Combined reference-free and multi-reference approaches uncover cryptic variation underlying rapid adaptation in microbial pathogens

Anik Dutta[1,#], Bruce A. McDonald[1], Daniel Croll[2, *]

[1] Plant Pathology, Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland

[2] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, 2000 Neuchâtel, Switzerland

[#] Current address: Institute of Phytopathology, Kiel University, 24118 Kiel, Germany

*Author for correspondence: daniel.croll@unine.ch

Abstract

Background: Microbial species often harbor substantial functional diversity driven by structural genetic variation. Rapid adaptation from such standing variation in pathogens threatens global food security and human health. Genome wide association studies (GWAS) provide a powerful approach to identify genetic variants underlying recent pathogen evolution. However, the reliance on single reference genomes and single nucleotide polymorphisms (SNPs) obscures the true extent of adaptive genetic variation. Here, we show quantitatively how a combination of multiple reference genomes and reference-free approaches captures substantially more relevant genetic variation compared to single reference mapping.

Results: We performed reference-genome based association mapping across 19 reference-quality genomes covering the diversity of the species. We contrasted the results with a reference-free (i.e., K-mer) approach using raw whole genome sequencing data. We assessed the relative power of these GWAS approaches in a panel of 145 strains collected across the global distribution range of the fungal wheat pathogen *Zymoseptoria tritici*. We mapped the genetic architecture of 49 life history traits including virulence, reproduction and growth in multiple stressful environments. The inclusion of additional reference genome SNP datasets provides a nearly linear increase in additional loci mapped through GWAS. Variants detected through the K-mer approach explained a higher proportion of phenotypic variation than a reference genome based approach, illustrating the benefits of including genetic variants beyond SNPs.

Conclusions: Our study demonstrates how the power of GWAS in microbial species can be significantly enhanced by comprehensively capturing functional genetic variation. Our approach is generalizable to a large number of microbial species and will uncover novel mechanisms driving rapid adaptation in microbial populations.

Keywords: genome-wide association mapping, single nucleotide polymorphisms, K-mer, multiple-reference-genome, *Zymoseptoria tritici*

Introduction

Rapid genetic change in microbial pathogens has led to significant damage to agricultural production as well as to human health over recent decades (Casadevall et al. 2011; Fisher et al. 2012; Figueroa et al. 2018). The rapid evolution in pathogen populations of virulence and resistance to anti-microbial drugs are key concerns in plant, animal and human health. There is an urgent need to identify the precise genetic determinants in pathogens that underlie differences in virulence and evasion of control mechanisms. Vast genomic datasets can now be exploited to retrace evolutionary pathways of pathogen adaptation. Association mapping can be used to establish relationships between genetic and phenotypic variation using field collections of pathogens (Bartoli and Roux, 2017; Sánchez-Vallet et al. 2018). The genetic variation relevant for trait evolution is often more complex than the commonly used single nucleotide polymorphisms (SNPs). Structural variants (SVs) such as insertions-deletions (indels), copy number variants, chromosomal rearrangements, inversions and duplications can also be major facilitators of microbial adaptation (Dutilh et al. 2013: Plaumann et al. 2018; Zeevi et al. 2019; Allen et al. 2021; Langner et al. 2021). For plant studies, powerful approaches were recently proposed to associate SVs to causal genes controlling trait variation (Todesco et al. 2020; Guo et al. 2020). However, our understanding of SVs governing trait variation in microorganisms is limited by approaches focused on SNPs (Laabei et al. 2014; Pereira et al. 2020b; Singh et al. 2021). Microbial genomes are highly plastic in terms of gene content and associated SVs. GWAS based on a single reference-genome can only capture the gene content described in that single genome (Lees et al. 2016). Using a compilation of reference genomes to construct a pangenome resource that integrates a more comprehensive set of the genes present in a pathogen species shows substantial promise (Badet and Croll, 2020). The ability to integrate various types of SVs while performing association mapping will also substantially expand our understanding of microbial adaptation.

Pathogen adaptation is frequently governed by genetic determinants termed accessory genes that are not shared among all individuals of a species. Accessory genes were found to affect defense responses, virulence, drug resistance and environmental adaptation (Holt et al. 2015; Sánchez-Vallet et al. 2018; Wu et al. 2018; Zou et al. 2019). The detection of such adaptive accessory genes can be accelerated by expanding GWAS to include multiple reference genomes covering distinct segments of the gene space of a species. Additionally, single reference genome based GWAS can

be confounded by gene presence/absence variation as such variation is challenging to account for (Gage et al. 2019). These shortcomings of a GWAS based on a single reference genome can be overcome by repeating the mapping across multiple reference genomes representing the pangenome of a species (Tettelin et al. 2005; Bayer et al. 2020; Gupta, 2021). Recent advances in genomics are rapidly expanding the number of microbial pathogens with such pangenome resources (Baddam et al. 2014; Liu et al. 2014; Badet et al. 2020). These resources can facilitate the identification of pathogen virulence factors as well as previously unknown anti-microbial resistance factors emerging after the application of newly designed chemical control agents (Golicz et al. 2020; Allen et al. 2021). In particular, SVs in highly repetitive regions are unlikely to be captured. This can be overcome by adopting an alignment-free approach where short reads are screened for subsequences of specific length, *i.e.* K-mers (Sheppard et al. 2013; Weinert et al. 2015). A major advantage of K-mer based analyses is the ability to capture genetic variation without depending on a reference genome, avoid SNP calling ascertainment biases or allow identifying sequence segments absent from a reference genome (Lees et al. 2016; Jaillard et al. 2018). Capturing complex SVs is particularly relevant because significant genetic variation, sometimes referred to as the "missing heritability" problem, can go undetected using traditional reference-based GWAS (Zuk et al. 2012; Rahman et al. 2018). Though their potential advantages are clear, reference-free methods to capture adaptive genetic variation remain largely unexplored in pathogenic microorganisms.

The fungal pathogen *Zymoseptoria tritici* causes septoria tritici blotch (STB), a disease that has a significant impact on global wheat production (Fones and Gurr, 2015; Torriani et al. 2015). *Z. tritici* has a highly plastic genome with 13 core chromosomes and 8 accessory chromosomes that exhibit presence-absence variation among isolates (Goodwin et al. 2011). Large effective population sizes, high gene flow and high recombination rates facilitate rapid evolution of resistance toward fungicides and virulence on resistant hosts (Croll et al. 2015; Hartmann et al. 2018, 2021; Singh et al. 2020). The pathogen population harbors substantial variation for many life history traits including growth rates, stress tolerance, melanization and reproduction on the wheat host (Dutta et al. 2021). Structural rearrangements and deletion events were found to be associated with host adaptation (Hartmann et al. 2017; Meile et al. 2018). GWAS based on single reference genomes was successful in discerning the genetic underpinnings of pathogen virulence and fungicide resistance (Hartmann et al. 2021; Singh et al. 2021). The recent pangenome

103 constructed for *Z. tritici* based on 19 different isolates from six continents showed that the pathogen
104 harbors a substantially larger gene repertoire than the canonical reference genome (Badet et al.
105 2020). Accessory genes within the species encode diverse but largely unknown functions and were
106 likely missed in previous analyses that relied on a single reference genome. Thus, expanding
107 GWAS beyond one reference genome will likely capture a larger fraction of genes underlying
108 recent adaptation.

109 Here, we assess the performance of both reference-free and multi-reference GWAS by conducting
110 a comprehensive mapping analysis based on a global set of *Z. tritici* populations. We screened for
111 sources of genetic variation affecting 49 biotic and abiotic traits. Both GWAS conducted on SNP
112 datasets mapped to 19 different reference genomes and k-mer based GWAS revealed a large
113 number of previously missed loci contributing to trait variation. Our study provides quantitative
114 insights how improved GWAS approaches can identify genetic variants underpinning adaptation
115 in rapidly evolving microbial pathogens.

116

117

118 ## Results

119 ### *A generalizable framework for conducting microbial GWAS*

120 We performed comprehensive association mapping analyses to detect genetic variants of varying
121 complexity underlying pathogen adaptation to different hosts and environments (**Figure 1**). We
122 analyzed genetically diverse pathogen populations spanning the global distribution of wheat and
123 recapitulating host diversity and climatic gradients. Isolates were phenotyped under greenhouse
124 and laboratory conditions to assess both pathogenicity-related traits (e.g., degree of host damage,
125 amount of spore production) and responses to abiotic stresses (e.g., fungicide, low temperature)
126 (Dutta et al. 2021). Genetic variation in the mapping panel was assessed in two complementary
127 ways. (1) Whole-genome sequence datasets were used to generate SNP calls on multiple reference
128 genomes. A total of 19 telomere-to-telomere reference genomes have been assembled to capture
129 the global diversity in structural variation (Badet et al. 2020). (2) Short reads were also used to
130 generate 25-bp K-mer profiles for each isolate. These presence/absence K-mer tables applied to

mapping populations are highly effective in capturing structural variation independent of a reference genome (Voichek and Weigel, 2020).

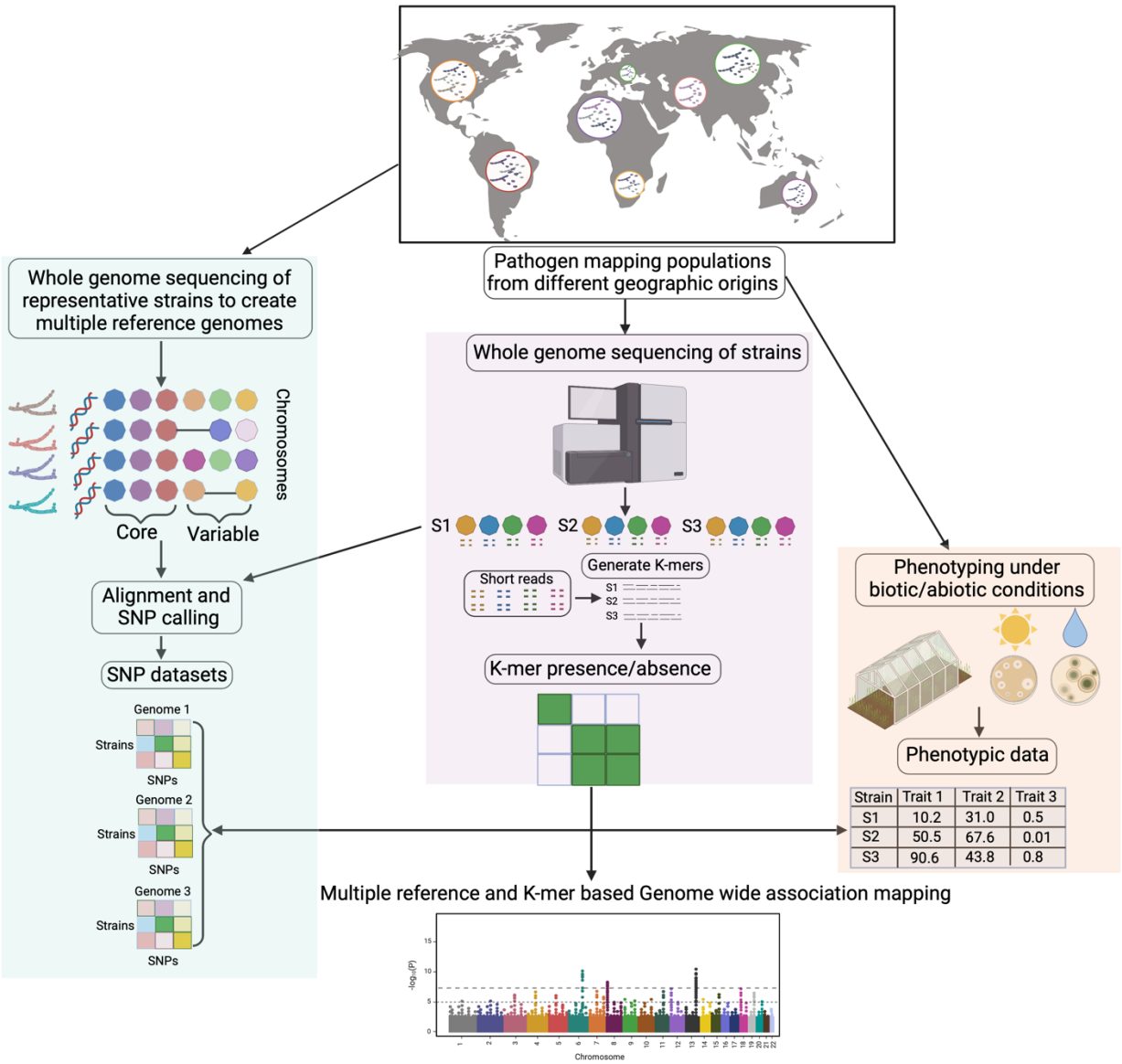


Figure 1. A comprehensive workflow for conducting microbial genome wide association studies (GWAS) using multiple reference genomes and K-mer data from mapping populations. Genetically diverse pathogen populations from different geographic locations are sampled to construct an association panel followed by greenhouse and laboratory phenotyping to assess heritable trait variation (right panel; Dutta et al. 2021). Chromosome-level genome assemblies of representative isolates is performed to generate reference genomes and establish a species pangenome (left panel; Badet el al. 2020). Whole genome sequencing of the association panel enables single nucleotide polymorphism (SNP) calling against multiple

142 reference genomes and creation of K-mer presence/absence tables (middle panel). GWAS can be performed
143 simultaneously to take advantage of SNP datasets or K-mer presence/absence tables.
144

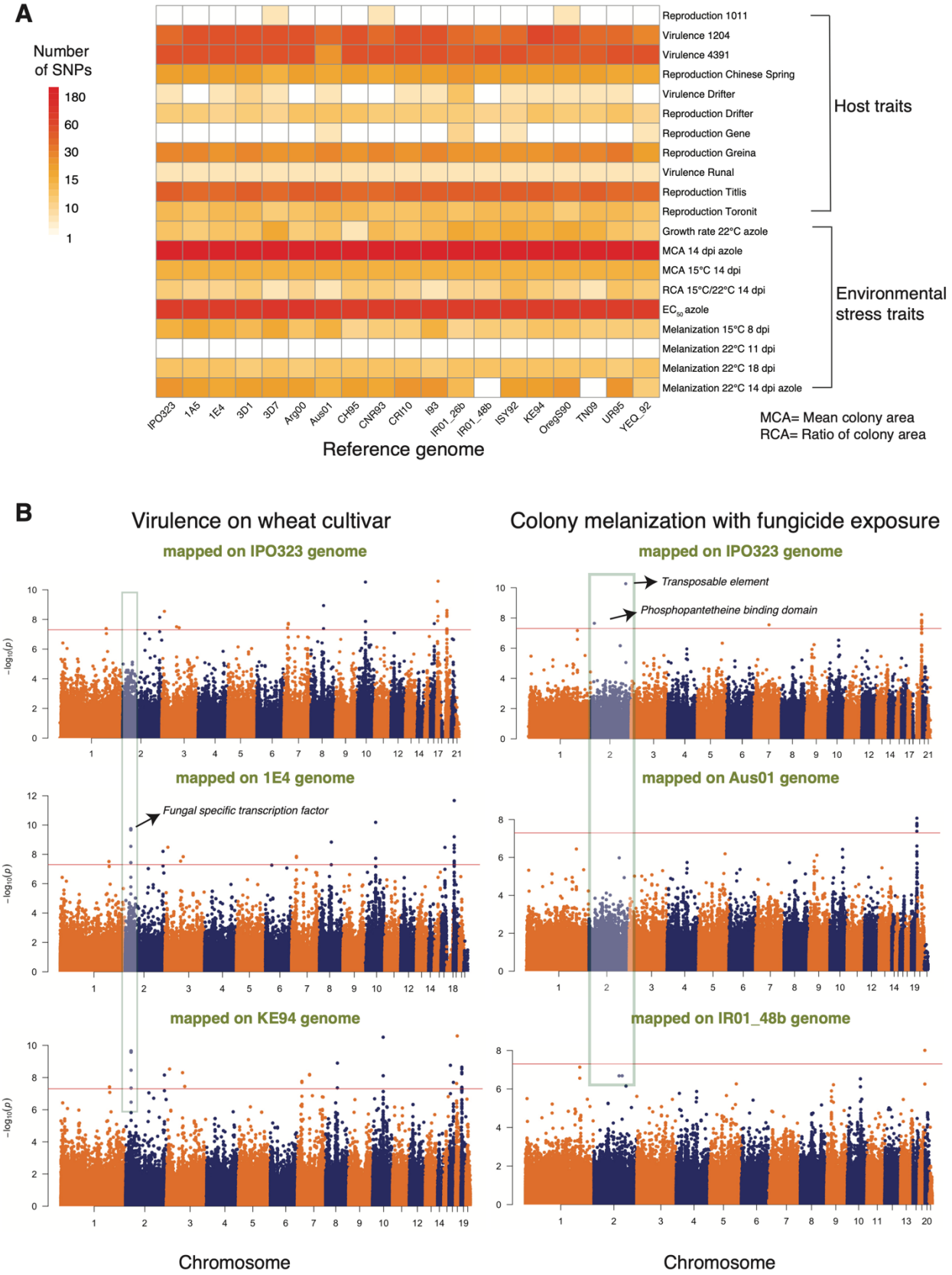
145

146 ***Multiple reference genome based GWAS***

147 We performed association mapping for a total of 49 traits including measures of virulence and
148 reproduction on twelve wheat cultivars and growth and melanization under various stress
149 conditions such as different temperature regimes and fungicide exposure. The mapping was
150 performed independently for SNP panels generated from each of the 19 reference genomes based
151 on mixed linear models. We estimated the genomic inflation factor (GIF; λ), which ranged from
152 0.91 to 1.09 without principal components as a random effect controlling for population
153 substructure, and from 0.70 to 1.36 when including principal components (**Supplementary Figure**
154 **S1**). The multiple reference-based GWAS detected a range of significant marker-trait associations
155 above the Bonferroni threshold ($\alpha = 0.05$) for a total of 20 traits related to virulence, reproduction,
156 growth rate, fungicide resistance and melanization (**Figure 2A**). We found high variability in the
157 number of significant SNPs for the same trait depending on the choice of the reference genome
158 SNP panel (**Figure 2A, Supplementary Table S5**). The number of significant SNPs ranged from
159 1-55 for pathogen virulence and reproduction on different wheat hosts depending on the reference
160 genome. The highest number of significant SNPs were identified for virulence on landrace 1204
161 with the alternative reference genome KE94 (**Figure 2B**). This trait also showed the highest
162 variance in the number of significant associations among the 19 reference genomes
163 (**Supplementary Table S5**). The number of significant SNPs for environmental stresses ranged
164 from 1-180 with the azole resistance trait showing the largest and most variable number of SNPs
165 among the 19 reference genomes. The most significant SNPs for each trait explained 3-15% of the
166 phenotypic trait variation (**Supplementary Table S6**). This suggests that numerous genes affect
167 most trait variation in most environments, consistent with polygenic architectures for most of these
168 traits.

169 A substantial fraction of all significant associations could not be mapped with the canonical
170 reference genome IPO323 (**Figure 2B**). Also, significant associations for several traits mapped in
171 to the canonical reference genome were not found using alternative reference genomes **(Figure**

172 **2B)**. This shows that multiple reference genome SNP panels can overcome limitations due to
173 presence-absence variation and challenges in SNP calling. To analyze putative gene functions
174 contributing to phenotypic trait variation, we extracted all the genes in close physical proximity to
175 each SNP (< 1 kb). *Z. tritici* populations show rapid decay in linkage disequilibrium within this
176 distance and the average distance between genes is ~1 kb (Goodwin et al. 2011; Hartmann et al.
177 2017). We identified a variable number of associated genes depending on the reference genome
178 SNP panel. The number of associated genes ranged from 54 when mapping was performed on the
179 reference genome Aus01 to 79 on IPO323 for pathogen virulence and reproduction on different
180 wheat hosts. The number of genes ranged from 88 (reference genome TN09) to 102 (reference
181 genome CRI10) for environmental stress traits (*i.e.* fungicide resistance, growth rate and
182 melanization; **Supplementary Table S7**). Based on the annotation of the canonical reference
183 genome IPO323, the identified genes encoded a broad range of functions including major
184 facilitator superfamily (MFS) transporters, fungal-specific transcription factors, zinc finger and
185 protein kinase domains (**Supplementary Table S7**). Such gene functions may have specific
186 metabolic and regulatory functions underlying pathogen adaptation (Shelest, 2008; Krishnan et al.
187 2018; Pereira et al. 2020b). Importantly, we detected significant SNPs near three genes encoding
188 predicted virulence factors (*i.e.* effectors) on chromosomes 2, 5, and 7 associated with reproduction
189 on the wheat cultivars Greina, Titlis and Chinese Spring, respectively (**Supplementary Table S7**).
190 We also detected numerous significant SNPs for azole resistance tagging the *CYP51* gene that is
191 known to underlie resistance to azole fungicides (Cools and Fraaije, 2012).


A
Number of SNPs
180
60
30
15
10
1
Reproduction 1011
Virulence 1204
Virulence 4391
Reproduction Chinese Spring
Virulence Drifter
Reproduction Drifter
Reproduction Gene
Reproduction Greina
Virulence Runal
Reproduction Titlis
Reproduction Toronit
Growth rate 22°C azole
MCA 14 dpi azole
MCA 15°C 14 dpi
RCA 15°C/22°C 14 dpi
EC50 azole
Melanization 15°C 8 dpi
Melanization 22°C 11 dpi
Melanization 22°C 18 dpi
Melanization 22°C 14 dpi azole
Host traits
Environmental stress traits
IPO323
1A5
1E4
3D1
3D7
Arg00
Aus01
CH95
CNR93
CRI10
I93
IR01_26b
IR01_48b
ISY92
KE94
OregS90
TN09
UR95
YEQ_92
Reference genome
MCA= Mean colony area
RCA= Ratio of colony area
B
Virulence on wheat cultivar
Colony melanization with fungicide exposure
mapped on IPO323 genome
Transposable element
Phosphopantetheine binding domain
mapped on 1E4 genome
Fungal specific transcription factor
mapped on Aus01 genome
mapped on KE94 genome
mapped on IR01_48b genome
−log10(p)
Chromosome
Chromosome

192

193 **Figure 2. Genome wide association mapping based on 19 reference genomes for 49 pathogen traits**
194 **measured under different host and abiotic conditions in *Zymoseptoria tritici*.** (A) Heatmap showing
195 differences in the number of significantly associated SNPs for each trait obtained for each reference
196 genome. Pathogen virulence (percentage of the leaf surface covered by necrotic lesions) and reproduction
197 (pycnidia density within lesions) were measured on 12 genetically diverse wheat lines. (B) Manhattan plots
198 showing SNP *p*-values for two traits (pathogen virulence in the left panel and melanization in presence of
199 fungicide in the right panel) on multiple reference genomes. The shaded gray boxes highlight differences
200 in significant associations found when using different reference genomes. The red line indicates the
201 Bonferroni threshold at a 5% significance level. Pathogen virulence and reproduction were measured on 12
202 genetically diverse wheat lines.
203

204 A challenge associated with performing multiple reference genome GWAS is to identify redundant
205 associations across SNP panels. To estimate the extent of novel gene functions discovered through
206 the expansion of the reference genome SNP panels, we performed a saturation analysis based on
207 orthology information. For each gene with a significant association, we assessed whether any
208 ortholog identified in a different reference genome was already tagged (*i.e.* is a member of the
209 same orthogroup). We randomly selected subsets of the reference genome SNP panels and counted
210 the number of unique orthogroups with significant associations for groups of traits. We observed
211 a near-linear increase in the number of unique orthogroups with significant associations with an
212 increasing number of reference genome panels (**Figure 3**). The most substantial increase was
213 observed by including a second reference genome panel. Beyond two reference genome panels,
214 the benefits for each additional reference genome SNP panel decreased slightly. This shows that a
215 substantial fraction of the genetic factors contributing to adaptation to host, and environmental
216 stress factors cannot be identified from a single reference genome SNP panel. Fungicide resistance
217 related traits show the highest number and fastest gain in significantly associated orthogroups with
218 additional reference genome SNP panels. Pathogen virulence and reproduction showed
219 intermediate increases in significantly associated orthogroups and melanization showed the
220 slowest increase in significantly associated orthogroups. Overall, including multiple reference
221 genome SNP panels substantially expands the spectrum of identifiable genetic factors
222 (**Supplementary Figure S3**).

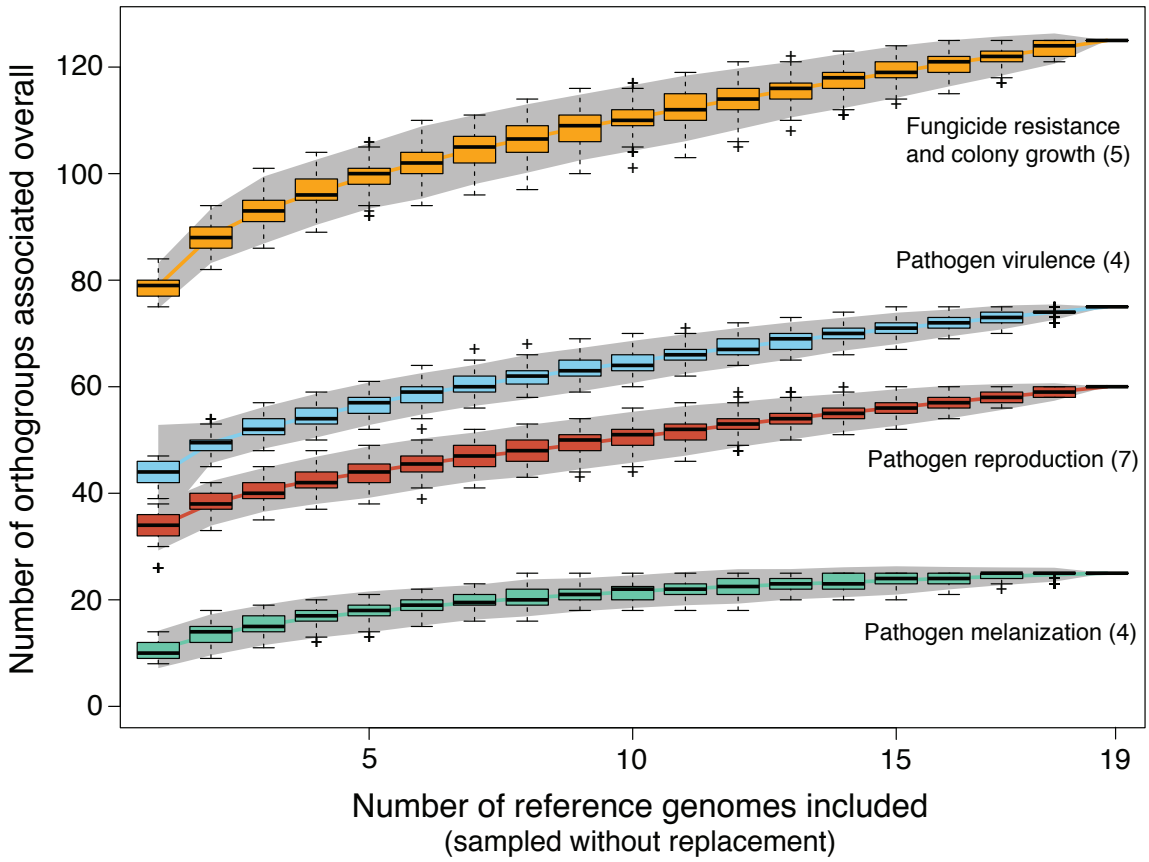


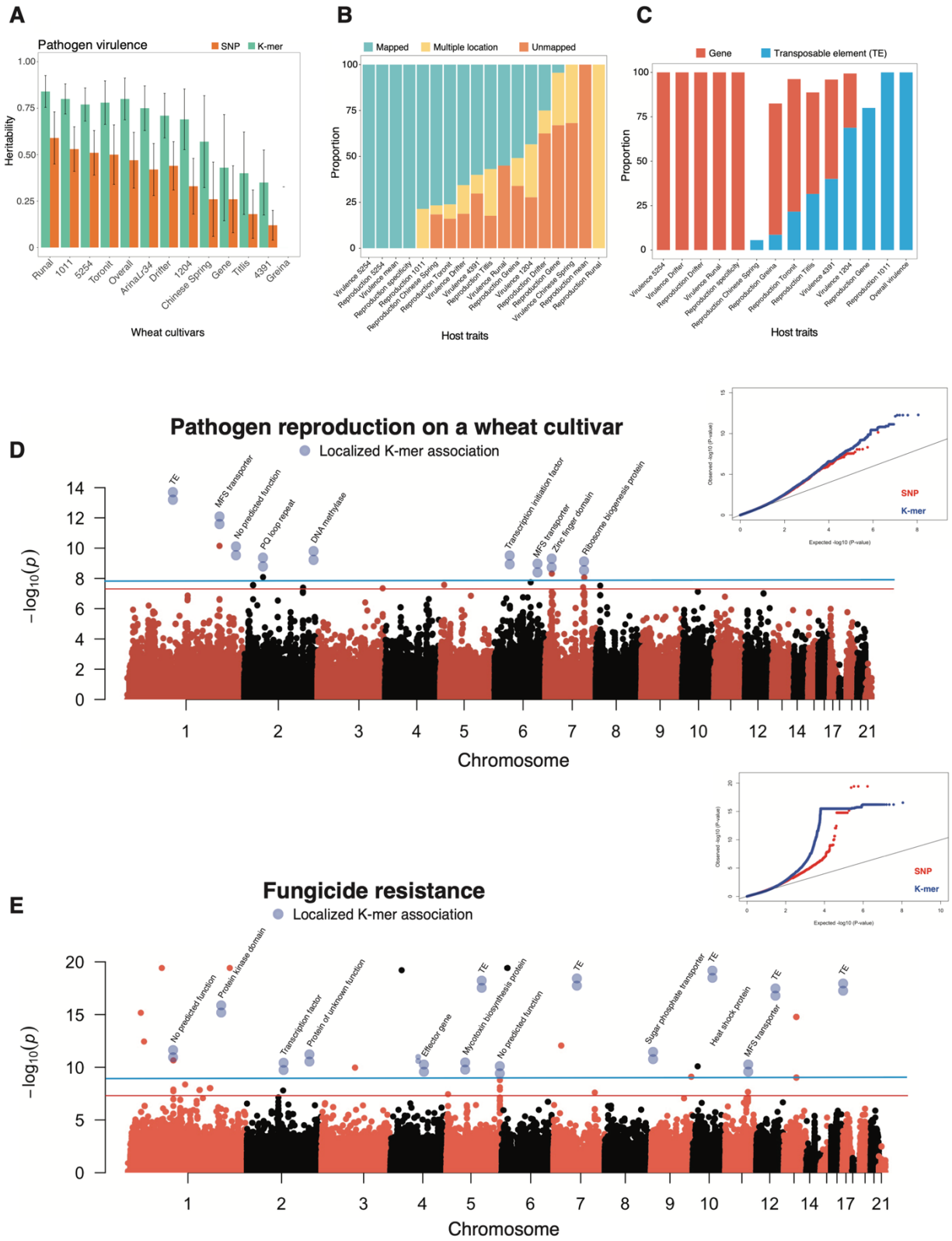
Figure 3. Accumulation curves for the total number of distinct genes (identified by orthogroups within the species) associated with GWAS for different traits as a function of the number of reference genomes analyzed. Mapping outcomes are shown for different groups of traits. The numbers in parentheses indicate the number of traits included in each category. Pathogen virulence (percentage of the leaf surface covered by necrotic lesions) and reproduction (pycnidia density within lesions) were measured on 12 genetically diverse wheat lines.

K-mer approach to uncover additional sources of genetic variation

To further expand our survey of structural variation potentially associated with trait variation, we performed reference-free GWAS on the same trait dataset using 25-bp K-mers generated from whole genome sequencing data. We identified a total of ~55 million K-mers of which 7,111,640 were detected in at least five isolates. We estimated K-mer based heritability to contrast with SNP-based heritability from Dutta et al. (2021). For pathogen virulence, K-mers explained a higher proportion of phenotypic variance compared to the SNP-based estimates (**Figure 4A**). A similar trend of increased heritability accounted by K-mers was observed for all other traits as well

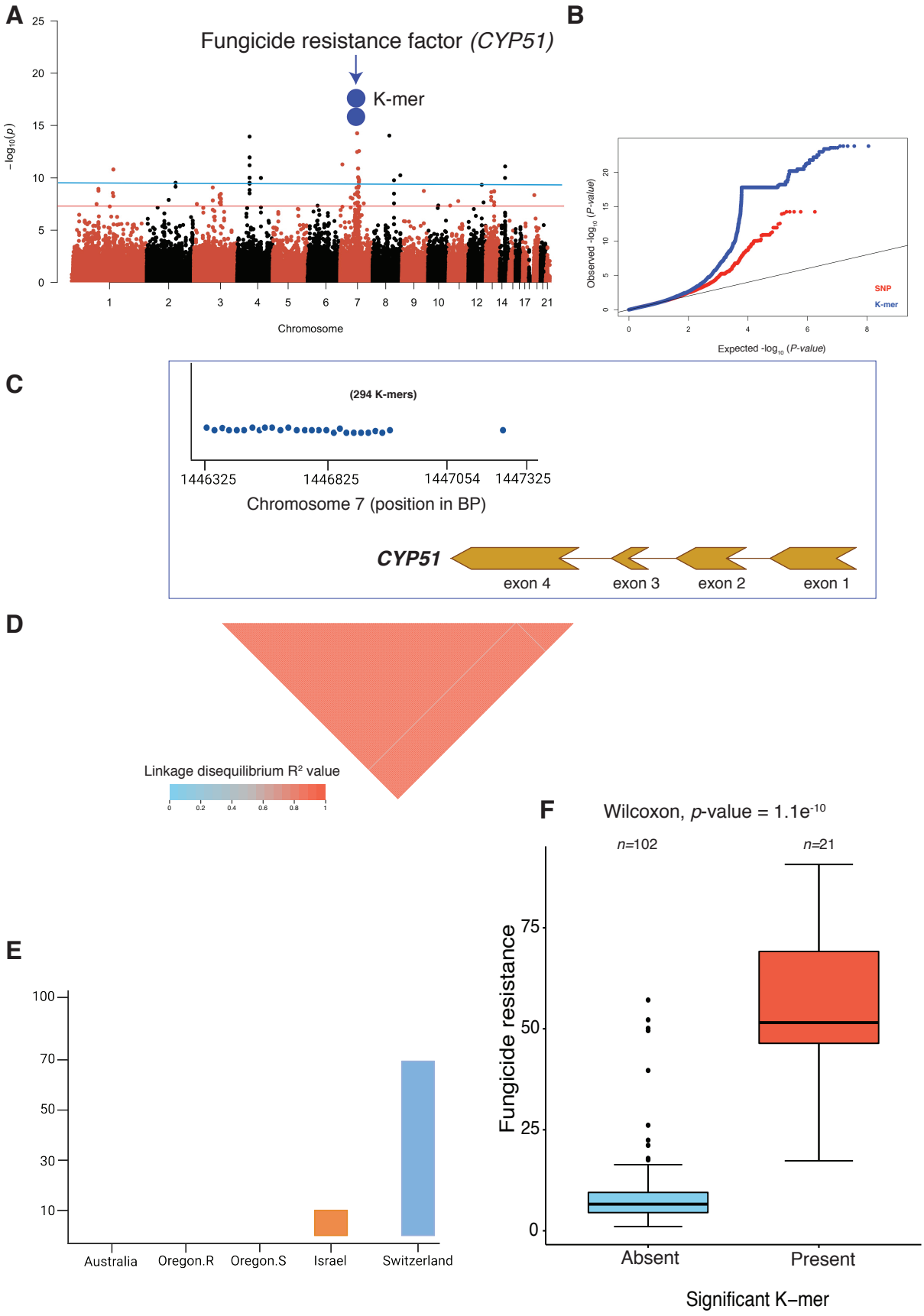
239 (**Supplementary Figure S2A, 2B, 2C**). The heritability for virulence ranged from 0 to 0.84
240 (standard error, SE=0.08) with an average of 0.6 (SE=0.16) compared to 0.35 (SE=0.14) based on
241 SNPs. Heritability for reproduction traits ranged from 0.73 (SE=0.13) to 0.96 (SE=0.01) with an
242 average of 0.86 (SE=0.06) compared to SNP-based heritabilities with an average of 0.65 (SE=0.1).
243 The average heritability for environmental stress factors (i.e., fungicide resistance, growth rate and
244 melanization at different temperatures) was 0.7 (SE=0.18) compared to 0.51 based on SNPs
245 (SE=0.18). Consistent with the high heritability estimates, the K-mer GWAS yielded numerous
246 K-mers above the permutation-based significance threshold ($\alpha = 0.05$) for 33 out of 49 phenotypic
247 traits. The number of significant K-mers ranged from 3-2066 for pathogen virulence, from 3-640
248 for pathogen reproduction, from 3-166 for pathogen melanization, and from 9-3606 for fungicide
249 resistance and growth-related traits.

250 To identify gene functions mapped through K-mer GWAS, we searched K-mer sequences in the
251 canonical reference genome IPO323 (**Figure 4B, Supplementary Figure S2D**). We found a
252 substantial fraction of significant K-mers tagging either a transposable element (TE) or a gene in
253 the *Z. tritici* genome (**Figure 4C, Supplementary Figure S2E**). For host-related traits **(Figure**
254 **4B)**, an average of 63.6% of all significant K-mers tagged a gene compared to 32.1% tagging a
255 TE. In contrast, the proportions of significant K-mers tagging a TE or a gene were roughly inverted
256 (59.17% *vs.* 34.6%) for environmental stress traits (**Supplementary Figure S2D)**. Furthermore,
257 for the majority of the traits, the K-mer with the highest *p*-value tagged a TE (**Figure 4D, 4E**). The
258 high proportion of K-mers mapping to a TE suggests that active transposition has contributed
259 significantly to phenotypic variation in *Z. tritici*. Additionally, the K-mer GWAS discovered a
260 large number of not previously identified genes associated with both host-related and
261 environmental stress traits (**Figure 4D, 4E; Supplementary Figure S3**). The K-mer tagged genes
262 encoded a broad range of functions including a transcription factor, MFS transporters, and
263 peptidases as well as effector candidates (**Figure 4D, 4E, Supplementary Table S8)**.


A
Pathogen virulence
SNP
K-mer
Heritability
Wheat cultivars
B
Mapped
Multiple location
Unmapped
Proportion
Host traits
C
Gene
Transposable element (TE)
Proportion
Host traits
D
Pathogen reproduction on a wheat cultivar
Localized K-mer association
TE
MFS transporter
No predicted function
PQ loop repeat
DNA methylase
Transcription initiation factor
MFS transporter
Zinc finger domain
Ribosome biogenesis protein
SNP
K-mer
Chromosome
E
Fungicide resistance
Localized K-mer association
No predicted function
Protein kinase domain
Transcription factor
Protein of unknown function
Effector gene
TE
Mycotoxin biosynthesis protein
No predicted function
TE
Sugar phosphate transporter
TE
Heat shock protein
TE
MFS transporter
TE
SNP
K-mer
Chromosome

264

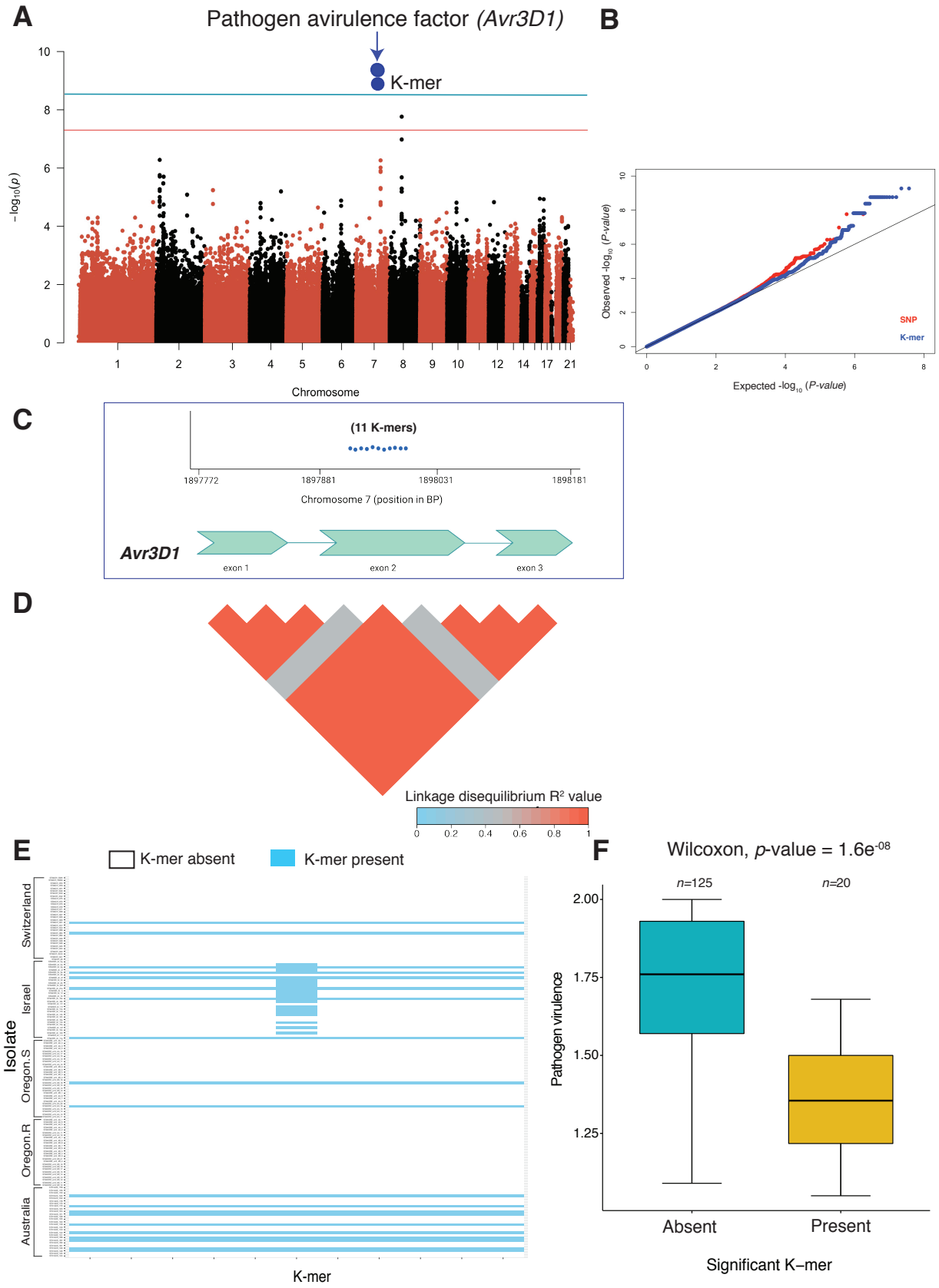
265 **Figure 4. K-mer GWAS on 49 life-history traits based on a K-mer presence/absence table for all 145**
266 ***Zymoseptoria tritici* isolates.** (A) Comparison of heritability estimates for pathogen virulence (percentage
267 of the leaf surface covered by necrotic lesions) based on SNPs (for the reference genome IPO323) and K-
268 mers. Both SNP-based and K-mer-based heritability were estimated by following a genome-based restricted
269 maximum likelihood (GREML) approach. Standard errors are indicated by error bars (B) Alignment of
270 significantly associated K-mers against the reference genome (IPO323) show the proportion of K-mers
271 having a unique mapping position, multiple locations, or no unambiguous mapping position in host-related
272 traits *i.e.* pathogen virulence and reproduction (pycnidia density within lesions). (C) Proportion of
273 significant K-mers with a unique mapping position in the reference genome either tagging a gene or a
274 transposable element for host-related traits. (D, E) Manhattan plots showing significant K-mer associations
275 with pathogen reproduction and fungicide resistance together with quantile-quantile plots for *p*-value
276 comparisons. Manhattan plots were created from SNP-based GWAS and blue dots represents the significant
277 K-mer associations with the K-mers being uniquely mapped to a location in the reference genome. The two
278 blue dots represent individual K-mers with significant associations. The red and blue lines indicate the
279 Bonferroni and permutation-based significance threshold at 5% level for SNPs and K-mers, respectively.
280 Pathogen virulence and reproduction were measured on 12 genetically diverse wheat lines. Overall
281 virulence and reproduction represent the average value of the respective trait measured on 12 genetically
282 diverse wheat lines. Reproduction specificity was estimated based on the adjusted coefficient of variation
283 of mean reproduction across 12 genetically diverse wheat lines. Higher specificity suggests affinity to
284 certain hosts for maximizing reproductive fitness.


A
Fungicide resistance factor (CYP51)
K-mer
−log10(p)
Chromosome
B
Observed -log10 (P-value)
Expected -log10 (P-value)
SNP
K-mer
C
(294 K-mers)
1446325
1446825
1447054
1447325
Chromosome 7 (position in BP)
CYP51
exon 4
exon 3
exon 2
exon 1
D
Linkage disequilibrium R² value
E
Australia
Oregon.R
Oregon.S
Israel
Switzerland
F
Wilcoxon, p-value = 1.1e-10
n=102
n=21
Fungicide resistance
Absent
Present
Significant K−mer

285

286 **Figure 5. Analysis of K-mer GWAS identifying causal genes underlying major phenotypes in**
287 ***Zymoseptoria tritici*.** (A) Manhattan plot showing significant K-mers associated with fungicide resistance.
288 The two blue dots represent all 294 significant K-mers with a unique genomic position on chromosome
289 seven tagging the *CYP51* gene encoding the target of azole fungicides. The red and blue lines show the
290 Bonferroni and permutation-based significance threshold (α=0.05) for SNP and K-mer GWAS,
291 respectively. (B) Quantile-Quantile plot showing the *p*-value comparison between SNPs and K-mer based
292 GWAS. (C) Physical position of 294 significant K-mers mapped to unique positions on chromosome seven
293 associated with the fungicide resistance gene *CYP51*. (D) Linkage disequilibrium (LD) heatmap showing
294 the pairwise r^2 value among 294 significant K-mer presence/absence genotypes associated with the *CYP51*
295 gene. (E) Proportion of isolates from different populations carrying significant K-mers that tagged *CYP51*.
296 (F) Boxplot showing fungicide resistance levels in isolates with presence of the K-mers associated with the
297 *CYP51* gene.
298

299 We analyzed in detail how the K-mer approach expanded the discovery of loci compared to SNP-
300 based GWAS. We focused on the key azole resistance gene *CYP51* (**Figure 5A**). We found 294
301 K-mers above the 5% significance threshold on chromosome 7 associated with *CYP51*
302 (*Zt09_07_00450*) for the resistance trait EAM_14_dpi_azole. All the K-mers could be located to
303 a unique position on the chromosome. The K-mer *p*-values tagging this gene were lower than the
304 SNP *p*-values (**Figure 5B**). Nearly all (293/294) K-mers were located in the upstream region of
305 the gene spanning between the positions 1,446,325 and 1,446,893 bp. The K-mer presence/absence
306 among isolates were in full linkage disequilibrium (**Figure 5C, 5D**; r^2=1). One additional K-mer
307 localized (1,447,308 bp) to the fourth and largest exon of the gene and showed lower linkage
308 disequilibrium (r^2=0.48) with the other K-mers. Most of the isolates from Switzerland (71.1%)
309 and a few from Israel (10%) carried the K-mers associated with increased azole resistance (**Figure**
310 **5E**; **5F**). We expanded our analyses of K-mer associations to virulence traits (**Figure 6A**). We
311 discovered 11 significant K-mers on chromosome 7 (from 1,897,941-1,897,951 bp) for virulence
312 on the cultivar Runal. The tagged gene was previously identified through QTL mapping and
313 encodes a virulence factor termed *Avr3D1*. No SNPs in the same region passed the Bonferroni
314 significance threshold (**Figure 6B**). All K-mers were located in the largest exon and all but one
315 was in full linkage disequilibrium with each other (**Figure 6C, 6D**). The K-mer with lower linkage
316 disequilibrium to the other K-mers was primarily detected in isolates of the Israel population
317 (**Figure 6E**). The isolates carrying the significant K-mers produced less leaf damage (**Figure 6F**).


A
Pathogen avirulence factor (Avr3D1)
K-mer
−log10(p)
Chromosome
B
Observed -log10 (P-value)
Expected -log10 (P-value)
SNP
K-mer
C
(11 K-mers)
1897772
1897881
1898031
1898181
Chromosome 7 (position in BP)
Avr3D1
exon 1
exon 2
exon 3
D
Linkage disequilibrium R² value
E
K-mer absent
K-mer present
Switzerland
Israel
Oregon.S
Oregon.R
Australia
Isolate
K-mer
F
Wilcoxon, p-value = 1.6e-08
n=125
n=20
Pathogen virulence
Absent
Present
Significant K−mer

318

319 **Figure 6. K-mer based GWAS recovered a known effector gene in *Zymoseptoria tritici* with a higher**
320 **statistical power than SNP-based GWAS.** (A) Manhattan plot showing significant K-mers associated
321 with pathogen virulence on the wheat cultivar Runal. The two blue dots represent all 11 K-mers uniquely
322 mapping to positions on chromosome seven and tagging the avirulence gene *Avr3D1* encoding an effector
323 protein. The red and blue lines indicate the Bonferroni and permutation-based significance threshold
324 (α=0.05) for SNP and K-mer GWAS, respectively. (B) Quantile-Quantile plot showing the *p*-value
325 comparison between SNPs and K-mers. (C) Physical position of 11 uniquely mapped K-mers on
326 chromosome seven associated with *Avr3D1*. (D) Linkage disequilibrium (LD) heatmap showing the
327 pairwise r^2 value among 11 significant K-mers associated with *Avr3D1*. (E) Presence/absence pattern of 11
328 significant K-mers associated with *Avr3D1* in five *Z. tritici* populations. The continuous horizontal blue
329 line indicates isolates containing all the significant K-mers. (F) Boxplot showing pathogen virulence
330 (percentage of the leaf surface covered by necrotic lesions) on the wheat cultivar Runal in isolates with or
331 without the significant K-mers associated with *Avr3D1*.
332

333

334 ## Discussion

335 Here, we report the most comprehensive assessment of association mapping performance to date
336 for microbial pathogens to unravel genetic determinants of phenotypic trait variation. We find that
337 expanding association mapping to include multiple reference genome SNP datasets provides a near
338 linear increase in the number of additional loci detected by GWAS. Performing a reference-free
339 GWAS approach using K-mers similarly boosted the power to uncover genetic variation
340 underlying important traits. The extensive gains in the power of GWAS analyses that take into
341 account structural variation reveals a greater proportion of the complexity inherent in adaptive
342 genetic variation within microbial species.

343 SNP-based GWAS based on a single reference genome dataset have been successful in describing
344 the genetic basis of complex pathogen traits (Mohd-Assaad et al. 2016; Pereira et al. 2020b; Caseys
345 et al. 2021; Singh et al. 2021). By expanding the number of reference genome SNP datasets used
346 for GWAS, we identified substantially more independent loci than what was previously identified
347 using the same phenotype dataset (Dutta et al. 2021). The number of loci associated with most trait
348 variation increased almost linearly with the addition of reference genome SNP datasets. Such an
349 increase is striking given the fact that most traits are thought to be significantly constrained by
350 stabilizing selection and have a conserved genetic basis (*e.g.* growth, melanization, reproduction;
351 Steffansson et al. 2014, Qin et al. 2016, Pereira et al. 2020a). Stabilizing selection tends to reduce

352 shared additive genetic variation between populations and closely related species, which ultimately
353 reduces phenotypic variation (Yair and Coop, 2021). Pathogen trait expression is expected to
354 stabilize at an optimal level due to genetic trade-offs (Bonneaud et al. 2020; Dutta et al. 2021).
355 Climatic conditions and host genotype turnover may lead to rapid shifts in selection pressures.
356 Hence, there should also be turnover in the genes underlying adaptation to previous environmental
357 conditions. The *Z. tritici* pathogen model may be an outlier given the maintenance of very large
358 population sizes, high gene flow and extensive chromosomal polymorphism (Hartmann et al.
359 2018; Badet and Croll, 2020). The near-linear increase in associated loci may also be explained,
360 at least in part, by the use of a highly diverse, global panel of reference genomes. The reference
361 genome isolates originating from six continents stem from populations that likely experienced
362 divergent selection pressure from locally adapted hosts and local climatic conditions. Overall, we
363 show that including a broad set of reference genome SNP datasets efficiently overcomes
364 limitations imposed by using a single reference genome. Such limitations often stem from
365 ascertainment bias in SNP calling and genetic distance between the reference genome and mapping
366 populations (Valiente-Mullor et al. 2021). A particular concern is that a single reference may not
367 represent the full catalog of gene functions relevant for adaptation in the species pool (Golicz et
368 al. 2020). For instance, missed associations for genes that are absent from a reference genome may
369 underpin an adaptive advantage in a specific ecological context and/or geographic region (Lassalle
370 et al. 2015; Gori et al. 2020).

371 We find that accounting for genetic variation using K-mers instead of SNPs explains more genetic
372 variation (*i.e.* gives a higher heritability). This implies that significant phenotypic variation is
373 explained by genetic factors located in genomic regions that are difficult to access using SNPs.
374 Such genetic variants are likely to be found in non-coding and TE-rich regions. Such variants may
375 be in accessory genomic regions absent from the reference genome and not easily assessed through
376 SNP calling. Missing heritability in human traits has been recovered by including rare genetic
377 variants (Wainschtein et al. 2021). We show that incorporating genetic variants other than SNPs
378 in plant pathogen GWAS increases trait heritability as well. We also found K-mers in extremely
379 polymorphic regions of the core genome such as the regions surrounding the genes *CYP51* and
380 *Avr3D1*. Recent studies have shown that SVs such as chromosomal rearrangements and copy
381 number variations contribute to adaptive evolution in pathogens (Peter et al. 2018; Firrao et al.
382 2018; Badet et al. 2021). The K-mer approach broadly revealed three classes of loci: (1) loci

383 previously identified by SNP-based GWAS, (2) gene functions that were not identified through
384 SNP-based GWAS but have independent evidence for their contribution to phenotypic trait
385 variation (*i.e. CYP51* and *Avr3D1*) and (3) previously unknown gene functions including genes
386 encoding effector candidates for host manipulation and genes encoding detoxification functions
387 (*e.g.* MFS transporters). The K-mer approach for GWAS has been successfully implemented for
388 plants (Voichek and Weigel, 2020) and bacteria (Lees et al. 2016; Young et al. 2019). Here we
389 provide strong evidence that such reference-free GWAS can also be successfully performed in
390 eukaryotic microbial pathogens.

391 Genetic variation in plant pathogens is characterized by high degrees of functionally relevant
392 polymorphism as well as genomic plasticity underpinning accessory genes (Ehrlich et al. 2005;
393 Hammond et al. 2020; Badet and Croll, 2020). Beyond this, we found substantial complexity in
394 the genes underlying the expression of the same trait under different environmental conditions.
395 Working with such highly diverse pathogen populations poses serious challenges for selecting
396 appropriate reference genome resources. Here we show that GWAS conducted on multiple
397 reference genome SNP datasets and using reference-free approaches effectively compensates for
398 this genetic diversity. This is supported by our recovery of known causal loci for specific
399 phenotypes, including loci missed by previous GWAS, as well as a general improvement in
400 heritability for all traits. Further refinements of our approach should integrate recent developments
401 such as pangenome graphs that might alleviate limitations of studies based on SNPs and single
402 reference genomes. Leveraging a multitude of GWAS signals following our combinatorial
403 approach is likely to significantly advance our mechanistic understanding of pathogen emergence
404 and adaptation.

405

406

407 # Methods

408 ***Fungal material***

409 A collection of 145 *Z. tritici* isolates sampled independently from four different wheat fields was
410 used in this study. The field isolates were sampled between 1990 and 2001 from four different
411 countries (Zhan et al. 2005): Australia (*n*=27), Israel (*n*=30), Switzerland (*n*=32) and USA
412 (Oregon.R, *n*=26; Oregon.S, *n*=30). The two Oregon populations were sampled from the wheat
413 cultivar Madsen (moderately resistant) and Stephens (susceptible), growing simultaneously in the
414 same field. Clones were removed from the field populations so that the analyzed panel comprises
415 only strains with unique genotypes. Blastospores of each isolate were preserved in either 50%
416 glycerol or anhydrous silica at −80°C.

417

418 ***Phenotyping for host infection traits***

419 Datasets on virulence and reproduction for each pathogen strain were previously established by
420 Dutta et al. (2020) (**Supplementary Table S1**). Virulence and reproduction were measured on 12
421 genetically different wheat cultivars displaying varying degrees of resistance and susceptibility to
422 STB. The wheat panel included six commercial varieties (Drifter, Gene, Greina, Runal, Titlis,
423 Toronit), a back-cross line (Arina*Lr34*) and five landraces (1011, 1204, 4391, 5254, Chinese
424 Spring). Four of the landraces (1011, 1204, 4391, 5254) came from the Swiss National Gene Bank
425 (www.bdn.ch). Detailed phenotyping protocols are described in Dutta et al. (2020). Briefly, three
426 seeds of each cultivar were planted in a six-pot strip arrayed in a 2 × 3 pattern. Due to space
427 limitations, the experiment was conducted in two stages, each including six cultivars. All plants
428 were maintained in a greenhouse chamber at 22 °C (day) and 18 °C (night) with 70% relative
429 humidity (RH) and a 16-h photoperiod. Blastospores of each isolate were inoculated using an
430 airbrush spray gun until run-off on two-week-old seedlings to initiate the infection process. In both
431 stages, the inoculations were repeated separately three times to generate three biological
432 replications in separate greenhouse chambers. All inoculated second leaves were collected
433 between 19-26 days post inoculation (dpi) and fixed on QR-coded A4 paper for scanning. The
434 scanned images were analyzed using automated image analysis (AIA; Karisto et al. 2018) to

435 generate quantitative data on the amount of damaged leaf tissue (*i.e.* virulence) and the density of
436 pathogen fruiting bodies called pycnidia produced within the damage area (*i.e.* reproduction).

437 ***Phenotyping for growth and stress-related traits***

438 In vitro traits comprised fungal growth rate (mm per day), thermal sensitivity, mean colony area,
439 fungicide resistance and melanization measured at different temperatures with or without fungicide
440 (**Supplementary Table S2)** following previously described phenotyping protocols (Lendenmann
441 et al. 2014, 2015, 2016; Mohd-Assaad et al. 2016). Briefly, after revival from long-term storage,
442 each isolate was cultured on Petri dishes filled with yeast malt sucrose agar (4 g/L yeast extract,
443 4 g/L malt extract, 4 g/L sucrose, 50 mg/L kanamycin) for 4-5 days at 18 °C. Blastospore solutions
444 were diluted using sterile water to a final concentration of 200 spores/ml using KOVA counting
445 slides (Hycor Biomedical, Inc., Garden Grove, CA, USA). Petri dishes containing potato dextrose
446 agar (PDA, 4 g/L potato starch, 20 g/L dextrose, 15 g/L agar) were inoculated with 500 µl of the
447 spore solution. Inoculated plates were maintained at 15 °C (cold treatment) or 22 °C (control
448 treatment) at 70% RH. Images were captured with a digital camera at 8, 11 and 14 days post
449 inoculation (dpi) to generate five technical replicates. The photographs were analyzed using AIA
450 macros in ImageJ as described in Lendenmann et al. (2014) to measure colony growth. The
451 estimates of colony growth rate for each isolate were obtained by fitting a general linear model
452 over three time points by taking the mean colony radii from 45 colonies. The growth rate ratio
453 between colonies growing at 15 °C or 22 °C, or on 22 °C PDA plates with or without propiconazole
454 (Syngenta, Basel, Switzerland; 0.05 ppm) were expressed as temperature and fungicide sensitivity
455 at 14 dpi, respectively. Fungicide resistance was also quantified on microtiter plates by growing
456 100 µl spore solutions at a concentration of 2.5×10^4 spores/ml of each isolate on 100 µl
457 Sabouraud-dextrose liquid media (SDLM; 20 g/L dextrose, 5 g/L pancreatic digest of casein, 5 g/L
458 peptic digest of animal tissue; Oxoid, Basingstoke, UK) with 12 different concentrations of
459 propiconazole (0, 0.00006, 0.00017, 0.0051, 0.0086, 0.015, 0.025, 0.042, 0.072, 0.20, 0.55,
460 1.5 ppm propiconazole). Plates containing fungal spores amended with the fungicide of each
461 isolate were gently shaken for one minute, sealed and incubated in the dark for four days at 22 °C
462 with 80% RH. Three technical replicates of each isolate were performed. Fungal growth was
463 estimated with an ELISA plate reader (MR5000, Dynatech) by examining the optical density (OD)
464 at 605 nm wavelength. We estimated the EC_{50} value (concentration at which the growth was

465 reduced by 50%) for each isolate using dose-response curves across the varying fungicide
466 concentrations using the drc v.3.0-1 package (Ritz et al. 2015) in the R-studio (R Core Team,
467 2014). Melanization of each isolate was measured at 8, 11, 14 and 18 dpi during growth at 15°C,
468 22°C and at 22°C with 0.05 ppm propiconazole. We measured the mean gray value of fungal
469 colonies from replicated plates for each isolate ranging from 0 (black) to 255 (white) for each time
470 point. To provide a more intuitive interpretation of melanization, each mean gray value was
471 subtracted from 255 to transform the original melanization scale to range from 0 (white) to 255
472 (black).

473 ***Read mapping and single nucleotide polymorphism calling***

474 We used publicly available raw Illumina whole genome sequences of 145 *Z. tritici* isolates
475 (**Supplementary Table S3**; Dutta et al. 2021). Trimmomatic v.0.36 (Bolger et al. 2014) was used
476 with the following settings (illuminaclip = TruSeq3-PE.fa:2:30:10, leading = 10, trailing = 10,
477 slidingwindow = 5:10, minlen = 50) to trim off low-quality reads and remove adapter
478 contamination from each isolate. Trimmed sequence data from all isolates were aligned to the *Z.*
479 *tritici* reference genome IPO323 (Goodwin et al. 2011) using Bowtie2 v.2.3.3 with the option "--
480 very-sensitive-local" (Langmead et al. 2009). We removed PCR duplicates from the alignment
481 (.bam) files by using the MarkDuplicates module in Picard tools v.1.118
482 (http://broadinstitute.github.io/picard). Single nucleotide polymorphism (SNP) calling and variant
483 filtration steps were performed using the Genome Analysis Toolkit (GATK) v.4.0.1.2 (McKenna
484 et al. 2010). We performed SNP calling for all 145 *Z. tritici* isolates independently using the GATK
485 HaplotypeCaller with the command "-emitRefConfidence GVCF; -sample_ploidy 1" (*Z. tritici* is
486 haploid). Then, GenotypeGVCFs was used to conduct joint variant calls on a merged gvcf variant
487 file with the command -maxAltAlleles 2. SNPs found only in the joint variant call file were
488 retained. As recommended by GATK Best Practices, we performed hard filtering of SNPs based
489 on quality cut-offs using the GATK VariantFiltration and SelectVariants tools. Variants matching
490 any of the following criteria were removed: QUAL < 250 (overall quality filter); QD < 20.0
491 (avoiding quality inflation in high-coverage regions); MQ < 30.0 (avoid calls from ambiguously
492 mapped reads); −2 > BaseQRankSum > 2; −2 > MQRankSum > 2; −2 > ReadPosRankSum > 2;
493 FS > 0.1. Using this procedure, the genotyping accuracy was shown to be high and congruent with
494 an alternative SNP caller (Hartmann et al. 2018). We retained a genotypic call rate of ≥80% and

495 minor allele frequency (MAF) > 5% to generate a final SNP dataset containing 883,207 biallelic
496 SNPs based on the reference genome IPO323. We repeated the SNP calling and filtering procedure
497 separately for 18 additional fully assembled *Z. tritici* genomes from Badet et al. (2020). The
498 number of biallelic SNPs called on the 18 additional reference genomes ranged from 827,851
499 (genome TN09) to 883,119 (genome I93; **Supplementary Table S4**).

500 ***SNP-based genome-wide association mapping***

501 Log-transformed least-square means for each isolate × environment combination including 49
502 traits were obtained from Dutta et al. (2021) to conduct genome-wide association (GWAS)
503 mapping. We used a mixed linear model (MLM) approach implemented in the program GEMMA
504 v.0.98 (Zhou and Stephens, 2012) to perform GWAS on all the traits. MLMs control for genetic
505 relatedness and population structure (Kang et al. 2008; Zhang et al. 2012). Prior to GWAS, we
506 converted all 19 SNP datasets (one per reference genome) into PLINK ".bed" format to perform
507 principal component analyses (PCA) using the "--pca" command in PLINK v.1.90 (Purcell et al.
508 2007). To account for genetic relatedness among isolates, a centered genetic relatedness matrix
509 (GRM) for each SNP dataset was constructed using the option "-gk 1" in GEMMA by considering
510 all genome-wide SNPs. As both PCA and GRM can efficiently control for *p*-value inflation, we
511 estimated genomic inflation factors (GIF, λ; Devlin and Roeder, 1999) to make decisions on
512 whether PCs should be included in the GWAS models as covariates or not. The GIF for each trait
513 was estimated as λ = M/E, where M is the median of the observed chi-squared test statistics and E
514 is the expected median of the chi-squared distribution (Yang et al. 2011). The distribution of all
515 SNP effects follows a one degree of freedom chi-square distribution under the null hypothesis with
516 a median of ~0.455, which can be inflated by discrepancies in allele frequencies caused by
517 population structure, genetic relatedness, and genotyping errors. The inflation is proportional to
518 the deviation from the null hypothesis. When the fitted GWAS model efficiently accounts for such
519 systematic deviations, the λ value is close to 1. Therefore, depending on the λ value, the reference
520 genome based GWAS were performed using either LMM+K or LMM+K+PC, where K is the
521 GRM used as a random effect and the first three PCs were used as fixed covariates. We used the
522 following LMM model in GEMMA:

523 $$y = W\alpha + x\beta + u + \varepsilon;\ u \sim MVN_n(0, \lambda\tau^{-1}K),\ \varepsilon \sim MVN_n(0, \tau^{-1}I_n)$$

524 where y represents a vector of phenotypic values for n individuals; W is a matrix of covariates
525 (fixed effects with a column vector of 1 and the first three PCs), α is a vector of the corresponding
526 coefficients including the intercept; x is a vector of the genotypes of the SNP marker, β is the effect
527 size of the marker; u is a vector of random individual effects; ε is a vector of random error; τ^{-1} is
528 the variance of the residual errors; λ is the ratio between the two variance components; K is the n
529 $\times$ n genetic relatedness matrix and I_n is an $n \times n$ identity matrix and MVN_n represents the
530 multivariate normal distribution. We set the MAF to 5% with a maximum of 50% missing values
531 with the option "-miss 0.5". SNP p-values were estimated following a likelihood ratio test in
532 GEMMA. We used the stringent Bonferroni threshold ($\alpha = 0.05$; $p = \alpha$ / total number of SNPs) to
533 define a SNP significantly associated with a phenotype. The proportion of phenotypic variance
534 explained by the most significant SNPs was estimated by $2f(1-f)a^2$, where f is the minor allele
535 frequency and a is the standardized coefficient (Gudbjartsson et al. 2008). To obtain the
536 standardized coefficient for each SNP, we estimated the standardized regression coefficient
537 applying a linear regression model with the "standard-beta" option implemented in PLINK v.1.9.
538 We restricted this analysis only to the canonical reference genome IPO323. To identify genes close
539 to significantly associated SNPs in one of the reference genomes (Badet et al. 2020), we used the
540 BEDtools v.2.29.0 (Quinlan and Hall, 2010) *closest* command. We further investigated patterns of
541 linkage disequilibrium (LD) in the genomic regions with the most significantly associated SNPs.
542 All possible SNP pairs in 5 kb windows were analyzed using the "--hap-r2" command in vcftools.
543 To visualize the r^2 values, heatmaps for each locus were generated using the R package LDheatmap
544 v.0.99-7 (Shin et al. 2006). We created a heatmap summarizing the number of significant SNPs
545 passing the Bonferroni threshold for each trait and each genome using the R package *pheatmap*
546 (Kolde, 2012).

547

548 ***K-mer based genome-wide association mapping***

549 We performed K-mer based GWAS on all 49 traits in the panel of 145 *Z. tritici* isolates following
550 the methodology described in Voichek and Weigel (2020). This approach uses raw sequencing
551 reads of specific length and was designed for settings where a reference genome is lacking or to
552 account for structural variation. K-mers of 25 bp length were counted with and without
553 canonization, sorted and listed in a textual format for each isolate separately. K-mer canonization

554 refers to storing K-mers and their reverse-complement for generating presence/absence patterns
555 since these sequences are indistinguishable (Voichek and Weigel, 2020). K-mer length has an
556 impact on the number and accuracy of K-mers. For small genomes of the size of *Z. tritici,* 25-bp
557 K-mers are recommended (Voichek and Weigel, 2020). K-mers were filtered based on the
558 presence/absence patterns among isolates with a 5% MAF and compressed into a presence/absence
559 table for running GWAS. There were 55,758,186 unique K-mers generated from 145 isolates. Prior
560 to GWAS, a GRM was estimated with EMMA (Efficient Mixed-Model Association) that
561 comprised an identity-by-state (IBS) matrix under the assumption that each K-mer has a small,
562 random effect on the phenotype. GWAS was performed by using an LMM+K model in GEMMA
563 with the likelihood ratio test to estimate *p*-values. A K-mer was considered to be significant when
564 the *p*-value passed the permutation-based threshold as described in Voichek and Weigel (2020).
565 The pairwise LD among significant K-mers for each trait was estimated by converting the K-mer
566 presence/absence table containing all the K-mers into PLINK format and using the command "--
567 r2" in PLINK. We attempted to map all the significant K-mers for each trait to the *Z. tritici*
568 reference genome IPO323 using the short-read aligner bowtie v1.2.2 (Langmead and Salzberg,
569 2012) with the command "-a --best –strata". We used the center position of the K-mer alignment
570 to the reference genome as a coordinate to inspect nearby features using BEDtools. If no significant
571 K-mer could be mapped to the reference genome, we retrieved the isolates carrying the specific
572 K-mer and used the paired-end raw sequencing reads to detect the origin of the K-mer. These
573 paired-end reads were then aligned to the canonical reference genome IPO323 using Bowtie2
574 v.2.3.3 (Langmead and Salzberg, 2009).

575

576 ***Heritability estimation using SNPs and K-mers***

577 We estimated SNP-based heritability on multiple reference genomes and K-mer-based heritability
578 following the same procedure described in Dutta et al. (2021). Briefly, the phenotypic data of each
579 trait and the GRM representing the additive effect of all genome-wide SNPs from the canonical
580 reference genome IPO323 and K-mers were included in a genome-based restricted maximum
581 likelihood (GREML) approach using the genome-wide complex trait analysis (GCTA) tool
582 v.1.93.0 (Yang et al. 2011) to estimate heritability. GRMs for reference genome SNP datasets and
583 the K-mer presence/absence table (converted into PLINK format) were estimated following a

584 normalized identity-by-state method and fitted as a random factor in the model to estimate the
585 proportion of phenotypic variance for each trait. The following formula from Yang et al. (2011)
586 was used to estimate the relatedness between two individuals:

587 $$A_{jk}=\frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij}-2p_i)(x_{jk}-2p_i)}{2p_i(1-p_i)}$$

588 Where x_{ij} is the number of copies of the reference allele for the ith SNP of the jth individual and p_i
589 is the frequency of the reference allele and N is the number of SNPs. Here, the GRMs were
590 constructed using all genome-wide SNPs and K-mers irrespective of the nature of their relationship
591 with the phenotype, thus indicating the approximated genetic similarities at causal loci and the
592 accuracy of the heritability estimates.

593 ***Pangenome analyses***

594 We generated accumulation curves to estimate the gain in additional loci from performing GWAS
595 on more than one reference genome. For this, we retrieved for each GWAS based on SNPs mapped
596 to a particular reference genome the set of genes within 1 kb distance with significantly associated
597 SNPs. Then, we matched the set of associated genes among genomes using within-species gene
598 orthology information (Badet et al. 2020) to determine whether genes belong to the same
599 orthogroup. We used a sampling procedure (without replacement) among reference genomes to
600 assess the total number of distinct orthogroups with a significantly associated gene. The
601 accumulation curves for 1-19 genomes were produced using the "specaccum" function in the R
602 package *vegan* (Oksanen et al. 2011). We fitted an Arrhenius nonlinear model to the gene
603 accumulation curve to visualize the distribution using the "random" and "fitspecaccum"
604 commands. UpSetR package (Lex et al. 2014) was used to visualize the number of significantly
605 associated genes identified by the multiple reference-based GWAS and K-mer GWAS. All other
606 figures were generated using the R packages *qqman* (Turner, 2014) and *ggplot2* v.3.1.0 (Wickham,
607 2016).

608

609

610 **Data availability**

611 All genome sequences are available from the NCBI Sequence Read Archive (BioProject
612 accessions PRJNA327615, PRJNA596434, and PRJNA178194).

613 **Author contributions**

614 AD and DC conceived the research. AD conducted experiments, performed data analyses, and
615 wrote the manuscript with DC. BAM provided funding. All co-authors edited the manuscript.

616 **Competing interests**

617 We declare that we have no competing interests

618 **Acknowledgments**

619 Emile Gluck-Thaler provided helpful comments on a previous version of the manuscript. This
620 work was supported by the Swiss Federal Office for Agriculture (BLW) in the framework of the
621 NAP-PGREL Project Nr. 627000640.

622

References

Allen JP, Snitkin E, Pincus NB, Hauser AR. Forest and Trees: Exploring Bacterial Virulence with Genome-wide Association Studies and Machine Learning. Trends in Microbiology. 2021 Jan 14.

Baddam R, Kumar N, Shaik S, Lankapalli AK, Ahmed N. Genome dynamics and evolution of *Salmonella Typhi* strains from the typhoid-endemic zones. Scientific reports. 2014 Dec 12;4(1):1-9.

Badet T, Croll D. The rise and fall of genes: origins and functions of plant pathogen pangenomes. Current opinion in plant biology. 2020 Aug 1;56:65-73.

Badet T, Fouché S, Hartmann FE, Zala M, Croll D. Machine-learning predicts genomic determinants of meiosis-driven structural variation in a eukaryotic pathogen. Nature communications. 2021 Jun 10;12(1):1-4.

Badet T, Oggenfuss U, Abraham L, McDonald BA, Croll D. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen *Zymoseptoria tritici*. BMC biology. 2020 Dec;18(1):1-8.

Bartoli C, Roux F. Genome-wide association studies in plant pathosystems: toward an ecological genomics approach. Frontiers in plant science. 2017 May 23;8:763.

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. Nature plants. 2020 Aug;6(8):914-20.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

Bonneaud C, Tardy L, Hill GE, McGraw KJ, Wilson AJ, Giraudeau M. Experimental evidence for stabilizing selection on virulence in a bacterial pathogen. Evolution Letters. 2020 Dec;4(6):491-501.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K et al. (2009). BLAST+: architecture and applications. BMC Bioinformatics 10: 421–429.

Casadevall A, Fang FC, Pirofski LA. Microbial virulence as an emergent property: consequences and opportunities. PLoS Pathog. 2011 Jul 21;7(7):e1002136.

Caseys C, Shi G, Soltis N, Gwinner R, Corwin J, Atwell S, Kliebenstein DJ. Quantitative interactions: the disease outcome of *Botrytis cinerea* across the plant kingdom. G3. 2021 Aug;11(8):jkab175.

Cools HJ, Fraaije BA. Update on mechanisms of azole resistance in *Mycosphaerella graminicola* and implications for future control. Pest management science. 2013 Feb;69(2):150-5.

Croll D, Lendenmann MH, Stewart E, McDonald BA. The impact of recombination hotspots on genome evolution of a fungal plant pathogen. Genetics. 2015 Nov 1;201(3):1213-28.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156-8.

Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55(4):997–1004.

Dutilh BE, Backus L, Edwards RA, Wels M, Bayjanov JR, van Hijum SA. Explaining microbial phenotypes on a genomic scale: GWAS for microbes. Briefings in functional genomics. 2013 Jul 1;12(4):366-80.

Dutta A, Croll D, McDonald BA, Barrett LG. Maintenance of variation in virulence and reproduction in populations of an agricultural plant pathogen. Evolutionary applications. 2021 Feb;14(2):335-47.

Dutta A, Hartmann FE, Francisco CS, McDonald BA, Croll D. Mapping the adaptive landscape of a major agricultural pathogen reveals evolutionary constraints across heterogeneous environments. The ISME journal. 2021 May;15(5):1402-19.

Ehrlich GD, Hu FZ, Shen K, Stoodley P, Post JC. Bacterial plurality as a general mechanism driving persistence in chronic infections. Clinical orthopaedics and related research. 2005 Aug(437):20.

Engle LJ, Simpson CL, Landers JE. Using high-throughput SNP technologies to study cancer. Oncogene. 2006 Mar;25(11):1594-601.

Figueroa M, Hammond-Kosack KE, Solomon PS. A review of wheat diseases—a field perspective. Molecular plant pathology. 2018 Jun;19(6):1523-36.

Firrao G, Torelli E, Polano C, Ferrante P, Ferrini F, Martini M, Marcelletti S, Scortichini M, Ermacora P. Genomic structural variations affecting virulence during clonal expansion of *Pseudomonas syringae* pv. actinidiae biovar 3 in Europe. Frontiers in microbiology. 2018 Apr 5;9:656.

Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, McCraw SL, Gurr SJ. Emerging fungal threats to animal, plant and ecosystem health. Nature. 2012 Apr;484(7393):186-94.

Fones H, Gurr S. The impact of Septoria tritici Blotch disease on wheat: An EU perspective. Fungal Genet Biol. 2015;79:3–7.

Gage JL, Vaillancourt B, Hamilton JP, Manrique-Carpintero NC, Gustafson TJ, Barry K, Lipzen A, Tracy WF, Mikel MA, Kaeppler SM, Buell CR. Multiple maize reference genomes impact the identification of variants by genome-wide association study in a diverse inbred panel. The plant genome. 2019 Jun 1;12(2).

714
715 Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. Pangenomics comes of age: from bacteria
716 to plant and animal applications. Trends in Genetics. 2020 Feb 1;36(2):132-45.
717
718 Goodwin SB, Ben M'Barek S, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, et al. Finished
719 genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure,
720 chromosome plasticity, and stealth pathogenesis. Malik HS, editor. PLoS Genet.
721 2011;7(6):e1002070.
722
723 Gori A, Harrison OB, Mlia E, Nishihara Y, Chan JM, Msefula J, Mallewa M, Dube Q, Swarthout
724 TD, Nobbs AH, Maiden MC. Pan-GWAS of *Streptococcus agalactiae* highlights lineage-specific
725 genes associated with virulence and niche adaptation. MBio. 2020 Jun 9;11(3):e00728-20.
726
727 Grau-Bové X, Lucas E, Pipini D, Rippon E, van 't Hof AE, Constant E, Dadzie S, Egyir-Yawson
728 A, Essandoh J, Chabi J, Djogbénou L. Resistance to pirimiphos-methyl in West African Anopheles
729 is spreading via duplication and introgression of the Ace1 locus. PLoS Genetics. 2021 Jan
730 21;17(1):e1009253.
731
732 Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P,
733 Sulem P, Thorlacius S, Gylfason A, Steinberg S, Helgadottir A. Many sequence variants affecting
734 diversity of adult human height. Nature genetics. 2008 May;40(5):609-15.
735
736 Guo J, Cao K, Deng C, Li Y, Zhu G, Fang W, Chen C, Wang X, Wu J, Guan L, Wu S. An integrated
737 peach genome structural variation map uncovers genes associated with fruit traits. Genome
738 biology. 2020 Dec;21(1):1-9.
739
740 Gupta PK. Quantitative genetics: pan-genomes, SVs, and k-mers for GWAS. Trends in Genetics.
741 2021 Jun 25.
742
743 Hammond JA, Gordon EA, Socarras KM, Chang Mell J, Ehrlich GD. Beyond the pan-genome:
744 current perspectives on the functional and practical outcomes of the distributed genome
745 hypothesis. Biochemical Society Transactions. 2020 Dec 18;48(6):2437-55.
746
747 Hartmann FE, McDonald BA, Croll D. Genome-wide evidence for divergent selection between
748 populations of a major agricultural pathogen. Molecular Ecology. 2018;27:2725–41.
749
750 Hartmann FE, Sánchez-Vallet A, McDonald BA, Croll D. A fungal wheat pathogen evolved host
751 specialization by extensive chromosomal rearrangements. ISME J. 2017;11(5):1189–204.
752
753 Hartmann FE, Vonlanthen T, Singh NK, McDonald MC, Milgate A, Croll D. The complex
754 genomic basis of rapid convergent adaptation to pesticides across continents in a fungal plant
755 pathogen. Molecular Ecology. 2020 Jan 1.
756
757 Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, Jenney A, Connor TR,
758 Hsu LY, Severin J, Brisse S. Genomic analysis of diversity, population structure, virulence, and

759 antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. Proceedings
760 of the National Academy of Sciences. 2015 Jul 7;112(27):E3574-81.
761
762 Jaillard, M., Lima, L., Tournoud, M., Mahé, P., Van Belkum, A., Lacroix, V. and Jacob, L., 2018.
763 A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap
764 between k-mers and genetic events. PLoS genetics, 14(11), p.e1007758.
765
766 Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of
767 population structure in model organism association mapping. Genetics. 2008 Mar 1;178(3):1709-
768 23.
769
770 Karisto P, Hund A, Yu K, Anderegg J, Walter A, Mascher F, et al. Ranking quantitative resistance
771 to Septoria tritici blotch in elite wheat cultivars using automated image analysis. Phytopathology.
772 2018;108:568–81.
773
774 Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements
775 in performance and usability. Molecular biology and evolution. 2013 Jan 16;30(4):772-80.
776 Kolde, R. Pheatmap: pretty heatmaps, R package v. 16 (R Foundation for Statistical Computing,
777 2012).
778
779 Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL, Bayles
780 KW, Fey PD, Yajjala VK. Predicting the virulence of MRSA from its genome sequence. Genome
781 research. 2014 May 1;24(5):839-49.
782
783 Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short
784 DNA sequences to the human genome. Genome Biol. 2009;10:R25.
785
786 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–
787 359 (2012).
788
789 Langner T, Harant A, Gomez-Luciano LB, Shrestha RK, Malmgren A, Latorre SM, Burbano HA,
790 Win J, Kamoun S. Genomic rearrangements generate hypervariable mini-chromosomes in host-
791 specific isolates of the blast fungus. PLoS genetics. 2021 Feb 16;17(2):e1009386.
792
793 Lassalle F, Muller D, Nesme X. Ecological speciation in bacteria: reverse ecology approaches
794 reveal the adaptive part of bacterial cladogenesis. Research in microbiology. 2015 Dec
795 1;166(10):729-41.
796
797 Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, Marttinen P, Davies
798 MR, Steer AC, Tong SY, Honkela A. Sequence element enrichment analysis to determine the
799 genetic basis of bacterial phenotypes. Nature communications. 2016 Sep 16;7(1):1-8.
800
801 Lendenmann MH, Croll D, McDonald BA. QTL mapping of fungicide sensitivity reveals novel
802 genes and pleiotropy with melanization in the pathogen *Zymoseptoria tritici*. Fungal Genet Biol.
803 2015;80:53–67.
804

Lendenmann MH, Croll D, Palma-Guerrero J, Stewart EL, McDonald BA. QTL mapping of temperature sensitivity reveals candidate genes for thermal adaptation and growth morphology in the plant pathogenic fungus *Zymoseptoria tritici*. Heredity. 2016;116:384–94.

Lendenmann MH, Croll D, Stewart EL, McDonald BA. Quantitative trait locus mapping of melanization in the plant pathogenic fungus *Zymoseptoria tritici*. G3: Genes, Genomes. Genetics. 2014;4:2519–33.

Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: visualization of intersecting sets. IEEE transactions on visualization and computer graphics. 2014 Nov 6;20(12):1983-92.

Liu F, Zhu Y, Yi Y, Lu N, Zhu B, Hu Y. Comparative genomic analysis of *Acinetobacter baumannii* clinical isolates reveals extensive genomic variation and diverse antibiotic resistance determinants. BMC genomics. 2014 Dec;15(1):1-4.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

Meile L, Croll D, Brunner PC, Plissonneau C, Hartmann FE, McDonald BA, et al. A fungal avirulence factor encoded in a highly plastic genomic region triggers partial resistance to septoria tritici blotch. New Phytol. 2018;219(3):1048–61.

Mohd-Assaad N, McDonald BA, Croll D. Multilocus resistance evolution to azole fungicides in fungal plant pathogen populations. Mol Ecol. 2016;25:6124–42.

OGGENFUSS, Ursula, et al. A population-level invasion by transposable elements triggers genome expansion in a fungal pathogen. bioRxiv , 2021, S. 2020.02. 11.944652.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, et al. vegan: Community Ecology Package. 2011 R package version 2.0-2. 2011.

Pereira D, Croll D, Brunner PC, McDonald BA. Natural selection drives population divergence for local adaptation in a wheat pathogen. Fungal Genetics and Biology. 2020a Aug 1;141:103398.

Pereira D, McDonald BA, Croll D. The genetic architecture of emerging fungicide resistance in populations of a global wheat pathogen. Genome biology and evolution. 2020b Dec;12(12):2231-44.

Peter J, De Chiara M, Friedrich A, Yue JX, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A, Cruaud C. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. Nature. 2018 Apr;556(7701):339-44.

Plaumann PL, Schmidpeter J, Dahl M, Taher L, Koch C. A dispensable chromosome is required for virulence in the hemibiotrophic plant pathogen *Colletotrichum higginsianum*. Frontiers in microbiology. 2018 May 18;9:1005.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics. 2007 Sep 1;81(3):559-75.

Qin CF, He MH, Chen FP, Zhu W, Yang LN, Wu EJ, Guo ZL, Shang LP, Zhan J. Comparative analyses of fungicide sensitivity and SSR marker variations indicate a low risk of developing azoxystrobin resistance in *Phytophthora infestans*. Scientific reports. 2016 Feb 8;6(1):1-0.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. http://www.R-project.org/.

Rahman A, Hallgrímsdóttir I, Eisen M, Pachter L. Association mapping from sequencing reads using k-mers. Elife. 2018 Jun 13;7:e32920.

Ritz C, Baty F, Streibig JC, Gerhard D. Dose-response analysis using R. PloS One. 2015;10:12.

Sánchez-Vallet A, Hartmann FE, Marcel TC, Croll D. Nature's genetic screens: using genome-wide association studies for effector discovery. Molecular plant pathology. 2018 Jan;19(1):3.

Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. Proceedings of the national academy of sciences. 2013 Jul 16;110(29):11923-7.

Shin JH, Blay S, McNeney B, Graham J. LDheatmap: An R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. J Stat Softw. 2006;16:1–9

Singh NK, Chanclud E, Croll D. Population-level deep sequencing reveals the interplay of clonal and sexual reproduction in the fungal wheat pathogen *Zymoseptoria tritici*. bioRxiv. 2020 Jan 1.

Singh NK, Badet T, Abraham L, Croll D. Rapid sequence evolution driven by transposable elements at a virulence locus in a fungal wheat pathogen. BMC genomics. 2021 Dec;22(1):1-6.

Stefansson TS, Willi Y, Croll D, McDonald BA. An assay for quantitative virulence in *Rhynchosporium commune* reveals an association between effector genotype and virulence. Plant Pathology. 2014 Apr;63(2):405-14.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”. Proceedings of the National Academy of Sciences. 2005 Sep 27;102(39):13950-5.

Todesco M, Owens GL, Bercovich N, Légaré JS, Soudi S, Burge DO, Huang K, Ostevik KL, Drummond EB, Imerovski I, Lande K. Massive haplotypes underlie ecotypic differentiation in sunflowers. Nature. 2020 Aug;584(7822):602-7.

Torriani SF, Melichar JP, Mills C, Pain N, Sierotzki H, Courbot M. *Zymoseptoria tritici*: a major threat to wheat production, integrated approaches to control. Fungal Genet Biol. 2015;79:8–12.

Turner SD. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. Biorxiv. 2014 Jan 1:005165.

Valiente-Mullor C, Beamud B, Ansari I, Francés-Cuesta C, et al. One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. PLoS computational biology. 2021 Jan 27;17(1):e1008678.

Voichek Y, Weigel D. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nature genetics. 2020 May;52(5):534-40.

Wainschtein P, Jain D, Zheng Z, Cupples LA, Shadyab AH, McKnight B, Shoemaker BM, Mitchell BD, Psaty BM, Kooperberg C, Liu CT. Recovery of trait heritability from whole genome sequence data. BioRxiv. 2021 Jan 1:588020.

Weinert LA, Chaudhuri RR, Wang J, Peters SE, Corander J, Jombart T, Baig A, Howell KJ, Vehkala M, Välimäki N, Harris D. Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*. Nature communications. 2015 Mar 31;6(1):1-0.

Wickham H. Ggplot2 : elegant graphics for data analysis. New York: Springer-Verlag; 2016. https://tidyverse.github.io/ggplot2-docs/authors.html. Accessed 27 May 2021.

Wu Y, Zaiden N, Cao B. The core-and pan-genomic analyses of the genus Comamonas: from environmental adaptation to potential virulence. Frontiers in microbiology. 2018 Dec 12;9:3096.

Yair S, Coop G. Population differentiation of polygenic score predictions under stabilizing selection. bioRxiv. 2021 Jan 1.

Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.

Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, Smith AV, Ingelsson E, O'connell JR, Mangino M, Mägi R. Genomic inflation factors under polygenic inheritance. European Journal of Human Genetics. 2011 Jul;19(7):807-12.

Young BC, Earle SG, Soeng S, Sar P, Kumar V, Hor S, Sar V, Bousfield R, Sanderson ND, Barker L, Stoesser N. Panton–Valentine leucocidin is the key determinant of *Staphylococcus aureus* pyomyositis in a bacterial GWAS. Elife. 2019 Feb 22;8:e42486.

Zeevi D, Korem T, Godneva A, Bar N, Kurilshikov A, Lotan-Pompan M, et al. Structural variation in the gut microbiome associates with host health. Nature. 2019 Apr;568(7750):43-8.

Zhan J, Linde CC, Jürgens T, Merz U, Steinebrunner F, McDonald BA. Variation for neutral markers is correlated with variation for quantitative traits in the plant pathogenic fungus *Mycosphaerella graminicola*. Mol Ecol. 2005;14:2683–93.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. Mixed linear model approach adapted for genome-wide association studies. Nature genetics. 2010 Apr;42(4):355-60.

Zhong Z, Marcel TC, Hartmann FE, Ma X, Plissonneau C, Zala M, Ducasse A, et al. A small secreted protein in *Zymoseptoria tritici* is responsible for avirulence on wheat cultivars carrying the *Stb6* resistance gene. New Phytologist. 2017 Apr;214(2):619-31.

Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature genetics. 2012 Jul;44(7):821.

Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, Sun H, Xia Y, Liang S, Dai Y, Wan D. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nature biotechnology. 2019 Feb;37(2):179-85.

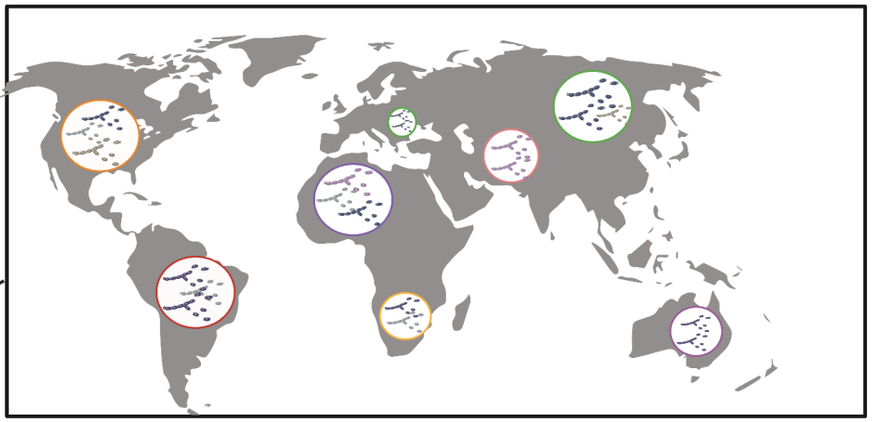
Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences. 2012 Jan 24;109(4):1193-8.

986 **Figure legends**
987

988 **Figure 1. A comprehensive workflow for conducting microbial genome wide association studies**
989 **(GWAS) using multiple reference genomes and K-mer data from mapping populations.** Genetically
990 diverse pathogen populations from different geographic locations are sampled to construct an association
991 panel followed by greenhouse and laboratory phenotyping to assess heritable trait variation (right panel;
992 Dutta et al. 2021). Chromosome-level genome assemblies of representative isolates is performed to generate
993 reference genomes and establish a species pangenome (left panel; Badet el al. 2020). Whole genome
994 sequencing of the association panel enables single nucleotide polymorphism (SNP) calling against multiple
995 reference genomes and creation of K-mer presence/absence tables (middle panel). GWAS can be performed
996 simultaneously to take advantage of SNP datasets or K-mer presence/absence tables.
997

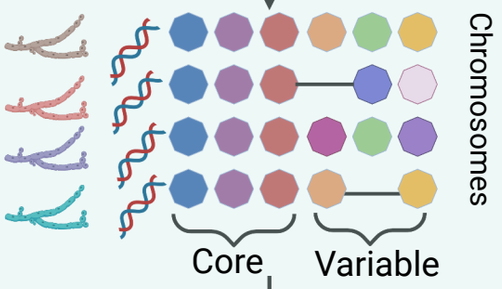
998 **Figure 2. Genome wide association mapping based on 19 reference genomes for 49 pathogen traits**
999 **measured under different host and abiotic conditions in *Zymoseptoria tritici*.** (A) Heatmap showing
1000 differences in the number of significantly associated SNPs for each trait obtained for each reference
1001 genome. Pathogen virulence (percentage of the leaf surface covered by necrotic lesions) and reproduction
1002 (pycnidia density within lesions) were measured on 12 genetically diverse wheat lines. (B) Manhattan plots
1003 showing SNP *p*-values for two traits (pathogen virulence in the left panel and melanization in presence of
1004 fungicide in the right panel) on multiple reference genomes. The shaded gray boxes highlight differences
1005 in significant associations found when using different reference genomes. The red line indicates the
1006 Bonferroni threshold at a 5% significance level. Pathogen virulence and reproduction were measured on 12
1007 genetically diverse wheat lines.
1008
1009 **Figure 3. Accumulation curves for the total number of distinct genes (identified by orthogroups**
1010 **within the species) associated with GWAS for different traits as a function of the number of reference**
1011 **genomes analyzed.** Mapping outcomes are shown for different groups of traits. The numbers in parentheses
1012 indicate the number of traits included in each category. Pathogen virulence (percentage of the leaf surface
1013 covered by necrotic lesions) and reproduction (pycnidia density within lesions) were measured on 12
1014 genetically diverse wheat lines.
1015
1016 **Figure 4. K-mer GWAS on 49 life-history traits based on a K-mer presence/absence table for all 145**
1017 ***Zymoseptoria tritici* isolates.** (A) Comparison of heritability estimates for pathogen virulence (percentage
1018 of the leaf surface covered by necrotic lesions) based on SNPs (for the reference genome IPO323) and K-
1019 mers. Both SNP-based and K-mer-based heritability were estimated by following a genome-based restricted
1020 maximum likelihood (GREML) approach. Standard errors are indicated by error bars (B) Alignment of
1021 significantly associated K-mers against the reference genome (IPO323) show the proportion of K-mers
1022 having a unique mapping position, multiple locations, or no unambiguous mapping position in host-related
1023 traits *i.e.* pathogen virulence and reproduction (pycnidia density within lesions). (C) Proportion of
1024 significant K-mers with a unique mapping position in the reference genome either tagging a gene or a
1025 transposable element for host-related traits. (D, E) Manhattan plots showing significant K-mer associations
1026 with pathogen reproduction and fungicide resistance together with quantile-quantile plots for *p*-value
1027 comparisons. Manhattan plots were created from SNP-based GWAS and blue dots represents the significant

1028 K-mer associations with the K-mers being uniquely mapped to a location in the reference genome. The two
1029 blue dots represent individual K-mers with significant associations. The red and blue lines indicate the
1030 Bonferroni and permutation-based significance threshold at 5% level for SNPs and K-mers, respectively.
1031 Pathogen virulence and reproduction were measured on 12 genetically diverse wheat lines. Overall
1032 virulence and reproduction represent the average value of the respective trait measured on 12 genetically
1033 diverse wheat lines. Reproduction specificity was estimated based on the adjusted coefficient of variation
1034 of mean reproduction across 12 genetically diverse wheat lines. Higher specificity suggests affinity to
1035 certain hosts for maximizing reproductive fitness.
1036
1037 **Figure 5. Analysis of K-mer GWAS identifying causal genes underlying major phenotypes in**
1038 ***Zymoseptoria tritici*.** (A) Manhattan plot showing significant K-mers associated with fungicide resistance.
1039 The two blue dots represent all 294 significant K-mers with a unique genomic position on chromosome
1040 seven tagging the *CYP51* gene encoding the target of azole fungicides. The red and blue lines show the
1041 Bonferroni and permutation-based significance threshold (α=0.05) for SNP and K-mer GWAS,
1042 respectively. (B) Quantile-Quantile plot showing the *p*-value comparison between SNPs and K-mer based
1043 GWAS. (C) Physical position of 294 significant K-mers mapped to unique positions on chromosome seven
1044 associated with the fungicide resistance gene *CYP51*. (D) Linkage disequilibrium (LD) heatmap showing
1045 the pairwise r^2 value among 294 significant K-mer presence/absence genotypes associated with the *CYP51*
1046 gene. (E) Proportion of isolates from different populations carrying significant K-mers that tagged *CYP51*.
1047 (F) Boxplot showing fungicide resistance levels in isolates with presence of the K-mers associated with the
1048 *CYP51* gene.
1049
1050 **Figure 6. K-mer based GWAS recovered a known effector gene in *Zymoseptoria tritici* with a higher**
1051 **statistical power than SNP-based GWAS.** (A) Manhattan plot showing significant K-mers associated
1052 with pathogen virulence on the wheat cultivar Runal. The two blue dots represent all 11 K-mers uniquely
1053 mapping to positions on chromosome seven and tagging the avirulence gene *Avr3D1* encoding an effector
1054 protein. The red and blue lines indicate the Bonferroni and permutation-based significance threshold
1055 (α=0.05) for SNP and K-mer GWAS, respectively. (B) Quantile-Quantile plot showing the *p*-value
1056 comparison between SNPs and K-mers. (C) Physical position of 11 uniquely mapped K-mers on
1057 chromosome seven associated with *Avr3D1*. (D) Linkage disequilibrium (LD) heatmap showing the
1058 pairwise r^2 value among 11 significant K-mers associated with *Avr3D1*. (E) Presence/absence pattern of 11
1059 significant K-mers associated with *Avr3D1* in five *Z. tritici* populations. The continuous horizontal blue
1060 line indicates isolates containing all the significant K-mers. (F) Boxplot showing pathogen virulence
1061 (percentage of the leaf surface covered by necrotic lesions) on the wheat cultivar Runal in isolates with or
1062 without the significant K-mers associated with *Avr3D1*.
1063



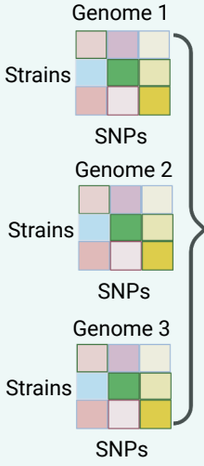
Pathogen mapping populations from different geographic origins

Whole genome sequencing of representative strains to create multiple reference genomes

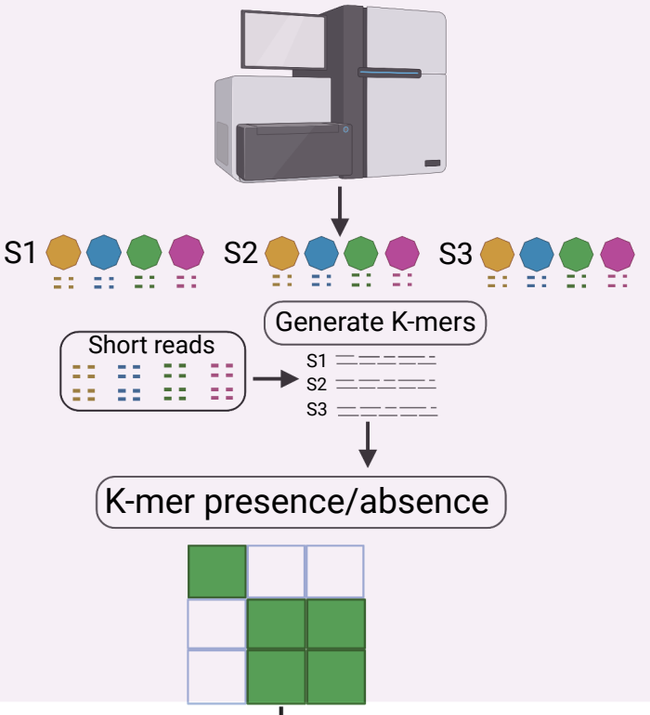


Alignment and SNP calling

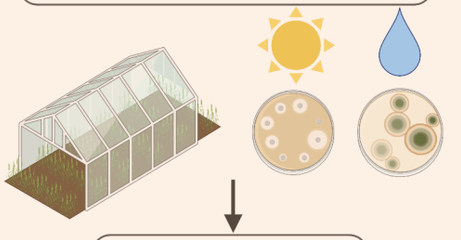
SNP datasets



Whole genome sequencing of strains



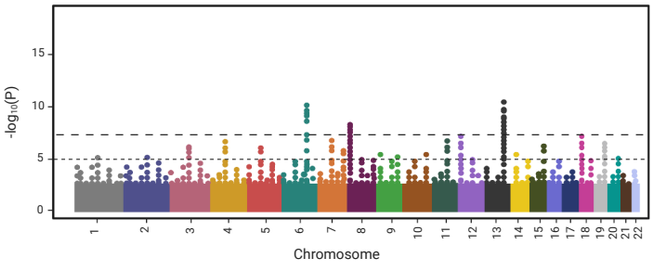
Phenotyping under biotic/abiotic conditions

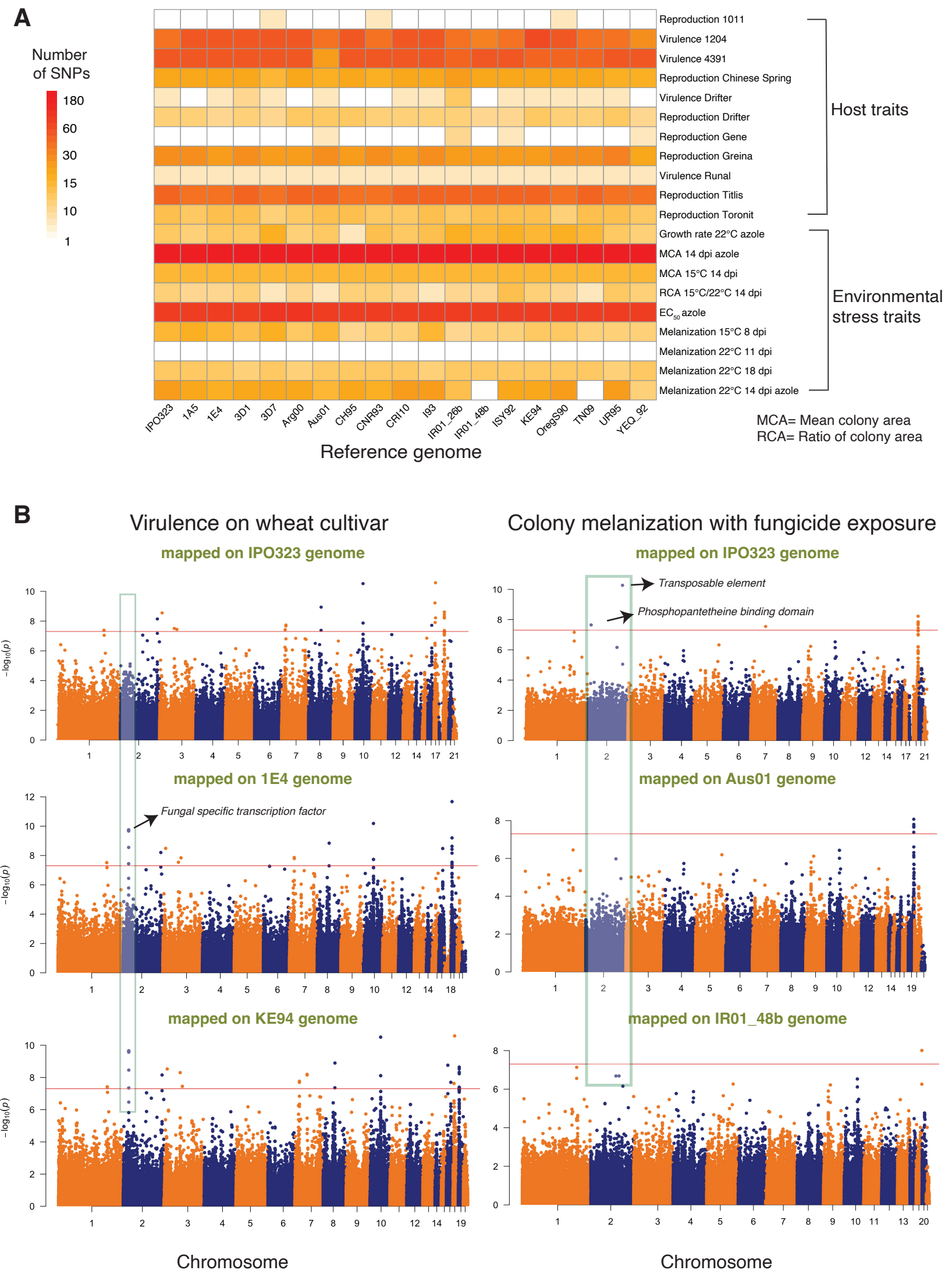


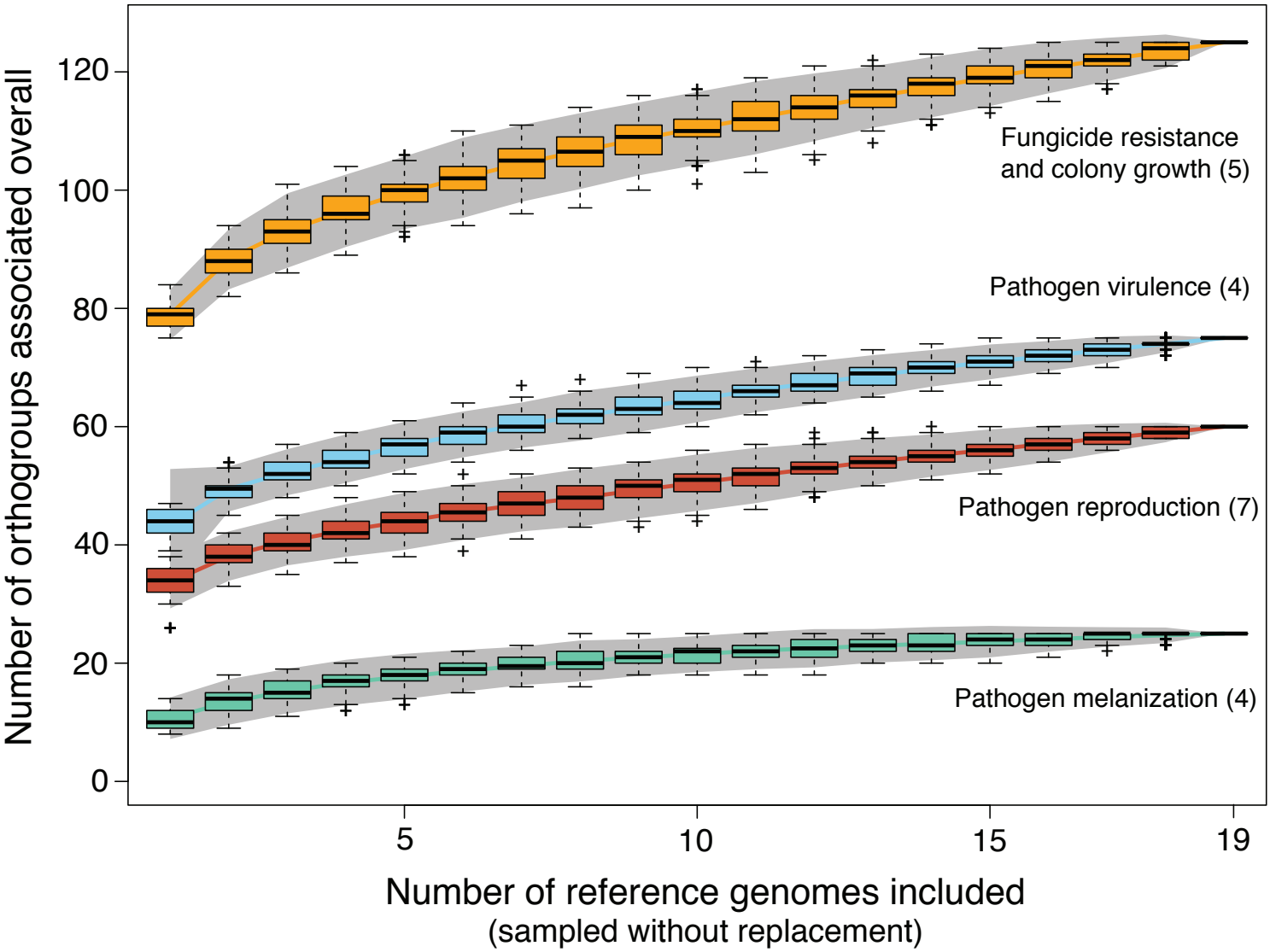
Phenotypic data

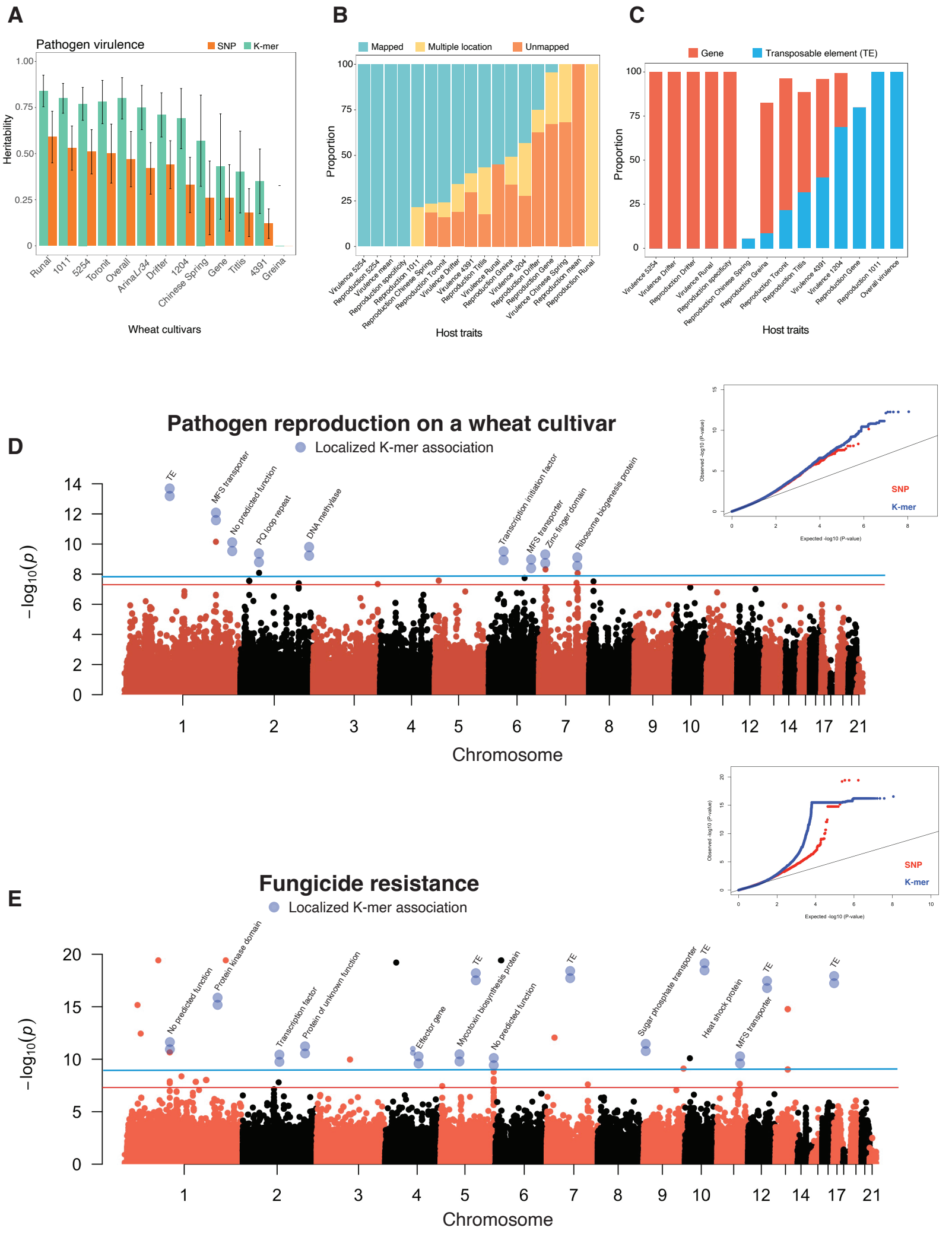
Strain	Trait 1	Trait 2	Trait 3
S1	10.2	31.0	0.5
S2	50.5	67.6	0.01
S3	90.6	43.8	0.8

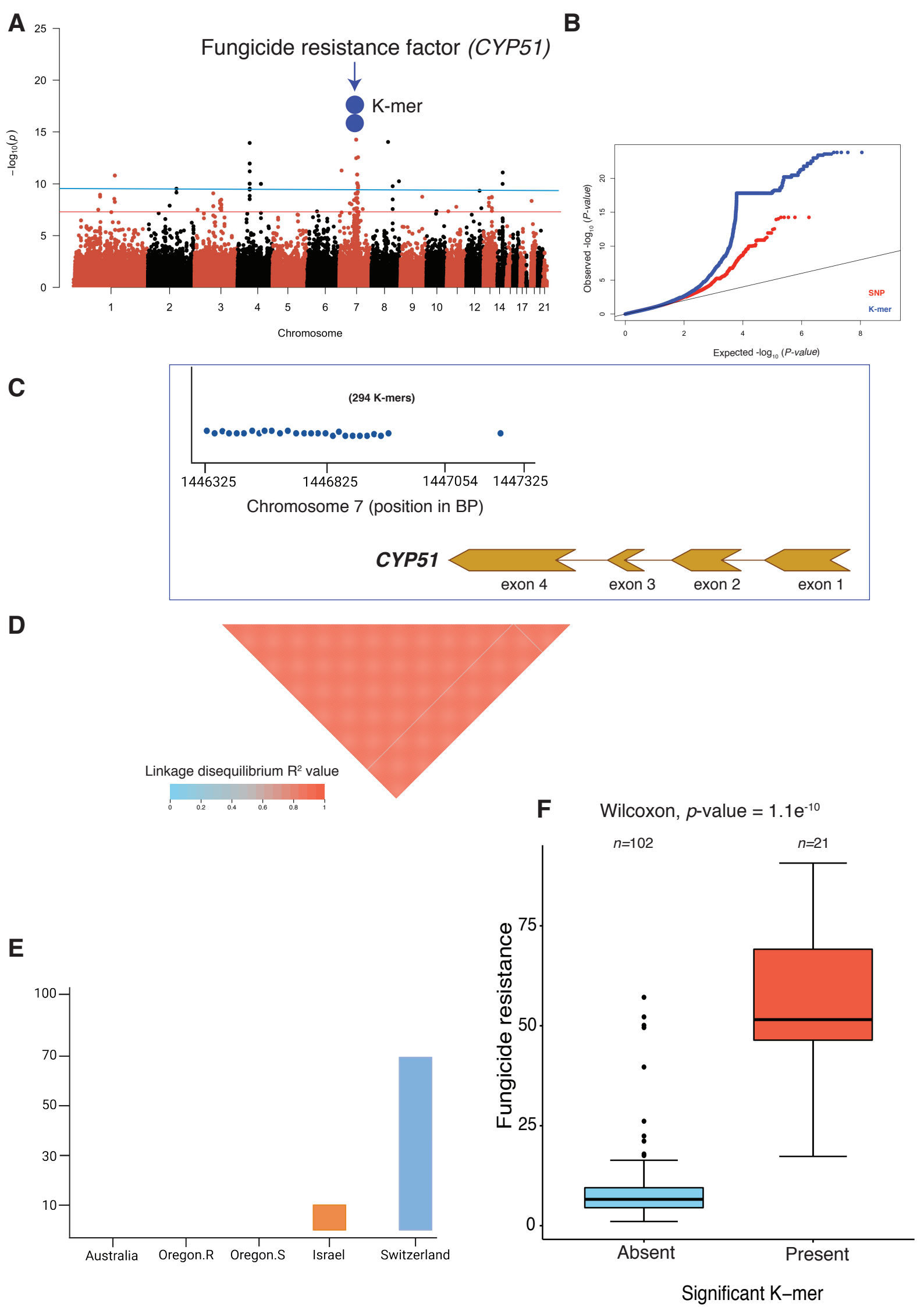
Multiple reference and K-mer based Genome wide association mapping

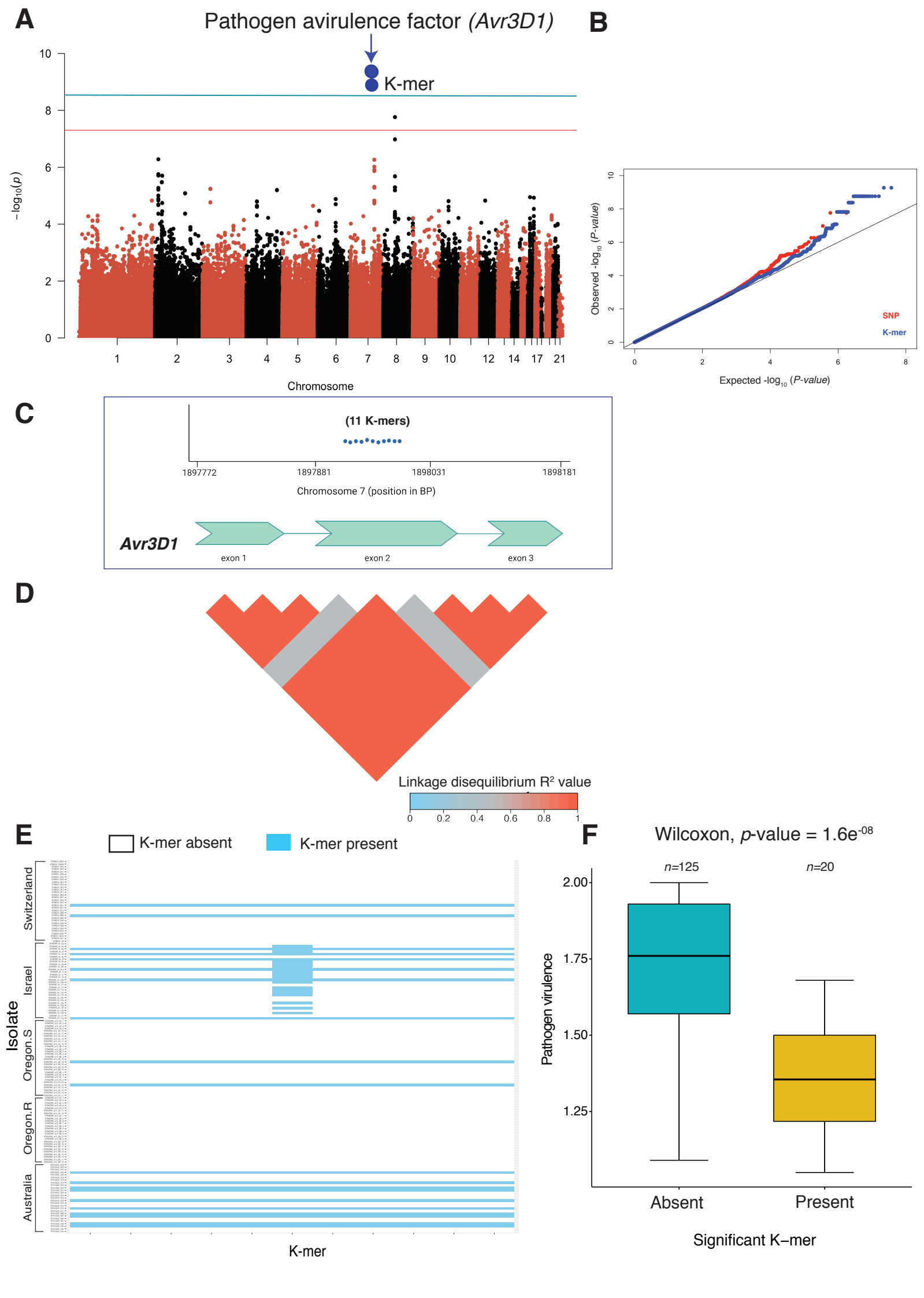



A
Number of SNPs
180
60
30
15
10
1
Reproduction 1011
Virulence 1204
Virulence 4391
Reproduction Chinese Spring
Virulence Drifter
Reproduction Drifter
Reproduction Gene
Reproduction Greina
Virulence Runal
Reproduction Titlis
Reproduction Toronit
Growth rate 22°C azole
MCA 14 dpi azole
MCA 15°C 14 dpi
RCA 15°C/22°C 14 dpi
EC50 azole
Melanization 15°C 8 dpi
Melanization 22°C 11 dpi
Melanization 22°C 18 dpi
Melanization 22°C 14 dpi azole
Host traits
Environmental stress traits
IPO323
1A5
1E4
3D1
3D7
Arg00
Aus01
CH95
CNR93
CRI10
I93
IR01_26b
IR01_48b
ISY92
KE94
OregS90
TN09
UR95
YEQ_92
Reference genome
MCA= Mean colony area
RCA= Ratio of colony area
B
Virulence on wheat cultivar
Colony melanization with fungicide exposure
mapped on IPO323 genome
mapped on IPO323 genome
Transposable element
Phosphopantetheine binding domain
mapped on 1E4 genome
Fungal specific transcription factor
mapped on Aus01 genome
mapped on KE94 genome
mapped on IR01_48b genome
−log10(p)
Chromosome
Chromosome


Number of orthogroups associated overall
0
20
40
60
80
100
120
Fungicide resistance
and colony growth (5)
Pathogen virulence (4)
Pathogen reproduction (7)
Pathogen melanization (4)
5
10
15
19
Number of reference genomes included
(sampled without replacement)


A
Pathogen virulence
SNP
K-mer
Heritability
1.00
0.75
0.50
0.25
0
Runal
1011
5254
Toronit
Overall
ArinaLr34
Drifter
1204
Chinese Spring
Gene
Titlis
4391
Greina
Wheat cultivars
B
Mapped
Multiple location
Unmapped
Proportion
100
75
50
25
0
Virulence 5254
Reproduction 5254
Virulence mean
Reproduction specificity
Reproduction 1011
Reproduction Chinese Spring
Reproduction Toronit
Virulence Drifter
Virulence 4391
Reproduction Titlis
Virulence Runal
Reproduction Greina
Virulence 1204
Reproduction Drifter
Reproduction Gene
Virulence Chinese Spring
Reproduction mean
Reproduction Runal
Host traits
C
Gene
Transposable element (TE)
Proportion
100
75
50
25
0
Virulence 5254
Virulence Drifter
Reproduction Drifter
Virulence Runal
Reproduction specificity
Reproduction Chinese Spring
Reproduction Greina
Reproduction Toronit
Reproduction Titlis
Virulence 4391
Virulence 1204
Reproduction Gene
Reproduction 1011
Overall virulence
Host traits
D
Pathogen reproduction on a wheat cultivar
Localized K-mer association
TE
MFS transporter
No predicted function
PQ loop repeat
DNA methylase
Transcription initiation factor
MFS transporter
Zinc finger domain
Ribosome biogenesis protein
$-\log_{10}(p)$
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
12
14
17
21
Chromosome
Observed -log10 (P-value)
Expected -log10 (P-value)
SNP
K-mer
E
Fungicide resistance
Localized K-mer association
No predicted function
Protein kinase domain
Transcription factor
Protein of unknown function
Effector gene
TE
Mycotoxin biosynthesis protein
No predicted function
TE
Sugar phosphate transporter
TE
Heat shock protein
TE
MFS transporter
TE
$-\log_{10}(p)$
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
12
14
17
21
Chromosome
Observed -log10 (P-value)
Expected -log10 (P-value)
SNP
K-mer


A
Fungicide resistance factor (CYP51)
K-mer
−log10(p)
Chromosome
B
Observed -log10 (P-value)
Expected -log10 (P-value)
SNP
K-mer
C
(294 K-mers)
1446325
1446825
1447054
1447325
Chromosome 7 (position in BP)
CYP51
exon 4
exon 3
exon 2
exon 1
D
Linkage disequilibrium R2 value
E
Australia
Oregon.R
Oregon.S
Israel
Switzerland
F
Wilcoxon, p-value = 1.1e-10
n=102
n=21
Fungicide resistance
Absent
Present
Significant K−mer


A
Pathogen avirulence factor (Avr3D1)
K-mer
−log10(p)
Chromosome
B
Observed -log10 (P-value)
Expected -log10 (P-value)
SNP
K-mer
C
(11 K-mers)
1897772
1897881
1898031
1898181
Chromosome 7 (position in BP)
Avr3D1
exon 1
exon 2
exon 3
D
Linkage disequilibrium R² value
E
K-mer absent
K-mer present
Switzerland
Israel
Oregon.S
Oregon.R
Australia
Isolate
K-mer
F
Wilcoxon, p-value = 1.6e-08
n=125
n=20
Pathogen virulence
Absent
Present
Significant K−mer