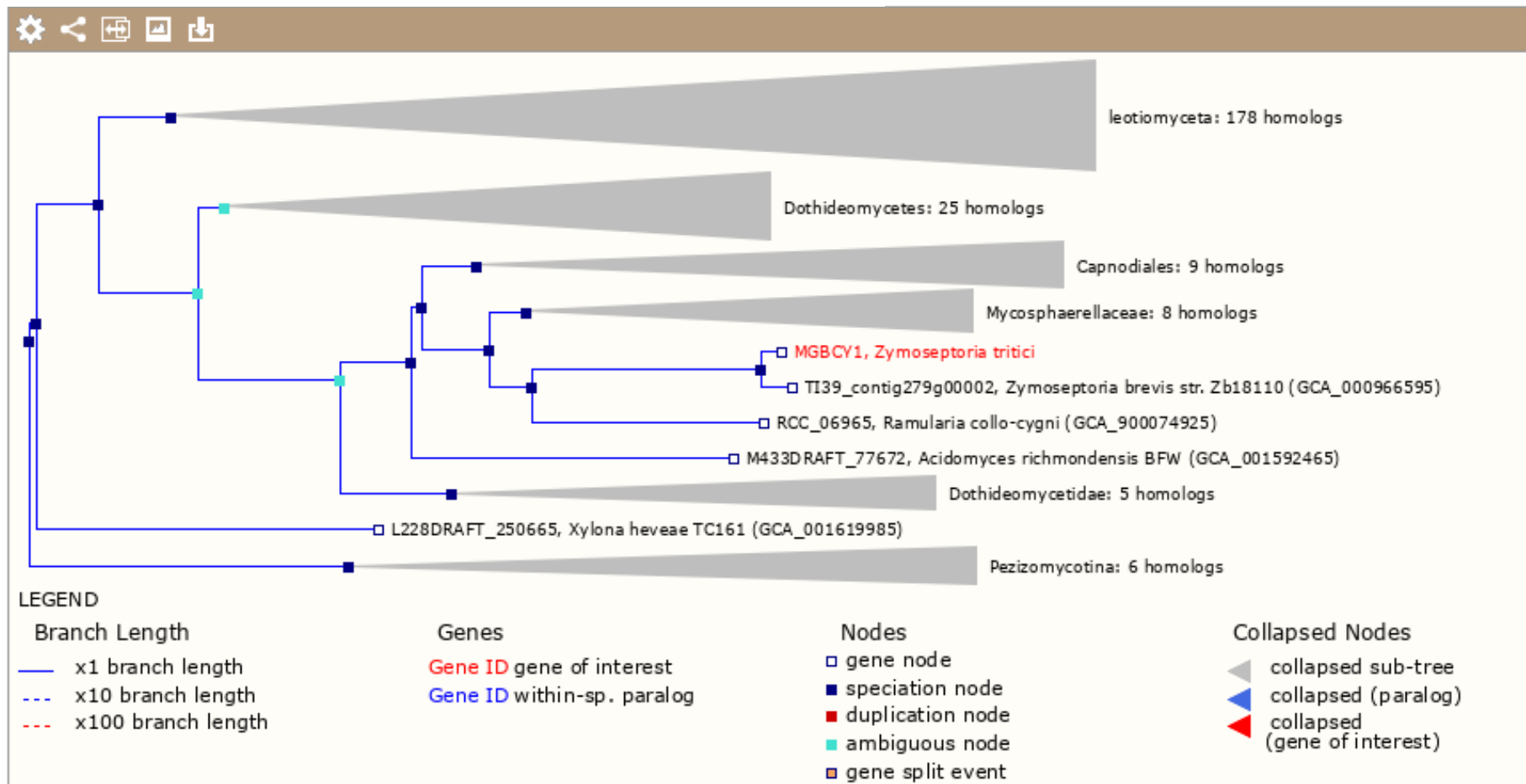
Two new features for PHI-base since 2017

1. Improved author curation tool: PHI-Canto

2. Curation of the first host targets in plants of pathogen effectors

Species expert checks
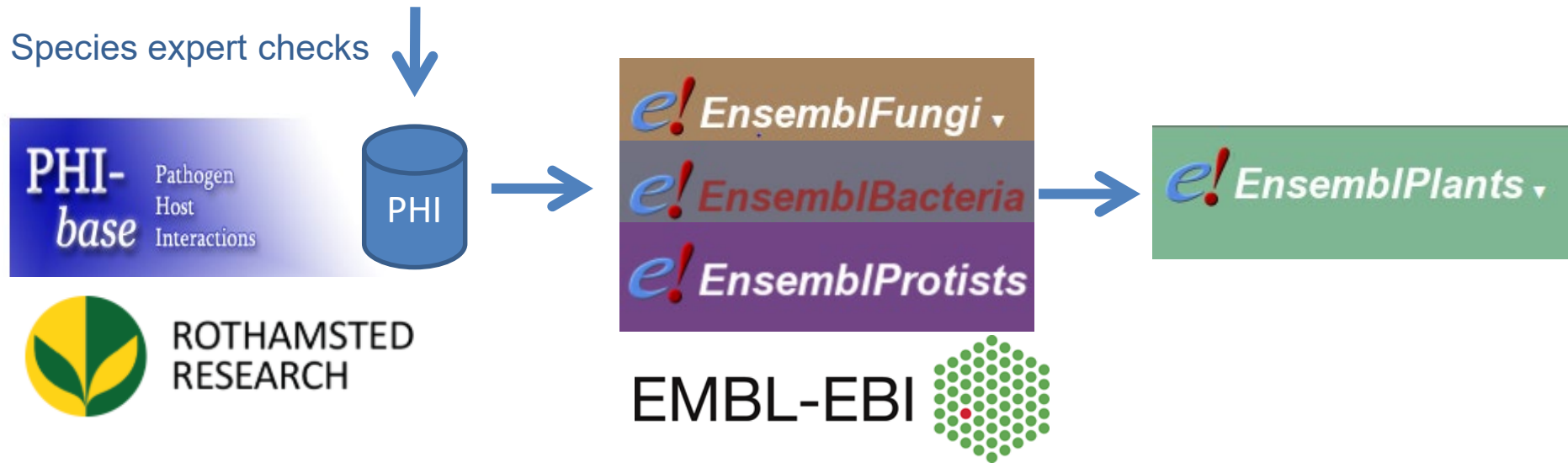
Further integration of PHI-base data with Ensembl Genomes and the BioMart tools

Community multi-species curation tool for authors PHI-Canto

http://curation.phi-base.org (beta version)

Start with Manuscript ID,

enter details including gene IDs (UniProt),
phenotypes, submit

Anticipated curation time per article: 15-30 min

New BBSRC-BBR grant (July 2019 for 3 years)
Focus: Further PHI-Canto, metagenotype and ontology developments

PHI-base uses that are relevant to EPPO / Euphresco mission in the era of metagenomics

Emerging infectious diseases (EIDs)

Typical workflow

➡ Identify the problematic pathogen species taxa

    ➡ Full genome sequencing, assembly and ORF predictions

        ➡ **Annotation**

        Limited gene set
        Full genome annotation

PHI-base - hypervirulence category (4.6%) - ability to cause more disease
    635 interactions   (529 genes) - 108 pathogenic species
    244 interactions   (198 genes) - 49 plant pathogenic species
    29 fungi, 15 bacteria
    4 protists, 1 nematode

    Rice and grape vine hosts have the highest number of entries

Magnaporthe oryzae (rice blast pathogen) – 10 hypervirulence entries
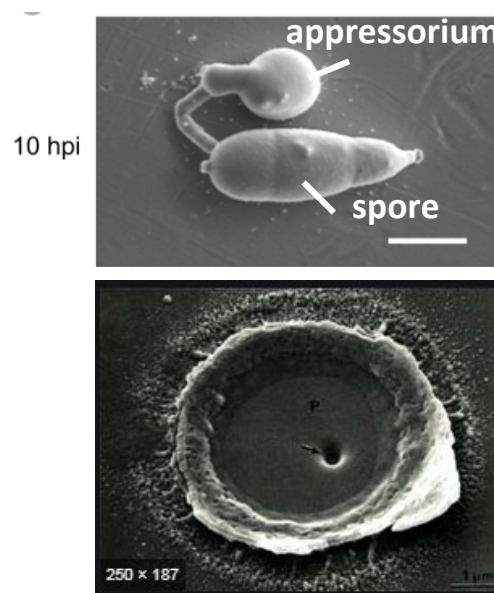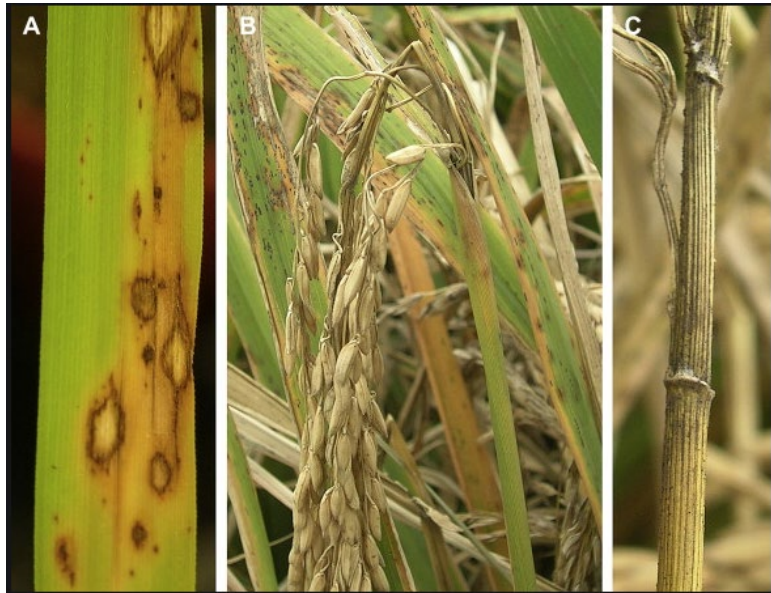
Negative regulator – RGS1 (intracellular signalling)

appressorium

spore

10 hpi

250 × 187

WT rgs1Δ

Non-inductive

Germ tube

Conidium
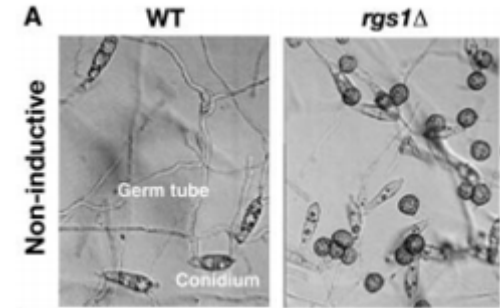
more appressorium

rgs1 mutant produces more asexual spore

Many fungal pathogens need to produce an appressorium to penetrate plant cell surfaces including:

Powdery mildews, the rust and Colletotrichum species

Exploiting knowledge on a generic 'lifestyle' theme

Xylella fastidiosa (Grape vine infecting bacteria) - 13 hypervirulence entries

2013 1st recorded in Europe
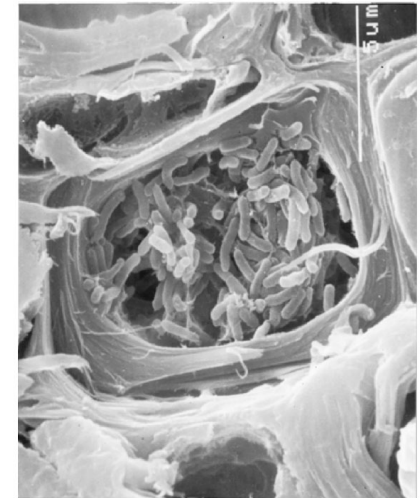
Gram-negative bacteria in xylem vessels

Outer membrane vesicles (OMVs) produced by Xylella, block its interaction with various surfaces including xylem vessel walls. **OMVs stops bacterial adhesion**

Low OMVs (bacteria are static)
- insect vector acquisition ➡ **High OMVs – systemic bacteria spread in host**

A X. fastidiosa **ΔrpfF mutant** in **which quorum signalling is disrupted**, has increased OMV release (5 fold), and as a result **adheres less to plant cell surfaces**, becomes **more systemic** within the olive plant and **causes more disease.**

Exploiting knowledge on a generic 'lifestyle' theme

Emergence of a new disease as a result of interspecific virulence gene transfer

Parastagonospora nodorum causes Septoria nodorum blotch (SNB) on wheat

Pyrenophora tritici-repentis causes Tan Spot on wheat
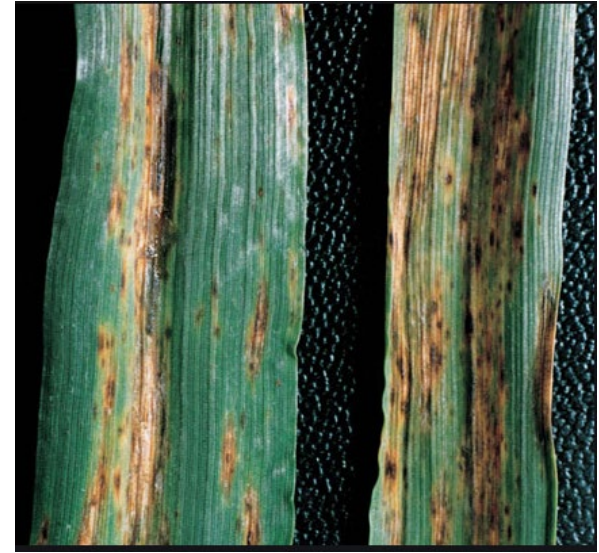
ca ~ 1940

~11kb DNA Tox A gene

Wheat target is Tsn1 (susceptibility locus)

Emergence of new diseases as a result of acquisition or modification of a secondary metabolite cluster

Tracking pathogen effectors and their 1ˢᵗ host targets

An effector is an entity produced by a pathogenic species to facilitate the infection process
- activate or suppress host defences or other host responses
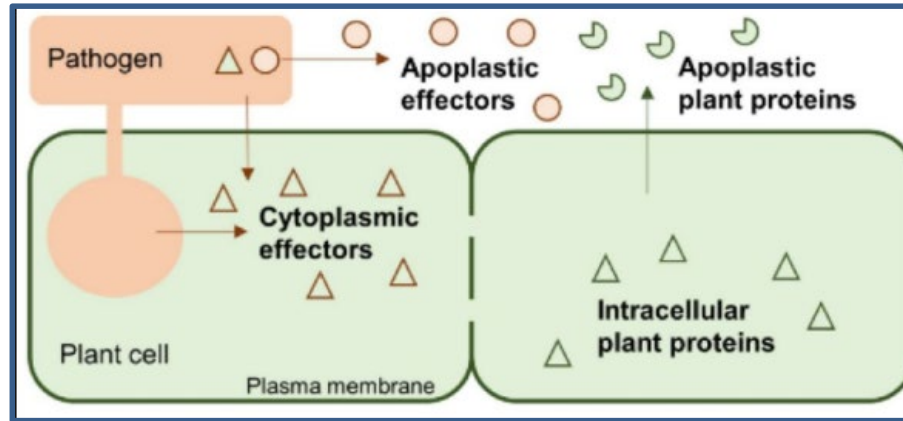
PHI-base data – Version 4.8
2,261 effector entries (16%) coming from 83 species, mostly plant pathogens

PHI-base started to curate 1ˢᵗ host targets - soon to be displayed in ENSEMBL

Most effectors are species / strain specific
– maximum discovery by exploring the variome

But a few are generic and essential for pathogenicity   - chitin binding
Ecp6, 3LysM

PHI-base entries can contribute to plant breeding via effectoromics

Effectoromics, a high-throughput functional genomics approach that uses effectors to probe plant germplasm to detect disease resistance (*R*) genes.
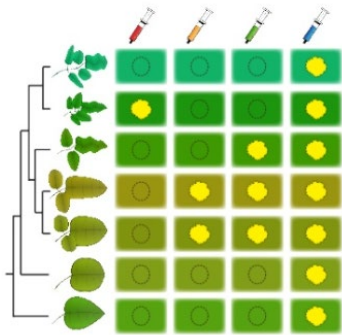
Effectoromics
- **accelerates *R* gene identification**
- **distinguishes functional redundancy**
- **pinpoints recognition specificities**
- **assists *R* gene deployment**

Application particularly successful

Phytophthora infestans **(late blight) – potato /** *Solanum tuberosum* **model /** *Nicotiana benthamiana*

Experimental techniques - agroinfiltration and agroinfection

Breeders' aim to maintain the widest possible repertoire of disease resistance genes in elite germplasm

Discovery of novel anti-infective targets

Lethals **183 entries in PHI-base (1.3%) - 101 are plant pathogen entries**

4 fungi, 3 bacteria

Experimental evidence - no transformants recovered when deleting a single gene

Use this new information about 'essential for life' genes to nominate novel intervention targets to be explored by bioscientists and the AgChem /AgBiotech industries

'Essential for life' proteins to be included in chemical screens – natural products and / or synthetic chemistries

High likelihood that 'essential for life' genes are conserved between species

Pan-genome analyses

Core genomes vis variable genome

Variable genome — Genome strain 1 — Core genome

Genome strain 2 — Genome strain 3
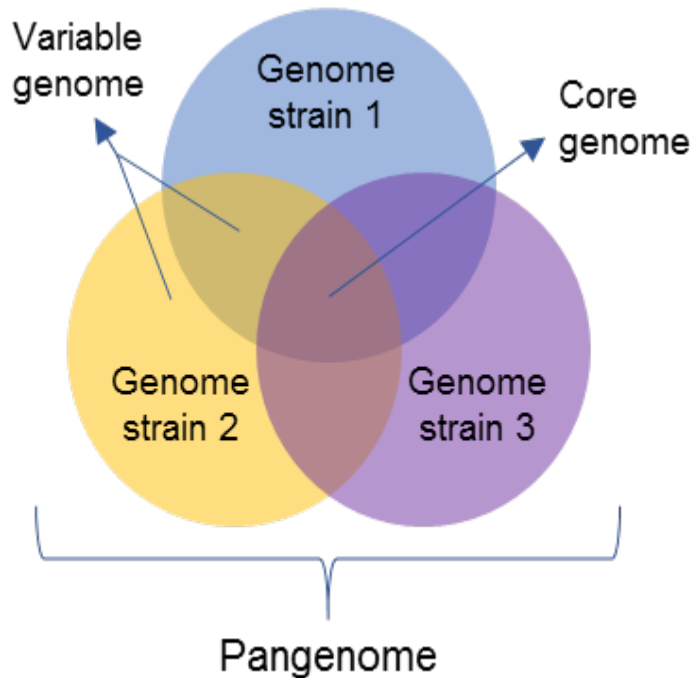
Pangenome

Historic isolates compared to **newer problematic isolates**

**Previous control practices** compared to **existing / new control practices**

Fungicide use ➡ Biocontrol options plus *R* gene use

Pinpoint whole genome evolutionary changes

- **Identify which genome locations and gene sets need to be monitored**

Monitoring specific genome locations for changes

Pathogen Species A – minimal repetitive DNA or transposons

SNPs

recombination frequencies (cM)

genes not found in related taxa
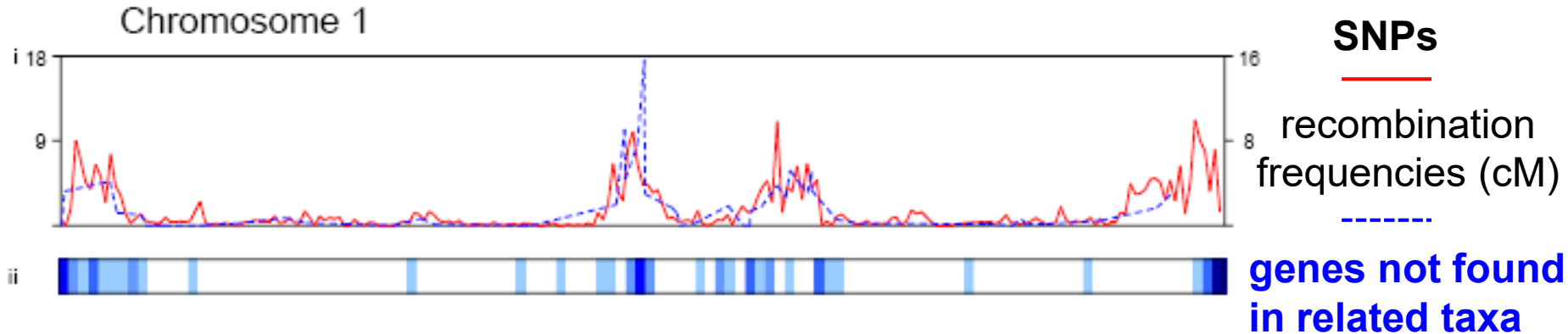
Pathogen Species B - rich in repetitive DNA and transposons

■ = blocks of simple sequence repeats and / or transposons

| = **species specific gene and / or small effector gene**

PROBLEM : Repeat masker software obscures these important sequences

Pathogenicity and virulence networks

**Using well explored pathogens to extrapolate knowledge
onto the under-explored species**

Taxonomic and / or lifestyle relatedness

Cereal infecting

Fusarium graminearum – **1571** interactions

Fusarium verticillioides – 55 interactions

5 other Fusarium species – 21 interactions

Non- Cereal infecting

Fusarium oxysporum – **209** interactions

Fusarium solani – 6 interactions

2 other Fusarium species – 5 interactions

**New types of interspecies comparative analyses that also
take advantage of lifestyle similarities or differences**

Predictive biology - Network analyses

Graphical visualisation analyses of protein-protein interaction network analyses to predict potential Achilles' heels for particularly problematic species
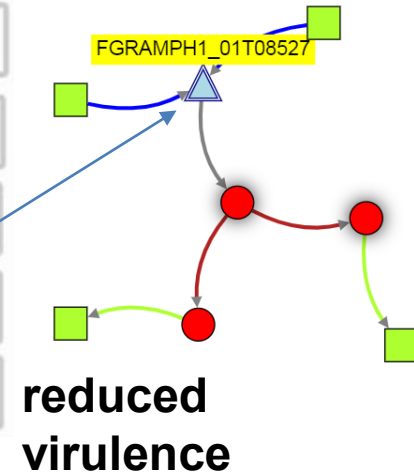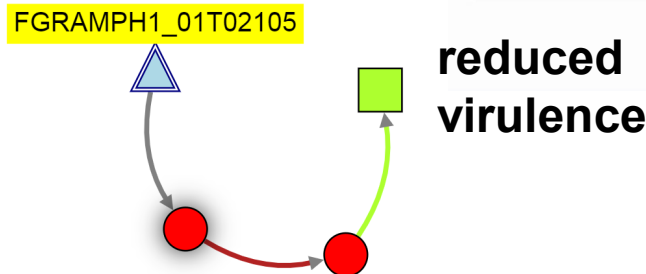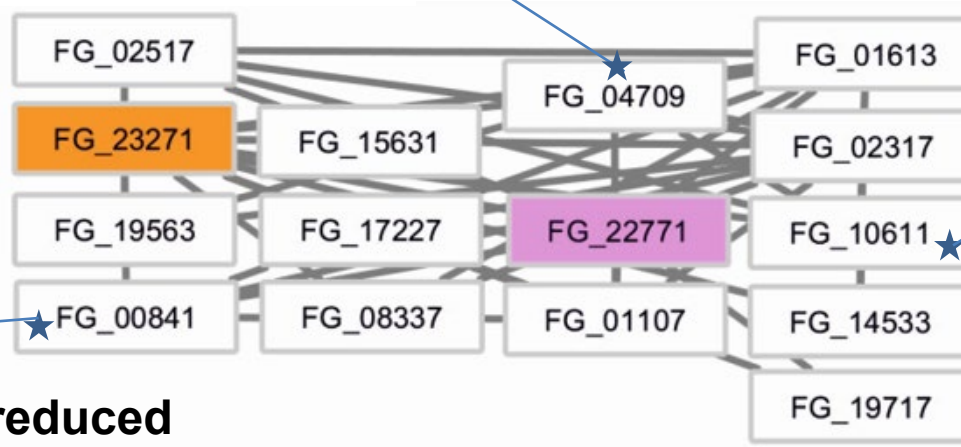
Knetminer
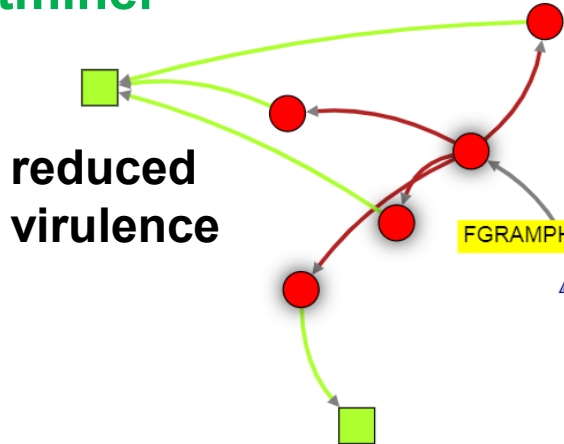
reduced virulence

PPI networks based on interlogs and domains
3,800 – 8,300 nodes / 150,000 to 450,000 edges
PHI-Nets - visualised using Cytoscape

Fusarium graminearum – sub-network

= predicted siRNA effector target site
= reduced virulence in *Fg*

FGRAMPH1_01T16125

FGRAMPH1_01T08527

FGRAMPH1_01T02105

FG_02517 FG_01613
FG_04709
FG_23271 FG_15631 FG_02317
FG_19563 FG_17227 FG_22771 FG_10611
FG_00841 FG_08337 FG_01107 FG_14533
FG_19717

reduced virulence

reduced virulence

PHI-base Universe - FAIR Data Principles

PHI-base users/associated databases

Findable
Accessible
Interoperable
Reuseable
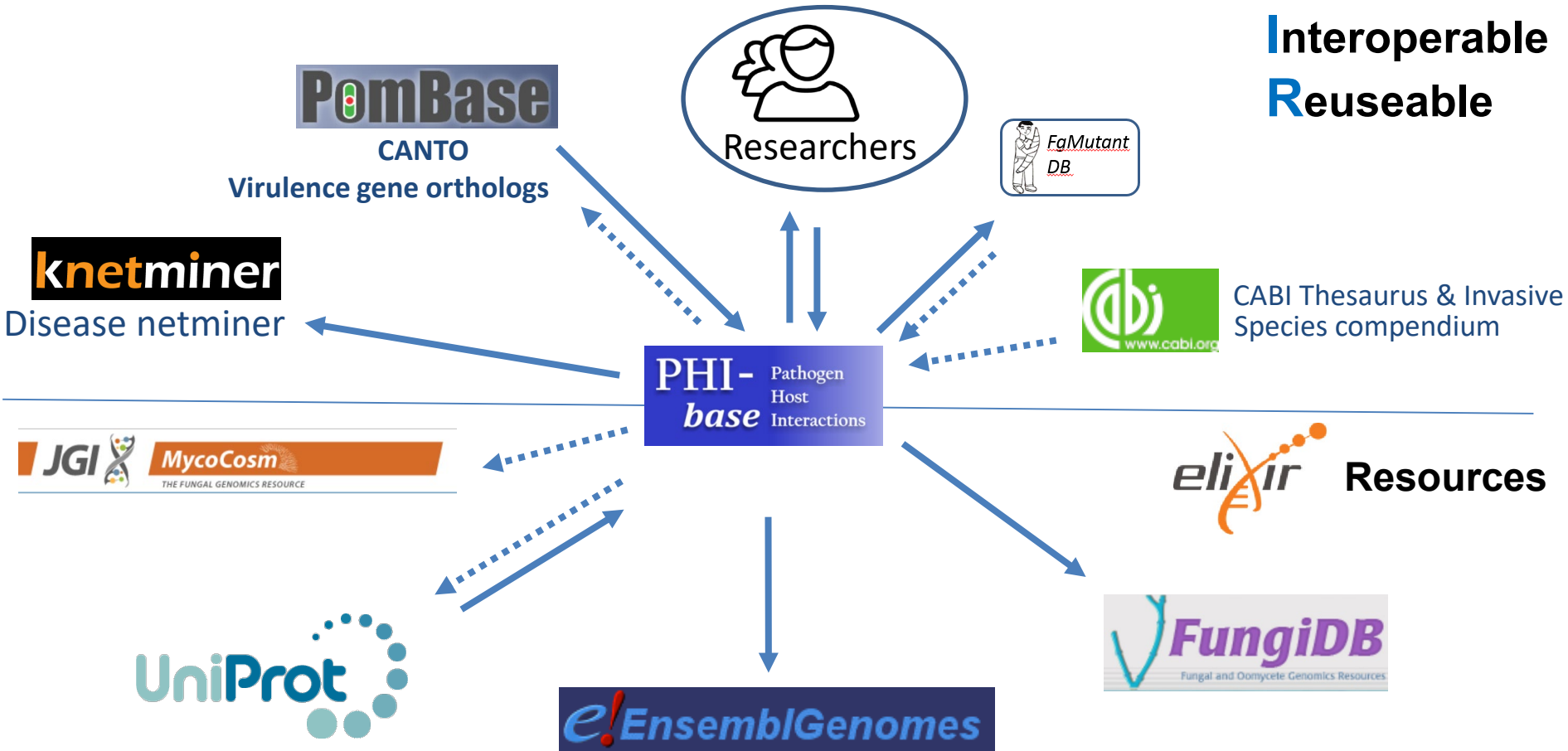
Acknowledgements

PHI- Pathogen Host base Interactions

From genes to mutant phenotypes

www.phi-base.org

ROTHAMSTED RESEARCH

DFW

MOLECULAR CONNECTIONS
Knowledge Solutions for Accelerated Research

BBSRC
bioscience for the future

EMBL-EBI

elixir
UNITED KINGDOM

Rothamsted Research

Kim Hammond-Kosack

Martin Urban

Alayne Cuzick

James Seager

Barnaby Norman

Dan Smith

Chris Rawlings

William Brown

Elzbieta Janowska-Sejda – PHI-Nets

Artem Lysenko –PHI-Nets

Keywan Hassani-Pak - Knetminer

Collaborators

Individual species experts globally (~50)

Phytopath team @ EBI

Paul Kersey   Nishadi De Silva

Helder Pedro   Manuel Carbajo

Uma Mareshwari

PomBase *The scientific resource for fission yeast* **[ ] Canto**

Valerie Wood

Kim Rutherford

Midori Harris

Thank you