1 **Young genes to the front - a strategy for future resistance against powdery mildew?**

2

3 Markus Boenn^{1,2,3}, François Buscot^{2,3}, Marcel Quint⁴, Ivo Grosse^{1,3}

4

5 1 Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle/Saale, Germany

6

7 2 UFZ - Helmholtz Centre for Environmental Research, Dept. Soil Ecology, Theodor-Lieser-Str. 4,

8 06120 Halle/Saale, Germany

9

10 3 German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz

11 5e, 04103 Leipzig, Germany

12

13 4 Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg,

14 Halle/Saale, Germany

15

16

17 * Author for Correspondence: Markus Boenn, Institute of Computer Science, Martin Luther

18 University Halle-Wittenberg, Halle (Saale), Germany, markus.boenn.p@gmail.com

19

20

21

22

23

24

25

26 **Abstract**

27 Nonhost resistance of a plant against a microbial pathogen can be the result of a long-lasting
 28 coevolutionary optimization of resource allocation in both host and pathogen. Although this has
 29 been suggested for years, coevolutionary aspects leading to nonhost resistance in plants are not fully
 30 understood yet. Instead, most studies focus on limited subsets of genes which are differentially
 31 expressed in infected plants to describe details of defense strategies and symptoms of diseases.

32
 33 Here, we exploit publicly available whole genome gene expression data and combine them with
 34 evolutionary characteristics of genes to uncover a mechanism of host-pathogen coevolution. Our
 35 results suggest that metabolic efficiency in gene regulation is a key aspect leading to nonhost
 36 resistance. In addition, we find that progressing host-pathogen coevolution is accompanied by
 37 subtle, but systematic overexpression of recently founded genes. In support of our plant-specific
 38 data, we observe similar effects in animal species.

39
 40
 41
 42
 43 Key words: coevolution, phylotranscriptomics, age index, nonhost resistance, trial and error,
 44 efficiency

45

46

47

48

49

50

51 **Introduction**

52 Susceptibility of complex organisms to a small number of host-specific microbial pathogens causes
 53 diseases and, in plants, significant crop failure and yield losses each year [Oerke et al. 2012].
 54 However, in their natural habitats, organisms are exposed to a much broader range of pathogenic
 55 microbes without showing symptoms of disease for the majority of their life cycle. The mechanism
 56 responsible for this phenomenon in plants is called nonhost resistance (NHR), the details of its
 57 mechanism and evolution are not entirely understood yet [Foley et al. 2013].

58

59 NHR is part of the innate immune system of plants. Different from systemic acquired
 60 resistance in plants or the adaptive immune system in vertebrates, NHR is not established for only
 61 some individuals having been exposed to a certain pathogen. Instead, NHR has been inherited from
 62 ancestral plants and affects the entire host plant species [Heath 2000].

63

64 A number of studies investigate the plant innate immune system and NHR (reviewed in
 65 Ausubel [Ausubel 2005], Jones and Dangl [Jones & Dangl 2006] or Bettgenhaeuser et al.
 66 [Bettgenhaeuser et al. 2014], for instance), frequently due to its importance for crop yields.
 67 Accordingly, numerous genes and strategies are known to be involved in NHR, causing interaction-
 68 specific types of resistance.

69

70 Most studies focus on the description of symptoms and the exploration of small subsets of
 71 genes, which are significantly modulated in infected plants. However, as NHR is not learned during
 72 an individual's life cycle, evolutionary characteristics of genes must provide explanations for
 73 presence and absence of NHR. A process describing this phenomenon is the arms race [Hulbert et al.
 74 2001]. In the context of hosts and pathogens, the arms race describes host-pathogen coevolution
 75 [Jones & Dangl 2006] [Bettgenhaeuser et al. 2014] and is, in addition to coevolution of flowers and

76 their pollinators, probably one of the best studied coevolutionary processes of all living organisms,
77 not only plants.

78

79 Immune responses, including NHR, are energetically costly [Segerstrom 2007]. Typically,
80 two types of immune response are distinguished: resistance against a pathogen or disease tolerance.
81 In the latter case, the host focusses on maintainance and repair of the damage caused by the
82 pathogen in order to survive. As suggested by McNamara and Buchanan [McNamara & Buchanan
83 2005] as well as Segerstrom [Segerstrom 2007], resources consumed for one of the two tasks are
84 not available for other tasks.

85

86 Having been established by (co-)evolutionary selection, the interplay of immune response
87 and maintainance should be highly optimized, i.e. ensure survival and ongoing reproduction of the
88 host, while staying efficient in terms of minimal utilization of available resources [Beilharz et al.
89 1993].

90

91 A recently established method to uncover evolutionary trends in entire genomes is
92 phylostratigraphy [Domazet-Lošo et al. 2007]. Using extensive BLAST searches against sequence
93 databases comprising reference protein sequences from a large number of species, this approach is
94 able to assign an approximate evolutionary age to each protein-coding gene of a target organism,
95 based on sequence homology.

96

97 In a phylotranscriptomic approach, a combination of phylostratigraphy and gene expression
98 data has been applied to explore various scientific questions, mainly regarding developmental
99 processes like the developmental hourglass in animals [Domazet-Lošo & Tautz 2010], plants [Quint
100 et al. 2012] and also in fungi [Cheng et al. 2015]. Here, amount and complexity of data necessiate to

101 sum up the data to a scalar value. The proposed weighted mean, called Transcriptome Age Index
102 (TAI) [Domazet-Lošo & Tautz 2010], combines gene age and gene expression and is interpreted as
103 the mean evolutionary age of the transcriptome. A complementary measure incorporating
104 information about gene divergence instead of age, the Transcriptome Divergence Index (TDI), has
105 also been proposed [Quint et al. 2012].

106

107 In this work we apply phylotranscriptomic methods to investigate how presence of
108 resistance (NHR) differs from absence of resistance. We find that NHR is associated with
109 coevolutionary optimization of the immune response and mainly based on favoured recruitment of
110 older genes. In contrast to this, we further suggest that the host makes significant use of recently
111 founded genes to escape from susceptibility for a host-specific pathogen.

112

113

114

115 **Material and Methods**

116 *Phylostratigraphy*

117 Phylostratigraphy is a method to estimate the phylogenetic age of the entire set of an organisms
118 genes and is described in detail in [Domazet-Lošo et al. 2007]. For a given target organism,
119 phylostratigraphy splits the tree of life into age classes, the phylostrata. Phylostrata are identified by
120 ps1, ps2, ..., psK with psK representing the set of youngest, recently founded, genes and ps1
121 representing the set of oldest genes, of which domains are conserved in species of all living species.
122 In this work, phylostrata were selected along the lineage of *Arabidopsis thaliana*, according to the
123 NCBI taxonomy database, with ps15 being the set of youngest genes.

124

125 Protein-coding genes were assigned to phylostrata using the method of Domazet-Lošo et al.

126 [Domazet-Lošo et al. 2007]. In brief, protein sequences of representative gene models of *A.*
 127 *thaliana* were downloaded from arabidopsis.org (release TAIR10). We used BLAST+ [Camacho et
 128 al. 2009] to perform protein-protein searches against the NCBI-NR database (E-value < 1e-5).
 129 Retrieved sequences were assigned to a phylostratum, according to the Last Common Ancestor
 130 (LCA) of *A. thaliana* and the species the retrieved sequence originates from. Prior to application of
 131 BLAST, sequences originating from viruses and environmental samples were removed from NCBI-
 132 NR database.

133

134 Unless the protein is specific to *A. thaliana*, each query protein has hits in various recent and
 135 distant phylostrata, like Arabidopsis, dicots, or eukaryotes. Each query protein was assigned to the
 136 most distant phylostratum with a BLAST hit. Query proteins without hit were assigned to the
 137 youngest phylostratum ps15.

138

139

140 *Divergence*

141 Estimates of divergence between *A. thaliana* and related species were downloaded from Ensembl-
 142 database via biomaRt-package [Durinck et al. 2005] and have been derived by the codeml-function
 143 of the PAML package [Yang 1997]. These estimates comprise the number of synonymous
 144 substitutions (dS) and nonsynonymous substitutions (dN) per site, of which the ratio dN/dS was
 145 calculated. According to classic evolutionary biology, small dN/dS ratios are indicative of negative
 146 selection, whereas large dN/dS ratios are associated with genes under positive selection. However,
 147 as the absolute dN/dS value is sufficient for the analyses presented here, it is not necessary to test
 148 for specific signatures of selection.

149

150 To allow for better comparison between age of genes (represented by discrete phylostrata)

151 and divergence (represented by continuous values), dN/dS ratios were distributed into 5% quantiles
 152 (discrete representation), where each 5% quantile is a divergence stratum, identified by ds1,
 153 ds2,...,ds20. Here, ds20 represents the set of divergent, fast-evolving, genes and ds1 represents the
 154 set of genes being highly conserved in *A. thaliana* and the reference species. In total, 17651 genes
 155 were considered for the reference species *Arabidopsis lyrata*.

156

157

158 *Divergence times*

159 Unless otherwise stated, estimates of divergence times between species related to *A. thaliana* and
 160 geological times covered by phylostrata were taken from the TimeTree database [Hedges et al.
 161 2006].

162

163

164 *Microarray data & Filtering*

165 We used previously published microarray data (Affymetrix Arabidopsis ATH1 Genome Arrays)
 166 investigating NHR in *A. thaliana*. In the experiment the authors challenged plants by two fungal
 167 pathogens [Stein et al. 2006]. In the host-specific treatment (H) plants were inoculated with the
 168 powdery mildew *Erysiphe cichoracearum* (synonym for *Golovinomyces cichoracearum*). In the
 169 nonhost-specific treatment (NH) plants were inoculated with the grass mildew *Blumeria graminis*
 170 f.sp. *hordei*; its natural host is barley. Material for four biological replicates per condition (12
 171 samples in total) has been collected from rosettes one day after inoculation. Normalized data was
 172 downloaded from NCBI-GEO database (accession GSE3220).

173

174 From microarray data only probesets of genes present in the current release of TAIR
 175 (TAIR10) were kept. Probesets representing multiple genes were removed. If a gene is represented

176 by multiple probesets (167 genes), expression values of corresponding probesets were summarized
177 by the arithmetic mean.

178

179 Together, expression values of 20096 genes were considered for further analyses.

180

181

182 *Additional datasets*

183 To support findings of the main text, we investigated additional datasets based on microarrays and
184 RNA-Seq experiments. For a brief description of design, preprocessing and assignment of
185 phylostrata to genes we refer to Supporting Datasets.

186

187

188 *Regulation strength*

189 We applied fold-change (FC) as a typical measure to assess strength and direction of gene
190 regulation. The FC of each gene was defined as the ratio of mean of raw expression values in
191 treatment and mean of raw expression values in control.

192

193

194 *Regulatome based indices*

195 In analogy to the Transcriptome Age Index (TAI) introduced by Domazet-Lošo and Tautz
196 [Domazet-Lošo & Tautz 2010] (see Supporting Note), the Regulatome Age Index (RAI) was
197 obtained by substituting expression values of each condition in the formula for the TAI by FCs
198 between conditions. The term 'regulatome' describes the set of genes modulated in an experiment
199 [Ponomarev et al. 2010]. To enable focussing on the direction of modulation, two types of RAI were
200 defined, one for induction and one for repression.

201

202 For a given comparison c and a set of N genes the RAI for up-regulated (induced) genes was
203 defined as

$$204 \quad RAI_c^{up} := \frac{\sum_{n=1}^N p s_n f c_n}{\sum_{n'=1}^N f c_{n'}}$$

205 with $f c_n$ being the FC of gene n. For down-regulated (repressed) genes, the inverted FC was used:

$$206 \quad RAI_c^{down} := \frac{\sum_{n=1}^N p s_n 1/f c_n}{\sum_{n'=1}^N 1/f c_{n'}}$$

207

208 Quint et al. [Quint et al. 2012] introduced a complementary measure, the Transcriptome
209 Divergence Index (TDI), using information about gene divergence instead of gene age. In analogy
210 to the TAI-TDI relationship, phylostrata were substituted by divergence strata in the definition of
211 the RAI. The resulting measure is called Regulatome Divergence Index (RDI). While the RAI
212 provides information about the mean age of gene regulation, the RDI quantifies selective pressure
213 on gene regulation, providing information about selective pressure of modulated genes and possible
214 evolutionary constraints affecting gene regulation.

215

216

217 *Comparison of regulation strengths*

218 To compare the two regulation strengths per phylostratum or divergence stratum, we calculated the
219 difference of FC_n^{NH} and FC_n^H , where n is the gene and H and NH are the comparisons. This was
220 done for each direction of modulation separately. The calculated difference can be rewritten as
221 $(e_n^{NH} - e_n^H)/e_n^{Co}$, i.e. the difference between expression values, normalized by the expression in
222 control conditions. From resulting stratum-wise distributions, we took the median as a

223 representative value.

224

225 Compared to the arithmetic mean, the median has the advantage that it allows for
 226 quantifying interpretations, as it divides the set of genes in two groups of equal size. E.g.,
 227 considering the toy example of gene expression values (1,1,2,2,3,4,4,4,100) we can say that more
 228 'genes' have a value greater than the median ($m=3$), which is not possible for the arithmetic mean
 229 ($M=13.4$) due to the outlier. In the same sense, considering the toy example of differences between
 230 gene expression values (-5,-3.5,-4,-2,-1,7,10) we can say that more 'genes' have a difference greater
 231 than the median ($m=-2$), which is not possible for the arithmetic mean ($M=0.21$).

232

233 Standard errors for each stratum were obtained by applying a two-sample bootstrap
 234 approach within the stratum, given the direction of regulation. In detail, we took a random sample
 235 from pairwise differences between treatment-wise FC (with repetition) and calculate the median.
 236 This procedure was repeated 1000 times. From the resulting distribution of medians, the standard
 237 deviation represents the standard error of the observed median difference between treatments.

238

239

240 *Occupation of metabolic resources*

241 In the absence of an established estimate for metabolic resources used along the entire process of
 242 gene expression, from transcription to translation, we used the transcript concentration, i.e. the
 243 expression value obtained from microarray data, as an approximation.

244

245 Occupation (or release) of resources in treatments with pathogens were estimated by the
 246 difference between the expression value in a treatment and the expression value in uninoculated
 247 samples.

248

249 To assess the amount of expressed transcripts on average, for curves representing host-
250 specific treatment H a segmented regression approach was applied to fit a line to the steady increase
251 accross young genes and to find the point where dominance of resource occupation stops. For the
252 same set of genes a regression line was fitted to the curve representing nonhost-specific treatment.

253

254

255 *Availability of data and methods*

256 To perform analyses, the statistical programming language R was used. Routines were summarized
257 in the R package 'phyintom', currently deposited at <https://sourceforge.net/projects/phyintom/>. The
258 package comprises the routines as well as manuals and a vignette to reproduce essential findings
259 presented in the main text of this work.

260

261

262

263 **Results**

264 *Choice of data*

265 To understand NHR, it is imperative to also understand what happens when plants are not resistant
266 against a pathogen and are in a stage of coevolutionary optimization. Thus, careful selection of a
267 dataset with a proper experimental design is critical. We choose the dataset from Stein et al. [Stein
268 et al. 2006] with a design as depicted in Figure 1a.

269

270 This dataset covers two essential, mutually exclusive treatments. Plants are inoculated with a
271 host-specific pathogen (H). To compare this state of susceptibility with the opposite state,
272 resistance, plants are treated with a nonhost-specific pathogen (NH) in an independent experiment.

273 Both treatments are compared to control, i.e. uninoculated, samples (Co).

274

275 Two directions of modulation are considered. Induced genes are stronger expressed in
276 treatments, compared to the control (Figure 1b). Repressed genes are stronger expressed in control,
277 compared to the treatment (Figure 1c). The large overlap between comparisons indicates high
278 agreement of both treatments, considering only the direction but not the strength of modulation.

279

280

281 *Age and divergence correlate with expression and regulation*

282 To determine phylostrata [Domazet-Lošo et al. 2007], protein-coding genes of *A. thaliana* were
283 assigned to a set of 15 distinct phylostrata (see Table 1), each representing the evolutionary age of a
284 certain set of genes. According to previous findings [Wolf et al. 2009], we confirm that expression
285 values of these genes increase with age (Supplementary Figure S1).

286

287 We used the log₂-transformed fold change (log(FC)) to create a plot in analogy to
288 Supplementary Figure S1 for gene regulation, shown in Supplementary Figure S2. As the FC
289 provides information about the direction of the modulation, phylostratum-wise distributions of
290 transformed FC are shown for induced and repressed genes separately.

291

292 We find that induction of genes systematically decreases with gene age, i.e. the younger the
293 genes are, the larger is their FC. This is observed for both treatments, but the relationship is stronger
294 for treatment H. For repressed genes, no systematic dependence on age is visible for treatment H
295 (see Supplementary Table S1), but for treatment NH. Here, Kendall's rank-based correlation
296 coefficient indicates stronger repression of young genes.

297

It is possible, however, that correlations between age and magnitude of modulation are artificial. Gene expression data used in this work has been normalized using MAS 5.0 for Affymetrix microarrays [Stein et al. 2006]. The FC calculated from such data tends to be biased towards larger values when low transcript concentrations are involved, compared to a FC from high transcript concentrations [Wu et al. 2004]. Hence, recalling that young genes exhibit low expression values (Supplementary Figure S1), slopes shown for induced genes are less surprising.

From a technical point of view, these observations have significant impact on further analyses. Typically, 1.5- or 2-fold modulation of genes is considered to be meaningful. Applying these cutoffs to the data (horizontal gray lines in Supplementary Figure S2), a large proportion of genes assigned to distant (old) phylostrata would be excluded from further analyses.

To analyse the relationship between expression values and selection pressure on genes as well, we assigned dN/dS ratios (in the context of this study synonymously used with the term *sequence conservation*) to 20 divergence strata. According to previous studies [Quint et al. 2012] [Drost et al. 2015] we confirm that age and sequence conservation exhibit only weak dependence according to Kendall's rank-based correlation coefficient and Cramer's V (Supplementary Figure S3), suggesting that they are complementary measures for evolutionary studies.

We find that gene expression is not independent from sequence conservation (Supplementary Figure S4). In all conditions, very conserved genes exhibit high expression values. This is in line with the suggestion of Drummond et al. [Drummond et al. 2005] that highly expressed genes evolve slowly to avoid cytotoxic protein misfolding.

In analogy to Supplementary Figure S2, we also explore the relationship between FC and

sequence conservation. For treatment with H we find that conserved induced genes (low divergence strata) exhibit lower FCs, while no dependence on sequence conservation for repressed genes can be detected (Supplementary Figure S5, Supplementary Table S1). Dependence on sequence conservation in treatment NH is significantly weaker for induced genes and indicates that conserved genes exhibit lower FCs. In contrast to this, repression affects conserved genes only mildly in this comparison.

Despite the possibility of a bias introduced by microarray normalization, regarding dependence on age and divergence there are distinct differences between treatments. These differences are likely to be of biological rather than technical origin. However, based on these observations we can not apply one of the traditional cutoffs (or any other global cutoff). Instead, we consider induced genes as genes having $FC > 1$ and repressed genes as genes having $FC < 1$.

Resistance is achieved efficiently

Induced defenses are accepted as effective strategies of plants to fight against attacks by herbivores and pathogens [Karban & Myers 1989] [Kessler & Baldwin 2004]. Even more, in terms of bioenergetics induced strategies for resistance are suggested to be cost-saving as they are not activated when resistance expression is not required [Karban et al. 1997].

Hence, following the cost-benefit paradigm, resistance against the nonhost-specific pathogen NH should be efficient and achieved at minimum usage of resources. In this case the number of induced genes should be much smaller when the plant is treated with NH than with H. Otherwise, the plant takes similarly high efforts in presence of H although immune response is insufficient and ineffective.

348

349 Figure 1b reveals induction of numerous genes in both treatments, contradicting a
350 constitutive defense strategy. We find that the number of genes induced in the presence of the
351 nonhost-specific pathogen NH (N=8710, Figure 1b) is significantly smaller than one would expect
352 by chance (Binomial test, $P < 2e-16$). In contrast to this, the null hypothesis of similar numbers of
353 induced and repressed genes cannot be rejected when the plant is inoculated with pathogen H
354 (N=10145, $P = 0.17$). Accordingly, the number of genes induced in NH is significantly smaller than
355 the number of induced genes in H (McNemar's test, $P < 2e-16$).

356

357 Vice versa, Figure 1c trivially reveals repression of large numbers of genes in presence of
358 NH. This supports the idea of a cost-saving immune response and is in line with findings of
359 previous NHR studies, focussing on much smaller sets of genes [Zimmerli et al. 2004] [Stein et al.
360 2006] [Foley et al. 2013].

361

362

363 *Regulatome based indices reveal a general pattern of NHR*

364 To get a general evolutionary pattern describing NHR in plants, we applied two regulation based
365 indices, the Regulatome Age Index (RAI) and the Regulatome Divergence Index (RDI). These
366 indices combine relative gene expression and gene age/conservation and focus on gene induction
367 and repression. They outperform typical transcriptome based indices like TAI and TDI under
368 several aspects (see Supporting Note).

369

370 Both directions of modulation were taken into account. Accordingly, Figure 2 shows results
371 for RAI^{up} and RAI^{down} as well as RDI^{up} and RDI^{down} . Considering gene induction, we find that
372 modulated genes are less susceptible to evolutionary changes in treatment NH (red lines). Here,

induction affects most notably older genes and genes under weak selection (low values of RAI and RDI, respectively), compared to treatments with the host-specific fungus H. Vice versa, repression affects older genes and genes with high dN/dS ratios when the plant is inoculated with H. Reliability of observed differences between comparisons is indicated by standard errors and confirmed for the RAI by a z-test ($P < 0.01$ for induction as well as for repression). For the RDI, differences between comparisons are not significant (RDI^{up} , $P < 0.11$) or only weakly significant (RDI^{down} , $P < 0.05$), respectively.

380

For a fair comparison with the TAI, the RAI was calculated accross all genes. However, considering subsets of genes is more intuitive. Accordingly, Supplementary Figure S6 shows RAI profiles for (1) taking into account only all induced genes (10139 for H, 8714 for NH), (2) 6906 commonly induced genes, as well as (3) genes which are exclusively induced in each treatment (3229 for H, 1808 for NH) and harbour genes which are very specific to the corresponding type of interaction. For repressed genes RAI values are calculated for analogous subsets.

387

Although the dominating shape of the 'full' RAI profile is not changed in any case, consideration of subsets of genes clearly increases significance of observed differences in cases (1) and (3). However, reliability of indicated differences between comparisons in case (2) is weak, in particular for commonly repressed genes. Together, Supplementary Figure S6 indicates that the NHR pattern is mainly caused by the relatively small number of genes which are highly specific to the type of interaction, being modulated to the opposite direction in the other treatment (see Figure 1).

395

396

Regulation strengths of commonly modulated genes

398 Although observed differences between treatments are not significant in RAI profiles, due to their
399 sizes sets of commonly modulated genes are likely to harbour not only noise, but genes being
400 relevant for both types of treatments. As immune responses can be expected to be induced, we focus
401 on commonly induced genes first.

402

403 We want to investigate, if specific phylostrata ranges are responsible for the shape of the
404 RAI profile. For this, we calculated gene-wise differences of regulation strengths as expressed by
405 FCs. From the resulting distribution we take the median as a representative value. This is done for
406 each phylostratum. The median allows to draw conclusions about the number of genes exhibiting
407 higher or lower induction in one of the treatments, hence combining number of induced genes and
408 strength of induction. Medians are greater than zero, when modulation is stronger and affects more
409 genes in the nonhost-specific treatment. Otherwise, medians are less than zero.

410

411 NHR is complex due to the involvement of numerous pathways [Gill et al. 2015], requiring
412 strict control of gene expression by likewise complex regulatory mechanisms, which are most
413 notably observed for old genes [Warnefors & Eyre-Walker 2011]. Accordingly, Figure 3a reveals
414 that strength of induction is systematically higher for old genes when considering plants treated
415 with NH. By construction of phylostrata [Domazet-Lošo et al. 2007], functions of expression
416 products of these genes tend to base on evolutionarily optimized domains without significant
417 modifications since their first appearance. In contrast to this, plants treated with the host-specific
418 pathogen H systematically exhibit stronger induction of younger, recently founded genes. Together,
419 this results in a sigmoidal arrangement of phylostratum-wise medians.

420

421 Applying the same procedure to divergence strata as well (Figure 3b) reveals weak evidence
422 that genes under strongest selective pressure are stronger induced in treatment H, while conserved

423 genes tend to be stronger induced in NH.

424

425 Supplementary Figure S7 reveals that most commonly repressed genes are stronger
426 repressed in presence of NH, no matter the stratum.

427

428

429 *Systematic induction of young genes*

430 Results obtained so far were derived from FCs (Figure 2) and differences between FCs (Figure 3).
431 They suggest treatment-specific favour of age ranges regarding strength and direction of modulation
432 by condensing data or consideration of a large subset of genes.

433

434 However, the FC provides information about relative changes in transcript concentrations.
435 Hence, it is unable to distinguish between inductions from, e.g., 1 to 10 and 100 to 1000 transcripts
436 (10-fold in both cases). To further understand how absolute changes in transcript concentrations
437 affect NHR, we next consider absolute differences between transcript concentrations of control and
438 treatment groups.

439

440 Using absolute differences has the advantage that also information in terms of occupation of
441 resources is provided; in this case numbers of transcripts serve as an approximation for resources.

442

443 For a comprehensive view on the data, genes were sorted according to their phylostratum,
444 young genes first. Within each phylostratum, genes were sorted by transcript concentration in
445 control conditions, lowest values first. Then, differences between treatment and control were
446 cumulatively added to each other, accross all genes, beginning with young genes. This is motivated
447 by the finding that large numbers of young genes exhibit distinct and treatment-specific behaviour,

as they are strongest induced in H (Figure 3). When values increase, induction is indicated. Otherwise, genes are repressed. To focus on young genes, curves covering recent phylostrata are shown in Figure 4. Curves for the entire dataset are given in Supplementary Figure S8.

While results obtained so far suggest that numerous young genes are stronger induced in H, the immediate and steady increase of the orange curve in Figure 4 suggests that in fact all young genes either (i) are induced in H or (ii) do not experience significant repression. Extending the conclusion that young genes are stronger induced in H (Figure 3), this systematic induction not only comprises younger phylostrata ps15 to ps11, covering 1190 genes, but reaches back even to the evolutionarily old phylostratum ps5 (Embryophyta), covering 7324 genes in total (36.4% of all genes). Indeed, 3872 young genes in these phylostrata exhibit a $FC > 1$.

To get an impression of the number of transcripts added by each gene, a regression line was fitted to the curve for genes contained in ps15 to ps5. We find that, on average and compared to the control, gene expression increases by nine (rounded from slope=9.08, $R^2=0.86$) transcripts per gene in H.

In contrast to this, the slope of the curve representing NH for the same range of phylostrata points to the opposite direction, revealing that gene expression decreases by five (-4.87 , $R^2=0.41$) transcripts per gene. Indeed, only 2854 young genes exhibit a $FC > 1$.

From these observations we extract two pieces of information. First, young genes are less likely to play important roles in an induced immune response against NH. Second, in a cell with limited resources (in terms of free nucleotides, for instance), compared to control conditions young genes release resources, which are likely reallocated to increase expression of old genes.

473

474 The curves furthermore suggest that very highly expressed genes (located at right borders of
475 phylostrata) are repressed in each treatment. This is, however, an artifact caused by the arrangement
476 of data points. Genes exhibiting a very high expression value in control condition are outliers in the
477 corresponding stratum and are likely to exhibit a much smaller expression value in any treatment.
478 This consistently results in systematic drops at the end of each stratum.

479

480 To confirm that this does not affect our findings significantly, we recreate Figure 4 without
481 sorting genes according to their transcript concentration. Instead, within each phylostratum genes
482 are randomly permuted, followed by computation of the cumulative curve. This procedure is
483 repeated ten times and the mean cumulative curve is considered (Supplementary Figure S9). We
484 find that young genes still exhibit an immediate and steady increase in H, which is not visible for
485 NH.

486

487 However, the slope stops at the slightly younger phylostratum ps6 (Tracheophyta). This also
488 affects the number of transcripts expressed on average, now being about 14 for H, with $R^2 > 0.95$.
489 Permutation also affects the curve representing NH. Here, the negative trend is increased, i.e. more
490 negative, resulting in about ten transcripts for which no resources in terms of free nucleotides or
491 ribosomes have to be occupied ($R^2 > 0.85$).

492

493 We applied the same analyses for divergence strata as well (Figure 4 and Supplementary
494 Figure S8). We find that treatment with NH results in immediate and steady decrease of the curve,
495 indicating that significant repression of genes dominates this treatment. In contrast to this, the 50 %
496 of genes being under lowest negative selection (meaning dN/dS close to one) are dominated by
497 induction, when the host-specific treatment is considered. This is indicated by the curve for H,

which is located above the zero line for divergence strata ds20-ds11. For the remaining 50 % of genes being more conserved, repression dominates. Again, this general impression does not change by permutation and averaging (Supplementary Figure S10).

501

502

503 *Additonal datasets*

Application of the same approach to a second NHR experiment (again *A. thaliana* challenged with H and NH fungi, see Supporting Datasets for details) confirmed most patterns derived for the experiment of Stein et al. [Stein et al. 2006] (see Supplementary Figure 11 and 12). However, dataset-specific differences are visible. E.g. the NHR pattern derived by the RAI is not significant. For commonly induced genes, consideration of divergence strata reveals that more weakly conserved genes are involved in interaction with H. Furthermore, systematic resource occupation does not affect fast-evolving genes in H, but in NH.

511

We also investigated a dataset dealing with rice as well as datasets dealing with animal hosts, which are designed in a fashion that is comparable to Figure 1a (see Supporting Datasets for details). For this, again we assigned genes to phylostrata using the method introduced by Domazet-Lošo et al. [Domazet-Lošo et al. 2007]. Numbers of genes per phylostratum can be found in Supplementary Table S2. Next, we computed the RAI and accumulated transcript concentrations for each species. Considering Supplementary Figure 13 to 16, we find that in treatment H young genes accumulate transcripts in these datasets as well, mirroring the findings of Figure 4. However, the general NHR pattern exhibited by the RAI for *A. thaliana* (Figure 2) is visible in only some cases. Interestingly, in particular the NHR pattern is not visible when considering mice (Supplementary Figure 15 and 16), which rely on both innate and an evolutionarily much younger acquired immune system [Zhu et al. 2013].

522

523

524

525

526 **Discussion**

527 Applying phylotranscriptomic methods, we have observed a systematic activation of thousands of
528 young genes during a compatible interaction between a host and a microbial pathogen H. Moreover,
529 we observed that this activation is specific to the compatible interaction. In contrast to this, during
530 an incompatible interaction (NH) recruitment of old genes is favoured.

531

532 We presume that activation of thousands of young genes is a sophisticated coevolutionary
533 strategy of the host and a key element of the arms race. Here, functions of induced young genes,
534 which, by construction of phylostrata [Domazet-Lošo et al. 2007], harbour previously not
535 established domains, are combined with functions of induced old genes. This might generate new
536 ways for detection of microbial effectors and proper responses, lowering susceptibility for the
537 pathogen.

538

539 However, as complex regulatory mechanisms are rare for young genes [Warnefors & Eyre-
540 Walker 2011], a directed activation of large amounts of young genes appears to be unlikely. Instead,
541 from our point of view the results propose an undirected and trial-and-error-based strategy (TES) of
542 the host.

543

544 Young genes are usually short and consist of a low number of exons. This has been found
545 for animals by Neme and Tautz [Neme & Tautz 2013] and can be confirmed for *A. thaliana*
546 (Supplementary Figure S17 and S18). Further, their regulation requires fewer transcription factors
547 and they harbour lower numbers of other regulatory and structural elements [Warnefors & Eyre-

Walker 2011]. These characteristics of young genes indicate rapid transcription and post-transcriptional processing. Subsequently, initiation and elongation during translation of short genes tend to be faster [Ding et al. 2012]. Hence, irrespective of their originally intended biological function, expression products of young genes are rapidly available for the immune response.

In addition to this and in line with the argumentation of Drummond et al. [Drummond et al. 2005], undirected induction of young genes is less risky than undirected induction of old genes, which is a further benefit. As recently founded genes tend not to be involved in complex regulatory pathways [Warnefors & Eyre-Walker 2011], their induction is rarely expected to accidentally have negative impacts on well established and essential pathways controlling growth and metabolism, for instance.

On the other hand, the systematic induction of thousands of young genes is a metabolic challenge. The biosynthesis of nine transcripts and proteins on average occupies significant amounts of resources, ranging from free nucleotides and RNA polymerases for transcription to free ribosomes and t-RNAs for translation. As resources in cells are limited and genes are competing for them [Brewster et al. 2014], they have to be reallocated towards younger genes. Vice versa, when used for this task, they cannot be used for other tasks [Segerstrom 2007] [McNamara & Buchanan 2005], e.g. expression of old genes. Hence, the observed slightly lower induction of old genes (Figure 3) might, at least in part, be a consequence of the systematic induction of young genes.

An obvious contradiction in this scenario is that the host induces young genes at the cost of old genes, potentially lowering the effectiveness of the part of the immune response against H, which is based on old genes. However, tolerating disease by an only partially efficient immune response can be sufficient to survive pathogen attack and has been suggested to increase fitness of

573 the host [Rauw 2012]. At the coevolutionary stage of susceptibility the host is lacking mechanisms
574 to detect and respond to all effectors elicited into the cell by the pathogen. Hence, with an alternating
575 arms race in mind, we suggest that the TES is an investment into future defense strategies and is to
576 be preferred over investment of too many resources for an unpromising defense response.

577

578 We analyzed datasets from additional independent studies involving compatible interactions
579 (like H) and incompatible interactions (like NH) to confirm the presence of a TES. Here, we took
580 into account a second dataset dealing with the host *A. thaliana*, and datasets dealing with *Oryza*
581 *sativa* and the animal hosts *Mus musculus* and *Caenorhabditis elegans*.

582

583 Beside the diversity of hosts representing two eukaryotic kingdoms, setups of these
584 experiments are highly heterogenic regarding the utilized high-throughput platform (microarrays
585 and RNA-Seq) and preprocessing of data as well as types of pathogens (bacterial, eukaryotic and
586 viral) used.

587

588 Surprisingly, although these and other differences usually result in lower comparability
589 between experiments and agreement in their outcomes [Ingersoll et al. 2010] [Wang et al. 2014]
590 [Maboreke et al. 2016], we repeatedly find that curves representing H exhibit a steeper positive
591 slope across significant amounts of young genes, compared to curves representing NH.

592

593 This similar overall logic of the immune response supplements other examples of
594 convergent evolution towards similar components of the immune systems of plants and animals
595 [Ausubel 2005], indicating a re-invention of the same idea to overcome susceptibility for a
596 challenging pathogen.

597

598 Important questions remain as to how the host initiates and controls the undirected
599 expression of young genes during a TES. Phylotranscriptomic methods provide the opportunity to
600 apply powerful analyses to address these questions, provided the availability or generation of proper
601 datasets.

602

603

604

605 **Conclusions**

606 In this phylotranscriptomic study we explored publicly available data and propose that successful
607 resistance of a plant against a nonhost-specific pathogen is caused by efficient use of old genes,
608 indicating that NHR is less susceptible to evolutionary changes. We uncovered a potential approach
609 how coevolution between pathogens and hosts works and found hints that this mechanism is also
610 established in animals, indicating a re-invention of the same idea across eukaryotic kingdoms.

611

612 In this work we analysed data of the host. However, it is possible, that microbial pathogens
613 similarly rely on induction of recently founded genes to overcome defense strategies of the host.
614 This will be subject of further investigations.

615

616 Taken together, our results supplement currently existing knowledge about evolutionary
617 aspects of NHR and coevolution of hosts and pathogens.

618

619

620 **Acknowledgments**

621 We thank Robert Paxton for helpful comments on animal hosts. This work was partially supported
622 by a Flexible Pool Grant (number 50170649_#20) of the German Centre for Integrative

623 Biodiversity Research (iDiv) Halle – Jena – Leipzig.

624

625

626 References

627 **Ausubel, F. M. (2005).** *Are innate immune signaling pathways in plants and animals conserved?*,
628 *Nat. Immunol.* 6 : 973-979.

629 **Beilharz, R. G.; Luxford, B. G. and Wilkinson, J. L. (1993).** *Quantitative genetics and*
630 *evolution: Is our understanding of genetics sufficient to explain evolution?*, *J. Anim. Breed. Genet.*
631 110 : 161-170.

632 **Bettgenhaeuser, J.; Gilbert, B.; Ayliffe, M. and Moscou, M. J. (2014).** *Nonhost resistance to rust*
633 *pathogens - a continuation of continua.*, *Front Plant Sci* 5 : 664.

634 **Brewster, R. et al. (2014).** *The Transcription Factor Titration Effect Dictates Level of Gene*
635 *Expression*, *Cell* 156 : 1312 - 1323.

636 **Camacho, C. et al. (2009).** *BLAST+: architecture and applications.*, *BMC Bioinformatics* 10 : 421.

637 **Cheng, X.; Hui, J. H. L.; Lee, Y. Y.; Wan Law, P. T. and Kwan, H. S. (2015).** *A “Developmental*
638 *Hourglass” in Fungi*, *Molecular Biology and Evolution* 32 : 1556.

639 **Ding, Y.; Shah, P. and Plotkin, J. B. (2012).** *Weak 5'-mRNA secondary structures in short*
640 *eukaryotic genes.*, *Genome Biol Evol* 4 : 1046-1053.

641 **Domazet-Lošo, T.; Brajković, J. and Tautz, D. (2007).** *A phylostratigraphy approach to uncover*
642 *the genomic history of major adaptations in metazoan lineages.*, *Trends Genet.* 23 : 533-539.

643 **Domazet-Lošo, T. and Tautz, D. (2010).** *A phylogenetically based transcriptome age index*
644 *mirrors ontogenetic divergence patterns.*, *Nature* 468 : 815-818.

645 **Drost, H.-G.; Gabel, A.; Grosse, I. and Quint, M. (2015).** *Evidence for active maintenance of*
646 *phylotranscriptomic hourglass patterns in animal and plant embryogenesis.*, *Mol. Biol. Evol.* 32 :
647 1221-1231.

648 **Drummond, D. A.; Bloom, J. D.; Adami, C.; Wilke, C. O. and Arnold, F. H. (2005).** *Why highly*
649 *expressed proteins evolve slowly.*, *Proc. Natl. Acad. Sci. U.S.A.* 102 : 14338-14343.

650 **Durinck, S. et al. (2005).** *BioMart and Bioconductor: a powerful link between biological*
651 *databases and microarray data analysis.*, *Bioinformatics* 21 : 3439-3440.

652 **Foley, R. C.; Gleason, C. A.; Anderson, J. P.; Hamann, T. and Singh, K. B. (2013).** *Genetic and*
653 *genomic analysis of Rhizoctonia solani interactions with Arabidopsis; evidence of resistance*
654 *mediated through NADPH oxidases.*, *PLoS ONE* 8 : e56814.

655 **Gill, U. S.; Lee, S. and Mysore, K. S. (2015).** *Host versus nonhost resistance: distinct wars with*
656 *similar arsenals.*, *Phytopathology* 105 : 580-587.

657 **Heath, M. C. (2000).** *Nonhost resistance and nonspecific plant defenses.*, *Curr. Opin. Plant Biol.* 3 :
658 315-319.

659 **Hedges, S. B.; Dudley, J. and Kumar, S. (2006).** *TimeTree: a public knowledge-base of*
660 *divergence times among organisms.*, *Bioinformatics* 22 : 2971-2972.

661 **Hulbert, S. H.; Webb, C. A.; Smith, S. M. and Sun, Q. (2001).** *Resistance gene complexes:*

- 662 *evolution and utilization.*, Annu Rev Phytopathol 39 : 285-312.
- 663 **Ingersoll, M. A. et al. (2010).** *Comparison of gene expression profiles between human and mouse*
664 *monocyte subsets*, Blood 115 : e10-e19.
- 665 **Jones, J. D. G. and Dangl, J. L. (2006).** *The plant immune system.*, Nature 444 : 323-329.
- 666 **Karban, R.; Agrawal, A. A. and Mangel, M. (1997).** *The benefits of induced defenses against*
667 *herbivores*, Ecology 78 : 1351-1355.
- 668 **Karban, R. and Myers, J. H. (1989).** *Induced plant responses to herbivory*, Annual Review of
669 Ecology and Systematics : 331-348.
- 670 **Kessler, A. and Baldwin, I. T. (2004).** *Herbivore-induced plant vaccination. Part I. The*
671 *orchestration of plant defenses in nature and their fitness consequences in the wild tobacco*
672 *Nicotiana attenuata.*, Plant J. 38 : 639-649.
- 673 **Maboreke, H. R. et al. (2016).** *Transcriptome analysis in oak uncovers a strong impact of*
674 *endogenous rhythmic growth on the interaction with plant-parasitic nematodes.*, BMC Genomics
675 17 : 627.
- 676 **McNamara, J. and Buchanan, K. (2005).** *Stress, resource allocation, and mortality*, Behavioral
677 Ecology 16 : 1008-1017.
- 678 **Neme, R. and Tautz, D. (2013).** *Phylogenetic patterns of emergence of new genes support a model*
679 *of frequent de novo evolution.*, BMC Genomics 14 : 117.
- 680 Oerke, E.-C.; Dehne, H.-W.; Schönbeck, F. and Weber, A., **2012.** *Crop production and crop*
681 *protection: estimated losses in major food and cash crops.* Elsevier, .
- 682 **Ponomarev, I.; Rau, V.; Eger, E. I.; Harris, R. A. and Fanselow, M. S. (2010).** *Amygdala*
683 *transcriptome and cellular mechanisms underlying stress-enhanced fear learning in a rat model of*
684 *posttraumatic stress disorder.*, Neuropsychopharmacology 35 : 1402-1411.
- 685 **Quint, M. et al. (2012).** *A transcriptomic hourglass in plant embryogenesis.*, Nature 490 : 98-9101.
- 686 **Rauw, W. M. (2012).** *Immune response from a resource allocation perspective.*, Front Genet 3 :
687 267.
- 688 **Segerstrom, S. C. (2007).** *Stress, Energy, and Immunity: An Ecological View.*, Curr Dir Psychol Sci
689 16 : 326-330.
- 690 **Stein, M. et al. (2006).** *Arabidopsis PEN3/PDR8, an ATP binding cassette transporter, contributes*
691 *to nonhost resistance to inappropriate pathogens that enter by direct penetration.*, Plant Cell 18 :
692 731-746.
- 693 **Wang, C. et al. (2014).** *The concordance between RNA-seq and microarray data depends on*
694 *chemical treatment and transcript abundance*, Nature biotechnology 32 : 926-932.
- 695 **Warnefors, M. and Eyre-Walker, A. (2011).** *The accumulation of gene regulation through time.*,
696 Genome Biol Evol 3 : 667-673.
- 697 **Wolf, Y. I.; Novichkov, P. S.; Karev, G. P.; Koonin, E. V. and Lipman, D. J. (2009).** *The*
698 *universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes*
699 *of different apparent ages.*, Proc. Natl. Acad. Sci. U.S.A. 106 : 7273-7280.
- 700 **Wu, Z.; Irizarry, R. A.; Gentleman, R.; Martinez-Murillo, F. and Spencer, F. (2004).** *A model-*
701 *based background adjustment for oligonucleotide expression arrays*, Journal of the American
702 statistical Association 99 : 909-917.
- 703 **Yang, Z. (1997).** *PAML: a program package for phylogenetic analysis by maximum likelihood.*,

704 Comput. Appl. Biosci. 13 : 555-556.

705 **Zhu, M. et al. (2013).** *A Silurian placoderm with osteichthyan-like marginal jaw bones.*, Nature
706 502 : 188-193.

707 **Zimmerli, L.; Stein, M.; Lipka, V.; Schulze-Lefert, P. and Somerville, S. (2004).** *Host and non-*
708 *host pathogens elicit different jasmonate/ethylene responses in Arabidopsis.*, Plant J. 40 : 633-646.

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732 **Table 1:** Phylostrata along the lineage of *A. thaliana*

ps	taxon	genes
15	Arabidopsis thaliana	393
14	Arabidopsis	119
13	Camelineae	119
12	Brassicales	310
11	rosids	249
10	dicots	398
9	Mesangiospermae	566
8	Magnoliophyta	695
7	Spermatophyta	432
6	Tracheophyta	699
5	Embryophyta	3344
4	Streptophyta	228
3	Chlorobionta	647
2	eukaryotes	5996
1	cellular organisms	5901
	total	20096

733

734

735

736

737

738

739

740

741

742

743

744

745 **Figure legends**

746

747 **Fig. 1** Design of the study. (a): The NHR-experiment considers three groups: an uninoculated
748 control (Co), a group with plants challenged by the host-specific pathogen (H) and a group with
749 plants challenged by a non-host-specific pathogen (NH), which is not compatible to the plant under
750 investigation, but to plants from more distant phyla. (b) and (c): The two treatments are compared to
751 the control group. Genes are considered induced if the fold-change is greater than one (b). Genes
752 are considered repressed if the fold-change is less than one (c). Seven genes exhibit a fold-change
753 of exactly one in the comparison with treatment NH. To ensure that numbers of genes sum up to
754 20096, they are included in the sets regarding modulation in treatment H.

755

756 **Fig. 2** Regulome Age Index (RAI) and Regulome Divergence Index (RDI) for the data.
757 Subfigures (a) and (b) represent RAI and RDI, respectively. The RDI is given for the related species
758 *Arabidopsis lyrata*. Measures distinguish between the two directions of modulation, RAI^{up} (or
759 RDI^{up}) and RAI^{down} (or RDI^{down}) for induction and repression of genes, respectively.

760

761 **Fig. 3** Comparison of commonly induced genes. Medians of pairwise differences of genes in each
762 phylostratum and divergence stratum are shown for commonly induced genes. Horizontal zero lines
763 determine whether regulation in treatment NH (points are above line), or treatment H (points are
764 below line) is stronger. Strata comprising young or fast-evolving genes, respectively, are located on
765 the left. Vertical gray lines separate points at right-most stratum having points below zero line,
766 allowing for a single outlier. Stars indicate significance of enrichment of points below zero line
767 (Fisher's test, *: $P < 0.05$, ***: $P < 0.001$).

768

769

770 **Fig. 4** Accumulation of transcript expression accross genes. Figure shows cumulative curves of
 771 differences between expression levels in control and treatment groups. For this, genes are ordered
 772 by their phylostratum (or divergence stratum), so that young (divergent) genes are on the left and
 773 old (conserved) genes are on the right. Borders between strata are given by vertical lines. Within
 774 each stratum, genes were sorted according to their expression value in uninoculated plants (lowest
 775 first). Each point corresponds to a gene. Curves are given for comparison with host-specific
 776 treatment H (orange points) and nonhost-specific treatment NH (green points). When genes are
 777 induced or repressed, the curve increases or decreases, respectively. Curves are given for youngest
 778 (most divergent) genes only.

779

a **b**

a **b**

Co vs NH (8714)

c

Co vs NH (11375)