Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften an der Fakultät für
Biologie der Ludwig-Maximilians-Universität München

EVOLUTION OF DISEASE RESISTANCE GENES IN WILD TOMATO SPECIES

Anja Christina Hörger

München 2011

EVOLUTION OF DISEASE RESISTANCE GENES IN WILD TOMATO SPECIES

Dissertation
der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

vorgelegt von
Anja Christina Hörger
aus München

München, den 05.04.2011

1. Gutachter: Prof. Dr. Wolfgang Stephan

2. Gutachter: Prof. Dr. John Parsch

Datum der mündlichen Prüfung: 30. Mai 2011

Erklärung:

Diese Dissertation wurde im Sinne von §12 der Promotionsordnung von Prof. Dr. Wolfgang Stephan betreut. Ich erkläre hiermit, dass die Dissertation nicht einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich nicht anderweitig einer Doktorprüfung ohne Erfolg unterzogen habe.

Ehrenwörtliche Versicherung:

Ich versichere ferner hiermit ehrenwörtlich, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt wurde.

München, den 05.04.2011

Anja Hörger

NOTE

In this thesis, I present my doctoral research, all of which has been done by myself except for the following: The population-based study on five resistance genes was done in collaboration with Lukasz Grzeskowiak, who sequenced *Pto* and *Fen*, and with Martin Groth, who sequenced *Prf*. The species wide sampling approach was done in collaboration with Katharina Böndel, who kindly provided *Pto* and partly *Rin4* sequence data. Hilde Lainer and Gisela Brinkmann provided excellent technical assisstance (sequencing of the reference loci and partly *Pfi*). Aurélien Tellier contributed to the gene conversion analysis for the *Rcr3* gene.

The results of my thesis have contributed to the following publication:

Rose, L. E., L. Grzeskowiak*, A. C. Hörger*, M. Groth, and W. Stephan (2011). Targets of selection in a disease resistance network in wild tomatoes, *Molecular Plant Pathology* (accepted pending minor revisions), *These authors contributed equally.

ACKNOWLEDGEMENTS

First, I would like to thank Prof. Wolfgang Stephan for convincing me to join the Population Genetic group at the end of my Diplom studies and for offering me the opportunity to do my PhD in his group. He supported and motivated me whenever it was needed and he always had a sympathetic ear for me. I enjoyed my time in the group very much and learned a lot. I also thank Laura Rose for the opportunity of doing my PhD in her lab, for providing everything I needed to complete my work and for motivating discussions. Both, Wolfgang and Laura, contributed largely to the improvement of my scientific skills.

Further, I am very grateful to Aurélien Tellier for continued encouragement, mental and scientific support, plenty of fruitful discussions, awsome ideas and so much more.

I would not have been able to perform an interesting part of my project – the Rcr3 functional studies – if it had not been for Renier van der Hoorn. He welcomed me in the Plant Chemetics Lab in the Max Planck Institute in Cologne and provided everything necessary to complete the functional experiments. His encouragement contributed largely to the successful completion of this project. Further, I would like to thank Muhammad Ilyas for giving me a great start in the world of plant chemetics and for tolerating my attempts to stress out the whole group.

A huge „Thank you“ goes to the whole Munich Evolutionary Biology Group. I had a great, memorable, funny, inspiring time with you in, but also outside the lab. Hilde Lainer for excellent technical support and for making the lab a warm place, Gisela Brinkmann for excellent technical support and getting the nearly impossible working, Simone Lange for the crazy time in the lab (involving Gummibärchen, Dornenvögel, dry ice…), Iris Fischer for a memorable journey to Egypt, Claus Kemkemer for his constant and nevertheless appreciated visits in the lab, Winfried Hense for political and magical discussions, Miri Linnenbrink for her taste of music and her sunny personality, Katharina Böndel for being a model Diplom student, Lukasz Grzeskowiak for fruitful collaboration, Francesco Paparazzo for entertaining episodes in the lab, Mamadou Mboup for inspiring discussions, a great time in the office and his occasional confusion, Anne Wilken and Hedwig Gebhart for their helpfulness, Anica Vrljic for being the heart and spirit of the group, Ingrid Kroiss for her irreplaceable support whenever one has to deal with German administration, John Parsch and Dirk Metzler for helpful comments during my seminar talks, Lena Müller for her pleasant company, Martin Hutzenthaler for the great skiing trips, Rayna Stambolyiska for her patience and a great time in the office, Astrid Stück for a great time in the office as well, Sonja Grath for the relaxing barbecue sessions, Carlos Merino for never getting bored because of him, Nicolas Svetec for the nice hiking and skiing trips, Kim Steige for her pleasant personality, Robert Piskol, Ana Catalán, Susi Voigt, Annegret Werzner and Ricardo Wilches for entertaining discussions and Lisha Naduvilezhath, Stefan Laurent, Daniel Zivkovic, Meike Wittmann, Pavlos Pavlidis, Stephan Hutter, Pablo Duchén for sporty hours in our football team. And not to forget: W. Schnitzel, who provided a good piece of entertainment.

I would like to thank everybody in the Plant Chemetics Lab in Cologne for the great working atmosphere: Selva Kumari for help with the protein purification and great bench neighbourship, Izabella Kolodziejek (the Polish Princess) for entertaining and encouraging discussions, Takayuki Shindo for useful hints and the course in Japanese, Farnusch Kaschani, Kerstin Richau and Christian Gu for useful comments and good mood in the lab, Johana Misas for encouraging lunch time discussions, Bikram Pandey for help with the Typhoon scanner and not feeling bored when waiting for the results. I am also very grateful to Madlen Vetter for generously offering me a home.

Thank you to Hans Nitschko, who encouraged me to “run head first through the wall – if it is not too solid”. I did so and it was the right decision.

A big thanks goes to my dear friends for legendary times and conserving my mental health: Thomas Kunze, Martina Eder, Corinna Steinmeyer, Andy Stöger, Wolfgang Hell, Susanne Spieß, Andrea Brandmeier, Steffi Tews, Achim Gütlein and Xaver Steemann.

Last but not least, I owe very special thanks to my family, especially my parents and my sister, who are always there for me and support me in every respect. Without you, this would not have been possible.

SUMMARY

The coevolutionary arms race between hosts and pathogens is often described as a recurrent struggle for increased resistance in hosts and evasion of recognition by pathogens. These coevolutionary dynamics dominated by balancing selection lead to the maintenance of allelic diversity at genes involved in interactions between hosts and pathogens. In plant-pathogen interactions, the current paradigm posits that the specific defence response is activated upon recognition of a specific pathogen effector through the corresponding resistance (R) gene in the host. Numerous studies demonstrated that balancing selection acts on these *R* genes. However, little is known about the evolutionary mechanisms shaping other molecules not directly involved in pathogen recognition, but nevertheless playing an important role in defence signal activation. In this thesis I investigate the evolutionary forces acting at these genes in wild tomato species (*Solanum* sp.).

First, I focus on Rcr3, a ‘guardee’, *i.e.* target of pathogen effectors secreted by the fungus *Cladosporium fulvum* and the oomycete *Phytophthora infestans* in tomato plants. Specific activation of the defence response occurs when *R* genes (the ‘guards’) sense the modification of the ‘guardee’ by pathogen effectors (‘Guard-Hypothesis’). These interactions between effector, ‘guardee’ and ‘guard’, are expected to favour contrasting evolutionary forces acting on the guardee. I study the pattern of sequence evolution and functional consequences of natural sequence variation on host resistance and show that the evolution of *Rcr3* is characterized by gene duplication, gene conversion and balancing selection in wild tomato species. Investigating the functional characteristics of 54 natural variants through *in vitro* and *in planta* assays, I reveal differences in the strength of the defence response, but not in pathogen recognition specificity. These results suggest that functional diversity may be maintained at the ‘guardee’ (*Rcr3*) through the coevolution with its ‘guard’ because natural selection favours improved transduction of the defence signal or avoidance of auto-immune response.

Second, I study the pattern of polymorphism in one population of the wild tomato species *S. peruvianum* at five genes (*Pto, Fen, Rin4, Prf* and *Pfi*) involved in a common defence signalling network. This network contributes to resistance against the bacterium *Pseudomonas syringae*. Two of these genes, *Pto* and *Pfi,* exhibit a signature of balancing selection but only *Pto* is known to directly interact with pathogen ligands in pathogen recognition. *Pfi* however was found to function further ‘downstream’ in the network. These results suggest that pathogens may target genes at different positions of the resistance networks to manipulate or nullify host resistance. I further investigate the evolution of these two genes in three recently diverged sister species (*S. peruvianum*, *S. corneliomulleri* and *S. chilense*) using a species wide sampling approach. Both genes exhibit trans-species polymorphism, but it is shown that at the *Pto* gene this is most likely due to recent introgression of favourable alleles, where *Pfi1* exhibits ancestral trans-species polymorphism. Interestingly, *Pfi* shows signature of enhanced divergence between species, suggesting that this gene may represent a potential example of Dobzhansky-Muller incompatibility.

Altogether, these results suggest that coevolution occurs not only at genes of interaction between hosts and pathogens, but as well at genes indirectly involved in recognition (guardees) or signal transduction. Understanding the evolution of the plant immune system requires therefore extending the scope of functional and population genetics studies to signalling molecules in defence networks.

ZUSAMMENFASSUNG

Coevolution zwischen Wirt und Krankheitserreger beruht auf einem periodischen Wettstreit um verbesserte Immunität des Wirtes und Vermeidung der Erkennung des Krankheitserregers durch das wirtsspezifische Immunsystem. Balanzierende Selektion ist die dominierende evolutionäre Kraft, die diese Dynamik beeinflusst. Diese Art der natürlichen Selektion trägt zum Erhalt von Variabilität an den beteiligten Genen in Wirt und Pathogen bei. In Pflanzen wird gemäß der geläufigen Annahme eine spezifische Immunantwort ausgelöst, sobald ein sogenanntes Effektormolekül, das vom betreffenden Pathogen abgesondert wird, durch das passende pflanzliche Resistenz-(R)-Gen erkannt wird. Zahlreiche Studien haben bislang gezeigt, dass diese Gene balanzierender Selektion unterliegen. Im Gegensatz dazu ist immer noch wenig über Mechanismen bekannt, die eine Rolle in der Evolution anderer bedeutender Moleküle mit vielleicht eher indirekter Beteiligung in der Aktivierung des Immunsignals spielen. In dieser Arbeit untersuche ich die evolutionäre Geschichte derartiger Moleküle am Beispiel von Wildtomaten (*Solanum* sp.).

Meine Arbeit konzentriert sich zunächst auf das Rcr3-Molekül aus der Tomate. Dieses Molekül ist ein sog. ‚guardee'-Molekül (ein „bewachtes" Molekül), d.h. es stellt ein pflanzliches Zielmolekül für Effektoren, die von hauptsächlich zwei Pathogen sekretiert werden dar: vom Pilz *Cladosporium fulvum* und vom Oomyceten *Phytophthora infestans*. Guardees spielen eine besondere Rolle bei der Aktivierung der pflanzlichen Immunantwort. Sie werden durch pathogene Effektormoleküle während der Infektion modifiziert und so in ihrer eigentlichen Funktion beeinträchtigt. Der sogenannten ‚Guard-Hypothese' zu Folge kann diese Modifikation durch das R-Gen, das den Zustand des Guardees überwacht und somit die Rolle eines ‚Guards' (Wächter-Moleküls) übernimmt, gefühlt werden. Die Interaktionen zwischen Effektor, Guardee und Guard sind von komplizierter Natur und können das Wirken kontrastierender evolutionärer Kräfte auf das Guardee-Molekül begünstigen. In dieser Arbeit untersuche ich Sequenzvariation am *Rcr3* Gen und deren Auswirkungen auf den Phänotypen der Pflanze. Meine Ergebnisse zeigen, dass die evolutionäre Geschichte des *Rcr3* Gens in Tomaten hauptsächlich durch Genduplikation, Genkonversion und balanzierende Selektion bestimmt wird. Ich führe *in vitro* und *in planta* Experimente mit insgesamt 54 natürlichen Varianten dieses Gens durch und zeige, dass diese sich in der Ausprägung der Immunantwort unterscheiden. Diese Beobachtung lässt den Schluss zu, dass funktionale Diversität von Guardee-Molekülen durch Coevolution mit dem

Guard-Molekül erhalten wird. Natürliche Selektion könnte dabei entweder die Aktivierung der Immunantwort verbessern oder ungewünschte Autoimmunantworten vermeiden.

Desweiteren beschäftigt sich meine Arbeit mit der evolutionären Geschichte von fünf Resistenzgenen (*Pto*, *Fen*, *Prf*, *Pfi* and *Rin4*), die alle im gleichen Resistenznetzwerk zusammenarbeiten. Dieses Netzwerk ist an der Immunabwehr des bakteriellen Pathogens *Pseudomonas syringae* in Tomaten beteiligt. Zwei der untersuchten Gene wiesen Hinweise auf balanzierende Selektion auf. Dabei ist aber nur eines dieser Gene, *Pto*, nachweislich an der direkten Interaktion mit pathogenen Liganden beteiligt. Das andere Gen, *Pfi*, hat jedoch vermutlich eine nachgeschaltete Funktion in diesem Netzwerk. Diese Ergebnisse deuten darauf hin, dass Pathogene dazu in der Lage sind, Moleküle an verschiedenen Stellen in Signalnetzwerken des Wirtes zu manipulieren und dadurch dessen Immunabwehr zu schwächen. Um diese Ergebnisse zu vertiefen, untersuche ich die evolutionäre Geschichte dieser Resistenzgene in drei nah verwandten Tomatenarten, die erst kürzlich voneinander abgespalten wurden. Dabei wende ich eine Beprobungsstrategie an, die den gesamten Verbreitungsraum der drei Arten abdeckt. Diese Strategie legt Trans-Spezies Polymorphismen in beiden Genen zwischen den untersuchten Arten dar. Jedoch ist dies im *Pto*-Gen eine Folge von adaptativem Genfluss zwischen den Arten und im Falle des *Pfi*-Gens ein Überbleibsel von ursprünglichen Polymorphismen, die bereits im gemeinsamen Vorfahren der Arten vorhanden waren. Interessanterweise, gibt es am *Pfi*-Gen auch Hinweise darauf, dass die untersuchten Tomatenarten eine verstärkte Divergenz aufweisen. Dies lässt darauf schließen, dass es sich bei diesem Gen um ein Beispiel für Dobzhansky-Muller Inkompatibilität zwischen Tomatenarten handeln könnte.

Zusammenfassend deuten meine Ergebnisse darauf hin, dass Coevolution zwischen Wirt und Pathogen nicht nur diejenigen Gene beeinflusst, die direkt an der Interaktion beteiligt sind, sondern auch solche Moleküle, die entweder indirekt an der Pathogenerkennung (Guardees) oder an der Weiterleitung des Immunsignals beteiligt sind. Als Fazit lässt sich sagen, dass zukünftige Studien, die sich mit der Aufklärung des pflanzlichen Immunsystems befassen, auch auf andere Moleküle innerhalb des Immunnetzwerks ausweiten sollten.

TABLE OF CONTENTS

TABLE OF CONTENTS 1

LIST OF TABLES 4

LIST OF FIGURES 5

ABBREVIATIONS 7

CHAPTER 1: INTRODUCTION **9**

1.1 The plant immune system 9

1.2 Two modes of pathogen recognition 13

1.3 Host-pathogen coevolution 15

1.4 Pathway evolution 17

1.5 The *Pto* signalling pathway 18

1.6 *Rcr3* – an example for the Guard-Hypothesis 22

1.7 Tomato as a model system 25

1.8 The use of different sampling schemes 27

1.9 Aims of this thesis 28

CHAPTER 2: MATERIALS AND METHODS **33**

2.1. Sequence evolution at the *Rcr3* gene family 33

2.1.1 Plant material and DNA sequencing 33

2.1.2 Analysis of *Rcr3* flanking regions 34

2.1.3 Long Range PCR 35

2.1.4 Population genetic analyses of the *Rcr3* gene family 35

2.2 Functional consequences of sequence variation at the *Rcr3* locus 40

2.2.1 Cloning procedure and *Agrobacterium*-mediated transient expression 40

2.2.2 Activity-based protein profiling and inhibition assays 41

2.2.3 Protein purification 41

2.2.4 Inhibition assays 42

2.2.5 HR-assays 42

2.2.6 RT-PCR 43

2.2.7 Association of genotypic and observed phenotypic data 43

2.2.8 Structural model of the Rcr3 protein 43

2.3 Evolution of the *Pto*- disease resistance pathway 44

2.3.1 Plant materials and sequencing 44

2.3.2 Population genetic analysis of the pathway 47
2.3.3 Phylogenetic inference 47
2.4 Evolution of resistance genes beyond species boundaries 48
2.4.1 Plant materials and sequencing 48
2.4.2 Population genetic analysis of the species wide sample 52
2.4.3 Phylogenetic inference 53

CHAPTER 3: RESULTS 55
3.1 Evolutionary history of the *Rcr3* gene family 55
3.1.1 The *Rcr3* locus is duplicated in *S. peruvianum* and its sister species 55
3.1.2 Two differentiated sequence types are maintained in the *Rcr3* gene family 58
3.1.3 The evolutionary history of *Rcr3* is characterized by balancing selection 60
3.1.4 The two *Rcr3* loci are affected by frequent gene conversion 63
3.2 Functional consequences of sequence variation at the *Rcr3* locus 65
3.2.1 Natural variation at the *Rcr3* locus has effects on the interaction with effectors 65
3.2.2 Amino acid variation at the Rcr3 locus translates into differential strength of HR 71
3.3. Contrasting evolutionary patterns at a disease resistance pathway 77
3.3.1 Elevated level of polymorphisms at *Pto* and *Pfi* 77
3.3.2 Mixed pattern of polymorphisms at *Fen* and *Prf* 80
3.3.3 Peculiar haplotype structure at *Rin4* 81
3.4 Evolution of resistance genes beyond species boundaries: adaptive introgression vs. ancestral polymorphism 85
3.4.1 Phylogenetic relationships between species 85
3.4.2 Within species diversity 90
3.4.3 Haplotype structure at the *Pto* locus 97
3.4.4 Across species diversity 99
3.4.5 Divergence to outgroups 103
3.4.6 Between species diversity 105

CHAPTER 4: DISCUSSION 125
4.1 Complex evolutionary history at the *Rcr3* gene family 125
4.1.1 *Rcr3* forms a young gene family 125
4.1.2 Two sequence types are maintained at the *Rcr3* locus 126
4.2 *Rcr3* – a multi-tasking disease resistance gene? 128
4.2.1 Phenotypic basis of balanced polymorphism is not due to interaction with Avr2 128

4.2.2 Interspecific functional diversification at the *Rcr3* locus is likely, but the balanced polymorphism is not based on recognition of differential pathogen effectors 129

4.2.3 The balanced polymorphism is based on differences in activation of the defence response 132

4.2.4 Does *Rcr3* evolution follow a Birth-and-Death process? 135

4.3 Contrasting evolutionary forces act on the *Pto*-mediated disease resistance pathway 136

4.4 Repeated adaptive introgression or ancestral polymorphism: signatures of balancing selection at resistance genes in two closely related wild tomato species 141

4.4.1 Information on the species' history 143

4.4.2 Genetic diversity reflects the habitat range 144

4.4.3 The reference dataset is affected by demography and purifying selection 145

4.4.4 Resistant and susceptible alleles are maintained at *Pto* by balancing selection 146

4.4.5 Evidence for adaptive introgression at the *Pto* locus 148

4.4.6 Strong purifying selection at the *Rin4* gene 150

4.4.7 Signature of balancing selection at *Pfi* and high divergence between species 152

4.4.8 Possible counter-selection of gene-flow at *Pfi* 153

4.4.9 Comparison to the population-based study 154

4.5 Conclusion and outlook: There is more in immunity beyond the *R* gene 157

BIBLIOGRAPHY 161

APPENDIX A: MATERIALS AND METHODS 179

APPENDIX B: RESULTS 181

CURRICULUM VITAE 187

LIST OF TABLES

Table 1.1: Overview of the different scenarios of resistance gene evolution 31

Table 2.1. Reference loci used in this study and their predicted gene products 46
Table 2.2: Geographic information for tomato accessions used in this study 49

Table 3.1: Summary statistics and neutrality test results 61
Table 3.2: Summary of all phenotypic results of the different Rcr3 constructs 75
Table 3.3: Summary statistics measured at the five resistance genes 78
Table 3.4: Results of coalescent simulations at the *Pto* and *Pfi* locus 79
Table 3.5: Results of the McDonald-Kreitman test on *Pfi* 79
Table 3.6. Overview of summary statistics for all genes in all three species 96
Table 3.7: Overview of summary statistics for all genes in the combined dataset 102
Table 3.8: Divergence to the outgroup *S. ochranthum* 104
Table 3.9: Divergence between trans-specifically shared alleles 107
Table 3.10: Summary of the JSFS between *S. peruvianum* and *S. chilense* 117

Table 4.1: Summary of the results obtained with the species wide sample approach 141

Table A1: List of primers and annealing temperatures used in the *Rcr3* project 179
Table A2: List of primers and annealing temperatures used to amplify *Pfi* and *Rin4* 180

LIST OF FIGURES

Figure 1.1: Schematic overview of the two layers in plant innate immunity 12
Figure 1.2: Schematic overview of the genes involved in Pto-mediated signalling 22
Figure 1.3: Schematic overview of the Avr2-Rcr3-Cf-2 interaction 24

Figure 2.1. Schematic summary of the idea behind the Long Range PCR approach 35
Figure 2.2. Map with geographic distribution of all individuals analyzed in this study 50

Figure 3.1: One of 1,000 most parsimonious trees of the 3'flanking region of the *Rcr3* gene 57
Figure 3.2: Divergence between *Locus A* and *Locus B* to *S. lycopersicum Rcr3* 58
Figure 3.3: One of 1,000 most parsimonious gene trees of all assigned *Rcr3* alleles 59
Figure 3.4: Distribution of neutral expectation of Tajima's *D* 62
Figure 3.5: Tajima's *D* across the two *Rcr3* loci 62
Figure 3.6: Frequency spectrum of derived shared and private polymorphisms at the two *Rcr3* loci 63
Figure 3.7: RT-PCR with Rcr3 constructs failing to accumulate in *N. benthamiana* AFs 66
Figure 3.8: Association of SNPs along the *Rcr3* locus with inhibition by Avr2 67
Figure 3.9: Association of SNPs along the *Rcr3* locus with inhibition by Epic1 68
Figure 3.10: Structural model of the Rcr3 protease domain 69
Figure 3.11: Association of SNPs along the *Rcr3* locus with inhibition by Epic2B 70
Figure 3.12: Association of SNPs along the *Rcr3* locus with inhibition by Rip1 71
Figure 3.13: Phenotypic evaluation of a subset of Rcr3 alleles 72
Figure 3.14: Association of the weak HR phenotype with sequence polymorphism 73
Figure 3.15: Structural models of the Rcr3 protease domain 74
Figure 3.16: Gene tree of most tested *Rcr3* alleles 75
Figure 3.17: Comparison of average nucleotide diversity π at the five resistance genes 79
Figure 3.18: Sliding window of silent and nonsynonymous nucleotide diversity π 80
Figure 3.19: Sliding window of silent and nonsynonymous nucleotide diversity π 81
Figure 3.20: Distribution of non-singleton polymorphisms at the *Rin4* gene 82
Figure 3.21: Pattern of linkage disequilibrium among *Rin4* alleles 83
Figure 3.22: Maximum parsimony tree of *Rin4* alleles from *S. peruvianum* 84
Figure 3.23: Maximum parsimony gene tree of all studied individuals at the reference loci 86
Figure 3.24: Maximum parsimony gene tree of all studied individuals at the *Pto* locus 87
Figure 3.25: Maximum parsimony gene tree of all studied individuals at the *Rin4* locus 88
Figure 3.26: Maximum parsimony gene tree of all studied individuals at the *Pfi* locus 89
Figure 3.27: Ratios of π_a to π_s at all reference and candidate genes 91
Figure 3.28: Tajima's *D* values at all reference and candidate genes 94

Figure 3.29: Site frequency spectra of the three resistance genes 95
Figure 3.30: Protein haplotypes at the Pto locus 97
Figure 3.31: Linkage Disequilibrium across the *Pto* gene 98
Figure 3.32: Structural model of the Pto protein 99
Figure 3.33: π_a/π_s ratios 100
Figure 3.34: Tajima's *D* values at all and synonymous sites at the reference and resistance genes 101
Figure 3.35: Site frequency spectrum of mutations of the three resistance genes 101
Figure 3.36: Mismatch distributions of the reference loci 106
Figure 3.37: Mismatch distributions at the *Pto* gene 108
Figure 3.38: Maximum parsimony protein tree of all studied individuals at the *Pto* gene 109
Figure 3.39: Mismatch distributions at the *Rin4* gene 110
Figure 3.40: Mismatch distributions at the *Pfi* gene 112
Figure 3.41: Maximum parsimony protein tree of all studied individuals at the *Pfi* gene 113
Figure 3.42: Mismatch distributions between *S. peruvianum* and *S. chilense* at nonsyn sites 114
Figure 3.43: Frequency spectra of polymorphisms, which are private to *S. peruvianum* 118
Figure 3.44: Frequency spectra of shared polymorphisms in *S. peruvianum* 119
Figure 3.45: Frequency spectra of polymorphisms, which are private to *S. chilense* 120
Figure 3.46: Frequency spectra of shared polymorphisms in *S. chilense* 121
Figure 3.47: Heatmap of the JSFS between *S. peruvianum* and *S. chilense* at the reference loci 122
Figure 3.48: Heatmap of the JSFS between *S. peruvianum* and *S. chilense* at the *Pto* gene 122
Figure 3.49: Heatmap of the JSFS between *S. peruvianum* and *S. chilense* at the *Rin4* gene 123
Figure 3.50: Heatmap of the JSFS between *S. peruvianum* and *S. chilense* at the *Pfi* gene 123
Figure 3.51: F_{ST} values measured at all reference loci and resistance genes 124

Figure 4.1: Scheme of the proposed scenario of *Rcr3* gene family evolution 127
Figure 4.2: Phylogeny of the tomato clade showing differential behaviour of Rcr3 variants 132
Figure 4.3: Most likely scenarios explaining the evolutionary history of the *Pto* locus 150
Figure 4.4: Possible scenario explaining the evolutionary history at the *Pfi* locus 154

Figure B1: ABC estimates of parameters for Model 2 with gene conversion (for *Rcr3* ORFs) 181
Figure B2: Posterior distributions of the parameters of the demographic model (Tarapaca) 181
Figure B3: Western Blot of all isolated AFs containing the expressed Rcr3 constructs 182
Figure B4: Protease activity profiling of all Rcr3 constructs 183
Figure B5: Protein haplotypes of different Rcr3 constructs 185
Figure B6: Inhibition assays with Epic1, Epic2B and Rip1 of a subset of Rcr3 constructs 185
Figure B7: Sequence polymorphisms of all assigned *Rcr3* alleles 186

ABBREVIATIONS

ABPP	activity-based protein profiling
AF	apoplastic fluid
Avr	avirulence factor
bHLH	basic Helix-Loop-Helix
bp	base pair
CC	coiled-coiled
cv.	cultivar
ETI	effector-triggered susceptibility
ETS	effector-triggered immunity
FLR	flanking region
HR	hypersensitive reaction
IptG	Isopropyl β-D-1-thiogalactopyranosidase
JSFS	joint site frequency spectrum
kb	kilo base pair
kDa	kilo Dalton
LB	lysogeny broth
LD	linkage disequilibrium
LPS	lipopolysaccharide
LRR	leucine rich repeat
MAMP	microbe associated molecular pattern
MAP	mitogen-activated protein
MES	2-(*N*-morpholino)ethanesulfonic acid
MHC	major histocompatibility complex
NBS	nucleotide binding site
NLS	nucleus localization signal
nt	nucleotide
OD	optical density
ORF	open reading frame
PAMP	pathogen associated molecular pattern
PCR	polymerase chain reaction
PD	Protease domain
pro	Pro-domain
PRR	pattern recognition receptor
Pst	*Pseudomonas syringae* pv. *tomato*
PTI	PAMP-triggered immunity
pv.	pathovar
R	resistance-
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SFS	site frequency spectrum
SNP	single nucleotide polymorphism

CHAPTER 1: INTRODUCTION

Pathogens have a negative impact on the fitness of their host, and are responsible for drastic epidemics in human, animal or plant populations. Understanding the architecture of the immune system and how specific immunity evolves is of key importance to improving pathogen resistance in various crop species (Gust *et al.* 2010; Lacombe *et al.* 2010) as well as human health (Lambrechts 2010). In plants, the molecular perception of pathogens and activation of defence are well understood and provide an ideal means to study coevolutionary processes between host and pathogen (reviewed in Hammond-Kosack & Jones 1997; Jones & Dangl 2006; Nishimura & Dangl 2010). Previous studies on the evolution of plant immunity focused mainly on the evolution of pathogen recognition (*e.g.* Bakker *et al.* 2006b; Rose *et al.* 2004; Stahl *et al.* 1999), while little is known about evolution of other components in the plant immune system (but see Bakker *et al.* 2008; Caldwell & Michelmore 2009). This work combines evolutionary and functional molecular biology to investigate the evolutionary history of complex interactions between plant immune genes and attempts to explain the molecular causes of physiological variation in immunity.

1.1 The plant immune system

Several hundreds of prokaryotic, eukaryotic and viral species are known to be pathogenic (Madigan & Martinko 2005). Some of these are generalist pathogens and are able to infect large groups of potential host organisms, while others are specialized on only few particular host systems (Schulze-Lefert & Panstruga 2011). In combination, every living organism - no matter if plant or animal - can be a potential host for a particular set of pathogens. In spite of this constant threat of pathogen infection, disease is rather the exceptional state during an organism's life cycle (Nürnberger *et al.* 2004). The reason for this is that most living organisms express effective defences against pathogens to enhance their probability of survival and reproduction. These defences can simply be protective layers like the cuticle in plants or chitin in insects, which serve as constitutive barriers against pathogen penetration, mechanical protections such as cilial mucus movement (*e.g.* in the mammalian respiratory tract) or constant secretion of antimicrobial enzymes (*e.g.* lysozyme) as for example found in the mammalian lacrimal fluid (Murphy *et al.* 2008). However, if these first barriers are overcome by the pathogen, other defences come into play. These usually distinguish between self and non-self and are induced after recognition of non-self molecules. The vertebrate immune system, which efficiently combines innate and adaptive components, is commonly well known among these defences (Murphy *et al.* 2008; Nürnberger *et al.* 2004). In vertebrates, the innate component serves as a first line of defence. General molecular patterns, which are common to groups of pathogens (or microbes in general), so-called PAMPs (pathogen associated molecular patterns) or MAMPs (microbe associated molecular patterns) are recognized by pattern recognition receptors (PRRs), which then recruit defensive agents to the site of infection. The recognized patterns are alien to the host, usually common to a large group of pathogens (or microbes in general), indispensable for the pathogen and predicted to be evolutionary conserved (Medzhitov & Janeway 1997). Examples include bacterial flagellin or lipopolysaccharide (LPS) or fungal chitin (Aderem & Ulevitch 2000; Baureithel *et al.* 1994; Felix *et al.* 1999). Therefore, PRR-mediated defence works unspecifically against large groups of pathogens and confers broad-spectrum resistance. When pathogens have the means to overcome this unspecific defence and to proliferate in the host, the adaptive component of the immune system is induced. Here, recognition happens specifically through cells (*e.g.* B- and T-lymphocytes), which are adapted to particular pathogens and can even adapt further during the course of infection through processes like somatic recombination (reviewed in Schatz & Ji 2011). Since immune cells can take advantage of different circulatory systems (the blood system or the lymphatic system) in the vertebrate body, a rapid and directed proceeding against the pathogen is ensured (Murphy *et al.* 2008; Nürnberger *et al.* 2004).

Plants employ effective immunity as well, which differs substantially from the vertebrate immune system due to some physiological and physical restrictions. For example, the adaptive component found in the vertebrate immune system is absent in plants. Additionally, since plants do not possess a circulatory system and each cell is an autonomous unit, mobile defence response does not occur like in vertebrates. Nevertheless, it is common knowledge that most plants are resistant to most pathogens (Nürnberger *et al.* 2004). There are several reasons for this. Because of their rigid cell wall, the usually thick cuticle and occasional support by wax layers or antibiotic enzymes on the surface, plants are naturally equipped with a first set of constitutive barriers against pathogen invasion. The vast majority of potential pathogens are not capable of penetrating these. Those, which can overcome these barriers, immediately encounter the first inducible layer of plant innate immunity (Figure 1.1).

This first layer presents molecular and functional similarities to the innate immunity in many animals (vertebrates and invertebrates) and together with the mechanical barriers forms the co-called non-host resistance (against non-adapted pathogens) (Nürnberger & Lipka 2005; Thordal-Christensen 2003). Similarly as for animals, a large number of different PAMPs is recognized by a large set of particular, membrane associated PRRs (Boller & Felix 2009, Figure 1.1). Recognition of these patterns subsequently leads to a basal defence response (PAMP-triggered immunity, PTI), which includes recruitment of antimicrobial compounds (*e.g.* proteases, chitinases, phytoalexins) to the site of infection, burst of reactive oxygen species, ethylene biosynthesis, salicylic acid accumulation and callose deposition in the cell wall close to the infection site (Alvarez 2000; Dangl & Jones 2001; Heath 2000; Scheel 1998). All these processes stop invasion or proliferation of biotrophic pathogens. This first layer of innate immunity in plants exhibits striking similarities to the innate immunity in vertebrates or invertebrates (Ausubel 2005; Nürnberger *et al.* 2004). For example, PRRs in both systems recognize similar sets of PAMPs (Zipfel & Felix 2005) and these receptors exhibit functional and structural similarities as they are both composed of extracellular leucin-rich repeat (LRR) domains involved in recognition and intracellular kinase domains transducing the signal (Ausubel 2005; Nürnberger & Brunner 2002; Nürnberger *et al.* 2004). Furthermore, downstream signalling linking pathogen recognition and defence response employs MAP kinase cascades in both systems (Barton & Medzhitov 2003; Pitzschke *et al.* 2009). All these similarities suggested the idea that the innate immune system in animals and the first layer of innate immunity in plants are ancient and were already present in the eukaryotic common ancestor of both lineages. However, comparative genomic studies could not confirm sequence conservation between plant and animal PRRs and it is therefore likely, that the two systems do not share a common ancestry, but rather are the result of convergent evolution employing the same set of functional modules (Ausubel 2005; Nürnberger *et al.* 2004).

Many pathogens have evolved strategies to overcome this unspecific surveillance system. These strategies can involve evasion of recognition through altered PAMP structures (camouflage, Felix *et al.* 1999) or the delivery of toxins into the host cell (Melotto *et al.* 2006). A third strategy comprises the so-called effector-triggered susceptibility (ETS) of the host. Here, pathogens deliver specific effector molecules (sometimes also termed virulence factors) into the plant cell to overcome the resistance of the host and increase their survival probability. Most commonly, their function is to perturb signalling cascades or subsequent immune reactions in the plant (Boller & He 2009; Zhou & Chai 2008). However, it cannot be excluded that some of them function to support proliferation of the pathogen, mobilize host resources for pathogen nutrition or manipulate the development of the host (Dodds & Rathjen 2010). A well studied example is the pathogenic bacterium *Pseudomonas syringae*, which uses its type III secretion system to inject effectors into the host cell, which then directly interfere with PRR function (Goehre *et al.* 2008; Shan *et al.* 2008) or PRR signalling (Li *et al.* 2005; Zhang *et al.* 2007). It is believed that host plants have in turn evolved a second layer of the immune system to counteract these modifications by the pathogen, the effector-triggered immunity (ETI) (Jones & Dangl 2006). This second layer includes the recognition of the pathogen secreted effectors via resistance (*R*) genes (Chisholm *et al.* 2006; Jones & Dangl 2006, Figure 1.1). Specific *R* genes thereby recognize pathogen strain-specific effectors and not a large set of molecules showing a common pattern like PRRs. *R* genes typically encode intracellular proteins containing a nucleotide binding site (NBS) and a LRR domain. Recognition of the corresponding effector leads to activation of the R protein, which then induces the defence response through signalling cascades. The defence response can include measures such as the release of reactive oxygen species or recruitment of antimicrobial metabolites and most often, it results in a so-called hypersensitive response (HR), which involves localized cell death and stops pathogen proliferation (Dangl & Jones 2001; Hammond-Kosack & Jones 1997; Nürnberger *et al.* 2004). Alleles of pathogen virulence factors (or effectors), which are recognized by their corresponding R protein cannot promote infection in the plant and are therefore termed avirulence factors (Avr). Since it is now known that these (a)virulence alleles fulfil a certain modification in the host, the modern nomenclature mainly refers to them as effectors. To avoid confusion by using different names, I will refer to these molecules as effectors hereafter.

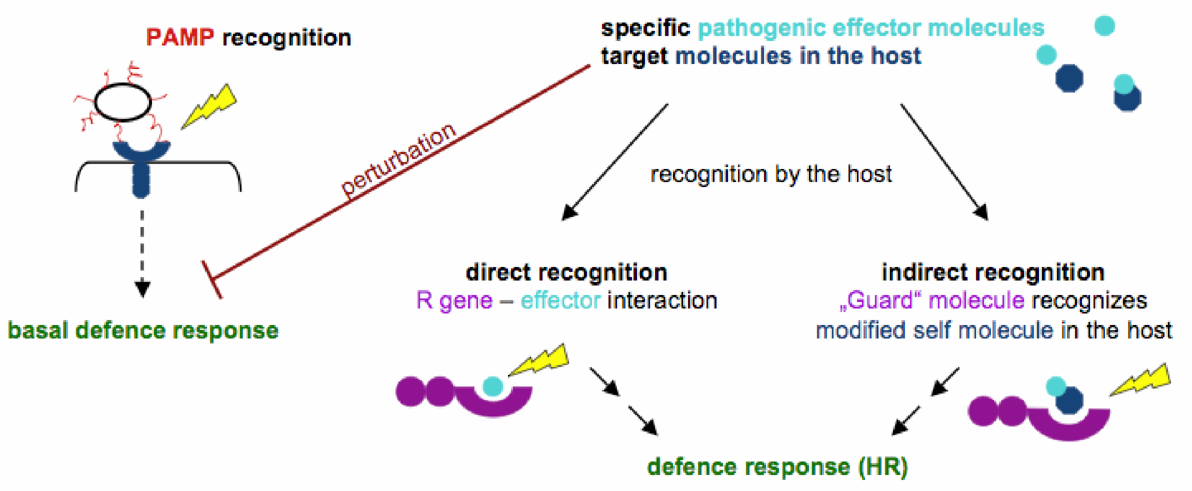

PAMP recognition
specific pathogenic effector molecules
target molecules in the host
perturbation
recognition by the host
basal defence response
direct recognition
R gene – effector interaction
indirect recognition
„Guard“ molecule recognizes
modified self molecule in the host
defence response (HR)

Figure 1.1: Schematic overview of the two layers in plant innate immunity.

1.2 Two modes of pathogen recognition

According to the ‘Gene-for-gene-Hypothesis’ (Flor 1956), the outcome of the interaction between host plant and pathogen is based on the allelic combination of *R* gene and effector. Only the resistant R gene can recognize the matching and therefore avirulent effector, while pathogens expressing ‘virulent’ effectors can infect all host genotypes. Thereby, R genes can recognize pathogen effectors either directly through physical interaction or indirectly through its activity in the host. According to the direct interaction model, a particular resistance gene product will physically bind the corresponding effector and recognition will thus be triggered through direct interaction between the molecules. There are two well-described examples for this type of interaction: the flax L-locus and the corresponding AvrL and AvrM proteins of the fungus *Melampsora lini* (Dodds *et al.* 2006) and the Pi-ta gene in rice, which directly binds and recognizes the corresponding effector AvrPita from the fungus *Magnaporthe grisea* (Jia *et al.* 2000).

Since evidence for direct interaction between effector and R protein from protein-protein interaction studies is sparse, a second model of R gene recognition and activation was proposed to explain the preponderance of indirect recognition: the ‘Guard-Hypothesis’ (Dangl & Jones 2001; van der Biezen & Jones 1998) is a mere extension and not contradiction to the Gene-for-gene-Model and the phenotypic outcome of both models is identical (Brown & Tellier 2011). In the Guard-Model, pathogen effectors target and modify a particular molecule in the host to increase the pathogen’s virulence. Specific modification of a host molecule by the effector creates a so-called ‘modified self’ target in the host. The R protein is then able to recognize this modified self molecule and activate resistance. Thus, the R protein guards the state of the target molecule and recognizes the pathogen by detecting its activity in the host rather than by direct interaction. The R protein accordingly adopts the role of the ‘guard’, while the target molecule has the role of the ‘guardee’. The Rps5-Rps1-AvrPphB and the Rpm1/Rps2-Rin4-AvrB/AvrRpm1/AvrRpt2 system from *Arabidopsis thaliana* and its pathogen *Pseudomonas syringae* are well-known examples for this kind of interaction (Coaker *et al.* 2005; Kim *et al.* 2005b; Shao *et al.* 2003). In support of this hypothesis, most plant species seem to acquire specific resistance to a large number of different pathogens, although the number of NBS-LRR proteins they encode is proportionally small (Jones & Dangl 2006; Nishimura & Dangl 2010). It has been shown that the effector-guardee-guard interaction can be multi-dimensional. Many pathogen effectors can modify more than one target molecule in the host (Gimenez-Ibanez *et al.* 2009; Kaschani *et al.* 2010; Song *et al.* 2009; Tian *et al.* 2007), many target molecules can be modified by several effectors (Coaker *et al.* 2005; Kaschani *et al.* 2010; Mackey *et al.* 2002; Rooney *et al.* 2005; Song *et al.* 2009; Tian *et al.* 2007) and there are examples for guardees, which are guarded by more than one R protein (Day *et al.* 2005; Kim *et al.* 2005b; Mackey *et al.* 2002). The combination of different guardees with different guards could contribute to the breadth of potentially recognized effectors in the plant cell and would explain the proportionally small number of *R* genes (Nishimura & Dangl 2010).

Recently, two extensions to the Guard-Hypothesis have been proposed. The first one attempts to resolve the potentially occurring evolutionary conflict at the guardee. In the presence of the pathogen carrying the corresponding effector(s), selective pressure would be expected to drive the guardee to avoid modification by the effector(s) when no guard is present. However, if the corresponding guard molecule is present, selection would enhance interaction between effector and guardee to improve pathogen recognition. Both selective pressures could then additionally interfere with the actual function of the molecule in the cell. It has been proposed that this evolutionary conflict can be resolved by a mimic protein of the target molecule (‘Decoy-Hypothesis’ by van der Hoorn & Kamoun (2008). This mimic could either evolve independently or through a duplication of the target molecule. Since its actual function in the cell would be redundant with that of the real pathogen target and selective constraints would be relaxed, the mimic would thus simply participate in effector recognition. According to this model, the effector target would only serve as target molecule without particular additional function. Evidence for the Decoy-Model remains to be provided (Dodds & Rathjen 2010; Song *et al.* 2009).

The second extension of the Guard-Hypothesis, the ‘Bait-and-switch-Model’, simplifies the interaction between R protein and effector (Collier & Moffett 2009). According to this model, the effector interacts with the bait, which is its target molecule. Subsequently, the R protein may recognize the effector molecule. Here, recognition of the effector is not based on modification of the host target. The step involving the host target simply facilitates the interaction between R protein and effector.

Based on the observation that a number of resistance genes are involved in hybrid necrosis between closely related plant species (Bomblies *et al.* 2007; Jeuken *et al.* 2009), an interesting implication of the Guard-Hypothesis has recently been suggested. Coevolution between proteins in plant immune complexes ensures efficient signalling to activate the defence response at the appropriate time point, *i.e.* after pathogen recognition, but also inhibition of defence signalling in absence of a pathogen. A mismatch between proteins in an immune complex, for instance when one of the components is introduced into a different, not coevolved genomic background through introgression, can potentially cause aberrant activation of the immune system. As a consequence, harmful constitutive autoimmune reactions can then contribute to the maintenance of postzygotic hybridization barriers and could therefore play a role in Dobzhansky-Muller incompatibility between plant species (Bomblies & Weigel 2007). Particularly, fine-tuned interactions – as must be the case in guard-guardee interactions – should favourably be involved in such incompatibilities (Bomblies *et al.* 2007; Bomblies & Weigel 2007; Ispolatov & Doebeli 2009).

1.3 Host-pathogen coevolution

Infectious diseases decrease the fitness of the host and impose thus strong selective pressure for traits conferring resistance. In turn, natural selection on the pathogen favours traits, which enhance its reproductive success and as a consequence its virulence. Hosts and pathogens therefore find themselves in a recurrent struggle for increased resistance in hosts involving improved recognition of pathogen molecules and evasion of recognition by pathogens through loss or mutation of these molecules (Chisholm *et al.* 2006; Dangl & Jones 2001; Dawkins & Krebs 1979; Schmid-Hempel 2008). These coevolutionary dynamics accelerate the evolutionary rate at involved loci and can have different outcomes for both host and pathogen (Michelmore & Meyers 1998; Paterson *et al.* 2010). In plants, this coevolutionary struggle is most likely centred on the specific recognition of pathogens. Host resistance induces the emergence of a virulent effector, while a virulent effector in turn causes the evolution of an *R* gene with matching recognition specificity. Theoretical models predict that the evolution of effectors and *R* genes is predominantly driven by negative frequency-dependent selection – a special mode of balancing selection (Holub 2001; Tellier & Brown 2007b). According to these models, two main coevolutionary scenarios are possible:

In the classical arms-race scenario, indirect frequency-dependent selection of effector and *R* gene alleles is strong (Frank 1992; Woolhouse *et al.* 2002). In other words, the frequency of a particular effector allele depends on the frequency of the corresponding *R* allele and vice-versa. This indirect dependence allows new, beneficial alleles in both host and pathogen to rapidly rise in frequency and completely replace older variants at these loci. These dynamics result in a series of recurrent selective sweeps and usually lead to low intraspecific variation and high interspecific divergence at these loci (Holub 2001). Since the replacement process happens rapidly, alleles are usually young. Up to date, only weak evidence for this type of plant-pathogen coevolution was found in nature (Bakker *et al.* 2006b; Bergelson *et al.* 2001b).

Inconsistently with this scenario, however, high allelic diversity is typically found at genes controlling host-pathogen coevolution in plants as well as in animals such as the major histocompatibility complex (MHC) in vertebrates (Apanius *et al.* 1997), the *Rpm1* gene in *Arabidopsis* (Stahl *et al.* 1999), the *Rpp13* gene in *Arabidopsis* (Rose *et al.* 2004), the *Pto* gene in tomato (Rose *et al.* 2005; Rose *et al.* 2007), the *Rps2* gene in *Arabidopsis* (Axtell & Staskawicz 2003; Mackey *et al.* 2003; Mauricio *et al.* 2003) or effector genes in plant pathogens (Stukenbrock & McDonald 2009). Furthermore, polymorphism at these loci is ancestral and even predates speciation in many cases (Gyllensten & Erlich 1989; Lawlor *et al.* 1988; Mayer *et al.* 1988; Stahl *et al.* 1999). These observations cannot be explained by the arms-race scenario. It has been proposed that fitness costs may account for the observed polymorphism (Leonard 1977). For instance, if resistance exhibits considerable levels of fitness cost, then alleles conferring resistance would be beneficial if the corresponding pathogen is present and deleterious in absence of the pathogen (Bergelson *et al.* 2001b). The same interaction can be expected for the pathogen if virulence is costly. Cost of resistance or virulence as an additional factor can therefore help decrease the strength of indirect frequency-dependent selection and favour the maintenance of resistant/virulent and susceptible/avirulent alleles (or presence/absence genotypes) in the population. Up to date, numerous studies investigated costs in host-pathogen interactions. Some studies could experimentally demonstrate such costs (Tian *et al.* 2003), while others could not (Bergelson & Purrington 1996; Brown 2003). It is therefore unlikely that costs are the only causative agents of the observed diversity at loci involved in host-pathogen coevolution (Tellier & Brown 2007b).

The scenario of indirect frequency-dependence simply describes the interaction between allele frequencies and neglects other factors, which can potentially interfere with the coevolutionary dynamics. However, in natural populations, these dynamics are likely perturbed by a variety of ecological and epidemiological factors such as polycyclicity of the pathogen, spatial or temporal population structure or seed banks in the host (Tellier & Brown 2007b; Tellier & Brown 2009). These factors can cause direct frequency-dependence of host and pathogen alleles; that is the frequency of alleles is auto self-limiting. Rare host or pathogen alleles therefore have a selective advantage, since the counterpart in the host or pathogen could not yet adapt to this allele. As soon as a beneficial allele rises in frequency and becomes too common in the population its selective advantage will decrease and become negatively selected until it is rare again. These dynamics can maintain considerable allelic diversity at coevolving host and pathogen loci over long time periods (Bergelson *et al.* 2001b; Tellier & Brown 2007b; Thompson & Burdon 1992). In general, these dynamics are referred to as the 'Trench-warfare-Hypothesis' (Stahl *et al.* 1999).

Coevolution between hosts and pathogens not only contributes to the maintenance of allelic diversity within species, but can also shape the genome architecture of the interacting species via gene duplication, transposition and deletion of disease resistance and virulence genes (Michelmore & Meyers 1998; Raffaele *et al.* 2010). The 'Birth-and-Death-Hypothesis' proposed by (Michelmore & Meyers 1998) describes the evolutionary dynamics of resistance gene family evolution. According to this hypothesis, evolution of *R* genes is largely characterized by duplication events. As a consequence, functional redundancy can lead to rapid functional divergence of duplicates resulting in the birth of new recognition specificities or pseudogenization (death) of duplicates (Hammond-Kosack & Jones 1997; Parniske *et al.* 1997).

1.4 Pathway evolution

Most proteins do not operate in isolation, but as components of complex pathways or networks. The 'connectivity' of a protein (*i.e.* the number of interactions of a protein with the other components of a pathway or network) may determine the level of constraint and hence the rate of molecular evolution. Indeed, in yeast the connectivity of proteins in the network is correlated with their rate of evolution (Costanzo *et al.* 2010). Similarly, the position in the pathway or network can affect the evolutionary constraint on the protein. For example, downstream proteins that serve as convergence points of a diverse group of signalling upstream molecules may be subject to greater evolutionary constraint than the upstream molecules (Alvarez-Ponce *et al.* 2009). This can be viewed in terms of the extent of pleiotropic effects amino acid substitutions may have in proteins which serve as convergence points for different signalling molecules. Mathematically, it has been shown that highly pleiotropic genes ought to show much reduced molecular variation (Waxman & Peck 1998). Another type of constraint arises due to the effect of linkage between genes. Genetic linkage as well as physical interaction may affect the evolution of associated loci. Theoretical studies have shown that tight linkage greatly facilitates compensatory evolution (*e.g.* Phillips 1996), meaning that linked genes evolve in a correlated fashion. Finally, the level of constraint may be affected by the degree of redundancy of genes in a pathway, which – in turn – may depend on whether the proteins are encoded by single copy genes or by duplicate genes with overlapping functions (Costanzo *et al.* 2010; Wagner 2001).

In plants, many defence response signalling pathways or networks are known and several genes involved in these networks have been characterized and cloned (Katagiri & Tsuda 2010; Oh & Martin 2011). Several case studies have been performed to reveal selective constraints acting on components of these networks, which interact with pathogen ligands (*e.g.* Rose *et al.* 2007). However, little is known about the evolution of 'downstream' components. Population genetic analyses of genes involved in different parts of signalling networks will reveal their evolutionary history and the selective pressures acting on different positions in the network. In this thesis, sequence variation at genes operating at different points in a pathway controlling disease resistance in wild tomatoes - the Pto-Prf pathway - is described. Such case studies complement analyses of large protein databases, because in case studies, the forces underlying evolutionary constraints can be analyzed in much greater detail and can potentially capture evolutionary constraints over different evolutionary timescales (*e.g.* Wagner 2000; Wagner 2001).

1.5 The *Pto* signalling pathway

The interaction between *Solanum* section *Lycopersicon* (wild tomatoes) and the bacterial pathogen, *Pseudomonas syringae* pv. *tomato* (*Pst*), is ideal for evolutionary studies because both the pathogen ligands and resistance genes have been extensively characterized at the molecular level (Figure 1.2, reviewed in Bogdanove 2002; Sessa & Martin 2000). Furthermore, this is one of the few plant-pathogen interactions in which it has been demonstrated that resistant plants possess receptors for specific pathogen ligand molecules and that these molecules must physically interact for the plant to activate the disease resistance response. One of these receptors is the *Pto* gene, which confers resistance to strains of *Pst* expressing *AvrPto* and was introgressed into the cultivated species, *S. lycopersicum*, from the sister species *S. pimpinellifolium* (Martin *et al.* 1993; Pilowsky & Zutra 1982). *Pto* is a small gene; the open reading frame (ORF) consists of 963-966 nucleotides, has no introns and encodes a functional serine-threonine kinase (Loh & Martin 1995; Martin *et al.* 1993). Protein kinases are well-studied, integral components of many cellular signalling pathways. Currently, three different models describe the activation of the Pto-mediated defence response. All three models assume that Pto activation involves Pto binding to the pathogen ligand, AvrPto or AvrPtoB, in the plant cell. 1) Pto then becomes activated, possibly by a change in protein conformation induced through ligand binding. The activated Pto protein then transduces the pathogen ligand signal, which is dependent on kinase activity and functional downstream genes (Bogdanove 2002; Rathjen *et al.* 1999; Sessa & Martin 2000). The observation that activation of the Pto induced defence response depends on the presence of another molecule, the LRR-protein Prf, contradicts this model. 2) Following the Guard-Hypothesis this dependence is included and it is assumed that Pto is guarded by Prf. Pto is seen as target protein of AvrPto or AvrPtoB and the modification of Pto involves interaction with these effectors. Conformational changes, which are caused by this interaction, finally lead to activation of Prf and the downstream signalling components. This model has not been validated yet, since so far no role for Pto rendering it an effector target (*e.g.* a role in the basal defence) – a prediction of the Guard-Hypothesis – has been shown. 3) Pto acts as a decoy for other kinase receptors, which are involved in the basal defence. Therefore, Pto evolved to distract the effector AvrPto from their real targets and to activate the Prf-dependent defence pathway. According to this model, the decoy molecule does not have an initial function in the cell and only acts as target mimic. The fact that Pto possesses a functional kinase domain contradicts this prediction.

The downstream response after activation of this pathway includes in any case the synthesis of anti-microbial compounds and results in localized cell death at the site of infection. The sequence variation of *Pto* and the functional consequences of this variation within and between populations of seven *Solanum* species have previously been investigated (Rose *et al.* 2005; Rose *et al.* 2007). There is evidence for elevated levels of amino acid polymorphism at this *R* gene consistent with balancing selection at this locus. Furthermore, resistant as well as susceptible alleles seem to be maintained in the population (Rose *et al.* 2005).

The *Pto* resistance gene belongs to a small multigene family of five to six family members in *Solanum* sp. (Martin *et al.* 1993). One of the paralogs, *Fen*, is a functional kinase and confers sensitivity to the insecticide fenthion (Chang *et al.* 2002; Martin *et al.* 1994). This paralog can also recognize and activate defence responses to versions of the *Pst* effector molecule, AvrPtoB lacking E3 ligase activity (Rosebrock *et al.* 2007). However, wild type forms of AvrPtoB ubiquitinate Fen which leads to its degradation in plant cells. One possible scenario posits that ancestral forms of AvrPtoB (or possibly related molecules from other pathogens) lacked the E3 ligase domain and thus were recognized by Fen alleles (Rosebrock *et al.* 2007). Acquisition of the E3 ligase domain and concomitant ability to ubiquitinate Fen was advantageous to the pathogen because it nullified recognition of AvrPtoB by Fen, allowing the pathogen to go undetected in plants expressing the *Fen* gene. In contrast to Fen, Pto is not sensitive to ubiquitination by AvrPtoB, and instead actively phosphorylates AvrPtoB (Ntoukakis *et al.* 2009). Phosphorylation by Pto inactivates AvrPtoB and leads to the activation of the defence signalling pathway in response to pathogens expressing AvrPtoB. Phylogenetic analyses indicate that *Fen* is much older than the *Pto* gene, fitting with sequential bouts of adaptation and counteradaptation between members of this gene family and pathogen effector molecules (Riely & Martin 2001).

Pto and Fen require a second protein, Prf, for activating defence responses. *Prf* is located in the same cluster as the *Pto* gene family, although it is phylogenetically unrelated to *Pto* and its paralogs. The coding region of the *Prf* gene is nearly 6 kb long and encodes a 209.7 kDa protein with regions showing homology to nucleotide-binding sites (NBS), coiled-coil domains (CC) and leucine-rich repeats (LRRs). Recently, it was demonstrated that both of these two kinases, Pto and Fen, physically interact with the same N terminal portion of Prf (Mucyn *et al.* 2006; Ntoukakis *et al.* 2009). Silencing of Prf prevents signalling by Fen or Pto, indicating that Prf acts epistatically to Fen and Pto. Since Pto and Fen, but not Prf, bind pathogen ligands such as AvrPto, or intact or modified versions of AvrPtoB, it is likely that Pto and Fen are located at the proximal portion of this signalling pathway, with Prf being one of the first proteins involved in downstream signalling.

Another gene in the pathway is Prf-interactor or *Pfi*. This gene (originally named Prf-interactor 30137) was cloned via yeast-two-hybrid analyses with different portions of the Prf protein as a bait (Tai 2004). The gene encodes a cytoplasmatic protein, 740 amino acids long and was identified in a screen with a portion of the coiled-coil region and NBS of Prf. There is only little knowledge about its function in the cell or in the Pto/Prf pathway. However, functional testing of this gene indicated that overexpression in tomato suppresses the hypersensitive response (HR), while viral induced gene silencing of *Pfi* showed no phenotypic response (Tai 2004). As such, this gene appears to be a negative regulator of the hypersensitive response. Controls using other elicitors of HR, including the constitutively active form of LeMEK2 (involved in elicitin recognition) or the pathogen proteins AvrRpm1, AvrB, AvrRpt2, and elicitin, indicated that the observed HR suppression was specific to the Pto pathway. Bioinformatic analyses of the Pfi protein structure (http://smart.embl-heidelberg.de) revealed three regions of interest: a domain with high similarity to the class of basic Helix-Loop-Helix (bHLH) type transcription factors, a putative nucleus localization signal (NLS) and a region that shows homology to hydrolases. Recent studies show that some proteins that are involved in pathogen recognition can also migrate to the nucleus and contribute to either activation or suppression of defence related gene transcription (Deslandes *et al.* 2003). Therefore, it might be possible that Pfi functions to directly activate or suppress genes needed for resistance.

The final gene in this study is *Rin4*. *Rin4* was originally identified in *Arabidopsis* and plays a role in several different *R* gene signalling pathways (Axtell & Staskawicz 2003; Kim *et al.* 2005b; Mackey *et al.* 2002). In *Arabidopsis thaliana*, Rin4 is a membrane associated, acetylated protein with a length of 211 amino acids. Its primary function in the cell is not fully understood, but it has been shown that it is involved in the cytokinin pathway of the plant (Igari *et al.* 2008). It also exhibits a negative regulatory activity during the basal defence response of the plant immune system (Kim *et al.* 2005a; Kim *et al.* 2005b; Mackey *et al.* 2002). Due to these two functions, Rin4 is a perfect target for pathogen effectors. In fact, *Arabidopsis* Rin4 is targeted and modified by various effectors including AvrB, AvrRpt2, AvrRpm1. Additionally, there are a handful of endogenous resistance proteins in *A. thaliana* including Rpm1 and Rps2 that are able to recognize modifications of Rin4 and activate defence signalling pathways. Thus, *Arabidopsis* Rin4 is an outstanding example for the Guard-Hypothesis. Identification and functional studies of the *Rin4* gene in tomato indicate that it also functions in the Pto/Prf pathway. Here, it is degraded by an endogenous protease in the presence of different pathogen effector molecules including AvrPto, AvrPtoB, and AvrRpt2 (Luo *et al.* 2009). Rin4 interacts with both Pto and AvrPto in yeast-two-hybrid assays and the degradation of Rin4 in the presence of AvrPto depends on Pto and Prf (Luo *et al.* 2009). Since Rin4 is believed to be a negative regulator of the basal defence, its degradation is predicted to activate these defences. In this way, Rin4 may also play a role in the Pto-Prf signalling pathway, possibly enhancing the resistance response through its specific degradation in the presence of AvrPto.

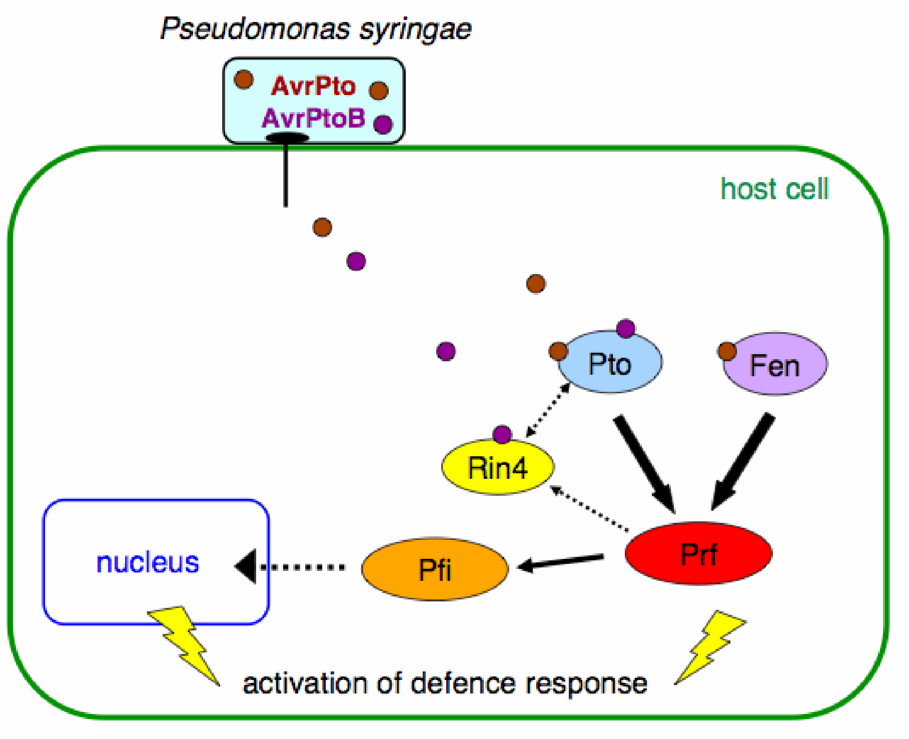


Figure 1.2: Schematic overview of the genes involved in Pto-mediated signalling, which are investigated in this study. The pathogen *P. syringae* secretes the effectors AvrPto and AvrPtoB into the host cell. The effectors can interact with Pto, Fen or Rin4. Pto- and Fen-downstream signalling is transmitted through Prf. Subsequently, the signal might be transferred via Pfi either directly or through other unknown molecules to the nucleus where the defence response is activated. Solid arrows indicate protein interactions with proven function in the network. Dotted arrows indicate putative interactions.

1.6 *Rcr3* – an example for the Guard-Hypothesis

The tomato R gene *Cf-2* confers resistance to the leaf mold pathogen *Cladosporium fulvum* by recognizing the fungal effector Avr2 (Figure 1.3, Luderer *et al.* 2002). The recognition of Avr2 by Cf-2 is dependent on the plant papain-like cystein endoprotease Rcr3 (required for *C. fulvum* resistance 3). This molecule consists of 344 amino acids and one portion of the protein – the protease domain – is secreted into the plant apoplast. Here, it fulfils a putative function in the basal defence and is targeted by the protease inhibitor Avr2 (Rooney *et al.* 2005; van Esse *et al.* 2008). The Avr2-Rcr3 interaction causes conformational changes of the Rcr3 protease, which lead to inhibition of its protease function. These conformational changes or the inhibition of Rcr3 function are putatively detected by the Cf-2 R protein and lead to the activation of the Cf-2 mediated defence response (Krüger *et al.* 2002), which typically involves HR. The Cf-2 protein therefore recognizes the presence of the pathogen indirectly by monitoring the status of Rcr3. Thus, the Cf-2-Rcr3 interaction is an example for the Guard-hypothesis. Additionally, the Cf-2-Rcr3 system is also an example of coadaptation between components in the same signalling pathway. Pairs of Rcr3 and Cf-2 originating from closely related, but different *Solanum* species are incompatible and cause an autonecrotic response (Krüger *et al.* 2002). This incompatibility is caused by only six amino acid substitutions and one amino acid deletion differentiating the Rcr3 molecules between species. The Rcr3-Cf-2 interaction seems to be extremely fine-tuned and only coevolved partners are able to interact effectively.

The Rcr3 molecule is the target of additional effectors, namely Epic1, Epic2B and Rip1 (Song *et al.* 2009, R.A.L. van der Hoorn personal communication). Similarly to Avr2, these three effectors function as protease inhibitors as well. Epic1 and Epic2B are upregulated and secreted by the oomycete *Phytophthora infestans* during infection of tomato plants (Tian *et al.* 2007). They strongly bind and inhibit the tomato papain-like cysteine protease Pip1, which is closely related to Rcr3 and located in the same genomic region. Additionally, both effectors have recently been shown to weakly inhibit Rcr3 as well (Song *et al.* 2009). It could not be demonstrated that Rcr3 inhibition by Epic1 or Epic2B causes Cf-2 dependent HR in tomato plants. However, plants without functional Rcr3 protein are more susceptible to *P. infestans* strains expressing Epic1 and Epic2B compared to plants with functional Rcr3 protein (Song *et al.* 2009). Therefore, Rcr3 seems to play a role in resistance to the oomycete *P. infestans* – possibly in a Cf-2 independent fashion. Rip1 is an effector secreted by the bacterium *P. syringae* and has an inhibitory effect on protease function of Rcr3 (R.A.L. van der Hoorn personal communication). It remains to be demonstrated whether this interaction has an effect on infection of tomato plants by *P. syringae* or in turn on resistance of tomato plants to this pathogen.

In the genome of the cultivated tomato *S. lycopersicum*, the *Rcr3* gene is located on chromosome 2 near the centromer within a cluster of five papain-like cysteine endoproteases (Tian *et al.* 2007). Previous studies revealed that *Rcr3* itself forms a gene family within the wild tomato species *S. peruvianum* and *S. corneliomulleri* (Hörger 2007). This gene family seems to be very young, exhibits copy number variation and consists of paralogs, which are more closely related to one another than to these other proteases. Intriguingly, the Rcr3 interacting partner, Cf-2, also forms a gene family (Dixon *et al.* 1996). Paralogs of Cf-2 are diverse, show copy number variation and belong to different size classes, which are defined by differences in LRR number (Caicedo 2008; Caicedo & Schaal 2004).

Studies that investigated sequence variation and the evolutionary history of this gene family revealed a weak pattern of balancing selection at the 5' end of the gene and presence/absence polymorphism at the *Cf*-2 locus. In the wild tomato species *S. pimpinellifolium*, *Cf-2* variation follows a latitudinal cline (Caicedo 2008). This pattern may simply be due to the species' demography, but the action of natural selection (perhaps through the interacting pathogen *C. fulvum* as selective agent) cannot be excluded (Caicedo 2008). Note, that there has been no study investigating the evolution of the *Rcr3* gene so far. However, Shabab *et al.* (2008) analyzed sequence diversity at the *Rcr3* locus between different tomato species. They revealed high interspecific genetic diversity throughout the tomato clade at this locus and could demonstrate that some of the observed variation has effects on the interaction with the effector Avr2.

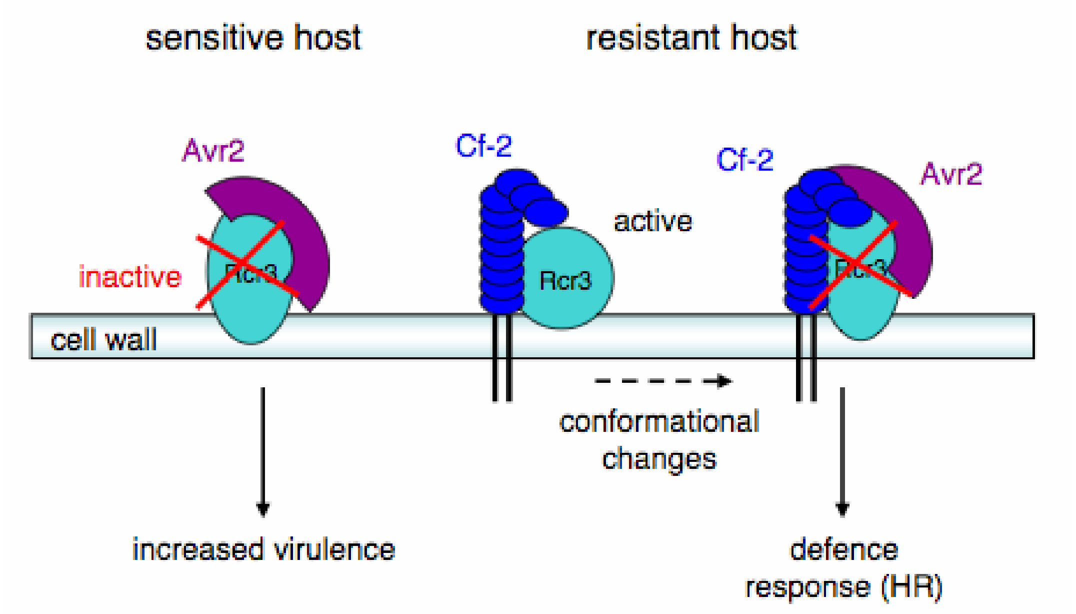


Figure 1.3: Schematic overview of the Avr2-Rcr3-Cf-2 interaction. The pathogen *C. fulvum* secretes the effector Avr2 into the tomato apoplast. This protease inhibitor targets the plant protease Rcr3, which is thereby inactivated. In resistant hosts, the LRR protein Cf-2 gets activated upon putative conformational changes in Rcr3, which are caused by Avr2-binding.

1.7 Tomato as a model system

Tomatoes (*Solanum* section *Lycopersicon*) form a monophyletic clade within the *Solanaceae* family. The section *Lycopersicon* includes a total of 13 species representing all described wild tomato species and the cultivated tomato *S. lycopersicum*, which all diverged within the last 6 million years (Peralta *et al.* 2008; Rodriguez *et al.* 2009). Tomato species show a variety of different mating systems ranging from selfing to obligate outcrossing. Their native geographical distributions range from Ecuador to northern Chile and stretch along the Andean mountains and the pacific coast. Two species are endemic to the Galapagos Islands (Peralta *et al.* 2008).

This geographical distribution covers a wide range of diverse habitats including temperate deserts, seasonable highlands in the Andes or tropical rainforests in the Amazon basin (Young *et al.* 2002). Wild tomato species therefore occupy diverse habitats, which are characterized by substantially varying environmental conditions such as temperature, altitude, soil composition and annual precipitation. Each species displays a characteristic geographical distribution pattern, which is defined by its habitat preference (Nakazato *et al.* 2010; Rick 1973). The different species harbour distinctive morphological characteristics and many of these morphological traits may have evolved as adaptation to local environmental conditions each species encounters. Examples are *S. chilense*, which is restricted to very arid areas and develops extremely deep roots or *S. cheesmaniae*, which is distributed along the coastline of the Galapagos Islands and has developed high salt tolerance (Rick 1973). Depending on the habitat distribution, there is adaptation to abiotic, but also to biotic stress (like pathogen attack). For example, the adaptation of *S. lycopersicum* var. *cerasiforme* to high humidity does not only involve high tolerance for water-logging, but also increased resistance to fungal pathogens, which are favoured in humid conditions (Rick 1973). All these observations suggest that adaptation to environmental conditions play an important role in the evolution of this clade (Bloom *et al.* 2004; Nakazato *et al.* 2008; Rick 1973; Rick *et al.* 1976). Hence, wild tomatoes are suitable model organisms to study adaptation to biotic and abiotic stress. Additionally, the clade of tomato species has many useful attributes for population genetic studies. *S. lycopersicum* is one of the model plant species and detailed genetic maps exist for this and several other species in this genus (Chen & Foolad 1999; Haanstra *et al.* 1999; Monforte & Tanksley 2000). In addition, extensive well-documented collections from natural populations exist for all tomato species (http://tgrc.ucdavis.edu).

Numerous studies have investigated through functional and population genetic approaches how adaptation to environmental factors has shaped the species range of wild tomatoes. These include studies on drought tolerance (Bloom *et al.* 2004; Fischer *et al.* 2011; Xia *et al.* 2010) or pathogen resistance (Caicedo 2008; Caicedo & Schaal 2004; Legnani *et al.* 1996; Rose *et al.* 2005; Rose *et al.* 2007).

The focal species in this thesis are *S. peruvianum*, *S. corneliomulleri* and *S. chilense*. These three species are obligately outcrossing, closely related sister species. *S. corneliomulleri* is a recently proposed species and was previously imbedded in the species *S. peruvianum* (Peralta *et al.* 2005; Peralta *et al.* 2008). *S. corneliomulleri* shares its habitat largely with *S. peruvianum* (Nakazato *et al.* 2010; Zuriaga *et al.* 2009) and is distinguished from this species only by altitude. Previous phylogenetic studies using 19 COSII markers and morphological comparisons proposed the distinction between these two species (Peralta *et al.* 2008; Rodriguez *et al.* 2009). However, these studies present the potential caveat of being based on genetic material from single individuals from both species. Zuriaga *et al.* (2009) sampled 20 *S. corneliomulleri* and 28 *S. peruvianum* individuals and performed phylogenetic analyses using a combination of AFLP markers and two nuclear genes. This study revealed the existence of considerable amount of variation within both clades and could not distinguish between *S. peruvianum* (*sensu strictu*) and *S. corneliomulleri*. It remains to be demonstrated on the molecular level whether these two species are indeed distinct or merely phenotypic variants of the same species.

The phylogenetic relationship between *S. peruvianum* and *S. chilense* on the other hand is resolved. They form distinct species and diverged approximately 0.55 million years ago (Städler *et al.* 2008). These species differ morphologically, but display partly overlapping habitat ranges in the arid coastal regions of southern Peru and northern Chile (Rick 1986; Rick & Lamm 1955). However, evidence for hybridization between the two species is controversial. Crossing experiments revealed strong intrinsic postzygotic incompatibilities between the two species (Rick 1986; Rick & Lamm 1955), while population genetic studies found evidence for postdivergence gene-flow (Städler *et al.* 2008). The two species differ substantially in their habitat preference and diversity. Within the tomato clade, *S. peruvianum* exhibits the highest level of morphological and genetic diversity and has the largest habitat range. Adaptation to biotic factors seems to play an important role in evolution of this species and is more pronounced than adaptation to abiotic stress (Fischer *et al.* 2011; Nakazato *et al.* 2010; Rose *et al.* 2005; Rose *et al.* 2007; Xia *et al.* 2010). In contrast, the habitat distribution of *S. chilense* is restricted to arid regions (Chetelat *et al.* 2009; Nakazato *et al.* 2008; Nakazato *et al.* 2010; Rick & Lamm 1955) and numerous studies demonstrated adaptation to dry climate (Fischer *et al.* 2011; Nakazato *et al.* 2010; Rick 1973; Xia *et al.* 2010).

In this thesis, these three tomato species are studied in context of their coevolution with mainly two different pathogens: the bacterium *P. syringae* and the fungus *C. fulvum*. These two systems are very suitable for population genetic and functional studies, since both are well-characterized on the molecular level (reviewed in Bogdanove 2002; Oh & Martin 2011; Sessa & Martin 2000; Wulff *et al.* 2009). The focal genes involved in the tomato-*P. syringae* interaction all function in the Pto-mediated disease resistance. The *Pto* gene was the first *R* gene to be cloned and since then more than 25 other interacting genes were discovered and extensively described in tomato (Oh & Martin 2011). *C. fulvum* is a host specific pathogen in the tomato clade (Bond 1938; Caicedo 2008). Therefore, this pathosystem is well suited to investigate the evolution of specific immunity. Furthermore, the focal gene involved in the tomato – *C. fulvum* interaction, the *Rcr3* gene, is one of the best studied examples for the Guard-Hypothesis.

1.8 The use of different sampling schemes

All evolutionary processes whether stochastic (mutation, drift) or deterministic (selection) leave signatures in the genome. Therefore, patterns of genetic variation in addition to the structure of this variation (*e.g.* the frequency of polymorphisms in a given population or species) can help us understand the evolutionary history of single genes or even whole genomes. It has to be noted that demography affects patterns of polymorphism over the full genome, while selection acts only on particular genomic regions. Comparing the evolutionary history of genes of interest to the genomic background (*e.g.* a set of reference loci) can therefore help to disentangle signatures of natural selection, from demographic artefacts (Pavlidis *et al.* 2008).

One approach, which can be applied, is the coalescent theory (Kingman 1982). The coalescent is the genealogy of a given sequence sample describing the relationship between the lineages backwards in time. A coalescent event occurs at the time point when two lineages find their common ancestor. The shape of a coalescent tree and the distribution of mutations depend on the evolutionary history of the sample and can be influenced by different evolutionary forces including selection or demography. Long external branches, for instance, can be observed in cases of population expansion or purifying selection (Charlesworth *et al.* 1993; Tajima 1989). Like most other species, wild tomatoes exist as spatially structured populations (metapopulations) with many subpopulations (or demes), which are linked by migration and subjected to extinction/recolonization (Wakeley & Aliacar 2001).

The coalescent tree of substructured populations is composed of two different phases (Pannell 2003; Wakeley & Aliacar 2001). The scattering phase describes the short time scale of coalescent events within the single demes until the common ancestor of the deme is found including weak migration between demes. Note that migrants will not coalesce with the common ancestor of their host deme, but share the ancestor with their deme of origin. The second phase of the coalescent in metapopulations, the collecting phase, describes the long time scale until the lineages from all demes coalesce. Both phases together cover thus the evolutionary history of the entire species.

When addressing evolutionary questions concerning a metapopulation, the mode of sampling has to be considered carefully (Städler *et al.* 2009). On the one hand, local adaptation or balancing selection can only be detected, if sufficient numbers of alleles within single demes are sampled (Pannell & Charlesworth 2000). On the other hand, for the investigation of species wide processes, the sampling has to cover the collecting phase well (Städler *et al.* 2009). This can be achieved by sampling one or more alleles from sufficient numbers of demes distributed over the species range. Accordingly, sampling sufficient numbers of alleles from many demes will give a reliable picture of the entire species' evolutionary history.

1.9 Aims of this thesis

In this work, a set of questions all concerning the evolution of specific immunity in plants is addressed. The main focus is to study classes of immunity genes and their interactions, which are extensively analyzed on the molecular level, but have not received much attention by evolutionary biologists yet. These are genes, which are involved in indirect pathogen recognition - the guardees - and genes, which are involved in signal transduction after pathogen perception. Molecules such as *R* genes, which directly encounter the pathogen or pathogen derived molecules and therefore are assumed to be in the centre of host-pathogen coevolution, have always attracted much interest by evolutionary biologists (*e.g.* Bergelson *et al.* 2001b; Caicedo & Schaal 2004; Rose *et al.* 2004; Rose *et al.* 2005; Rose *et al.* 2007; Stahl *et al.* 1999). In contrast, genes, which act in more 'downstream' positions, were mainly expected to maintain conserved functions and to lack interesting evolutionary potential. However, increasing knowledge about the importance of these other components in plant immunity and the strategies pathogens and hosts employ to be ahead in the coevolutionary arms race are challenging this view (Katagiri & Tsuda 2010).

This thesis therefore addresses the question of the role of the guardee in the evolution of the specific plant immune system. The guardee can be seen as a bridge between pathogen recognition and signal transduction and is expected to be subject to contrasting evolutionary forces (Caldwell & Michelmore 2009; van der Hoorn & Kamoun 2008). In this work, these evolutionary forces are investigated using the *Rcr3* gene in wild tomatoes as an example and their impact on the function of the guardee is resolved. Furthermore, mechanisms of genome evolution such as gene duplication or gene conversion, which can potentially have an impact on specificity in immunity, are studied with respect to *Rcr3* evolution. Additionally, complex evolutionary interactions between genes involved in signal transduction following pathogen recognition are studied using the Pto-mediated signalling network as an example. First, the evolutionary history of genes involved in this network is investigated on a short evolutionary scale, *i.e.* on the population level. Second, genes from this network displaying an interesting pattern on the population level are further studied on a long-term evolutionary scale, *i.e.* at the species wide level and beyond species divergence.

Evolution and functional evaluation of *Rcr3*

While natural selection may favour multiple protein variants at *R* genes, pathogens may specifically target conserved proteins in hosts and thus create an evolutionary dilemma for the host plant. It is expected that complex interactions occur between the guardee, its guard and the pathogen effector, and that the guardee is subject to contrasting evolutionary forces (Caldwell & Michelmore 2009; van der Hoorn & Kamoun 2008). For example if pathogen pressure is high, positive selection on the effector-guardee interface could improve either the detection of the effector in presence of the guard, or reduce the damage done by the effector. Alternatively, positive selection on the guardee-guard interface may improve pathogen related activation and/or prevent autoactivation of the defence response (Bomblies & Weigel 2007; Ispolatov & Doebeli 2009). Purifying selection is expected to remove mutations, which have a negative impact on pathogen recognition or activation of the defence response. If pathogen pressure (or the allele frequency of the corresponding effector) varies, balancing selection is expected to act on the guardee-effector or guard-guardee interface. Up to now, relatively few studies have investigated the evolutionary history of host targets serving as guardees (Caldwell & Michelmore 2009; Shabab *et al.* 2008). These studies have revealed that typically such effector targets show high inter- and intraspecific diversity, but the evolutionary and functional basis of this diversity is still unknown and is addressed in this work using the tomato *Rcr3* gene as an example.

Evolution of the *Pto*-pathway

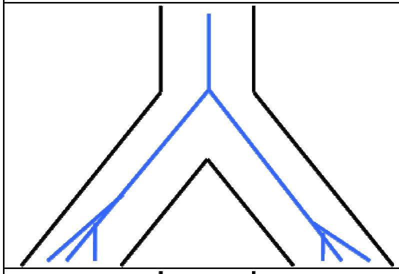
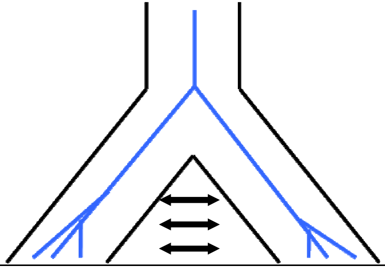
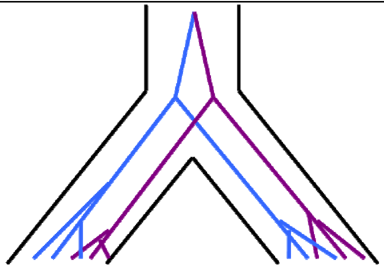
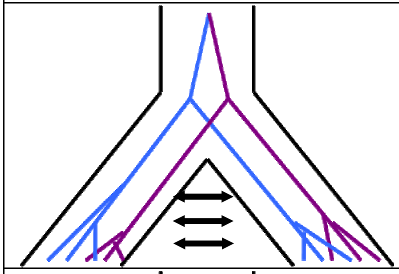
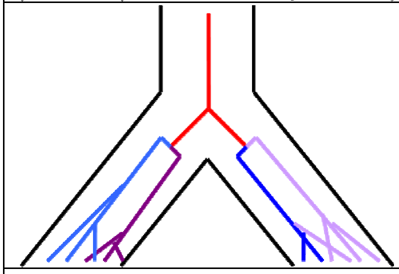
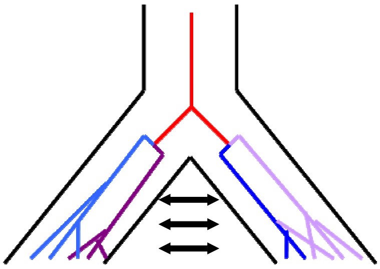
Based on the current understanding, the Pto-mediated signalling network consists of multiple signalling molecules, at least including Pto and Fen, as potential inputs (Figure 1.2). Prf interprets and transduces these signals and this may lead to the activation of the 'downstream' gene *Pfi*. Since *Pfi* has domains that are typical for transcription factors, this protein may move to the nucleus, via its nuclear localization signal, to activate 'downstream' defence genes. Based on this pathway structure, one might expect that mutations in the 'downstream' genes *Prf* and *Pfi* will have the greatest pleiotropic effects and hence these genes may experience the greatest evolutionary constraint. On the other hand, the genes interacting with pathogen ligands, *Pto*, *Fen* and possibly *Rin4*, may show a signature of relaxed constraint because they are less pleiotropic or be subject to adaptive changes because they directly interact with pathogen molecules. In this study, population genetic methods are applied to evaluate the sequence variation and evolutionary history of these five genes to identify the strength and magnitude of natural selection at different points in a defence signalling pathway.

Evolution of resistance genes beyond species boundaries

Host-pathogen coevolution may not be a transient process, but can continue over thousands of generations and even extend beyond divergence of different species. Previous studies report trans-species polymorphism at immune genes in closely related sister species, possibly predating speciation and being maintained in the species ever since divergence. Well-known examples are MHC polymorphisms, which are found in both chimpanzees and humans and were most likely already present in the common ancestor (Gyllensten & Erlich 1989; Lawlor *et al.* 1988; Mayer *et al.* 1988). Coevolutionary dynamics such as direct frequency-dependent selection are likely the cause for these ancient trans-species polymorphisms. Alternatively, adaptive introgression of beneficial alleles, which are involved in coevolutionary processes, may explain trans-species polymorphism, which emerged recently, frequency-dependent selection being the likely condition facilitating adaptive introgression. This has been demonstrated in the case of the selfincompatibility-(S)-locus in *Arabidopsis* species (Castric *et al.* 2008). Polymorphism at this locus is maintained by balancing selection and it has been shown that migration of beneficial alleles between recently diverged species is enhanced compared to the genomic background. Since resistance gene evolution resembles S-locus evolution, this scenario may also apply in that case.

Coevolution most likely results in two different scenarios, the arms-race and the trench-warfare scenario. These propose different predictions concerning the pattern visible on the sequence level at involved genes. The outcome of each scenario also depends on whether selection was present prior to species divergence, whether there is migration between the species or whether selection is present in one or both of the species. Predictions on the outcome of the different scenarios apply to the locus under selection in comparison to the genome average and are reflected by the genetic variation within the species, among the species and the level of gene flow between the species (Table 1.1). In this work, genes that are most likely subject to coevolutionary dynamics as identified in the study at the population level are tested for different evolutionary scenarios extending beyond species divergence: *Pto*, *Pfi* and *Rin4*. Therefore, the level of genetic diversity at the three genes within and among the wild tomato species *S. peruvianum*, *S. chilense* and *S. corneliomulleri* is analyzed in comparison to the genomic background and the amount of trans-species polymorphism at these three loci is assessed.

Table 1.1: Overview of the different scenarios, which can apply to resistance genes in recently diverged species. The first column shows the coalescent tree of each scenario for two recently diverged species. Black lines indicate the demographic history of two species with a recent common ancestor. Coloured lines indicate the evolutionary history of the gene of interest. Arrows indicate postdivergence gene-flow.

scenario	**description of scenario**	**expectation between species**	**expectation within species**
	arms-race in one or both species, can predate speciation or be recent	divergent alleles (neutral and nonsynonymous polymorphisms)	low variation, signature of positive selection
	arms-race in one or both species, can predate speciation or be recent introgression	divergent or shared alleles (neutral and nonsynonymous polymorphisms), higher or lower rate of introgression than genomic average	low variation, signature of positive selection
	balancing selection scenario, predates speciation	trans-species polymorphism (nonsynonymous), private polymorphism (neutral), signature of balancing selection (nonsynonymous polymorphisms)	high variation, signature of balancing selection
	balancing selection scenario, predates speciation, migration	trans-species polymorphism (nonsynonymous and neutral due to introgression), signature of balancing selection	high variation, signature of balancing selection
	balancing selection scenario (occurs after speciation)	high divergence, private nonsynonymous and neutral polymorphisms	high variation, signature of balancing selection
	balancing selection scenario (occurs after speciation), migration	trans-species polymorphism (nonsynonymous and neutral due to introgression), higher introgression rate at locus under selection compared to genomic average, signature of balancing selection	high variation, signature of balancing selection

CHAPTER 2: MATERIALS AND METHODS

2.1. Sequence evolution at the *Rcr3* gene family

The aim of this project was to assess natural sequence variation at the *Rcr3* locus in wild tomato species and to decipher the evolutionary history of this putatively young gene family. To understand the structure of this gene family, an exhaustive sequencing approach was applied to amplify as many paralogous and allelic *Rcr3* sequences possible. Eleven individuals from one population (Tarapaca, LA2744) from the wild tomato species *S. peruvianum* and additional individuals from other tomato species were used.

2.1.1 Plant material and DNA sequencing

The ORF of the *Rcr3* gene was amplified from genomic DNA from multiple individuals of *S. peruvianum* (accession LA 2744 from Tarapaca, Chile), collected by Charles Rick. Seeds from eleven different field collected plants were grown under standard greenhouse conditions in Davis, CA. DNA was isolated using the CTAB method (Doyle & Doyle 1987) from 2 g of leaf tissue collected from each plant. The DNA was resuspended in 300 to 1000 µl TE depending on yield. Alleles from single individuals from seven additional species of *Solanum* were sequenced. These species included: *S. peruvianum* (accessions LA1954, LA3636 and LA0446), *S. chilense* (accessions LA2748, LA 1930 and LA1958), *S. corneliomulleri* (accessions LA1274 and LA1973), *S. pimpinellifolium* (accession LA0400), *S. lycopersicum* (cv. VFNT Cherry and cv. Rio Grande), *S. chmielewskii* (accession LA3653), *S. habrochaites* (accession LA1777), *S. pennellii* (accessions LA0716 and LA 3791). For outgroup comparisons, the *Rcr3* gene from *S. lycopersicoides* (accession LA2951) was sequenced. Plant growth conditions and DNA extraction for these accessions (with exception of LA1954, LA3636, LA0446, LA 2748, LA1930, LA1958, LA 1274 and LA 1973) were identical as for *S. peruvianum* from Tarapaca (LA2744). DNA from these other accessions was extracted using the Dneasy DNA Extraction Kit (Qiagen). The *Rcr3* gene was identified using the *Rcr3* reference sequence from *S. lycopersicum* cv. '*Mogeor*' (GenBank, accession number AF493234). Restriction sites for cloning (see below) were introduced into the primer sequences, which were designed to cover the start and stop codon. The gene was PCR amplified using the Phusion proofreading polymerase (Finnzymes, Espoo, Finland), cloned into Zero Blunt TOPO vectors (Invitrogen, Carlsbad, CA) and sequenced from all individuals. Direct sequencing of PCR products and sequencing of miniprepped plasmid DNA (obtained by the QIAquick Spin Miniprep Kit Protocol, Qiagen) from clones were conducted in parallel (Big Dye Terminator v 1.1, Applied Biosystems). Sequencing was performed according to the Sanger sequencing protocol using the DNA analyzer ABI 3730 (Applied Biosystems & Hitachi). Multiple clones per gene per individual were sequenced and ambiguous positions were compared to the direct sequences from the original PCR products. When necessary, independent rounds of PCRs, cloning and sequencing were conducted to resolve ambiguities. Raw sequences were edited and aligned in Sequencher 4.8 (1991 – 2007 Gene Codes Corporation) and alignments were refined by hand with MacClade (Version 4.0, Maddison and Maddison 2000, Sinauer Associates).

2.1.2 Analysis of *Rcr3* flanking regions

Previous analysis of the *Rcr3* gene from *S. peruvianum* had revealed that it is member of a very young gene family (composed of closely related members) (Hörger 2007). Although as many clones as possible were sequenced per individual, it was not possible to distinguish allelic and paralogous sequences. Paralogs and orthologs may be distinguished by their flanking sequences since allelic sequences originate from the same locus in a genome and should possess the same (or very similar) flanking sequences. Paralogs, which are located at different positions in the genome should have different flanking sequences. To distinguish between paralogs and orthologs, fragments of 500-2000 bp of *Rcr3* flanking DNA (with a minimum of 200 bp overlap with the gene) were amplified, cloned and sequenced from the Tarapaca population of *S. peruvianum.* A three-step Tail-PCR protocol with a set of random and nested *Rcr3* specific primers was used (Table A1, Liu *et al.* 1995). The location of the amplified *Rcr3* flanking regions in the tomato genome was assessed using BLASTn searches (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and phylogenetic reconstruction (PAUP v. 4.0b10, Swofford 1999, Sinauer Associates). Flanking regions were assigned to the different allele sequences obtained for the *Rcr3* gene. Assignment was performed stringently and only unambiguous pairs of alleles and flanking regions were matched. Subsequently, the genomic origin of alleles with matching flanking region was defined according to the preceding BLAST search. Only *Rcr3* alleles, which could be matched unambiguously to a certain flanking region, were used for population genetic analysis.

2.1.3 Long Range PCR

To test whether *Rcr3* paralogs are located in the same genomic region, a Long Range PCR approach was performed. Therefore, different combinations of primers at the 5' and 3' end of the gene pointing towards the intergenic region were used in a PCR reaction with Crimson LongAmp®Taq DNA polymerase. This polymerase allows the amplification of genomic fragments of up to 20 kb length. If *Rcr3* paralogs are located within this distance and in the according orientation from each other, this approach will permit the amplification of the intergenic region between them. The size of the PCR amplicon reflects the distance between these paralogs.

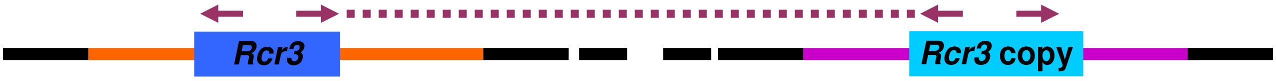


Figure 2.1. Schematic summary of the idea behind the Long Range PCR approach. Primers (in purple) at the 5' and 3' ends of the gene copies, which point towards each other, will permit the amplification of the intergenic region between the two copies whereas primers in the opposite direction will not produce an amplicon.

2.1.4 Population genetic analyses of the *Rcr3* gene family

Summary statistics

The analysis of sequence variation is a convenient method to discover which selective forces act upon genes. The use of standard summary statistics π (average pairwise difference) and θ (population mutation parameter) allows for comparisons of sequence variation among loci. The sequence evolution at *R* genes and their components can be compared to putative neutrally evolving, non-resistance genes to determine if patterns of polymorphism observed in the resistance genes are different than those observed for neutrally evolving loci. Tests of neutrality can be applied to determine whether the observed nucleotide polymorphisms can be explained by neutral evolution or selection. One such test is the McDonald-Kreitman test (McDonald & Kreitman 1991). This test is based on the analysis of several alleles from one species and at least one allele from an outgroup species. It compares the number of 'polymorphisms' (sites that are segregating in one species) to the number of 'fixed differences' (changes that are fixed and monomorphic between species). These polymorphisms and fixed differences are further sub-divided into replacement differences or silent differences. Under neutrality, the ratio of polymorphisms to fixed differences for silent and replacement differences should be the same. An excess of amino acid polymorphism is consistent with a history of balancing or negative frequency-dependent selection acting on the locus. Conversely, an excess of fixed replacement differences indicates strong directional selection. Another test of neutrality that can be used is the Hudson-Kreitman-Aguade test (Hudson *et al.* 1987). It compares the level of intraspecific polymorphism to the level of divergence between species at two or more loci and determines whether polymorphism and divergence are decoupled. In this test, the resistance genes will be compared to a set of non-resistance genes. Tajima's *D* test statistic (Tajima 1989) is based on the fact that under the standard neutral model estimates of θ_W and π are identical. The test measures the skew in the frequency spectrum. A negative *D* value indicates an excess of rare polymorphisms and a positive Tajima's *D* suggests an excess of intermediate frequency polymorphisms. The significance of the test is conservative for testing the departures from neutral equilibrium. For the Fu and Li's *D* test (Fu & Li 1993), θ is estimated by the number of singleton mutations η_e and by the total number of mutations η. Under the infinite-site model, η should equal S (= number of segregating sites) and the number of singleton mutations is expected to reflect the number of mutations occurring on the young, external branches of the underlying genealogy. Directional selection is consistent with an excess of mutations on external branches and causes Fu and Li's *D* to be negative. Fu and Li's *D* will be positive, if most mutations occurred on the old, internal branches. In this case, polymorphisms might be old and have been kept in the population over a long time. While the statistical tests and summary statistics described above provide information about the recent evolution of a gene as a whole, further details of variation and divergence within a gene may be explored using sliding-window analyses. For this analysis, polymorphism and divergence parameters are calculated within defined regions (= window) of the sequence. The window is then shifted across the gene by a defined step size.

The standard summary statistics including π, divergence, Tajima's *D* and Fu and Li's *D* test statistics were calculated at the *Rcr3* genes and the 3' flanking regions using DnaSP v. 5.10 (Librado & Rozas 2009). Phylogenetic analyses were completed using PAUP v. 4.0b10 (Swofford 1999, Sinauer Associates). The phylogenetic relationships between the sequences (ORFs) were determined using maximum parsimony (MP) considering gaps as a fifth state. To rule out demographic effects, which could potentially perturb the pattern of sequence variation at the *Rcr3* locus, all population genetic summary statistics were compared to values obtained from the 14 reference loci described above.

Inference of gene conversion using Approximate Bayesian Computation (ABC)

To test whether gene conversion was present between *Rcr3* copies simulations were performed to simulate the observed data assuming a recent gene duplication event, copy number variation at the duplicated locus and different gene conversion rates using software by K.R. Thornton (Thornton 2007) (program *cnvcoal* available on the K.R. Thornton webpage, http://www.molpopgen.org/software/coalescent.html). An Approximate Bayesian Computation method was developed using ABCest (Excoffier *et al.* 2005) to identify the simulated dataset (and gene conversion rate), which fits the observed dataset best. The use of the ABC algorithm was composed of three steps: simulation of datasets, model choice, and parameter estimation. The simulation step consisted in simulating, for every evolutionary scenario 100,000 datasets with identical features to the *Rcr3* loci. The length of the loci was fixed to 1,100 bp for the coding region. The population intergenic recombination rate was assumed to be equal to the population mutation rate ($\rho = \theta$) following previous observations (Städler *et al.* 2008; Stephan & Langley 1998), and the distance between the two loci was fixed at 9,000 bp (based on the available data from the *S. lycopersicum* genome). The simulated sample sizes matched the observed data: 14 and 9 sequences for *Locus A* and *Locus B* respectively. Pseudogenized alleles were considered in this analysis as they potentially contribute to gene conversion. Note that the population mutation rate ($\theta = 4N\mu$) was chosen to vary uniformly between 8.8 and 9.8 (based on the observed θ_W, identical results were obtained when it varied between 4 and 12). The population rate of gene conversion is defined as $C = 4Nc$ where N is the population effective size, and c the rate of gene conversion per nucleotide per generation (Innan 2003b). Each evolutionary scenario was defined by a set of parameters characterized by uniform prior distributions, from which was sampled to perform coalescent-based simulations, and to compute summary statistics using the GSL C++ library and the libsequence C++ library as implemented in the software *summstats* (Thornton 2003; Thornton 2007).

The model choice procedure was based on a weighted multinomial logistic regression (Fagundes *et al.* 2007) computed on the best 500 simulations (over the 100,000 performed per model) for which δ, the Euclidean distance between the observed summarized dataset and the summarized datasets, is smallest (Beaumont *et al.* 2002). Bayes factors were calculated as the ratio of the posterior probabilities for the tested models (Kass & Raftery 1995). Three models were tested: Model 1 assumes ancestral gene duplication without subsequent gene conversion ($4Nc = 0$). Model 2 assumes ancestral gene duplication with subsequent gene conversion, with two parameters: the mean length of gene conversion track which varies uniformly between 10 and 1,000 bp, and the gene conversion rate ($C = 4Nc$) varying from 0 to 10. Model 3 assumes ancestral gene duplication with subsequent gene conversion, but has only the gene conversion rate $C = 4Nc$ as a parameter. In Model 3, the mean length of the gene conversion track was fixed to 395 bp, the value obtained with *Geneconv* (Sawyer 1989).

The following six summary statistics were chosen with the value indicated for the observed data:

- F_{ST} between the two loci (Hudson, Boos, Kaplan) = 0.03036
- $\pi_{between}$, the mean pairwise difference between the two loci = 10.604
- Number of fixed differences between loci = 0
- Number of shared polymorphisms between loci = 20
- Number of private polymorphisms in *Locus A* = 10
- Number of private polymorphisms in *Locus B* = 8

These statistics are supposed to be good indicators of gene conversion (Thornton 2007), and the addition of four more summary statistics (number of segregating sites per locus and π per locus) yielded identical results (data not shown). The model choice procedure revealed that Model 2 is clearly favoured with a Bayes factor > 1,000 compared to the other two models. This demonstrates that gene conversion is necessary to account for the observed number of fixed differences and shared polymorphisms. Finally, the posterior distributions (mode and 95% credibility intervals) of each parameter of Model 2 were estimated by applying the locally weighted multivariate regression method (Beaumont *et al.* 2002) implemented in the ABCest program (Excoffier *et al.* 2005).

Inference of gene conversion using the site frequency spectrum of shared and private polymorphisms

To test whether intergenic gene conversion occurs between the *Rcr3* ORFs and 3’ flanking regions, the frequency spectrum of polymorphisms shared between the duplicated loci or private to one of them was surveyed (Innan 2003a). Derived polymorphisms were identified using *S. lycopersicum* as outgroup and the number of shared, private and fixed polymorphisms was counted for each mutational class.

Test for natural selection at the *Rcr3* ORFs

To test whether the *Rcr3* locus (ORFs) deviates from neutral expectations a neutral model of evolution was derived for the studied population of *S. peruvianum*. Therefore, the site frequency spectrum of each locus was studied under the assumption of gene conversion occurring between copies. A total of 2,000 coalescent simulations under the best model found above (Model 2 with gene duplication and gene conversion) was simulated drawing parameter values from the 95% probability intervals (for *C* and mean track length of gene conversion). This dataset represents the neutral evolution expectation for each of the two loci under gene conversion in a population with constant size. The expected distribution of Tajima's *D* under neutrality at the *Rcr3* ORFs was estimated and observed values of Tajima's *D* were compared to this distribution.

Test for natural selection at the *Rcr3* 3' flanking regions

To derive the neutral model for the *Rcr3* 3'FLRs data from the 14 reference loci was used. Based on previous studies (Tellier *et al.*, unpublished results), two demographic models for the Tarapaca population were tested: Model 1 with constant population size and Model 2 with a past expansion. Model 1 has for parameter the population mutation rate (θ). Model 2 has three parameters: the present population mutation rate ($\theta = 4N\mu$), the factor of expansion (ratio of past over present population size) and the time of expansion (scaled in $4N$). In both models there is intra-locus recombination with the population recombination rate being fixed equal to the mutation rate ($\rho = \theta$). The 14 loci were simulated as concatenated for a total of 19,053 bp using Hudon's ms. The following observed summary statistics were used: total number of segregating sites (713), the average π_s per nucleotide over 14 loci (0.0231) and the average Tajima's *D* over 14 loci (-0.3425). The model with demographic expansion (model 2) was clearly favoured (Bayes factor > 1,000) based on 200,000 simulated datasets (retaining the best 500). The final aim was to reveal whether the 3'FLR are under natural selection (purifying, positive or balancing). Therefore, 2,000 coalescent simulations were computed under the demographic model with expansion drawing parameter values from the 95% probability intervals for θ, the expansion factor and time of expansion. This simulated dataset represents the neutral evolution expectation under the expansion scenario for each of the two 3'FLR. The expected distribution of Tajima's *D* under neutrality at the *Rcr3* 3'FLRs was estimated and observed values of Tajima's *D* were compared to this distribution.

2.2 Functional consequences of sequence variation at the *Rcr3* locus

The aim of this project was to link the observed variation at the *Rcr3* locus to functional diversity in disease resistance. For this purpose, 54 alleles from multiple individuals in several wild tomato species were cloned into a binary expression vector and transiently expressed in *Nicotiana benthamiana* plants. Afterwards, apoplastic fluids (AFs) containing the expressed Rcr3 alleles were isolated. In vitro assays using activity based protein profiling were performed to determine every allele's ability to interact with different pathogen effector molecules (Avr2, Epic1, Epic2B and Rip1). Afterwards, all recovered Rcr3 alleles were co-infiltrated with Avr2 or buffer into *rcr3* mutant tomato plants to assess their ability to elicit an Avr2 specific defense response. All phenotypic data were associated with the nucleotide sequence.

2.2.1 Cloning procedure and *Agrobacterium*-mediated transient expression

A total of 54 *Rcr3* variants, which had been cloned into TOPO Zero Blunt for sequence analyses, were selected for functional testing. Cloning procedures of these variants were conducted according to the protocol described in Shabab *et al.* (2008). Each *Rcr3* variant to be functionally tested was excised from the Zero Blunt TOPO vector using the restriction enzymes XhoI and NcoI, for which restriction sites resided in the PCR primers. Excised fragments were gel purified using the Gel Extraction Kit (Qiagen) and cloned into the pFK26 vector carrying the 35S overexpression promoter. 35S::Rcr3 cassettes were shuttled into the binary vector pTP05 (Shabab *et al.* 2008) using the restriction enzymes XbaI and SalI. All clones were verified by sequencing and electroporated into *Agrobacterium tumefaciens* strain GV3101. *A. tumefaciens* carrying the binary vectors were grown overnight at 28°C in LB media containing 50 µg/ml Kanamycin and 50 µg/ml Rifampicin. The bacteria were centrifuged (10 min, 3000g) and the pellets resuspended in 10 mM MES pH 5, 10 mM MgCl2, 0.2 µM acetosyringone to a final OD_{600} of 2. Each bacterial culture was mixed at an equal volume with an *Agrobacterium* culture containing a binary vector expressing the silencing inhibitor p19 (Voinnet *et al.* 2003). *Agrobacterium* cultures were infiltrated into leaves of 3-4 week old *Nicotiana benthamiana* plants using 1-ml syringes without needle. Infiltrated *N. benthamiana* leaves were harvested 72 h post inoculation. The apoplastic fluid (AF) of all infiltrated leaves was isolated as described previously (Shabab *et al*. 2008). Leaves were vacuum infiltrated with ice-cold water and dried on filter paper. Surface-dry leaves were centrifuged in a tube with holes in the bottom (10 min., 1600g). The AF was collected in a collection tube, aliquoted and stored at –20°C. Equal volumes of AF were used for all further experiments. Western Blot analysis was used to confirm the expression of Rcr3 using Rcr3 specific antibodies described previously (Rooney *et al.* 2005).

2.2.2 Activity-based protein profiling and inhibition assays

Activity-based protein profiling (ABPP) using fluorescent DCG-04 was used to detect Rcr3 activity in the isolated AFs. 45 µl of AF were labelled with 2 µM fluorescent TMR-DCG-04 at pH 5.5 in the presence of 1 mM DTT for 5 h as described previously (Shabab *et al.* 2008). DCG-04 is an inhibitor of papain-like cysteine proteases and reacts irreversibly and covalently to the active cite cysteine of active proteases (Greenbaum *et al.* 2000). Inhibition studies were performed by pre-incubation with 100 nM affinity-purified Avr2 (Shabab *et al.* 2008), followed by ABPP. Proteins were separated via sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) and fluorescently-labeled proteins were detected by in-gel fluorescence scanning using a Typhoon 8600 scanner (GE Healthcare Life Sciences, http://www.gelifesciences.com) at ex/em 580 nm.

2.2.3 Protein purification

To test whether the selective pressure acting on the *Rcr3* locus is due to the interaction between Rcr3 and other pathogen effector molecules, inhibition assays were performed with several effectors, which are secreted by different pathogen species and are all protease inhibitors with high similarity to Avr2. Two oomycete effectors, Epic1 and Epic2B (secreted by *Phythophthora infestans*), and the bacterial effector Rip1 (secreted by *Pseudomonas syringae*) were tested in addition to the fungal effector Avr2. These proteins were affinity-purified prior to use. Transformed bacterial cultures expressing the effector of interest were incubated at 37°C (*Escherichia coli* expressing Epic1 and Rip1) or 28°C (*E. coli* expressing Epic2B) overnight. At the next day, overexpression of the effector was induced by adding 0.6 mM Isopropyl β-D-1-thiogalactopyranosidase (IptG) and subsequent incubation at 37 or 28°C overnight. The supernatant of the bacterial cultures, which contained most of the expressed protein, was harvested at the following day. Ammonium sulfate powder (380 g per effector) was added to each supernatant and dialysis was performed overnight in dialysis buffer using semipermeable tubings. Affinity purification of the overexpressed proteins was performed using a 50% Resin (Ni-NTA-Agarose) Superflow column (Qiagen). The samples were eluted from the column using different concentrations of Imidazol Buffer and the different concentration phases were collected separately. Quantity and purity of the proteins were checked by protein gel electrophoresis and only pure samples containing the protein of interest were kept at –80°C.

2.2.4 Inhibition assays

Inhibition-assays were performed based on competitive ABPP. The affinity purified effectors Avr2, Epic1, Epic2B and Rip1 were pre-incubated with AFs containing overexpressed *Rcr3*-constructs for 30 min. at a final concentration of 100nM (all effectors) or 1 µM (Epic1, Epic2B and Rip1). For each construct a negative control without effector and a positive control with 40 µM E-64 was added. Subsequently, labelling with fluorescent TMR-DCG-04 was performed as described above. Detection was performed via sodium dodecyl sulphate polyacrylamide gel electrophoresis and fluorescent scanning. Only non-inhibited Rcr3 proteases could be labelled by DCG-04.

2.2.5 HR-assays

I investigated whether the different Rcr3 constructs could activate the hypersensitive response upon exposure to Avr2 in tomato plants (*S. lycopersicum* cv. Money Maker). For this purpose, 100 µl of AF containing equal concentrations of expressed Rcr3 were mixed with Avr2 to a final concentration of 100 nM or PBS (negative control) respectively. These mixtures were infiltrated into Cf2Rcr3-3 and Cf0Rcr3^{esc} tomato leaves using sterile syringes. The following controls were infiltrated onto Cf-2rcr3-3, Cf0Rcr3^{esc} and Cf-2Rcr3^{pim} tomato leaves: Rcr3^{esc} (from *S. lycopersicum*) with/without Avr2, Rcr3^{pim} (from *S. pimpinellifolium*) with/without Avr2, p19 with/without Avr2, only Avr2 and only PBS. Additionally, to investigate the effect of inhibition of Rcr3 by the other three effectors on disease resistance, 100 µl of AFs containing equal concentrations of expressed Rcr3 were infiltrated into tomato leaves with or without 1µM of Epic1, Epic2B or Rip1. The procedure including controls was performed as described for co-infiltration with Avr2. To test, whether Cf-2/Rcr3 dependent HR can be activated in wild tomato plants with known genotype, Avr2 (or PBS) was infiltrated into leaves of wild tomato plants from some accessions used in the preceding functional assays.

Tissue collapse was monitored daily until five days post inoculation (dpi) in all three experiments and recorded photographically.

2.2.6 RT-PCR

Seven Rcr3-constructs failed to be expressed in *N. benthamiana* leaves. To confirm that the construct was designed correctly and that the agroinfiltration process was successful, RNA of infiltrated leaves was isolated and RT-PCR with *Rcr3*-specific primers was performed. The extraction of RNA was conducted using the Rneasy RNA Extraction Kit (Qiagen) starting with 40-80 mg of plant material. cDNA-banks were created by reverse transcription using the SuperScriptTM Reverse Transcriptase (Invitrogen). RT-PCR was performed to amplify the transcribed *Rcr3*-gene and a part of the Ribulose-bisphosphate-carboxylase-oxigenase as a RNA-extraction control. As negative control, RNA from AF without expressed Rcr3 (only p19) was extracted. For details of these PCR-protocols see Table A1.

2.2.7 Association of genotypic and observed phenotypic data

A general linear model algorithm implemented in TASSEL v. 3.0 was used to evaluate correlations between phenotypic variation and sequence polymorphism at the *Rcr3* locus (http://www.maizegenetics.net/). The genotypic data was filtered such that only mutations, which occurred in frequencies greater than 25 %, were included. Resulting *P*-values were Bonferroni-corrected for multiple testing.

2.2.8 Structural model of the Rcr3 protein

The crystal structure of Rcr3 is not yet resolved. Therefore, to create a reliable structural model of this protein, which allows visualizing interesting amino acid changes within the protein structure, a template search was performed using SWISS Model (Arnold *et al.* 2006). The best-fitting template obtained via this search was papain (PDB code 9papA). This template was used to create a putative structural model of Rcr3 using SWISS Model. To visualize the protein structure and to highlight interesting amino acid changes, the program Depp View (SWISS-PdbViewer) was used.

2.3 Evolution of the *Pto*- disease resistance pathway

The aim of this project was to study sequence variation at five genes (*Pto*, *Prf*, *Fen*, *Pfi* and *Rin4*), which are involved in one disease resistance pathway in the wild tomato species *S. peruvianum*. For this purpose, individuals from one population of this species (LA2744), which had been shown to harbour a high proportion of all reported variation in this species (Rose *et al.* 2005; Rose *et al.* 2007), were chosen for analysis. All five genes were amplified, cloned and sequenced from ten individuals in this population. Both alleles were recovered from every individual.

2.3.1 Plant materials and sequencing

Plants of *S. peruvianum* were grown from seed collected from a single, large population in Tarapaca, Chile by Dr. Charles R. Rick. Seeds were stored at the Tomato Genetics Resource Center (TGRC; http://tgrc.ucdavis.edu) until 1996, at which time 10 seeds from different field collected plants were grown under standard greenhouse conditions in Davis, CA. DNA was isolated using the CTAB method (Doyle & Doyle 1987) from 2 g of leaf tissue collected from each plant. The DNA was resuspended in 300 to 1000 µl TE depending on yield. For outgroup comparisons, individuals of the following species were used: *S. hirsutum* from Ancash, Peru (LA 1775), *S. pennellii* from Arequipa, Peru (LA3791) and *S. lycopersicoides* from Tarapaca, Chile (LA 2951). Plant growth conditions and DNA extraction for these outgroups were identical as used for *S. peruvianum*.

PCR amplification, cloning and sequencing strategies differed slightly for each gene. However, the entire coding region of each gene was amplified using a proofreading polymerase, either Pfu polymerase (Stratagene, LaJolla, CA) or Phusion (Finnzymes, Espoo, Finnland). PCR fragments were cloned into pCR-Blunt or Zero Blunt TOPO (Invitrogen, Carlsbad, CA). Direct sequencing of PCR products and sequencing of minipreped plasmid DNA from clones (Big Dye Terminator v 1.1, Applied Biosystems) were conducted in parallel for each gene. Multiple clones per gene per individual were sequenced and ambiguous positions were compared to the direct sequences from the original PCR products. When necessary, independent rounds of PCRs, cloning and sequencing were conducted to resolve ambiguities. Specific amplification and cloning strategy for each gene are described below.

Sequencing of *Pto*

The primers SSP17 and JCP32 were initially used to amplify alleles of *Pto*. These primers also amplify to a lesser degree two paralogs of *Pto*, namely *Pth3* and *Pth5*. Plasmids containing *Pto* were discriminated from the other paralogs by restriction digest. The restriction enzyme BstXI specifically digests alleles of *Pth3* and *Pth5*, but not *Pto*. To circumvent non-specific amplification of *Pto* alleles and to facilitate direct sequencing of *Pto* for confirmation of homozygosity/heterozygosity respectively, two *Pto* specific primers in the upstream region of *Pto* were developed. These primers, FromPth5A and FromPth5B, were used in combination with the JCP32 primer, which anneals at the 3' end of *Pto*.

Sequencing of *Prf*

Prf is a large gene (5587 from start to stop codon), so it was divided into two overlapping halves for PCR and these were sequenced separately. The first half of *Prf* is well-known for being recalcitrant to cloning, so a direct sequencing strategy, combined with allele specific primers to resolve phase, was used on this half. Both direct sequencing of PCR products and cloning were employed to generate the data for the second half of the gene (approximately 58% of the gene). A large number of primers (>90) were designed for sequencing and allele specific amplification.

Sequencing of *Pfi*

Pfi is also a large gene (5428 bp from start to stop codon), so a similar sequencing strategy as used for *Prf* was applied to *Pfi*. The gene was divided into two to three overlapping fragments for PCR and these were sequenced and cloned separately. Primers were designed based upon the GenBank mRNA sequence AY662518 from *S. lycopersicum* cv. Rio Grande 76R and can be found in Table A2 in the appendix.

Sequencing of *Fen*

The primers SSP17 and SSP19 were used initially to amplify alleles of *Fen*. Cloning of these PCR products revealed that these primers did not specifically amplify alleles of *Fen*. Ultimately two additional *Fen*-specific primers were designed, one upstream of *Fen* and one downstream of *Fen*, based upon the GenBank sequence AF220602 of this region from the Rio Grande 76R haplotype. These two intergenic primers, FenFor and FenRev, were used in combination or with SSP19 or SSP17, respectively.

Sequencing of *Rin4*

Rin4 was originally described and cloned from *A. thaliana* (Mackey *et al.* 2002). To identify the putative tomato *Rin4* homolog, BLAST was used to search the tomato BAC database on the SOL Genomics Network website (http://www.sgn.cornell.edu). The gene prediction program GeneMark (http://exon.gatech.edu/GeneMark) was used to predict the open reading frame of the putative tomato *Rin4* gene. Primers were designed based upon the tomato genomic sequence and the incorporated gene prediction information. Two primers (Rin4For3 and Rin4Rev5) were used to amplify the entire coding sequence of *Rin4*. A combination of internal primers was used for sequencing of miniprepped clones.

Reference loci

For comparisons between loci, the sequences of alleles of 14 other loci (*CT066, CT093, CT099, CT114, CT143, CT148, CT166, CT179, CT189, CT198, CT208, CT251, CT268* and *sucr*) were obtained from (Baudry *et al.* 2001) and (Roselius *et al.* 2005) (Table 2.1). These reference genes were amplified from five of the same individuals of this population of *S. peruvianum*. These genes are single-copy cDNA markers previously developed and mapped in (Tanksley *et al.* 1992).

Table 2.1. Reference loci used in this study and their predicted gene products.

Locus	Chromosome	Length [bp]	Putative encoded protein
CT066	10	1346	Arginine decarboxylase
CT093	5	1415	S-adenosylmethionine decarboxylase proenzyme
CT099	12	1354	Copper binding protein
CT114	7	1169	Phospho-glycerate kinase
CT143	9	1821	Sterol C-14 reductase
CT148	8	1497	Copper/zinc superoxide dismutase
CT166	2	2673	Ferredoxin-NADP reductase
CT179	3	995	Tonoplast intrinsic protein Δ-type
CT189	12	1463	40S ribosomal protein S19
CT198	9	779	Submergence induced protein 2-like
CT208	9	1767	Alcohol dehydrogenase, class III
CT251	2	1779	At5g37260-like gene (transcription factor involved in circadian regulation)
CT268	1	1887	Receptor-like protein kinase
Sucr	3	1575	Vacuolar invertase

2.3.2 Population genetic analysis of the pathway

The standard summary statistics including π, Tajima's *D*, Fu and Li's *D*, Fu's *F* test statistics were calculated using DnaSP v. 5.10 (Librado & Rozas 2009). Coalescent simulations were used to examine whether the pattern of substitutions at synonymous and nonsynonymous sites at the resistance genes differed from the 14 other genes from these same individuals. For synonymous sites, the arithmetic mean of π of the 14 non *R* genes was used as the estimate of theta for the simulations. A total of 1,000 simulations were executed in DnaSP and subsequently it was determined whether the value of π observed at the resistance genes fell within the 95% confidence interval of the simulations based on theta estimated from the 14 non *R* genes. For these simulations, no recombination was assumed, the most conservative assumption. The same approach was also used to test if π at nonsynonymous sites (π_a) was different for the resistance genes versus the arithmetic mean across these 14 non *R* genes. McDonald-Kreitman tests and the sliding window analyses were also conducted using DnaSP. Linkage disequilibrium was evaluated using TASSEL v. 2.1 (http://www.maizegenetics.net/).

2.3.3 Phylogenetic inference

To evaluate the genes' phylogenetic relationships within the population sample and between the sample and other *Solanum* species, phylogenies can be reconstructed for each gene. This analysis included sequence data from the *S. peruvianum* population as well as sequence data from other *Solanum* species: *S. chmielewskii* LA3653, *S. lycopersicum* LA3343, *S. lycopersicum* LA1221, *S. hirsutum* LA1777, *S. pennellii* LA0716 and *S. pimpinellifolium* LA0400. Phylogenetic analyses were completed using PAUP v. 4.0b10 (Swofford, 1999). The phylogenetic relationships between these sequences were determined using maximum parsimony (MP) and neighbor-joining (NJ) and these methods yielded similar topologies.

2.4 Evolution of resistance genes beyond species boundaries

The aim of this project was to compare the evolution of three disease resistance genes (*Pto*, *Rin4* and *Pfi*) which exhibited an interesting evolutionary history in the population based study to a set of reference loci in species wild samples of three closely related wild tomato species (*S. peruvianum*, *S. corneliomulleri* and *S. chilense*). For this purpose multiple populations representing the whole species distribution were chosen. One allele per population was selected for sequence analysis. These alleles were recovered through PCR-amplification, cloning and sequencing.

2.4.1 Plant materials and sequencing

The genomic DNA of a total of 38 individuals from three wild tomato species (*Solanum* section *lycopersicum*) was used: 13 from the *S. chilense*, 19 from *S. peruvianum* and 6 from *S. corneliomulleri*). For outgroup comparisons, genomic DNA of *S. ochranthum* (LA 2682*)*, *S. lycopersicoides* (LA 2951) and *S. pennellii* (LA 0716) was used. The plant material was provided by several sources. Most accessions are from the Tomato Genetics Resource Center (TGRC) at the University of California at Davis (http://tgrc.ucdavis.edu). The accession PI 128654 is from the U. S. Department of Agriculture, Agricultural Research Service Plant Genetic Resources Unit in Geneva, New York. The individuals ARE244, NAZ251, CAN262, TAC101, MOQ111 and QUI126 were collected in Peru by Thomas Städler and Tobias Marczewski (Arunyawat *et al.* 2007). Plants were grown under standard greenhouse conditions. Genomic DNA from most plants was extracted from tomato leaves using the DNeasy Plant Mini Kit (Qiagen GmbH, Hilden, Germany). The DNA extraction procedure for the individuals from accessions LA 3355, LA 2744, LA 3218, LA 3636, LA 3666 and PI 128654 followed the CTAB protocol as described in paragraph 2.1 (Doyle & Doyle 1987). All accessions with geographical information are listed in table 2.2 and visualized on the map in Figure 2.2.

Table 2.2: Geographic information for tomato accessions used in this study. Numbers in brackets indicate the individual that was included in the sample. If not noted otherwise, accessions are from the Tomato Genetics Resource Center (TGRC) at the University of California at Davis (http://tgrc.ucdavis.edu).

accession	collection site	location
S. chilense		
LA 1930	Quebrada Calapampa	Province Arequipa, Southern Peru
LA 1958	Pampa de la Clemesi	Province Moquegua, Southern Peru
LA 1960	Rio Osmore	Province Moquegua, Southern Peru
LA 1969	Estique Pampa	Province Tacna, Southern Peru
LA 2748	Soledad	Province Tarapaca, Northern Chile
LA 2750	Mina La Desperciada	Province Antofagasta, Northern Chile
LA 2778	Chapiquina	Province Tarapaca, Northern Chile
LA 2930	Quebrada Taltal	Province Antofagasta, Northern Chile
LA 2932	Quebrada Gatico, Mina Escalera	Province Antofagasta, Northern Chile
LA 3355 (7186)	Cacique de Ara	Province Tacna, Southern Peru
TAC101[a]	Tacna	Southern Peru
MOQ111[a]	Moquegua	Southern Peru
QUI126[a]	Quicacha	Southern Peru
S. peruvianum		
LA 0111	Supe	Province Lima, Central Peru
LA 0153	Culebras	Province Ancash, Peru
LA 0446	Atiquipa	Province Arequipa, Southern Peru
LA 1333	Loma Camana	Province Arequipa, Southern Peru
LA 1336	Atico	Province Arequipa, Southern Peru
LA 1616	La Rinconada	Province Lima, Central Peru
LA 1913	Tinguiayog	Province Ica, Southern Peru
LA 1951	Ocona	Province Arequipa, Southern Peru
LA 1954	Mollendo	Province Arequipa, Southern Peru
LA 2732	Moquella	Province Tarapaca, Northern Chile
LA 2744 (7232)	Sobraya	Province Tarapaca, Northern Chile
LA 2834	Hacienda Asiento	Province Ica, Southern Peru
LA 2964	Quebrada de Burros	Province Tacna, Southern Peru
LA 3218 (7242)	Quebrada Guerrero	Province Arequipa, Southern Peru
LA 4125	Camina	Province Tarapaca, Northern Chile
PI 128654[b]	Azapa Valley	Northern Chile
ARE244[a]	Arequipa	Southern Peru
NAZ251[a]	Nazca	Southern Peru
CAN262[a]	Canta	Central Peru
S. corneliomulleri		
LA 1274	Pacaibamba	Province Lima, Central Peru
LA 1283	Santa Cruz de Laya	Province Lima, Central Peru
LA 1973	Yura	Province Arequipa, Southern Peru
LA 2759	Mamina	Province Tarapaca, Northern Chile
LA 3636 (7257)	Coayllo	Province Lima, Central Peru
LA 3666 (7267)	La Yapa	Province Ica, Southern Peru

[a] These individuals were collected by Thomas Städler and Tobias Marczewski (Arunyawat *et al.* 2007)

[b] This accession was provided by the U. S. Department of Agriculture, Agricultural Research Service Plant Genetic Resources Unit in Geneva, New York

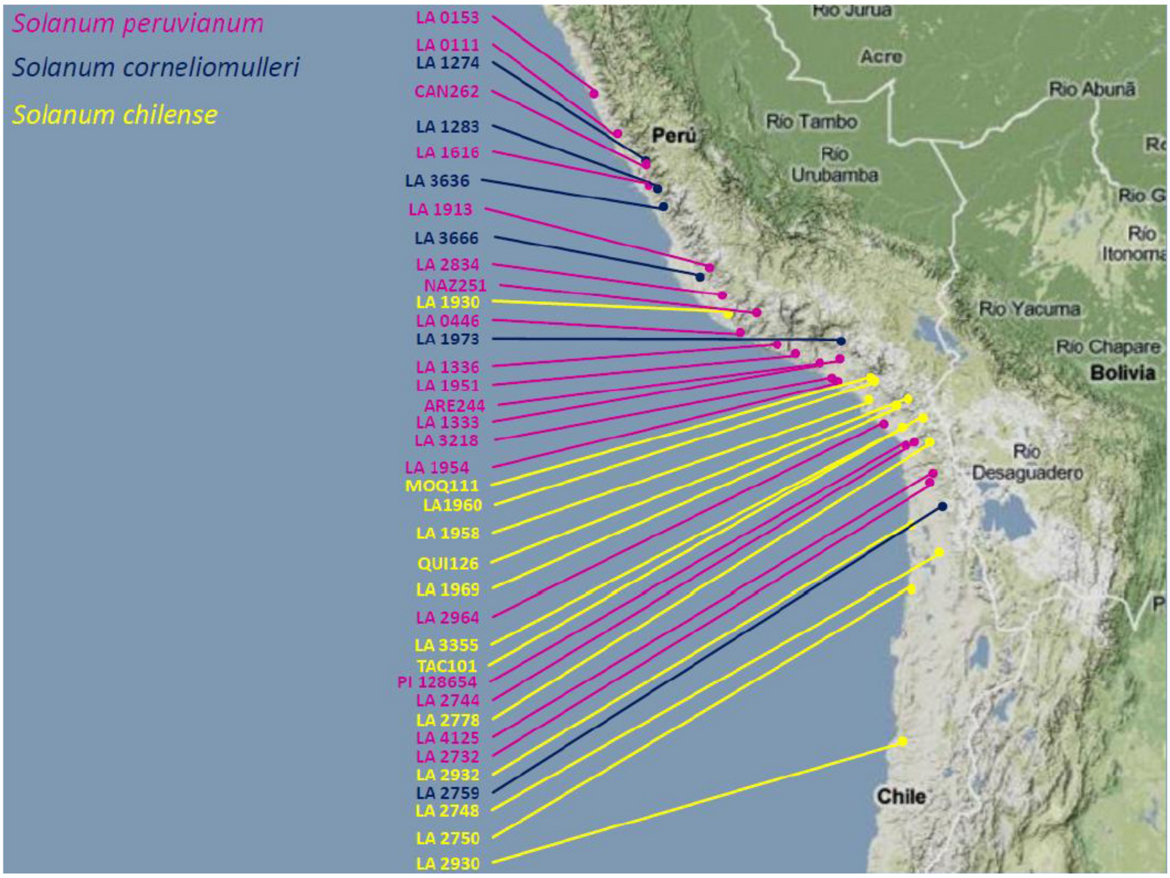


Figure 2.2. Map with geographic distribution of all individuals analyzed in this study. *S. peruvianum* individuals are labelled in pink, *S. chilense* individuals in yellow and *S. corneliomulleri* individuals in blue. This map was adopted from Böndel (2010).

Sequencing *Pto*

In case of the *Pto* gene, both alleles per individual were recovered to create datasets with putatively resistant and suceptible alleles and a random dataset like for the other genes. Altogether five different 5’ primers and two different 3’ primers were used to specifically recover both alleles of every individual, but not other *Pto* homologs. The PCRs were performed in the following order of 5’ – 3’ primer combinations: Pth5B and JCP32, Pth5A and JCP32, PtoFor-99 and JCP32, SSP17 and JCP32, PtoFor+4 and JCP32, Pth5B and JCP6, Pth5A and JCP6. The annealing temperature varied between 54.4 °C and 59.0 °C depending on the primer combination (Böndel 2010). As soon as both alleles of an individual were obtained amplification was stopped and the recovered PCR products were used for all further procedures. All recovered sequences were loaded into a phylogeny of the *Pto* gene family (Rose et al. 2007, unpublished results) to verify them as *Pto* alleles. Details of this procedure can be found in (Böndel 2010).

Sequencing *Pfi*

Since the results from the population based study (see Chapter 3.3) revealed a putative history of balancing selection for only the second half of the *Pfi* gene, only this portion was amplified, cloned and sequenced in the species wide sample. The primers Prfint30137For1200 and Prfint30137Rev2374 were used for amplification of this portion. A combination of internal primers was used for sequencing of miniprepped clones (Table A2). The first allele recovered in at least three clones was chosen for analysis.

Sequencing *Rin4*

Sequencing of the *Rin4* gene was performed as described in paragraph 2.1.2. The first allele recovered in at least three clones was chosen for analysis.

Sequencing of reference loci

Nine unlinked nuclear loci are used in this study: *CT066*, *CT093*, *CT166*, *CT179*, *CT189*, *CT198*, *CT208*, *CT251* and *CT268* (Table 2.1). These loci are single-copy cDNA markers originally mapped by Tanksley *et al.* (1992) (Tanksley *et al.* 1992) in genomic regions with different estimated recombination rates (Stephan & Langley 1998). PCR amplification was performed with High Fidelity Phusion Polymerase (Finnzymes, Espoo, Finland), and all PCR products were examined with 1% agarose gel electrophoresis. Generally, direct sequencing was performed on PCR products to identify homozygotes and obtain their corresponding sequences. For heterozygotes, a dual approach of both cloning before sequencing and direct sequencing was used to obtain the sequences of both alleles. The first allele recovered in at least three clones was chosen. Sequencing reactions were run on an ABI 3730 DNA analyser (Applied Biosystems and HITACHI, Foster City, USA). One allele was sequenced for each individual, and a total of 19 (*S. peruvianum*), 13 (*S. chilense*) and six (*S. corneliomulleri*) sequences were obtained for each locus/species combination. Contigs of each locus were built and edited using the Sequencher program 4.8 (Gene Codes, Ann Arbor, USA) and adjusted manually in MacClade (Version 4.0, Maddison and Maddison 2000, Sinauer Associates).

2.4.2 Population genetic analysis of the species wide sample

Summary statistics

Population genetic summary statistics were obtained as described in paragraph 2.1.3.

The site frequency spectra of all reference loci and the three resistance genes were calculated using the program SITES by Jody Hey (http://genfaculty.rutgers.edu/hey/software). The site frequency spectrum of mutations can be influenced by two evolutionary forces: demography of a species or population and selection. Demography usually affects all genomic regions and loci simultaneously, while selection only acts on particular genes or genomic regions. Comparison of the candidate genes to reference genes can help rule out effects of demography. Since it has been shown that purifying selection dominates at the reference genes, the site frequency spectrum (based on synonymous sites only) was calculated for the reference genes (Tellier *et al.* 2011). Additionally, the expected frequency spectrum under complete neutrality was calculated using the following equation

$$E[S_i] = \frac{\theta_W}{i}.$$

Frequency spectra of all three resistance genes were compared to both the synonymous site spectrum at the reference loci and the neutral expectation.

Mismatch distributions

The number of differences between pairs of sequences in a given data set can be visualized by the mismatch distribution. Thereby, the number of pairs exhibiting a certain amount of differences can be blotted for each class of differences. The differences can be measured at intra- and interspecific sequence pairs. A mismatch distribution with interspecific pairs shows the degree of differentiation between species. Furthermore, this method can be used to identify shared alleles between species and putatively introgressed alleles (Castric *et al.* 2008). The program SITES was used to compile the mismatch distributions within and between the species used in this study. Alignment gaps and multiple hits were not considered. Numbers of differences per locus were adjusted to a length of 1,000 bp to allow comparisons between loci of different length.

Joint site frequency spectrum (JSFS)

The JSFS can be calculated to assess the frequency and number of shared and private polymorphisms between two populations or species (Wakeley & Hey 1997). It is an array S of dimension $(n_1 +1) \times (n_2 +1) -2$ where entry $S_{i,j}$ is the number of polymorphic sites for which the derived state is found i times in the sample from population 1 and j times in the sample from population 2. For example, $S_{2,3} = 10$ if 10 polymorphisms are found as doubletons in population 1 and as tripletons in population 2. For parameter estimation, Wakeley and Hey summarized the JSFS by a vector $W=(W_1,W_2,W_3,W_4)$ containing the number of private polymorphisms in species 1 and 2, respectively (W_1, W_2), fixed differences between species (W_3), and shared ancestral polymorphisms (W_4). The JSFS of all reference loci and the three resistance genes was calculated using an Awk script (B. Haubold, personal communication) and an R code (R Development Core Team 2005).

2.4.3 Phylogenetic inference

To assess the phylogenetic relationships between individuals used in this study and to detect phylogenetic differences between the different candidate genes and the reference dataset, phylogenetic analyses were completed using PAUP v. 4.0b10 (Swofford 1999, Sinauer Associates). The phylogenetic relationships between these sequences were determined using maximum parsimony (MP) and neighbor-joining (NJ) and these methods yielded similar topologies. The genealogy for the reference dataset was compiled using a concatenated sequence file containing all nine reference loci. The candidate genes were analysed independently.

CHAPTER 3: RESULTS

3.1 Evolutionary history of the *Rcr3* gene family

3.1.1 The *Rcr3* locus is duplicated in *S. peruvianum* and its sister species

To investigate the evolutionary history of the *Rcr3* locus I cloned and sequenced *Rcr3* alleles from accessions of the wild tomato species *S. peruvianum*. This approach revealed that *Rcr3* does not segregate as a single locus in this species, but forms a small gene family with closely related paralogs. The duplication of the *Rcr3* locus appears to be restricted to *S. peruvianum* and its sister species, *S. corneliomulleri,* since no evidence for a duplication event was found in the cultivated tomato or in the other tomato species investigated in this study. The *Rcr3* paralogs detected in this study could not be unambiguously distinguished from one another based on sequence divergence in the *Rcr3* ORF. Therefore, I sequenced the flanking regions (FLRs) of the alleles to define their genomic origin. BLAST and phylogenetic analyses of the *Rcr3* flanking regions were used to assign the different *Rcr3* alleles to their genomic origin relative to the genome of the cultivated tomato (*S. lycopersicum*). These analyses showed consistent results: Flanking regions, which corresponded to the orthologous *Rcr3* containing region of the cultivated tomato based on significant BLAST hits, clustered together in the phylogenetic tree, while flanking regions which mapped to other genomic locations based on the genome sequence of the cultivated tomato formed distinct clusters (Figure 3.1). The analyses of the *Rcr3* flanking regions revealed that the *Rcr3* gene was duplicated at least twice in *S. peruvianum* – the duplicates are named *Locus A*, *Locus B* and *Locus C* hereafter. All 5' flanking regions matched the *Rcr3* locus from cultivated tomato over the full sequenced length reaching 400 to 2000 bp upstream of the gene. This indicates that the duplicated region extends far upstream of the *Rcr3* gene. In contrast, only a portion of the 3' flanking regions matched the *Rcr3* locus from *S. lycopersicum.* At approximately 600 bp downstream of the stop codon, *Locus B* diverges from both *Locus A* and the *S. lycopersicum* sequence (Figure 3.2). This marks the likely insertion point of the duplicated *Rcr3* segments into a novel genomic location at the time of origin of these new duplicates. The 3' flanking regions of alleles originating from *Locus B* matched sequences 8.2 kb downstream from the *Rcr3* locus or other chromosomes in the tomato genome. The flanking regions of alleles originating from *Locus C* are characterized by a large deletion.

Long-Range-PCR results in *S. peruvianum* indicate that at least two copies of *Rcr3* are located in the same genomic region in a distance of approximately 10 kb from one another. Since the phylogenetic and BLAST analyses of the flanking sequences indicate that alleles from *Locus A* have the highest sequence similarity to *Rcr3* from other *Solanum* species, it is likely that alleles from *Locus A* are orthologous to the *Rcr3* gene in the other species, in which the *Rcr3* gene is not duplicated. This implies that *Locus B* and *Locus C* are more recently derived duplicates of *Locus A* in *S. peruvianum*.

In total, I was able to assign 27 of 43 *Rcr3* sequences obtained from *S. peruvianum* (Tarapaca population) unambiguously to a locus based on their flanking regions. Of these, 14 alleles were assigned to *Locus A*, nine alleles to *Locus B* and four to *Locus C*. All tested individuals had alleles that were assigned to two different *Rcr3* loci. In most cases, at least one allele originated from *Locus A* (Figure 3.1). For population genetic analyses, only alleles, which could be unambiguously assigned to their corresponding locus were used. Alleles originating from *Locus C* were excluded from the analysis as a sample size of four alleles is not sufficient for population genetic studies.

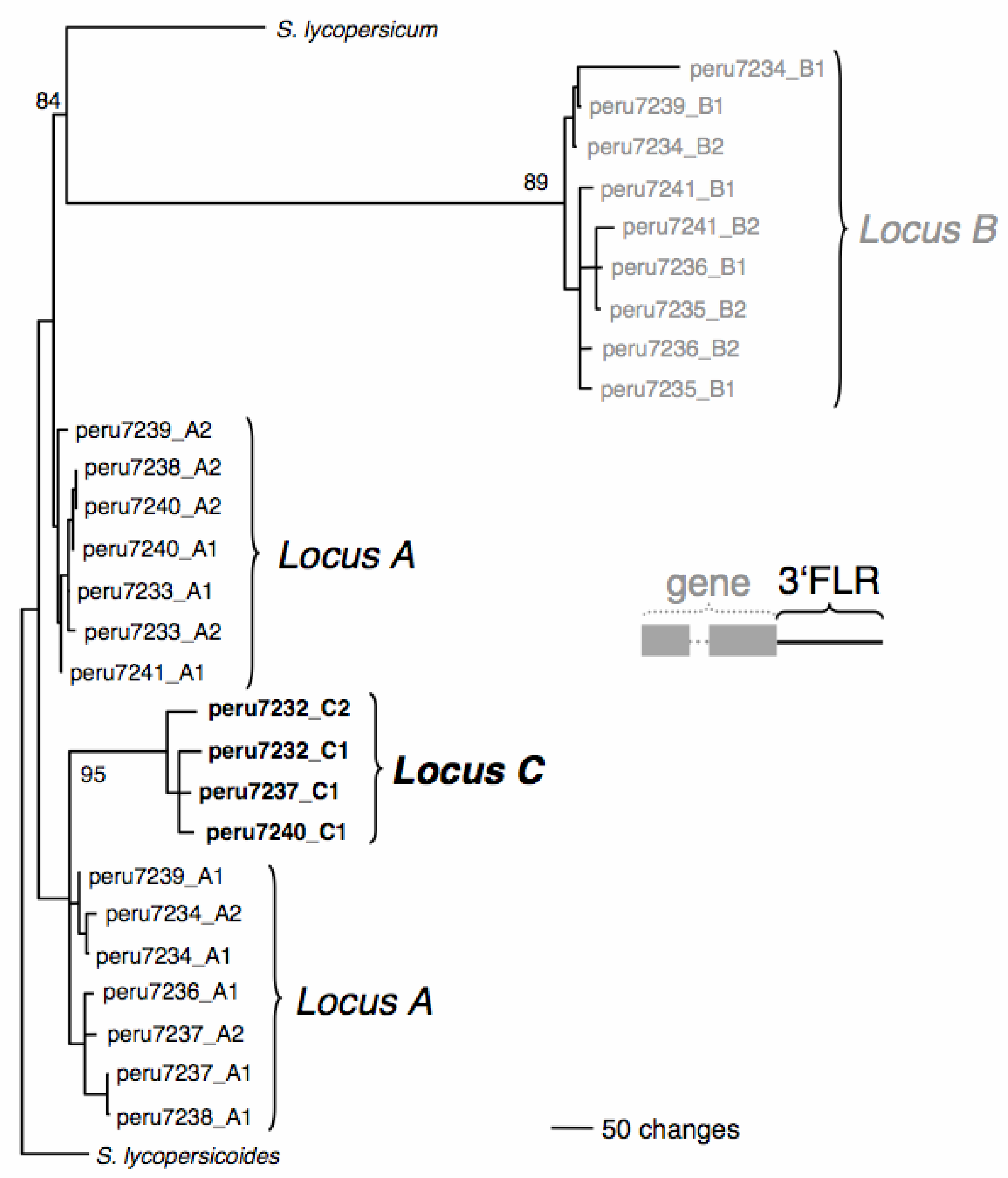


Figure 3.1: One of 1,000 most parsimonious trees of the 3'flanking region of the *Rcr3* gene (indicated in black in the sketch of the *Rcr3* locus). This tree was obtained by heuristic search with bootstrap support. *S. lycopersicoides* was used as outgroup.

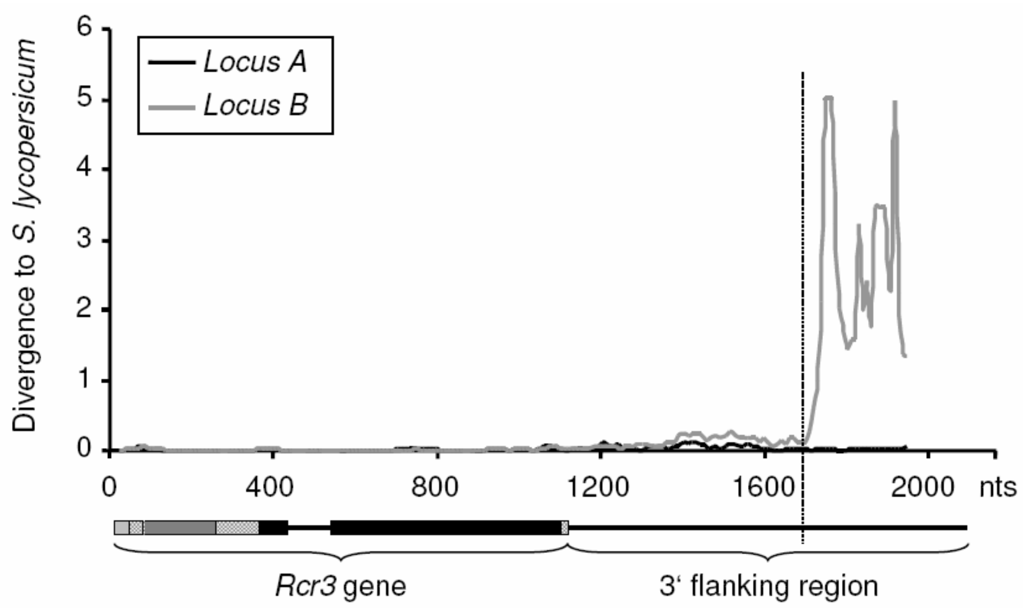


Figure 3.2: Divergence between *Locus A* and *Locus B* to *S. lycopersicum Rcr3*. The grey and black boxes represent the exons of the *Rcr3* gene, while the horizontal black line represents the intron within the *Rcr3* gene and the 3' flanking region. The vertical dashed line approximately 600 bp downstream of the stop codon and corresponding to the position at which *Locus B* diverges from both *Locus A* and the outgroup sequence, indicates the likely insertion point of the duplicated *Rcr3* segment into a novel genomic location at the time of origin of *Locus B*.

3.1.2 Two differentiated sequence types are maintained in the *Rcr3* gene family

Phylogenetic analyses of the coding sequence of all assigned *Rcr3* alleles revealed two well-separated clades, corresponding to differentiated sequence types (Figure 3.3). However, these two sequence types do not correspond to *Locus A, B and C* described above. Based on their classification according to their flanking regions, both *Rcr3* coding sequence types segregate at *Loci A* and *B*. The haplotype structure of the sequence types is mainly due to two different intronic types and variation linked to this intron (Figure B7). The two sequence types are highly differentiated from one another (fixation index $F_{ST} = 0.311$, which is larger than the average F_{ST} within *S. peruvianum* of 0.2, (Arunyawat *et al.* 2007), but exhibit an intermediate level of diversity within each sequence type ($\pi_{\text{sequence type 1}} = 0.007$, $\pi_{\text{sequence type 2}} = 0.005$). One possible explanation for this distinct haplotype structure could be long-term balancing selection as discussed in the following.

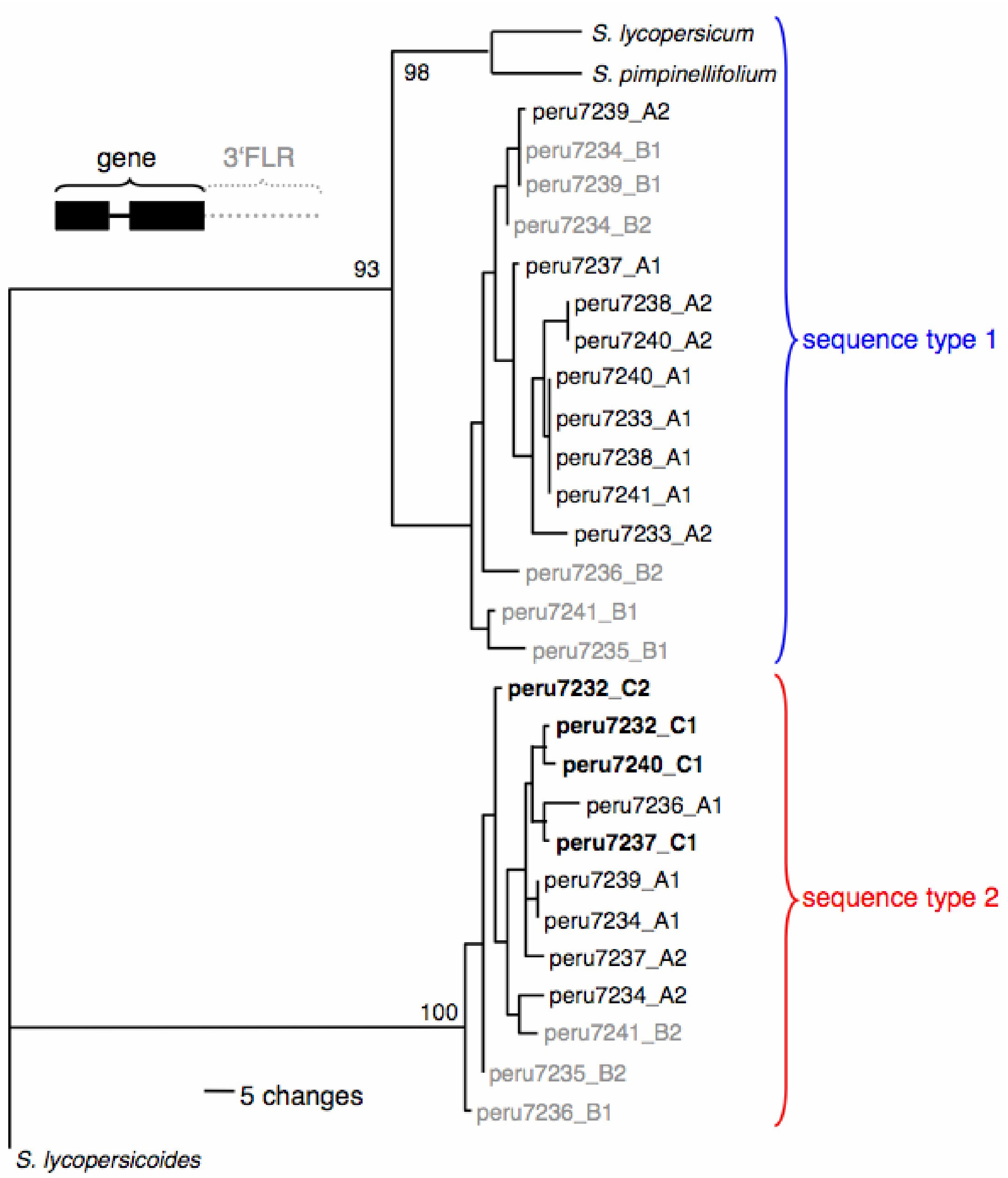


Figure 3.3: One of 1,000 most parsimonious gene trees of all assigned *Rcr3* alleles, obtained by heuristic search of the coding sequence (indicated in black in the sketch of the locus) of the *Rcr3* gene. Gaps were considered as a fifth state. Bootstrap proportions of 1,000 bootstrap replicates > 500 are indicated on the branches. The *Rcr3* sequence of the outgroup, *S. lycopersicoides,* was used to root the tree.

3.1.3 The evolutionary history of the *Rcr3* locus is characterized by balancing selection

To evaluate whether natural selection contributed to the maintenance of the distinct sequence types at the *Rcr3* locus, several population genetic statistics were calculated for the alleles of *Rcr3 Locus A* and *Locus B*. Putative pseudogenes (see below) were excluded from these analyses. To rule out demographic effects, which could interfere with the signature of natural selection acting at the *Rcr3* locus, all statistics were compared to a set of 14 reference loci, which had previously been sequenced in the same individuals of *S. peruvianum* (Baudry *et al.* 2001; Roselius *et al.* 2005; Städler *et al.* 2005)

Values of average nucleotide diversity π at the two analyzed *Rcr3* loci are 0.006 (*Locus A*) and 0.008 (*Locus B*) and are therefore lower than the genome average of this population of S. peruvianum (0.013) based on the set of 14 reference loci (Table 3.1). Synonymous nucleotide diversity π_s (0.008 at *Locus A* and 0.021 at *Locus B*) is also lower than the average π_s (0.023) in this population, while the nonsynonymous nucleotide diversity π_a (0.004 at *Locus A* and 0.004 at *Locus B*) is twice that at the reference loci (0.002). The ratio π_a to π_s is 0.513 for *Locus A* and 0.179 for *Locus B*, which is elevated compared to the average ratio for the reference loci (0.09).

Divergence to the outgroup species *S. lycopersicoides* at both *Rcr3* loci is 0.034 and is similar to the average divergence at the reference loci (average $K = 0.038$). Divergence at the *Rcr3* loci at synonymous and nonsynonymous sites exceeds the genomic mean two to threefold (average $K_s = 0.064$, average $K_a = 0.008$). Divergence across the *Rcr3* coding and 3' flanking region compared to the genome sequence of the cultivated tomato *S. lycopersicum* differs between *Locus A* and *Locus B* (Figure 3.2). The amount of divergence of *Locus A* and its flanking region compared to the *Rcr3* locus of the cultivated tomato is within the range of average divergence between *S. peruvianum* and *S. lycopersicum*. On the other hand, while *Locus B* behaves similarly until approximately 600 bp downstream of the *Rcr3* coding region, divergence between *Locus B* and the *Rcr3* locus of *S. lycopersicum* drastically increases beyond this point. This region of elevated divergence to the cultivated tomato coincides with the distal portion of the *Rcr3 Locus B* flanking region producing BLAST hits which do not match the *Rcr3* locus from *S. lycopersicum*.

Although I failed to detect a significant deviation from neutrality based on the standard tests of neutrality, both Tajima's *D* and Fu and Li's *D* values were elevated compared to the genome average at the *Locus B* ORF and Fu and Li's *D* values were elevated at the *Locus A* ORF (Table 3.1). Positive Tajima's *D* values indicate an excess of polymorphism at intermediate frequency, a pattern consistent with balancing selection.

Sliding window analyses depicting Tajima's *D* across the entire *Rcr3* coding region reveal that significantly elevated Tajima's *D* values are located at the intron and at some positions in the coding region consistent with the observation that these regions distinguish the two sequence types and may be maintained in the population by balancing selection (Figures 3.5 and B7). Tajima's *D* at the 3' flanking regions of the two *Rcr3* loci shows a different pattern. Tajima's *D* is highly positive at both 3' flanking regions (*Locus A* $T_D = 1.278$, *Locus B* $T_D = 1.462$) compared to the genome average. These values are not significantly positive based on neutral assumptions. However, in comparison to the expected neutral distribution of Tajima's *D* in this population estimated via coalescent simulations based on the set of reference loci, these values show up as highly significant (outside the 0.01 confidence interval, Figure 3.4).

Table 3.1: Summary statistics and neutrality test results calculated at the *Rcr3 Locus A* and *Locus B* and their 3' flanking regions in the *S. peruvianum* population Tarapaca.

	Rcr3* *Locus A	***Rcr3* *Locus B***	**3' FLR *Locus A***	**3' FLR *Locus B***	**mean at reference loci§**
analyzed length	1106	1106	511	527	1361
S	22	20	21	12	51
θ	0.006	0.007	0.015	0.008	0.014
π	0.006	0.008	0.017	0.011	0.013
π_s	0.008	0.021	-	-	0.023
π_a	0.004	0.004	-	-	0.0024
π_a/π_s	0.513	0.179	-	-	0.09
K	0.034[1]	0.034[1]	0.045[2]	0.116[2]	0.038[3]
K_s	0.084[1]	0.089[1]	-	-	0.064[3]
K_a	0.021[1]	0.021[1]	-	-	0.008[3]
K_a/K_s	0.241[1]	0.220[1]	-	-	0.177[3]
Tajima's *D*	-0.490	0.362	1.278*	1.462*	-0.386
Fu and Li's *D*	-0.063	-0.033	0.496	1.201	-0.49

§ The mean at the reference loci was estimated using a set of 14 reference loci.
[1] outgroup *S. lycopersicoides*, [2] outgroup *S. lycopersicum,* [3] outgroup *S. ochranthum*
* significantly outside the simulated distribution of Tajima's *D* in this population ($P < 0.01$) (Figure 3.4)

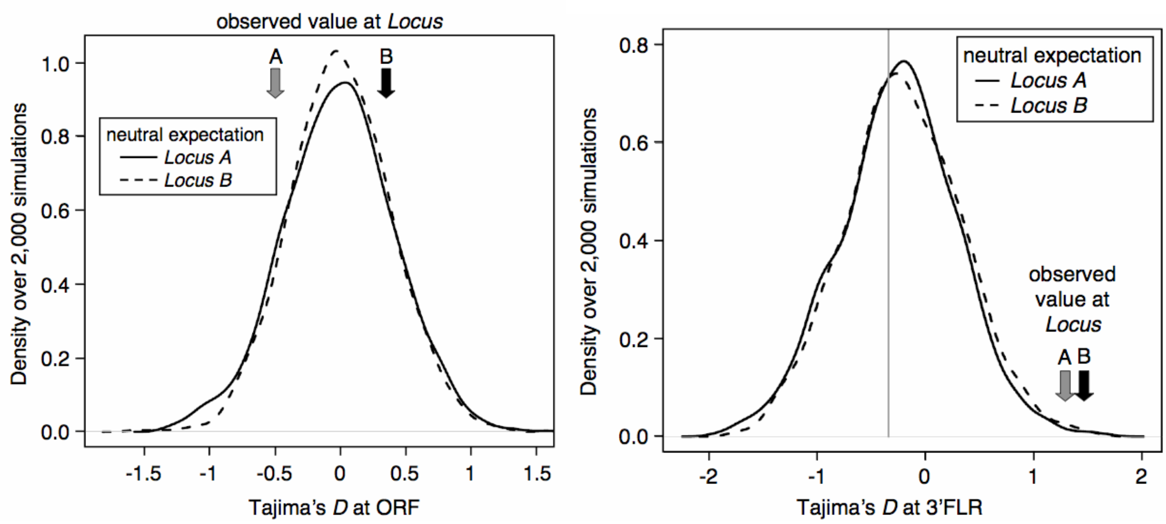


Figure 3.4: Distribution of neutral expectation of Tajima's *D* based on 2,000 coalescent simulations. Left: Distribution at the ORFs. Simulated distributions consider gene conversion at the rate observed at the *Rcr3* gene. The observed values at the *Rcr3* ORFs are within the 95% confidence interval. Right: Distribution at the 3'FLRs. Simulations are based on the 14 reference loci and assume population expansion. The grey vertical line indicates the observed average at the reference loci. The observed values at the *Rcr3* FLRs are outside the 99% confidence interval.

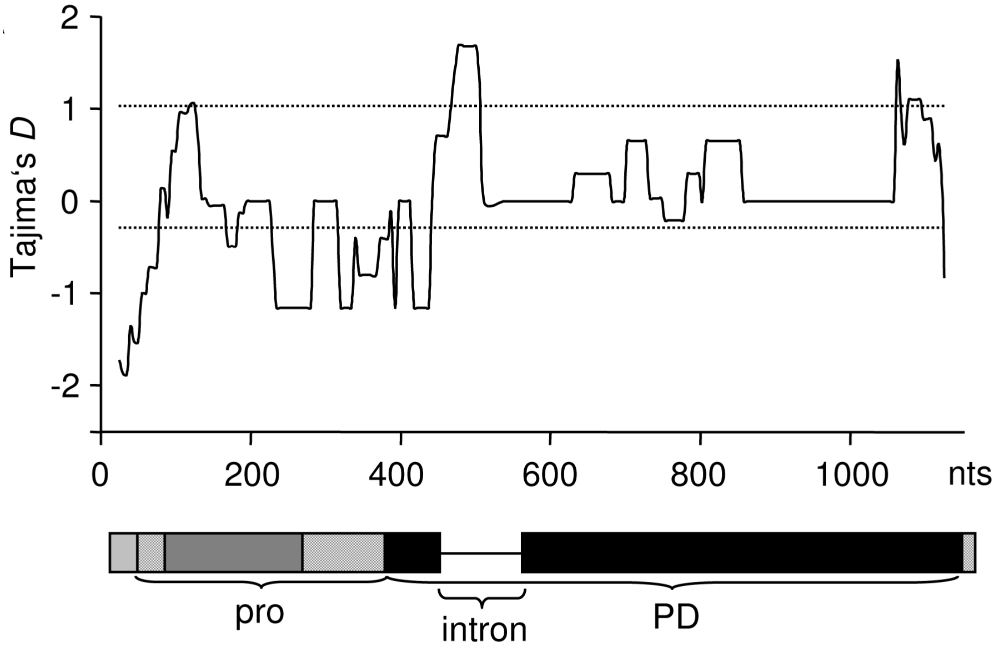


Figure 3.5: Tajima's *D* across the two *Rcr3* loci. The dotted lines indicate the mean and maximum of D_T measured at 14 reference loci. D_T is elevated at the intron and at other positions in the coding region. A schematic of the gene is indicated below the x-axis (grey and black boxes are the exons, pro = pro-domain, PD = protease domain).

3.1.4 The two *Rcr3* loci are affected by frequent gene conversion

The observations that alleles originating from different *Rcr3* duplicates could be amplified with the same set of PCR primers, that nucleotide diversity across loci is low and that the different *Rcr3* duplicates still seem to be segregating in the population are consistent with the recent origin of this small gene family in *S. peruvianum*. Young gene duplicates, which exhibit high sequence similarity, can undergo frequent intergenic gene conversion. I applied three different methods to test for gene conversion at the *Rcr3* locus.

First, I surveyed the site frequency spectrum of polymorphisms, which are shared between the two loci and private to only one of them (Innan 2003a). I found substantial shared polymorphism between alleles of the two *Rcr3* loci (Figure 3.6). Polymorphisms in the coding region, which are private to one of the loci, occur mainly in low frequency. No fixed differences differentiate alleles between the coding regions of the two loci. Taken together, these observations suggest a history of frequent gene conversion between the ORFs of the two *Rcr3* loci. However, the pattern at the 3' flanking region of the *Rcr3* gene is different. Here fewer polymorphisms are shared between the two loci and more polymorphisms are unique to one of them (Figure 3.6). These private polymorphisms are also found at intermediate to high frequencies. Also, in contrast to the coding region, a large number of fixed differences differentiate the two loci in the 3' flanking regions. These findings suggest that gene conversion does not happen as frequently in the 3' flanking region of the *Rcr3* gene compared to the coding region of *Rcr3*.

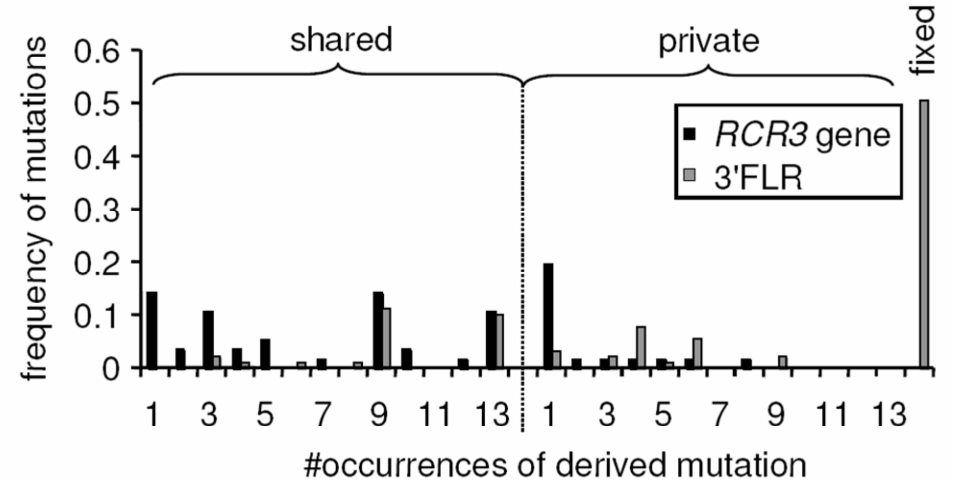


Figure 3.6: Frequency spectrum of derived shared and private polymorphisms at the two *Rcr3* loci and their 3' flanking regions (3'FLR). The outgroup sequence (*S. lycopersicum*) was used to define derived polymorphisms: shared polymorphisms occur in alleles from both *Rcr3* loci, private polymorphisms occur in only one locus, and fixed polymorphisms are fixed in one of the loci and do not occur in the other one.

Second, an Approximate Bayesian Computation method (Beaumont 2010) was developed to fit data, which had been simulated assuming different gene conversion rates (Thornton 2007), to the data observed at the *Rcr3* loci. Using this approach, the gene conversion rate between the *Rcr3* ORFs could be estimated to be different from zero, with a mode of 1.08 (Figure B1). This is more than 100 times larger than the mutation rate estimated in *S. peruvianum* of 0.014 (0.0085 at the *Rcr3* locus) confirming a high rate of gene conversion between the two loci. The model fitting procedure for the 3' flanking regions produced results, which were not as accurate as for the *Rcr3* ORFs. However here, the model without gene conversion fitted best. This is in accordance with the results obtained from the frequency spectrum and indicates a low rate of gene conversion at the *Rcr3* 3' flanking regions.

Using the program Geneconv, I performed a third independent analysis to evaluate the occurrence of gene conversion between the *Rcr3* loci. This analysis revealed that gene conversion occurs both within the gene and, to an even greater extent, in the 3' flanking regions. Innan (2003a) reported that Geneconv can underestimate frequent gene conversion events. Therefore, it is possible that the actual rate of gene conversion is estimated accurately in the 3' flanking regions by Geneconv, while the rate of gene conversion in the coding region is underestimated.

3.2 Functional consequences of sequence variation at the *Rcr3* locus

3.2.1 Natural variation at the *Rcr3* locus has effects on the interaction with effectors

Based on the presence of positive Tajima's *D* values concentrated in the intron of the *Rcr3* gene, three potential targets of selection can be envisioned: 1) selection on different regulatory motifs in the intron, 2) selection for different splicing variants, or 3) selection on one or more amino acid polymorphism(s) in linkage with the intron. To test the first scenario, I performed i*n silico* analysis to survey the two different intronic sequence types for differences in regulatory motifs (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/). This analysis did not reveal different regulatory motifs between the two intronic sequence types. To test the second possible scenario, I sequenced the mRNA from the two sequence types. Sequence comparisons along the two sequence types allowed me to exclude the existence of different splicing variants at the *Rcr3* locus. Therefore, it is most likely that the third scenario - balancing selection acting on amino acid polymorphism(s) linked to the intron – applies in the case of the *Rcr3* locus.

To investigate the functional consequences of natural variation at the *Rcr3* locus, I overexpressed 54 *Rcr3* alleles transiently in *N. benthamiana* by agroinfiltration. I chose these alleles from the data set used in the population genetic analyses and from additional individuals of *S. peruvianum* and closely related tomato species to maximize the amount of amino acid variation assayed. Of the total number of tested alleles, 47 were detected in apoplastic fluids (AFs) by Western blotting (Table 3.2, Figure B3). The remaining seven *Rcr3* alleles did not accumulate in multiple expression assays, although the accumulation of mRNA was confirmed by RT-PCR (Figure 3.7). Five of these *Rcr3* alleles that failed to accumulate have frameshift mutations, which lead to premature stop codons, potentially explaining their protein instability. The remaining two alleles that failed to accumulate each carry a single mutation, which distinguishes them from expressed alleles. These mutations likely interfere with protein stability or trafficking of the protein into the apoplast. Since these seven alleles appear to be pseudogenes, they were excluded from population genetic analyses described above.

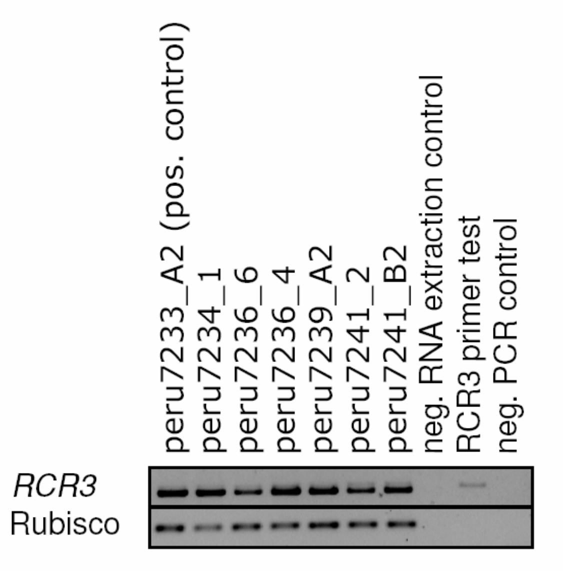


Figure 3.7: RT-PCR with Rcr3 constructs failing to accumulate in *N. benthamiana* AFs. RT-PCR was conducted for the *Rcr3* gene and a portion of the Ribulose-bisphosphate-carboxylase-oxigenase, as RNA-extraction control. PCR from genomic DNA was used to test if splicing of the *Rcr3* intron had occurred. AFs not expressing any *Rcr3* construct were used as negative control for RNA-extraction.

The activity of the Rcr3 proteins in AFs was detected by Activity-based Protein Profiling (ABPP) using fluorescent DCG-04. DCG-04 is an inhibitor of papain-like cysteine proteases and reacts irreversibly and covalently to the active site cysteine of proteases in an activity-dependent manner (Greenbaum *et al.* 2000). This assay has been used frequently to detect Rcr3 activity and its inhibition by Avr2 (Kaschani *et al.* 2010; Rooney *et al.* 2005; Shabab *et al.* 2008; van Esse *et al.* 2008). All expressed 47 Rcr3 proteins could be labelled by DCG-04 to similar levels, confirming that they all encode active proteases (Table 3.2, Figure B4). Inhibition assays based on competitive ABPP were performed to determine which Rcr3 can be inhibited by the pathogen protease inhibitors Avr2, Epic1, Epic2B and Rip1.

Inhibition by Avr2

Of the 47 tested Rcr3 alleles, 41 were inhibited by Avr2 (Table 3.2, Figure 3.13). The six alleles, which failed to be inhibited by Avr2 were isolated from individuals of *S. peruvianum* and *S. chilense*. One nonsynonymous substitution at position 692, resulting in a change from asparagine (N) to aspartic acid (D) at position 194 in the protein (N194D), is significantly associated with this phenotypic difference (R^2 = 0.842, P-value = 1.05 x 10^{-26}, Table 3.2, Figures 3.8 and B5). This finding is substantiated by previous structure-function studies using site-directed mutagenesis, which demonstrated that the N194D mutation in Rcr3 causes insensitivity to inhibition by Avr2 (Shabab *et al.* 2008). A single allele having the N194D substitution (peru1954_1) was inhibited by Avr2, however this allele shared additional differences to alleles insensitive to inhibition by Avr2 (Table 3.2, Figure B5). In addition to the N194D polymorphism, nucleotide polymorphisms at positions 717 (synonymous mutation) and 750 (causes amino acid variance R213S) were associated with insensitivity to inhibition by Avr2 (R^2 = 0.254, P-value = 2.9 x 10^{-6}; R^2 = 0.336, P-value = 9.8 x 10^{-8}).

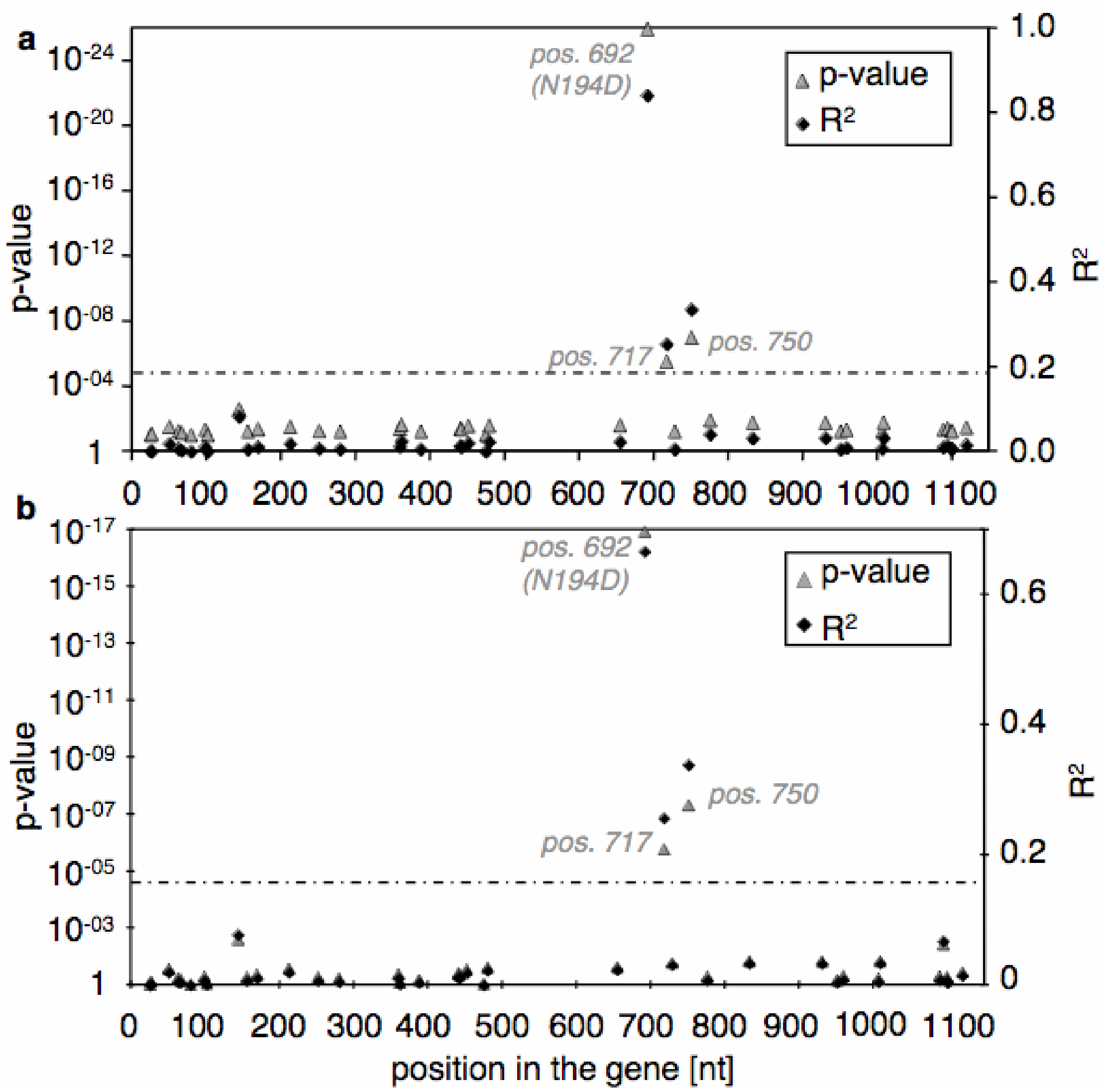


Figure 3.8: Association of SNPs along the *Rcr3* locus with inhibition by Avr2 *in vitro* (a) and *in planta* (b). SNPs were correlated with the observed phenotype using a general linear model. The y-axes on the left hand side show the *P*-values of the correlation. The y-axes on the right hand side show the correlation coefficient. Values were corrected by the Bonferroni method. The dashed line indicates the significance threshold after Bonferroni correction (0.01). **a**, Association with insensitivity to inhibition by Avr2 *in vitro*. **b**, Association with inability to elicit HR after co-infiltration with Avr2 into Cf-2/rcr3-3 tomato plants.

Inhibition by Epic1

Of all tested Rcr3 constructs, only nine constructs were inhibited by 1µM Epic1 (Table 3.2, Figure B6). Association of this phenotype with the Rcr3 genotype revealed five SNPs, which are significantly associated after Bonferroni-correction (Figure 3.9). These are nonsynonymous changes at nucleotide positions 280 ($R^2 = 0.275$, $P = 1.23 \times 10^{-6}$), 442 ($R^2 = 0.427$, $P = 8.62 \times 10^{-10}$), 452 ($R^2 = 0.534$, $P = 1.59\ 10^{-12}$) and 958 ($R^2 = 0.294$, $P = 5.30 \times 10^{-7}$) causing the amino acid changes Q94E, H148N, R151Q and D283N and one synonymous change at position 441 ($R^2 = 0.449$, $P = 2.59 \times 10^{-10}$). Nearly all constructs, which were inhibited by Epic1, carry the amino acid substitutions H148N and R151Q, which are significantly associated with Epic1 inhibition. Constructs, which carried these substitutions, but additional substitutions such as N174K were not inhibited by Epic1 (Table 3.2, Figure B5). A structural model of the Rcr3 protease domain shows that the significantly associated amino acid substitutions are located around the catalytic centre of the protease (Figure 3.10). These amino acid substitutions were found in this dataset in *S. peruvianum*, *S. chilense*, *S. habrochaites* and *S. lycopersicoides*.

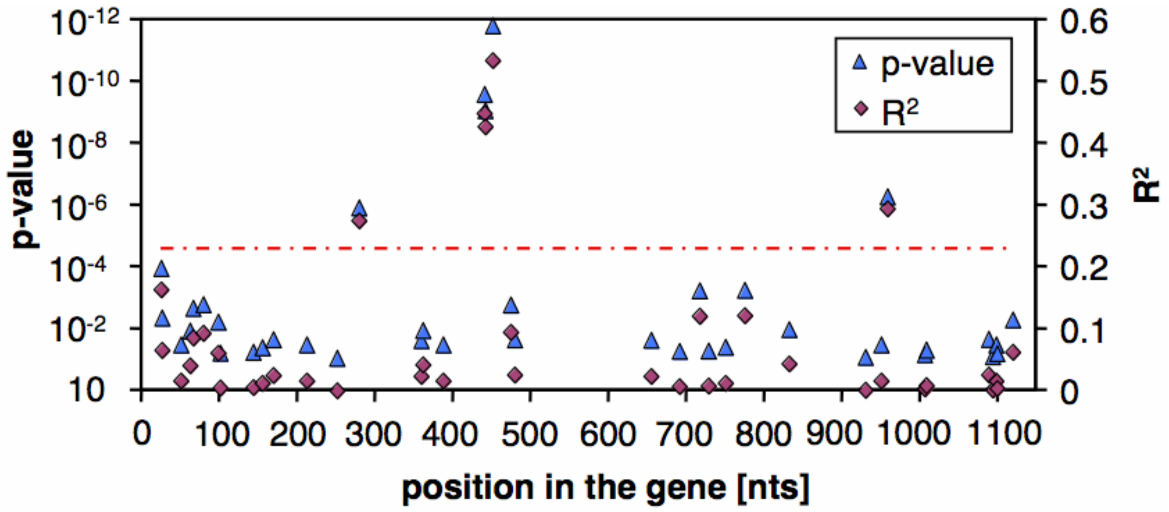


Figure 3.9: Association of SNPs along the *Rcr3* locus with inhibition by Epic1 *in vitro*. SNPs were correlated with the observed phenotype using a general linear model. The y-axis on the left hand side shows the *P*-values of the correlation. The y-axis on the right hand side shows the correlation coefficient. Values were corrected by the Bonferroni method. The dashed line indicates the significance threshold after Bonferroni correction (1%).

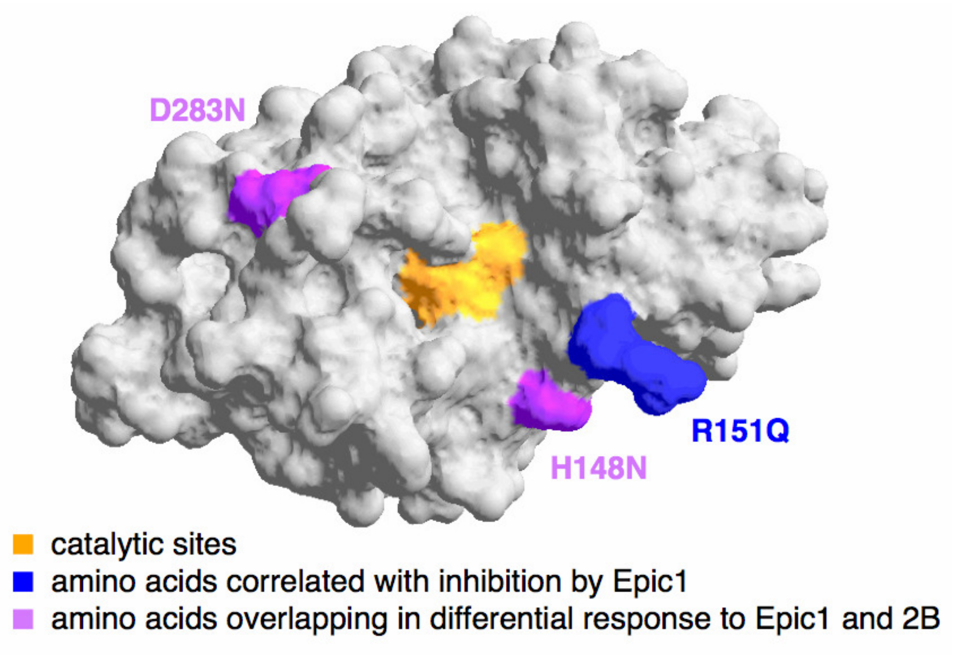

D283N
R151Q
H148N
catalytic sites
amino acids correlated with inhibition by Epic1
amino acids overlapping in differential response to Epic1 and 2B

Figure 3.10: Structural model of the Rcr3 protease domain. Amino acids associated with inhibition by Epic1 or insensitivity to inhibition by Epic2B are highlighted. The two amino acids, which are associated with insensitivity to inhibition by Epic2B exhibit overlapping associations.

Inhibition by Epic2B

The pattern regarding inhibition of Rcr3 constructs by Epic2B looks different. This effector inhibited nearly all tested constructs. Only four constructs were insensitive to inhibition by Epic2B (Table 3.2, Figure B6). Association of this phenotype with the *Rcr3* genotype revealed two nonsynonymous SNPs, which are significantly associated after Bonferroni-correction (Figure 3.11). These are at the nucleotide positions 26 ($R^2 = 0.174$, $P = 0.00101$) and 958 ($R^2 = 0.215$, $P = 2.17 \times 10^{-5}$) causing the amino acid changes N9S and D283N. The positions 80 ($R^2 = 0.152$, $P = 0.0022$, amino acid change G27A), 441 ($R^2 = 0.126$, $P = 0.0057$, synonymous) and 442 ($R^2 = 0.162$, $P = 0.0016$ amino acid change H148N) were marginally significant after Bonferroni-correction. The amino acid substitutions H148N and D283N showing significant correlation are centered around the catalytic site of the Rcr3 protease domain and are identical with amino acid substitutions, which are significantly correlated with inhibition by Epic1.

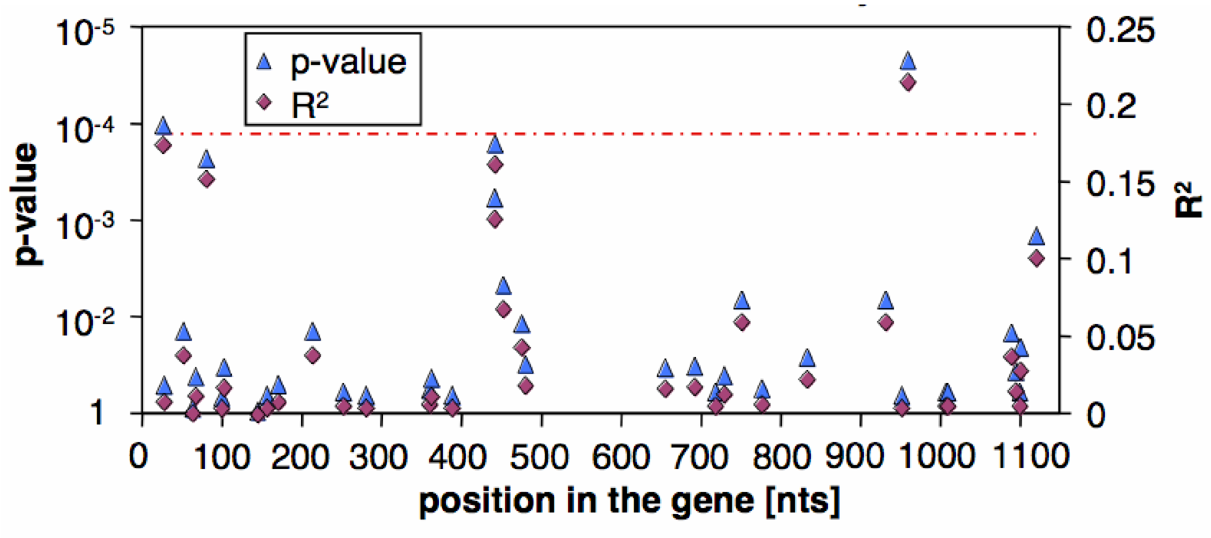


Figure 3.11: Association of SNPs along the *Rcr3* locus with inhibition by Epic2B *in vitro*. SNPs were correlated with the observed phenotype using a general linear model. The y-axis on the left hand side shows the *P*-values of the correlation. The y-axis on the right hand side shows the correlation coefficient. Values were corrected by the Bonferroni method. The dashed line indicates the significance threshold after Bonferroni correction (5%).

Inhibition by Rip1

Most of the tested constructs were not inhibited by Rip1. This effector only had an inhibitory effect on the protease function for eleven constructs (Table 3.2, Figure B6). Association of this phenotype did not reveal any polymorphic sites, which are significantly associated with this phenotypic behaviour (Figure 3.12).

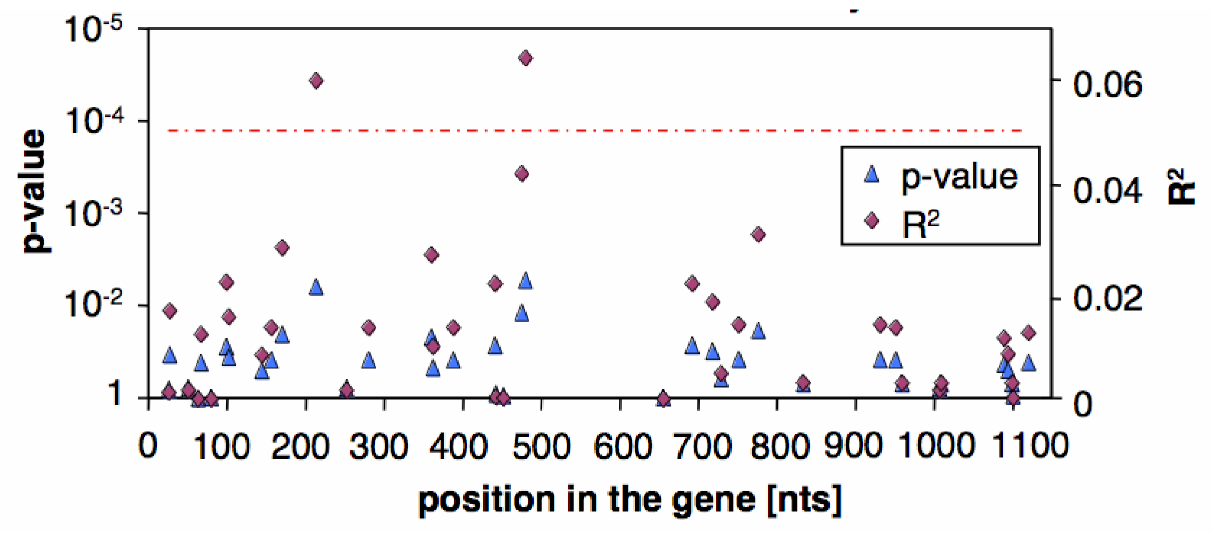
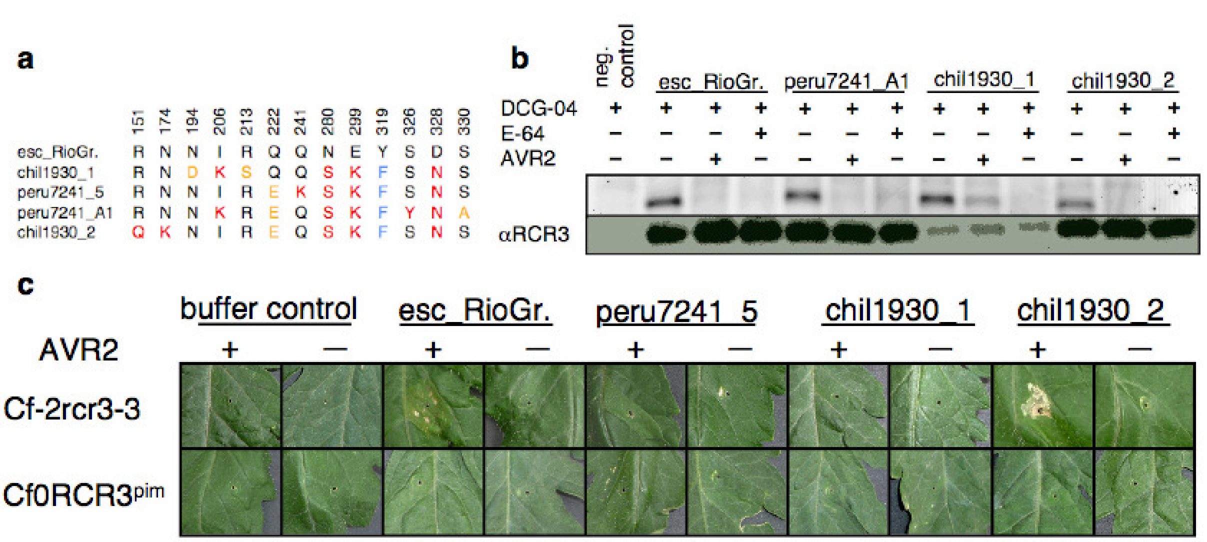


Figure 3.12: Association of SNPs along the *Rcr3* locus with inhibition by Rip1 *in vitro*. SNPs were correlated with the observed phenotype using a general linear model. The y-axis on the left hand side shows the *P*-values of the correlation. The y-axis on the right hand side shows the correlation coefficient. Values were corrected by the Bonferroni method. The dashed line indicates the significance threshold after Bonferroni correction (5%).

3.2.2 Amino acid variation at the Rcr3 locus translates into differential strength of HR

Rcr3 variants were evaluated for their ability to elicit the hypersensitive response in tomato plants. All Rcr3 alleles confirmed to be active proteases were co-infiltrated into Rcr3-mutant/Cf-2 (Cf-2Rcr3-3) tomato plants with either Avr2, Epic1, Epic2B or Rip1 or infiltration buffer as a negative control (Figure 3.13). Out of the 47 tested alleles, the 40 alleles, which can be inhibited by Avr2 in the inhibition assays specifically induced HR upon co-infiltration with Avr2. The remaining seven Rcr3 alleles, which were insensitive to inhibition by Avr2 did not induce HR upon co-infiltration with Avr2. Consistently, this phenotypic observation is significantly associated with the N194D polymorphism as well (Figures 3.8 and 3.13). Among all tested alleles, which do not carry the N194D substitution, only a single allele, peru7233_2, did not induce HR. This allele has one amino acid difference compared to other alleles, which induced HR upon co-infiltration with Avr2: R138I.



	151	174	194	206	213	222	241	280	299	319	326	328	330
esc_RioGr.	R	N	N	I	R	Q	Q	N	E	Y	S	D	S
chil1930_1	R	N	D	K	S	Q	Q	S	K	F	S	N	S
peru7241_5	R	N	N	I	R	E	K	S	K	F	S	N	S
peru7241_A1	R	N	N	K	R	E	Q	S	K	F	Y	N	A
chil1930_2	Q	K	N	I	R	E	Q	S	K	F	S	N	S

Figure 3.13: Phenotypic evaluation of a subset of Rcr3 alleles. One representative result out of at least three independent replicates is shown. **a**, Variable amino acids in the protease domain of the shown alleles: red = nonsimilar amino acid, blue = similar amino acid, orange = functionally relevant amino acids **b**, Inhibition assays with Avr2. AF without overexpressed Rcr3 was used as a negative control. Expression of each Rcr3 construct was confirmed by protein blots using αRcr3 for detection. Despite lower concentration, chil1930_1 was less inhibited by Avr2. **c,** *In planta* assays of Rcr3 alleles. All active Rcr3 constructs were co-infiltrated into Cf-2/rcr3-3 and Cf0/RCR3^{pim} tomato plants with Avr2 or buffer. Necrotic lesions indicate HR. Yellow discolouration of the leave tissue indicates weak HR.

Interestingly, Rcr3 alleles showed differences in the strength of the HR response with several alleles showing weaker HR when being co-infiltrated with Avr2 compared to others. Five SNPs are significantly correlated with phenotypic variation in the strength of the HR, one of which is statistically significant after Bonferroni-correction (Figure 3.14). These nucleotide positions are: 102 (R^2 = 0.298, P-value = 7.8 x 10^{-7}), 144 (R^2 = 0.113, P-value = 0.0086), 728 (R^2 = 0.1, P = 0.015) causing the amino acid change I206K, 775 (R^2 = 0.132, P-value = 0.0044) causing the amino acid change Q222E and 1099 (R^2 = 0.146, P-value = 0.0026) causing the amino acid change S330A. In a structural model of the Rcr3 protease domain, these three substitutions are located in close proximity to positions with known functional consequences on compatibility between Rcr3 and Cf-2 (Figure 3.15). Together with the intron, all five mutations are located in the regions of positive Tajima's *D* values in the sliding window analysis (Figure 3.14). No defence response involving necrotic tissue collapse was reported after up to 10 days post inoculation when Rcr3 was co-infiltrated with Epic1, Epic2B or Rip1.

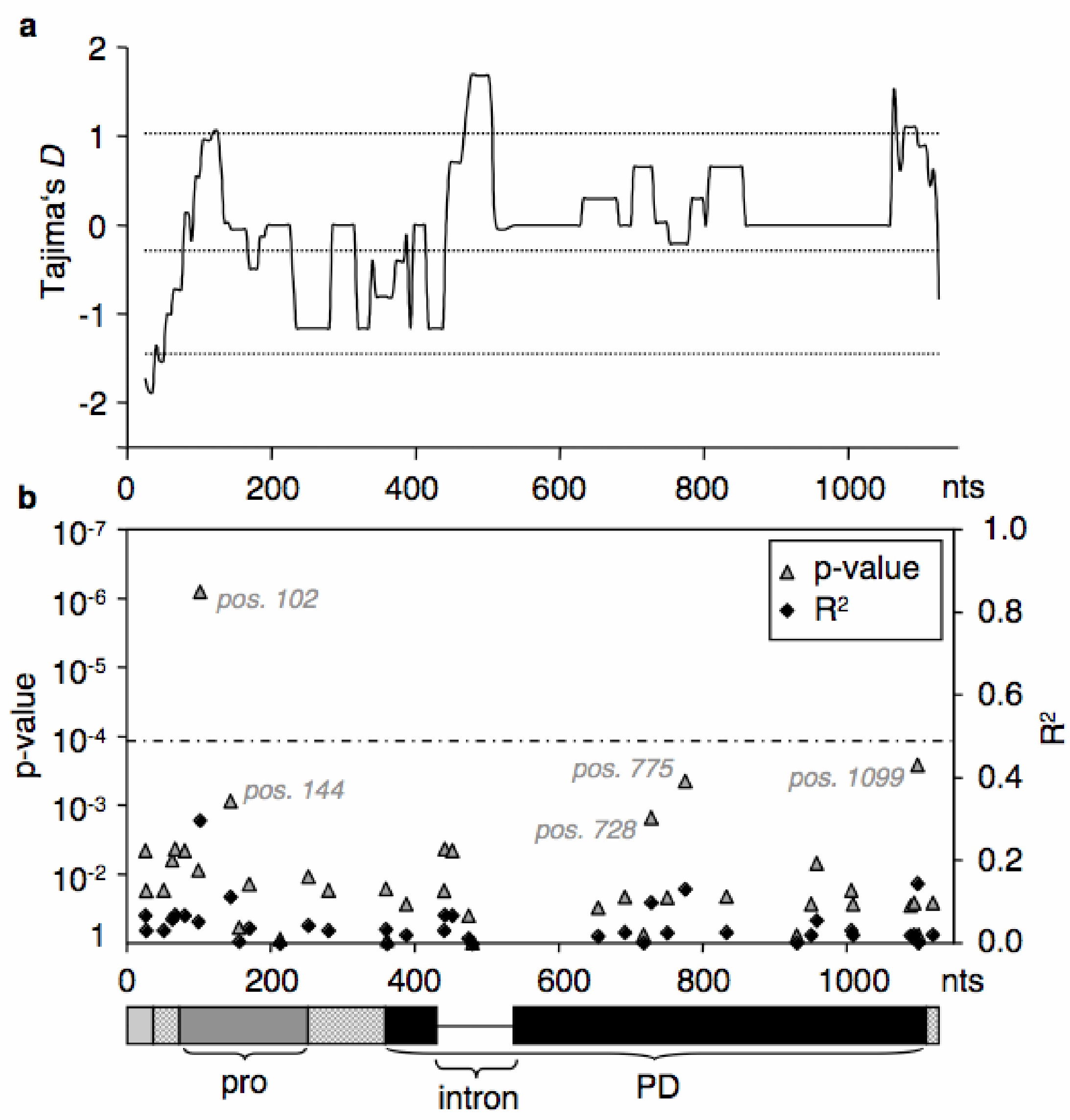


Figure 3.14: Association of the weak HR phenotype with sequence polymorphism. a, Tajima's *D* across the two *Rcr3* loci. The dotted lines indicate the minimum, mean and maximum of Tajima's *D* measured at 14 reference loci. Tajima's *D* is elevated at sites which are associated with the weak HR compared to the neutral expectation (= 0) and even more so compared to the genomic average. **b**, Association of SNPs and phenotypic variation in HR response with *P*-value of the correlation (left y-axis) and correlation coefficient (right y-axis). Values were corrected using the Bonferroni-method (dashed line indicates the 5% significance threshold). A schematic of the gene is indicated below the x-axis (grey and black boxes are the exons, pro = pro-domain, PD = protease domain).

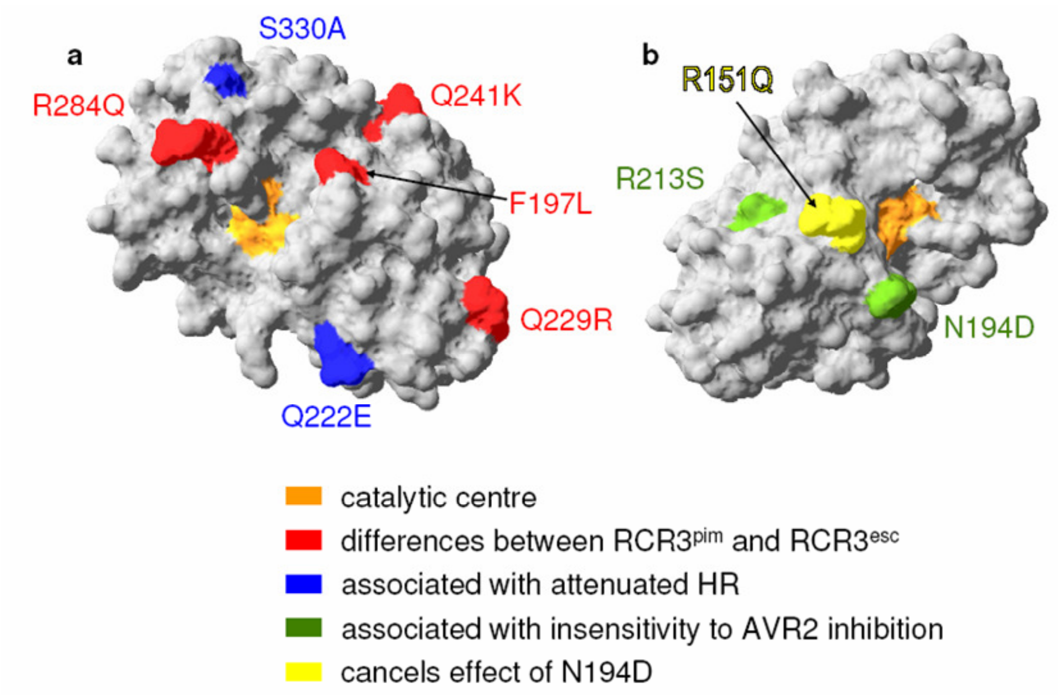

a
S330A
R284Q
Q241K
F197L
Q229R
Q222E
b
R151Q
R213S
N194D
🟧 catalytic centre
🟥 differences between RCR3pim and RCR3esc
🟦 associated with attenuated HR
🟩 associated with insensitivity to AVR2 inhibition
🟨 cancels effect of N194D

Figure 3.15: Structural models of the Rcr3 protease domain. a, amino acids, which are associated with the weak HR response or incompatibility between Rcr3 and Cf-2 are highlighted. **b**, amino acids, which are associated with insensitivity to Avr2 inhibition are highlighted.

Occurrence throughout phylogeny

Blotting the different observed phenotypic traits of Rcr3 constructs on a phylogeny of the *Rcr3* gene reveals that Avr2 and Epic2B can inhibit most of the Rcr3 alleles, while Epic1 and Rip1 only have an inhibitory effect on single alleles (Figure 3.16). Insensitivity to inhibition by Avr2 and Epic2B occurs only in few alleles. This differential interaction of *Rcr3* alleles with different effectors is spread out all over the phylogeny without special pattern. It does neither correspond to the peculiar haplotype structure observed at the *Rcr3* locus, nor the different *Rcr3* copies. However interestingly, most individuals carry alleles, which are able to interact with several effectors and may in combination confer more than one recognition specificity to the individual. Alleles, which can be inhibited by Epic1 tend to be insensitive to inhibition by Epic2B.

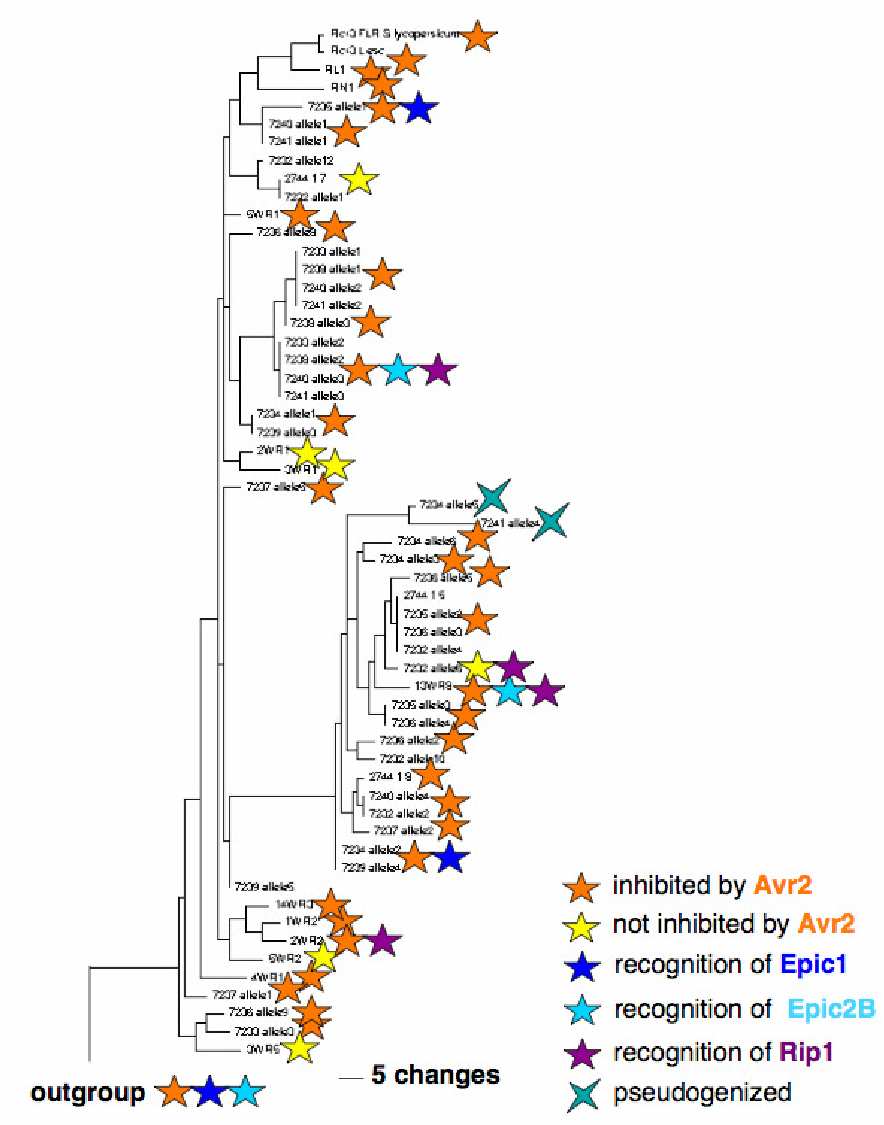


Figure 3.16: Gene tree of most tested *Rcr3* alleles. The tree was obtained by heuristic search using *S. lycopersicoides* as outgroup. The coloured stars indicate the phenotypic behaviour of each construct. The clade containing the two pseudogenized alleles corresponds to sequence type 2 in Figure 3.3.

Table 3.2: Summary of all phenotypic results of the different Rcr3 constructs. All phenotypic results including protein accumulation in AFs, activity-based protein profiling, inhibition by all four effectors and HR-response are shown. Constructs are named according to their species, their accession or individual number and their origin from *Locus A*, *B* or *C* in those cases for which unambiguous assignment was possible. [a]identical on the protein level to peru7236_A1, [b]identical to peru7234_A1, [c]identical to peru7234_B1 and peru7234_B2, [d]identical to peru7233_A1, peru7238_A1 and peru7240A1, [e]identical to peru7232_C2, [f]identical to peru7238_A2, [g]identical to peru7235_B2, peru7236_B1 and peru7241_B1

+ = phenotype present, **-** = phenotype absent, **(+)** = weak response, n.t. = not tested

Rcr3 construct	Protein accumulation	Protease acitvity					HR-response	
		-effector	+Avr2	+Epic1	+Epic2B	+Rip1	-AVR2	+AVR2
esc_RioGrande	+	+	-	+	-	+	-	+
peru7233_2	+	+	-	+	-	-	-	-
chil1930_3	+	+	+	+	-	+	-	-
peru7232_5	+	+	+	+	-	-	-	-
peru7232_1	+	+	+	+	-	-	-	-
chil1930_1	+	+	+	-	-	+	-	-
peru1954_2	+	+	+	+	+	+	-	-
peru0446_2	+	+	+	-	-	+	-	-
peru7241_2[a]	-	-	n.t.	+	-	+	-	-
peru7233_3	+	+	-	+	-	+	-	(+)
peru7241_5	+	+	-	+	-	+	-	(+)
peru7234_2	+	+	-	+	-	-	-	(+)
peru2744_3	+	+	-	+	-	-	-	(+)
peru7232_2	+	+	-	+	-	+	-	(+)
peru7234_3	+	+	-	+	-	+	-	(+)
peru7236_3	+	+	-	+	-	+	-	(+)
peru7237_2	+	+	-	+	-	+	-	(+)
peru7235_B1	+	+	-	-	-	-	-	(+)
peru7239_A1[b]	+	+	-	-	-	-	-	(+)
peru7239_B1[c]	+	+	-	+	-	+	-	(+)
peru1954_1	+	+	-	+	-	+	-	(+)
peru7237_A1	+	+	-	+	-	-	-	(+)
peru7241_A1[d]	+	+	-	+	-	+	-	+
peru7233_1	+	+	-	-	-	+	-	+
peru7237_C1	+	+	-	+	-	-	-	+
peru7241_3	+	+	-	+	-	+	-	+
corn1973_1	+	+	-	+	+	+	-	+
chil2748_1	+	+	-	+	-	+	-	+
peru0446_1	+	+	-	+	-	+	-	+
peru2744_1[e]	+	+	-	+	-	-	-	+
hab1777_1	+	+	-	+	-	+	-	+
chil1958_1	+	+	-	+	-	+	-	+
peru7234_A2	+	+	-	+	-	+	-	+
peru7236_4	+	+	-	+	-	+	-	+
peru7236_5	+	+	-	+	-	-	-	+
peru7238_1	+	+	-	+	-	+	-	+
peru7232_4	+	+	-	+	-	+	-	+
lyco2951_1	+	+	-	-	-	+	-	+
peru7240_A2[f]	+	+	-	+	-	+	-	+
corn1274_3	+	+	-	+	-	+	-	+
chil1930_2	+	+	-	+	-	+	-	+
pimp0400_1	+	+	-	+	-	+	-	+
hab1777_3	+	+	-	-	-	+	-	+
chm3653_1	+	+	-	-	-	-	-	+
pen0716_1	+	+	-	+	+	+	-	+
pen3791_2	+	+	-	-	+	+	-	+
peru7233_A2	+	+	-	+	-	+	-	+
peru7232_3	+	+	-	+	-	+	-	+
peru7240_1[g]	+	+	-	+	-	+	-	+
corn1274_1	-	-	n.t.	+	-	+	n.t.	n.t.
peru7241_B2	-	-	n.t.	+	-	+	n.t.	n.t.
peru7234_1	-	-	n.t.	+	-	+	n.t.	n.t.
peru7236_6	-	-	n.t.	+	-	+	n.t.	n.t.
peru7239_A2	-	-	n.t.	+	-	+	n.t.	n.t.

3.3. Contrasting evolutionary patterns at a disease resistance pathway

In this project, patterns of sequence evolution at five genes involved in the *Pto*-disease resistance (*Pto*, *Fen*, *Prf*, *Pfi* and *Rin4*) were assessed in one population of the wild tomato species *S. peruvianum* and compared to a set of reference genes.

3.3.1 Elevated level of polymorphisms at *Pto* and *Pfi*

Polymorphism, as quantified by average pairwise differences across all sites (π), in the five resistance genes ranges from 0.006 (*Prf*) to 0.016 (*Pfi*) (Table 3.3). For comparison, the mean across the set of 14 reference genes for this same population is 0.013. *Pfi* and *Pto* showed the highest polymorphism at synonymous sites π_s (0.022 and 0.02, respectively), as well as at nonsynonymous sites π_a (0.013 at both loci) (Table 3.3, Figure 3.17). The ratio of π_a to π_s was 0.57 for *Pfi* and 0.62 for *Pto*, while this ratio was consistently much lower at the 14 reference loci (mean π_a to $\pi_s = 0.099$).

Neutral coalescent simulations were used to test if the value of π observed at nonsynonymous and synonymous sites fell within the 95% confidence interval of simulations in which the population mutation parameter theta (θ) was estimated from the average π across 14 non *R*-genes from these same individuals. These coalescent simulations indicated that both *Pfi* and *Pto* show excess variation, specifically at nonsynonymous sites (*P*-value < 0.001), while at synonymous sites the observed level of variation at *Pfi* and *Pto* is within the 95% confidence interval based on the 14 reference genes (Table 3.4). The other three resistance genes did not show deviations from the expected distribution. A significant departure from neutrality at *Pfi* is also captured in the McDonald-Kreitman test (Table 3.5, McDonald & Kreitman 1991). According to this test, *Pfi* displays significantly more variation at nonsynonymous positions than expected under neutrality. A closer inspection of the distribution of variation across this large gene reveals that the putative hydrolase region and nuclear localization signal harbour substantial amounts of nonsynonymous variation ($\pi_{non} = 0.0216$; Figure 3.18). In contrast, nonsynonymous variation for the remainder of the gene is 0.00423. Other neutrality tests such as Tajima's *D* (T_D) and Fu and Li's *D* (FL_D) did not reveal significant departures from neutrality at any of the five genes (Table 3.3). However, at the *Pto* locus, both Tajima's *D* and Fu and Li's *D* are positive (0.064 and 0.129) compared to negative reference loci means of -0.417 (T_D) and -0.490 (FL_D). At the *Pfi* gene, only the Fu and Li's *D* test resulted in a positive value (0.232), which is even elevated when only the putative hydrolase domain is analyzed (0.618).

Table 3.3: Summary statistics measured at the five resistance genes and the 14 reference loci in the Tarapaca population. The mean values were calculated using the 14 reference loci.

Locus	π_{total}	π_s	π_a	π_a/π_s	T_D	FL_D
CT066	0.00984	0.03366	0.00204	0.06	-0.307	-0.441
CT093	0.00568	0.01763	0.00105	0.06	-0.141	0.006
CT099	0.01827	0.02138	0.00709	0.33	-0.815	-0.854
CT114	0.00820	0.01605	0.00000	0.00	0.189	-0.122
CT143	0.01839	0.01652	0.00000	0.00	0.250	0.232
CT148	0.01433	0.02010	0.00524	0.26	-1.059	-1.104
CT166	0.01429	0.00699	0.00078	0.11	-0.034	-0.485
CT179	0.01069	0.03457	0.00000	0.00	-0.163	-0.085
CT189	0.01010	0.00726	0.00000	0.00	-1.578	-1.784
CT198	0.02924	0.05648	0.00182	0.03	0.069	0.024
CT208	0.00746	0.00674	0.00000	0.00	-0.598	-0.737
CT251	0.01400	0.03448	0.00721	0.21	-0.250	-0.568
CT268	0.00941	0.02587	0.00446	0.17	-0.735	-0.521
sucr	0.01406	0.02692	0.00401	0.15	-0.669	-0.416
mean	**0.013**	**0.023**	**0.0024**	**0.099**	**-0.417**	**-0.490**
Pto	**0.01450**	**0.02038**	**0.01278**	**0.62**	**0.064**	**0.129**
Fen	**0.00871**	**0.0156**	**0.00676**	**0.43**	**-0.551**	**-0.672**
Prf	**0.00667**	**0.01386**	**0.00448**	**0.32**	**-0.272**	**-0.473**
Pfi	**0.01662**	**0.02233**	**0.01277**	**0.57**	**-0.592**	**0.232**
Pfi hyd	**0.01978**	**0.0282**	**0.02119**	**0.75**	**-0.928**	**0.618**
Rin4	**0.00924**	**0.01984**	**0.00320**	**0.16**	**-0.881**	**-0.689**

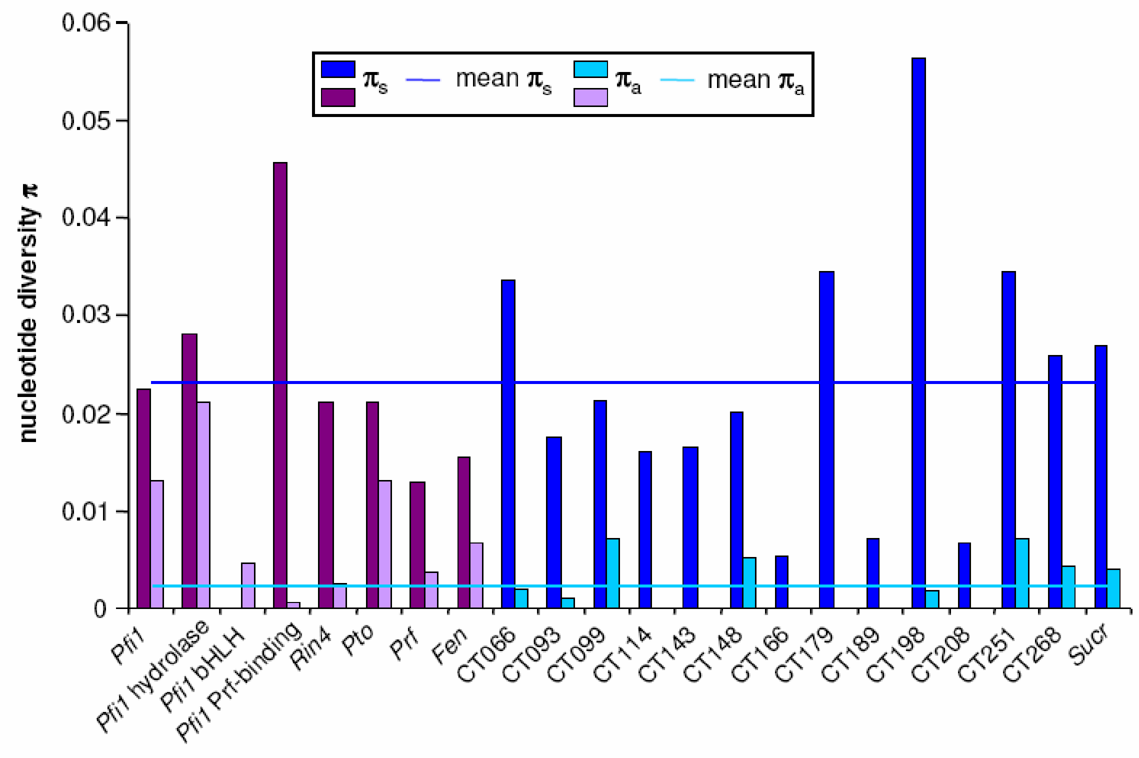


Figure 3.17: Comparison of average nucleotide diversity π at the five resistance (purple bars) and the reference genes (blue bars). The mean values were calculated using the 14 reference loci.

Table 3.4: Results of coalescent simulations at the *Pto* and *Pfi* locus.

Locus	π_s	P(π exp > π obs[b])[c]	π_a	P(π exp > π obs[b])[c]
reference genes[a]	0.023		0.0024	
Fen	0.016	0.668	0.0068	0.015*
Pfi	0.022	0.413	0.013	0.0001**
Prf	0.014	0.787	0.0045	0.085
Pto	0.020	0.390	0.013	0.0001**
Rin4	0.020	0.520	0.0032	0.191

[a] Arithmetic mean of π at synonymous sites from 14 genes (CT066, CT093, CT099, CT114, CT143, CT148, CT166, CT179, CT189, CT198, CT208, CT251, CT268, and *Sucr*).

[b] π obs is the arithmetic mean of π at nonsynonymous sites from 14 genes.

[c] Probability of observing a value of π greater than that observed at the reference loci in 1,000 coalescent simulations, conditioned on the π values of the non *R*-genes.

Table 3.5: Results of the McDonald-Kreitman test on *Pfi*. *S. lycopersicoides* was used as outgroup. *P*-value = 0.0026

	fixed differences	polymorphisms
silent	131	252
nonsynonymous	18	90

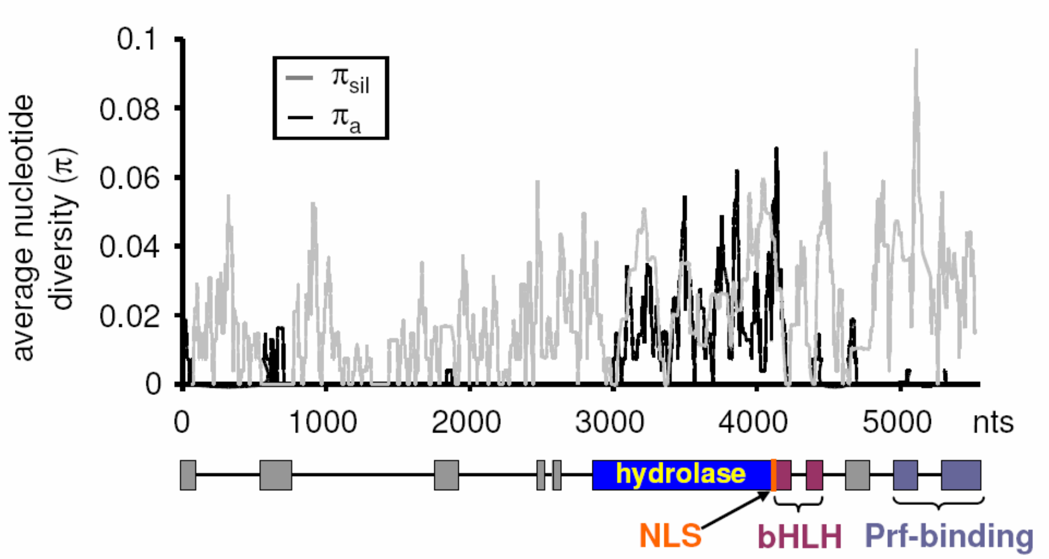


Figure 3.18: Sliding window of silent and nonsynonymous nucleotide diversity π in the Tarapaca population at the *Pfi* gene. Values are midpoints of 30 bp windows. The sketch underneath represents the structure of the gene. Grey and coloured boxes symbolize exons, black lines symbolize Introns. NLS = nucleus localization signal

3.3.2 Mixed pattern of polymorphisms at *Fen* and *Prf*

Fen and *Prf* show the lowest levels of polymorphism of these five loci and intermediate values for the ratio of π_a versus π_s (0.43 and 0.32, Table 3.3, Figure 3.17). *Fen*, like *Pto*, is a small gene (966 nucleotides) and encodes a functional protein kinase. In contrast, *Prf* is a large gene, made up of both well-defined and poorly-defined domains. These different domains show different evolutionary histories, as captured in the sliding window analyses (Figure. 3.19). In contrast to many other *R*-genes, the LRR region of *Prf* does not show an excess of amino acid polymorphism. Instead, two peaks of amino acid polymorphism are located in the N-terminal portion of the protein which binds to the host proteins including *Pto* and *Fen*. Neutrality test results were within the range of the reference loci at the *Fen* gene and slightly elevated, but not positive at the *Prf* locus (Table 3.3).

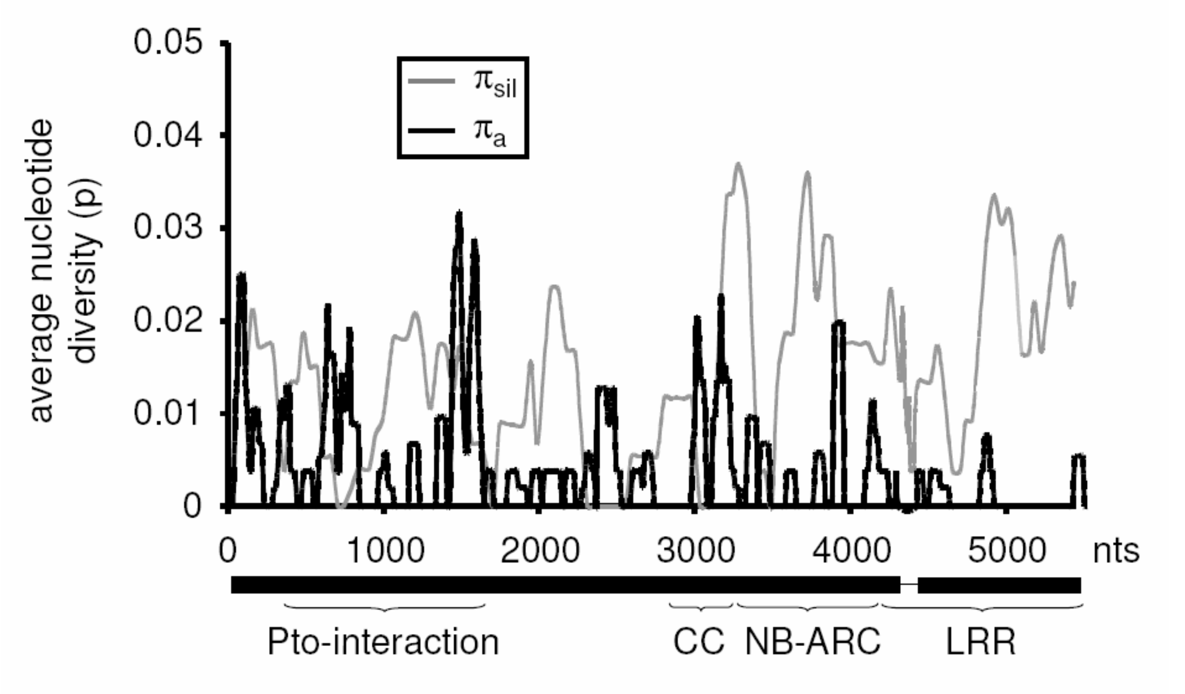
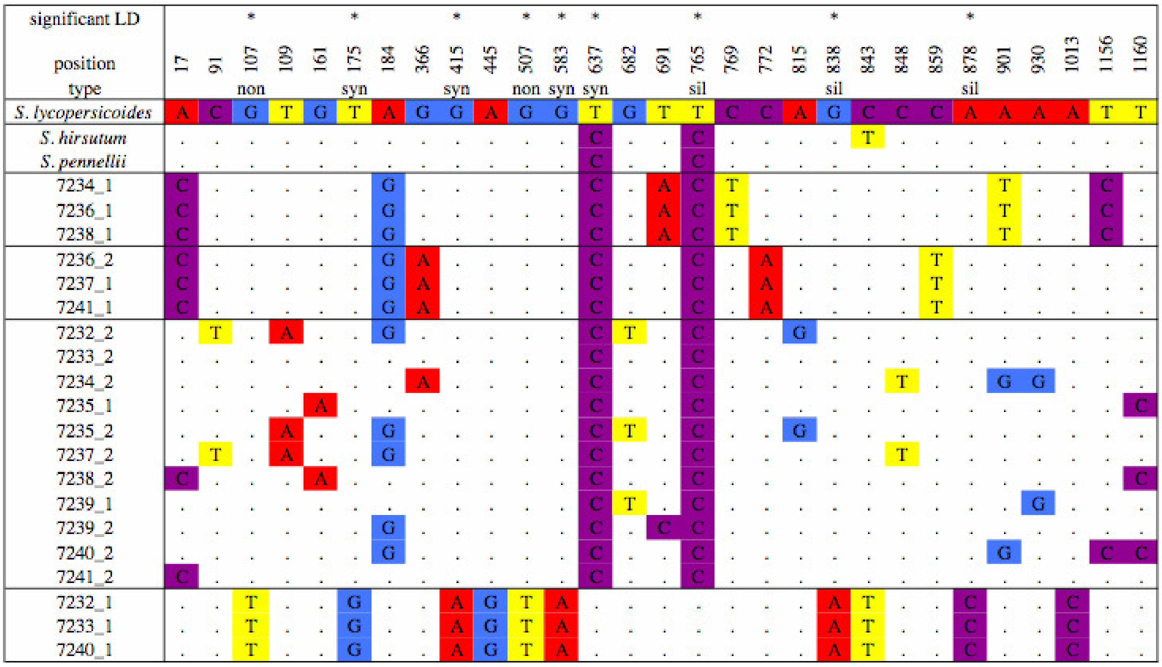


Figure 3.19: Sliding window of silent and nonsynonymous nucleotide diversity π in the Tarapaca population at the *Prf* gene. Values are midpoints of 50 bp windows. The sketch underneath represents the structure of the gene. Black boxes symbolize exons, black lines symbolize Introns. The putative functional regions are indicated below the appropriate exons.

3.3.3 Peculiar haplotype structure at *Rin4*

The gene showing the greatest level of evolutionary constraint is *Rin4*. This gene has the lowest level of nonsynonymous polymorphism and the lowest ratio of π_a to π_s of these five genes (0.16, Table 3.3, Figure 3.17). In fact, based on the distribution and levels of polymorphism, this gene appears indistinguishable from the 14 reference loci. However, in contrast to the set of reference loci, LD is strong at this locus (Figure 3.21). Elevated LD is caused in part by the presence of a mixture of sequence types found either only a single time in the sample or found in three different individuals. Each individual in this sample was heterozygous at *Rin4* and the majority of the individuals (8/10) have one allele that is common (present three times in the sample) and one allele that is found only once in the sample (Figures 3.20 and 3.22). Collectively, these groups containing identical sequence types show multiple fixed differences with respect to the other alleles. In particular the group of alleles, 7232.1, 7233.1, and 7240.1, show nine fixed differences relative to the other alleles. When excluding singleton polymorphisms, these nine positions are in significant linkage disequilibrium and, are derived state in alleles 7232.1, 7233.1 and 7240.1 relative to the outgroup species, *S. hirsutum* and *S. pennellii*. Seven of nine of these changes are derived relative to the more distantly related outgroup, *S. lycopersicoides*. These nine fixed differences are distributed throughout the *Rin4* coding sequence. Two out of nine of these fixed, derived differences are nonsynonymous, while the others are either synonymous or silent. The absence of evidence of recombination between this sequence type and the others, the strong pattern of linkage disequilibrium involving derived changes, two of which are nonsynonymous, and the low to moderate frequency of this sequence type, is consistent with the potential presence of a partial or ongoing sweep at *Rin4*. This assumption is supported by the finding that both Tajima's *D* and Fu and Li's *D* are lower than the reference loci means at the *Rin4* gene.



significant LD			*			*			*		*	*	*			*				*				*					
position	17	91	107	109	161	175	184	366	415	445	507	583	637	682	691	765	769	772	815	838	843	848	859	878	901	930	1013	1156	1160
type			non			syn			syn		non	syn	syn			sil				sil				sil					
S. lycopersicoides	A	C	G	T	G	T	A	G	G	A	G	G	T	G	T	T	C	C	A	G	C	C	C	A	A	A	A	T	T
S. hirsutum	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	.	.	.	T	.	.	.	.	.	.	.	.
S. pennellii	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.
7234_1	C	.	.	.	.	.	G	.	.	.	.	.	C	.	A	C	T	.	.	.	.	.	.	.	T	.	.	C	.
7236_1	C	.	.	.	.	.	G	.	.	.	.	.	C	.	A	C	T	.	.	.	.	.	.	.	T	.	.	C	.
7238_1	C	.	.	.	.	.	G	.	.	.	.	.	C	.	A	C	T	.	.	.	.	.	.	.	T	.	.	C	.
7236_2	C	.	.	.	.	.	G	A	.	.	.	.	C	.	.	C	.	A	.	.	.	.	T	.	.	.	.	.	.
7237_1	C	.	.	.	.	.	G	A	.	.	.	.	C	.	.	C	.	A	.	.	.	.	T	.	.	.	.	.	.
7241_1	C	.	.	.	.	.	G	A	.	.	.	.	C	.	.	C	.	A	.	.	.	.	T	.	.	.	.	.	.
7232_2	.	T	.	A	.	.	G	.	.	.	.	.	C	T	.	C	.	.	G	.	.	.	.	.	.	.	.	.	.
7233_2	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.
7234_2	.	.	.	.	.	.	.	A	.	.	.	.	C	.	.	C	.	.	.	.	.	T	.	.	G	G	.	.	.
7235_1	.	.	.	.	A	.	.	.	.	.	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	C
7235_2	.	.	.	A	.	.	G	.	.	.	.	.	C	T	.	C	.	.	G	.	.	.	.	.	.	.	.	.	.
7237_2	.	T	.	A	.	.	G	.	.	.	.	.	C	.	.	C	.	.	.	.	.	T	.	.	.	.	.	.	.
7238_2	C	.	.	.	A	.	.	.	.	.	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	C
7239_1	.	.	.	.	.	.	.	.	.	.	.	.	C	T	.	C	.	.	.	.	.	.	.	.	.	G	.	.	.
7239_2	.	.	.	.	.	.	G	.	.	.	.	.	C	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.
7240_2	.	.	.	.	.	.	G	.	.	.	.	.	C	.	.	C	.	.	.	.	.	.	.	.	G	.	.	C	C
7241_2	C	.	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.
7232_1	.	.	T	.	.	G	.	.	A	G	T	A	.	.	.	.	.	.	.	A	T	.	.	C	.	.	C	.	.
7233_1	.	.	T	.	.	G	.	.	A	G	T	A	.	.	.	.	.	.	.	A	T	.	.	C	.	.	C	.	.
7240_1	.	.	T	.	.	G	.	.	A	G	T	A	.	.	.	.	.	.	.	A	T	.	.	C	.	.	C	.	.

Figure 3.20: Distribution of non-singleton polymorphisms at the *Rin4* gene among individuals of *S. peruvianum*. Dots indicate positions matching the reference allele from *S. lycopersicoides*. Positions showing statistically significant LD (see Figure 3.21) are indicated along the top row with an asterix. The type of mutation (*i.e.* synonymous or nonsynonymous) of these positions is indicated in the third row. The upper three rows contain the nucleotide states of three outgroups of *S. peruvianum* at these same positions. Note that nearly all positions in significant linkage equilibrium are found in the 'derived' state for the first three alleles in the table (7232_1, 7233_1 and 7240_1).

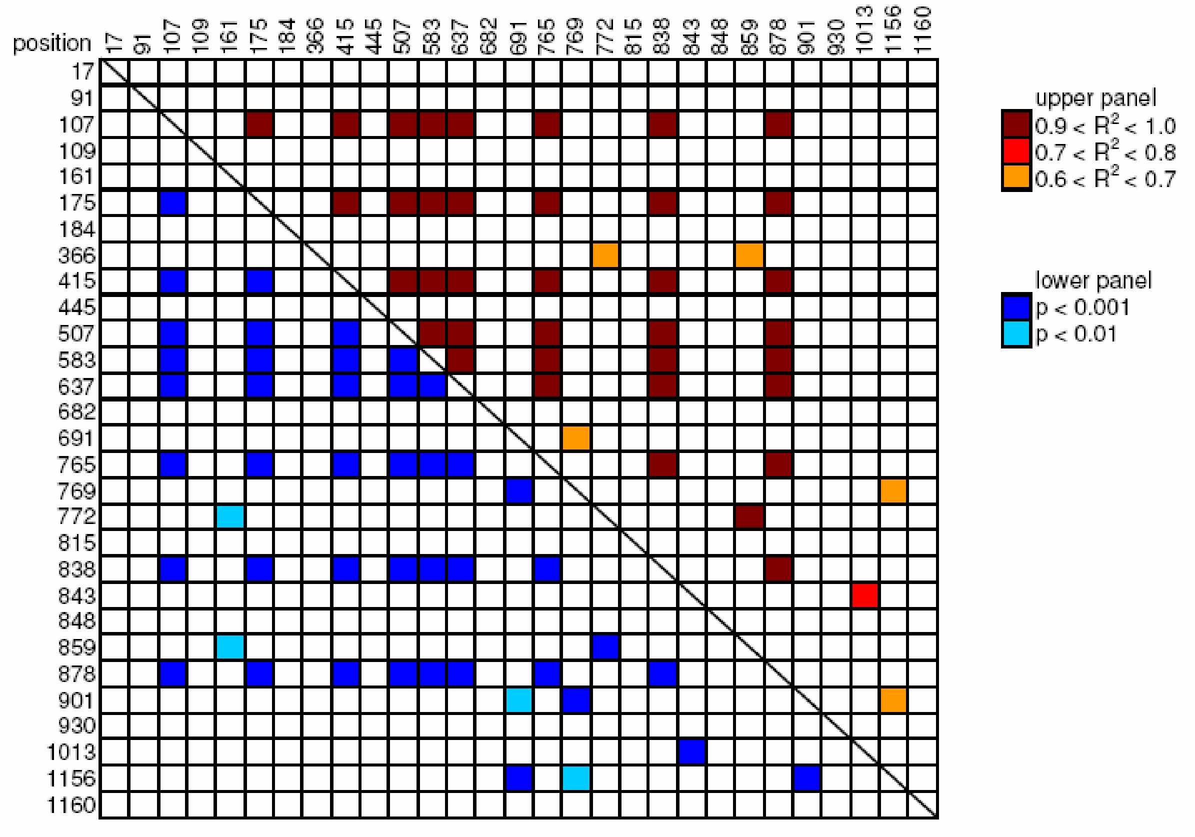


Figure 3.21: Pattern of linkage disequilibrium among *Rin4* alleles. R^2 values greater than 0.6 are highlighted in the upper right panel. The statistical significance (associated *P*-values) for these R^2 values are highlighted in the lower left panel.

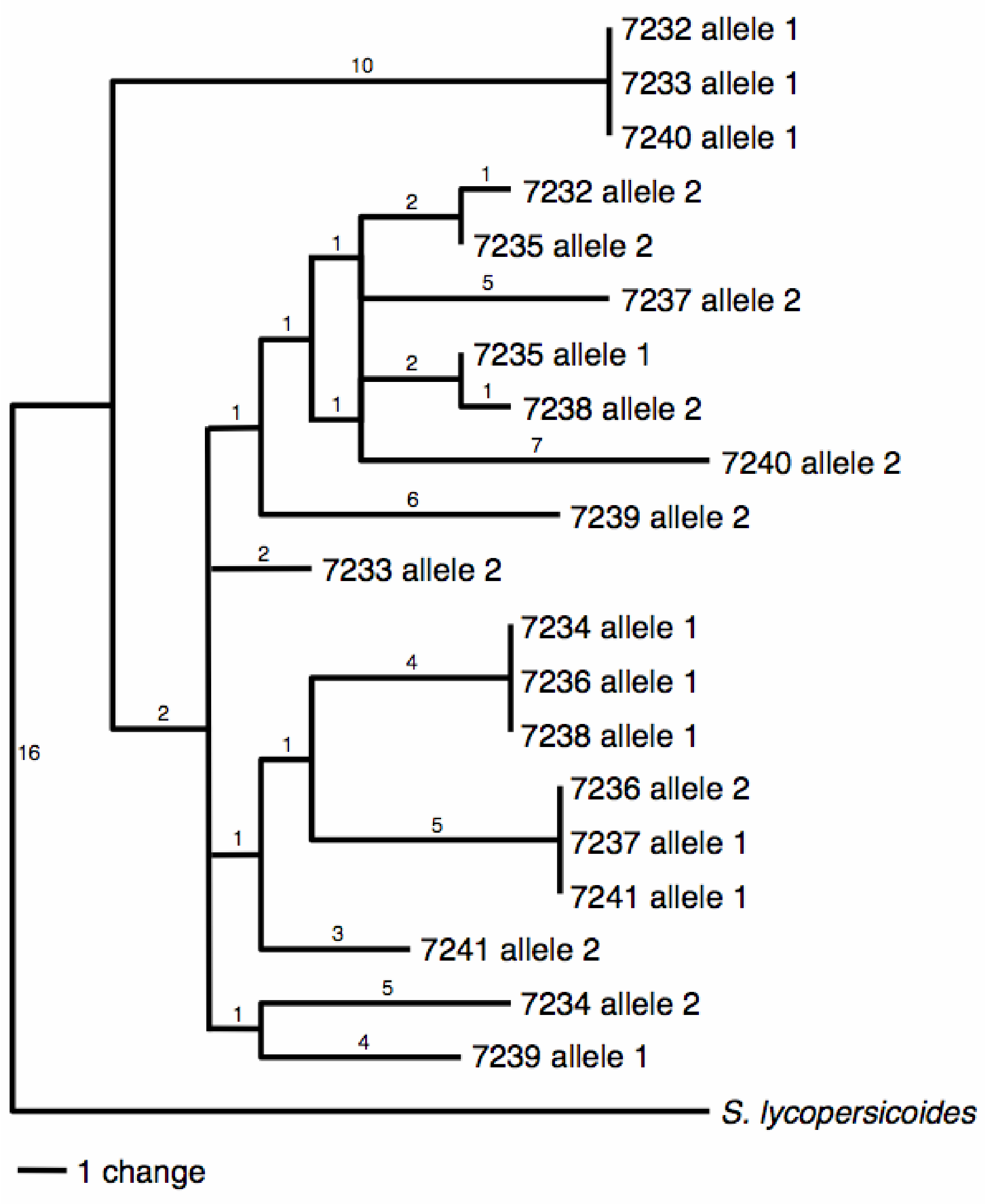


Figure 3.22: Maximum parsimony tree based on nucleotide sequences of *Rin4* alleles from *S. peruvianum*. The tree was rooted with the *Rin4* sequence from the outgroup *S. lycopersicoides*. Branch lengths are indicated above the branches.

3.4 Evolution of resistance genes beyond species boundaries: adaptive introgression vs. ancestral polymorphism

Genes displaying an interesting pattern in the population-based study (*Pto*, *Pfi* and *Rin4*), were further investigated using a species wide sample approach. For this purpose, single alleles from several accessions covering the whole species range of the three wild tomato species *S. peruvianum*, *S. chilense* and *S. corneliomulleri* were sequenced and analyzed.

3.4.1 Phylogenetic relationships between species

The phylogenetic analysis of all nine reference loci concatenated into one single alignment reveals a separation between *S. peruvianum* and *S. chilense* (bootstrap support of 90 and 100%, Figure 3.23). Alleles originating from *S. corneliomulleri* however cluster mainly together with *S. peruvianum*, but can occasionally also fall into a clade with *S. chilense* (LA 2759 and LA 3666). Phylogenetic relationships at the *Pto* gene are different from the reference loci, as no clear distinction between *S. peruvianum* and *S. chilense* is visible (Figure 3.24). This pattern becomes even stronger in a phylogeny of the protein sequence (Figure 3.38). Alleles from *S. peruvianum* form two main groups, only one of which is shared by alleles originating from *S. chilense* and *S. corneliomulleri*. This group is again subdivided into two smaller clusters, which share alleles from all three species. The phylogeny of the *Rin4* gene does not reveal any special pattern (Figure 3.25). Although alleles from the same species tend to be more closely related to one another than to alleles from another species, no significant distinction is visible between the different species, and identical alleles can be shared between species. Alleles at the *Pfi* locus are dispersed over two clades. Each of these clades consists of alleles from all three species. Again, the pattern is stronger in a phylogeny considering the *Pfi* protein sequence (Figure 3.26 and 3.41).

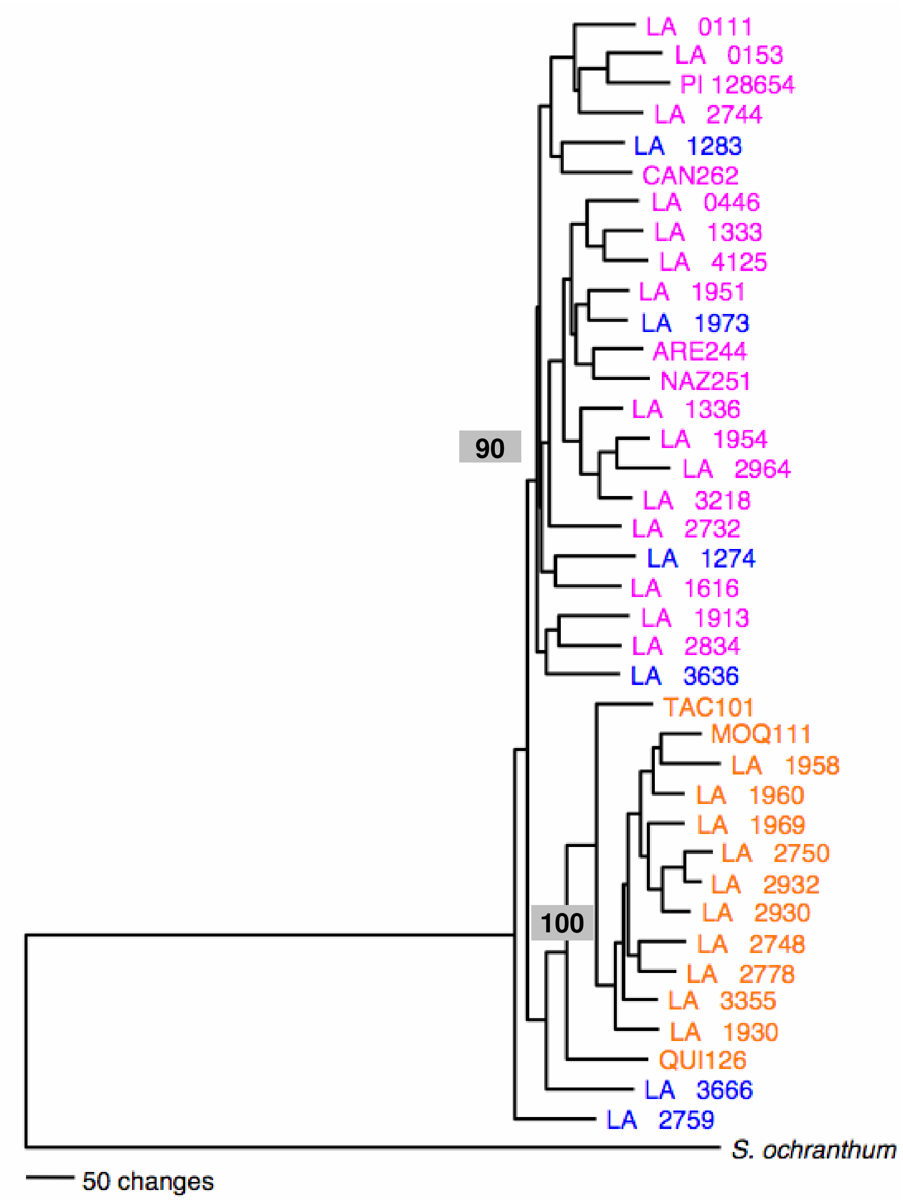


Figure 3.23: Maximum parsimony gene tree of all studied individuals at the reference loci. The phylogeny is based on all nine reference loci, which have been concatenated into one large alignment. The root was defined by midpoint rooting. Pink taxon names indicate *S. peruvianum* alleles, orange taxon names indicate *S. chilense* alleles, blue taxon names indicate *S. corneliomulleri* alleles. Numbers of changes are indicated on the branches.

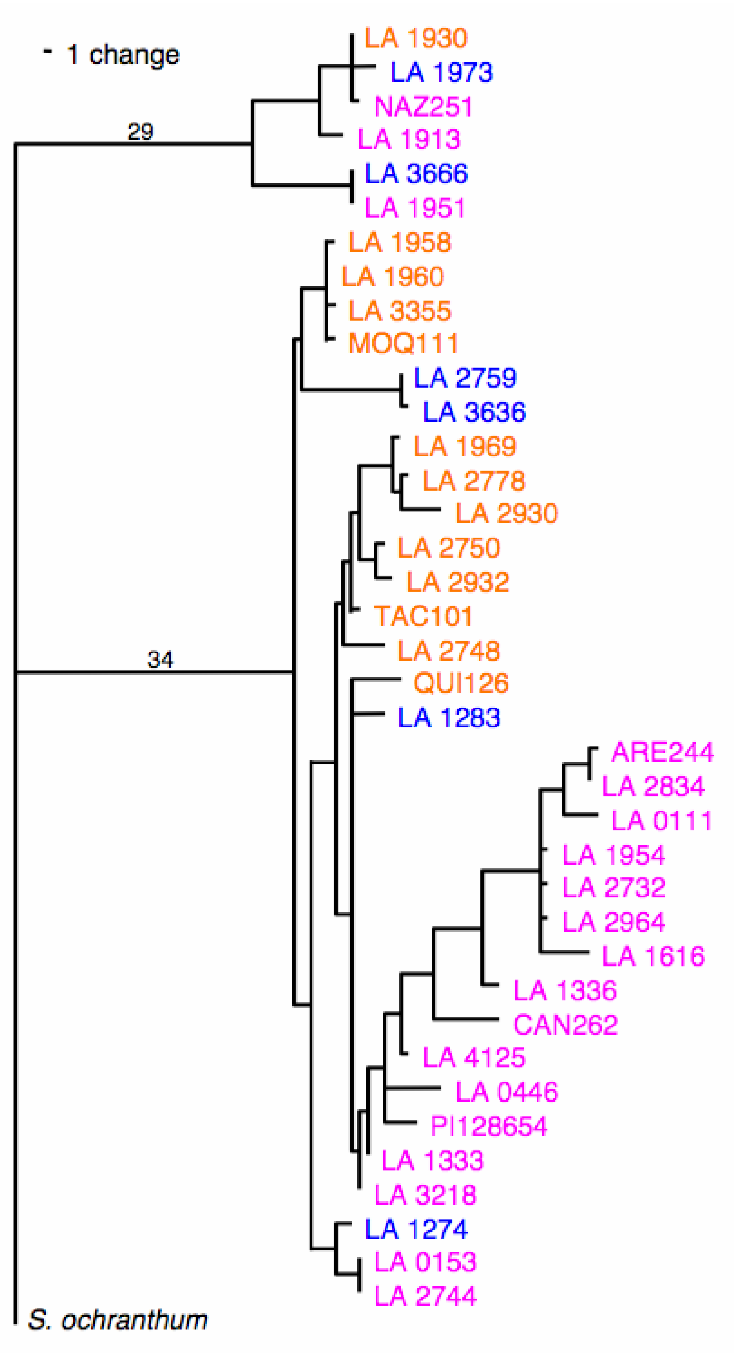


Figure 3.24: Maximum parsimony gene tree of all studied individuals at the *Pto* locus. The outgroup *S. ochranthum* was used to root the tree. Pink taxon names indicate *S. peruvianum* alleles, orange taxon names indicate *S. chilense* alleles, blue taxon names indicate *S. corneliomulleri* alleles. Numbers of changes are indicated on the branches.

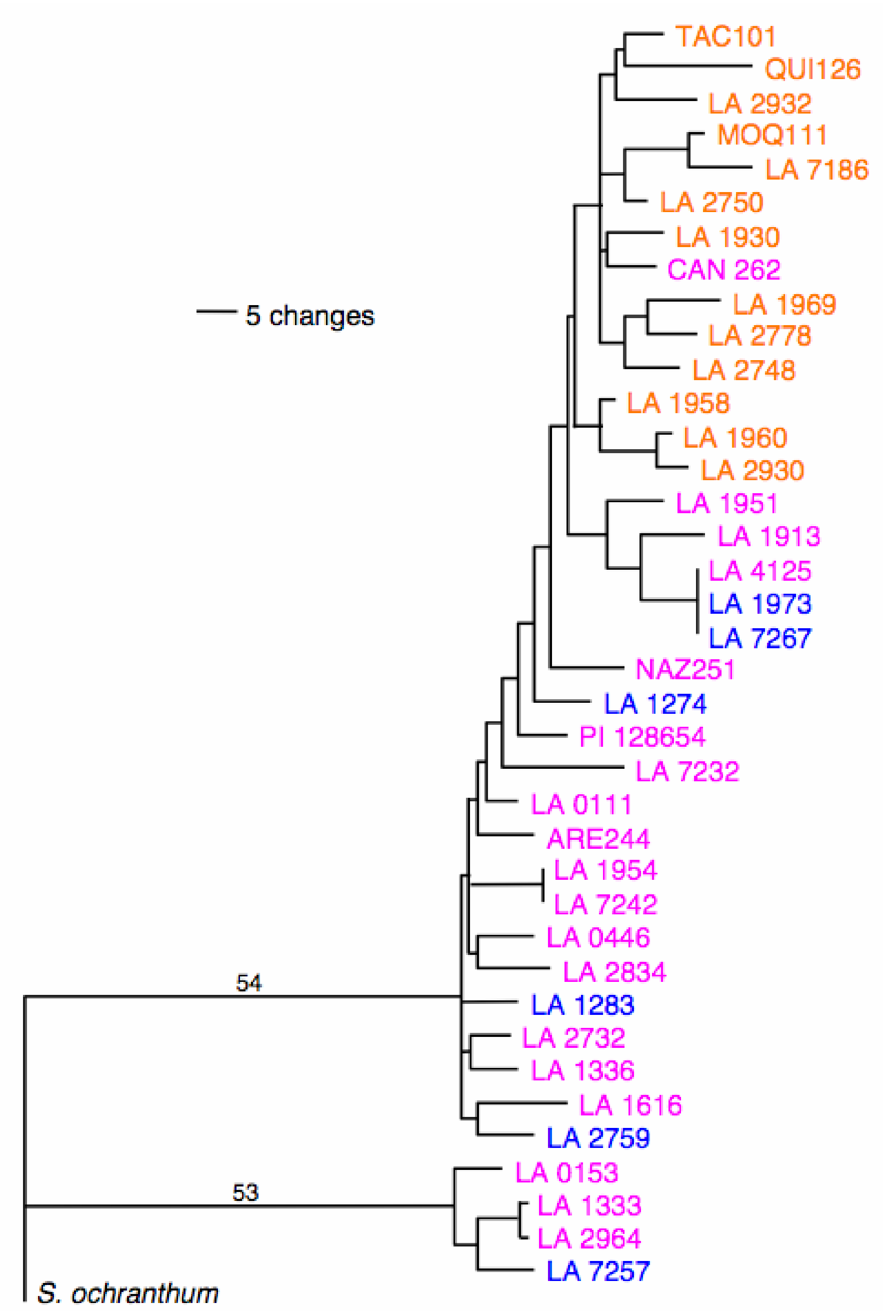


Figure 3.25: Maximum parsimony gene tree of all studied individuals at the *Rin4* locus. The outgroup *S. ochranthum* was used to root the tree. Pink taxon names indicate *S. peruvianum* alleles, orange taxon names indicate *S. chilense* alleles, blue taxon names indicate *S. corneliomulleri* alleles. Numbers of changes are indicated on the branches.

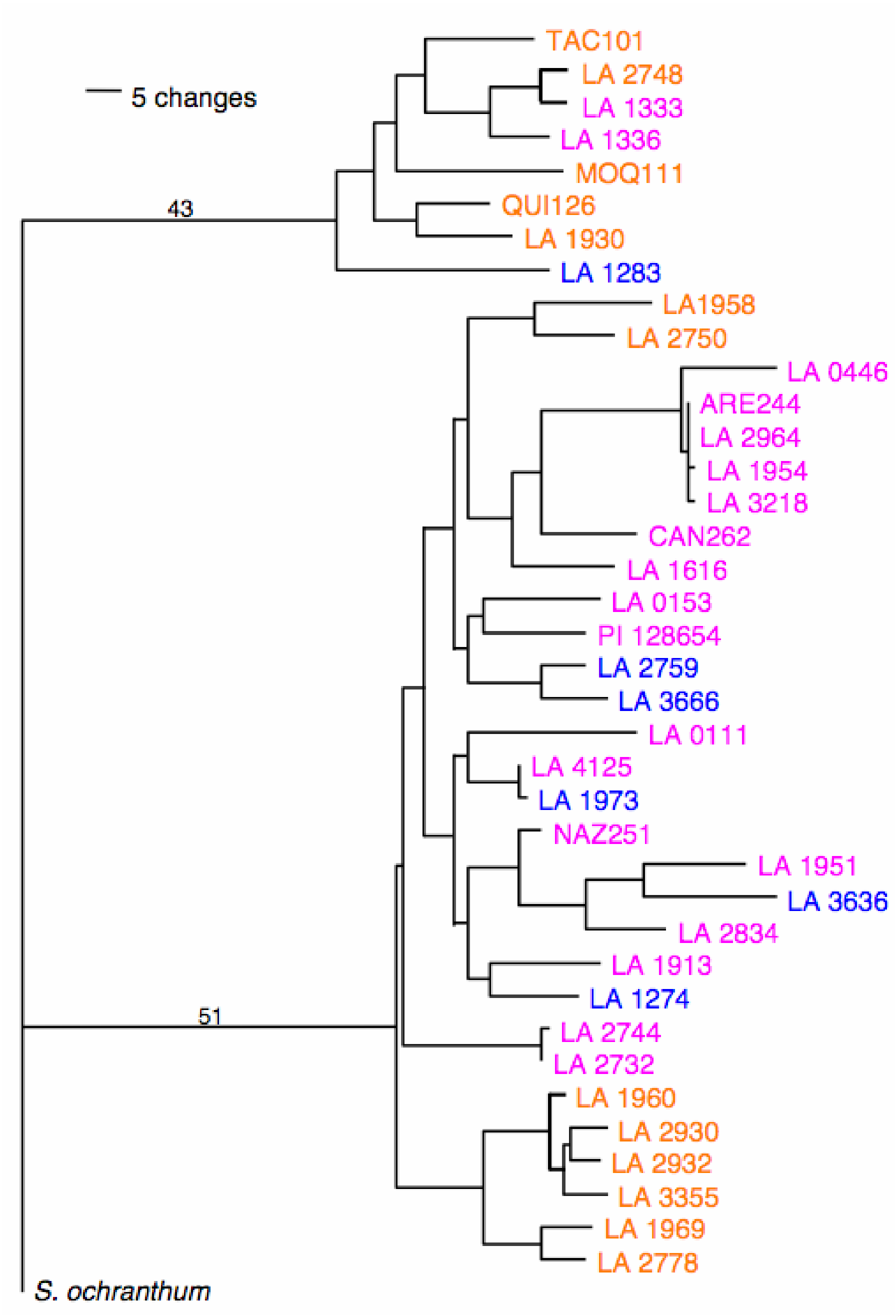


Figure 3.26: Maximum parsimony gene tree of all studied individuals at the *Pfi* locus. The outgroup *S. ochranthum* was used to root the tree. Pink taxon names indicate *S. peruvianum* alleles, orange taxon names indicate *S. chilense* alleles, blue taxon names indicate *S. corneliomulleri* alleles. Numbers of changes are indicated on the branches.

3.4.2 Within species diversity

Nucleotide diversity at the reference genes

The average pairwise nucleotide diversity π measured over the nine reference loci is 0.00954 in *S. chilense* (Table 3.6). With an average pairwise nucleotide diversity π of 0.013 at the reference loci, *S. peruvianum* is more polymorphic than *S. chilense*. In *S. corneliomulleri*, π of 0.015 at the reference loci is similar to the value measured in *S. peruvianum*. The synonymous nucleotide diversity π_s lies usually between 0.02 and 0.03, while the nonsynonymous nucleotide diversity π_a is much lower (between 0.002 and 0.003). The ratio of π_a to π_s is therefore low with values comprised between 0.085 and 0.115 (Figure 3.27).

Nucleotide diversity at the *Pto* locus

In *S. chilense*, the *Pto* gene exhibits similar levels of variability to the reference loci (Table 3.6). The synonymous nucleotide diversity π_s at the *Pto* gene behaves like the mean π_s at the reference loci, while the mean π_a is four times the value at the reference loci. The ratio of π_a to π_s is therefore four times elevated compared to the reference. A similar pattern is found in *S. peruvianum*. π_s at the *Pto* locus behaves like the mean π_s at the reference loci, while the mean π_a is six times π_a at the reference loci. In *S. corneliomulleri*, diversity at the *Pto* locus is comparable to *S. peruvianum*: π_s is nearly halved compared to π_s at the reference loci, π_a is ten times elevated compared to the reference, and therefore the ratio of π_a to π_s is higher than one (Figure 3.27).

Nucleotide diversity at the *Rin4* locus

The *Rin4* gene exhibits similar levels of variability compared to the reference loci in all three species (Table 3.6, Figure 3.27). Merely, π_a is slightly elevated compared to the reference loci, while π_s behaves like the genomic mean, leading to the ratio being slightly elevated.

Nucleotide diversity at the *Pfi* locus

In *S. chilense*, nucleotide diversity at the *Pfi* gene is approximately twice the genomic average (Table 3.6), with π_a being seven fold increased compared to the genomic mean. The ratio is therefore five times the ratio measured at the reference loci (Figure 3.27). In *S. peruvianum* and in *S. corneliomulleri*, nucleotide diversity at the *Pfi* gene is also elevated compared to the genomic mean: π_s is elevated in *S. peruvianum*, and π_a is five fold increased compared to the genomic mean. The ratio is therefore five times the ratio at the reference loci. The pattern observed at the *Pfi* gene in *S. corneliomulleri* is similar to the observations in *S. peruvianum*.

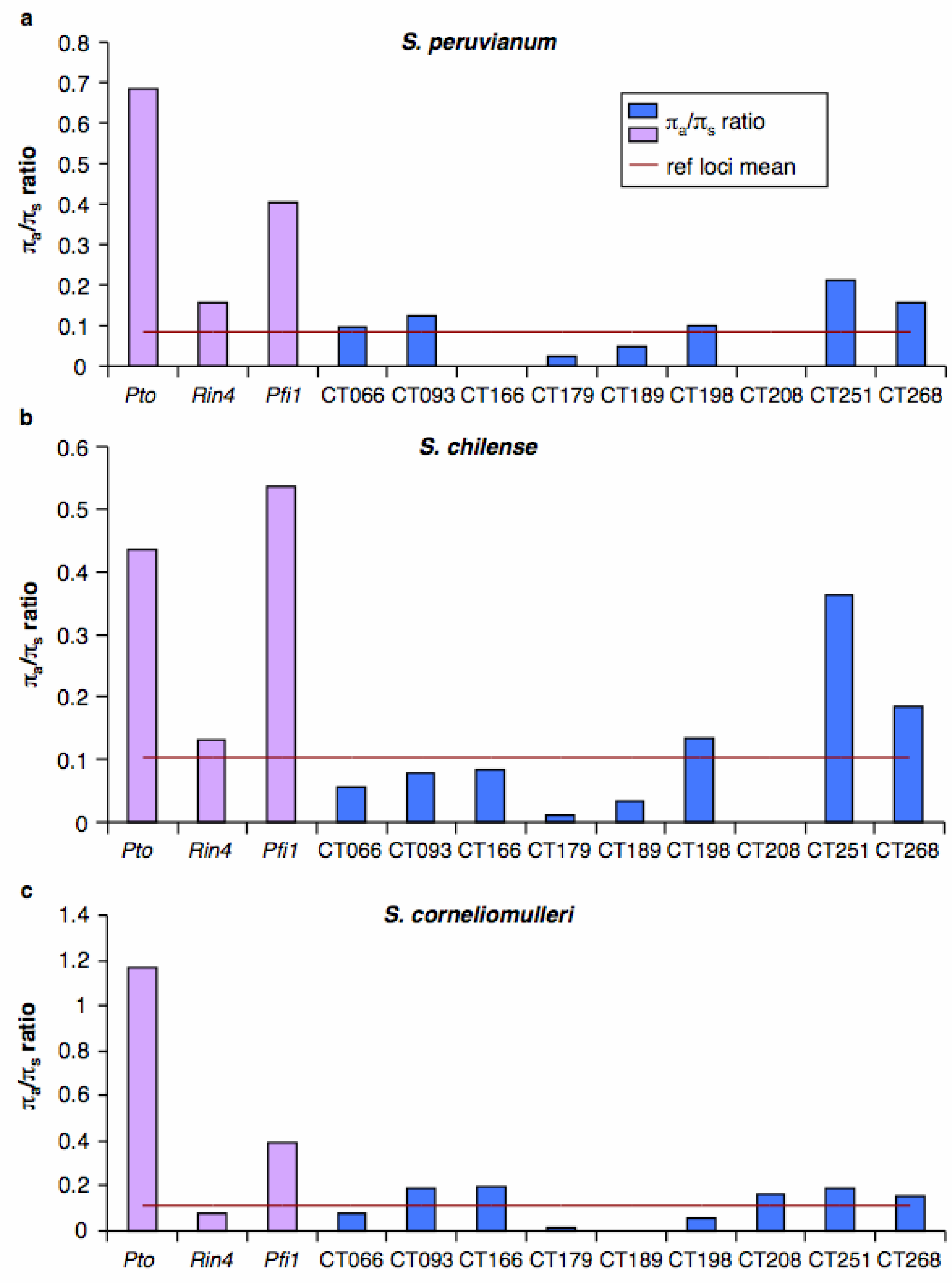


Figure 3.27: Ratios of π_a to π_s at all reference (blue) and candidate genes (purple). The mean (red line) was calculated using the nine reference loci. **a**, Ratios in *S. peruvianum*. **b**, Ratios in *S. chilense*, **c**, Ratios in *S. corneliomulleri*.

Neutrality tests at the reference loci

In *S. chilense*, the average Tajima's *D* measured at the reference loci over all sites is -0.828, while the value measured at synonymous sites is -0.298. Fu and Li's *D* is -0.811 (Table 3.6, Figures 3.28). According to the synonymous site frequency spectrum at the reference genes, there are fewer polymorphisms in all mutational classes than expected under a neutral scenario (Figure 3.29b). The average Tajima's *D* measured at the reference loci over all sites is -1.226 and the value measured over synonymous sites is -1.129 in *S. peruvianum*. Fu and Li's *D* is -1.770. According to the synonymous site frequency spectrum at the reference genes, there is an excess of low and high frequency polymorphisms compared to the expectation under a neutral scenario (Figure 3.29a). In *S. corneliomulleri*, the average Tajima's *D* measured at the reference loci over all sites is -0.665 and over synonymous sites -0.407. Fu and Li's *D* is -0.733. According to the synonymous site frequency spectrum at the reference genes, there are fewer polymorphisms in all mutational classes than expected under a neutral scenario (Figure 3.29c).

Neutrality tests at the *Pto* locus

In *S. chilense*, *Pto* shows elevated Tajima's *D* and Fu and Li's *D* values with -0.336 and -0.643 at all sites compared to the genomic mean (Table 3.6, Figure 3.28). The site frequency spectrum reveals that *Pto* exhibits an excess of intermediate frequency polymorphisms compared to the mean at synonymous sites measured at the reference loci (Figure 3.29b). This pattern of polymorphism frequencies at the *Pto* locus is more pronounced in *S. peruvianum*, where Tajima's *D* and Fu and Li's *D* are positive (0.436 and 0.119) and the site frequency spectrum reveals an excess of intermediate frequency polymorphisms compared to the reference loci (Figure 3.29a). The pattern in *S. corneliomulleri* is similar to the observed pattern in *S. peruvianum* ($T_D = 0.584$, $FL_D = 0.527$). McDonald-Kreitman test results were not significant and did not reveal any excess of polymorphic sites within species compared to both outgroups *S. ochranthum* and *S. pennellii*.

Neutrality tests at the *Pfi* locus

In *S. chilense*, *Pfi* shows elevated Tajima's *D* and Fu and Li's *D* values compared to the genomic mean with -0.297 and -0.610 (Table 3.6, Figure 3.28). The site frequency spectrum reveals that this gene exhibits an excess of intermediate frequency polymorphisms compared to the mean at synonymous sites measured at the reference loci (Figure 3.29b). In *S. peruvianum*, Tajima's *D* and Fu and Li's *D* are only slightly elevated compared to the reference dataset (-0.814 and -1.266). However, the frequency spectrum reveals an excess of intermediate frequency polymorphisms here as well (Figure 3.29a). The pattern in *S. corneliomulleri* is again similar to the pattern observed in *S. peruvianum* (T_D = -0.384, FL_D = -0.471). McDonald-Kreitman test results were not significant at the *Pfi* locus. However, amongst polymorphisms in *S. chilense*, which are derived relative to the outgroup *S. ochranthum*, there is a tendency towards an excess of nonsynonymous compared to synonymous polymorphisms.

Neutrality tests at the *Rin4* locus

Tajima's *D* and Fu and Li's *D* at the *Rin4* gene are more negative in *S. chilense* compared to the genome average (-1.104 and -0.995) (Table 3.6, Figure 3.28). The frequency spectrum behaves similarly to the neutral expectation in this species, but shows an excess of singletons (Figure 3.29b). Tajima's *D* and Fu and Li's *D* in *S. peruvianum* are more negative than the genome average as well (-1.290 and -1.337). The frequency spectrum reveals an excess of low and high frequency variance compared to the frequency spectrum measured at the reference genes (Figure 3.29a). In *S. corneliomulleri* however, *Rin4* exhibits slightly elevated values of T_D = -0.305 and FL_D = -0.261 compared to the reference dataset. The frequency spectrum behaves similarly to the neutral expectation in this species (Figure 3.29c). McDonald-Kreitman test results were not significant at the *Rin4* locus. However, when comparing *S. peruvianum* to the outgroup *S. ochranthum*, a greater proportion of nonsynonymous than synonymous polymorphisms are found (P = 0.052).

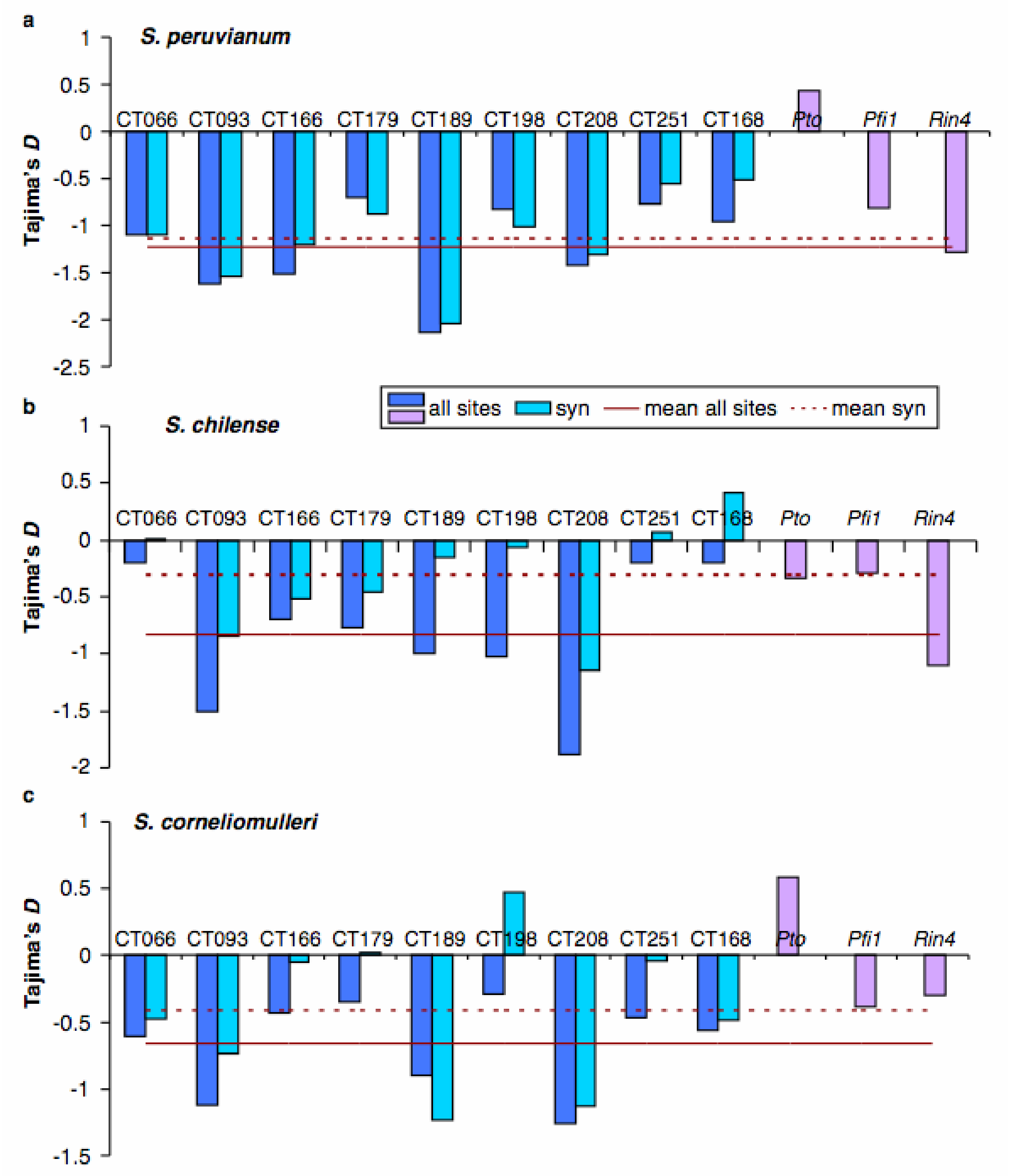


Figure 3.28: Tajima's *D* values at all reference (blue bars) and candidate genes (purple bars). For the reference loci, Tajima's *D* is shown at all and synonymous sites. At the candidate genes only at all sites. Mean values over the nine reference loci are shown for all (red solid line) and synonymous sites (red dashed line). **a**, Tajima's *D* in *S. peruvianum*. **b**, Tajima's *D* in *S. chilense*. **c**, Tajima's *D* in *S. corneliomulleri*.

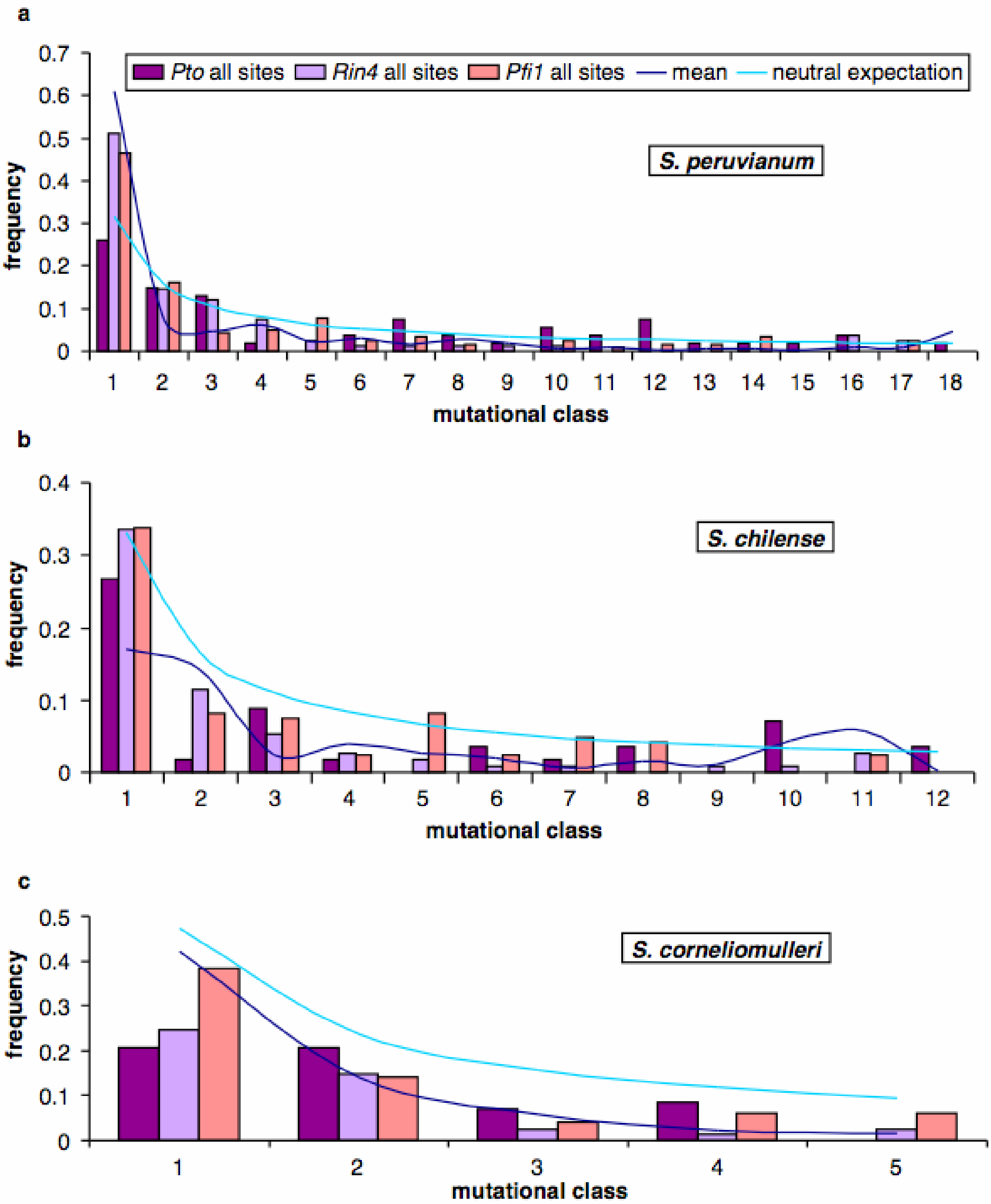


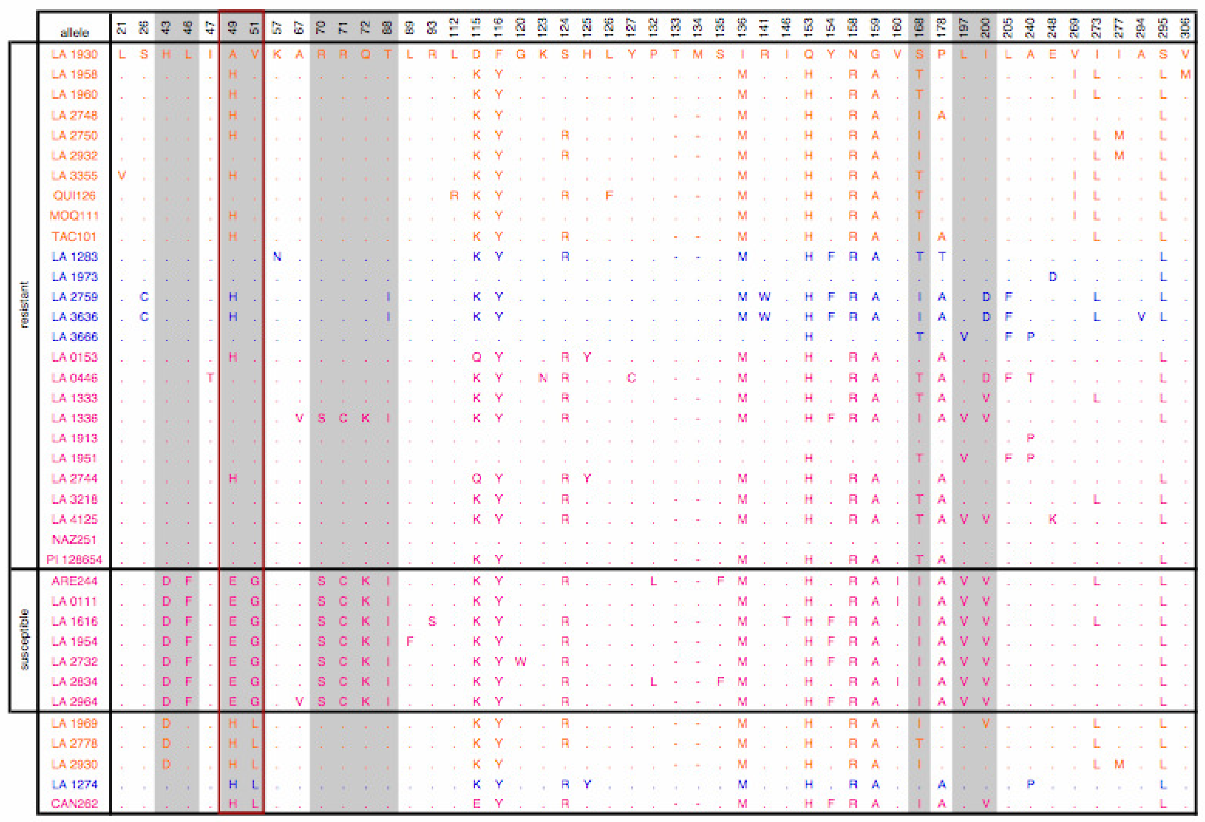
Figure 3.29: Site frequency spectra of the three resistance genes in comparison to the neutral expectation and the frequency spectrum at synonymous sites at the reference loci. The frequency of mutations occurring at a certain number in the dataset (mutational class) is blotted for each mutational class. The light blue line indicates the expectation under complete neutrality. The dark blue line represents the mean frequency spectrum at synonymous sites measured over all nine reference loci and thus reflects only demography in the given species. **a**, Frequency spectrum in *S. peruvianum*. **b**, Frequency spectrum in *S. chilense*. **c**, Frequency spectrum in *S. corneliomulleri*.

Table 3.6. Overview of summary statistics for all genes in all three species. The mean was calculated using the nine reference loci. T_D = Tajima's *D*, FL_D = Fu and Li's *D*

Locus	**θ**	**π**	**π_s**	**π_a**	**π_a/π_s**	**T_D**		**FL_D**
						all sites	**syn sites**	
S. peruvianum								
CT066	0.016	0.011	0.035	0.003	0.096	-1.089	-1.097	-1.869
CT093	0.011	0.006	0.015	0.002	0.124	-1.626	-1.536	-2.255
CT166	0.017	0.010	0.011	0.000	0.000	-1.515	-1.203	-1.986
CT179	0.026	0.020	0.053	0.001	0.023	-0.698	-0.885	-1.182
CT189	0.010	0.005	0.008	0.000	0.050	-2.130	-2.046	-3.292
CT198	0.033	0.025	0.044	0.004	0.099	-0.825	-1.014	-0.998
CT208	0.023	0.015	0.017	0.000	0.000	-1.415	-1.314	-1.353
CT251	0.019	0.015	0.036	0.008	0.211	-0.776	-0.552	-1.319
CT268	0.017	0.013	0.035	0.006	0.158	-0.957	-0.512	-1.678
reference loci mean	**0.019**	**0.013**	**0.028**	**0.003**	**0.085**	**-1.226**	**-1.129**	**-1.770**
Pto	**0.020**	**0.020**	**0.027**	**0.018**	**0.684**	**0.436**	**0.111**	**0.119**
Rin4	**0.019**	**0.013**	**0.026**	**0.004**	**0.157**	**-1.290**	**-0.792**	**-1.337**
Pfi	**0.028**	**0.022**	**0.033**	**0.013**	**0.403**	**-0.814**	**-0.795**	**-1.266**
S. chilense								
CT066	0.012	0.011	0.038	0.002	0.055	-0.194	0.010	-0.859
CT093	0.006	0.004	0.010	0.001	0.077	-1.502	-0.848	-1.810
CT166	0.017	0.014	0.018	0.002	0.085	-0.688	-0.510	-0.982
CT179	0.018	0.015	0.035	0.000	0.011	-0.766	-0.448	-1.004
CT189	0.010	0.007	0.017	0.001	0.033	-0.994	-0.161	-0.038
CT198	0.011	0.009	0.012	0.002	0.134	-1.032	-0.059	-0.10
CT208	0.008	0.005	0.001	0.000	0.000	-1.880	-1.149	-2.349
CT251	0.011	0.011	0.021	0.007	0.362	-0.197	0.066	0.450
CT268	0.011	0.011	0.028	0.005	0.184	-0.201	0.418	-0.608
reference loci mean	**0.012**	**0.010**	**0.020**	**0.002**	**0.105**	**-0.828**	**-0.298**	**-0.811**
Pto	**0.013**	**0.011**	**0.020**	**0.009**	**0.434**	**-0.336**	**0.662**	**-0.643**
Rin4	**0.018**	**0.013**	**0.028**	**0.004**	**0.131**	**-1.104**	**-0.560**	**-0.995**
Pfi	**0.024**	**0.022**	**0.029**	**0.016**	**0.536**	**-0.297**	**0.379**	**-0.610**
S. corneliomulleri								
CT066	0.012	0.011	0.035	0.003	0.074	-0.608	-0.476	-0.689
CT093	0.008	0.006	0.012	0.002	0.192	-1.126	-0.735	-1.116
CT166	0.019	0.017	0.011	0.002	0.197	-0.427	-0.050	-0.635
CT179	0.020	0.018	0.053	0.001	0.015	-0.348	0.018	-0.436
CT189	0.013	0.011	0.012	0.000	0.000	-0.901	-1.233	-0.859
CT198	0.030	0.028	0.080	0.005	0.055	-0.291	0.472	-0.398
CT208	0.016	0.013	0.006	0.001	0.158	-1.257	-1.132	-1.317
CT251	0.015	0.014	0.030	0.006	0.192	-0.471	-0.046	-0.410
CT268	0.015	0.013	0.038	0.006	0.152	-0.557	-0.482	-0.733
reference loci mean	**0.016**	**0.015**	**0.031**	**0.003**	**0.115**	**-0.665**	**-0.407**	**-0.733**
Pto	**0.019**	**0.020**	**0.018**	**0.021**	**1.170**	**0.584**	**-0.106**	**0.527**
Rin4	**0.013**	**0.012**	**0.024**	**0.002**	**0.076**	**-0.305**	**-0.387**	**-0.261**
Pfi	**0.025**	**0.023**	**0.035**	**0.014**	**0.389**	**-0.384**	**-0.637**	**-0.471**

3.4.3 Haplotype structure at the *Pto* locus

A closer inspection of the *Pto* alleles reveals a distinct haplotype structure, which correlates with a classification of *Pto* alleles proposed by Xing *et al.* (2007) based on putative functional polymorphisms. Xing *et al.* (2007) describe two amino acid residues which are important for recognition of AvrPto. According to their study, Pto alleles carrying the amino acids 49A/H and 51V are resistant, while alleles carrying 49E and 51G are susceptible (for details see Böndel (2010). The Pto sequence alignment reveals additional amino acid changes that are responsible for the haplotype structure and are in significant LD with one another and with these functionally important amino acid positions (Figures 3.30 and 3.31). Furthermore, these amino acid substitutions are located in close proximity to the AvrPto-interacting interface of Pto or structurally important domains of the protein (Figure 3.32).



	allele	21	26	43	46	47	49	51	57	67	70	71	72	88	89	93	112	115	116	120	123	124	125	126	127	132	133	134	135	136	141	148	153	154	158	159	160	168	178	197	200	205	240	248	269	273	277	294	295	306
resistant	LA 1930	L	S	H	L	I	A	V	K	A	R	R	Q	T	L	R	L	D	F	G	K	S	H	L	Y	P	T	M	S	I	R	I	Q	Y	N	G	V	S	P	L	I	L	A	E	V	I	I	A	S	V
	LA 1958	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	K	Y	.	.	.	.	.	.	.	.	.	.	M	.	.	H	.	R	A	.	T	.	.	.	.	.	.	I	L	.	.	L	M
	LA 1960	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	K	Y	.	.	.	.	.	.	.	.	.	.	M	.	.	H	.	R	A	.	T	.	.	.	.	.	.	I	L	.	.	L	.
	LA 2748	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	K	Y	.	.	.	.	.	.	.	-	-	.	M	.	.	H	.	R	A	.	I	A	.	.	.	.	.	.	.	.	.	L	.
	LA 2750	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	.	R	A	.	I	.	.	.	.	.	.	.	L	M	.	L	.
	LA 2932	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	.	R	A	.	I	.	.	.	.	.	.	.	L	M	.	L	.
	LA 3355	V	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	K	Y	.	.	.	.	.	.	.	.	.	.	M	.	.	H	.	R	A	.	T	.	.	.	.	.	.	I	L	.	.	L	.
	QUI126	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	R	K	Y	.	.	R	.	F	.	.	-	-	.	M	.	.	H	.	R	A	.	T	.	.	.	.	.	.	I	L	.	.	L	.
	MOQ111	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	K	Y	.	.	.	.	.	.	.	.	.	.	M	.	.	H	.	R	A	.	T	.	.	.	.	.	.	I	L	.	.	L	.
	TAC101	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	.	R	A	.	I	A	.	.	.	.	.	.	L	.	.	L	.
	LA 1283	.	.	.	.	.	.	.	N	.	.	.	.	.	.	.	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	F	R	A	.	T	T	.	.	.	.	.	.	.	.	.	L	.
	LA 1973	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	D	.	.	.	.	L	.
	LA 2759	.	C	.	.	.	H	.	.	.	.	.	.	I	.	.	.	K	Y	.	.	.	.	.	.	.	.	.	.	M	W	.	H	F	R	A	.	I	A	.	D	F	.	.	.	L	.	.	L	.
	LA 3636	.	C	.	.	.	H	.	.	.	.	.	.	I	.	.	.	K	Y	.	.	.	.	.	.	.	.	.	.	M	W	.	H	F	R	A	.	I	A	.	D	F	.	.	.	L	.	V	L	.
	LA 3666	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	.	.	.	.	T	.	V	.	F	P	.	.	.	.	.	.	.
	LA 0153	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	Q	Y	.	.	R	Y	.	.	.	.	.	.	M	.	.	H	.	R	A	.	.	A	.	.	.	.	.	.	.	.	.	L	.
	LA 0446	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	K	Y	.	N	R	.	.	C	.	-	-	.	M	.	.	H	.	R	A	.	T	A	.	D	F	T	.	.	.	.	.	L	.
	LA 1333	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	.	R	A	.	T	A	.	V	.	.	.	.	L	.	.	L	.
	LA 1336	.	.	.	.	.	.	.	.	V	S	C	K	I	.	.	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	F	R	A	.	I	A	V	V	.	.	.	.	.	.	.	L	.
	LA 1913	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	P	.	.	.	.	.	.	.
	LA 1951	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H	.	.	.	.	T	.	V	.	F	P	.	.	.	.	.	.	.
	LA 2744	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	Q	Y	.	.	R	Y	.	.	.	.	.	.	M	.	.	H	.	R	A	.	.	A	.	.	.	.	.	.	.	.	.	L	.
	LA 3218	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	.	R	A	.	T	A	.	.	.	.	.	.	L	.	.	L	.
	LA 4125	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	.	R	A	.	T	A	V	V	.	.	K	.	.	.	.	L	.
	NAZ251	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	PI 128654	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	K	Y	.	.	.	.	.	.	.	-	-	.	M	.	.	H	.	R	A	.	T	A	.	.	.	.	.	.	.	.	.	L	.
susceptible	ARE244	.	.	D	F	.	E	G	.	.	S	C	K	I	.	.	.	K	Y	.	.	R	.	.	.	L	-	-	F	M	.	.	H	.	R	A	I	I	A	V	V	.	.	.	.	L	.	.	L	.
	LA 0111	.	.	D	F	.	E	G	.	.	S	C	K	I	.	.	.	K	Y	.	.	.	.	.	.	.	.	.	.	M	.	.	H	.	R	A	I	I	A	V	V	.	.	.	.	.	.	.	L	.
	LA 1616	.	.	D	F	.	E	G	.	.	S	C	K	I	.	S	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	T	H	F	R	A	.	I	A	V	V	.	.	.	.	L	.	.	L	.
	LA 1954	.	.	D	F	.	E	G	.	.	S	C	K	I	F	.	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	F	R	A	.	I	A	V	V	.	.	.	.	.	.	.	L	.
	LA 2732	.	.	D	F	.	E	G	.	.	S	C	K	I	.	.	.	K	Y	W	.	R	.	.	.	.	-	-	.	M	.	.	H	F	R	A	.	I	A	V	V	.	.	.	.	.	.	.	L	.
	LA 2834	.	.	D	F	.	E	G	.	.	S	C	K	I	.	.	.	K	Y	.	.	R	.	.	.	L	-	-	F	M	.	.	H	.	R	A	I	I	A	V	V	.	.	.	.	.	.	.	L	.
	LA 2964	.	.	D	F	.	E	G	.	V	S	C	K	I	.	.	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	F	R	A	.	I	A	V	V	.	.	.	.	.	.	.	L	.
	LA 1969	.	.	D	.	.	H	L	.	.	.	.	.	.	.	.	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	.	R	A	.	I	.	.	V	.	.	.	.	L	.	.	L	.
	LA 2778	.	.	D	.	.	H	L	.	.	.	.	.	.	.	.	.	K	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	.	R	A	.	T	.	.	.	.	.	.	.	L	.	.	L	.
	LA 2930	.	.	D	.	.	H	L	.	.	.	.	.	.	.	.	.	K	Y	.	.	.	.	.	.	.	-	-	.	M	.	.	H	.	R	A	.	I	.	.	.	.	.	.	.	L	M	.	L	.
	LA 1274	.	.	.	.	.	H	L	.	.	.	.	.	.	.	.	.	K	Y	.	.	R	Y	.	.	.	.	.	.	M	.	.	H	.	R	A	.	.	A	.	.	.	P	.	.	.	.	.	L	.
	CAN262	.	.	.	.	.	H	L	.	.	.	.	.	.	.	.	.	E	Y	.	.	R	.	.	.	.	-	-	.	M	.	.	H	F	R	A	.	I	A	.	V	.	.	.	.	.	.	.	L	.

Figure 3.30: Protein haplotypes at the Pto locus. Sequences are ordered according to their putative resistant or susceptible phenotype based on the amino acids at positions 49 and 51 (red box, Böndel 2010). Amino acid positions coloured in grey are in significant LD with these two functionally important sites. pink = *S. peruvianum*, blue = *S. corneliomulleri*, orange = *S. chilense*

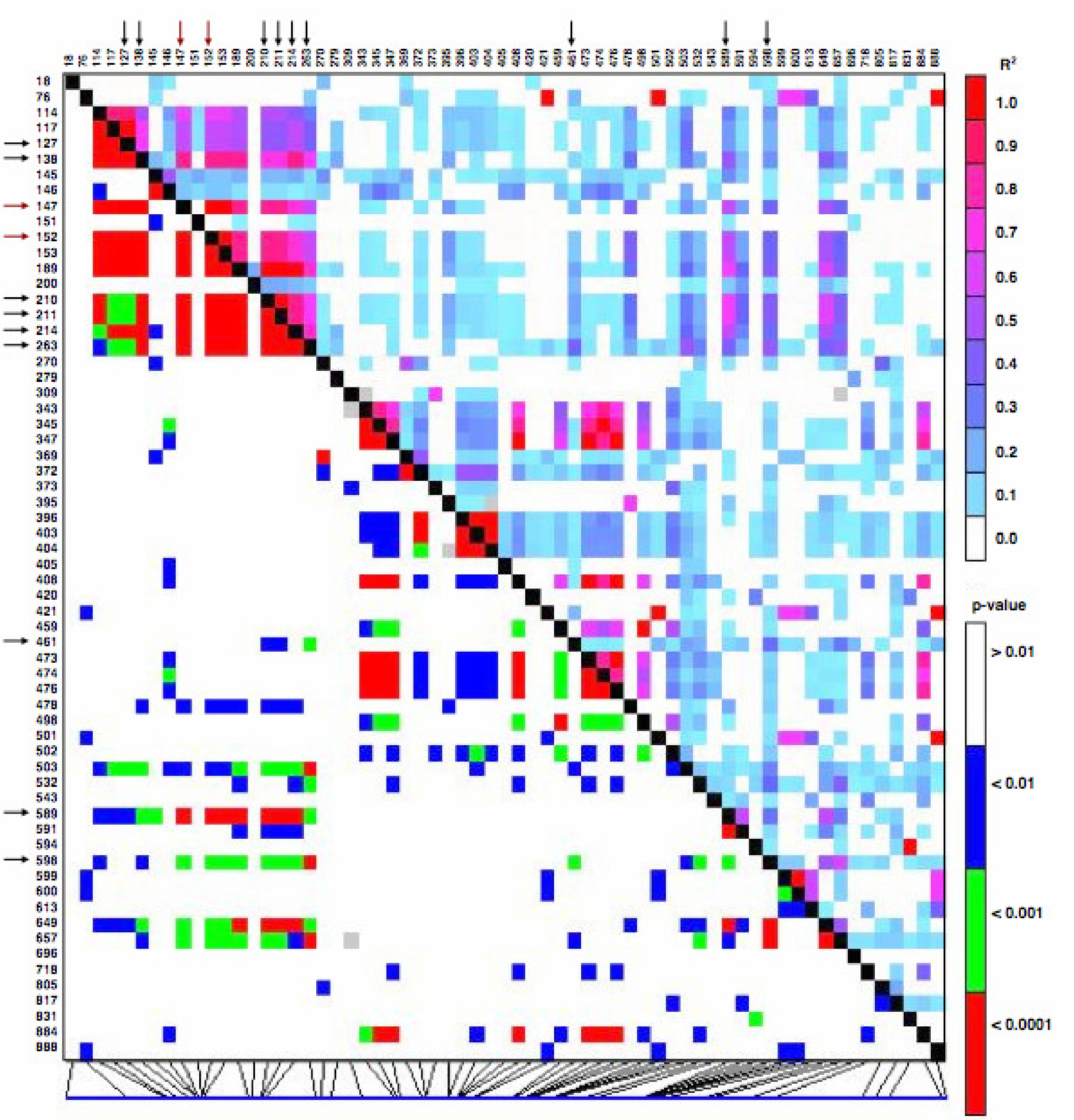


Figure 3.31: Linkage Disequilibrium across the *Pto* gene. The upper panel gives the correlation coefficient between nucleotide positions (R^2) and the lower panel gives the significance of each correlation (*P*-value). The nucleotide positions 147 and 152 (amino acid positions 49 and 51), which determine the resistant/susceptible phenotype of an allele are highlighted by red arrows. Black arrows indicate amino acid changes in significant LD with these two functionally important positions.

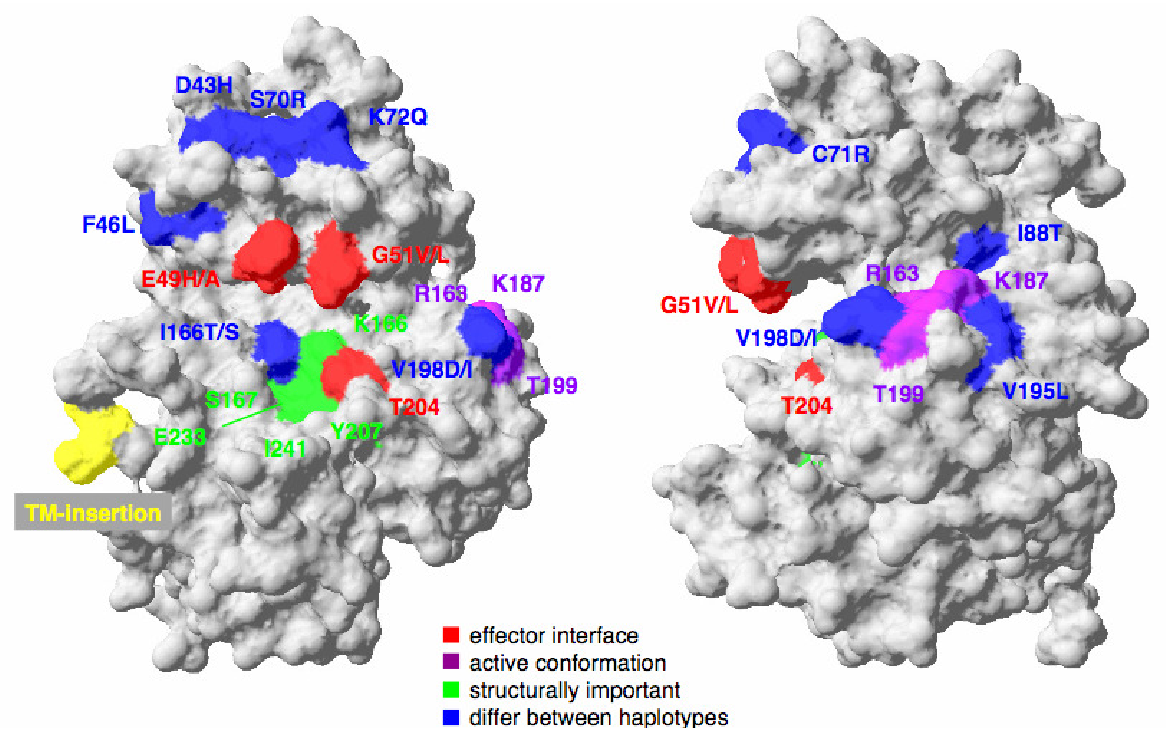


Figure 3.32: Structural model of the Pto protein. Amino acid positions with importance in AvrPto recognition are labelled in red. Positions with structural importance are labelled in green and purple. Amino acid changes labelled in blue or yellow are found to differ between the two haplotypes. **a**, View from the front of the protein. **b**, View from the side of the protein.

3.4.4 Across species diversity

Nucleotide diversity

When analyzing a dataset combining *S. peruvianum* and *S. chilense*, the mean nucleotide diversity π at the reference loci is 0.0145, which is similar to the level of diversity in only *S. peruvianum* (Table 3.7). Values of π_s (0.029), π_a (0.0028) and the ratio of π_a to π_s (0.086) all are similar to those in *S. peruvianum* (Figure 3.33). At the *Pto* locus, π_s is slightly elevated compared to the value observed in either species, but lower compared to the genomic average. π_a is lower than in *S. peruvianum*, but increased compared to *S. chilense* and compared to the reference dataset. The π_a to π_s ratio of 0.608 is elevated compared to *S. chilense* and the genomic average across the two species, but smaller than that in *S. peruvianum* alone. At the *Pfi* gene, both π_a and π_s are elevated compared to the genomic average, and the ratio of 0.444, is found to be elevated compared to the reference loci. π_a at the *Rin4* gene is slightly elevated compared to the reference dataset, while π_s is decreased. The ratio of π_a to π_s is therefore slightly elevated compared to the genomic mean.

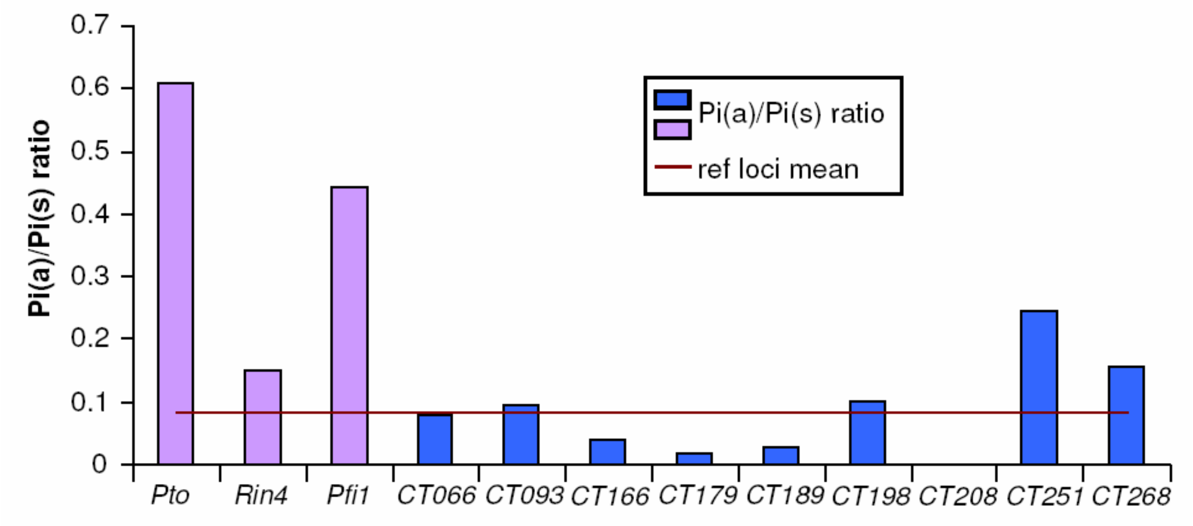


Figure 3.33: π_a/π_s ratios measured at all reference (blue bars) and resistance genes (purple bars) in the dataset combining *S. peruvianum* and *S. chilense*. The mean value (red line) was calculated over all nine reference loci.

Neutrality tests

When pooling both species together, Tajima's *D* and Fu and Li's *D* values are negative throughout the reference loci (Table 3.7, Figure 3.34). The mean values over all reference loci are -1.218 and -2.041. Tajima's *D* at synonymous sites is -1.196. The synonymous site frequency spectrum at the reference loci reveals an excess of singletons compared to neutral expectations (Figure 3.35). At the *Pto* locus, both Tajima's *D* and Fu and Li's *D* are elevated (T_D is even positive) compared to the genomic mean. This is caused by an excess of intermediate frequency polymorphisms compared to the synonymous reference frequency spectrum and the expectation under neutrality. At the *Pfi* gene, both values are only slightly elevated (-1.026 and -1.983), while both Tajima's *D* and Fu and Li's *D* are more negative than the genomic average at the *Rin4* gene. The frequency spectrum reveals a higher number of intermediate frequency polymorphisms at the *Pfi* gene and an excess of high and low frequency variants at the *Rin4* gene compared to the synonymous mean at the reference dataset. No significant McDonald-Kreitman test results were found at all three resistance genes.

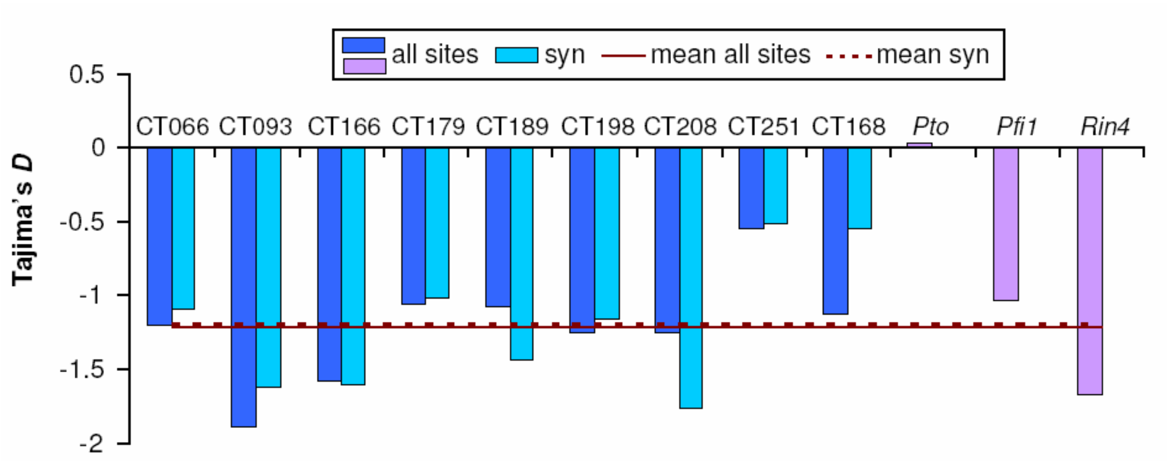


Figure 3.34: Tajima's *D* values at all and synonymous sites at the reference (blue bars) and resistance genes (purple bars). The mean values at all and synonymous sites were calculated using the nine reference loci.

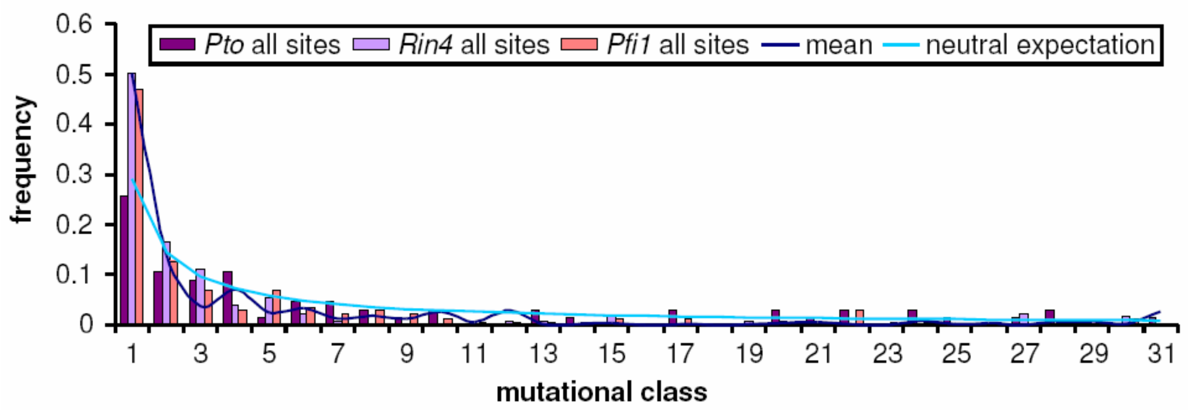


Figure 3.35: Site frequency spectrum of mutations of the three resistance genes in comparison to the neutral expectation and the frequency spectrum at synonymous sites at the reference loci for the dataset combining *S. peruvianum* and *S. chilense*. The frequency of mutations occurring at a certain number in the dataset (mutational class) is blotted for each mutational class. The light blue line indicates the expectation under complete neutrality. The dark blue line represents the mean frequency spectrum at synonymous sites measured over all nine reference loci and thus reflects only demography in the given species.

Table 3.7: Overview of summary statistics for all genes in the combined dataset *S. peruvianum* and *S. chilense*. The mean was calculated using the nine reference loci. T_D = Tajima's *D*, FL_D = Fu and Li's *D*

Locus	θ	π	π_s	π_a	π_a/π_s	T_D		FL_D
						all sites	**syn sites**	
S. peruvianum and S. chilense								
CT066	0.020	0.013	0.042	0.004	0.081	-1.199	-1.098	-2.217
CT093	0.012	0.006	0.015	0.001	0.096	-1.90	-1.630	-3.245
CT166	0.024	0.013	0.015	0.001	0.040	-1.582	-1.609	-2.691
CT179	0.027	0.019	0.049	0.001	0.020	-1.056	-1.008	-1.390
CT189	0.017	0.012	0.015	0.000	0.029	-1.070	-1.439	-2.159
CT198	0.032	0.020	0.037	0.004	0.101	-1.244	-1.150	-1.738
CT208	0.025	0.016	0.011	0.000	0.000	-1.252	-1.771	-1.771
CT251	0.020	0.017	0.037	0.009	0.246	-0.541	-0.516	-1.156
CT268	0.020	0.013	0.038	0.006	0.159	-1.124	-0.539	-2.001
reference loci mean	**0.022**	**0.014**	**0.029**	**0.003**	**0.086**	**-1.218**	**-1.196**	**-2.041**
Pto	**0.020**	**0.019**	**0.027**	**0.017**	**0.608**	**0.025**	**-0.234**	**-0.247**
Rin4	**0.026**	**0.014**	**0.027**	**0.004**	**0.152**	**-1.669**	**-1.296**	**-2.137**
Pfi	**0.035**	**0.024**	**0.035**	**0.016**	**0.444**	**-1.026**	**-0.778**	**-1.983**

3.4.5 Divergence to outgroups

Reference loci

In *S. chilense*, average divergence *K* at the reference loci to the outgroup *S. ochranthum* is 0.0347 (Table 3.8). Divergence at these loci to *S. ochranthum* is similar in *S. peruvianum* and *S. corneliomulleri* (0.0317 and 0.0332). At all three candidate genes and in all three species, divergence to *S. ochranthum* is slightly elevated, but always within the range of the reference loci.

Pto gene

In all three species, divergence to the putative pseudogene *S. ochranthum* at the *Pto* locus is lower than divergence to the more closely related *Pto* gene from *S. pennellii*. Divergence to *S. ochranthum* shows a similar pattern of divergence than at the reference loci – except for K_a, which is elevated at *Pto*. This finding can be due to the fact that *Pto* in *S. ochranthum* is indeed pseudogenized or to the fact that also π_a is elevated at this locus. In this case, the putative *Pto* pseudogene in *S. ochranthum* could still be functional and therefore not under relaxed constraint or pseudogenization could have occurred recently and new mutations have not yet accumulated. For this reason, *S. ochranthum* is – with caution – used as outgroup at the *Pto* locus in few further analyses requiring an outgroup. The synonymous divergence K_s at the *Pto* locus to *S. ochranthum* is lower than at the reference loci, while the nonsynonymous divergence K_a is elevated. This increases the ratio of K_a to K_s to nearly 1. The pattern in *S. peruvianum* is similar. In *S. corneliomulleri*, K_s is lower than at the reference loci, while K_a is elevated. This increases the ratio to over 1.

Rin4 gene

K_s at the *Rin4* gene is slightly higher than the genomic average, while K_a is slightly lower in all three species. Therefore, the ratio is decreased compared to the genome average.

Pfi gene

At the *Pfi* gene, K_s behaves like the average K_s, while K_a is elevated in all three species. This doubles the ratio of K_a to K_s.

Table 3.8: Divergence to the outgroup *S. ochranthum* at the three resistance and the nine reference genes. Additionally, divergence to *S. pennellii* was calculated at the *Pto* locus as *S. ochranthum* putatively carries a pseudogene.

Locus	K	K_s	K_a	K_a/K_s
S. peruvianum				
CT066	0.016	0.098	0.007	0.066
CT093	0.015	0.031	0.007	0.226
CT166	0.033	0.043	0.000	0.000
CT179	0.043	0.102	0.006	0.053
CT189	0.037	0.015	0.008	0.491
CT198	0.044	0.065	0.017	0.247
CT208	0.034	0.083	0.000	0.000
CT251	0.040	0.084	0.025	0.280
CT268	0.024	0.056	0.014	0.244
reference loci mean	**0.032**	**0.064**	**0.009**	**0.179**
Pto	**0.044**	**0.046**	**0.043**	**0.947**
Pto penn	**0.056**	**0.065**	**0.054**	**0.824**
Rin4	**0.043**	**0.073**	**0.008**	**0.101**
Pfi	**0.041**	**0.063**	**0.025**	**0.381**
S. chilense				
CT066	0.030	0.102	0.006	0.052
CT093	0.016	0.034	0.007	0.190
CT166	0.035	0.049	0.001	0.015
CT179	0.047	0.093	0.005	0.049
CT189	0.034	0.030	0.008	0.252
CT198	0.042	0.053	0.019	0.347
CT208	0.039	0.075	0.000	0.000
CT251	0.046	0.088	0.030	0.329
CT268	0.024	0.056	0.014	0.249
reference loci mean	**0.035**	**0.064**	**0.010**	**0.165**
Pto	**0.041**	**0.041**	**0.041**	**0.987**
Pto penn	**0.058**	**0.070**	**0.054**	**0.765**
Rin4	**0.044**	**0.076**	**0.008**	**0.099**
Pfi	**0.043**	**0.060**	**0.026**	**0.418**
S. corneliomulleri				
CT066	0.028	0.097	0.006	0.054
CT093	0.017	0.034	0.007	0.208
CT166	0.034	0.044	0.001	0.024
CT179	0.046	0.108	0.005	0.047
CT189	0.035	0.018	0.007	0.402
CT198	0.043	0.074	0.018	0.233
CT208	0.032	0.077	0.000	0.006
CT251	0.040	0.081	0.024	0.285
CT268	0.024	0.056	0.014	0.246
reference loci mean	**0.033**	**0.066**	**0.009**	**0.167**
Pto	**0.044**	**0.039**	**0.045**	**1.162**
Pto penn	**0.059**	**0.069**	**0.056**	**0.800**
Rin4	**0.042**	**0.073**	**0.007**	**0.085**
Pfi	**0.041**	**0.067**	**0.025**	**0.363**

3.4.6 Between species diversity

Mismatch distribution at the reference loci

The mismatch distribution comparing interspecific pairs of sequences was calculated for all reference loci. Afterwards, the mean number of pairwise differences between species was calculated and averaged over a locus length of 1,000 bp. The comparison of *S. peruvianum* and *S. chilense* reveals a distribution of average pairwise differences from mainly five to 43 differences between species pairs with a maximum of 18 differences (Figure 3.36). There are only few species pairs, with fewer than five or more than 43 differences. These pairs are found at loci CT066 and CT179. Divergence at fourfold degenerate sites between these pairs is on average lower than divergence between the same pairs at the other reference loci (Table 3.9). It is expected that divergence at fourfold degenerate sites depends only on the mutation rate and therefore reflects the age since divergence between sequences. Therefore, it can be assumed that the similarities between these shared sequence pairs are not due to segregating ancestral polymorphism, but rather evidence for recent introgression between *S. peruvianum* and *S. chilense*. The distribution of *S. peruvianum* - *S. corneliomulleri* comparisons appears to be different. More species pairs exhibit a low number of pairwise differences, while only few pairs differ by more than 30 differences from one another. The age of these pairs was not estimated, since the species identity of *S. corneliomulleri* is not clarified yet. It is even hypothesized that *S. peruvianum* and *S. corneliomulleri* are not separate species (Zuriaga *et al.* 2009). Comparisons between *S. chilense* and *S. corneliomulleri* reveal a similar pattern as *S. peruvianum* – *S. chilense* comparisons.

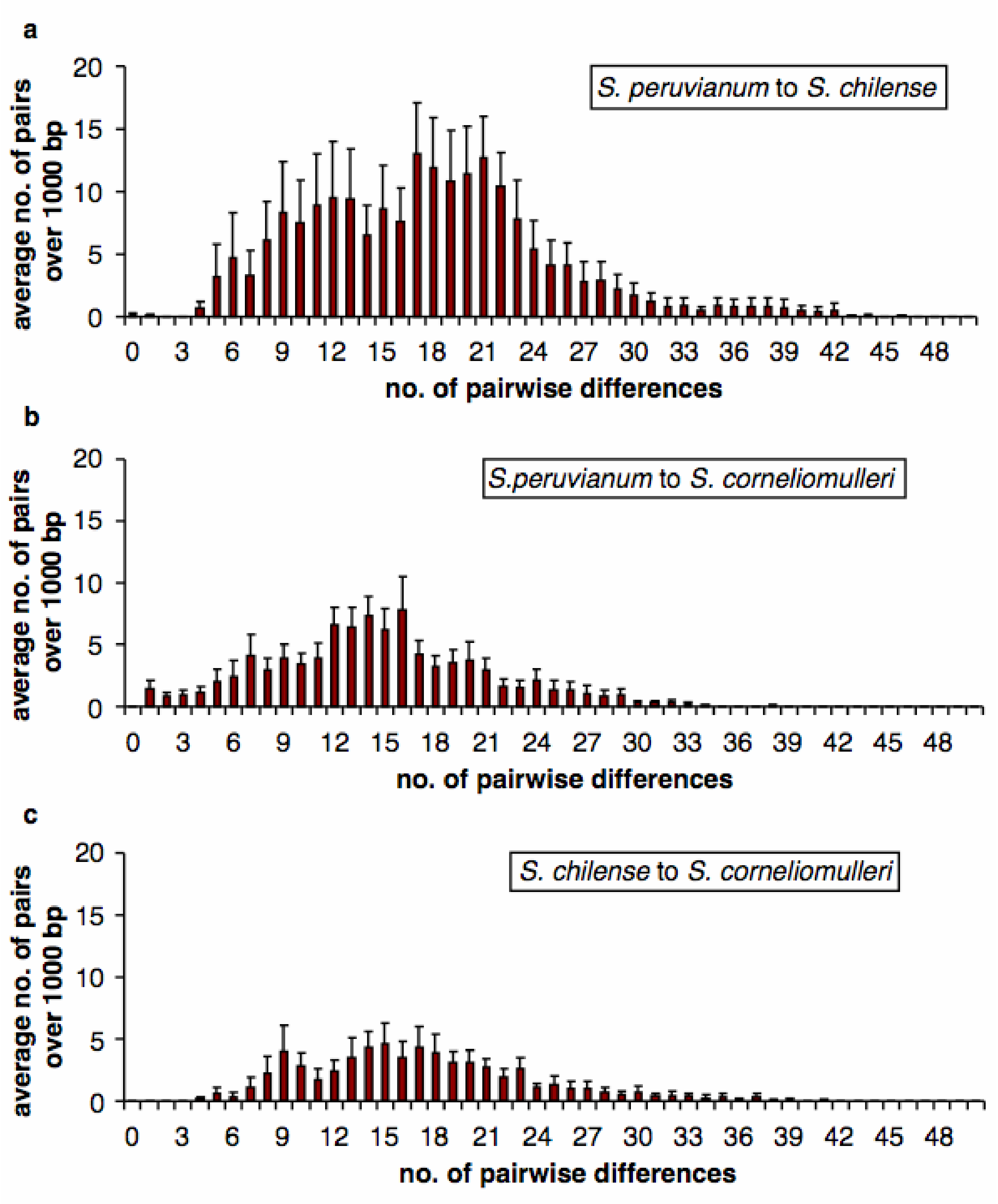


Figure 3.36: Mismatch distributions of the reference loci. Interspecific sequence pairs were screened for the number of differences they exhibit (x-axes) and the occurrence of pairs exhibiting the same number of differences was counted (y-axes). Bars show the mean value calculated for all nine reference loci extrapolated to a sequence length of 1,000 bp. Standard errors are indicated by the black error bars. **a**, *S. peruvianum* to *S. chilense* comparisons. **b**, *S. peruvianum* to *S. corneliomulleri* comparisons. **c**, *S. chilense* to *S. corneliomulleri* comparisons.

Table 3.9: Divergence between trans-specifically shared alleles (*Pto*, CT066 and CT179) Measured at synonymous K_s, nonsynonymous K_a, and fourfold degenerate sites K_{4f} in comparison to the divergence of the same pairs at the (remaining) reference loci.

	shared sequence pair			**average at reference loci**		
sequence pair	K_s	K_a	K_{4f}	K_s	K_a	K_{4f}
CT066[a]						
LA 2964 - TAC101	0.00317	0	0.00472	0.02497	0.00393	0.02327
LA 2964 - QUI126	0.01266	0	0.01415	0.02875	0.00343	0.02438
LA 2964 - LA 1958	0	0	0	0.02810	0.00236	0.02405
LA 2964 - LA 2748	0.019	0	0.01887	0.02238	0.00173	0.02179
LA 2964 - LA 2778	0.01583	0	0.01415	0.02113	0.00213	0.01923
LA 2964 - LA 3355	0.01584	0	0.01415	0.02109	0.00201	0.01923
average	**0.01108**	**0**	**0.01101**	**0.02441**	**0.00260**	**0.02199**
CT179[b]						
LA 2748 - PI 128654	0	0	0	0.02776	0.00383	0.02007
Pto[c]						
LA 1930 - LA 1913	0	0.0013	0	0.03145	0.00420	0.02829
LA 1930 - NAZ251	0.0046	0	0.0126	0.03151	0.00433	0.02710
LA 2932 - LA 3218	0.0136	0.0041	0.0075	0.02488	0.00376	0.02011
average	**0.0061**	**0.0018**	**0.0067**	**0.02928**	**0.00410**	**0.02517**

[a] The average at the reference loci was measured using all reference loci except CT066.
[b] The average at the reference loci was measured using all reference loci except CT179.
[c] The average at the reference loci was measured using all reference loci.

Mismatch distribution at the *Pto* gene

The mismatch distribution between *S. peruvianum* and *S. chilense* at the *Pto* locus reveals in general fewer pairwise differences compared to the mean of the nine reference loci (Figure 3.37). There are only few species pairs showing more than 25 pairwise differences. The pattern for *S. peruvianum* – *S. corneliomulleri* comparisons is similar, but here, more pairs with more than 25 differences can be observed. The analysis between *S. chilense* and *S. corneliomulleri* pairs reveals a narrow distribution ranging from nine to 26 pairwise differences. Intraspecific comparisons create distributions, which are shifted towards a low number of pairwise differences. There are three trans-specifically shared sequence pairs exhibiting very low numbers of differences (Table 3.9) and even more when the full *Pto* dataset is analyzed (Böndel 2010). These alleles are less diverged from one another compared to alleles at the reference loci in the same sequence pairs suggesting a recent common ancestor (Table 3.9). This finding is also supported by phylogenies obtained for the *Pto* locus on the nucleotide and protein level (Figures 3.24 and 3.38) and by the mismatch distribution obtained for nonsynonymous sites (Figure 3.42). Interspecific sequence pairs, which are similar on the protein level, exhibit only few differences on the nucleotide level as well suggesting recent divergence.

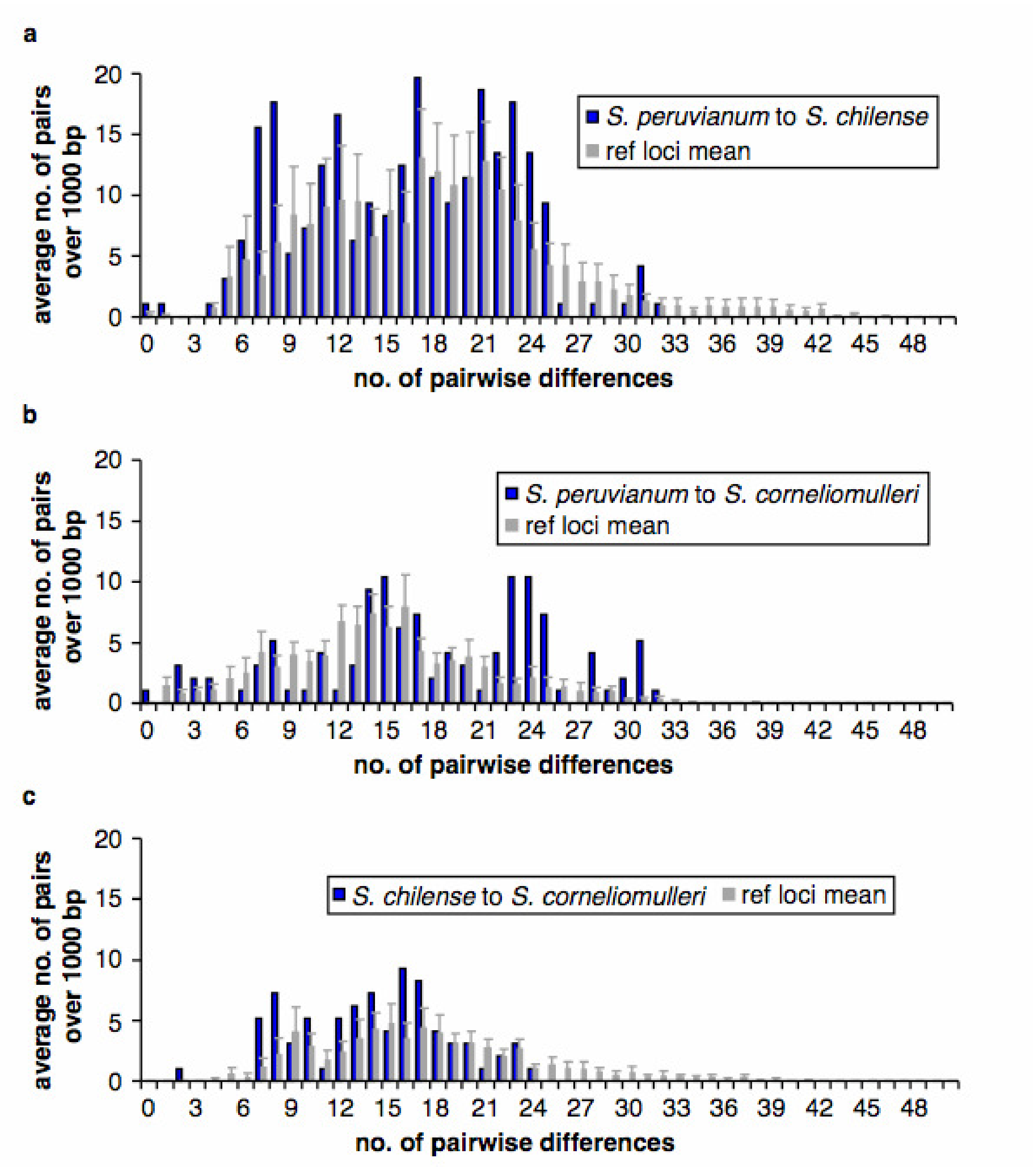


Figure 3.37: Mismatch distributions at the *Pto* gene. Interspecific sequence pairs were screened for the number of differences they exhibit (x-axes) and the occurrence of pairs exhibiting the same number of differences was counted (y-axes). Blue bars show the mean value calculated at the *Pto* locus extrapolated to a sequence length of 1,000 bp, while grey bars indicate the mean distribution and standard error observed at the reference loci (see Figure 3.36). **a**, *S. peruvianum* to *S. chilense* comparisons. **b**, *S. peruvianum* to *S. corneliomulleri* comparisons. **c**, *S. chilense* to *S. corneliomulleri* comparisons.

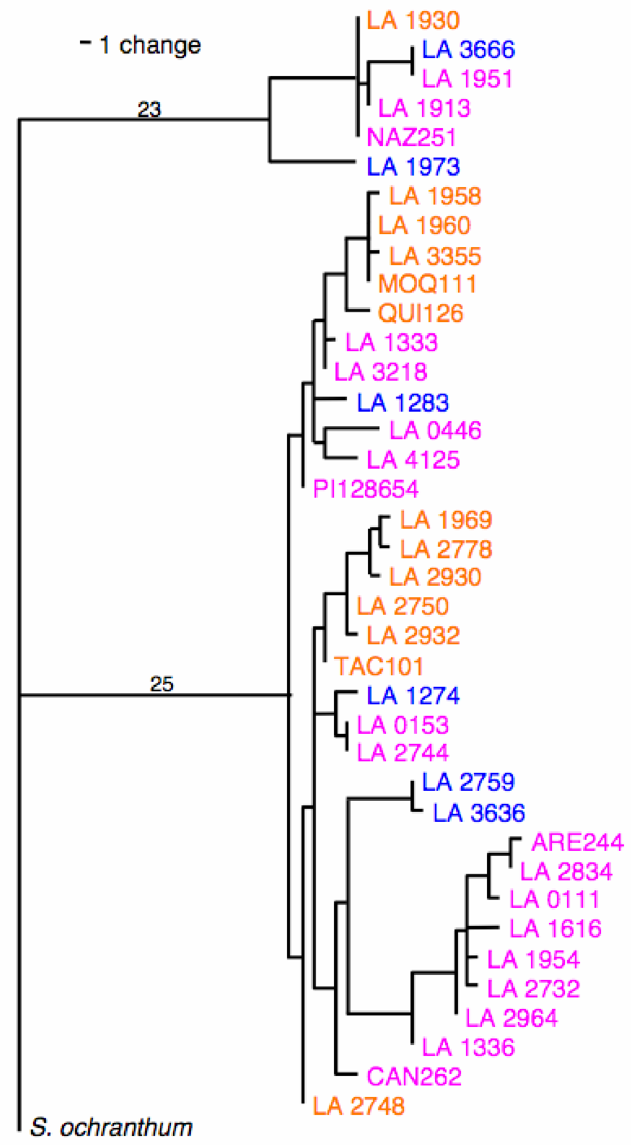


Figure 3.38: Maximum parsimony protein tree of all studied individuals at the *Pto* gene. The outgroup *S. ochranthum* was used to root the tree. Pink taxon names indicate *S. peruvianum* alleles, orange taxon names indicate *S. chilense* alleles, blue taxon names indicate *S. corneliomulleri* alleles. Numbers of changes are indicated on the branches.

Mismatch distribution at the *Rin4* gene

At the *Rin4* gene narrow distributions can be observed in all species comparisons (Figure 3.39). The distributions range in general from ten to 25 differences with a mean around 20 pairwise differences. Comparisons involving *S. chilense* are shifted towards higher numbers of pairwise differences. At intraspecific comparisons, in general fewer differences can be observed. When comparing sequence pairs at nonsynonymous sites, all distributions are shifted towards few pairwise differences with a maximum of eight differences in *S. peruvianum* – *S. chilense* comparisons.

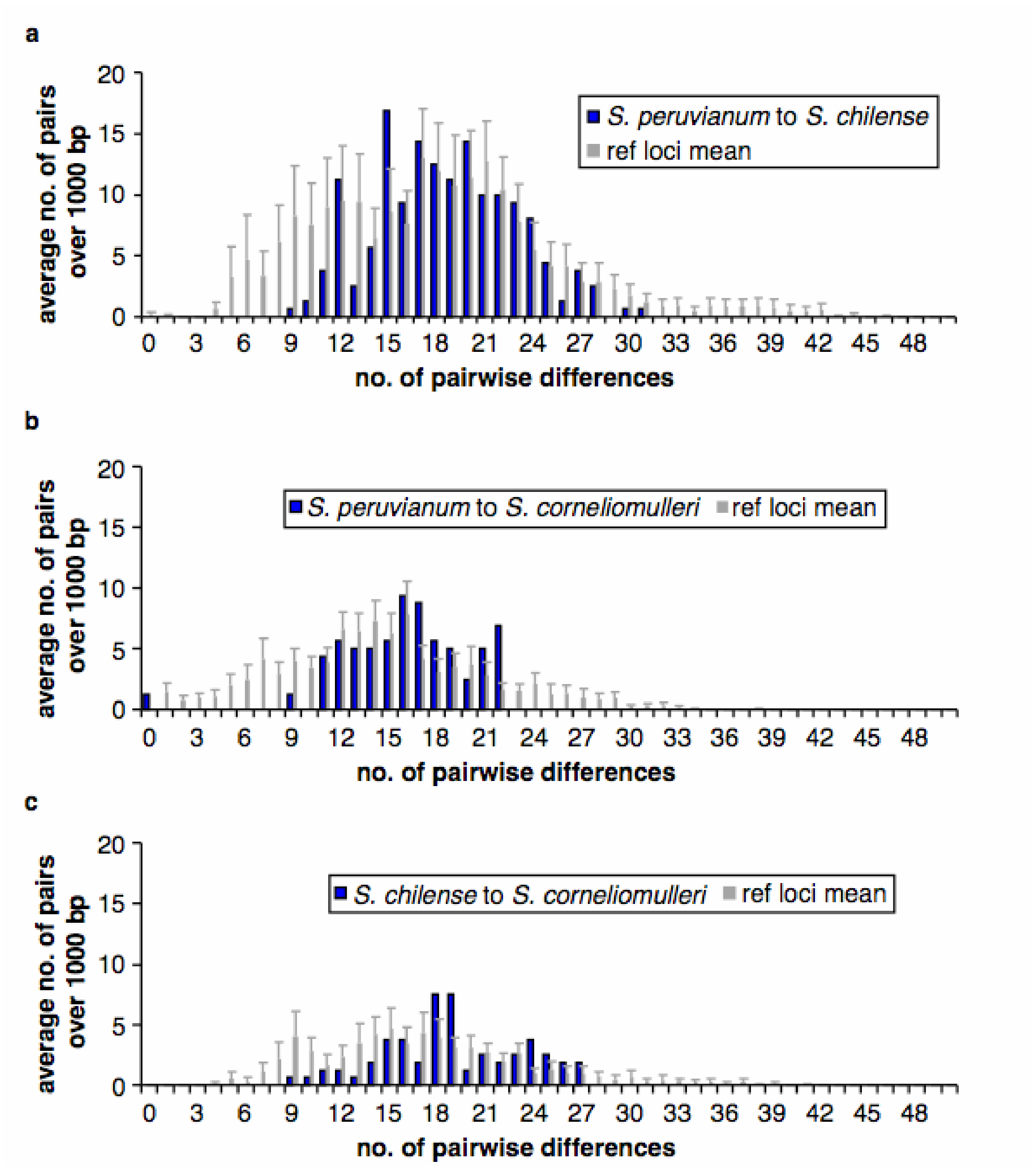


Figure 3.39: Mismatch distributions at the *Rin4* gene. Interspecific sequence pairs were screened for the number of differences they exhibit (x-axes) and the occurrence of pairs exhibiting the same number of differences was counted (y-axes). Blue bars show the mean value calculated at the *Rin4* locus extrapolated to a sequence length of 1,000 bp, while grey bars indicate the mean distribution and standard error observed at the reference loci (see Figure 3.36). **a**, *S. peruvianum* to *S. chilense* comparisons. **b**, *S. peruvianum* to *S. corneliomulleri* comparisons. **c**, *S. chilense* to *S. corneliomulleri* comparisons.

Mismatch distribution at the *Pfi* gene

Comparisons between the three different species at the *Pfi* gene are similar to each other and reveal increased numbers of pairwise differences than observed at the reference loci (Figure 3.40). Distributions range in general from 16 to 50 differences with a maximum at 30-35 differences. Intraspecific mismatch distributions are slightly shifted towards lower numbers of pairwise differences. Comparison of the mismatch distributions and phylogenies obtained at the *Pfi* nucleotide and protein sequence (Figures 3.26 and 3.41) reveals that the large number of interspecific differences between *S. peruvianum* and *S. chilense* is mainly caused by synonymous or silent polymorphisms. There are several interspecific sequence pairs, which only differ from each other by few amino acids. However, on the nucleotide level, these pairs exhibit many more differences (Figure 3.42).

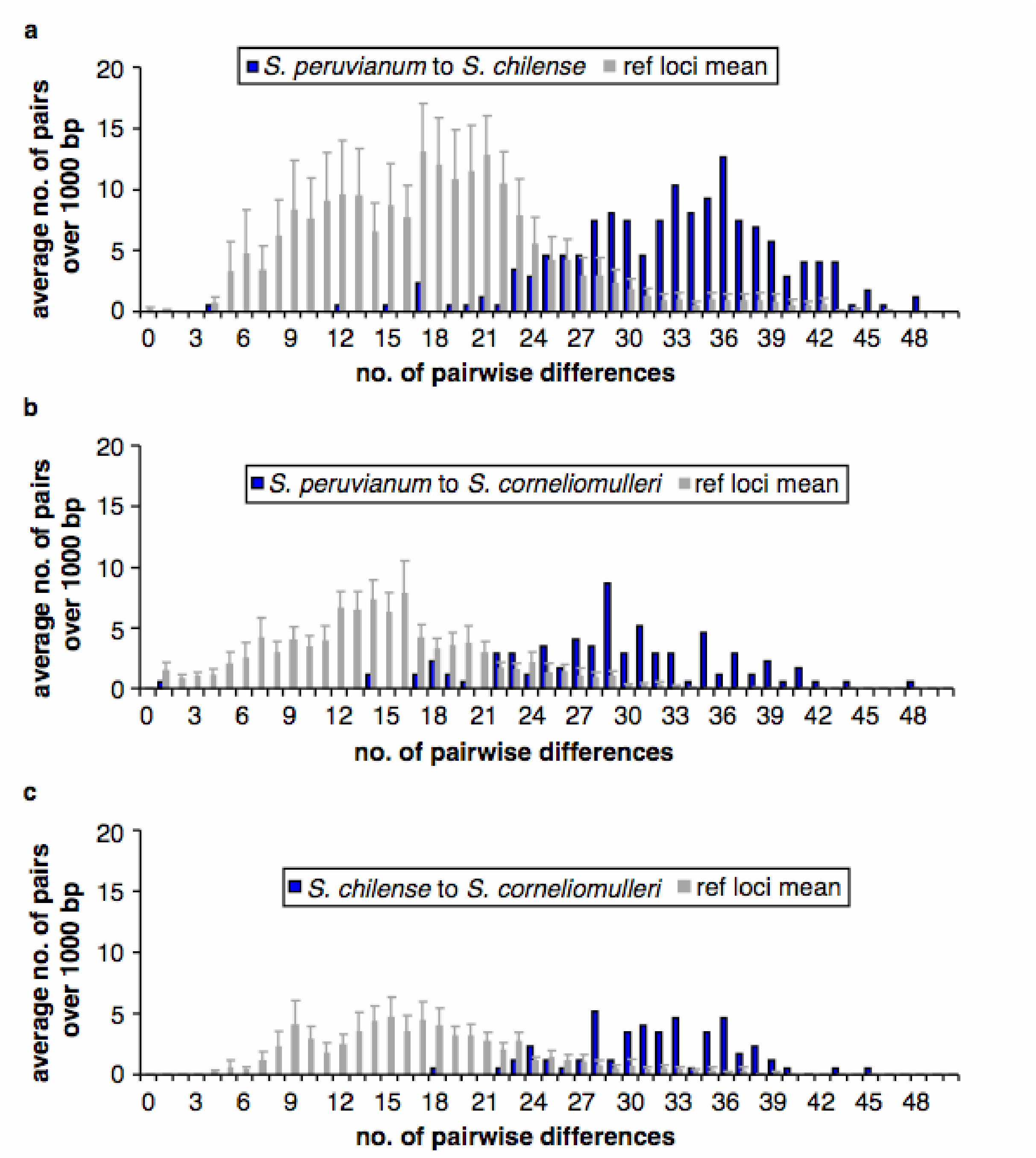


Figure 3.40: Mismatch distributions at the *Pfi* gene. Interspecific sequence pairs were screened for the number of differences they exhibit (x-axes) and the occurrence of pairs exhibiting the same number of differences was counted (y-axes). Blue bars show the mean value calculated at the *Pfi* locus extrapolated to a sequence length of 1,000 bp, while grey bars indicate the mean distribution and standard error observed at the reference loci (see Figure 3.36). **a**, *S. peruvianum* to *S. chilense* comparisons. **b**, *S. peruvianum* to *S. corneliomulleri* comparisons. **c**, *S. chilense* to *S. corneliomulleri* comparisons.

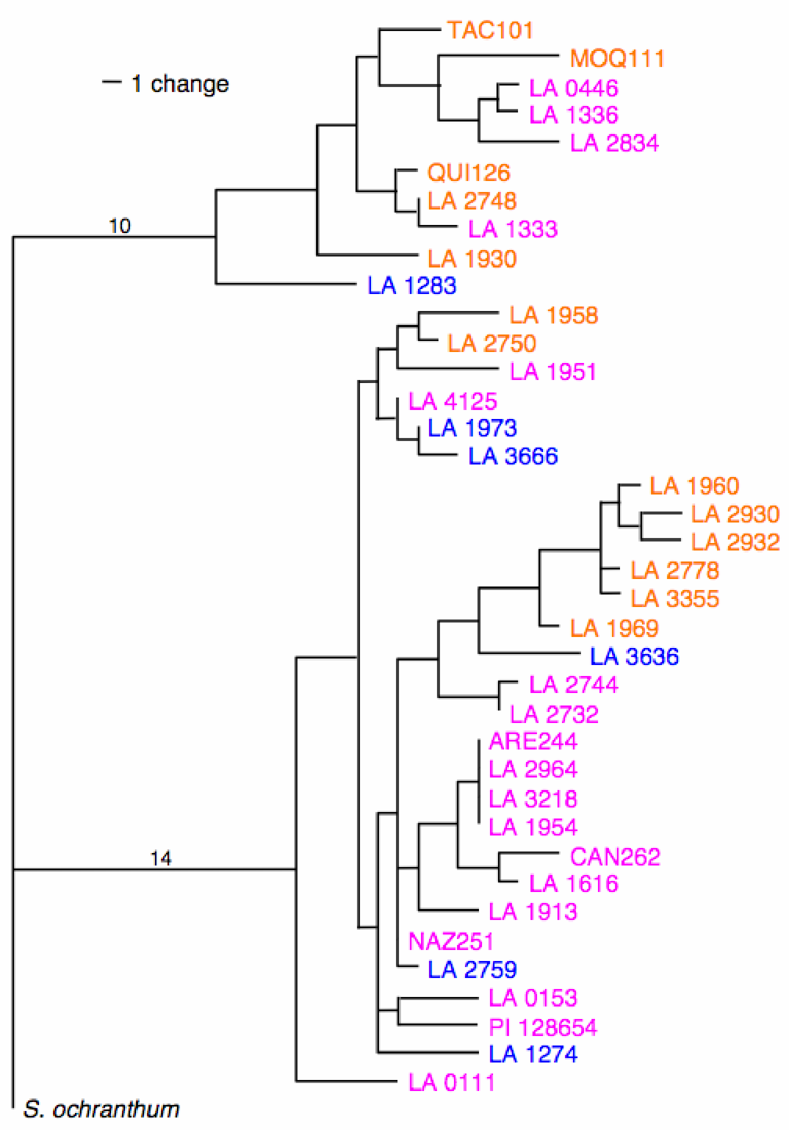


Figure 3.41: Maximum parsimony protein tree of all studied individuals at the *Pfi* gene. The outgroup *S. ochranthum* was used to root the tree. Pink taxon names indicate *S. peruvianum* alleles, orange taxon names indicate *S. chilense* alleles, blue taxon names indicate *S. corneliomulleri* alleles. Numbers of changes are indicated on the branches.

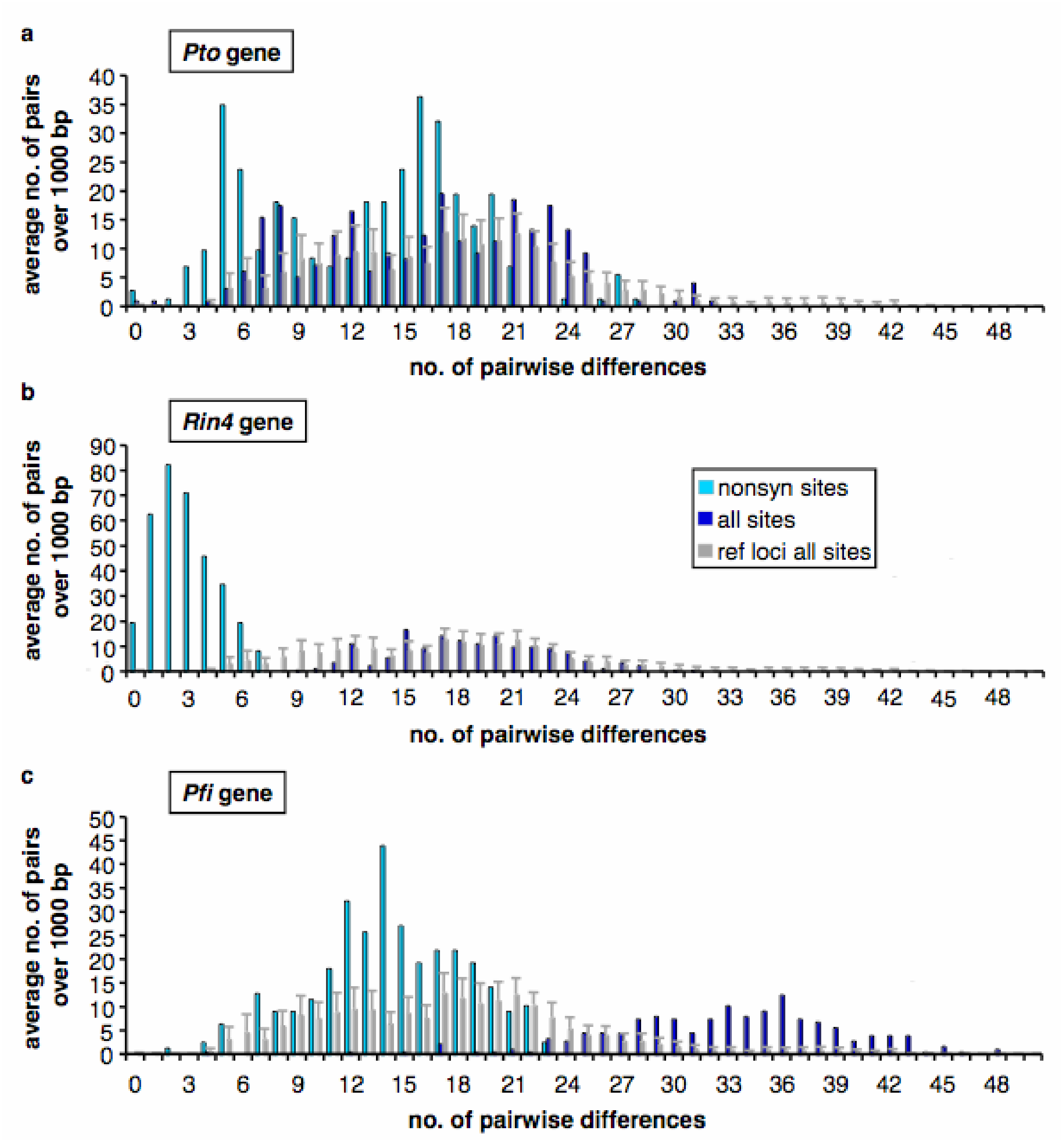


Figure 3.42: Mismatch distributions between *S. peruvianum* and *S. chilense* at nonsynonymous sites at the three resistance genes. Interspecific sequence pairs were screened for the number of differences they exhibit (x-axes) and the occurrence of pairs exhibiting the same number of differences was counted (y-axes). Blue bars show the value calculated at the resistance genes extrapolated to a sequence length of 1,000 bp (dark blue = all sites, light blue = nonsynonymous sites), while grey bars indicate the mean distribution and standard error observed at the reference loci (see Figure 3.36).

Joint site frequency spectrum at the reference loci

Since the sample size of *S. corneliomulleri* (n = 6) is very small and the identity of this species is controversial (see Chapter 4.4), the Joint Site Frequency Spectrum (JSFS) was only calculated between *S. peruvianum* and *S. chilense*. On average, 33 synonymous and 16 nonsynonymous sites are polymorphic in the JSFS of *S. peruvianum* and *S. chilense* at the nine reference loci (Table 3.10). Of all synonymous polymorphisms, 13.5% are shared between the two species, while 63.3% are private to *S. peruvianum* and 23.2% are private to *S. chilense*. Only 8.7% of all nonsynonymous polymorphisms are shared, while 55.8% are private to *S. peruvianum* and 24.4% are private to *S. chilense*. In *S. peruvianum*, the JSFS at the reference loci reveals an excess of singletons for both shared synonymous and nonsynonymous polymorphisms (Figures 3.44 and 3.47). Few shared polymorphisms are found in the other mutational classes. A similar pattern is shown for private synonymous and nonsynonymous substitutions (Figures 3.43 and 3.47). The JSFS reveals a different picture in *S. chilense*. While there is an excess of private singletons as observed in *S. peruvianum* (Figures 3.45 and 3.47), there is a larger proportion of high frequency shared polymorphisms (Figures 3.46 and 3.47).

Joint site frequency spectrum at the *Pto* gene

At the *Pto* locus, 20 synonymous and 37 nonsynonymous polymorphisms are observed in the JSFS (Table 3.10). Surprisingly, a greater proportion of nonsynonymous (35.1%) polymorphisms are shared between the two species than synonymous (25.0%) polymorphisms. Amongst the private nonsynonymous polymorphisms, a greater proportion is private to *S. peruvianum* (51.4%) than to *S. chilense* (13.5%), while the difference is not as extreme at synonymous sites (45.0% private to *S. peruvianum* and 30.0% private to *S. chilense*). There are few or no shared synonymous or nonsynonymous singletons at *Pto* compared to the reference loci in *S. peruvianum* (Figures 3.44 and 3.48), because polymorphisms occur mainly in intermediate frequency. Private synonymous and nonsynonymous substitutions occur mainly in intermediate frequency as well (Figures 3.43 and 3.48). In *S. chilense*, an excess of singletons and intermediate frequency polymorphisms is revealed at shared synonymous and nonsynonymous mutations (Figures 3.46 and 3.48). At private synonymous positions, fewer singletons and more intermediate frequency variants can be observed than at the reference loci, while at private nonsynonymous positions only an excess of intermediate frequency variants is shown (mutational class 3, Figures 3.45 and 3.48).

Joint site frequency spectrum at the *Rin4* gene

The JSFS at the *Rin4* gene appears to be different from the other resistance genes, and 35 synonymous and 27 nonsynonymous polymorphisms are observed (Table 3.10). There is an absence of shared nonsynonymous polymorphisms between species. Two thirds of all nonsynonymous mutations are private to *S. peruvianum*, while the remaining SNPs are private to *S. chilense*. Of all synonymous substitutions 31.4% are shared and 34.4% are private to *S. peruvianum* and *S. chilense* respectively. In *S. peruvianum*, private nonsynonymous polymorphisms occur only in low frequency, while a slight excess of intermediate frequency variants can be observed in *S. chilense* (Figures 3.43, 3.45 and 3.49). There are fewer shared synonymous singletons and more high frequency variants compared to the reference loci in *S. peruvianum* (Figures 3.44 and 3.49). The frequency spectrum of private synonymous mutations behaves similarly to the reference loci, but fewer intermediate and high frequency polymorphisms can be reported (Figures 3.43 and 3.49). Observations at synonymous polymorphisms in *S. chilense* are similar to the reference loci (Figures 3.45, 3.46 and 3.49). However, fewer high frequency variants can be observed at private synonymous polymorphisms.

Joint site frequency spectrum at the *Pfi* gene

At the *Pfi* gene, 36 synonymous and 65 nonsynonymous polymorphisms are found in the JSFS (Table 3.10). Slightly more nonsynonymous than synonymous mutations are shared between the two species (24.6% and 19.4% respectively). Fewer nonsynonymous and synonymous polymorphisms are private to *S. peruvianum* than at the reference loci (43.1% and 58.3%), while as many or more synonymous and nonsynonymous mutations are private to *S. chilense* compared to the reference loci (22.2% and 32.2%). Shared substitutions (synonymous and nonsynonymous) at the *Pfi* gene show fewer singletons and an excess of intermediate frequency variants in comparison with the reference loci in *S. peruvianum* (Figures 3.44 and 3.50). Nonsynonymous polymorphisms private to *S. peruvianum*, exhibit more polymorphisms in intermediate frequency compared to reference loci, while private synonymous polymorphism exhibit an excess of singletons (Figures 3.43 and 3.50). In *S. chilense*, nonsynonymous polymorphisms (private and shared) occur mainly in intermediate frequency (Figures 3.45, 3.46 and 3.50). Synonymous substitutions, which are private to *S. chilense*, exhibit fewer singletons and high frequency polymorphisms, but more intermediate frequency polymorphisms than the reference loci (Figures 3.45 and 3.50).

Synonymous polymorphisms, which are shared, occur mainly in intermediate frequency (Figures 3.46 and 3.50).

Table 3.10: Summary of the Joint Site Frequency Spectrum between *S. peruvianum* and *S. chilense*. The total number of synonymous and nonsynonymous polymorphisms is given. The frequency of shared and private polymorphisms is based on this total number. The mean value was calculated using all nine reference loci.

Locus	**total no.**	**nonsynonymous polymorphisms frequency**			**total no.**	**synonymous polymorphisms frequency**		
		shared	**private *S. peruvianum***	**private *S. chilense***		**shared**	**private *S. peruvianum***	**private *S. chilense***
CT066	20	0.050	0.550	0.400	72	0.153	0.528	0.319
CT093	10	0.100	0.700	0.200	27	0.148	0.630	0.222
CT166	3	0	0	1.000	10	0	0.500	0.500
CT179	2	0	0	0	35	0.257	0.600	0.143
CT189	1	0	1.000	0	8	0.125	0.500	0.375
CT198	4	0.250	0.750	0	17	0.118	0.824	0.059
CT208	0	0	0	0	10	0	0.900	0.100
CT251	41	0.268	0.488	0.244	46	0.196	0.652	0.152
CT268	62	0.113	0.532	0.355	73	0.219	0.562	0.219
mean	**15.889**	**0.087**	**0.558**	**0.244**	**33.11**	**0.135**	**0.633**	**0.232**
Pto	**37**	**0.351**	**0.514**	**0.135**	**20**	**0.250**	**0.450**	**0.300**
Rin4	**27**	**0**	**0.667**	**0.333**	**35**	**0.314**	**0.343**	**0.343**
Pfi	**65**	**0.246**	**0.431**	**0.323**	**36**	**0.194**	**0.583**	**0.222**

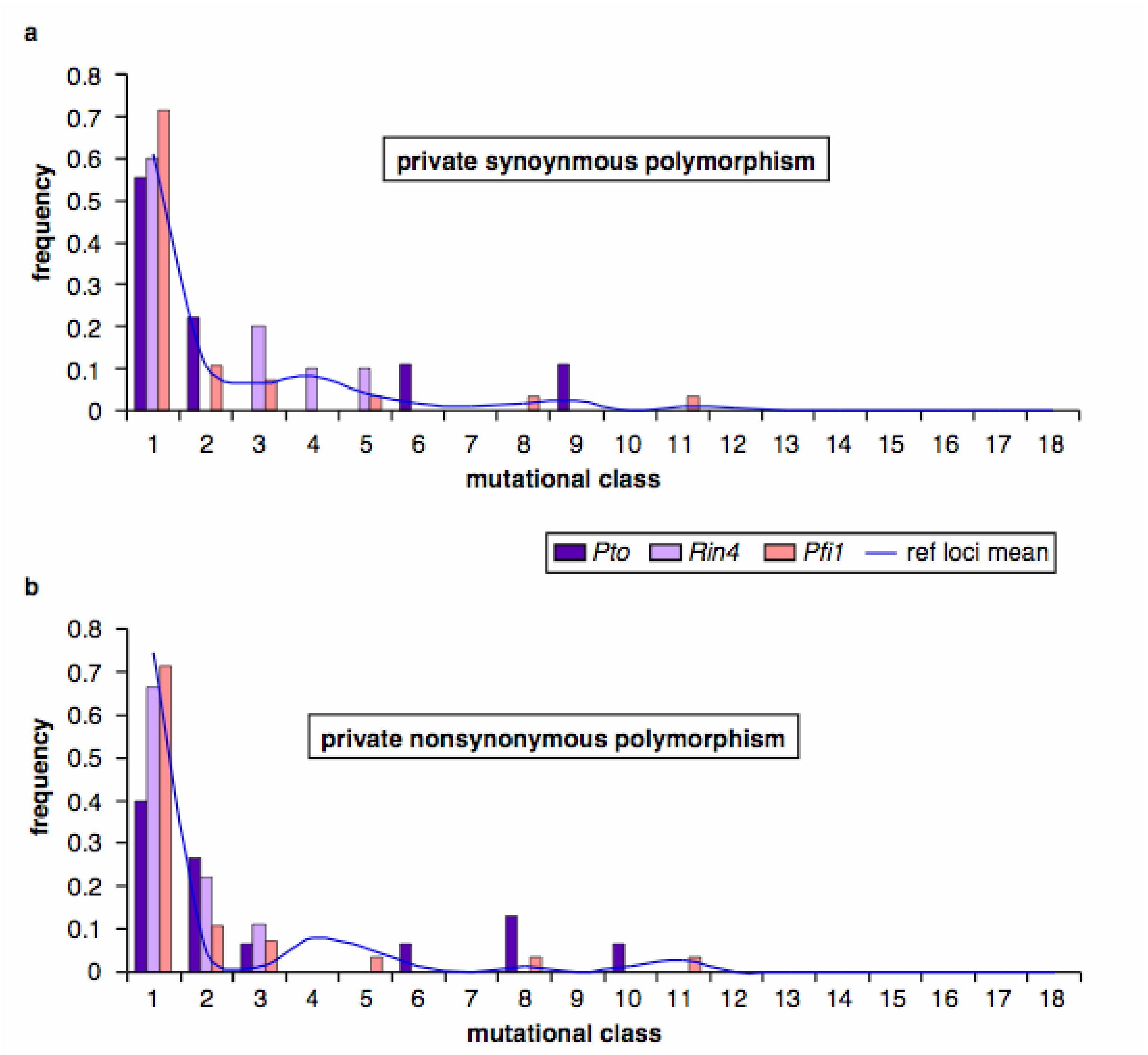


Figure 3.43: Frequency spectra of polymorphisms, which are private to *S. peruvianum*. The reference loci mean was averaged over all nine reference loci. **a**, Private synonymous polymorphisms. **b**, Private nonsynonymous polymorphisms

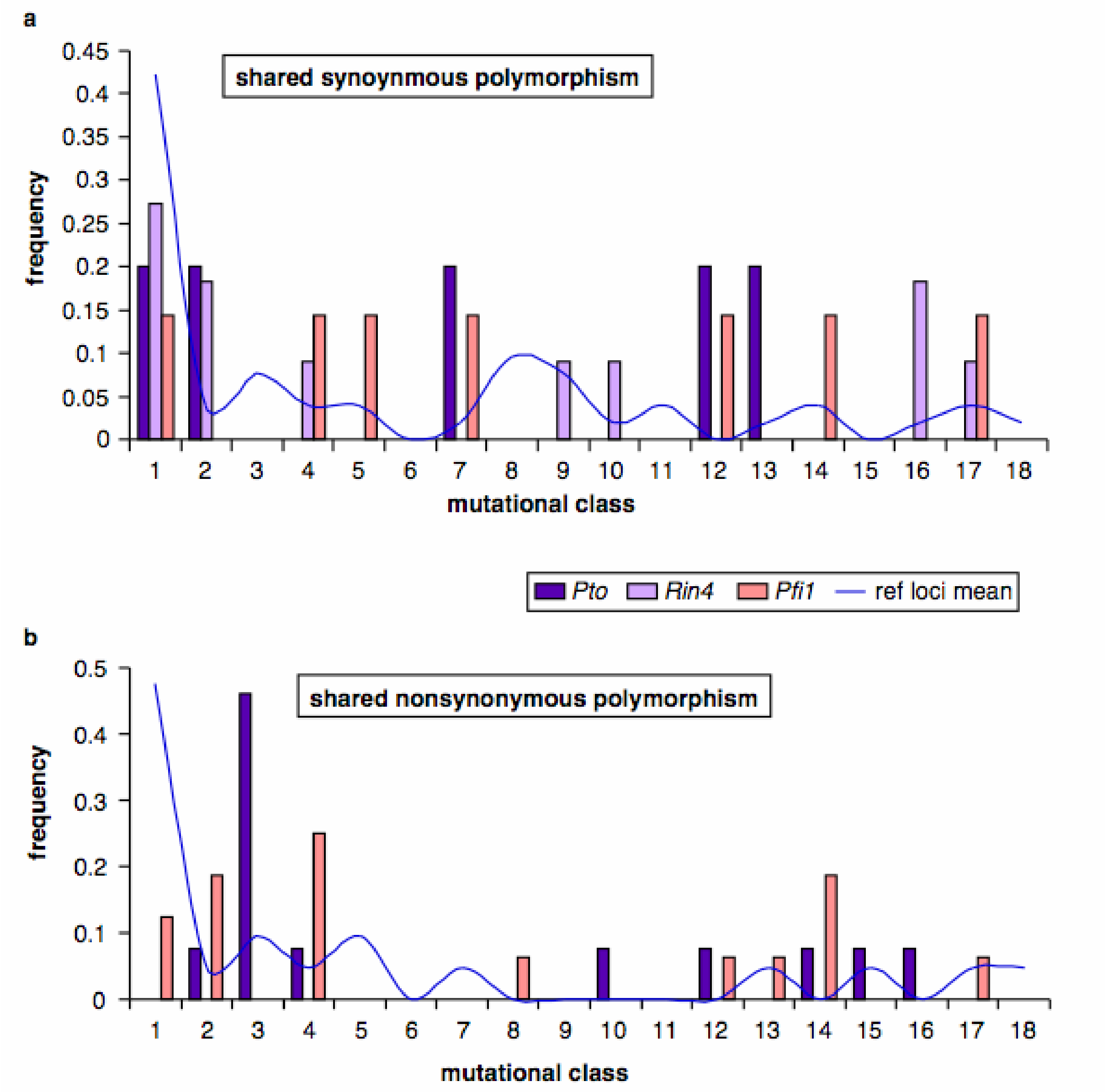


Figure 3.44: Frequency spectra of polymorphisms in *S. peruvianum*, which are shared between *S. peruvianum* and *S. chilense*. The reference loci mean was averaged over all nine reference loci. **a**, Shared synonymous polymorphisms. **b**, Shared nonsynonymous polymorphisms

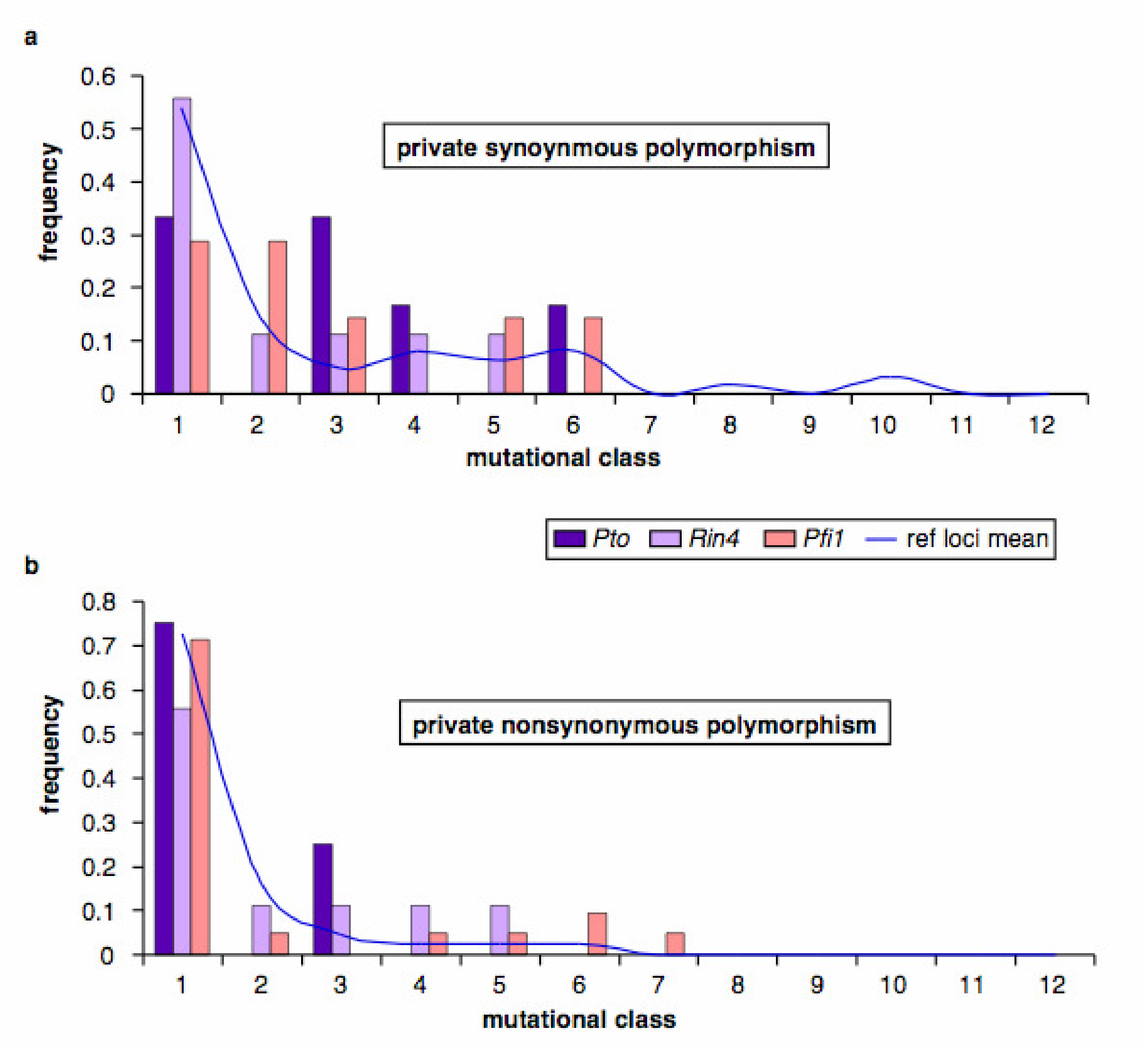


Figure 3.45: Frequency spectra of polymorphisms, which are private to *S. chilense*. The reference loci mean was averaged over all nine reference loci. **a**, Private synonymous polymorphisms. **b**, Private nonsynonymous polymorphisms

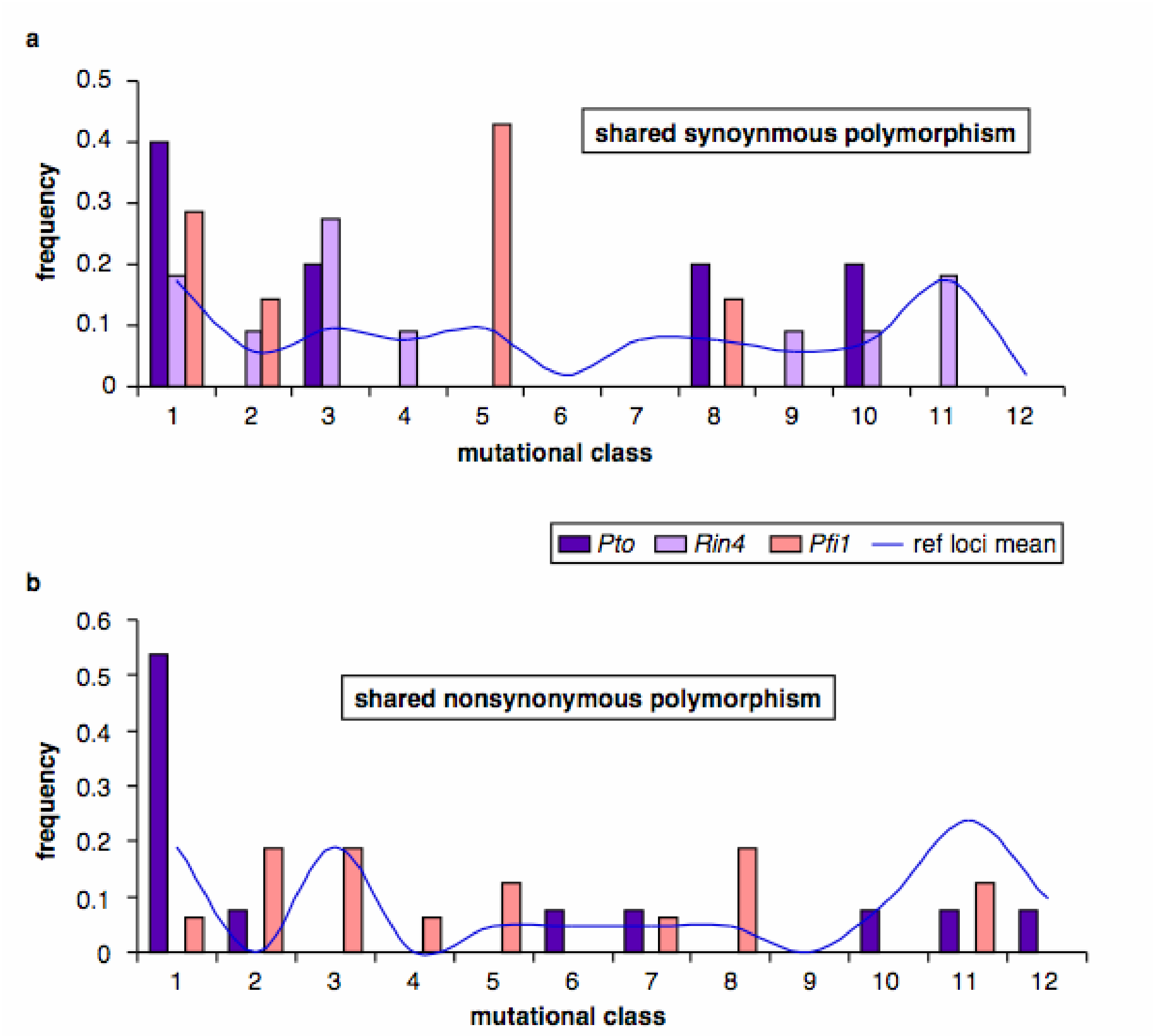


Figure 3.46: Frequency spectra of polymorphisms in *S. chilense*, which are shared between *S. peruvianum* and *S. chilense*. The reference loci mean was averaged over all nine reference loci. **a**, Shared synonymous polymorphisms. **b**, Shared nonsynonymous polymorphisms

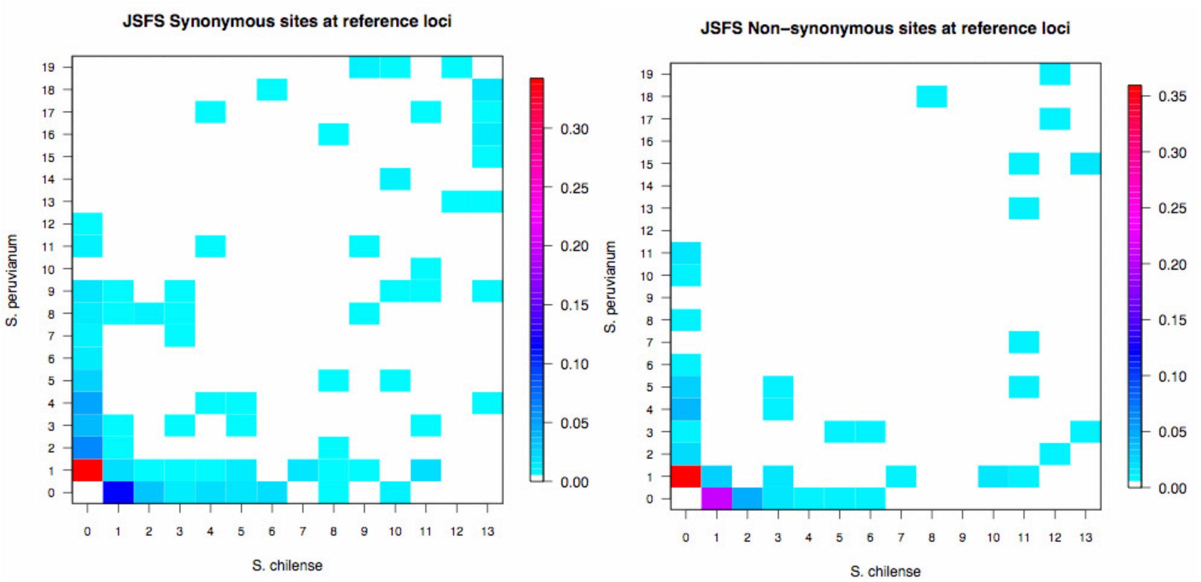


Figure 3.47: Heatmap of the JSFS between *S. peruvianum* and *S. chilense* at the nine reference loci. The frequency of polymorphisms as found in each mutational class in the two species is indicated by the colour code explained in the legend on the right of each heatmap. Left: synonymous sites, right: nonsynonymous sites

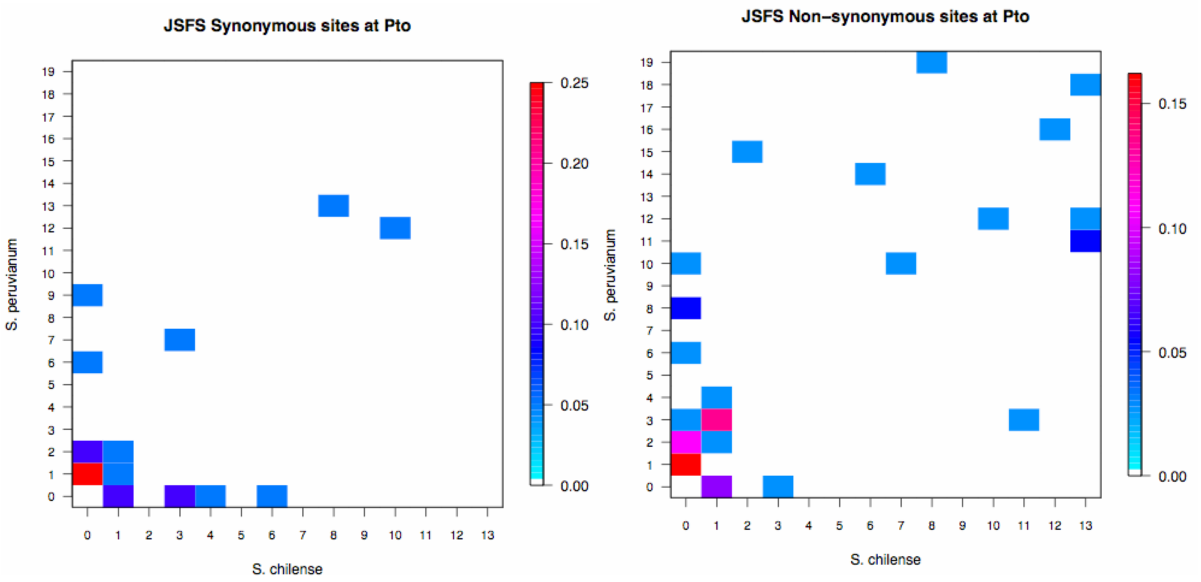


Figure 3.48: Heatmap of the JSFS between *S. peruvianum* and *S. chilense* at the *Pto* gene. The frequency of polymorphisms as found in each mutational class in the two species is indicated by the colour code explained in the legend on the right of each heatmap. Left: synonymous sites, right: nonsynonymous sites

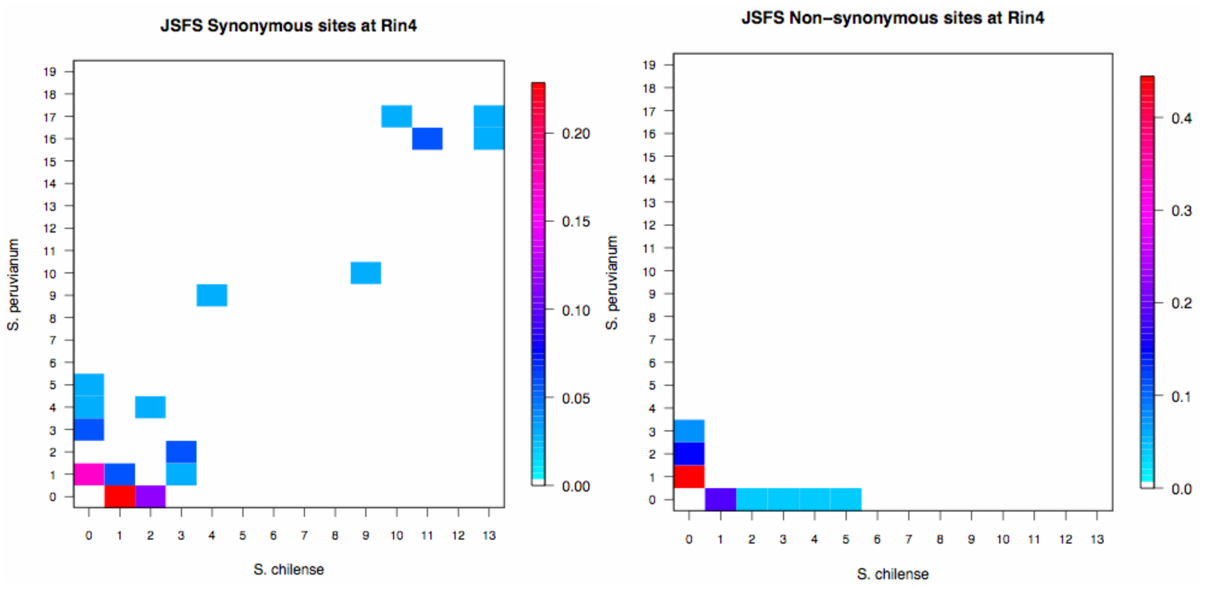


Figure 3.49: Heatmap of the JSFS between *S. peruvianum* and *S. chilense* at the *Rin4* gene. The frequency of polymorphisms as found in each mutational class in the two species is indicated by the colour code explained in the legend on the right of each heatmap. Left: synonymous sites, right: nonsynonymous sites

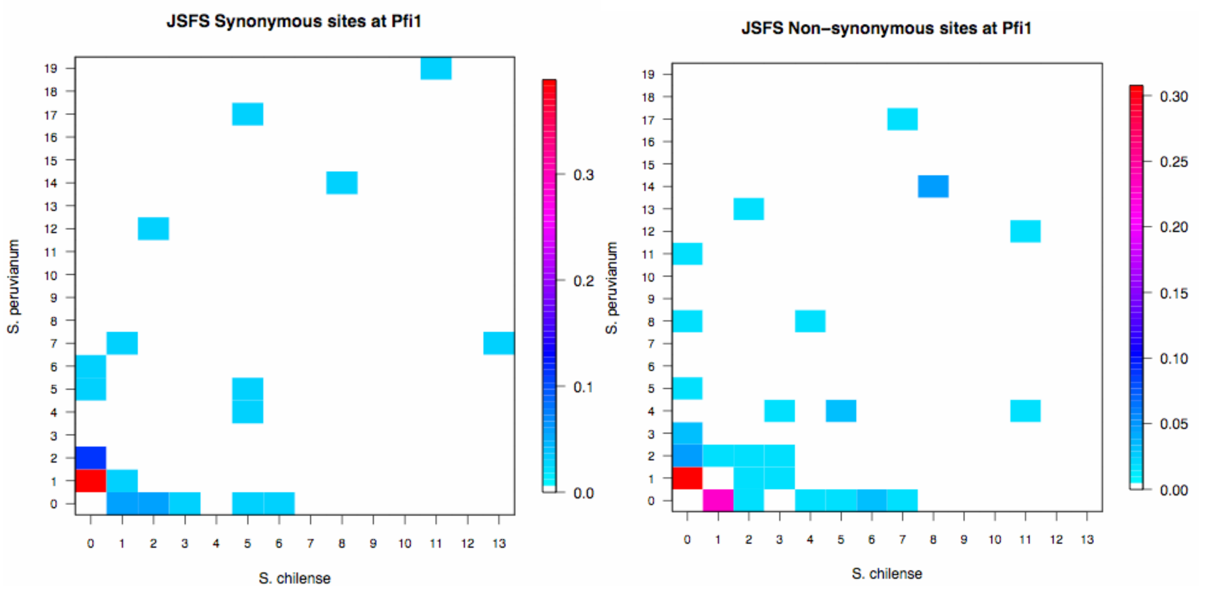


Figure 3.50: Heatmap of the JSFS between *S. peruvianum* and *S. chilense* at the *Pfi* gene. The frequency of polymorphisms as found in each mutational class in the two species is indicated by the colour code explained in the legend on the right of each heatmap. Left: synonymous sites, right: nonsynonymous sites

F_{ST}

Differentiation between the species using the index of fixation F_{ST}

The index of fixation (F_{ST}) between *S. peruvianum* and *S. chilense* varies at the reference loci between 0.18 (CT198) and 0.671 (CT189) (Figure 3.51), the mean value being 0.326. At all three candidate loci, F_{ST} between these two species is lower than the genome average. F_{ST} values at the *Rin4* and *Pfi* locus are 0.096 and 0.165 and are therefore outside the range observed at the reference loci. At the *Pto* locus, the value however falls within the distribution of the reference loci (0.221). The level of genetic differentiation observed between *S. chilense* and *S. corneliomulleri* is similar to that observed between *S. peruvianum* and *S. chilense*. The mean value at the reference loci is only slightly lower (0.283). Values at the candidate genes are 0.135 (*Pto* and *Rin4*) and 0.137 (*Pfi*) and therefore lower than the genomic average. Between *S. peruvianum* and *S. corneliomulleri*, no genetic differentiation can be observed. F_{ST} values at the reference loci range from -0.018 (CT251 and CT268) to 0.06 (CT166) with a mean of 0.021. Values at the *Rin4* and *Pfi* gene lie within this range. F_{ST} between *S. peruvianum* and *S. corneliomulleri* at the *Pto* locus however is elevated (0.114).

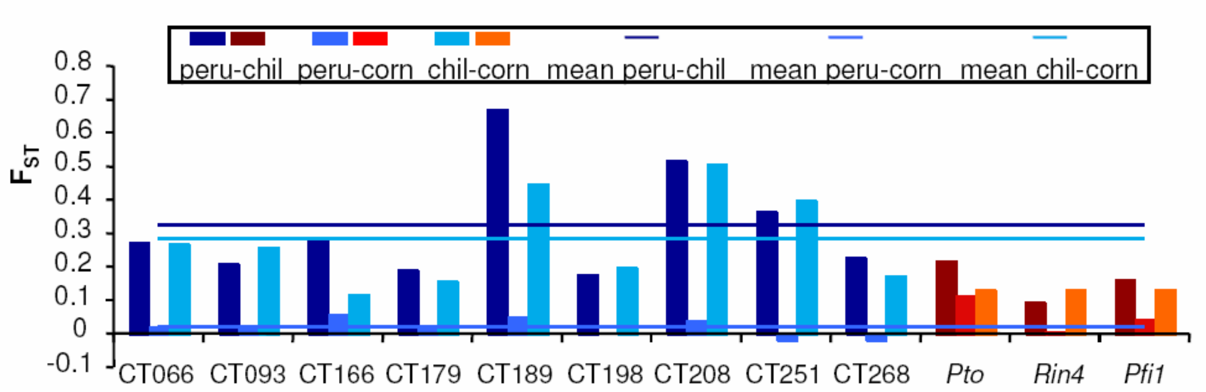


Figure 3.51: F_{ST} values measured at all reference loci and resistance genes. The mean values were calculated based on the nine reference genes.

CHAPTER 4: DISCUSSION

4.1 Complex evolutionary history at the *Rcr3* gene family

To better understand the role of guardees in plant immunity, I analyzed alleles of the effector target *Rcr3* from 26 individuals of multiple wild tomato species with particular focus on the species *S. peruvianum*. The *Rcr3* gene forms a young gene family in *S. peruvianum* and its closely related sister species *S. corneliomulleri*. Two differentiated sequence types are maintained, potentially by balancing selection, within and across *Rcr3* loci. In contrast to previous studies, which found variation in pathogen recognition at resistance loci, this study provides evidence for variation in activation of the defence response (Rose *et al.* 2004; Rose *et al.* 2005; Rose *et al.* 2007).

4.1.1 *Rcr3* forms a young gene family

Sequence analysis revealed that the *Rcr3* gene is not a single-copy locus in *S. peruvianum* and its closely related sister species *S. corneliomulleri*, but forms a small gene family with at least three closely related paralogs, *Locus A*, *Locus B* and *Locus C*. Based on the following four observations: 1) alleles originating from different *Rcr3* loci could be amplified with the same set of PCR primers, 2) allelic diversity within and across loci is relatively low, 3) *Rcr3* duplicates still segregate within the population, and 4) gene conversion occurs frequently, it can be concluded that the duplication at the *Rcr3* locus must have happened recently. Since the 3’ flanking region of *Locus A* exhibits the greatest sequence similarity to the *Rcr3* locus from the cultivated tomato *S. lycopersicum* and most individuals used in the population genetic analyses carry alleles originating from *Locus A*, but not necessarily alleles from *Locus B* or *C,* it is likely that *Locus A* is the original copy of *Rcr3* and orthologous to *Rcr3* from *S. lycopersicum.*

Gene duplication and subsequent (functional) divergence of duplicates are typical mechanisms contributing to diversity at genes involved in host-pathogen coevolution (Michelmore & Meyers 1998; Raffaele *et al.* 2010). However, young duplicates, which have not diverged from one another, can be homogenized by frequent gene conversion (Michelmore & Meyers 1998). The recent origin of the *Rcr3* gene family in *S. peruvianum*, as attested by the high sequence similarity between the *Rcr3* ORFs and the presence of copy number variants of *Rcr3* within populations, suggests that the *Rcr3* gene family is at the beginning of such a process.

The presence of *Rcr3* as a small gene family in *S. peruvianum* is particularly intriguing because its interacting molecular partner, *Cf-2*, also exists as a gene family in wild tomatoes (Caicedo & Schaal 2004; Dixon *et al.* 1996). Furthermore, this interaction between Rcr3 and Cf-2 requires a precise matching between allelic variants. A mismatch between these variants, such as when a variant of Rcr3 from one *Solanum* species is paired with a variant of Cf-2 from another but nevertheless closely related *Solanum* species, results in an autonecrotic response, and has been pinpointed as an example of Dobzhansky-Muller incompatibility between tomato species (Bomblies *et al.* 2007; Krüger *et al.* 2002). Thus, because coevolution between these two molecules is extremely fine-tuned, duplication of one of the partners could facilitate duplication of the other partner.

4.1.2 Two sequence types are maintained at the *Rcr3* locus

The phylogeny of the *Rcr3* coding regions reveals two distinct sequence types that are independent from the locus of origin. This structure is caused by variation at the intron and variation linked to the intron. Within sequence type variation is consistent with the maintenance of the two sequence types via long-term balancing selection, as ratios of π_a to π_s at *Rcr3* are elevated compared to the reference loci and neutrality test results, although not significant, tend to be increased compared to the genome average. Sliding window analyses of Tajima's *D* across the *Rcr3* coding region reveal positive values at the intron and at positions in the coding region (Figure 3.5).

Additionally, the flanking regions of both *Rcr3* loci exhibit highly significant Tajima's *D* values compared to the expected neutral distribution estimated via coalescent simulations based on a set of 14 reference loci. This may indicate that balancing selection was operating at the *Rcr3* locus prior to its duplication and formation of the gene family. Although initially the newly duplicated locus would have only copied a single sequence type (allele) from *Locus A*, it is likely that via gene conversion or recombination the second sequence type was 'exported' from *Locus A* to *Locus B* (Figure 4.1). The evidence for frequent gene conversion between loci and the maintenance of both sequence types at both loci is consistent with this scenario. However, assuming that balancing selection was operating to maintain two sequence variants, recurrent gene conversion between *Locus A* and *Locus B* may have effectively whittled down the region of linked variation, minimizing the signature of balancing selection within a locus. Consistently, the signature of balancing selection is stronger in the 3’ flanking region, where gene conversion occurs less frequently.

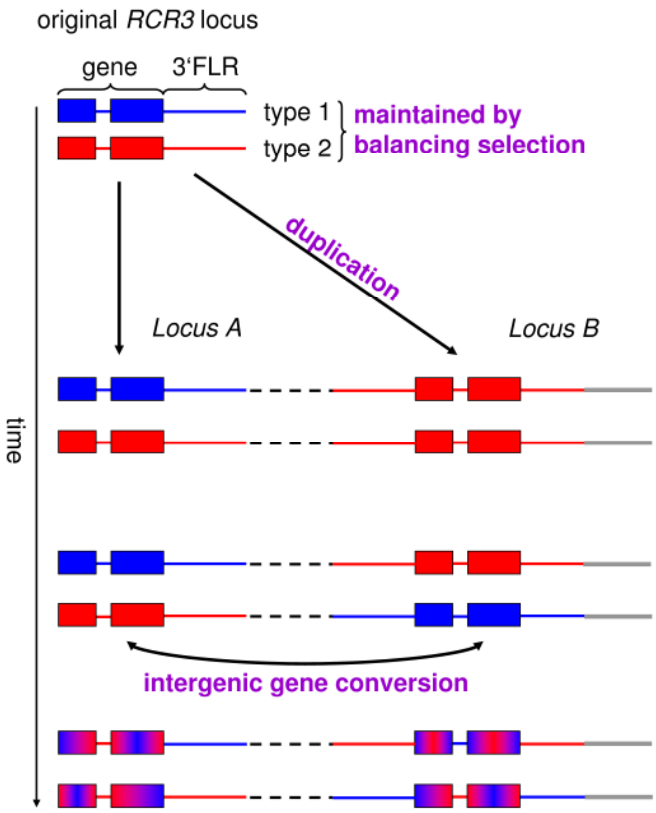


Figure 4.1: Scheme of the proposed scenario of *Rcr3* gene family evolution. It is assumed that two sequence types were maintained in the population by balancing selection prior duplication. The duplication event introduced one of these types (here the red type) into a new genomic location (*Locus B*, indicated by grey line). Subsequently, the second sequence type was exported from the original *Rcr3* locus (*Locus A*) to *Locus B* via gene conversion (black dotted line). Frequent gene conversion between the two duplicates homogenized the ORF, but not the 3’ flanking region (FLR).

4.2 *Rcr3* – a multi-tasking disease resistance gene?

4.2.1 Phenotypic basis of balanced polymorphism is not due to interaction with Avr2

Based on the presence of positive Tajima's *D* values concentrated in the intron of the *Rcr3* gene, three potential targets of selection can be envisioned: 1) selection on different regulatory motifs in the intron, 2) selection for different splicing variants or 3) selection on one or more amino acid polymorphism(s) in linkage with the intron. Since the first two alternatives could experimentally be excluded (see Chapter 3.2), it is most likely that balancing selection is acting on amino acid polymorphism(s) linked to the intron.

To evaluate functional differences between sequence types, I took a four-pronged approach. Using an over-expression vector *in planta*, I first evaluated whether protein accumulated equally from all sequence types. Of the 54 different alleles tested, protein was not detected for seven variants by Western Blots with Rcr3-specific antibodies. In all of these cases in which no protein accumulated, the candidate (putatively deleterious) substitutions could be identified. Since these seven alleles appear to be pseudogenes, they were excluded from population genetic analyses.

The second assay was a protease enzymatic assay. I found that all 47 alleles with detectable protein levels encoded functional proteases. The third and fourth functional assays were to detect differences among alleles in their sensitivity to Avr2 and in their ability to elicit HR upon co-infiltration with Avr2 into *Rcr3*-mutant tomato plants. Most of the tested alleles were inhibited by Avr2, resulting in the activation of the defence response *in planta*. The amino acid polymorphism (N194D) is significantly associated with variation in inhibition by Avr2. Alleles, which carry the N194D mutation, fail to activate the defence response *in planta,* in the presence of Avr2. This supports previous results using site directed mutagenesis by Shabab *et al.* (2008).

Due to the large sample size used in this study (54 alleles), I had sufficient power to detect epistatic interactions between amino acid variants. One such case is the presence of a second substitution (R151Q) in an allele carrying the N194D mutation. All other alleles with the D variant at site 194 failed to be inhibited by Avr2. However, the single allele with the Q variant at site 151 and D variant at site 194 was inhibited by Avr2, implicating potential epistatic interactions between these two polymorphisms.

Two additional polymorphisms were associated with variation in sensitivity to Avr2: a synonymous polymorphism at bp 717 and a nonsynonymous polymorphism at bp 750. However, since alleles, which have the polymorphism at bp 750 (R213S), but not N194D can be inhibited by Avr2 and elicit HR *in planta,* it is likely that the association between phenotype and sequence variation for this polymorphism is due to linkage disequilibrium.

In this data set encompassing nine *Solanum* species, the amino acid substitution N194D was found exclusively in individuals of *S. peruvianum* and *S. chilense*. Since the substitution was not detected in any other closely related species or the outgroup *S. lycopersicoides*, this suggests that this mutation is derived. However, as only few individuals carry this mutation, it is unlikely that the maintenance of these *Rcr3* variants accounts for the evidence of a balanced polymorphism at the *Rcr3* locus.

4.2.2 Interspecific functional diversification at the *Rcr3* locus is likely, but the balanced polymorphism is not based on recognition of differential pathogen effectors

Since Rcr3 is known to be targeted by several effectors secreted by different pathogens, one possibility could be that the balanced polymorphism at the *Rcr3* locus underlies differential recognition of different effectors. To test this hypothesis, I investigated the effect of three additional effectors on Rcr3 protease function and performance *in planta*: Epic1, Epic2B and Rip1. These additional effectors secreted by different pathogen species differed in their interaction with different Rcr3 variants.

Nearly all except nine Rcr3 variants were not inhibited by Epic1, while the opposite was true for Epic2B, *i.e.* most but four variants were inhibited. In these cases the potentially causative mutations could be pinpointed. There seems to be an overlap between the mutations causing sensitivity to inhibition by Epic1 and insensitivity to inhibition by Epic2B. In both cases, the substitutions H148N and D283N were found to be significantly associated with the alternate phenotype. Both effectors are secreted by the same pathogen, *P. infestans*. Since these effectors belong to the same effector gene family and are believed to originate from gene duplication, it is possible that they contribute to different host specificities (Tian *et al.* 2007). For example, it is possible that the two effectors employ complementary functions and are selectively maintained in the pathogen population. Perhaps, natural selection on the effector target in the host led to changes at the Epic interacting site, which in turn led to compensatory changes in the effector. Duplication of the effector gene in the pathogen and subsequent functional diversification in both interacting effector types would therefore be advantageous for the pathogen (Michelmore & Meyers 1998).

So far it has only been demonstrated that Rcr3 plays a role in tomato resistance to *P. infestans*. The mechanism of this resistance, however, has not been clarified yet. In this study, I show that inhibition of Rcr3 by Epic1 or Epic2B does not cause HR in tomato plants expressing Cf-2. The reason for this result may be that inhibition of Rcr3 by Epic1/2B does not activate the defence response. As known from experiments using the protease inhibitor E-64, inhibition of Rcr3 alone is not sufficient to elicit the Cf-2 dependent defence response (Rooney *et al.* 2005). Putative conformational changes in Rcr3, which may be caused by inhibition through Avr2, elicit the Cf-2 dependent defence response (Krüger *et al.* 2002). It is possible that inhibition by Epic1/2B does not cause the same conformational changes and another mechanism between Rcr3 and Epic1/2B is responsible for the increased resistance of Rcr3-plants to *P. infestans*. Alternatively, the defence response activated by Epic1/2B may not be Cf-2 dependent, but could rely on other molecules, which were not present in the tested plant genotype.

Although most tested Rcr3 variants did not interact with Rip1, eleven Rcr3 variants were inhibited by this effector. However, a role of Rcr3 in resistance of tomatoes to *P. syringae* has not been demonstrated so far. Rip1 is a protease inhibitor and has been detected through interaction with the Rcr3-like protease C14, its host target in potato (M. Ilyas personal communication). In tomato C14 is one of the papain-like cysteine proteases residing in the tomato apoplast (like *Rcr3*) and has significant sequence homologies to *Rcr3* (Shabab *et al.* 2008). Therefore, it is likely that the Rip1 effect on Rcr3 is merely a side effect and happens because Rcr3 and C14 are related to one another. When Rcr3 alleles were coinfiltrated together with Rip1 in tomato plants expressing Cf-2, no HR phenotype could be observed. The reason for this could be that inhibition of Rcr3 by Rip1 does not activate the defence response because it does not cause the same conformational changes as Avr2 (see paragraph above). Alternatively, the defence response activated by Rip1 is not Cf-2 dependent, but relies on other molecules, which were not present in the tested plant genotype.

No nucleotide positions within the *Rcr3* gene could be significantly associated with inhibition by Rip1. The reason therefore may be that Rip1 functions differently from other protease inhibitors such as Avr2 or Epic1/2B. These three inhibitors seem to target very specific sites within the Rcr3 protease domain. Alteration of these sites will immediately prevent inhibition of protease activity by these effectors. Rip1 binding might be less specific (perhaps because Rcr3 is not its original target) and depend more on the entire structure of the protease. Many amino acid residues in the Rcr3 protease domain may interact epistatically and in different combinations and create a protein structure resembling the C14 structure, which by chance can be targeted by Rip1.

Mutations that are associated with differential behaviour of Rcr3 alleles regarding interaction with the different effectors were found among the different tomato species, but also within *S. peruvianum.* It is of course possible that this functional divergence between species is due to diversifying selection at the *Rcr3* locus as recently suggested (Shabab *et al.* 2008). Rcr3 would then adopt different functions in different lineages depending on the corresponding environmental conditions, *i.e.* the corresponding pathogen pressure and effector repertoire. This hypothesis may apply in this study to the tested aspects of Rcr3 function. For instance, differential sensitivity for inhibition by Avr2 was only found in individuals from the sister species *S. peruvianum* and *S. chilense*, but not in individuals from other tomato species or the outgroup *S. lycopersicoides* (Figure 4.2). The mutation was caused by the same nonsynonymous substitution and may likely have arisen in the common ancestor of the two species. Alternatively, the presence in both lineages could also be explained by convergent evolution. It is possible that this mutation confers selective advantage in the environmental conditions these two species encounter or has been introgressed by gene flow, but does not have the same effect in other species. Also, the fact that the mutation was only found in *S. peruvianum* and *S. chilense* could be due to the sampling effect, because more individuals from *S. peruvianum* and *S. chilense* were sampled than individuals from the other species. Thus it is not certain whether this mutation occurs in other species as well.

Concerning inhibition by Epic1, the alleles, which are inhibited, originate mostly from the species *S. habrochaites*, *S. pennellii*, *S. chmielewskii, S. chilense* and the outgroup *S. lycopersicoides*, but in general most Rcr3 alleles are not inhibited. Among the large dataset from *S. peruvianum*, only few alleles were inhibited by Epic1. Since this phenotypic characteristic is found in several species (from which a small sample size was recovered) and in the outgroup, but not abundant in the recently diverged tomato species, it is likely that the state of being inhibited is thus ancestral at *Rcr3* alleles. Insensitivity to Epic1 inhibition would then be the derived state. This may have arisen because it gives a potential selective advantage, perhaps in pathogen resistance. In summary, the functional tests with different Rcr3 alleles uncovered inter-species phenotypic variation regarding interaction with four different effectors, but this variation is unlikely the cause for the signature of balancing selection observed in *S. peruvianum.*

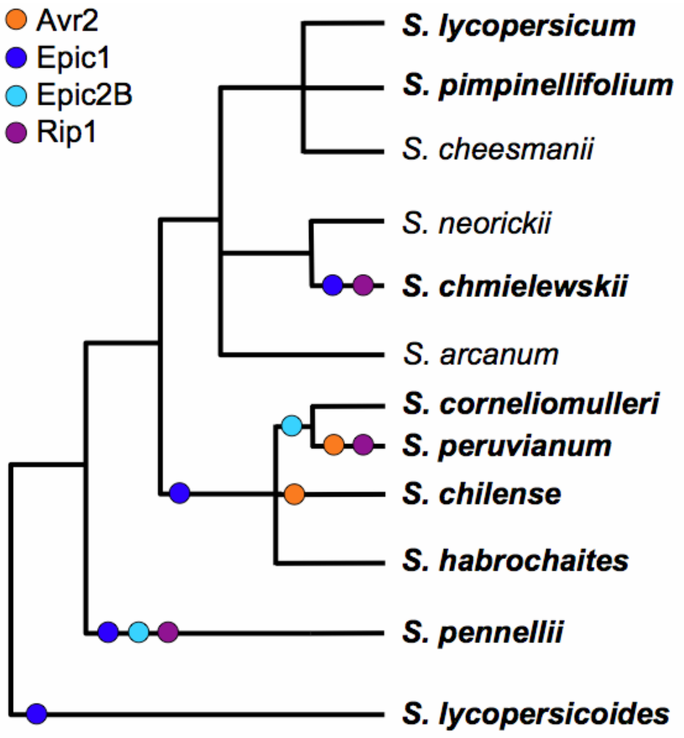


Figure 4.2: Phylogeny of the tomato clade showing differential behaviour of Rcr3 variants. The tree was adopted from Peralta *et al.* (2008). Alleles were tested from species indicated in bold. The coloured dots indicate differential phenotypic behaviour regarding the different tested effectors. Differential behaviour is defined as behaviour being different from the phenotype as known from the cultivated tomato *S. lycopersicum*.

4.2.3 The balanced polymorphism is based on differences in activation of the defence response

According to the Guard-Hypothesis the defence response relies upon two different events: modification of the guardee by the effector and activation of the defence signalling through the guard molecule (Dangl & Jones 2001; Jones & Dangl 2006; van der Biezen & Jones 1998). The *in vitro* results did not reveal substantial differences in the interaction between different effectors and guardee. However, the *in planta* assays enabled me to also investigate the second step of defence response activation. Indeed, *Rcr3* alleles differ substantially in the strength of the defence response they elicit *in planta*. Variation in this trait is significantly associated with three amino acid polymorphisms at nucleotide positions 728, 775 and 1099 encoding the amino acid changes I 206K, Q222E and S330A and two synonymous changes at positions 102 and 144 (Figure 3.14). In a structural model of the Rcr3 protease domain, these three substitutions are located in close proximity to positions with importance in protease function or compatibility between Rcr3 and Cf-2 (Figure 3.15, Krüger *et al.* 2002; Shabab *et al.* 2008). All five mutations correlated with the weak HR phenotype are in linkage with one another and with the intron, despite frequent gene conversion at the locus and are responsible for the peculiar sequence type structure (Figure 3.14, B7). These are the positions showing positive Tajima's *D* values in the sliding window analysis and may thus contribute to the signature of balancing selection at the *Rcr3* locus. It is possible that one or all of the amino acid changes is involved in the variation in the defence response. The other two polymorphisms (both at synonymous sites) are not likely the causative mutations and rather in linkage disequilibrium with the amino acid polymorphisms.

Additionally, to complete the link between genotype to phenotype I infiltrated the effector Avr2 into wild tomato plants grown from seeds, which were collected at the same sites as the plants I used for the population genetic and functional assays. In doing so, I tried to recover similar *Rcr3* genotypes to those that had previously been tested in the functional assays previously. In these assays, no HR phenotype was observed even at 10 days after inoculation with the effector. Several reasons might explain this observation. First, it is possible that Rcr3 was not expressed in these plants and therefore no Avr2-Rcr3 interaction could take place to elicit the defence response. Second, these plants may not carry or express functional Cf-2 molecules, which would prevent activation of the defence signalling. Third, other molecules in the pathway leading to HR are absent. Fourth, other stimuli such as PAMP-triggering are needed to activate the full defence machinery including ETI. Fifth, Rcr3 expressed in these plants is insensitive to the Avr2 used in this study and therefore no defence response is activated. It remains to be demonstrated which of these five explanations is the likely cause of these results.

Previous studies on *R* gene evolution demonstrated the maintenance of variation for pathogen recognition (for example in the case of *Pto*) or presence/absence polymorphism (for example in the case of *Rpm1*) (Bakker *et al.* 2006b; Bergelson *et al.* 2001b; Rose *et al.* 2004; Stahl *et al.* 1999). In contrast, the balanced polymorphism in the case of the *Rcr3* locus, – a guardee, not an *R* gene – seems to have an effect on the activation of the defence response. A possible explanation of this effect might involve a resistant/susceptible polymorphism at the *Rcr3* locus (Tellier & Brown 2007b), associated with a potential cost of resistance (Stahl *et al.* 1999; Tian *et al.* 2003). Alternatively, the diversity measured at the *Rcr3* locus might also be due to coevolution of *Rcr3* with allelic types of *Avr2* or of other pathogen effectors, which were not tested in this study. Based on the results revealing differences in the outcome of the defence response – the HR – the attenuated defence response may be a by-product of the interaction between Rcr3 and Cf-2. This interaction transmits the signal upon pathogen recognition and initiates the defence response. Members of the *Rcr3* and *Cf*-2 gene families seem to be tightly coevolving. This interaction requires a precise matching between allelic variants. A mismatch, such as when a variant of Rcr3 from one *Solanum* species is paired with a variant of Cf-2 from another but nevertheless closely related *Solanum* species, results in an autonecrotic response (Krüger *et al.* 2002) and could be an example of Dobzhansky-Muller incompatibility between tomato species (Bomblies *et al.* 2007). An attenuated response due to incompatibility between guard and guardee may decrease the risk of auto immune responses and can therefore be advantageous when the corresponding pathogen is absent (Ispolatov & Doebeli 2009). Since both genes evolve in gene families, it may be advantageous for a plant to have different alleles of both genes, which would form matching partners. One of the amino acid changes (Q222E) associated with the attenuated HR phenotype differs between Rcr3 from *S. lycopersicum* and *S. pimpinellifolium* and could potentially contribute to the reported incompatibility between Rcr3 and Cf-2 likely causing attenuated HR in incompatible genetic backgrounds (Krüger *et al.* 2002). In this study, most of the individuals carry both Rcr3 phenotypic types. Since I tested all Rcr3 alleles in identical genetic backgrounds, some may not be matched with their optimal Cf-2 partner, explaining attenuated response for some pairings of Rcr3 with Cf-2.

This combination of population genetic, computational and statistical methods and functional assays on the molecular level provided a more complete understanding of the role of the guardee in plant immune system evolution. I show that natural selection appears to maintain diversity at the effector target *Rcr3* through balancing selection and gene duplication. These are mechanisms, which have been shown to play a key role in *R* gene evolution (Michelmore & Meyers 1998; Stahl *et al.* 1999). This study reveals that in contrast to *R* gene evolution, evolutionary forces shaping the guardee rather act on the guard-guardee than on the guardee-effector interface and propose that diversity at genes involved in immunity is not only created by natural selection for pathogen recognition, but also for improved transduction of the defence signal or avoidance of autoimmune response.

4.2.4 Does *Rcr3* evolution follow a Birth-and-Death process?

As *Rcr3* evolution is characterized by gene duplication, one may expect functional divergence of different *Rcr3* copies following the Birth-and-Death Hypothesis. This would be the case, if differences in Rcr3 function were associated with the different gene copies. Even though there are functional differences between *Rcr3* alleles within the population, these do not display a copy-specific pattern. Considering the high sequence similarity between the copies and the presence of frequent gene conversion suppressing divergence of the copies, it is likely that *Rcr3* – if at all – is only at the beginning of such a Birth-and-Death process. Only then, when copies start diverging on the sequence level and gene conversion is less likely to occur, functional divergence of copies may be the case (Innan 2003b). This process may be accelerated by pseudogenization of single alleles, because these can rapidly acquire new mutations with potential functional effect (due to relaxed constraint) and these new mutations can be exported into functional copies via gene conversion. However, it has to be noted that all these processes (gene duplication, gene conversion, mutation) are stochastic and it is as likely that pseudogenization is followed by loss of entire copies. At the present state, it is not possible to predict the fate of the *Rcr3* gene family from these data, but I expect that considering the high rate of gene conversion the rate of divergence on the protein level of *Rcr3* copies would be very low.

Alternatively, functional divergence may occur on the gene expression level (Beisswanger & Stephan 2008). Since, the 3' flanking regions of the gene do not undergo frequent gene conversion, mutations in regulatory regions with an effect on gene expression may accumulate and get fixed in one of the copies. Differential gene expression may for example cause temporal variation in expression of the different copies, which might be advantageous during the time course of infection. It remains to be demonstrated if expression differences between *Rcr3* copies occur and if so, individuals with more than one expression type have fitness advantages.

4.3 Contrasting evolutionary forces act on the *Pto*-mediated disease resistance pathway

In the study of five genes involved in the *Pto* signalling pathway, two loci *Pto* and *Pfi* exhibit elevated amino acid polymorphism consistent with balancing selection. A third gene, *Prf*, shows signatures of both balancing selection and purifying selection, while two other genes, namely *Fen* and *Rin4*, show predominantly purifying selection. Previous studies have reported that *Pto* is subject to balancing selection within different wild tomato species and, given the substantial functional information available for *Pto*, the occurrence of balancing selection is not surprising (Rose *et al.* 2005; Rose *et al.* 2007). Pto binds and recognizes two different pathogen ligands and triggers a defence response in wild tomato. Thus, the maintenance of different host resistance proteins in natural populations is consistent with an on-going coadaptation between host and pathogen.

The second gene that shows elevated amino acid polymorphism relative to neutral expectations is *Pfi*. This gene is thought to be located further 'downstream' in the signalling sector and to act as a negative regulator of defence (Tai 2004). The protein product of *Pfi* physically interacts with Prf, has a putative nucleus localization signal and is predicted to encode a transcription factor. As such, it may respond upon activation of Prf by moving into the nucleus. There, it may mediate the downstream resistance responses including the hypersensitive response. As a component of the signalling sector, rather than a known pathogen target, it may seem surprising to uncover a signal of balancing selection at *Pfi*. The signature of balancing selection is located in a region that encodes a putative hydrolase domain, although enzymatic assays to confirm hydrolytic activity have yet to be conducted. Provided this molecule is enzymatically active, it is possible that natural selection operates directly on the enzymatic function and that protein variation is maintained in this region as a result of selection for different substrate specificities, perhaps involved in pathogen defence. Alternatively, this molecule could serve as a direct target by other tomato pathogens. Recent studies reveal that all proteins in plant (reviewed in Brodsky & Medzhitov 2009) or animal (Hajishengallis & Lambris 2011) immune signalling pathways can be vulnerable to pathogen manipulation. Pathogens may specifically secrete proteins (*i.e.* effector molecules) to target downstream points in the pathway to suppress host resistance (reviewed in Zhou & Chai 2008). Since Pfi is thought to be a negative regulator of defence, alteration of protein stability could result in suppression of the hypersensitive response. A third hypothesis is thus that balancing selection may not be specifically operating on enzymatic function, but rather on pathogen evasion. Alternative forms of Pfi found in these natural populations may vary in their ‘resistance’ to manipulation by pathogen molecules.

Prf, one of the central molecules of this pathway, shows two distinctive signals of natural selection. The region known to physically interact with Pto and Fen shows elevated amino acid polymorphism, providing the first hints that balancing selection at *Pto*, may be carrying over to its interacting partner, *Prf*. Such correlated selective histories suggest that more complex forms of selection, such as epistatic selection between molecules, may be an important force of immune pathway evolution. Evidence for epistatic selection between *Pto* and *Prf* has been found at candidate positions (Grzeskowiak 2009). In comparison, the C-terminus of this gene shows greater evolutionary constraint, consistent with its presumed role in signal transduction.

The *Fen* and *Rin4* genes show the greatest evolutionary constraint of these five genes. Although Fen is known to interact with some pathogen ligands, no resistance function similar to that of Pto has been assigned to this gene. If Fen is nevertheless involved in disease resistance, the strong evolutionary conservation observed at this locus would be consistent with a role of this molecule in basal rather than in isolate-specific defence. Molecules, which are involved in basal defence, are thought to be subject to different evolutionary forces than molecules known to be involved in isolate-specific defence, such as *Pto*, because of the putative conservation of pathogen perception in broad spectrum immunity. One such molecule that is known to contribute to basal defence and is involved in different resistance pathways (at least in *Arabidopsis thaliana*) is *Rin4* (Axtell & Staskawicz 2003; Kim *et al.* 2005a; Kim *et al.* 2005b; Mackey *et al.* 2002). Strong protein conservation at *Rin4* can be observed in the Tarapaca population. However, the frequency spectrum of mutations and pattern of LD among *Rin4* alleles reveal additional aspects of the history of *Rin4*, including the presence of a young, but divergent *Rin4* allelic type, carrying several derived mutations (Figures 3.20-22). One possible explanation for this pattern would be that this divergent *Rin4* allele is sweeping through the population as an advantageous allele. However, capturing a selective sweep in progress within a local population is quite unlikely because the sojourn times of advantageous alleles are generally too short. The fact that this *Rin4* allele with several derived changes is segregating with two other distinct alleles, all at moderate frequency, and none of these three allelic types show any evidence of recombination, indicates that the frequency spectrum of these alleles has been perturbed in the recent history of this plant population. This pattern of variation may be consistent with the ‘traffic hypothesis’ put forth by (Kirby & Stephan 1996) (1996). Here two or more sites experience positive selection, but are found on different haplotypes. The fixation process is slowed down until recombination can bring the adaptive mutations together into one haplotype. Competition between these alleles until a recombination event occurs will prolong the polymorphic phase and allow the detection of a sweep 'in progress'. If this is occurring at *Rin4*, it may be expected that the sweep would proceed once recombination takes place. Following a sweep, it may be possible to detect the fixation of an advantageous allele at *Rin4* through the elevation of amino acid substitutions along the lineage leading to *S. peruvianum*. Past sweeps at *Rin4* would be potentially detected if an elevated substitution rate at nonsynonymous sites between *S. peruvianum* and other species, in combination with a reduction of variation at *Rin4* within S. *peruvianum* would be observed. However, no evidence for recurrent selective events at *Rin4* in the history of this tomato population is found so far. This may indicate that sweeps at this locus are fairly rare and the predominant form of selection for *Rin4* is purifying selection, with the occasional sweep of a novel allele.

Evolutionary genetic approaches are now being applied more broadly to study groups of interacting genes, rather than single genes in isolation. Some of the first studies in plants indicated that genes located upstream in biochemical pathways showed the greatest protein conservation due to selective constraint, in comparison to downstream genes (Lu & Rausher 2003; Rausher *et al.* 2008; Rausher *et al.* 1999). Recent studies of 40 genes in the terpenoid pathway from a range of angiosperms also found slower evolutionary rates in upstream genes than in downstream genes (Ramsay *et al.* 2009). Although, signalling pathways and biochemical pathways may operate under similar rules regarding pleiotropy, the pleiotropy gradient in signalling pathways may be ‘inverted’ relative to what is observed in biochemical pathways: the genes with the greatest pleiotropy may be located further downstream rather than upstream, if they serve as convergence points for different host signals. Knowledge on signalling pathways involved in disease resistance has recently started to advance. Katagiri & Tsuda (2010) proposed that different signalling pathways in plants interact with each other and are connected by genes functioning as hubs. As a consequence the authors suggest referring to these pathways as signalling sectors of a network rather than as pathways because this would imply relative independence. Studies in *Arabidopsis* using knock-out mutants for genes located at different positions within defence signalling pathways, could demonstrate that different signalling sectors (*e.g.* jasmonate signalling, ethylene response, salicylate signalling, Pad4 sector) act in a synergistic or compensatory fashion. In addition, it is shown that many input sectors, which are involved in pathogen recognition in PTI as well as in ETI use the same signalling machinery (Sato *et al.* 2010). The outcome of the immune response, *i.e.* its strength or speed, therefore does not depend on the nature of the genes of the signalling machinery, but rather on how and when these are expressed and used and where exactly the input signal is fed into the network (Katagiri & Tsuda 2010). This may mean that the immune response network presents a core of most likely conserved genes because signals from the different input sectors converge and are processed by these key genes. The input sectors however, may evolve under different constraints depending on their actual role in pathogen recognition (*e.g.* PAMP or effector recognition).

Genes at proximal points of signalling pathways for pathogen defence, *i.e.* molecules at the interface with pathogen effectors, may thus well be expected to experience adaptive evolution. A few recent studies in *Arabidopsis* have evaluated a number of defence genes, some of which are known to operate together in specific signalling pathways (Bakker *et al.* 2006b; Bakker *et al.* 2008). Although these studies were not explicitly designed to test the effect of pathway position on evolutionary rates, the combined analysis of 27 *R* genes and 27 downstream defence genes in *Arabidopsis* by Bakker and colleagues (2006, 2008) revealed that while some *R* genes showed histories of transient balancing selection or partial selective sweeps, genes further downstream experienced almost exclusively purifying selection. At a broad scale, these results are consistent with expectations that genes in the core of defence signalling networks experience greater evolutionary constraint and genes interacting with pathogen ligands are subject to adaptive change. However, a subset of these same genes was recently evaluated more extensively by another team and they came to slightly different conclusions (Caldwell & Michelmore 2009). In a study of 10 downstream defence genes in *A. thaliana*, three genes (*NPR1, EDS1* and *PAD4*) showed interesting patterns of past adaptive evolution. This signature of balancing selection in these three genes may have been missed by Bakker and colleagues (2008) because in the original study only portions of the coding regions were analyzed, rather than the entire genes. Interestingly, a fourth gene in the Caldwell and Michelmore (2009) study overlapped with one in the present study, namely *Rin4*. In their initial analyses, *Rin4* was identified as a potential outlier based on HKA tests, but the results were inconclusive following correction for multiple testing. Nevertheless, the authors did highlight that *Rin4* harbours substantial silent polymorphism within *Arabidopsis*, displaying more genetic variation than found at 93.5% in a set of 355 reference loci (Caldwell & Michelmore 2009). To what degree this elevation in genetic diversity reflects past selective events, has not been investigated.

Compared to these other studies, this study does not find a strong correlation of selective constraint and pathway position. This may be a result of pathway length, since longer (linear) pathways usually result in stronger correlations with functional constraint (Ramsay *et al.* 2009), or more ‘network like’ organisation of the pathway. Perhaps, the present choice of genes captures only the very proximal part of the signalling sector and therefore does not include genes analogous to those reported in previous studies. As more genes ‘downstream’ in the *Pto* signalling pathway are identified, analyses could be extended to include these. Alternatively, the lack of correlation between pathway position and selective constraint may reflect the biological reality that genes at several points in defence networks can be targets of adaptive evolution and may in turn become proximal ends of input sectors, because they are targets of pathogen effectors. Consequently, population genetic studies such as the present one can uncover interesting candidates for future functional studies. For example, the consequences of Rin4 protein polymorphism on *Pto*-specific resistance and possibly basal defence responses could be tested using methods presented recently by Luo and colleagues (2009). Likewise, a better understanding of the functional consequences of protein polymorphism around the enzymatic core of the Pfi protein will likely reveal novel aspects of the defence repertoire of plants, since although this gene displays a signature of balancing polymorphism similar to other resistance genes in plants, this gene does not share the motifs of most other *R* genes.

4.4 Repeated adaptive introgression or ancestral polymorphism: signatures of balancing selection at resistance genes in two closely related wild tomato species

The species wide sampling approach provided insights into the general evolutionary history of the three studied tomato species *S. peruvianum*, *S. chilense* and *S. corneliomulleri* and into the evolution of three resistance genes (*Pto*, *Pfi* and *Rin4*) on a long-term evolutionary time scale, extending beyond species boundaries. This study reveals that *Pto* shows evidence of trans-species polymorphism most likely due to recent introgression of favourable alleles, where *Rin4* displays a strong pattern of purifying selection. However, *Pfi* exhibits a genomic signature of balancing selection within species, and enhanced divergence between species suggesting that this gene may represent a potential example of Dobzhansky-Muller incompatibility.

Table 4.1: Summary of the results obtained with the species wide sample approach.

	Pto	***Rin4***	***Pfi1***
pattern within species			
neutral diversity	like genomic mean or lower (*S. corneliomulleri*)	like genomic mean or elevated (*S. peruvianum*)	like genomic mean or elevated (*S. chilense*)
nonsynonymous diversity	high	like genomic mean	high
Tajima's *D*	- positive in *S. peruvianum* and *S. corneliomulleri* - elevated in *S. chilense* - excess of intermediate frequency polymorphism	- lower in *S. chilense* - like genomic mean in *S. peruvianum* - elevated in *S. corneliomulleri* - excess of low or high frequency polymorphism	- elevated in all species - excess of intermediate frequency polymorphism
divergence to outgroup			
neutral	lower than genomic mean	higher than genomic mean	like genomic mean
nonsynonymous	higher than genomic mean	lower than genomic mean	higher than genomic mean
between species			
F_{ST} *S. chilense*-*S. peruvianum*	lower than genomic mean	lower than genomic mean	lower than genomic mean
F_{ST} *S. chilense*-*S. corneliomulleri*	lower than genomic mean	lower than genomic mean	lower than genomic mean
F_{ST} *S. peruvianum*-*S.corneliomulleri*	higher than genomic mean	lower than genomic mean	like genomic mean
mismatch distribution	shifted towards fewer differences	like reference loci	shifted towards more differences
JSFS	many shared polymorphisms, especially nonsynonymous, private and shared polymorphisms are in intermediate frequency	no shared nonsynonymous polymorphisms, excess of singletons at shared and private polymorphisms	slightly more shared nonsynonymous polymorphisms than reference loci, shared and private polymorphisms mainly in intermediate frequency
across species			
diversity	high	low	high
Tajima's *D*	elevated compared to reference loci	slightly elevated	lower than reference loci

4.4.1 Information on the species' history

The analysis of the reference dataset, which was used in the study as genomic background, reveals general results about the studied species *S. peruvianum*, *S. chilense* and *S. corneliomulleri*.

First, based on the reference loci a clear distinction is seen between the two sister species *S. peruvianum* and *S. chilense*. The phylogeny reveals two separated clades for each of the two species, such that only few polymorphisms are shared between them and only few interspecific sequence pairs can be detected in the mismatch distribution. These findings are consistent with previous phylogenetic, population genetic and ecological studies, which demonstrated a monophyletic origin of the two sister species, but defined them as distinct species with differentiated habitat preferences (Nakazato *et al.* 2010; Peralta *et al.* 2008; Rodriguez *et al.* 2009; Städler *et al.* 2008; Städler *et al.* 2005). The present dataset reveals only few shared polymorphisms and few trans-specifically shared sequence pairs between the two species. According to age estimates based on fourfold degenerate sites, these sequence pairs have diverged only recently from one another suggesting a more recent common ancestor than the divergence time of the two species. This may reflect the presence of inter-specific gene flow (introgression) at a low level in accordance with results by Städler *et al.* (2008) and Städler *et al.* (2005).

Second, an intriguing pattern concerning the evolutionary history of *S. corneliomulleri* can be observed. In all phylogenies (not only at the reference loci, but also at the resistance genes), alleles from *S. corneliomulleri* do not form a distinct clade, but are always grouped together mostly with *S. peruvianum* alleles and only occasionally with *S. chilense* alleles. Within species patterns of diversity are similar in *S. peruvianum* and in *S. corneliomulleri* and the proxi for genetic differentiation between these two species measured through F_{ST} is close to zero at all studied loci. It is also always much lower than F_{ST} between *S. peruvianum* and *S. chilense* or *S. chilense* and *S. corneliomulleri*. Furthermore, all observed interspecific mismatch distributions of *S. peruvianum* – *S. corneliomulleri* comparisons are shifted towards lower numbers of pairwise differences compared to other comparisons. According to these findings, *S. corneliomulleri* would appear to be genetically indistinguishable from *S. peruvianum*. *S. corneliomulleri* was proposed as a new species only recently – mainly based on differences in geographical distribution and morphological traits (Peralta *et al.* 2005), though phylogenetic studies have yielded controversial results. According to Rodriguez *et al.* (2009), *S. corneliomulleri* forms a clearly separated clade from *S. peruvianum*, in opposition to (Zuriaga *et al.* 2009). Moreover, previous studies on the ecology of wild tomato species fail to demonstrate niche differentiation between the two species (Nakazato *et al.* 2010).

The present dataset supports the findings by Zuriaga *et al.* (2009) and Nakazato *et al.* (2010) and suggests three possible scenarios. 1) *S. peruvianum* and *S. corneliomulleri* are not distinct species (Zuriaga *et al.* 2009). 2) The rate of postdivergence gene-flow is high between *S. peruvianum* and *S. corneliomulleri*. This scenario would likely apply to species, which have diverged from one another very recently as it would be the case here. Evidence for hybridization between these two species has been provided by Spooner *et al.* (2005) through AFLP analyses. 3) The observed pattern of diversity is due to a large amount of ancestral polymorphism still shared between the two recently diverged sister species (Charlesworth 2010). Depending on the marker under investigation and on the age of the split, different levels of shared ancestral polymorphism can be found between species. It is more likely that genes, which have an important function for the plant and are thus subject to purifying selection, would exhibit more ancestral polymorphism for a longer period of time after divergence. As the studied reference genes seem to fulfil important housekeeping functions in the cell and evolve under purifying selection (Tellier *et al.* 2011), the observed intermixture of *S. peruvianum* and *S. corneliomulleri* could be due to maintained ancestral polymorphism. However, note that since it was not the main focus of this study to uncover the phylogenetic relationships within the *Solanum* genus, only six *S. corneliomulleri* alleles were analyzed and the current findings could result from a bias due to the low sample size.

4.4.2 Genetic diversity reflects the habitat range

Analysis of sequence variation at the reference loci revealed that *S. peruvianum* and *S. corneliomulleri* are more polymorphic than *S. chilense*, reflecting differences in their effective population sizes (Roselius *et al.* 2005; Städler *et al.* 2005). In fact it is suggested that the genetic structure of the two species reflects their ecological niche differentiation. *S. chilense* shows genetic, phenotypic and physiological adaptations to dry environments and ecological niche modelling revealed that its distribution is narrow and mainly determined by aridity (Fischer *et al.* 2011; Nakazato *et al.* 2010; Xia *et al.* 2010). In contrast, *S. peruvianum* is widely distributed and occupies a large range of variable habitats including very arid and rather mesic environments (Nakazato *et al.* 2010). The increased genetic variation (N_e) observed in *S. peruvianum* thus reflects its wider habitat range, *i.e.* a larger metapopulation (more demes). Other factors influencing the size of N_e, such as seed banks, are proposed to indirectly impact genetic diversity as well. It has been suggested that wild tomato species exhibit seed banks as a bet-hedging strategy to buffer environmental instability (Tellier *et al.* unpublished results). Seed banks are created through variability in germination rates of seeds, *i.e.* not all seeds germinate at the same moment and a certain proportion of seeds can remain in the soil over many generations until they germinate. If the germination rate is low, N_e and the variation in a population will increase, since genetic diversity is stored in the soil for a longer period of time. Additionally, seed banks can counteract habitat fragmentation by buffering against extinction of small and isolated populations, a phenomenon known as 'temporal rescue effect' (Honnay *et al.* 2008). This will in turn increase local N_e and contribute to variability in the metapopulation. An ABC-approach has revealed that *S. peruvianum* seeds exhibit prolonged germination rates compared to *S. chilense* (Tellier *et al.* unpublished results). This may be explained by the larger habitat variability and therefore more unstable environmental conditions of *S. peruvianum*.

4.4.3 The reference dataset is affected by demography and purifying selection

Analysis of synonymous and nonsynonymous nucleotide diversity at the reference loci revealed on average extremely low ratios of π_a to π_s. Under neutrality, the synonymous diversity should be as high as the nonsynonymous diversity and the ratio would be expected to equal one. (Tellier *et al.* 2011) showed recently that this low ratio is caused by purifying selection acting on local populations on the set of reference genes. Thereby in accordance with previous results, the selective constraints appear to be stronger in *S. chilense* and *S. corneliomulleri* than in *S. peruvianum*, where Tajima's *D* at synonymous sites is only slightly less negative than at all sites (Tellier *et al.* 2011). This finding has some implications for the present study. The reference dataset was initially chosen as genomic background to rule out demographic effects when studying genes putatively evolving under selection. Knowing that this reference dataset is also influenced by selection, comparisons to the reference genes cannot be done without prejudice. Values measured at the reference loci at synonymous sites should thus be used for comparisons because they should reflect neutral evolution with only demography. This comparison will then assure reliable detection of loci which deviate from neutral evolution under the observed demographic history.

When analyzing the set of reference loci at synonymous sites, Tajima's *D* values are usually below zero. This is reflected by an excess of singletons in the site frequency spectrum at synonymous sites compared to neutral expectations, which is particularly strong in *S.*

peruvianum. Since the influence of selection acting at synonymous sites is neglected here (but see (Parsch *et al.* 2010), it can be assumed that these results reflect pure demographic effects – in this case most likely a past population expansion. These results are in accordance with previous studies finding evidence for strong population expansion in *S. peruvianum*, but not or to a smaller degree in *S. chilense* (Städler *et al.* 2009, Tellier *et al*. unpublished results).

The analysis of the reference dataset for the species wide sample provided valuable information concerning the general evolutionary history of the studied species. It allows studying in depth the evolutionary history of genes of interest taking into account demographic history. For further analysis, three genes, which are key molecules in the Pto-mediated defence response and displayed interesting evolutionary patterns in the population based study (see Chapter 4.3) were chosen: *Pto*, *Rin4* and *Pfi*.

4.4.4 Resistant and susceptible alleles are maintained at *Pto* by balancing selection

Comparisons of summary statistics at the *Pto* locus to the reference loci support the findings of the population-based study (see summary in Table 4.1). First, π_a/π_s ratios at this locus are elevated compared to the reference loci in all three species and are even greater than one in *S. corneliomulleri*. This difference could simply be due to the fact that selective constraints on amino acid variation are not as strong at the *Pto* locus as at the reference genes. However, elevated ratios of π_a to π_s can also be a signature of balancing selection, as supported further by the results of neutrality tests. Tajima's *D* and Fu and Li's *D* values are elevated (sometimes positive) compared to the values estimated at synonymous sites in the reference dataset. The synonymous values at the reference loci reflecting only the demography are negative in all three species. Elevated Tajima's *D* and Fu and Li's *D* values at *Pto* even if they are not significant must therefore reflect the signature caused by selection. The *D* statistics applied here become positive when polymorphisms occur excessively in intermediate frequency, a signature, which is visible in the site frequency spectrum of *Pto* and can be caused by balancing selection. Results from (Rose *et al.* 2007) and the population study (see Chapter 4.3) suggested that *Pto* is affected by balancing selection in local populations of wild tomato species. Coevolution with the pathogen *P. syringae* was suggested as a likely cause of the balanced polymorphism (Rose *et al.* 2005; Rose *et al.* 2007, Chapter 4.3). The results obtained here show that the balanced polymorphism is not only due to local selective pressure, but seems to be active throughout *S. peruvianum*, suggesting that pathogen pressure may be homogenous across the whole species range.

Furthermore, the signature of balancing selection is also found in the other tomato species *S. corneliomulleri* and *S. chilense* suggesting coevolution with the pathogen also in these species. However, note that the signature of balancing selection acting at the *Pto* locus is strongest in *S. corneliomulleri* and *S. peruvianum*, but weaker in *S. chilense*. The geographical distribution of *S. chilense* is mainly defined by aridity and this species is thus characterized by adaptation to dry environments (Fischer *et al.* 2011; Nakazato *et al.* 2010; Xia *et al.* 2010). Dry habitats do not provide perfect conditions for pathogen growth and may thus reduce pathogen pressure on plants in these areas and result in attenuated levels of host-pathogen coevolution (Agrios 2005; Soubeyrand *et al.* 2009). Indeed, growth conditions for bacterial parasites such as *P. syringae* are optimal in temperate, humid regions, possibly excluding *S. chilense* as a perfect host (Agrios 2005). The weak pattern of balancing selection observed at the *Pto* gene in *S. chilense* could therefore also be an artefact of ancient selection (perhaps on the common ancestor of the two species). Additionally, the lower seed germination rate estimated in *S. peruvianum* may also contribute to the enhanced pattern of balancing selection in comparison to *S. chilense* (Tellier *et al.* unpublished results). Seed banks can cause direct frequency-dependence at genes involved in host-pathogen coevolution and can hence promote diversity at these loci (Tellier & Brown 2009). Balancing selection may therefore be more likely the case in species exhibiting low germination rates, or at least more observable in these species at the molecular level. Finally, considering that the habitat preference and genetic diversity of *S. corneliomulleri* is very similar to *S. peruvianum* and these two species have only – if at all – diverged recently, it is not surprising to find similar patterns of sequence evolution in this species. If altitude is indeed the main factor distinguishing these two species, these results may also suggest that this factor does not substantially influence pathogen prevalence. However, it must be noted that again the pattern observed in *S. corneliomulleri* could be biased by the small sample size.

Another evidence for the presence of balancing selection in this dataset comes from the phylogeny of the *Pto* gene. Both trees on the nucleotide and the protein level reveal a clustering of *Pto* alleles into two well-separated clades, which exhibit intermediate levels of polymorphism (Figures 3.24 and 3.38). The polymorphisms underlying this clustering seem to be shared between species, since alleles from all three species can be found in both clades. A closer look at the protein haplotypes at the *Pto* locus reveals a distinct haplotype structure, which is mainly present in *S. peruvianum* and might thus cause the stronger signature of balancing selection in this species (Figures 3.30-32). Interestingly, this haplotype structure underlies the amino acid variation, which is responsible for the recognition of AvrPto. The two functional positions are in significant linkage disequilibrium with all other amino acid positions causing the haplotype structure. These amino acid positions happen to be located nearby the functionally important P+1 loop in the Pto kinase domain or in close vicinity to other amino acid positions, which play a role in the maintenance of the protein structure or the active conformation of the kinase (Xing *et al.* 2007). Of course, it is possible that these positions cause the haplotype structure simply because they are in random linkage disequilibrium with the functionally important positions, which are putatively under selection. However, since these nine amino acids happen to be located around important residues in the Pto protein, it can be hypothesized that the other positions causing the haplotype structure are also functionally important and maintained in the population by balancing selection. Assuming *Pto* has a role as receptor for pathogen recognition, it is not surprising to find this genomic signature. Interestingly, the two observed main haplotypes in the *Pto* dataset correspond to two functional classes: 'recognizing AvrPto and 'not recognizing AvrPto' based on the molecular study by Xing *et al.* (2007). If recognition of AvrPto stands for resistance to *P. syringae*, one of the haplotypes would then potentially confer resistance and the other one would confer susceptibility. It is then possible that the observed balanced polymorphism is due to a resistant/susceptible polymorphism (Holub 2001; Stahl *et al.* 1999). Resistant and susceptible allele frequencies would fluctuate in the population depending on different factors such as prevalence of the pathogen carrying the corresponding effector, cost of resistance or cost of virulence (Bergelson *et al.* 2001a; Holub 2001; Tellier & Brown 2007b). Coevolution between *Pto* and various pathogen effector alleles could also promote the balanced polymorphism at the *Pto* locus.

4.4.5 Evidence for adaptive introgression at the *Pto* locus

Interestingly, the signature of balancing selection at the *Pto* locus is also visible across species when combining the datasets of *S. peruvianum* and *S. chilense*. Here, the π_a/π_s ratio and Tajima's *D* and Fu and Li's *D* values are elevated compared to the set of reference loci, the site frequency spectrum reveals an excess of intermediate frequency variants and there is clustering in the *Pto* phylogeny across the two species (Chapter 3.4). Two explanations for the finding that *Pto* is not only subject to balancing selection in one of the two sister species, but across both species, are possible: 1) Balancing selection at this locus is ancient and predates divergence of the two species. After speciation, the selective pressure on the locus might have been attenuated in *S. chilense* following its adaptation to dry environments. 2) Balancing selection is indeed acting on the *Pto* locus in both species in a similar fashion and adaptive introgression between the two species contributes to the observed pattern. The amount and type of variation shared between the two species can help to distinguish between these two scenarios. First, the rate of introgression at *Pto* seems to be increased compared to the reference loci. In the Joint Site Frequency Spectrum of *S. peruvianum* and *S. chilense*, a higher proportion of shared polymorphisms are observed and the mismatch distribution at the *Pt*o locus reveals more trans-specifically shared alleles than seen in the reference dataset. Second, according to the divergence at fourfold degenerate sites, these trans-species pairs have a very recent common ancestor and are therefore younger than the divergence time of the two species. This indicates that introgression of these alleles must have happened recently. Third, only few more nonsynonymous than synonymous polymorphisms are shared between *S. peruvianum* and *S. chilense* and alleles which are similar on the protein level (see protein tree of Pto and mismatch distribution at nonsynonymous sites, Figures 3.38 and 3.42) only differ by few substitutions at the nucleotide level. After an introgression event, the allele, which has been introgressed from one species into the other, is identical in both species at both neutral and synonymous sites. If the introgression was adaptive and the allele is kept in the population, private polymorphisms – mainly synonymous – will accumulate in either species. As few private synonymous polymorphisms have accumulated in either species, it can be assumed that the shared polymorphism is due to recent introgression rather than to ancestral polymorphism. Furthermore, the fact that shared polymorphisms occur excessively in intermediate frequency suggests that *Pto* alleles, which introgress into another species, are maintained in the population at intermediate frequencies. This scenario is similar to the case of the self-incompatibility-(S)-locus in *Arabidopsis* species, which evolves under balancing selection and undergoes adaptive introgression (Castric *et al.* 2008).

In summary, the results obtained in this study point to the *Pto* locus evolving following a balancing selection scenario probably in all three studied species. It cannot be concluded with certainty whether selection predates speciation or not. The pattern of polymorphism is affected by repeated adaptive introgression between the species (Figure 4.3).

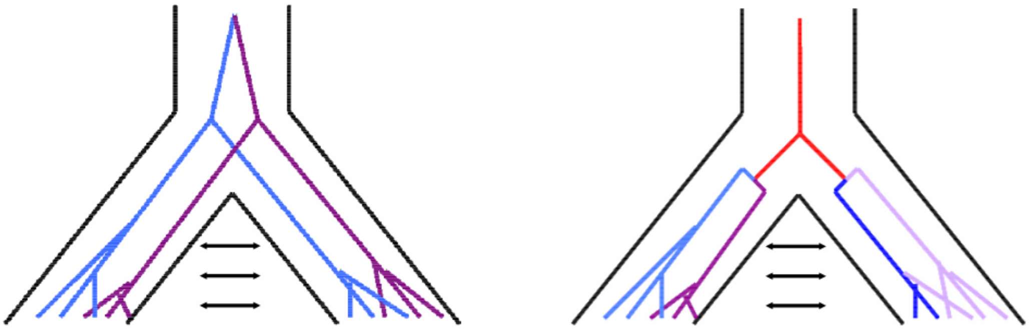


Figure 4.3: Likely scenarios explaining the evolutionary history of the *Pto* locus in *S. peruvianum* and *S. chilense*. Left: The balanced polymorphism was already present in the common ancestor of the two species. Right: Balancing selection was only active after speciation. In both scenarios, high rates of gene flow at the *Pto* locus explain the observed evolutionary pattern.

4.4.6 Strong purifying selection at the *Rin4* gene

The patterns of polymorphism observed at the *Rin4* gene are different from what can be discerned at the *Pto* and *Pfi* genes and at first glance, *Rin4* does not look very different from the reference loci based upon π_a/π_s ratios and neutrality test (*D*-)statistics (see summary in Table 4.1). These negative *D*-values are caused by an excess of low frequency polymorphisms and high frequency variants compared to neutral expectations and the synonymous site frequency spectrum measured at the reference loci. All these findings are consistent with a history of strong purifying selection at the *Rin4* locus. Deleterious mutations are quickly removed from the population and therefore kept in low frequency. Only polymorphisms, which are neutral, can reach high frequencies simply by drift (Charlesworth *et al.* 1993; Fay *et al.* 2001). According to the observed frequency spectrum and the *D*-statistics, the selective constraint on *Rin4* could be stronger than at most reference loci, as shown by the mismatch distributions which fit within that of the reference loci (slightly smaller range at *Rin4*, Figure 3.39). On the protein level, *Rin4* seems to be highly conserved, as all interspecific sequence pairs differ by only few differences, and we expect that many nonsynonymous positions should be shared between the species. Surprisingly, none of the shared nonsynonymous positions between *S. peruvianum* and *S. chilense* are derived compared to the outgroup *S. ochranthum*. This means that all nonsynonymous polymorphisms, which are derived compared to *S. ochranthum*, are private to one of the species. Private nonsynonymous polymorphisms in either species occur in low to moderate frequency, as expected under very strong purifying selection acting against most nonsynonymous mutations and keeping them in low frequency in one population (Charlesworth *et al.* 1993; Fay *et al.* 2001). It becomes then improbable that these mutations in low frequency would migrate to the other species by introgression. If such an unlikely event happens, purifying selection in the other species will most likely act against these introgressed deleterious mutations. Therefore, introgression between *S. peruvianum* and *S. chilense* is only visible at the *Rin4* locus through synonymous shared polymorphisms.

The signature of strong purifying selection is visible in all three studied species. It is also present in the dataset combining *S. peruvianum* and *S. chilense*. These findings together suggest that Rin4 has an important function in the cell and is therefore conserved across all three studied species. Indeed, Rin4 has a putatively important function as negative regulator of the basal defence response. The results obtained from the species wide sample are consistent with the results from the population-based study (see Chapter 4.3). However, in the Tarapaca population one haplotype may be potentially increasing in frequency. This haplotype is not present in the species wide sample. Although no signature consistent with a selective sweep scenario can be observed in the species wide sample, it is still possible that positive selection is only present in local samples depending on the local selective environment. Both the population-based and the species wide study presented here in this thesis reveal a similar evolutionary scenario for *Rin4* as the study by Caldwell & Michelmore (2009). The authors of this study could not reveal any evolutionary signature at *A. thaliana Rin4* differing from other nuclear genes. Caldwell & Michelmore (2009) explain this by the fact that *Arabidopsis* Rin4 is guarded by at least two different R genes (Rpm1 and Rps2), which evolve under frequency-dependent selection (Axtell & Staskawicz 2003; Mackey *et al.* 2003; Mauricio *et al.* 2003; Stahl *et al.* 1999). Coevolution between host and pathogen therefore could take place at the guard in this case, while the guarded effector target would be conserved. Note that, of course, a different set of molecules interacts with Rin4 in *Arabidopsis* than in tomato and the function of *Arabidopsis* Rin4 may be different from the function of the tomato Rin4. The evolutionary history of this gene may therefore differ between the two taxa. However, Rin4 seems to function as a negative regulator of the defence response in both *Arabidopsis* and tomato (Kim *et al.* 2005a; Kim *et al.* 2005b; Luo *et al.* 2009; Mackey *et al.* 2002). Strong conservation of this molecule may then be expected because of its regulatory function and may emphasize its importance for regulation of the defence pathway or in developmental processes in the plant rather than for coevolution with the pathogen (Igari *et al.* 2008). Therefore, none of the previously introduced scenarios (Chapter 1.9) seems to reflect the evolutionary history of the *Rin4* gene.

4.4.7 Signature of balancing selection at *Pfi* and high divergence between species

At the *Pfi* gene, observed patterns of sequence evolution are similar to those at the *Pto* locus (see summary in Table 4.1). The ratios of π_a/π_s, Tajima's *D* and Fu and Li's *D* values are elevated compared to the set of reference loci and the site frequency spectrum exhibits an excess of intermediate frequency polymorphism in all three species. The phylogeny of the *Pfi* gene and protein reveals two distinct clades, where each is formed by alleles from all three species (Figures 3.26 and 3.41). All these findings point to an evolutionary history under balancing selection at *Pfi* as also revealed in the population-based study (see Chapter 4.3). Unlike at *Pto*, the signature at *Pfi* is strong in *S. chilense* and weaker in *S. peruvianum* and *S. corneliomulleri*. This could be a consequence of differential pathogen pressure across the three species. Similarly to *Pto* however, the signature of balancing selection is visible in the dataset combining *S. peruvianum* and *S. chilense* suggesting that polymorphism has been maintained prior to species divergence or is currently being maintained in both species to a similar degree. A major difference when comparing patterns of sequence evolution at *Pfi* to *Pto* is revealed by the mismatch distribution, where interspecific sequence pairs exhibit in general more differences than sequence pairs at the reference loci (Figure 3.40). This indicates that the species – especially – *S. peruvianum* and *S. chilense* – are more differentiated from one another at the *Pfi* locus and gene flow between species occurs at a lower rate than at reference loci. This observed differentiation between species, however, is mainly due to neutral polymorphisms (synonymous sites), since in the mismatch distribution at nonsynonymous sites at the *Pfi* gene is similar to that of the reference loci at all sites (Figure 3.42). This pattern also becomes visible when looking at the phylogenies for the *Pfi* nucleotide and protein sequence. Similar alleles on the protein level between species exhibit a vast number of nucleotide differences. This means that trans-species polymorphism at the *Pfi* gene is found predominantly at the protein level and is less supported by neutral variation. In comparison to the reference dataset, more polymorphisms (nonsynonymous and synonymous) are shared between *S. peruvianum* and *S. chilense*. These shared as well as the observed private polymorphisms occur mainly in intermediate frequency. As suggested above, maintenance of polymorphism at the *Pfi* locus may be due to different factors. These factors could involve a function in the cell, which is independent from resistance, but nevertheless evolves under balancing selection. Alternatively, the signature of balancing selection at *Pfi* might be due to coevolution of this molecule with other molecules in the signalling network or with pathogen derived molecules such as effectors.

4.4.8 Possible counter-selection of gene-flow at *Pfi*

Since there is a substantial amount of shared variation at the *Pfi* locus at the protein level between species, this indicates that some of the observed variation predates speciation and has been maintained in both *S. peruvianum* and *S. chilense* ever since. The absence of trans-specifically shared alleles suggests that adaptive introgression at the *Pfi* locus occurs rarely, if at all. One explanation for this finding could be that after divergence of the two species, the *Pfi* gene accumulated private mutations in both species, which may not have substantially influenced the function of the protein or may even have been beneficial. This led to higher level of divergence at *Pfi* between the species than at the genomic average, even though functionally important polymorphisms may still be maintained in both species. Functional divergence at loci, which are involved in disease resistance, may be influenced by adaptation of each species to the species-specific pathogenic environment. Introgression of alleles originating from another species with different adaptations might prove to be maladapted in the second species' genomic background and pathogenic environment and therefore be less favoured – especially, if pathogen incidence is high. Alternatively, other molecules interacting with *Pfi*, may have coadapted to the *Pfi* alleles present in the corresponding population during divergence of the two species. *Pfi* alleles introgressed from one species to the other, may be in contact with incompatible molecules and may be thus disadvantageous for the hybrid. This effect is named a Bateson-Dobzhansky-Muller (BDM) incompatibility (Orr 1996). Since Pfi has the role of a negative regulator of the defence response, incompatibilities between Pfi and other defence related genes can potentially cause malfunctions in the immune response – *e.g.* autoimmune responses or blockage of the signal transmission. Autoimmune responses, especially when involving necrotic response, confer an immediate fitness disadvantage and have been highlighted as examples for BDM incompatibilities in plants (Bomblies *et al.* 2007; Bomblies & Weigel 2007; Ispolatov & Doebeli 2009; Wulff *et al.* 2004). Introgression of genes involved in those incompatibilities may therefore be unlikely or even counter-selected by natural selection (Ispolatov & Doebeli 2009).

In summary, the results obtained for the *Pfi* locus demonstrate that this gene evolves following a balancing selection scenario with shared ancestral polymorphism between the species. The species have diverged from each other on the nucleotide level over time. The observed sequence divergence may be due to counter-selection of gene-flow at this locus (Figure 4.4).

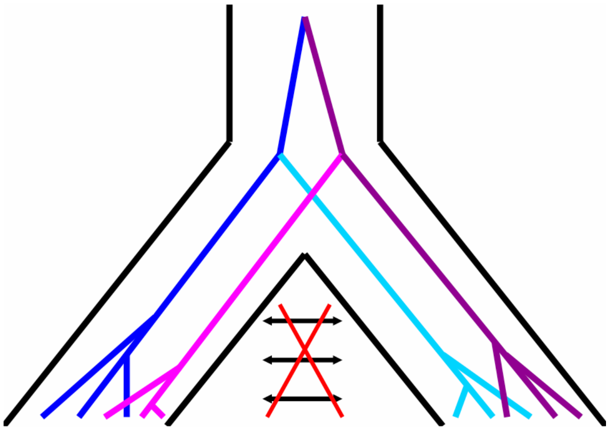


Figure 4.4: Possible scenario explaining the evolutionary history at the *Pfi* locus. The balanced polymorphism predates speciation of *S. peruvianum* and *S. chilense*. Alleles in both species have diverged following speciation. Gene flow between the species at this locus is rare or even counter-selected.

4.4.9 Comparison to the population-based study

I compare here the results of the species wide study to the results of the population-based study and to previous population-based studies of the same genes in the same set of species. Most signatures of selection, which were observed at the population level (purifying selection at the reference dataset, balancing selection at *Pto* and *Pfi*), were confirmed by the species wide study in *S. peruvianum* and could even be extended to *S. chilense* (or *S. corneliomulleri*). This may imply that selective pressures causing the observed signatures of purifying selection at reference loci are homogenous throughout the species, while acting locally within each population. This is not surprising considering that the reference loci are genes with known or likely housekeeping function (Baudry *et al.* 2001; Roselius *et al.* 2005; Städler *et al.* 2005). I suggest that it is improbable that selective constraints on these housekeeping functions differ substantially between populations of the same species and therefore the observed signature of selection within single populations should reflect the overall selection pattern within the whole species. Note however the difference in selective constraint observed between species (these results and Tellier *et al.* 2011).

Interestingly, at the *R* genes studied, the population-based study and the species wide sample yielded similar results, pointing to balancing selection as major force driving evolution at *Pto* and *Pfi*. These findings potentially indicate that the selective constraint resulting in the observed pattern is homogenous throughout the whole species or at least throughout the sampled range. However, pathogen incidence as well as parameters of coevolutionary dynamics such as parasite prevalence and disease severity are in general not a homogenous pressure, but usually harbour spatial and temporal variation within and among populations (Laine & Tellier 2008; Thompson 2005). This heterogeneity together with spatial structuring of host and parasite populations can cause direct frequency-dependent selection and as a consequence the pattern of balancing selection observed at coevolving genes in the entire metapopulation (Brown & Tellier 2011; Gavrilets & Michalakis 2008). Population structure is a key factor in promoting coevolutionary dynamics leaving the signature of balancing selection in the entire species (Burdon & Thrall 1999; Thrall & Burdon 2002). If coevolutionary dynamics occurring due to direct frequency-dependent selection are present for long periods of time, the genomic signature of balancing selection would be seen, even if local plant populations may coadapt to the local pathogenic genotypes. In this case, allele composition and allele frequencies would likely differ between populations.

Alternatively, it is also possible that the two sampling schemes do not capture the heterogeneity in pathogen pressure between populations. It is possible that in some populations and parts of the species range, pathogen incidence is high (hot spot of coevolution) but it is low or pathogens are absent in other part of the range (cold spots, Thompson 2005). If this heterogeneity subsists for a long enough period of time, the signature of balancing selection will be seen both in sequences at hot spots (*i.e.* Tarapaca population) and at the species wide level. Such possible geographic variation in pathogen pressure resulting in differential outcomes of host-pathogen coevolution has been suggested for example for the *Cf-2 R* gene in *S. pimpinellifolium* (Caicedo 2008). Note that as a corollary, balancing selection might not be detected both in samples at cold spots and the species wide level, depending on gene flow in the metapopulation.

Finally, a third plausible explanation for the observed genomic signatures could involve a selective constraint promoting polymorphism patterns resembling that of balancing selection, because it is not imposed by a pathogen but rather some cellular function of the two genes. This, however, is unlikely at least in the case of *Pto* for which a potential for host-pathogen coevolution has been demonstrated functionally (Bernal *et al.* 2005; Rose *et al.* 2005; Rose *et al.* 2007).

In the case of the *Rin4* gene, the two sampling schemes yielded different results, but these are not necessarily contradictory. A pattern of strong purifying selection was present in the species wide sample, while the pattern observed in the population-based study could have been due to either purifying or positive selection. I hypothesize that the *Rin4* gene is in general under strong selective constraint, but depending on the local environmental conditions beneficial mutations may occur occasionally and may sweep through the local population. If gene-flow between demes is low, this mutation might not immediately migrate to other demes and can as a consequence not immediately increase in frequency in the entire species (Charlesworth *et al.* 1997; Whitlock 2003). Alternatively, this new mutation may not confer a fitness advantage in other local environments and therefore be counter-selected in other populations. Both scenarios could lead to the pattern observed in the two studies: *Rin4* seems to be conserved in the entire species, but might experience positive selection in a local population. Of course, it is also possible that *Rin4* is conserved throughout the entire species and the pattern observed in the Tarapaca population is simply due to genomic peculiarities around this gene, such as low recombination rate, and/or close relationship between individuals within populations combined with low statistical power to detect outlier genes.

One great advantage of the species wide sample approach is that it allows investigating the evolutionary history of the three resistance genes beyond species boundaries. Analyzing the patterns of sequence variation between closely related species helps to understand details of their evolutionary history and to extend the evolutionary time scale under investigation. The population-based study revealed signatures of balancing selection at the *Pto* and *Pfi* genes as did the species wide sample. However, only the sampling of the collecting phase of the metapopulation coalescent tree in both *S. peruvianum* and *S. chilense* (and *S. corneliomulleri*) uncovered trans-species polymorphism at the two loci.

4.5 Conclusion and outlook: There is more in immunity beyond the *R* gene

In this thesis, I investigated genes, which are involved in the activation of disease resistance in tomatoes, but do not belong to the group of well-characterized *R* genes. The aim was to show whether the main force on the coevolutionary stage is indeed the interaction between effector and *R* gene as has been assumed, or if coevolution between plant and pathogen also takes place on other levels downstream of pathogen recognition. I addressed this question from the role of the guardee in indirect pathogen recognition and from the role of genes putatively being involved in a common signalling sector, which functions upon pathogen recognition. Results obtained through these different approaches reveal some interesting findings.

1) Balancing selection, which is assumed to be indicative of the host-pathogen interface does not only act on *R* genes, but also on other molecules, with important roles in pathogen defence. Although guardees are involved in the host-pathogen interface, it is surprising with regard to previous literature to find signatures of balancing selection, because of various constraints shaping these molecules. Similarly, other genes with supposed 'downstream function' in a signalling network also evolve under balancing selection. These findings are of importance for the understanding of (plant) immunity, and suggest that the immune system is not as simple as two levels of function: pathogen recognition and defence activation. It has long been believed that coevolution would occur mainly at genes involved in pathogen recognition (Woolhouse *et al.* 2002): in animals at the MHC (Apanius *et al.* 1997; Hughes & Yeager 1998), in invertebrates at PRR genes (Lemaitre & Hoffmann 2007; Sackton *et al.* 2007) or in plants (Bakker *et al.* 2006a; Bergelson *et al.* 2001b). Note that this paradigm has driven research to find functional similarities between plant and animal genes of recognition (Nürnberger *et al.* 2004). Recent advances, and result from this thesis, suggest that the immune system consists of different sectors that interact and are flexible, and are targeted by various parasite effectors. For instance in the case of the 'downstream' molecule Pfi, it may be possible that this molecule used to have a conserved function in signalling, but it may have become a pathogen target and consequently changed its role ending up at the host-pathogen interface. Furthermore, these findings suggest that it is not only the host-pathogen interface, which plays an important role in the entity of the immune system, but also protein-protein interactions within the host. Those can be important for fine-tuning of the defence response in either way. Overall, it can be concluded that the scientific view on the plant immune system has to

become more flexible. More studies with depth on whole defence signalling networks will help to understand the nature and flexibility of these complex interactions better (Katagiri & Tsuda 2010). Thereby, it will be necessary to approach the defence machinery on the functional level (for example in a manner like Sato *et al.* (2010) and on the population genetic level as recently performed by Obbard *et al.* (2009) in *Drosophila.* In combination, these studies will help revealing genes of importance in these networks and the mode of interaction they use, as well as the role of parasite effectors in targeting parts of the network (Hajishengallis & Lambris 2011).

2) Mechanisms of genome evolution, which are known to play a role in host *R* gene and pathogen effector evolution, are also important in guardee evolution. Here, I show that the *Rcr3* gene undergoes gene duplication and gene conversion, which may influence its evolutionary history. *R* genes can be subject to the same mechanisms and it is known that this is the case for the Rcr3 interacting partner Cf-2 (Caicedo & Schaal 2004; Dixon *et al.* 1996). This is intriguing because it suggests that coevolution between these two molecules, guard and guardee, may be extremely tight and may even extend to the level of concerted genome evolution. It should be of interest to screen different tomato species and populations for their repertoire of Rcr3 and Cf-2 molecules. This will shed light on whether mechanisms such as gene duplication are active in the coevolutionary process between these two molecules and on which evolutionary time scale (population or species level). Allelic variants and evolutionary history of the two molecules could then be associated with one another to understand the mode of interaction between the two molecules. Furthermore, combinatorial functional studies with the two forms of the protein as well as *in planta* assays will allow for a fine-scale understanding of this interaction.
3) Genes with similar signature of selection on the population level can have different evolutionary histories on a long-term evolutionary time scale. The evolutionary signature observed at the *Pfi* and *Pto* genes is similar in the population study. Only when extending the evolutionary time scale investigated beyond species divergence, great differences of the evolutionary history between the two genes are detected. For the *Pto* gene it is likely that the balanced polymorphism is directly due to coevolution with the pathogen - perhaps as resistant/susceptible polymorphism. Thereby, the evolutionary dynamics may resemble S-locus evolution. The S-locus is a well-characterized example for multi-allelic balancing selection and it has been shown that adaptive introgression is a process increasing polymorphism at this locus (Castric *et al.* 2008). Introgression may be a mechanism

promoting polymorphism at resistance loci linked to a direct fitness advantage of new alleles.

The pattern observed at the *Pfi* gene is substantially different and counterintuitive at first glance. However, long-term balancing selection does not necessarily mean that allele sharing between species is beneficial. As discussed above, the functional bases of the balanced polymorphism may be completely different from what is observed at the *Pto* locus. Overall, it can be concluded that analysis of single species can reveal interesting patterns, but one cannot be sure to uncover the complete story.

This can only be achieved through functional studies with the gene in question. For example, in a case like the *Pto* gene, it will be of interest to test different (for example putatively resistant and susceptible) *Pto* alleles together with the pathogen effectors in *Pto* lacking mutant plants as an extension of Rose *et al.* (2005). A screen for HR as performed in the study on *Rcr3* will help uncovering the basis of the balanced polymorphism at this locus. A similar approach could be applied for cases like the *Pfi* gene. Infiltration studies *in planta* using different *Pfi* alleles originating from different populations and species without pathogen effectors would for instance reveal incompatibilities between Pfi and other molecules in the plant and allow for a more complete understanding of its function.

4) In this study, comprising on total six different genes involved in pathogen resistance, at least one gene (*Rcr3*) may be involved in species incompatibility, and I suggest that *Pfi* may have a similar role, though no functional tests were performed with this gene. Taken together, this might reflect a general trend of highly concerted coevolution between components in plant immunity. Activation of ETI evolved to be rapid and efficient (Katagiri & Tsuda 2010) making it prone to potential overreaction (Bomblies & Weigel 2007). These incompatibilities are therefore merely an extension of host-pathogen coevolution. If pathogen pressure is high, the defence response is selected to be as efficient as possible. This efficiency in turn harbours the caveat of overreaction, *i.e.* autoimmune response. In absence of the pathogen, selection should therefore drive attenuation of this efficiency, which in turn will be disadvantageous if pathogen pressure increases again. In a population, this may therefore add an additional level of frequency-dependence on the interacting molecules caused by the pathogen prevalence. I suggest that theoretical studies could implement allele frequencies of the effector and the *R* gene (see Bergelson *et al.* 2001a; Tellier & Brown 2007a; Tellier & Brown 2007b; Tellier & Brown 2009), as well as those of the guardee. These models should then help to predict the outcome of this scenario depending on for example pathogen incidence at a given

geographical location. If this mechanism indeed plays a role in speciation, those models may also help to identify potential conditions for speciation, following the study of (Ispolatov & Doebeli 2009). In further steps, these predictions could be tested through sampling at geographical locations with high or low pathogen incidence (cold and hot spots in the geographic mosaic of coevolution) and genetic combinations of interacting genes could be associated to those environmental conditions. However, it is not clear yet if immune incompatibilities are rather the cause or the consequence of speciation and population divergence. If these incompatibilities are the cause of speciation they may drive the evolution of host ranges and cospeciation (Schulze-Lefert & Panstruga 2011), but if they are the consequence, they may just arise following the snow ball effect of accumulation of BDM incompatibilities (Orr 1996; Orr & Turelli 2001).

5) This work demonstrates that the combination of population genetic tools with functional assays on the molecular level allows for a more complete understanding of complex natural mechanisms. Here, it was thus possible to pinpoint single amino acids of functional importance what would not have been possible with functional studies or population genetic analyses alone. Such integrative studies give insight into various aspects of evolution and help to not only determine its mode, but also influencing factors and its outcome.

BIBLIOGRAPHY

Aderem, A. & Ulevitch, R. J. 2000 Toll-like receptors in the induction of the innate immune response. *Nature* **406**, 782-787.

Agrios, G. N. 2005 *Plant Pathology*: Elsevier Academic Press, New York, USA.

Alvarez, M. E. 2000 Salicylic acid in the machinery of hypersensitive cell death and disease resistance. *Plant Molecular Biology* **44**, 429-442.

Alvarez-Ponce, D., Aguade, M. & Rozas, J. 2009 Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. *Genome Research* **19**, 234-242.

Apanius, V., Penn, D., Slev, P. R., Ruff, L. R. & Potts, W. K. 1997 The nature of selection on the major histocompatibility complex. *Critical Reviews in Immunology* **17**, 179-224.

Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. 2006 The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195-201.

Arunyawat, U., Stephan, W. & Städler, T. 2007 Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Molecular Biology and Evolution* **24**, 2310-2322.

Ausubel, F. M. 2005 Are innate immune signaling pathways in plants and animals conserved? *Nature Immunology* **6**, 973-979.

Axtell, M. J. & Staskawicz, B. J. 2003 Initiation of RPS2-specified disease resistance in *Arabidopsis* is coupled to the AvrRpt2-directed elimination of RIN4. *Cell* **112**, 369-377.

Bakker, E. G., Stahl, E. A., Toomajian, C., Nordborg, M., Kreitman, M. & Bergelson, J. 2006a Distribution of genetic variation within and among local populations of *Arabidopsis thaliana* over its species range. *Molecular Ecology* **15**, 1405-1418.

Bakker, E. G., Toomajian, C., Kreitman, M. & Bergelson, J. 2006b A genome-wide survey of *R* gene polymorphisms in *Arabidopsis*. *Plant Cell* **18**, 1803-1818.

Bakker, E. G., Traw, M. B., Toomajian, C., Kreitman, M. & Bergelson, J. 2008 Low levels of polymorphism in genes that control the activation of defense response in *Arabidopsis thaliana*. *Genetics* **178**, 2031-2043.

Barton, G. M. & Medzhitov, R. 2003 Toll-like receptor signaling pathways. *Science* **300**, 1524-1525.

Baudry, E., Kerdelhue, C., Innan, H. & Stephan, W. 2001 Species and recombination effects on DNA variability in the tomato genus. *Genetics* **158**, 1725-1735.

Baureithel, K., Felix, G. & Boller, T. 1994 Specific, high-affinity binding of chitin fragments to tomato cells and membranes – competitive-inhibition of binding by derivatives of chitooligosaccharides and a NOD factor of *Rhizobium*. *Journal of Biological Chemistry* **269**, 17931-17938.

Beaumont, M. A. 2010 Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology and Systematics* **41**, 379-405.

Beaumont, M. A., Zhang, W. Y. & Balding, D. J. 2002 Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-2035.

Beisswanger, S. & Stephan, W. 2008 Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 5447-5452.

Bergelson, J., Dwyer, G. & Emerson, J. J. 2001a Models and data on plant-enemy coevolution. *Annual Review of Genetics* **35**, 469-499.

Bergelson, J., Kreitman, M., Stahl, E. A. & Tian, D. C. 2001b Evolutionary dynamics of plant *R*-genes. *Science* **292**, 2281-2285.

Bergelson, J. & Purrington, C. B. 1996 Surveying patterns in the cost of resistance in plants. *The American Naturalist* **148**, 536-558.

Bernal, A. J., Pan, Q. L., Pollack, J., Rose, L., Kozik, A., Willits, N., Luo, Y., Guittet, M., Kochetkova, E. & Michelmore, R. W. 2005 Functional analysis of the plant disease resistance gene *Pto* using DNA shuffling. *Journal of Biological Chemistry* **280**, 23073-23083.

Bloom, A. J., Zwieniecki, M. A., Passioura, J. B., Randall, L. B., Holbrook, N. M. & St Clair, D. A. 2004 Water relations under root chilling in a sensitive and tolerant tomato species. *Plant Cell and Environment* **27**, 971-979.

Bogdanove, A. J. 2002 *Pto* update: recent progress on an ancient plant defence response signalling pathway. *Molecular Plant Pathology* **3**, 283-288.

Boller, T. & Felix, G. 2009 A renaissance of elicitors: Perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annual Review of Plant Biology* **60**, 379-406.

Boller, T. & He, S. Y. 2009 Innate immunity in plants: An arms race between pattern recognition receptors in plants and effectors in microbial pathogens. *Science* **324**, 742-744.

Bomblies, K., Lempe, J., Epple, P., Warthmann, N., Lanz, C., Dangl, J. & Weigel, D. 2007 Autoimmune response as a mechanism for a Bateson-Dobzhansky-Muller-type incompatibility syndrome in plants. *PLoS Biology* **5**, e236.

Bomblies, K. & Weigel, D. 2007 Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nature Reviews Genetics* **8**, 382-393.

Bond, T. E. T. 1938 Infection experiments with *Cladosporium fulvum* cooke and related species. *Annals of Applied Biology* **25**, 277-307.

Böndel, K. 2010 Sequence evolution of resistance genes across species boundaries in wild tomatoes. In *Section of Evolutionary Biology*, vol. Diplom. Munich: Ludwig-Maximilians-University.

Brodsky, I. E. & Medzhitov, R. 2009 Targeting of immune signalling networks by bacterial pathogens. *Nature Cell Biology* **11**, 521-526.

Brown, J. K. M. 2003 A cost of disease resistance: paradigm or peculiarity? *Trends in Genetics* **19**, 667-671.

Brown, J. K. M. & Tellier, A. 2011 Plant-parasite coevolution: Bridging the gap between genetics and ecology. *Annual Review of Phytopathology* **49: in press**.

Burdon, J. J. & Thrall, P. H. 1999 Spatial and temporal patterns in coevolving plant and pathogen associations. *American Naturalist* **153**, S15-S33.

Caicedo, A. L. 2008 Geographic diversity cline of *R* gene homologs in wild populations of *Solanum pimpinellifolium* (Solanaceae). *American Journal of Botany* **95**, 393-398.

Caicedo, A. L. & Schaal, B. A. 2004 Heterogeneous evolutionary processes affect *R* gene diversity in natural populations of *Solanum pimpinellifolium. Proceedings of the National Academy of Sciences of the United States of America* **101**, 17444-17449.

Caldwell, K. S. & Michelmore, R. W. 2009 *Arabidopsis thaliana* genes encoding defense signaling and recognition proteins exhibit contrasting evolutionary dynamics. *Genetics* **181**, 671-684.

Castric, V., Bechsgaard, J., Schierup, M. H. & Vekemans, X. 2008 Repeated adaptive introgression at a gene under multiallelic balancing selection. *Plos Genetics* **4**.

Chang, J. H., Tai, Y. S., Bernal, A. J., Lavelle, D. T., Staskawicz, B. J. & Michelmore, R. W. 2002 Functional analyses of the *Pto* resistance gene family in tomato and the identification of a minor resistance determinant in a susceptible haplotype. *Molecular Plant-Microbe Interactions* **15**, 281-291.

Charlesworth, B., Morgan, M. T. & Charlesworth, D. 1993 The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289-1303.

Charlesworth, B., Nordborg, M. & Charlesworth, D. 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res. Camb.* **70**, 155-174.

Charlesworth, D. 2010 Don't forget the ancestral polymorphisms. *Heredity* **105**, 509-510.

Chen, F. Q. & Foolad, M. R. 1999 A molecular linkage map of tomato based on a cross between *Lycopersicon esculentum* and *L. pimpinellifolium* and its comparison with other molecular maps of tomato. *Genome* **42**, 94-103.

Chetelat, R. T., Pertuze, R. A., Faundez, L., Graham, E. B. & Jones, C. M. 2009 Distribution, ecology and reproductive biology of wild tomatoes and related nightshades from the Atacama Desert region of northern Chile. *Euphytica* **167**, 77-93.

Chisholm, S. T., Coaker, G., Day, B. & Staskawicz, B. J. 2006 Host-microbe interactions: Shaping the evolution of the plant immune response. *Cell* **124**, 803-814.

Coaker, G., Falick, A. & Staskawicz, B. 2005 Activation of a phytopathogenic bacterial effector protein by a eukaryotic cyclophilin. *Science* **308**, 548-550.

Collier, S. M. & Moffett, P. 2009 NB-LRRs work a ‘bait and switch’ on pathogens. *Trends in Plant Science* **14**, 521-529.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H. M., Koh, J. L. Y., Toufighi, K., Mostafavi, S., et al. 2010 The Genetic Landscape of a Cell. *Science* **327**, 425-431.

Dangl, J. L. & Jones, J. D. G. 2001 Plant pathogens and integrated defence responses to infection. *Nature* **411**, 826-833.

Dawkins, R. & Krebs, J. R. 1979 Arms Races between and within Species. *Proceedings of the Royal Society of London Series B-Biological Sciences* **205**, 489-511.

Day, B., Dahlbeck, D., Huang, J., Chisholm, S. T., Li, D. H. & Staskawicz, B. J. 2005 Molecular basis for the RIN4 negative regulation of RPS2 disease resistance. *Plant Cell* **17**, 1292-1305.

Deslandes, L., Olivier, J., Peeters, N., Feng, D. X., Khounlotham, M., Boucher, C., Somssich, I., Genin, S. & Marco, Y. 2003 Physical interaction between RRS1-R, a protein conferring resistance to bacterial wilt, and PopP2, a type III effector targeted to the plant nucleus. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 8024-8029.

Dixon, M. S., Jones, D. A., Keddie, J. S., Thomas, C. M., Harrison, K. & Jones, J. D. G. 1996 The tomato *Cf-2* disease resistance locus comprises two functional genes encoding leucine-rich repeat proteins. *Cell* **84**, 451-459.

Dodds, P. N., Lawrence, G. J., Catanzariti, A. M., Teh, T., Wang, C. I. A., Ayliffe, M. A., Kobe, B. & Ellis, J. G. 2006 Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 8888-8893.

Dodds, P. N. & Rathjen, J. P. 2010 Plant immunity: towards an integrated view of plant-pathogen interactions. *Nature Reviews Genetics* **11**, 539-548.

Doyle, J. J. & Doyle, J. L. 1987 A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11-15.

Excoffier, L., Estoup, A. & Cornuet, J. M. 2005 Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**, 1727-1738.

Fagundes, N. J. R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L. & Excoffier, L. 2007 Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 17614-17619.

Fay, J. C., Wyckoff, G. J. & Wu, C. I. 2001 Positive and negative selection on the human genome. *Genetics* **158**, 1227-1234.

Felix, G., Duran, J. D., Volko, S. & Boller, T. 1999 Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. *Plant Journal* **18**, 265-276.

Fischer, I., Camus-Kulandaivelu, L., Allal, F. & Stephan, W. 2011 Adaptation to drought in two wild tomato species: the evolution of the Asr gene family. *New Phytologist* **in press**.

Flor, H. H. 1956 The complementary genic systems in flax and flax rust. *Advances in Genetics Incorporating Molecular Genetic Medicine* **8**, 29-54.

Frank, S. A. 1992 Models of plant pathogen coevolution. *Trends in Genetics* **8**, 213-219.

Fu, Y. X. & Li, W. H. 1993 Statistical tests of neutrality of mutations. *Genetics* **133**, 693-709.

Gavrilets, S. & Michalakis, Y. 2008 Effects of environmental heterogeneity on victim-exploiter coevolution. *Evolution* **62**, 3100-3116.

Gimenez-Ibanez, S., Hann, D. R., Ntoukakls, V., Petutschnig, E., Lipka, V. & Rathjen, J. P. 2009 AvrPtoB targets the LysM receptor kinase CERK1 to promote bacterial virulence on plants. *Current Biology* **19**, 423-429.

Goehre, V., Spallek, T., Haeweker, H., Mersmann, S., Mentzel, T., Boller, T., de Torres, M., Mansfield, J. W. & Robatzek, S. 2008 Plant pattern-recognition receptor FLS2 is directed for degradation by the bacterial ubiquitin ligase AvrPtoB. *Current Biology* **18**, 1824-1832.

Greenbaum, D., Medzihradszky, K. F., Burlingame, A. & Bogyo, M. 2000 Epoxide electrophiles as activity-dependent cysteine protease profiling and discovery tools. *Chemistry & Biology* **7**, 569-581.

Grzeskowiak, L. 2009 The evolution of a disease resistance pathway in tomato. In *Section of Evolutionary Biology*, vol. PhD. Munich: Ludwig-Maximilians-University.

Gust, A. A., Brunner, F. & Nurnberger, T. 2010 Biotechnological concepts for improving plant innate immunity. *Current Opinion in Biotechnology* **21**, 204-210.

Gyllensten, U. B. & Erlich, H. A. 1989 Ancient roots for polymorphism at the HLA-DQ-ALPHA-locus in primates. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 9986-9990.

Haanstra, J. P. W., Wye, C., Verbakel, H., Meijer-Dekens, F., van den Berg, P., Odinot, P., van Heusden, A. W., Tanksley, S., Lindhout, P. & Peleman, J. 1999 An integrated high density RFLP-AFLP map of tomato based on two *Lycopersicon esculentum* x *L. pennellii* F-2 populations. *Theoretical and Applied Genetics* **99**, 254-271.

Hajishengallis, G. & Lambris, J. D. 2011 Microbial manipulation of receptor crosstalk in innate immunity. *Nature Reviews Immunology* **11**, 187-200.

Hammond-Kosack, K. E. & Jones, J. D. G. 1997 Plant disease resistance genes. *Annual Review of Plant Physiology and Plant Molecular Biology* **48**, 575-607.

Heath, M. C. 2000 Nonhost resistance and nonspecific plant defenses. *Current Opinion in Plant Biology* **3**, 315-319.

Holub, E. B. 2001 The arms race is ancient history in *Arabidopsis*, the wildflower. *Nature Reviews Genetics* **2**, 516-527.

Honnay, O., Bossuyt, B., Jacquemyn, H., Shimono, A. & Uchiyama, K. 2008 Can a seed bank maintain the genetic variation in the above ground plant population? *Oikos* **117**, 1-5.

Hörger, A. C. 2007 Sequenzevolution der Resistenzgene *Rcr3* und *Rin4* in Wildtomaten (*Lycopersicon peruvianum*). In *Section of Evolutionary Biology*, vol. Diplom. Munich: Ludwig-Maximilians-University.

Hudson, R. R., Kreitman, M. & Aguade, M. 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153-159.

Hughes, A. L. & Yeager, M. 1998 Natural selection at major histocompatibility complex loci of vertebrates. *Annual Review of Genetics* **32**, 415-435.

Igari, K., Endo, S., Hibara, K., Aida, M., Sakakibara, H., Kawasaki, T. & Tasaka, M. 2008 Constitutive activation of a CC-NB-LRR protein alters morphogenesis through the cytokinin pathway in *Arabidopsis*. *Plant Journal* **55**, 14-27.

Innan, H. 2003a The coalescent and infinite-site model of a small multigene family. *Genetics* **163**, 803-810.

Innan, H. 2003b A two-locus gene conversion model with selection and its application to the human *RHCE* and *RHD* genes. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 8793-8798.

Ispolatov, I. & Doebeli, M. 2009 Speciation due to hybrid necrosis in plant-pathogen models. *Evolution* **63**, 3076-3084.

Jeuken, M. J. W., Zhang, N. W., McHale, L. K., Pelgrom, K., den Boer, E., Lindhout, P., Michelmore, R. W., Visser, R. G. F. & Niks, R. E. 2009 Rin4 causes hybrid necrosis and race-specific resistance in an interspecific lettuce hybrid. *Plant Cell* **21**, 3368-3378.

Jia, Y., McAdams, S. A., Bryan, G. T., Hershey, H. P. & Valent, B. 2000 Direct interaction of resistance gene and avirulence gene products confers rice blast resistance. *Embo Journal* **19**, 4004-4014.

Jones, J. D. G. & Dangl, J. L. 2006 The plant immune system. *Nature* **444**, 323-329.

Kaschani, F., Shabab, M., Bozkurt, T., Shindo, T., Schornack, S., Gu, C., Ilyas, M., Win, J., Kamoun, S. & van der Hoorn, R. A. L. 2010 An effector-targeted protease contributes to defense against *Phytophthora infestans* and is under diversifying selection in natural hosts. *Plant Physiology* **154**, 1794-1804.

Kass, R. E. & Raftery, A. E. 1995 Bayes factors. *Journal of the American Statistical Association* **90**, 773-795.

Katagiri, F. & Tsuda, K. 2010 Understanding the plant immune system. *Molecular Plant-Microbe Interactions* **23**, 1531-1536.

Kim, H. S., Desveaux, D., Singer, A. U., Patel, P., Sondek, J. & Dangl, J. L. 2005a The *Pseudomonas syringae* effector AvrRpt2 cleaves its C-terminally acylated target, RIN4, from *Arabidopsis* membranes to block RPM1 activation. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 6496-6501.

Kim, M. G., da Cunha, L., McFall, A. J., Belkhadir, Y., DebRoy, S., Dangl, J. L. & Mackey, D. 2005b Two *Pseudomonas syringae* type III effectors inhibit RIN4-regulated basal defense in *Arabidopsis*. *Cell* **121**, 749-759.

Kingman, J. F. C. 1982 The coalescent. *Stochastic Processes and their Applications* **13**, 235-248.

Kirby, D. A. & Stephan, W. 1996 Multi-locus selection and the structure of variation at the white gene of *Drosophila melanogaster*. *Genetics* **144**, 635-645.

Krüger, J., Thomas, C. M., Golstein, C., Dixon, M. S., Smoker, M., Tang, S. K., Mulder, L. & Jones, J. D. G. 2002 A tomato cysteine protease required for Cf-2-dependent disease resistance and suppression of autonecrosis. *Science* **296**, 744-747.

Lacombe, S., Rougon-Cardoso, A., Sherwood, E., Peeters, N., Dahlbeck, D., van Esse, H. P., Smoker, M., Rallapalli, G., Thomma, B., Staskawicz, B., et al. 2010 Interfamily transfer of a plant pattern-recognition receptor confers broad-spectrum bacterial resistance. *Nature Biotechnology* **28**, 365-U94.

Laine, A. L. & Tellier, A. 2008 Heterogeneous selection promotes maintenance of polymorphism in host-parasite interactions. *Oikos* **117**, 1281-1288.

Lambrechts, L. 2010 Dissecting the genetic architecture of host-pathogen specificity. *PloS Pathogens* **6**, e1001019.

Lawlor, D. A., Ward, F. E., Ennis, P. D., Jackson, A. P. & Parham, P. 1988 HLA-A and HLA-B polymorphisms predate the divergence of humans and chimpanzees. *Nature* **335**, 268-271.

Legnani, R., Gognalons, P., Selassie, K. G., Marchoux, G., Moretti, A. & Laterrot, H. 1996 Identification and characterization of resistance to tobacco etch virus in *Lycopersicon* species. *Plant Disease* **80**, 306-309.

Lemaitre, B. & Hoffmann, J. 2007 The host defense of *Drosophila melanogaster*. *Annual Review of Immunology* **25**, 697-743.

Leonard, K. J. 1977 Selection pressures and plant pathogens. *Annals of the New York Academy of Sciences* **287**, 207-222.

Li, X. Y., Lin, H. Q., Zhang, W. G., Zou, Y., Zhang, J., Tang, X. Y. & Zhou, J. M. 2005 Flagellin induces innate immunity in nonhost interactions that is suppressed by *Pseudomonas syringae* effectors. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 12990-12995.

Librado, P. & Rozas, J. 2009 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452.

Liu, J., Elmore, J. M. & Coaker, G. 2009 Investigating the functions of the RIN4 protein complex during plant innate immune responses. *Plant Signaling and Behavior* **4**, 1107-10.

Liu, Y. G., Mitsukawa, N., Oosumi, T. & Whittier, R. F. 1995 Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insertjunctions by thermal assymmetric interlaced PCR. *Plant Journal* **8**, 457-463.

Loh, Y. T. & Martin, G. B. 1995 The *Pto* bacterial resistance gene and the *Fen* insecticide sensitivity gene encode functional protein-kinases with serine/threonine specificity. *Plant Physiology* **108**, 1735-1739.

Lu, Y. Q. & Rausher, M. D. 2003 Evolutionary rate variation in anthocyanin pathway genes. *Molecular Biology and Evolution* **20**, 1844-1853.

Luderer, R., Takken, F. L. W., de Wit, P. & Joosten, M. 2002 *Cladosporium fulvum* overcomes Cf-2-mediated resistance by producing truncated AVR2 elicitor proteins. *Molecular Microbiology* **45**, 875-884.

Luo, Y., Caldwell, K. S., Wroblewski, T., Wright, M. E. & Michelmore, R. W. 2009 Proteolysis of a negative regulator of innate immunity is dependent on resistance genes in tomato and *Nicotiana benthamiana* and induced by multiple bacterial effectors. *Plant Cell* **21**, 2458-2472.

Mackey, D., Belkhadir, Y., Alonso, J. M., Ecker, J. R. & Dangl, J. L. 2003 *Arabidopsis* RIN4 is a target of the type III virulence effector AvrRpt2 and modulates RPS2-mediated resistance. *Cell* **112**, 379-389.

Mackey, D., Holt, B. F., Wiig, A. & Dangl, J. L. 2002 RIN4 interacts with *Pseudomonas syringae* type III effector molecules and is required for RPM1-mediated resistance in *Arabidopsis*. *Cell* **108**, 743-754.

Madigan, M. T. & Martinko, J. M. 2005 *Brock Microbiology*. Prentice Hall, New Jersey, USA.

Martin, G. B., Brommonschenkel, S. H., Chunwongse, J., Frary, A., Ganal, M. W., Spivey, R., Wu, T. Y., Earle, E. D. & Tanksley, S. D. 1993 Map-based cloning of a protein-kinase gene conferring disease resistance in tomato. *Science* **262**, 1432-1436.

Martin, G. B., Frary, A., Wu, T. Y., Brommonschenkel, S., Chunwongse, J., Earle, E. D. & Tanksley, S. D. 1994 A member of the tomato *Pto* gene family confers sensitivity to fenthion resulting in rapid cell-death. *Plant Cell* **6**, 1543-1552.

Mauricio, R., Stahl, E. A., Korves, T., Tian, D. C., Kreitman, M. & Bergelson, J. 2003 Natural selection for polymorphism in the disease resistance gene *Rps2* of *Arabidopsis thaliana*. *Genetics* **163**, 735-746.

Mayer, W. E., Jonker, M., Klein, D., Ivanyi, P., Vanseventer, G. & Klein, J. 1988 Nucleotide-sequences of chimpanzee MHC class-I alleles – evidence for trans-species mode of evolution. *Embo Journal* **7**, 2765-2774.

McDonald, J. H. & Kreitman, M. 1991 Adaptive protein evolution at the *Adh* locus in drosophila. *Nature* **351**, 652-654.

Medzhitov, R. & Janeway, C. A. 1997 Innate immunity: The virtues of a nonclonal system of recognition. *Cell* **91**, 295-298.

Melotto, M., Underwood, W., Koczan, J., Nomura, K. & He, S. Y. 2006 Plant stomata function in innate immunity against bacterial invasion. *Cell* **126**, 969-980.

Michelmore, R. W. & Meyers, B. C. 1998 Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Research* **8**, 1113-1130.

Monforte, A. J. & Tanksley, S. D. 2000 Development of a set of near isogenic and backcross recombinant inbred lines containing most of the *Lycopersicon hirsutum* genome in a *L. esculentum* genetic background: A tool for gene mapping and gene discovery. *Genome* **43**, 803-813.

Mucyn, T. S., Clemente, A., Andriotis, V. M. E., Balmuth, A. L., Oldroyd, G. E. D., Staskawicz, B. J. & Rathjen, J. P. 2006 The tomato NBARC-LRR protein Prf interacts with Pto kinase in vivo to regulate specific plant immunity. *Plant Cell* **18**, 2792-2806.

Murphy, K., Travers, P. & Walport, M. 2008 *Janeway's Immunobiology*. Garland Science: New York and London.

Nakazato, T., Bogonovich, M. & Moyle, L. C. 2008 Environmental factors predict adaptive phenotypic differentiation within and between two wild andean tomatoes. *Evolution* **62**, 774-792.

Nakazato, T., Warren, D. L. & Moyle, L. C. 2010 Ecological and geographic modes of species divergence in wild tomatoes. *American Journal of Botany* **97**, 680-693.

Nishimura, M. T. & Dangl, J. L. 2010 *Arabidopsis* and the plant immune system. *Plant Journal* **61**, 1053-1066.

Ntoukakis, V., Mucyn, T. S., Gimenez-lbanez, S., Chapman, H. C., Gutierrez, J. R., Balmuth, A. L., Jones, A. M. E. & Rathjen, J. P. 2009 Host inhibition of a bacterial virulence effector triggers immunity to infection. *Science* **324**, 784-787.

Nürnberger, T. & Brunner, F. 2002 Innate immunity in plants and animals: emerging parallels between the recognition of general elicitors and pathogen-associated molecular patterns. *Current Opinion in Plant Biology* **5**, 318-324.

Nürnberger, T., Brunner, F., Kemmerling, B. & Piater, L. 2004 Innate immunity in plants and animals: striking similarities and obvious differences. *Immunological Reviews* **198**, 249-266.

Nürnberger, T. & Lipka, V. 2005 Non-host resistance in plants: new insights into an old phenomenon. *Molecular Plant Pathology* **6**, 335-345.

Obbard, D. J., Welch, J. J., Kim, K. W. & Jiggins, F. M. 2009 Quantifying adaptive evolution in the *Drosophila* immune system. *Plos Genetics* **5**.

Oh, C.-S. & Martin, G. B. 2011 Effector-triggered immunity mediated by the Pto kinase. *Trends in Plant Science* **16**, 132-140.

Orr, H. A. 1996 Dobzhansky, Bateson, and the genetics of speciation. *Genetics* **144**, 1331-1335.

Orr, H. A. & Turelli, M. 2001 The evolution of postzygotic isolation: Accumulating Dobzhansky-Muller incompatibilities. *Evolution* **55**, 1085-1094.

Pannell, J. R. 2003 Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution* **57**, 949-961.

Pannell, J. R. & Charlesworth, B. 2000 Effects of metapopulation processes on measures of genetic diversity. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **355**, 1851-1864.

Parniske, M., HammondKosack, K. E., Golstein, C., Thomas, C. M., Jones, D. A., Harrison, K., Wulff, B. B. H. & Jones, J. D. G. 1997 Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. *Cell* **91**, 821-832.

Parsch, J., Novozhilov, S., Saminadin-Peter, S. S., Wong, K. M. & Andolfatto, P. 2010 On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Molecular Biology and Evolution* **27**, 1226-1234.

Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A. J., Thomson, N. R., Quail, M., Smith, F., Walker, D., Libberton, B., et al. 2010 Antagonistic coevolution accelerates molecular evolution. *Nature* **464**, 275-278.

Pavlidis, P., Hutter, S. & Stephan, W. 2008 A population genomic approach to map recent positive selection in model species. *Molecular Ecology* **17**, 3585-3598.

Peralta, I. E., Knapp, S. K. & Spooner, D. M. 2005 New species of wild tomatoes (*Solanum* section *Lycopersicon*: Solanaceae) from Northern Peru. *Systematic Botany* **30**, 424-434.

Peralta, I. E., Spooner, D. M. & Knapp, S. 2008 The taxonomy of tomatoes: a revision of wild tomatoes (*Solanum* section *Lycopersicon*) and their outgroup relatives in sections *Juglandifolium* and *Lycopersicoides*. *Systematic Botany Monographs* **84**, 1-186.

Phillips, P. C. 1996 Waiting for a compensatory mutation: Phase zero of the shifting-balance process. *Genetical Research* **67**, 271-283.

Pilowsky, M. & Zutra, D. 1982 Screening wild tomatoes for resistance to bacterial speck pathogen (*Pseudomonas*-tomato). *Plant Disease* **66**, 46-47.

Pitzschke, A., Schikora, A. & Hirt, H. 2009 MAPK cascade signalling networks in plant defence. *Current Opinion in Plant Biology* **12**, 421-426.

R Development Core Team. 2005 R: A language and environment for statistical computing (ed. R Foundation for Statistical Computing). Vienna, Austria.

Raffaele, S., Farrer, R. A., Cano, L. M., Studholme, D. J., MacLean, D., Thines, M., Jiang, R. H. Y., Zody, M. C., Kunjeti, S. G., Donofrio, N. M., et al. 2010 Genome evolution following host jumps in the Irish Potato Famine pathogen lineage. *Science* **330**, 1540-1543.

Ramsay, H., Rieseberg, L. H. & Ritland, K. 2009 The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. *Molecular Biology and Evolution* **26**, 1045-1053.

Rathjen, J. P., Chang, J. H., Staskawicz, B. J. & Michelmore, R. W. 1999 Constitutively active Pto induces a Prf-dependent hypersensitive response in the absence of avrPto. *Embo Journal* **18**, 3232-3240.

Rausher, M. D., Lu, Y. Q. & Meyer, K. 2008 Variation in constraint versus positive selection as an explanation for evolutionary rate variation among anthocyanin genes. *Journal of Molecular Evolution* **67**, 137-144.

Rausher, M. D., Miller, R. E. & Tiffin, P. 1999 Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Molecular Biology and Evolution* **16**, 266-274.

Rick, C. M. 1973 Potential genetic resources in tomato species: clues from observations in native habitats. *Basic Life Science* **2**, 255-69.

Rick, C. M. 1986 Reproductive isolation in the *Lycopersicon peruvianum* complex. *D'arcy, W. G. (Ed.). Solanaceae: Biology and Systematics; Second International Symposium, St. Louis, Mo., USA, Aug. 3-6, 1983. Xiii+603p. Columbia University Press: New York, N.Y., USA. Illus*, 477-495.

Rick, C. M., Kesicki, E., Fobes, J. F. & Holle, M. 1976 Genetic and biosystematic studies on two new sibling species of *Lycopersicon* from inter-Andean Peru. *Theoretical and Applied Genetics* **47**, 55-68.

Rick, C. M. & Lamm, R. 1955 Biosystematic studies on the status of *Lycopersicon chilense*. *American Journal of Botany* **42**, 663-675.

Riely, B. K. & Martin, G. B. 2001 Ancient origin of pathogen recognition specificity conferred by the tomato disease resistance gene *Pto*. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 2059-2064.

Rodriguez, F., Wu, F. N., Ane, C., Tanksley, S. & Spooner, D. M. 2009 Do potatoes and tomatoes have a single evolutionary history, and what proportion of the genome supports this history? *Bmc Evolutionary Biology* **9:**191.

Rooney, H. C. E., van 't Klooster, J. W., van der Hoorn, R. A. L., Joosten, M., Jones, J. D. G. & de Wit, P. 2005 *Cladosporium* Avr2 inhibits tomato Rcr3 protease required for Cf-2-dependent disease resistance. *Science* **308**, 1783-1786.

Rose, L. E., Bittner-Eddy, P. D., Langley, C. H., Holub, E. B., Michelmore, R. W. & Beynon, J. L. 2004 The maintenance of extreme amino acid diversity at the disease resistance gene, *RPP13*, in *Arabidopsis thaliana*. *Genetics* **166**, 1517-1527.

Rose, L. E., Langley, C. H., Bernal, A. J. & Michelmore, R. W. 2005 Natural variation in the *Pto* pathogen resistance gene within species of wild tomato (*Lycopersicon*). I. Functional analysis of *Pto* alleles. *Genetics* **171**, 345-357.

Rose, L. E., Michelmore, R. W. & Langley, C. H. 2007 Natural variation in the *Pto* pathogen resistance gene within species of wild tomato (*Lycopersicon*). II. Population genetics of *Pto*. *Genetics* **175**, 1307-1319.

Rosebrock, T. R., Zeng, L. R., Brady, J. J., Abramovitch, R. B., Xiao, F. M. & Martin, G. B. 2007 A bacterial E3 ubiquitin ligase targets a host protein kinase to disrupt plant immunity. *Nature* **448**, 370-374.

Roselius, K., Stephan, W. & Städler, T. 2005 The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics* **171**, 753-763.

Sackton, T. B., Lazzaro, B. P., Schlenke, T. A., Evans, J. D., Hultmark, D. & Clark, A. G. 2007 Dynamic evolution of the innate immune system in *Drosophila*. *Nature Genetics* **39**, 1461-1468.

Sato, M., Tsuda, K., Wang, L., Coller, J., Watanabe, Y., Glazebrook, J. & Katagiri, F. 2010 Network modeling reveals prevalent negative regulatory relationships between signaling sectors in *Arabidopsis* immune signaling. *Plos Pathogens* **6**.

Sawyer, S. 1989 Statistical tests for detecting gene conversion. *Molecular Biology and Evolution* **6**, 526-538.

Schatz, D. G. & Ji, Y. 2011 Recombination centres and the orchestration of V(D)J recombination. *Nature Reviews Immunology* **11**, 251-263.

Scheel, D. 1998 Resistance response physiology and signal transduction. *Current Opinion in Plant Biology* **1**, 305-310.

Schmid-Hempel, P. 2008 Parasite immune evasion: a momentous molecular war. *Trends in Ecology & Evolution* **23**, 318-326.

Schulze-Lefert, P. & Panstruga, R. 2011 A molecular evolutionary concept connecting nonhost resistance, pathogen host range, and pathogen speciation. *Trends in Plant Science* **16**, 117-125.

Sessa, G. & Martin, G. B. 2000 Signal recognition and transduction mediated by the tomato Pto kinase: a paradigm of innate immunity in plants. *Microbes and Infection* **2**, 1591-1597.

Shabab, M., Shindo, T., Gu, C., Kaschani, F., Pansuriya, T., Chintha, R., Harzen, A., Colby, T., Kamoun, S. & van der Hoorn, R. A. L. 2008 Fungal effector protein AVR2 targets diversifying defense-related Cys proteases of tomato. *Plant Cell* **20**, 1169-1183.

Shan, L., Ping, H., Jianming, L., Antje, H., Scott, P., Thorsten, N., Gregory, M. & Jen, S. 2008 Bacterial effectors target BAK1 and disrupt MAMP receptor signaling complexes to impede plant innate immunity. *Plant Biology* **2008**, 166.

Shao, F., Golstein, C., Ade, J., Stoutemyer, M., Dixon, J. E. & Innes, R. W. 2003 Cleavage of *Arabidopsis* PBS1 by a bacterial type III effector. *Science* **301**, 1230-1233.

Song, J., Win, J., Tian, M. Y., Schornack, S., Kaschani, F., Ilyas, M., van der Hoorn, R. A. L. & Kamoun, S. 2009 Apoplastic effectors secreted by two unrelated eukaryotic plant pathogens target the tomato defense protease Rcr3. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 1654-1659.

Soubeyrand, S., Laine, A. L., Hanski, I. & Penttinen, A. 2009 Spatiotemporal structure of host-pathogen interactions in a metapopulation. *American Naturalist* **174**, 308-320.

Spooner, D. M., Peralta, I. E. & Knapp, S. 2005 Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [*Solanum* section *Lycopersicon* (Mill.) Wettst.]. *Taxon* **54**, 43-61.

Städler, T., Arunyawat, U. & Stephan, W. 2008 Population genetics of speciation in two closely related wild tomatoes (*Solanum* section *lycopersicon*). *Genetics* **178**, 339-350.

Städler, T., Haubold, B., Merino, C., Stephan, W. & Pfaffelhuber, P. 2009 The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**, 205-216.

Städler, T., Roselius, K. & Stephan, W. 2005 Genealogical footprints of speciation processes in wild tomatoes: Demography and evidence for historical gene flow. *Evolution* **59**, 1268-1279.

Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M. & Bergelson, J. 1999 Dynamics of disease resistance polymorphism at the *Rpm1* locus of *Arabidopsis*. *Nature* **400**, 667-671.

Stephan, W. & Langley, C. H. 1998 DNA polymorphism in *Lycopersicon* and crossing-over per physical length. *Genetics* **150**, 1585-1593.

Stukenbrock, E. H. & McDonald, B. A. 2009 Population genetics of fungal and oomycete effectors involved in gene-for-gene interactions. *Molecular Plant-Microbe Interactions* **22**, 371-380.

Tai, Y.-S. 2004 The role of Prf and its partners in resistance to *Pseudomonas syringae* pv. *tomato*, vol. PhD. Davis: University of California.

Tajima, F. 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595.

Tanksley, S. D., Ganal, M. W., Prince, J. P., Devicente, M. C., Bonierbale, M. W., Broun, P., Fulton, T. M., Giovannoni, J. J., Grandillo, S., Martin, G. B., et al. 1992 High-density molecular linkage maps of the tomato and potato genomes. *Genetics* **132**, 1141-1160.

Tellier, A. & Brown, J. K. M. 2007a Polymorphism in multilocus host-parasite coevolutionary interactions. *Genetics* **177**, 1777-1790.

Tellier, A. & Brown, J. K. M. 2007b Stability of genetic polymorphism in host-parasite interactions. *Proceedings of the Royal Society B-Biological Sciences* **274**, 809-817.

Tellier, A. & Brown, J. K. M. 2009 The influence of perenniality and seed banks on polymorphism in plant-parasite interactions. *The American Naturalist* **174**, 769-779.

Tellier, A., Fischer, I., Merino, C., Xia, H., Camus-Kulandaivelu, L., Städler, T. & Stephan, W. 2011 Fitness effects of derived deleterious mutations in four closely related wild tomato species with spatial structure. *Heredity* ***in press***.

Thompson, J. N. 2005 *The Geographic Mosaic of Coevolution*. Chicago, USA: University of Chicago Press.

Thompson, J. N. & Burdon, J. J. 1992 Gene-for-gene coevolution between plants and parasites. *Nature* **360**, 121-125.

Thordal-Christensen, H. 2003 Fresh insights into processes of nonhost resistance. *Current Opinion in Plant Biology* **6**, 351-357.

Thornton, K. 2003 libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**, 2325-2327.

Thornton, K. R. 2007 The neutral coalescent process for recent gene duplications and copy-number variants. *Genetics* **177**, 987-1000.

Thrall, P. H. & Burdon, J. J. 2002 Evolution of gene-for-gene systems in metapopulations: the effect of spatial scale of host and pathogen dispersal. *Plant Pathology* **51**, 169-184.

Tian, D., Traw, M. B., Chen, J. Q., Kreitman, M. & Bergelson, J. 2003 Fitness costs of *R*-gene-mediated resistance in *Arabidopsis thaliana*. *Nature* **423**, 74-77.

Tian, M. Y., Win, J., Song, J., van der Hoorn, R., van der Knaap, E. & Kamoun, S. 2007 A *Phytophthora infestans* cystatin-like protein targets a novel tomato papain-like apoplastic protease. *Plant Physiology* **143**, 364-377.

van der Biezen, E. A. & Jones, J. D. G. 1998 Plant disease-resistance proteins and the gene-for-gene concept. *Trends in Biochemical Sciences* **23**, 454-456.

van der Hoorn, R. A. L. & Kamoun, S. 2008 From Guard to Decoy: A new model for perception of plant pathogen effectors. *Plant Cell* **20**, 2009-2017.

van Esse, H. P., van't Klooster, J. W., Bolton, M. D., Yadeta, K. A., van Baarlen, P., Boeren, S., Vervoort, J., de Wit, P. & Thomma, B. 2008 The *Cladosporium fulvum* virulence protein Avr2 inhibits host proteases required for basal defense. *Plant Cell* **20**, 1948-1963.

Voinnet, O., Rivas, S., Mestre, P. & Baulcombe, D. 2003 An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus. *Plant Journal* **33**, 949-956.

Wagner, A. 2000 Robustness against mutations in genetic networks of yeast. *Nature Genetics* **24**, 355-361.

Wagner, A. 2001 The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution* **18**, 1283-1292.

Wakeley, J. & Aliacar, N. 2001 Gene genealogies in a metapopulation. *Genetics* **159**, 893-905.

Wakeley, J. & Hey, J. 1997 Estimating ancestral population parameters. *Genetics* **145**, 847-855.

Waxman, D. & Peck, J. R. 1998 Pleiotropy and the preservation of perfection. *Science* **279**, 1210-1213.

Whitlock, M. C. 2003 Fixation probability and time in subdivided populations. *Genetics* **164**, 767-779.

Woolhouse, M. E. J., Webster, J. P., Domingo, E., Charlesworth, B. & Levin, B. R. 2002 Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genetics* **32**, 569-577.

Wulff, B. B. H., Chakrabarti, A. & Jones, D. A. 2009 Recognitional specificity and evolution in the tomato-*Cladosporium fulvum* pathosystem. *Molecular Plant-Microbe Interactions* **22**, 1191-1202.

Wulff, B. B. H., Kruijt, M., Collins, P. L., Thomas, C. M., Ludwig, A. A., De Wit, P. & Jones, J. D. G. 2004 Gene shuffling generated and natural variants of the tomato resistance gene Cf-9 exhibit different auto-necrosis-inducing activities in *Nicotiana* species. *Plant Journal* **40**, 942-956.

Xia, H., Camus-Kulandaivelu, L., Stephan, W., Tellier, A. & Zhang, Z. 2010 Nucleotide diversity patterns of local adaptation at drought-related candidate genes in wild tomatoes. *Molecular Ecology* **19**, 4144-4154.

Xing, W., Zou, Y., Liu, Q., Liu, J. N., Luo, X., Huang, Q. Q., Chen, S., Zhu, L. H., Bi, R. C., Hao, Q., et al. 2007 The structural basis for activation of plant immunity by bacterial effector protein AvrPto. *Nature* **449**, 243-248.

Young, K. R., Ulloa, C. U., Luteyn, J. L. & Knapp, S. 2002 Plant evolution and endemism in Andean South America: An introduction. *Botanical Review* **68**, 4-21.

Zhang, J., Shao, F., Cui, H., Chen, L. J., Li, H. T., Zou, Y., Long, C. Z., Lan, L. F., Chai, J. J., Chen, S., et al. 2007 A *Pseudomonas syringae* effector inactivates MAPKs to suppress PAMP-induced immunity in plants. *Cell Host & Microbe* **1**, 175-185.

Zhou, J. M. & Chai, J. 2008 Plant pathogenic bacterial type III effectors subdue host responses. *Current Opinion in Microbiology* **11**, 179-185.

Zipfel, C. & Felix, G. 2005 Plants and animals: A different taste for microbes? *Current Opinion in Plant Biology* **8**, 353-360.

Zuriaga, E., Blanca, J. & Nuez, F. 2009 Classification and phylogenetic relationships in *Solanum* section *Lycopersicon* based on AFLP and two nuclear gene sequences. *Genetic Resources and Crop Evolution* **56**, 663-678.

APPENDIX A: MATERIALS AND METHODS

Table A1: List of primers and annealing temperatures used in the *Rcr3* project.

Primer	Sequence 5'→ 3'	Tm [°C]	use
Primers for amplification of flanking regions			
Rcr3 3'FLR For2	GGA GGT TTT ATG ACG AAT GC	56.0	Rcr3 specific in 1st Tail-PCR step
Rcr3 3'FLR For3	CAG TAC ACA TGC AGA AGC C	57.0	Rcr3 specific in 2nd Tail-PCR step
Rcr3 3'FLR For4	GTC CAT TGG AAT AGC TGC TAG	59.0	Rcr3 specific in 3rd Tail-PCR step
Rcr3 5'FLR Rev1	CTC TCC AGT CCA AGT TAG ACG	61.0	Rcr3 specific in 1st Tail-PCR step
Rcr3 5'FLR Rev2	CTC TTG TGA AGT AAT ATC TGC	55.0	Rcr3 specific in 2nd Tail-PCR step
Rcr3 5'FLR Rev3	CTC CTT TTT CTA CTT CGT CC	56.0	Rcr3 specific in 3rd Tail-PCR step
AD1	NGT CGA SWG ANA WGA A	46.0	random priming in the flanking region
Primers for cloning procedure			
Rcr3 start	AGC TCC ATG GCT ATG AAA GTT GAT TTG ATG	68.0	amplification of Rcr3
Rcr3 stop	AGC TCT CGA GCT ATG CTA TGT TTG GAT AAG AAG AC	73.0	amplification of Rcr3
F401	CGT TGT AAA ACG ACG GCC AGT	61.0	forward in pFK0026
F402	CAG GAA ACA GCT ATG ACC ATG	59.0	reverse in pFK0026
F403	AGG AAG TTC ATT TCA TTT GGA GAG G	63.0	forward in 35S promotor on pFK0026
F404	CAC ATT ATA GTG ATT AGC ATG TCA C	61.0	reverse in terminator on pFK0026
r114	TAG GTT TAC CCG CCA ATA TAT CCT GTC	67.0	forward in pTP05
r115	TTC TGT CAG TTC CAA ACG TAA AAC GGC	67.0	reverse in pTP05
Primers for RT-PCR			
Rcr3RTFor1	GCC AAA ACT CTC CGT GTC TG	60.0	forward in Rcr3 for RT-PCR
Rcr3RTRev1	AGA ATC TCT TAT AAT TTT CAT AAA CC	57.0	reverse in Rcr3 for RT-PCR
Rcr3RTRev2	CAG TGA ATA ATA TTT CAT GAG ACA G	59.0	reverse in Rcr3 for RT-PCR
RubiscoRTFor1	GTT CTC GAG GAG CTT ATC AAT GG	63.0	forward in tobacco Rubisco for RT-PCR
RubiscoRTRev1	CAG GGT CCC CAT TAT CGT C	59.0	reverse in tobacco Rubisco for RT-PCR

Table A2: List of primers and annealing temperatures used to amplify the *Pfi* and *Rin4* genes.

name	sequence (5'-3')	position [bp]	Tm[°C]
primers *Pfi* gene			
Prfint30137For1	GAAGAATGAGTGCTGCTTCCT	22	59
Prfint30137For77	AGCTTCAGCACCAGTGTCC	101	59
Prfint30137For253	GGCTTAGCCGTGGCTGAG	754	61
Prfint30137In2For1	CTAGCTCCGTGCTTGTTGG	1382	59
Prfint30137In2Rev1	CATAGTTATTGCCACTCTGATC	1459	58
Prfint30137In3For1	CTGATGTAGCTTGACGAAATGC	2028	60
Prfint30137Rev432	GCCAGCGACAAACTGAAGC	2498	59
Prfint30137For582	CCTGGTGGAGAAAGCTGTG	3059	59
Prfint30137Rev685	GCTACATCATCTTCGTTCACC	3122	59
Prfint30137For1200	CACAACTTGTCTGGCTTTAGTGC	3678	63
Prfint30137Rev1231	GATGGATTTGCACTAAAGCC	3666	56
Prfint30137For1698	CGAGATAGGCAGTTAATTCAGG	4175	60
Prfint30137For1799	GCACATGCTGTTTCTGCG	4353	56
Prfint30137Rev1825	GTAACACTTCGCAGAAACAGC	4340	59
Prfint30137Rev2374	AGCTTTCATTAATTCACATGGC	5396	57
	primers *Rin4* gene		
Rin4For3	GGCACTGTCTAAACTATGTTTGC	3	58
Rin4For5	CTTTCAGAACCAGTATGATTAGG	810	58
Rin4Rev3	ACGAGTCTGCTTCTCCTCTCG	1090	58
Rin4Rev5	CCAACAACTAATGGATGGCA	1800	58

APPENDIX B: RESULTS

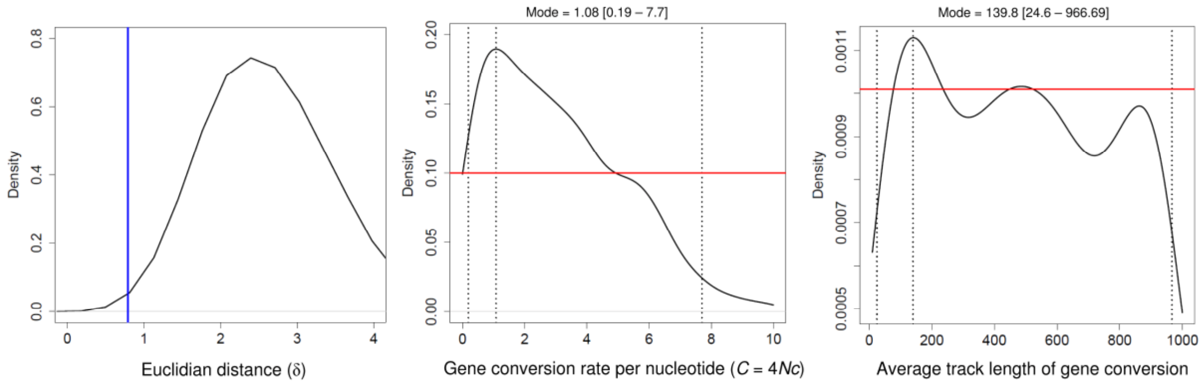


Figure B1: ABC estimates of parameters for Model 2 with gene conversion (for *Rcr3* ORFs). Left panel: Density of the distribution of Euclidian distances (δ) for all 100,000 simulated datasets. The blue line indicates the best 500 retained datasets after the rejection. Middle panel: Density of the posterior distribution for the gene conversion rate ($C = 4Nc$ per nucleotide), in red is the density of the uniform prior. Dotted lines indicate the 95% credibility intervals and the mode of the distribution. Right panel: Density of the posterior distribution for the mean length of the gene conversion track in bp, in red is the density of the uniform prior. Dotted lines indicate the 95% credibility intervals and the mode of the distribution.

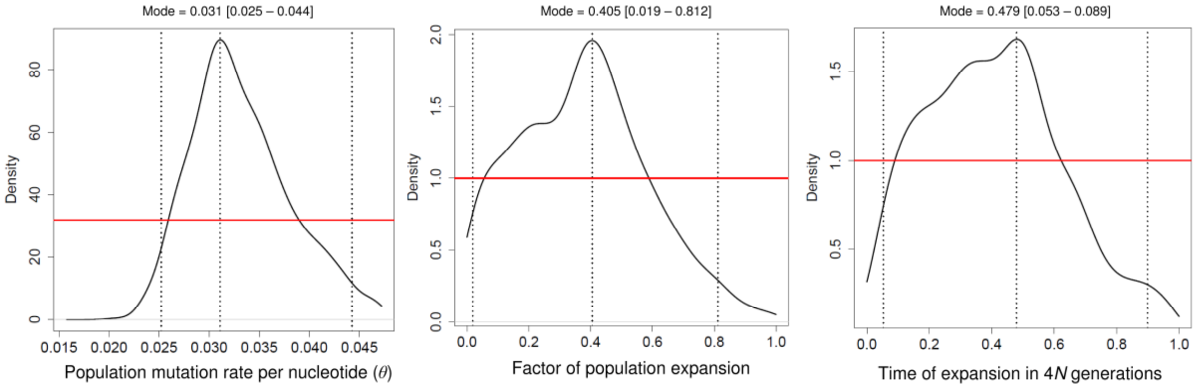


Figure B2: Posterior distributions of the parameters of the demographic model of the Tarapaca population with past expansion based on 14 reference loci (for *Rcr3* 3'FLRs). In red is the density of the uniform prior. Dotted lines indicate the 95% credibility intervals and the mode of the distribution. Left panel: Density of the posterior distribution for population mutation rate (θ per nucleotide). Middle panel: Density of the posterior distribution for the expansion factor. Right panel: Density of the posterior distribution for the time of the expansion (in 4*N* generations).

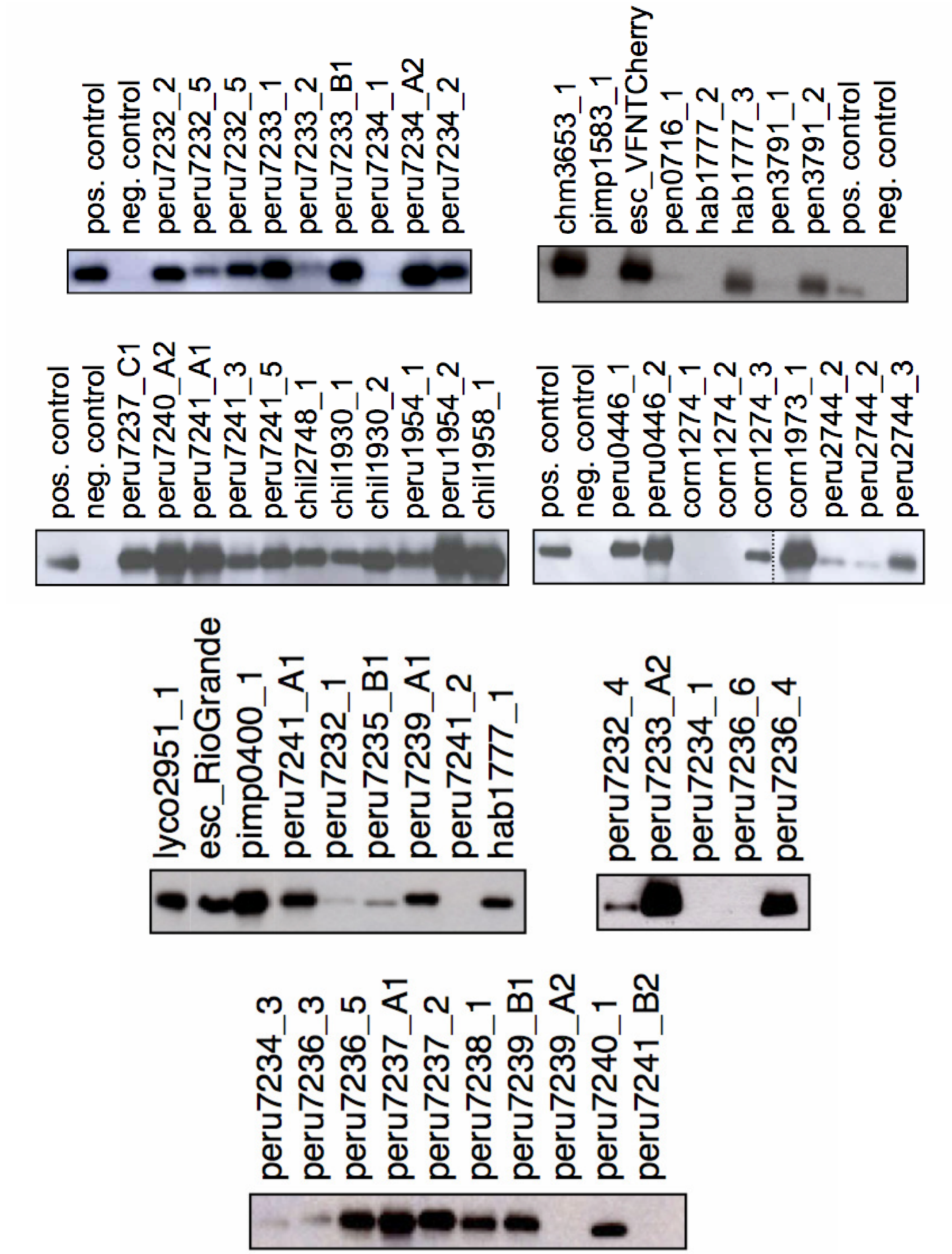


Figure B3: Western Blot of all isolated AFs containing the expressed Rcr3 constructs. Overexpressed Rcr3 constructs were confirmed using Rcr3 specific antibodies. Proteins were separated on 12% protein gels. AFs without overexpressed RCR3 were used as a negative control. AFs containing Rcr3 from *S. lycopersicum* (*cv*. Rio Grande) were used as a positive control.

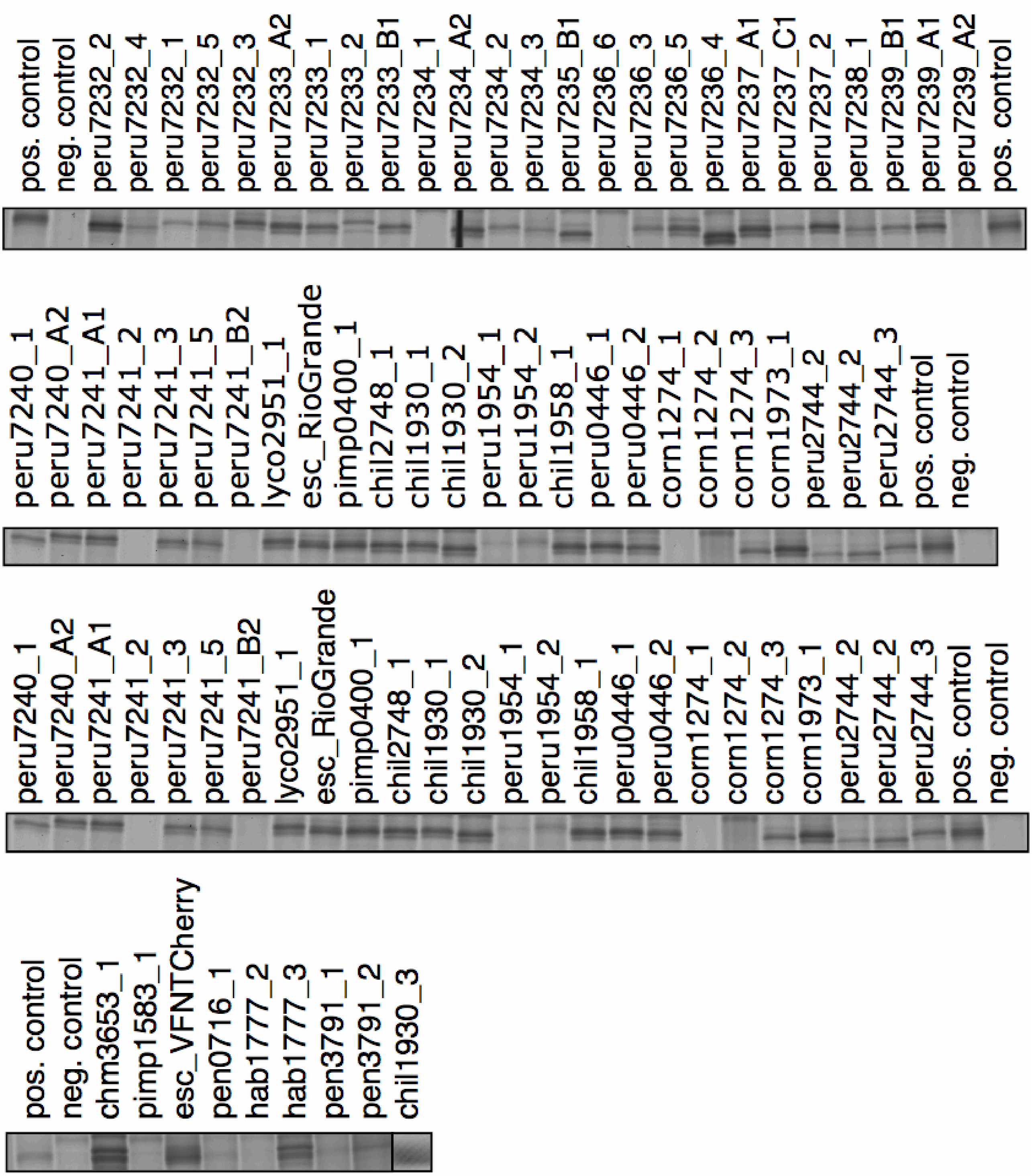
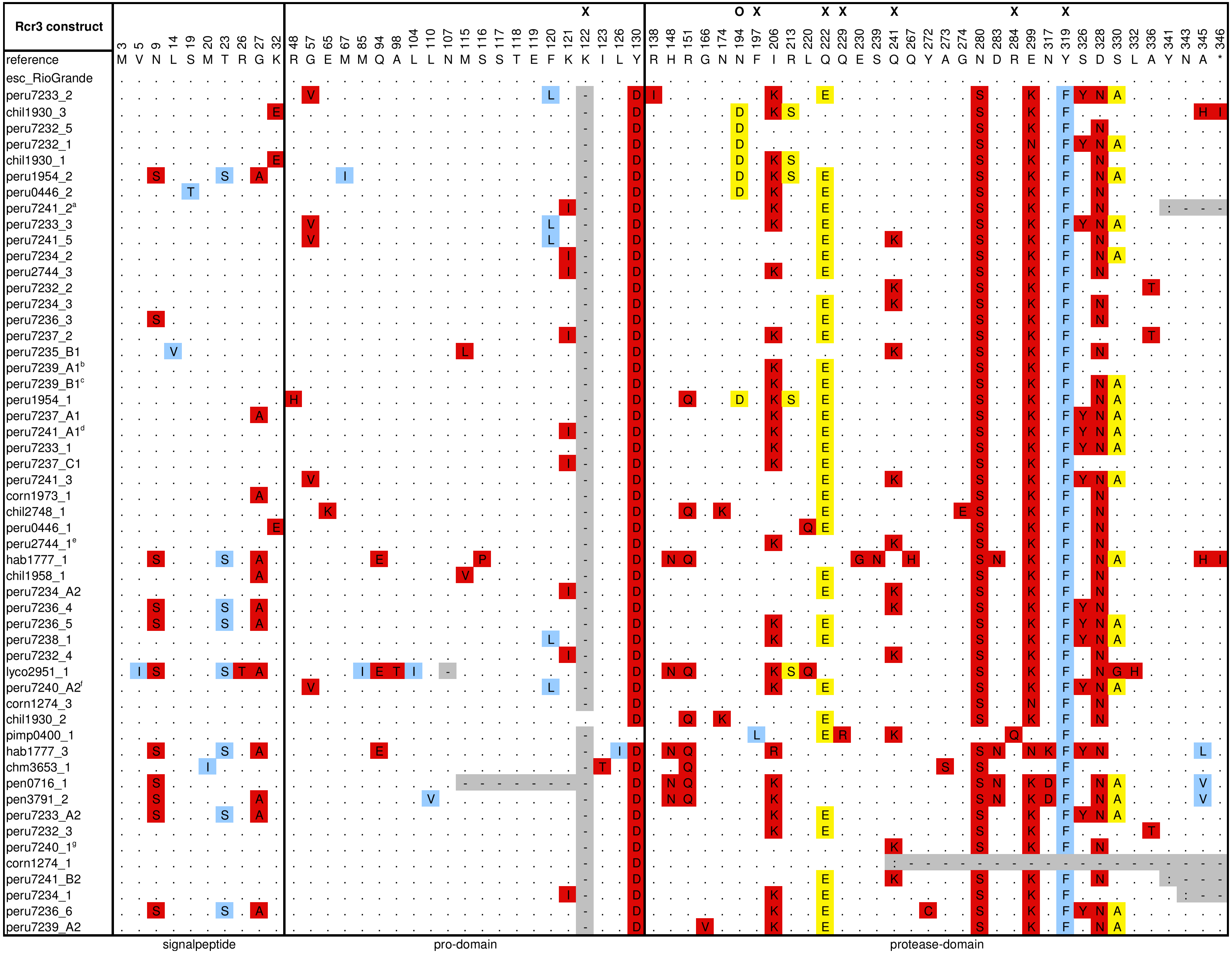


Figure B4: Protease activity profiling of all Rcr3 constructs. AFs that contained overexpressed Rcr3 constructs were labeled with DCG-04 at pH 5.5. Proteins were separated on 12% protein gels. AFs without overexpressed Rcr3 were used as a negative control. AFs containing Rcr3 from *S. lycopersicum* (*cv*. Rio Grande) were used as a positive control.



																											X									O	X				X	X			X							X			X									
Rcr3 construct	3	5	9	14	19	20	23	26	27	32	48	57	65	67	85	94	98	104	110	107	115	116	117	118	119	120	121	122	123	126	130	138	148	151	166	174	194	197	206	213	220	222	229	230	239	241	267	272	273	274	280	283	284	299	317	319	326	328	330	332	336	341	343	345	346
reference	M	V	N	L	S	M	T	R	G	K	R	G	E	M	M	Q	A	L	L	N	M	S	S	T	E	F	K	K	I	L	Y	R	H	R	G	N	N	F	I	R	L	Q	Q	E	S	Q	Q	Y	A	G	N	D	R	E	N	Y	S	D	S	L	A	Y	N	A	*
esc_RioGrande	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
peru7233_2	.	.	.	.	.	.	.	.	.	.	.	V	.	.	.	.	.	.	.	.	.	.	.	.	.	L	.	-	.	.	D	I	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	Y	N	A	.	.	.	.	.	.
chil1930_3	.	.	.	.	.	.	.	.	.	E	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	D	.	K	S	.	.	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	.	.	.	.	.	.	H	I
peru7232_5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	D	.	.	.	.	.	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
peru7232_1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	D	.	.	.	.	.	.	.	.	.	.	.	.	.	S	.	.	N	.	F	Y	N	A	.	.	.	.	.	.
chil1930_1	.	.	.	.	.	.	.	.	.	E	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	D	.	K	S	.	.	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
peru1954_2	.	.	S	.	.	.	S	.	A	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	D	.	K	S	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	A	.	.	.	.	.	.
peru0446_2	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	D	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
peru7241_2[a]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	I	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	:	-	-	-
peru7233_3	.	.	.	.	.	.	.	.	.	.	.	V	.	.	.	.	.	.	.	.	.	.	.	.	.	L	.	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	Y	N	A	.	.	.	.	.	.
peru7241_5	.	.	.	.	.	.	.	.	.	.	.	V	.	.	.	.	.	.	.	.	.	.	.	.	.	L	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	E	.	.	.	K	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
peru7234_2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	I	-	.	.	D	.	.	.	.	.	.	.	.	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	A	.	.	.	.	.	.
peru2744_3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	I	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
peru7232_2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	.	.	.	.	K	.	.	.	.	S	.	.	K	.	F	.	.	.	.	T	.	.	.	.
peru7234_3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	E	.	.	.	K	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
peru7236_3	.	.	S	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
peru7237_2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	I	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	.	.	.	T	.	.	.	.
peru7235_B1	.	.	.	V	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	L	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	.	.	.	.	K	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
peru7239_A1[b]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	.	.	.	.	.	.	.	.
peru7239_B1[c]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	A	.	.	.	.	.	.
peru1954_1	.	.	.	.	.	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	Q	.	.	D	.	K	S	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	A	.	.	.	.	.	.
peru7237_A1	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	Y	N	A	.	.	.	.	.	.
peru7241_A1[d]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	I	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	Y	N	A	.	.	.	.	.	.
peru7233_1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	Y	N	A	.	.	.	.	.	.
peru7237_C1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	I	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	.	.	.	.	.	.	.	.
peru7241_3	.	.	.	.	.	.	.	.	.	.	.	V	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	E	.	.	.	K	.	.	.	.	S	.	.	K	.	F	Y	N	A	.	.	.	.	.	.
corn1973_1	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
chil2748_1	.	.	.	.	.	.	.	.	.	.	.	.	K	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	Q	.	K	.	.	.	.	.	E	.	.	.	.	.	.	.	E	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
peru0446_1	.	.	.	.	.	.	.	.	.	E	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	Q	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
peru2744_1[e]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	K	.	.	.	.	.	.	K	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
hab1777_1	.	.	S	.	.	.	S	.	A	.	.	.	.	.	.	E	.	.	.	.	.	P	.	.	.	.	.	-	.	.	D	.	N	Q	.	.	.	.	.	.	.	.	.	G	N	.	H	.	.	.	S	N	.	K	.	F	.	N	A	.	.	.	.	H	I
chil1958_1	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	V	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
peru7234_A2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	I	-	.	.	D	.	.	.	.	.	.	.	.	.	.	E	.	.	.	K	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
peru7236_4	.	.	S	.	.	.	S	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	.	.	.	.	K	.	.	.	.	S	.	.	K	.	F	Y	N	.	.	.	.	.	.	.
peru7236_5	.	.	S	.	.	.	S	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	Y	N	A	.	.	.	.	.	.
peru7238_1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	L	.	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	Y	N	A	.	.	.	.	.	.
peru7232_4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	I	-	.	.	D	.	.	.	.	.	.	.	.	.	.	.	.	.	.	K	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
lyco2951_1	.	I	S	.	.	.	S	T	A	.	.	.	.	.	I	E	T	I	.	-	.	.	.	.	.	.	.	-	.	.	D	.	N	Q	.	.	.	.	K	S	Q	.	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	G	H	.	.	.	.	.
peru7240_A2[f]	.	.	.	.	.	.	.	.	.	.	.	V	.	.	.	.	.	.	.	.	.	.	.	.	.	L	.	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	Y	N	A	.	.	.	.	.	.
corn1274_3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	S	.	.	N	.	F	.	N	.	.	.	.	.	.	.
chil1930_2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	D	.	.	Q	.	K	.	.	.	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
pimp0400_1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	L	.	.	.	E	R	.	.	K	.	.	.	.	.	.	Q	.	.	F	.	.	.	.	.	.	.	.	.
hab1777_3	.	.	S	.	.	.	S	.	A	.	.	.	.	.	.	E	.	.	.	.	.	.	.	.	.	.	.	-	.	I	D	.	N	Q	.	.	.	.	R	.	.	.	.	.	.	.	.	.	.	.	S	N	.	N	K	F	Y	N	.	.	.	.	.	L	.
chm3653_1	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	T	.	D	.	.	Q	.	.	.	.	.	.	.	.	.	.	.	.	.	.	S	.	S	.	.	.	.	F	.	.	.	.	.	.	.	.	.
pen0716_1	.	.	S	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	-	-	-	-	-	-	.	.	D	.	N	Q	.	.	.	.	K	.	.	.	.	.	.	.	.	.	.	.	S	N	.	K	D	F	.	N	A	.	.	.	.	V	.
pen3791_2	.	.	S	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	V	.	.	.	.	.	.	.	.	-	.	.	D	.	N	Q	.	.	.	.	K	.	.	.	.	.	.	.	.	.	.	.	S	N	.	K	D	F	.	N	A	.	.	.	.	V	.
peru7233_A2	.	.	S	.	.	.	S	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	Y	N	A	.	.	.	.	.	.
peru7232_3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	.	.	.	T	.	.	.	.
peru7240_1[g]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	.	.	.	.	K	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	.	.	.	.
corn1274_1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	.	.	.	.	:	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
peru7241_B2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	.	.	.	E	.	.	.	K	.	.	.	.	S	.	.	K	.	F	.	N	.	.	.	:	-	-	-
peru7234_1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	I	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	.	.	.	.	.	:	-	-
peru7236_6	.	.	S	.	.	.	S	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	.	.	.	.	K	.	.	E	.	.	.	.	.	C	.	.	S	.	.	K	.	F	Y	N	A	.	.	.	.	.	.
peru7239_A2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	D	.	.	.	V	.	.	.	K	.	.	E	.	.	.	.	.	.	.	.	S	.	.	K	.	F	.	N	A	.	.	.	.	.	.
	signalpeptide										pro-domain																					protease-domain																																	

Figure B5: Protein haplotypes of different Rcr3 constructs. The protein sequence of *Solanum lycopersicum* (*esc*) is given in the one-letter code of amino acids and used as a reference. Only variable amino acid positions are shown. Amino acids, which are identical to the *esc* sequence are indicated with dots, similar amino acids with blue, nonsimilar amino acids with red, functionally relevant amino acid changes with yellow and deletions with grey. **X** = variant causing incompatibility with Cf-2, **O** = variant causing insensitivity to inhibition by AVR2; [a]identical on the protein level to peru7236_A1, [b]identical to peru7234_A1, [c]identical to peru7234_B1 and peru7234_B2, [d]identical to peru7233_A1, peru7238_A1 and peru7240A1, [e]identical to peru7232_C2, [f]identical to peru7238_A2, [g]identical to peru7235_B2, peru7236_B1 and peru7241_B1

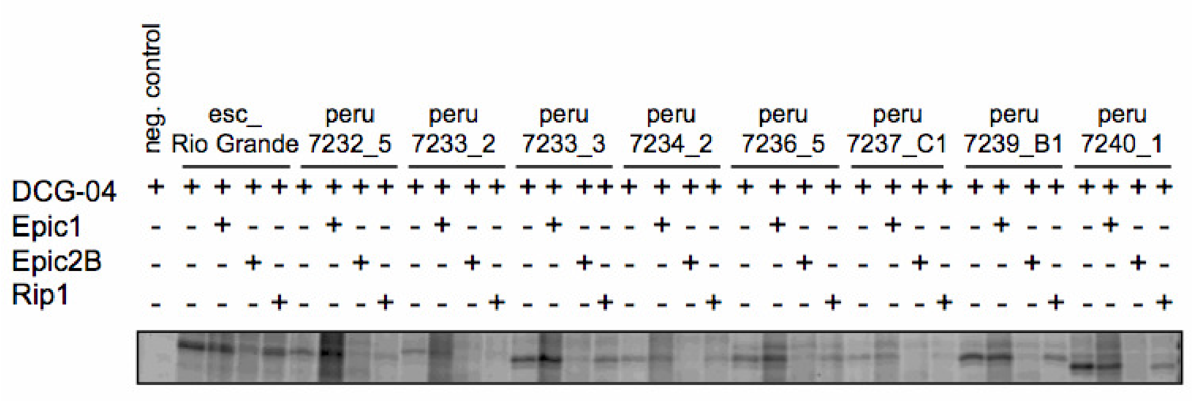
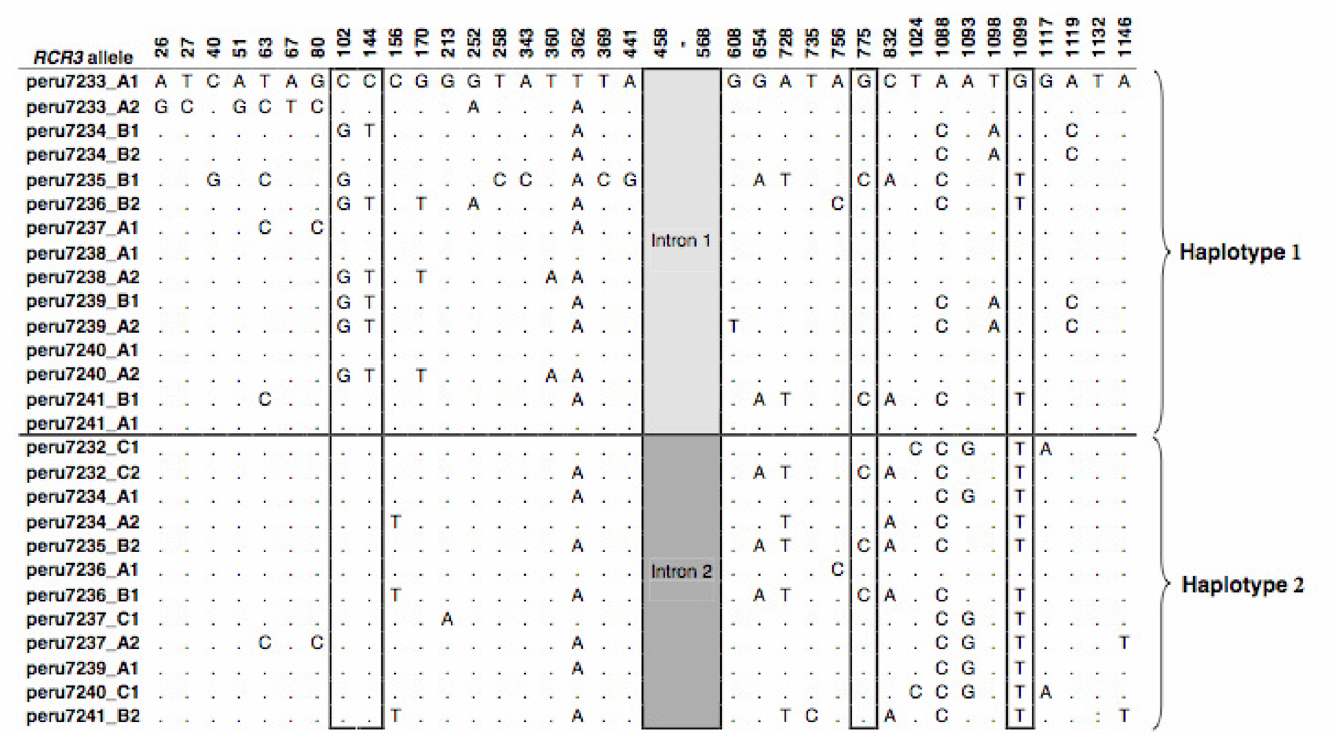

neg. control
esc_ Rio Grande
peru 7232_5
peru 7233_2
peru 7233_3
peru 7234_2
peru 7236_5
peru 7237_C1
peru 7239_B1
peru 7240_1
DCG-04 +
Epic1 - - + - - - + - - - + - - - + - - - + - - - + - - - + - - - + - - - + - -
Epic2B - - - + - - - + - - - + - - - + - - - + - - - + - - - + - - - + - - - + -
Rip1 - - - - + - - - + - - - + - - - + - - - + - - - + - - - + - - - + - - - +

Figure B6: Inhibition assays with Epic1, Epic2B and Rip1 of a subset of Rcr3 constructs. AFs that contained overexpressed Rcr3 constructs were labeled with DCG-04 at pH 5.5 after 30 min. pre-incubation with the corresponding effector. Proteins were separated on 12% protein gels. AFs without overexpressed Rcr3 were used as a negative control. AFs containing Rcr3 from *S. lycopersicum* (*cv*. Rio Grande) were used as a positive control.



RCR3 allele	26	27	40	51	63	67	80	102	144	156	170	213	252	258	343	360	362	369	441	458 - 568	608	654	728	735	756	775	832	1024	1088	1093	1098	1099	1117	1119	1132	1146	
peru7233_A1	A	T	C	A	T	A	G	C	C	C	G	G	G	T	A	T	T	T	A	Intron 1	G	G	A	T	A	G	C	T	A	A	T	G	G	A	T	A	Haplotype 1
peru7233_A2	G	C	.	G	C	T	C	.	.	.	.	.	A	.	.	.	A	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
peru7234_B1	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	A	.	.		.	.	.	.	.	.	.	.	C	.	A	.	.	C	.	.	
peru7234_B2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.		.	.	.	.	.	.	.	.	C	.	A	.	.	C	.	.	
peru7235_B1	.	.	G	.	C	.	.	G	.	.	.	.	.	C	C	.	A	C	G		.	A	T	.	.	C	A	.	C	.	.	T	.	.	.	.	
peru7236_B2	.	.	.	.	.	.	.	G	T	.	T	.	A	.	.	.	A	.	.		.	.	.	.	C	.	.	.	C	.	.	T	.	.	.	.	
peru7237_A1	.	.	.	.	C	.	C	.	.	.	.	.	.	.	.	.	A	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
peru7238_A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
peru7238_A2	.	.	.	.	.	.	.	G	T	.	T	.	.	.	.	A	A	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
peru7239_B1	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	A	.	.		.	.	.	.	.	.	.	.	C	.	A	.	.	C	.	.	
peru7239_A2	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	A	.	.		T	.	.	.	.	.	.	.	C	.	A	.	.	C	.	.	
peru7240_A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
peru7240_A2	.	.	.	.	.	.	.	G	T	.	T	.	.	.	.	A	A	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
peru7241_B1	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	A	.	.		.	A	T	.	.	C	A	.	C	.	.	T	.	.	.	.	
peru7241_A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
peru7232_C1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Intron 2	.	.	.	.	.	.	.	C	C	G	.	T	A	.	.	.	Haplotype 2
peru7232_C2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.		.	A	T	.	.	C	A	.	C	.	.	T	.	.	.	.	
peru7234_A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.		.	.	.	.	.	.	.	.	C	G	.	T	.	.	.	.	
peru7234_A2	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.		.	.	T	.	.	.	A	.	C	.	.	T	.	.	.	.	
peru7235_B2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.		.	A	T	.	.	C	A	.	C	.	.	T	.	.	.	.	
peru7236_A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	
peru7236_B1	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	A	.	.		.	A	T	.	.	C	A	.	C	.	.	T	.	.	.	.	
peru7237_C1	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	C	G	.	T	.	.	.	.	
peru7237_A2	.	.	.	.	C	.	C	.	.	.	.	.	.	.	.	.	A	.	.		.	.	.	.	.	.	.	.	C	G	.	T	.	.	.	T	
peru7239_A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.		.	.	.	.	.	.	.	.	C	G	.	T	.	.	.	.	
peru7240_C1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	C	C	G	.	T	A	.	.	.	
peru7241_B2	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	A	.	.		.	.	T	C	.	.	A	.	C	.	.	T	.	.	:	T	

Figure B7: Sequence polymorphisms of all *Rcr3* alleles, which could be assigned to their genomic origin. The alleles are ordered according to their intronic haplotype. Boxes indicate polymorphisms, which are correlated with the weak HR phenotype.

CURRICULUM VITAE

Place and date of birth:	18th September 1981 in Munich, Germany
Nationality:	German
Languages:	fluent in German (mother tongue) and English, good proficiency in French

EDUCATION

submission Apr 2011	**PhD in Biology** Ludwig-Maximilians-University (LMU), Munich, Germany Department of Evolutionary Biology
2007	**Diplom (Master of Science) in Biology** Ludwig-Maximilians-University (LMU), Munich, Germany Main Focus: Population Genetics, Evolutionary Biology, Microbiology, Genetics, Immunology, Anthropology and Human Genetics

RESEARCH EXPERIENCE

since 2007	**PhD Student of Biology** Department of Evolutionary Biology, LMU, Munich (supervision: Prof. Wolfgang Stephan and Dr. Laura Rose) "Evolution of disease resistance genes in wild tomato species"
2007	**Master Thesis** at the Department of Evolutionary Biology, LMU, Munich "Sequence evolution of the resistance genes *Rcr3* und *Rin4* in wild tomatoes (*Lycopersicon peruvianum*)"
2005	**Internship** at the GSF-National Research Center for Environment and Health, Munich, Department of Molecular Immunology "Role of immune cell surface protein expression during cancer and myocardial infarction"
2005	**Internship** at the Max-von-Pettenkofer Institute, Munich, Department of Virology "Establishing a real-time PCR based method for RNA virus (Hepatitis C) diagnostics in humans"

PUBLICATIONS

Rose, L. E., *Grzeskowiak L., ***Hörger A. C.**, Groth M. and Stephan W.
Targets of selection in a disease resistance network in wild tomatoes, 2011, *Molecular Plant Pathology* (accepted pending minor revisions), *These authors contributed equally.

Hörger A. C., Ilyas M., Stephan W., Tellier A., van der Hoorn R. A. L., and Rose L. E. Evolution of the tomato *Rcr3* resistance gene family is driven by balancing selection for activation of the defence response *Nature* (in review)

STUDENT SUPERVISION

2010	Katharina Böndel, MS student 2nd year
2009	Stephanie Weigl, MS student 2nd year

TEACHING EXPERIENCE

2009	**Course Instructor** Advanced Seminar: Modern Techniques in Evolutionary Biology Section of Evolutionary Biology, LMU, Munich
2008	**Course Instructor** Advanced Course: Molecular Evolution in Plants Section of Evolutionary Biology, LMU, Munich
2008	**Teaching Assistant** Basic Course Ecology and Evolutionary Biology Section of Evolutionary Biology, LMU, Munich
2008	**Course Instructor** Basic Course Evolutionary Biology Section of Evolutionary Biology, LMU, Munich
2004 *and* *2005*	**Teaching Assistant** Practical Courses: Microbiology for Students of Medicine and Chemistry Department of Microbiology, LMU, Munich
2004	**Teaching Assistant** Practical Course: Genetics for Students of Biology Department of Genetics, LMU, Munich

GRANTS AND PRIZES

2010	**EES Young Researchers Prize** for PhD students
2010	**EES travel fund** for attending the SMBE Meeting 2010
2008	**EES travel fund** for attending the UK PopGroup Meeting 2008

CONFERENCES

07/2010	**Annual Meeting of the Society for Molecular Biology and Evolution** in Lyon, France Talk: “Evolution of a resistance gene family in wild tomatoes: Relating nucleotide diversity to functional consequences“
05/2010	**VW Status Symposium in Evolutionary Biology** in Frauenchiemsee, Germany Poster: “Evolution of pathogen resistance pathways in wild tomato”
08/2009	**Congress of the European Society for Evolutionary Biology** in Turin, Italy Poster: “Evolution of pathogen resistance pathways in wild tomato”
03/2009	**Annual Evolutionary Biology PhD Meeting of the German Zoological Society (DZG)** in Munich, Germany Talk: “Evolution of pathogen resistance pathways in wild tomato”
12/2008	**PopGroup Meeting 2008** in Cardiff, UK Talk: “Evolution of pathogen resistance pathways in wild tomato”