The evolution of a disease resistance pathway in tomato

Lukasz Grzeskowiak

München 2009

The evolution of a disease resistance pathway in tomato

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

vorgelegt von
Lukasz Grzeskowiak

München, den 01. Oktober 2009

Erstgutachter: Prof. Dr. Wolfgang Stephan

Zweitgutachter: Prof. Dr. John Parsch

Tag der mündlichen Prüfung: 07. Dezember 2009

Summary

In this dissertation research I describe natural variation of five genes at different points in a signaling pathway controlling disease resistance to a bacterial pathogen of tomato, *Pseudomonas syringae*. Since these genes are involved in defense response to the same pathogen, I evaluate how position in the genetic network influences the selective constraint acting on these molecules. Three components of the pathway are encoded by resistance genes that are tightly linked in the tomato genome. Pto and Fen kinases, in complex with the Prf NBS-LRR protein, bind bacterial pathogen effectors and trigger a specific recognition event which initiates a signal leading to an immune response. Furthermore, these host proteins have multiple downstream interaction partners and experience posttranslational regulation such as phosphorylation and ubiquitination. Genes throughout signaling pathways controlling these different processes can be subject to natural selection. I use this system to address specific questions about evolution of a resistance gene complex. I analyze sequences of three resistance genes in natural populations of wild tomato species *Solanum peruvianum*, collected in South America at different altitudes and habitats. This outcrossing species shows the highest level of polymorphism among tomatoes. The patterns of nucleotide diversity and levels of genetic differentiation between populations suggest that these resistance genes have experienced a mixture of natural selection including not only purifying, but also balancing and positive selection. In addition to standard population genetic analyses, I evaluated the statistical associations between polymorphisms of the interacting proteins to determine whether epistatic selection has contributed to the observed patterns of balancing selection through the maintenance of particular combination of alleles. Using bioinformatic analyses of protein sequences, I found a set of significant associations, which could be due to the structural or functional coadaptation and accommodation between these interacting protein partners. I mapped these sites onto known and predicted structures of Pto, Fen and Prf to visualize putative coevolving regions between proteins. These specific positions are candidates for future functional studies.

Contents

CHAPTER 1

INTRODUCTION

1. Selective constraint and coevolution in protein pathways

The rate of evolution can differ radically among proteins (GILLESPIE 1991; LI 1997). This rate variation can be attributed to differences in selective constraint. Proteins subject to greater constraint should show lower rates of amino acid substitution while those that are less constrained should show higher rates. Some of the most variable proteins are those involved in pathogen resistance and self/non-self recognition (HUGHES and NEI 1989; TAKAHATA *et al.* 1992; HEDRICK 1999; CHARLESWORTH 2002; ROSE *et al.* 2004). This cannot be explained by lack of evolutionary constraint, but instead by natural selection maintaining variation. Understanding these differences in constraints and the forces determining them is one of the major challenges of modern biology.

Many proteins do not operate alone, but as components of complex pathways or metabolic networks. The protein connectivity (i.e. the number of protein interactions with the other components of a network) is determined by structural and physico-chemical properties of interacting partners. Thus, the specificity of interactions may determine the level of constraint and hence the rate of molecular evolution. Indeed, in yeast the connectivity of well-conserved proteins in the network is negatively correlated with their rate of evolution. Proteins that have many interactors generally evolve slowly as a greater proportion of their total length may be involved in functional interactions (FRASER *et al.* 2002). Likewise, the position in the pathway or network can affect the evolutionary constraint on the protein. For example, downstream proteins which serve as convergence points of a diverse group of signaling upstream molecules may be subject to greater evolutionary constraint than the upstream molecules. It has been shown that highly pleiotropic genes ought to display much reduced molecular variation (WAXMAN and PECK 1998). Thus, it 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 signaling molecules. Another type of constraint arises due to the degree of redundancy of a pathway, which may depend on whether the proteins are encoded by single copy genes or by duplicate genes with overlapping functions (WAGNER 2001). Finally, the level of constraint can be affected also by the effects that linkage among genes might impart on molecular evolution. In selfing species, linkage may play a significant role because the effective rate of recombination is reduced and selective forces operating on one locus may affect the evolution of associated loci (NORDBORG 2000). In outcrossing species, genetic linkage may create important constraint if the genes involved in the same pathway are physically close. Proteins need to be expressed in the cell in similar amounts at the same time to properly form complexes and perform their function (FRASER *et al.* 2004; BHARDWAJ and LU 2005). Genome-wide analyses in many model organisms show that coexpressed genes tend to be locally concentrated and have significantly stronger conservation of gene order than genes that are not coexpressed (HURST *et al.* 2002; LERCHER *et al.* 2003; STOLC *et al.* 2004; WILLIAMS and BOWLES 2004; SINGER *et al.* 2005; MEZEY *et al.* 2008). Since condensed chromatin could only be open in several places, linked genes are transcribed together more efficiently than non-clustered genes (DE LAAT and GROSVELD 2003; LEE and SONNHAMMER 2003; YI *et al.* 2007). Likewise, genes in functional modules have more similar rates of evolution than genes from different modules (CHEN and DOKHOLYAN 2006). Consequently, since molecules that share a functional relationship are subject to similar evolutionary pressure (for example control mechanisms), they seem to evolve at the same rate and share evolutionary history (FARES and TRAVERS 2006; HAKES *et al.* 2007). Proteins and RNA molecules under functional constraints show signs of correlated mutations and structural accommodation (CHEN and STEPHAN 2003; SUEL *et al.* 2003; GLOOR *et al.* 2005; SOCOLICH *et al.* 2005; WANG and POLLOCK 2007; WILLIAMS and LOVELL 2009). Genes of some interacting proteins have similar phylogenetic profiles or are eliminated together in a new species (PELLEGRINI *et al.*

1999; GOH *et al.* 2000). Moreover, proteins that interact require functional coadaptation, where change in one binding protein has a direct influence on change in the other protein (RAWSON and BURTON 2002). Therefore components of complexes and pathways are subject to constant “fine-tuning” to ensure that mutational perturbation (via natural selection or genetic drift) do not disrupt overall function. This indicates that coevolution – reciprocal selective pressure, is a universal feature of life, not only important between different organisms, such as host and pathogen, but also present at many levels of biological organization (THOMPSON 1994).

2. Epistatic selection

A mechanism that promotes coevolution is natural selection on cosegregating variants across loci. Epistatic selection, recognized by BATESON (1909) and WRIGHT (1932) to play an important role in the genotype to phenotype relationship, is a fundamental idea for understanding many aspects of adaptation, evolution in natural populations and complex genetic diseases (TEMPLETON 2000; MOORE and WILLIAMS 2005). Gene interaction effects are important when systematic associations between genes are created and maintained. The phenotypic effects and therefore the population genetic dynamics depend on the distribution and magnitude of interaction effects (GOODNIGHT 2000; KELLY 2000; WOLF 2000). Gene interaction effects are also important when many genes segregate in the population, such that several mutations have a chance to be collocated in the same genotype. This is the case with high genomic mutation rates. Epistasis may influence the evolutionary consequences of recurrent deleterious mutations (KONDRASHOV 1994) or the evolution of sex and recombination (CHARLESWORTH 1990; PETERS and LIVELY 2000). A mutation may be beneficial only in a specific genetic background and deleterious otherwise. Thus, epistasis is also the basis for the evolution of the genetic architecture of phenotypic traits (WAGNER and ALTENBERG 1996; RICE 2000, TEMPLETON 2000, CHEVERUD 2000).

There are several simple models explaining coevolutionary dynamics caused by epistatic selection in natural populations. For example, in the coadaptation model two mutations are individually neutral but together form a coadapted haplotype with selective advantage (DYKHUIZEN and HARTL 1980; ZHANG and ROSENBERG 2002; TAKAHASI and TAJIMA 2005). It was demonstrated by simulations that the fixation probability of a coadapted haplotype under finite-island assumptions critically depends on migration rate. The best condition for the fixation of the coadapted haplotype is moderate migration, so that the mutant alleles can spread across subpopulations, while at the same time preserving the favorable allelic combination established within each subpopulation. Moreover, the double fixation in a subdivided population should be expected only when the time difference between the two mutational events is short enough, suggesting the essential role played by epistatic selection when both mutations are segregating at low frequencies (TAKAHASI 2007).

In turn, the compensation model assumes that two individually deleterious mutations compensate each other when combined together. The first step is that a slightly deleterious mutation slowly gets fixed by random drift and then is compensated by selection on a compensatory mutation (KIMURA 1985; STEPHAN 1996; INNAN and STEPHAN 2001). In this model the dependence of fixation probability on migration rate would not be expected. If strong selection acts against the intermediate haplotypes in a subdivided population, mutant alleles would individually be kept at low frequencies in each subpopulation. Since it is very unlikely that the two mutations from different subpopulations will meet together in a single locality, individuals in each subpopulation have to wait for a new allele that compensates the deleterious mutation (PHILLIPS 1996).

In contrast to the above models, the model of epistatic selection proposed by SCHLOSSER and WAGNER (2008) assumes that the evolutionary dynamic driven by environmental factors and epistatic interactions includes only adaptive mutations and has no need for random drift. Therefore, it does not lead to a temporally homogeneous elevation of substitution rates in both loci but rather promotes coevolutionary “bursts of substitutions” – periods of elevated substitution rates in both loci which alternate with long periods of few or no substitutions. These coevolutionary bursts reflect the situation in modular networks of interacting genes where, after an environmental change, some adaptive substitutions at one locus may actually decrease the level of coadaptation of other loci in the same network and thus induce selection for compensatory change. The focus in this model is on long-scale evolutionary trajectories and it neglects population dynamics, assuming instead instant selective fixation of any advantageous mutation. Analysis of genes known to interact is required to investigate whether coevolution results in correlated bursts of substitutions.

3. Linkage disequilibrium

One way to detect epistatic selection is through the analysis of linkage disequilibrium (LD). LD is the nonrandom co-occurrence of alleles at different loci within a population (LEWONTIN and KOJIMA 1960; HILL and ROBERTSON 1968; HEDRICK 1987). If epistasis is synergistic, clustering of genes in the same chromosomal region as a coadapted complex or "supergene" would provide a large selective advantage and has been observed for example in genes controlling color patterns involved in butterfly mimicry (CHARLESWORTH and CHARLESWORTH 1975; JORON 2006). It is also possible that loci on different chromosomes, although unlinked, can show high levels of LD within a population. Thus, a significant deviation from random associations may be an indicator of gene interactions due to the linkage or fitness interactions among cosegregating variants. However, epistatic selection may be a weak effect relative to other factors such as mutation and genetic drift within a given population, recent admixture of subdivided populations (that have different allele frequencies) or founder effects (HILL and ROBERTSON 1968; OHTA and KIMURA 1969a, b; LI and NEI 1974; HILL 1975, 1976; AVERY and HILL 1979; OHTA 1982a, b). The stability of LD strongly depends on the recombination rate, especially when linkage is tight (KARLIN and FELDMAN 1978) and it is presumed that recombination would disrupt allelic associations, unless selection was strong enough (LEWONTIN 1974). Hence, both natural selection and demographic history can have large effects on the levels of LD observed in populations.

To analyze the influence of epistatic selection in populations with geographic structure, measures of LD between a pair of loci can be partitioned into contributions within and between populations. Such a partitioning overall LD is an appropriate first step when trying to determine if differences in LD result only from differences in allele frequency or from other factors that differ among populations. If epistatic selection maintains differences in allele frequencies at two or more loci among subpopulations, LD in each subpopulation will persist (LI and NEI 1974; SLATKIN 1975).

4. Case studies of epistatic selection

Although epistatic selection has been the subject of intense debate, its role received little attention in experimental research during the 20th century. For instance, it is known that when two associated loci are evolving under balancing selection, then LD can persist for a long time (LEWONTIN and KOJIMA 1960; KARLIN and FELDMAN 1970; FELDMAN *et al.* 1974). If many loci interact with each other, a large block of LD can be maintained by selection (FRANKLIN and LEWONTIN 1970). However, this theory was depreciated after studies showing that LD could not be detected between alleles of allozyme loci (CHARLESWORTH and CHARLESWORTH 1973; LANGLEY *et al.* 1974). Recent studies documenting interactions within and between genes have revived interest in epistatic selection.

There are several theoretical and experimental analyses of natural phenotypic variation mediated by epistasis. These studies try to link the functional epistasis in a classical sense (that is the result of physical interactions of molecules within gene regulatory networks or biochemical pathways in an individual) and statistical epistasis, arising from the multilocus composition of individuals within a population. This is especially challenging, because the presence of functional epistasis does not necessarily mean that there will be statistical epistasis. Moreover, the absence of statistical epistasis does not mean that there are no interactions between loci (CORDELL 2002; MOORE and WILLIAMS 2005). Gene interactions are just one class of interaction effects that may be relevant and need to be seen in a wider context. Genotype-by-genotype interactions in local populations due to ecological or social links among individuals, or interaction between traits rather than genes, are also significant. Likewise, it is necessary to consider interactions between other parts of epistatic networks, such as promoters, microRNAs and epigenetic modifications (FOLEY *et al.* 2009).

Epistatic interactions underlie regulatory gene networks, for example, in the flowering time in *Arabidopsis thaliana*. Genes in the vernalization pathway, including *FRI* (*FRIGIDA*) and *FLC* (*FLOWERING LOCUS C*) condition plant response to cold temperatures to induce flowering (JOHANSON *et al.* 2000; MICHAELS *et al.* 2004). The gene *FRI* is thought to be the major genetic factor determining flowering time in *A. thaliana* (STINCHCOMBE *et al.* 2004; LEMPE *et al.* 2005). The ability to flower at the appropriate time given local environmental conditions can strongly affect plant fitness

(O'NEILL 1999; EHRENREICH and PURUGGANAN 2006) and the molecular variation at *FRI* locus has been shaped by adaptive evolution (LE CORRE *et al.* 2002; HAGENBLAD and NORDBORG 2002; TOOMAJIAN *et al.* 2006). Several naturally occurring null *FRI* alleles have been described, and molecular analyses indicate that these independently derived alleles result from several deletions in the coding region (JOHANSON *et al.* 2000; LE CORRE *et al.* 2002). In addition, the *FRI* gene in *A. thaliana* modulates a latitudinal cline in flowering time (STINCHCOMBE *et al.* 2004). *FLC* is a MADS-box transcription factor that, together with *FRI*, controls the transition to flowering (SHELDON *et al.* 1999; MICHAELS *et al.* 2004). However, the *FRI* and *FLC* genes are not physically linked: *FRI* is located on chromosome 4, whereas *FLC* is on chromosome 5. Two major *FLC* haplogroups are associated with flowering time variation in *A. thaliana* and have a different geographical distribution. The study of CAICEDO *et al.* (2004) demonstrates that there is epistatic selection between *FRI* and *FLC* genes and this is partly responsible for the latitudinal cline in *A. thaliana* flowering time. First, there is significant linkage disequilibrium between *FRI* and *FLC* loci despite their location on separate chromosomes. The skewed genotypic associations between *FRI* and *FLC* alleles may suggest that the allele combinations are targeted by selection. Second, variation in flowering time associated with *FLC* is observed only in plants that have a putative functional *FRI* allele. This is in line with a model, in which *FRI* up-regulates *FLC* expression, so that ecotypes with deletion alleles of *FRI* have reduced *FLC* activity and cannot express phenotypes associated with *FLC* variation. Third, the data suggest also that *FLC* haplogroups display differences in latitudinal distribution only in ecotypes with putatively functional *FRI* alleles. Although each of these results alone may be insufficient to conclude the action of selection, together they present a compelling case to suggest that the epistatic regulatory gene interaction is maintained by selection for flowering time variation across the species range of *A. thaliana*. Latitudinal clines are classically regarded as strong evidence of selection associated with geographically structured climatic variables (ENDLER 1986; CAICEDO *et al.* 2004).

The molecular evolution of multiple interacting proteins was studied by PRESGRAVES and STEPHAN (2007), who investigated the evolution of six proteins from the nuclear pore complex (NPC). The NPC caught the attention when it was shown that Nup96, a component of the nuclear pore subcomplex Nup107, participates in the deleterious epistatic interactions that cause hybrid incompatibility between *Drosophila* *melanogaster* and *Drosophila simulans* (PRESGRAVES 2003). The authors investigated Nup96 and five other NPC proteins (two other proteins from the subcomplex Nup107 and three nucleoporins) that are known to interact and found that all of them experienced an excess of replacement substitutions in the relatively recent past. A lineage specific analysis of nucleotide substitutions showed that much of the differences are a result of coevolutionary bursts among interacting proteins. The extent and rate of evolution detected by Presgraves and Stephan is much higher than expected by external selection alone. In this case all six genes of the system show evidence of adaptive evolution, compared to 10% of genes genome-wide (*Drosophila* 12 Genomes Consortium 2007). The correlated rate and pattern of sequence evolution suggest that these bursts of substitutions are driven entirely by epistatic selection, rather than a mixture of genetic drift and selection, i.e. where one mutation drifts to fixation followed by epistatic selection on the compensatory mutation (SCHLOSSER and WAGNER 2008). Regardless of whether nucleoporin-based hybrid lethality has a simple or complex basis, the study of nuclear pore proteins suggests that the adaptive coevolution of a large multi-protein complex may have given rise to multiple hybrid-incompatibility genes. These finding may indicate also that divergent coevolution among the interacting partners of macromolecular complexes, particularly those prone to evolutionary conflicts, may drive the evolution of molecular incompatibilities that contribute to speciation (TANG and PRESGRAVES 2009).

A large proportion of studies that tested for loci interaction found epistasis between genes involved in the immune response, indicating that epistasis is an important component of genetic architecture of resistance (CARANTA *et al.* 1997a, b; KOVER and CAICEDO 2001; WILFERT and SCHMID-HEMPEL 2008). Furthermore, resistance genes in many species are clustered in the genome. These clusters comprise several copies of paralogs arising from a single gene family (simple clusters) or colocalized genes derived from two or more unrelated families (complex clusters), and may also contain unrelated single genes interspersed between the homologs (FRIEDMAN and BAKER 2007). For example, several *Drosophila* immune genes interact epistatically and some immune receptors display high levels of LD (LAZZARO *et al.* 2004, 2006; SCHLENKE and BEGUN 2005). The position of resistance genes along a chromosome is non-random and could potentially be explained by transcriptional regulation as well as selection for optimal recombination rate as a consequence of antagonistic host-parasite coevolution (WEGNER 2008).

Examples of epistatic selection were shown in association studies of human disease (WILTSHIRE *et al.* 2006; ABOU JAMRA *et al.* 2007; COUTINHO *et al.* 2007; TSAI *et al.* 2007), but in only a few cases was the functional basis of these potential interactions revealed. One of these involves the genetic interactions underlying multiple sclerosis, a chronic autoimmune disease of the central nervous system. Susceptibility to the disease is associated with different alleles of the polymorphic histocompatibility loci (MHC class II). The main association is with the DR2 haplotype comprising alleles of two genes, *DR2a* and *DR2b*, located 85 kb apart, which are always inherited together (LINCOLN *et al.* 2005; TROWSDALE 2006). GREGERSEN *et al.* (2006) found evidence that natural selection might be maintaining LD between *DR2a* and *DR2b*. To test the idea that epistatic selection is occurring between *DR2a* and *DR2b*, Gregersen and colleagues constructed genetically engineered mice that express the corresponding human immune proteins. They found that mice producing the protein encoded by *DR2b* were highly susceptible to disease, while those producing the *DR2a* protein did not progress towards disease. In the next test, mice expressing both alleles had an overall reduced susceptibility to disease, suggesting that *DR2a* modulates the impact of *DR2b*. One possible model for this interaction is that production of antigen sensitive T cells is stimulated by *DR2b* and the antigen induces multiple sclerosis, whereas *DR2a* suppresses or even leads to the death of these T cells. Such an interaction could help to explain why these negative effects could be segregating within human populations: under most conditions the influence of the two genes is balanced. Perhaps *DR2b* provides a vital function in controlling a real pathogen and the presence of *DR2a* alongside may keep self-harm to a minimum, like keeping an aggressive dog on a collar and chain (TROWSDALE 2006). However, multiple sclerosis is a complex disease with a rather weak genetic signal and the epistatic interaction of these two immune loci has yet to be tested in humans (SVEJGAARD 2008).

Autoimmune response due to epistatic interactions between alleles can also be observed in plants. Connection between disease resistance pathways and hybrid necrosis has been suggested in tomato, tobacco and *Arabidopsis*, indicating that this mechanism may act to create postzygotic gene flow barriers in diverse plant species. In tomato the polymorphic *Cf-2* gene, which was originally identified in *Solanum pimpinellifolium* as conferring resistance against the fungus *Cladosporium fulvum* expressing Avr2*,* can cause autonecrosis when introgressed into domesticated tomato, *S. lycopersicum* (KRUEGER *et al.* 2002). However, autonecrosis occurs only in plants homozygous for *S. lycopersicum* alleles at a second gene, *Rcr3.* This gene encodes a protease, the possible Avr2 target (ROONEY *et al.* 2005). In addition, two homologues of *C. fulvum* resistance gene *Cf-9* from wild tomato species caused autonecrosis when transiently expressed in *Nicotiana benthamiana.* Based on that, it was proposed that constitutive R-gene activation in hybrids might contribute to the maintenance of interspecific postzygotic hybridization barriers (WULFF *et al.* 2004). A possible scenario is that selection pressure on disease resistance genes involved in host-pathogen coevolution caused the genes to diverge in the parental lineages to the point where they became incompatible in the hybrid genome. Furthermore, proper regulation of resistance protein signaling depends on combinations with additional host proteins, with which they are genetically or/and physically linked (JONES and DANGL 2006, CHISHOLM *et al.* 2006; BOMBLIES *et al.* 2007).

5. Disease resistance in plants and animals

Plants, like other organisms, including animals, are constantly exposed to micro-organisms. They have coevolved with microbes since the first appearance on land and disease resistance is one of their evolutionary successes (GEHRIG *et al.* 1996; CHISHOLM *et al.* 2006). The majority of plant species are resistant to all isolates of any microbial species (DANGL and JONES 2001; NUERNBERGER and LIPKA 2005). However, unlike animals, plants lack comparable mobility and adaptive immune system with somatically generated new resistance specificities provided by B and T cells (VIVIER and MALISSEN 2005; MARTINON and TSCHOPP 2005). Instead, they rely only on innate, but equally effective, pathogen-recognition mechanisms. The ability to distinguish self from non-self is critical to mount an efficient immune response against potential pathogens.

Disease resistance in plants is conferred by recognition, signal transduction and defense activation. They use two major classes of innate immune receptors – pattern recognition receptors (PRRs) and disease resistance (R) proteins. PRRs detect highly conserved microbe-associated molecular patterns (MAMPs) present in many bacterial species. MAMPs include bacterial flagellin, specific nucleic acids, lipopolysaccharides, peptidoglycan and other molecules that are not produced by the host itself (DA CUNHA *et al.* 2006; ZIPFEL *et al.* 2004, 2006; ROBATZEK *et al.* 2007). The detection of MAMPs by PRRs results in MAMP-triggered immunity (MTI). MTI seems to be a highly conserved mechanism evolved in both plants and animals and defines the first layer of active defense that a pathogen must avoid or overcome for successful infection (NUERNBERGER *et al.* 2004). R-proteins recognize the structure or action of isolate-specific pathogen effectors encoded by so-called avirulence (*Avr*) genes. The detection of *Avr* gene products by R-proteins results in effector-triggered immunity (ETI) and represents a second layer of inducible defense. R-genes encode predominantly intracellular immune receptors containing a central nucleotide binding site domain (NBS) and C-terminal leucine-rich repeats (LRRs) and are structurally related to NOD-like receptors (NLRs) of the vertebrate innate immune system (SHEN and SCHULZE-LEFERT 2007). Virulence of most Gram-negative bacterial pathogens of plants and animals (like *Salmonella*, *Shigella* and *Yersinia)* depends on the type III secretion system (T3SS, a syringe needle-like pilis) that is encoded by hypersensitive response and pathogenicity (*hrp*) genes. Bacterial pathogens use the T3SS to secrete a wide range of effector proteins directly into host cells. This modifies host processes to establish a favorable cell environment for the bacteria.

Most isolated R-genes seem to activate common or overlapping sets of defense in local areas infected by pathogens. Those defense responses include fortification of the cell wall, transcriptional induction of pathogenesis-related genes, synthesis of antimicrobial compounds, production of reactive oxygen species (ROS, i.e. respiratory burst) and, in many cases, a hypersensitive response (HR) which is a form of plant programmed cell death (PCD), analogous to animal apoptosis (HAMMOND-KOSACK and JONES 1997; DANGL and JONES 2001; NUERNBERGER *et al.* 2004). The local resistance triggered by R-genes may also lead to activation of a defense termed systemic acquired resistance (SAR) in the uninfected tissues, which is a more long-lasting immune response against a broad range of pathogens throughout the entire plant (DURRANT and DONG 2004).

Different selective forces driving the evolution of specific R-genes at the molecular level have been documented (MICHELMORE and MEYERS 1998; MEYERS *et al.* 2005). A fundamental mechanism in R-gene evolution comes from disease pressure imposed on plants by pathogens. The type and strength of selection may vary, depending on the mechanisms by which plants recognize pathogens and the levels of pathogen virulence and host resistance. Both diversifying and balancing selection are typical types of selection between R and *Avr* genes. In nature, however the actual situation is far more complicated due to the coexistence of different pathogens and temporally and spatially variable environmental conditions that may favor or restrict plant or pathogen growth. Also, the strength of the diversifying selection or balancing selection may vary depending on the level of the cost associated with expression of the R-gene and the fitness penalties associated with loss of the *Avr* gene (BERGELSON and PURRINGTON 1996; MCDOWELL and SIMON 2006; SACRISTAN and GARCIA-ARENAL 2008).

A comparison of approximately 300 human disease-associated genes shows that almost 60% have an ortholog in *Arabidopsis*, compared to about 70% in nematode *Caenorhabditis elegans* and around 80% in *Drosophila melanogaster* (RUBIN *et al.* 2000; mips.gsf.de/proj/thal/db/tables/disease.html). The high percentage of shared genes is not surprising given that development and disease involve normal and abnormal activities of proteins. These include molecular structures of receptors involved in pathogen recognition, protein kinase-based downstream signaling pathways, use of ROS in direct defense and the production of antimicrobial peptides. One class of antimicrobial peptides similar across kingdom barriers is the defensins, which have been identified in plants, fungi, invertebrates and vertebrates. Such peptides are frequently induced upon infections in both mammals and plants as an important component of basal host defense in interactions with compatible microbial pathogens (THOMMA *et al.* 2002; JIN *et al.* 2004; MYGIND *et al.* 2005). The virulence factors that cause disease often interfere with cell membrane integrity and target fundamental cellular mechanisms. Some pathogens, sometimes referred to as cross-kingdom pathogens, can infect organisms from different kingdoms and use them successfully as hosts. Examples have been reported of mammalian pathogens that are also plant pathogens, and vice versa (AUSUBEL 2005; VAN BAARLEN *et al.* 2007).

The casual agent of plant disease, *Pseudomonas syringae* shares a key infection mechanism with other plant-infecting bacteria. Similar mechanisms of infection are found in human bacteria such as closely related *P. aeruginosa*, but also *Escherichia coli*, *Staphylococcus aureus, Salmonella* and *Yersinia*. For instance, *P. syringae* injects directly into plant cells the effector protein AvrPtoB, which has ubiquitin ligase activity and inhibits PCD initiated by the disease resistance proteins (JANJUSEVIC *et al.* 2006). Delayed cell death is crucial to successful host colonization and bacterial proliferation. The activity of protein ubiquitination in PCD has been revealed in many organisms (ZENG *et al.* 2006). Furthermore, AvrPtoB can also suppress PCD in *Saccharomyces cerevisiae* (ABRAMOVITCH *et al.* 2006). This indicates a high degree of similarity that helps to overcome the molecular defense processes in different hosts.

The structural similarity of R-proteins to animal immunity proteins and the similarity in the overall signaling structure of the defense reactions directly involved in attacking invading pathogens in both plants and animals, provoked speculation that the domains of defense proteins might have evolved in an ancient unicellular eukaryote predating the separation of the plant and animal kingdoms around 1.6 billion years ago (WANG *et al.* 1999; DANGL and JONES 2001; FLUHR and KAPLAN-LEVY 2002; NUERNBERGER and BRUNNER 2002; NUERNBERGER *et al.* 2004). After comparative study of the overall recognition and signaling mechanisms of animal immunity proteins and plant R-proteins, it was proposed that these apparently analogous regulatory modules used in innate immunity, evolved independently by convergent evolution and reflect inherent constraints on how an innate immune system can be constructed (AUSUBEL 2005).

Because of such similarity plants have been used as a system to study the important human pathogens (CAO *et al.* 2001; PRITHIVIRAJ *et al.* 2005). Plant models allow us to investigate the modes of action of microbial factors and their corresponding host targets using number of mutants in a cost-effective way without needing ethical clearance. The *Arabidopsis* mutants that differentially react to microbial virulence factors are publicly available to laboratory experiments involving bacterial and fungal pathogens (GREENBERG and AUSUBEL 1993; ASAI *et al.* 2000). Furthermore, databases dedicated to plant research and tools that enable cellular pathway reconstruction from plant gene expression profiles and protein functional data are valuable for comparative studies between plants and humans. Such tools definitely facilitate the discovery of novel signaling and metabolic pathways. In future research, the power of comparative and systems biology can completely be used to find and compare plant and human cellular pathways directly from microarray data (VAN BAARLEN *et al.* 2008).

6. Tomato as a model system to study evolution of disease and stress resistance

It is pertinent to mention why we should actually care about tomato plants in study of disease resistance pathways. This dissertation research represents just only one example of making use of tomato as a model system to understand the molecular basis of disease resistance. Even Charles Rick, the world's foremost authority on tomato genetics, recognizing that the tomato species offer unique advantages for certain investigations at the fundamental and more applied levels, said the following: "if *Arabidopsis* is the *Drosophila* of plant genetics, then the tomato has become the mouse" (RICK 1991). Indeed, these model plants not only have impact on plant biology, improving crop food security and reducing malnutrition, but also help to explain basic life processes and the evolutionary plasticity of cellular pathways and networks, which are important in human health. The broad adaptive diversity, wide species range and recent divergence make wild tomatoes a perfect model for evolutionary analysis at both the population and species level. Because of the strong environmental gradients and connection between climate adaptation and habitat isolation, understanding the genetic basis of adaptation can provide answers not only for the evolutionary origin of ecological barriers to reproduction but also responses to local and regional climatic differences (COYNE and ORR 2004).

Functional analyses in wild tomatoes have considered how abiotic and biotic stresses have affected the natural species ranges. These include studies of drought resistance (RUDICH and LUCHINSKY 1986; BLOOM *et al.* 2004), thermal tolerance (SCOTT and JONES 1982), salt tolerance (FOOLAD 2004), disease resistance (STEVENS and RICK 1986; LEGNANI *et al.* 1996) and the specific mechanisms that underlie these phenotypes. Many studies suggest that the substantial morphological and physiological trait variation observed in wild tomato species are adaptive responses to their native environmental context (RICK 1973, 1976a; VALLEJOS 1979; NAKAZATO *et al.* 2008). Geographic races within species also appear to exhibit environment-specific adaptive diversity. For example, *Solanum cheesmaniae* seems to have developed high salt tolerance in its coastline habitat on the Galapagos Islands. Long roots of *Solanum chilense* take up water from deep soil in the extremely dry environments of Northern Chile. In turn, adaptation of *Solanum lycopersicum* var. *cerasiforme* to high humidity has been noted in terms of its high tolerance of water-logging and resistance to various fungal infections (RICK 1973; TAYLOR 1986).

Since biotic stresses are considered as fundamental drivers of evolution, interactions of tomatoes with pathogens and herbivores could provide substantial insights into nature and evolution of disease resistance. For example, population genetics in combination with functional analyses of *Pto* within domesticated tomato and among wild tomatoes indicate that a large proportion of pathogen response can be attributed to sequence variation in this gene, and that a mixture of purifying and balancing selection appears to maintain replacements at this locus in wild species (ROSE *et al.* 2005, 2007; BERNAL *et al.* 2005). Furthermore, there is an evidence that pathogen and herbivore induced responses are similar to that caused by water, salt and UV radiation (SINGH *et al.* 2002; IZAGUIRRE *et al.* 2003; THALER and BOSTOCK 2004). These associations between biotic and abiotic stresses, suggests common mechanisms and correlated evolutionary history for these adaptive phenotypic responses.

Domesticated tomato *S. lycopersicum* is one of the most popular vegetables worldwide. Compared to wild tomatoes, it has equally complex population history, but with extensive impact of human. Several population bottlenecks through founder effects, and natural and artificial selection during domestication, reduced the variability within *S. lycopersicum* (RICK 1976b). It is estimated that only around 5% of the total genetic variation within clade *Lycopersicon* can be found in this species (RICK and FOBES 1975; MILLER and TANKSLEY 1990). As a result, domesticated tomato is susceptible to a large number of diseases caused by viruses, bacteria, fungi and nematodes. This diversity of pathogens emphasizes the importance of the tomato as a favorable model for studying plant-pathogen interactions. Most of the wild species are relatively resistant to diseases and have been used as a source of resistance genes in modern crop improvement (RICK and CHETELAT 1995; ARIE *et al.* 2007). The extremely diverse *Solanum peruvianum* is the source of several widely deployed R-genes, but it is mostly self incompatible and has various barriers present in sexual hybridization and gene transfer. Nevertheless it can be hybridized with *S. lycopersicum* using pollen mixture, embryo rescue or *S. chilense*-derived bridge lines (RICK 1979b; POYSA 1990; SANCHEZ-DONAIRE *et al.* 2000; PICO *et al.* 2000). This has been utilized, for example, to transfer in resistance to nematodes (*Mi)*, tobacco mosaic virus (*Tm-2*) and tomato spotted wilt virus (*Sw-5*) (STEVENS and RICK 1986; TIGCHELAAR 1986; PICO *et al.* 2002).

If resistance gene resources are unavailable from natural populations, an alternative strategy is to generate R-genes by directed mutagenesis or sequence shuffling with appropriate existing alleles and to screen for DNA clones that can induce pathogen-dependent HR through transient coexpression with cognate *Avr* genes in a suitable host. These synthetic genes may function as new R-genes in the native host to recognize the pathogens carrying new mutated versions of *Avr* genes and can be introduced to desirable plant cultivars through genetic transformation (WULFF *et al.* 2004; BERNAL *et al.* 2005). Such applications of evolutionary principles can be an alternative to traditional genetic modification methods, which introduce genes from foreign species to the host genome and in the long term have to face public opposition (even though no compelling evidence has been found to suggest that the genetically modified organisms utilization is likely to cause harm, e.g. HERITAGE 2005; GURR and RUSHTON 2005; BATISTA and OLIVEIRA 2009; DUNHAM 2009; WILLIAMS 2009). Further clarification of response to environmental factors will allow for precise genetic manipulations and lead to new strategies for improving disease and stress resistance.

7. *Solanum peruvianum*

The focal species in this study, *Solanum peruvianum,* belongs to a small monophyletic clade *Solanum* section *Lycopersicon*, within the large and diverse *Solanaceae* family. The clade *Lycopersicon* consists of 13 closely related species, including the cultivated tomato, *S. lycopersicum* (*L. esculentum*). In general, all members of the clade are diploids ($2n = 24$) with a small-to-medium sized genome (950 Mbp), share the same genomic structure and are intercrossable to some degree. The natural species range of wild tomatoes stretches from Ecuador to northern Bolivia and Chile, with two endemic species in the Galapagos Islands (RICK 1979a; PERALTA and SPOONER 2001; CHETELAT and JI 2007; PERALTA *et al.* 2008).

S. peruvianum is the most polymorphic species within the tomato clade, showing substantial morphological and molecular diversity within and between populations. Forty races or ecotypes have been identified in this species (RICK 1963). A large proportion of the variation found within and between species of the clade *Lycopersicon* is segregating in *S. peruvianum* (BAUDRY *et al.* 2001; ROSE *et al.* 2007).

This species is also the most widespread species and is distributed along the western side of the Andes from northern Peru to northern Chile. It occupies diverse habitats from sea level along the dry Pacific coast to the wet valleys up to 2,500 m (RICK 1973, 1979a, 1986; TAYLOR 1986; PERALTA *et al.* 2005; CHETELAT *et al.* 2009). The pattern of distribution suggests a single origin, spreading through the present range perhaps during the Tertiary period before the uplift of the Andes (RICK 1963). Population history of *S. peruvianum* was shaped significantly by the dynamic recent geological and climatic history of the region, including cyclical warm current events and ongoing tectonic movements. The geographic changes may have influenced the species demography and resulted in admixture events or spatial isolation. Hence, the interpretation of molecular diversity lies in the distinction of historical natural selection and demography from recent occurrence of mutations, genetic drift or migration.

8. The Pto signaling pathway

The focal genes of this thesis are involved in tomato signaling pathway, allowing the plant to overcome the bacterial speck disease, caused by strains of *Pseudomonas syringae* pathovar tomato (*Pst*) expressing the specific, sequence-unrelated ligands, AvrPto and AvrPtoB (Figure 1A). The interaction between tomato plants and the bacterial pathogen is ideal for evolutionary studies because both the resistance genes and the pathogen ligands have been extensively characterized at the molecular level (reviewed in VAN OOIJEN *et al.* 2007). 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 the host and pathogen molecules must physically interact to activate the disease resistance response.

The bacterial speck disease occurs throughout the world where conditions are cool (15–25°C) and wet (JONES 1991). The bacteria are spread by water or contaminated seeds and enter leaves through stomata or wounds where they multiply in the leaf apoplastic space (YUNIS *et al.* 1980; PRESTON 2000). Disease symptoms include small dark necrotic lesions (specks) that can become surrounded by chlorotic haloes, caused by the bacterial toxin. Infection may result in reduced photosynthetic ability, the loss of leaves and flowers (YUNIS *et al* 1980; MCCARTER *et al.* 1983; BENDER *et al.*

1987). Lesions also form on fruits, and this symptom can decrease marketability of the cultivated tomato (JONES 1991).

The bacterial effectors AvrPto and AvrPtoB are translocated and act inside the plant cell. In susceptible tomato plants and in *Arabidopsis*, these effectors contribute to bacterial virulence. Both ligands inhibit early basal defense signaling events triggered by MAMPs, suggesting that they act very close to PRRs (DE TORRES *et al.* 2006; SHAN *et al.* 2008; GOEHRE *et al.* 2008). AvrPto interacts with the basal defense receptor kinases, inhibiting their ability to autophosphorylation and activation of MAP kinase signaling cascades (XIANG *et al.* 2008). In turn, AvrPtoB is a modular protein with a longer N-terminal domain, able to suppress certain basal defense responses in *Arabidopsis*, and a C-terminal domain, which structurally and functionally mimics eukaryotic E3 ubiquitin ligase. This domain targets host proteins for degradation (JANJUSEVIC *et al.* 2006; GOEHRE *et al.* 2008; GIMENEZ-IBANEZ *et al.* 2009).

The *Pto* resistance gene belongs to a small multigene family of five to six family members in *Lycopersicon* clade (MARTIN *et al.* 1993), however functions have not been ascribed to all of these genes. The entire 60 kb region of chromosome 5 containing the *Pto* gene family has been sequenced from a susceptible *S. lycopersicum* cultivar and a resistant cultivar containing the *Pto* locus introgressed from the sister species *S. pimpinellifolium* (Figure 1B; GenBank accessions AF220602 and AF220603). The two haplotypes share five orthologous, clustered genes (*Fen*, *Pth2, Pth3, Pth4* and *Pth5*). Orthologous relationships of the *Pto* gene family members between the resistant and susceptible cultivars were determined based on positional information and sequence identity (D. LAVELLE and R. MICHELMORE, unpublished results).

Pto confers resistance to strains of *Pst* expressing either AvrPto or AvrPtoB. It was the first race-specific R-gene to be isolated (MARTIN *et al* 1993). This small gene without introns and the open reading frame (ORF) of 963 nucleotides encodes a functional serine/threonine kinase capable of autophosphorylation (LOH and MARTIN 1995). Protein kinases are well-studied, integral components of many cellular signaling pathways. Pto protein is in the same kinase class as the cytoplasmic domain of the *Brassica* self-incompatibility gene *SRK*, the *Drosophila* Pelle kinase, the mammalian signaling factor Raf and the human IRAK kinase (SHELTON and WASSERMAN 1993; BRAUN and WALKER 1996; CAO *et al.* 1996; STEIN *et al.* 1996). The current model for Pto activation involves Pto binding to the pathogen ligand in the plant cell and a change in protein conformation, induced through this physical interaction. The stabilization of the Pto molecule in the proper conformation is dependent on Pto kinase activity. Next the activated Pto protein transduces the signal, which is sensed by functional protein Prf to activate downstream plant immune responses. This includes the synthesis of antimicrobial compounds and results in localized cell death at the site of infection (RATHJEN *et al.* 1999; SESSA and MARTIN 2000; WU *et al.* 2004; MUCYN *et al.* 2006, XING *et al.* 2007).

Fen, one of the *Pto* family members, is a functional serine/threonine kinase and confers sensitivity to the insecticide fenthion (MARTIN *et al.* 1994; CHANG *et al.* 2002). The Fen protein shares 80% sequence identity with Pto, but does not confer AvrPto-dependent resistance (SCOFIELD *et al.* 1996; JIA *et al.* 1997; FREDERICK *et al.* 1998). However, this paralog can recognize and activate defense responses to variants of AvrPtoB effector lacking E3 ubiquitin ligase activity (ROSEBROCK *et al.* 2007). Nonetheless, wild type form of AvrPtoB ubiquitinates certain Fen alleles, which leads to their degradation in the plant cell. This suggests that the *Pto* cluster paralogs seem to have experienced a complex history of host–pathogen coevolution. One possible scenario posits that ancestral forms of AvrPtoB (or possibly related molecules from other pathogens) lacked the E3 ubiquitin 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 and to further inhibit basal defense. In contrast to Fen, Pto effectively phosphorylates AvrPtoB and is not sensitive to AvrPtoB-mediated degradation (NTOUKAKIS *et al.* 2009). Phosphorylation by Pto inactivates the E3 ligase and leads to activation of the defense signaling pathway in response to pathogens expressing AvrPtoB. Phylogenetic analyses indicate that Fen is much older than the Pto gene, fitting with a sequential bouts of adaptation and counter adaptation between this gene family and AvrPtoB (RIELY and MARTIN 2001; ROSE 2002).

Both Pto and Fen proteins do not act alone, but require a second protein, Prf, for the activation of disease resistance. *Prf* is a large gene embedded within the *Pto* gene cluster, although it is phylogenetically unrelated to *Pto* and its paralogs (Figure 1B). The complete transcribed region of *Prf* is almost 11 kb and contains five introns. In *S. pimpinellifolium* Rio Grande 76R, the 3' end of this gene is located about 500 bp from the ORF of *Fen* and 24 kb from the ORF of Pto. The protein coding region is 5.5 kb long. The resultant Prf protein is a large molecule (209.7 kDa) and contains NBS-LRR motifs, common to many other plant R-proteins (SALMERON *et al.* 1996). The Prf protein sequence can be classified into five domains: the N-terminal domain (amino acids 1–536), solanaceae domain (SD; 537–958), coiled-coil domain (CC, 959–1075); ATPase domain (NBARC, 1076–1430) and leucine-rich repeat domain (LRR, 1431–1824). It was demonstrated that both of the two kinases, Pto and Fen, physically interact with the same N-terminal portion of Prf (MUCYN *et al.* 2006, 2009; NTOUKAKIS *et al.* 2009). Silencing of Prf prevents signaling by Fen or Pto, indicating that Prf acts epistatically to Fen and Pto. Further support for this epistatic relationship is given by MUCYN *et al.* (2009), who observed that tomato Fen physically interacts with native form of Prf in *Nicotiana benthamiana*, but not with tobacco Prf. The authors propose that in the hybrid Fen-Prf complex, tomato Fen activity is suppressed or insufficiently regulated by tobacco Prf. This conflicting outcome could result from structural incompatibility of particular Fen and Prf alleles due to amino acid variation in the interaction surfaces between these protein partners. Thus, not only are *Pto* and *Fen* physically linked with *Prf*, which may indirectly affect their evolutionary history, but the physical protein interaction may require coadaptation between these molecules.

Other potential components of the Pto signaling pathway have been reported, such as the Pto- and Fen-interacting proteins. For example, Pti1 is a protein kinase thought to have a positive role in HR signaling, while Pti4, Pti5 and Pti6 activate transcription of defense-related genes (HALTERMAN 1999; BOGDANOVE 2002). In addition to Pto and Fen substrates, other proteins have been proposed to contribute to the Pto mediated resistance in tomato, such as Prf interacting proteins and a RIN4-like protein. Since Pto and Fen bind pathogen ligands, it is likely that they are located at the terminal portion of the pathway, with Prf being one of the first proteins involved in downstream signaling. The presence of a CC-NBS-LRR structure in Prf suggests its involvement in downstream protein-protein interactions. Using yeast two-hybrid screens with different portions of the Prf protein, five Prf interacting proteins were identified. After studying the biological relevance of these Prf interactors in the resistance to *Pst*, it was suggested that the Pto signaling pathway involves Prf combined with both positive and negative regulators (TAI 2004).

One of Prf interactors, *Pfi* (originally named Prf-interactor 30137) encodes a protein with homology to basic helix-loop-helix (bHLH) transcription factors and was identified in a screen with a portion of the CC-NBS region of Prf. Functional testing of this gene indicated that overexpression in tomato suppresses the HR, while viral induced gene silencing (VIGS) of *Pfi* showed no phenotypic response (TAI 2004). As such, this gene appears to be a negative regulator of the HR. Controls using other elicitors of HR, including the pathogen proteins AvrRpm1, AvrB, AvrRpt2 and elicitin, indicated that the observed HR suppression was specific to the Pto pathway. Beside a putative bHLH DNA binding domain, further bioinformatic analyses of *Pfi* revealed two additional regions of interest in this gene: a nuclear localization signal (NLS) and a region that shows some homology to hydrolases.

The other downstream gene in this study is *RIN4*, originally identified as a multifunctional regulator of resistance against *P. syringae* in *A. thaliana* (MACKEY *et al.* 2002). RIN4 plays a role in different R-gene signaling pathways, is a negative regulator of the basal defense response and a cleavage target of several bacterial virulence effectors (KIM *et al.* 2005). A search of EST databases revealed that *RIN4*-like gene is also present in tomato, potato, soybean and lettuce. For instance, *S. lycopersicum* homolog of RIN4 has 37% amino acid identity to the *Arabidopsis* RIN4 and conserved cleavage sites. It was proposed that the tomato RIN4 homolog is involved in the Pto signaling pathway, i.e. the recognition of AvrPto by Pto resulted in a Prf-dependent activation of a downstream proteolytic pathway that degrades RIN4. Since RIN4 is believed to be a negative regulator of basal defense, its degradation is predicted to activate these defenses. In this way, RIN4 may also play a role in the Pto-Prf pathway, possibly enhancing the resistance response through its specific degradation in the presence of AvrPto. This suggests that RIN4 in tomato, as in *A. thaliana*, could be a point of vulnerability, exploited by bacterial pathogen (LUO *et al.* 2009).

9. This research

In the present study I focus on sequence variation of five genes at different points in a signaling pathway controlling disease resistance in tomato: *Pto*, *Fen*, *Prf*, *Pfi* and *RIN4* (Figure 1A). This allows me to answer questions about the nature of evolutionary constraint in signaling pathways:

- Does natural selection differentially affect the evolution of genes in the same signal transduction pathway?
- Do proteins acting upstream in the pathway experience lower selective constraint that proteins acting downstream?

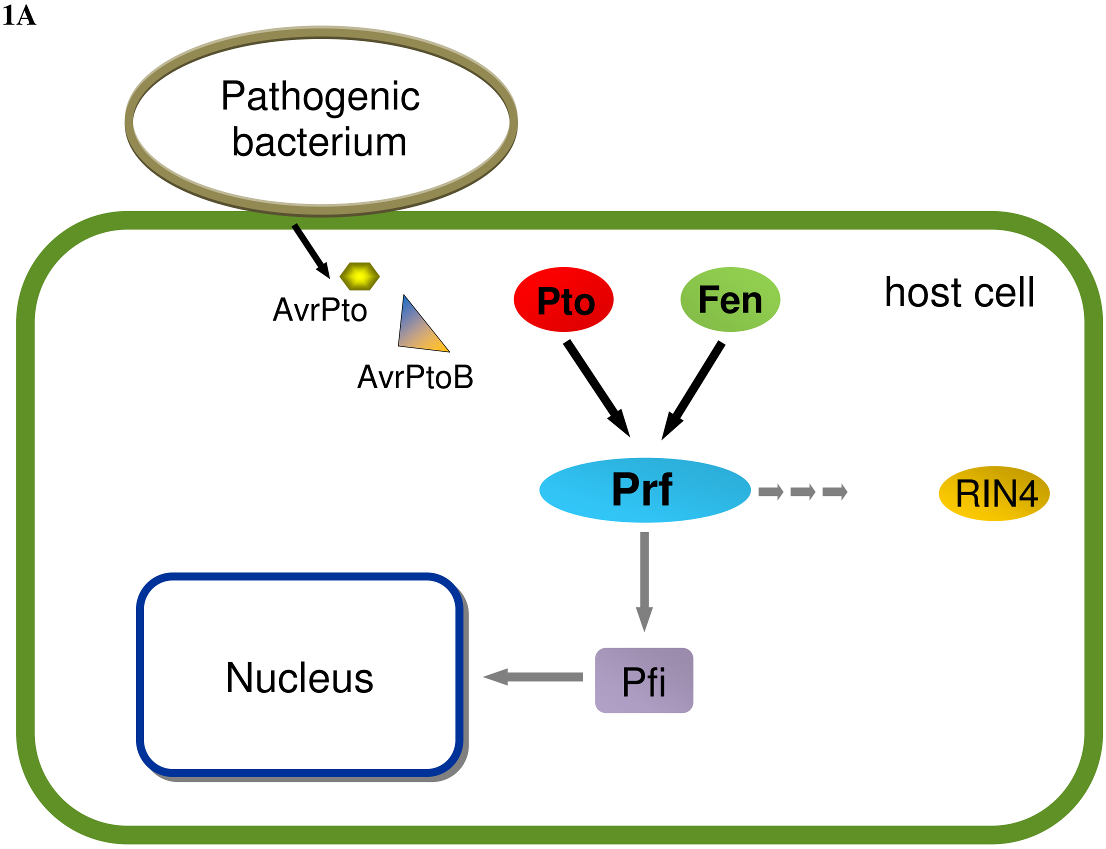
I analyze constraints that arise due to both the general requirements of pathways and the physical or functional linkage of the genes involved. Such case studies complement analyses of large protein databases, because in case studies the forces underlying selective constraint can be analyzed in much greater detail.

The evolutionary dynamics at *Pto* and *Fen* may well influence the evolutionary dynamics of *Prf*, through the indirect effects of linkage or the direct effects of coadaptation (Figure 1B). Previous studies have shown that activity of tomato Fen is suppressed in hybrid complex with tobacco Prf (MUCYN *et al.* 2009). In turn, *Pto* is subject to a mixture of balancing and purifying selection (ROSE *et al.* 2007). This suggests that some types of *Pto* and *Fen* may function best with certain types of *Prf*. I have been able to use this system to address more specific questions about evolution of a resistance gene complex:

- Do tightly linked genes *Pto*, *Fen* and *Prf* evolve in a correlated fashion?
- Does epistatic selection operate in the Pto signaling pathway?
- Does the maintenance of allelic variation at the *Pto* and *Fen* genes lead to the maintenance of allelic diversity at the *Prf* locus?
- How could the sequence variation at *Pto*, *Fen* and *Prf* affect resistance?

Following the study of compensatory evolution in the pre-mRNA of the *Drosophila* gene coding for alcohol dehydrogenase (SCHAEFFER and MILLER 1993; KIRBY *et al.* 1995; CHEN and STEPHAN 2003), I analyze correlated evolution in the *Pto* resistance gene complex. I identify amino acid positions that are candidates for coevolving sites between Pto/Fen and Prf using standard linkage disequilibrium, as well as partitioning of linkage disequilibrium components across populations and correlated substitution analysis. These candidates were mapped onto known and predicted structures of Pto, Fen and Prf to visualize putative coevolving regions between proteins. Functional significance of these coevolving pairs is discussed in the context of what is known from previous structure-function studies of Pto, Fen and Prf.

1A



1B

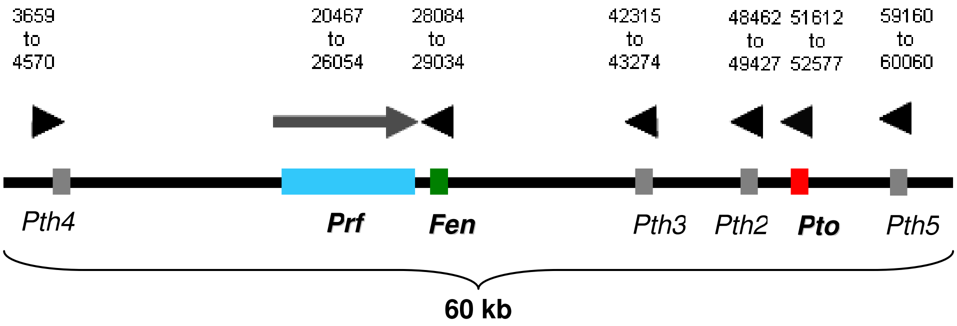


FIGURE 1. (A) Basic model of the signal transduction pathway characterized in this study (see text); (B) The *Pto* cluster in *S. pimpinellifolium*. Arrows indicate the direction of transcription (i.e. 5’ to 3’). Numbers indicate position of open reading frames in bp in the 60 kb region.

CHAPTER 2

MATERIALS AND METHODS

1. Plant materials

I sampled *Solanum peruvianum* from three different geographical locations: 1) Canta (Central Peru, 11°31'S, 76°41'W; 2050 m altitude) 2) Nazca (Southern Peru, 14°51'S, 74°44'W; 2130 m altitude), and 3) Tarapaca (Northern Chile, 18°33'S, 70°09'W: 400 m altitude). Samples from Nazca and Canta were gathered in May 2004 by T. Staedler and T. Marczewski. Six plants were collected per population. DNA was extracted from leaf material using the DNeasy Plant Mini Kit (Qiagen GmbH, Hilden, Germany). Seeds from the Tarapaca population were collected by C. Rick in April 1986 and stored at the Tomato Genetics Resource Center (TGRC) at the University of California, Davis (tgrc.ucdavis.edu; accession LA2744). This accession, exceedingly variable for many traits, is a member of the TGRC core collection. In 1996 seeds from ten different plants were germinated and grown under standard greenhouse conditions in Davis, CA. Genomic DNA was isolated using CTAB method (DOYLE and 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, I used an individual of *Solanum lycopersicoides* from Tarapaca, Chile (TGRC accession LA2951) or *Solanum habrochaites* from Ancash, Peru (LA1775). Plant growth conditions and DNA extraction were identical as used for the Tarapaca population.

2. Gene amplification and sequencing

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 (Stratagene, La Jolla, CA) or Phusion (Finnzymes, Espoo, Finland). 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 were conducted in parallel for each gene on an ABI 3730 DNA analyzer (Applied Biosystems/Hitachi, Foster City, CA). 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

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*.

Fen

A similar strategy as used for *Pto* was employed for sequencing of *Fen* alleles. 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 *S. lycopersicum* cv. Rio Grande 76R haplotype. These two intergenic primers, FenFor and FenRev, were used in combination and with SSP19 or SSP17, respectively.

Prf

Prf is a large gene (5587 bp from start to stop codon), therefore it was divided into two overlapping parts for PCR and these were sequenced separately. The first part of *Prf* is well-known for being recalcitrant to cloning, so here a direct sequencing strategy, combined with allele-specific primers to resolve phase was used. Both direct sequencing of PCR products and cloning were employed to generate the data for the second part of the gene (approximately 58% of *Prf*). A large number of primers (>90) were designed for sequencing and allele-specific amplification. For individuals from Nazca and Canta, the first 1701 bp of *Prf* was amplified. These PCR products were sequenced and phase was inferred using the ELB algorithm implemented in Arlequin (EXCOFFIER *et al.* 2003, 2005)

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.

RIN4-like gene

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 (sgn.cornell.edu). The gene prediction program GeneMark (BORODOVSKY and MCININCH 1993; exon.gatech.edu/GeneMark) was used to predict the ORF of the putative tomato *RIN4*-like gene. Primers were designed based upon the tomato genomic sequence and incorporated the gene prediction information. Two primers (Rin4For3 and Rin4Rev5) were used to amplify nearly the entire coding sequence of *RIN4*.

Reference genes

The sequences of 14 nuclear loci served as my reference gene set. These loci and the Pto pathway genes were sequenced from the same individuals. The reference loci were developed from cDNA markers used in the genetic map of tomato (TANKSLEY *et al.* 1992). They experience a range of recombination rates (STEPHAN and LANGLEY 1998) and have proposed putative functions (Table 1; sgn.cornell.edu).

TABLE 1. Reference genes from *S. peruvianum* used in this study.

Gene	Putative encoded protein	GenBank accession number		
		S. peruvianum populations		
		Tarapaca	Nazca	Canta
CT066	Arginine decarboxylase	AY941554– AY941563	EU077712– EU077723	EU077724– EU077735
CT093	S-adenosylmethionine decarboxylase proenzyme	AY941582– AY941591	EU077780– EU077791	EU077792– EU077803
CT166	Ferredoxin-NADP reductase	AY941690– AY941697	EU077849– EU077860	EU077861– EU077872
CT179	Tonoplast intrinsic protein Δ-type	AY941716– AY941725	EU077916– EU077927	EU077928– EU077939
CT198	Submergence induced protein 2-like	AY941744– AY941753	EU077980– EU077989	EU077990– EU077999
CT208	Alcohol dehydrogenase, class III	EU077614– EU077621	EU077632– EU077643	EU077644– EU077655
CT251	At5g37260-like protein (transcription factor involved in circadian regulation)	AY941415– AY941424	EU078040– EU078051	EU078052– EU078061
CT268	Receptor-like protein kinase	AY941461– AY941470	EU078108– EU078119	EU078120– EU078131
CT099	Copper binding protein	AY941610– AY941619	–	–
CT114	Phosphoglycerate kinase	AY941636– AY941645	–	–
CT143	Sterol C-14 reductase	AY941323– AY941332	–	–
CT148	Copper/zinc superoxide dismutase	AY941664– AY941673	–	–
CT189	40S ribosomal protein S19	DQ104648– DQ104657	–	–
Sucr	Vacuolar invertase	AY941509– AY941518	–	–

3. DNA sequence analyses

The standard summary statistics (including π, haplotype diversity, Tajima's *D*, Z_{nS}, F_{ST}) and McDonald-Kreitman (MK) test statistics were calculated using DnaSP (LIBRADO and ROZAS 2009). This program was also used to conduct coalescent simulations to examine whether the pattern of substitutions at synonymous and non-synonymous sites at *Pto* and *Pfi* differed from the 14 other genes from the same individuals. The population recombination parameter ρ was estimated using composite likelihood method of HUDSON (2001), implemented in the LDhat package (MCVEAN *et al.* 2002). The expected decay of linkage disequilibrium within resistance genes was modeled using the equation given by HILL and WEIR (1988) and fitted to the data in R statistical package (r-project.org). LD between pairs of sites of Pto and Prf or Fen and Prf was calculated using the composite-disequilibrium R^2 statistic (ZAYKIN *et al.* 2008). This method allows for greater than two alleles per site and can be applied to genotypic data (i.e. unphased data). This composite LD can be interpreted as the total correlation between a pair of loci (COCKERHAM and WEIR 1977; WEIR 1979, 1996). It is estimated directly from genotypic counts and is not biased by inbreeding or higher order departures from random assortment (i.e. Hardy–Weinberg equilibrium). The program MCLD was used to calculate both the approximate and exact (permutational, based on 30,000 permutations) p-values for R^2 tests (ZAYKIN *et al.* 2008). Individuals carrying pseudogenes were excluded from these analyses. Only two pseudogenes were observed among the 44 alleles sequenced from these three genes. Both pseudogenes were found in the Tarapaca population – one in *Pto* from individual 7232 and one in *Fen* from individual 7236. Therefore nine genotypes for each gene combination (*Pto* versus *Prf* or *Fen* versus *Prf*) were analyzed from Tarapaca, six genotypes from Nazca and six genotypes from Canta. Singleton polymorphisms were excluded from LD analyses.

Departures from linkage equilibrium may be caused by natural selection or stochastic (neutral) processes. I applied a method proposed by Ohta to partition the variance components of linkage disequilibrium to determine what fraction of the observed associations could be attributed to epistatic selection between these genes (OHTA 1982a, b). This LD partitioning is similar to Wright's *F*-statistics describing the partitioning of deviations from Hardy-Weinberg equilibrium frequencies into F_{ST} (the average deviation attributable to differences in allele frequency among populations) and F_{IS} (the average deviation within populations) (WRIGHT 1940).

Ohta's D-statistics consists of the within, the between subpopulations and total components of LD in a subdivided population: D_{IS}, D_{ST}, D'_{IS}, D'_{ST}, D_{IT}. The subscript "IS" stands for "individuals within subpopulations", the "ST" for "subpopulations within the total population" and the "IT" for "individuals within the total population". Thus, D_{IS} is the average LD measured within individuals within subpopulations, D_{ST} is the contribution to the overall LD caused by differences in allele frequencies among subpopulations, D'_{IS} is the variance in the observed frequency that a certain nucleotide combination appears within individuals within subpopulations, D'_{ST} is the variance of LD in the total population, while D_{IT} is the same measure made within all individuals irrespective of the subpopulation they come from.

Ohta's D-statistics discriminate between different sources of LD (OHTA 1982a,b). The use of two inequalities based on LD variance components allows us to characterize three patterns (WHITTAM *et al.* 1983; BLACK and KRAFSUR 1985). Three patterns correspond to three different types of LD: 1) If $D_{IS} < D_{ST}$ and $D'_{IS} > D'_{ST}$, LD is considered to be nonsystematic (i.e. LD is caused by random genetic drift and limited migration among subpopulations); 2) If $D_{IS} > D_{ST}$ and $D'_{IS} < D'_{ST}$, LD is considered to be systematic (epistatic selection) and 3) If $D_{IS} > D_{ST}$ and $D'_{IS} > D'_{ST}$, LD is considered to be unequal systematic (e.g. when epistatic selection does not operate across all subpopulations). Ohta's D-statistics were calculated using the Linkdos program (GARNIER-GERE and DILLMANN 1992).

While linkage disequilibrium analyses from pairs of genes across multiple populations allow us to determine the degree to which epistatic selection has shaped the evolution of these genes, F_{ST} analysis of these same genes across populations allows us to identify signatures of local adaptation or balancing selection operating at these loci individually. Loci showing significantly greater (or lesser) allelic differentiation than the genome wide average can be identified using the method of BEAUMONT and NICHOLS (1996). These loci are candidates for sites experiencing either strong directional selection (e.g. local adaptation, observed F_{ST} > expected) or balancing selection (observed F_{ST} < expected). I implemented this method in the program FDIST2 (BEAUMONT and NICHOLS 1996; FLINT *et al.* 1999), which calculates the F_{ST} estimator of WEIR and COCKERHAM (1984) for each gene in the sample. Coalescent simulations were then performed to generate data sets with a distribution of F_{ST} close to the empirical distribution. Based on this simulated distribution it is possible to calculate quantiles for outlier SNP loci. First I analyzed the eight reference genes from these populations to determine the appropriate mean F_{ST} for creating the expected distribution of F_{ST} and heterozygosity against which to test my resistance genes (*Pto*, *Fen* and *Prf*). Following this first pass, SNP loci falling outside of the 95% confidence intervals were discarded and the analysis was run again to calculate the mean ”neutral” F_{ST}. This procedure is recommended, since it lowers the bias on the estimation of the mean neutral F_{ST} by removing the most extreme loci from the estimation (BEAUMONT and NICHOLS 1996). Simulations were then run using 30,000 iterations, assuming 100 populations, 12 alleles per sample and an infinite mutation model. Simulated F_{ST} values were plotted against heterozygosity to yield a distribution for F_{ST} under a neutral model. Polymorphic sites with F_{ST} values for a given level of heterozygosity that fell outside the 0.95 quantile were considered candidates for directional positive selection. Conversely, loci with F_{ST} values that fell below the 0.05 quantile of the distribution were considered candidates for balancing selection. In addition, to confirm the robustness of the above frequentist method-of-moments approach, I used the Bayesian-based F_{ST} outlier detection method of FOLL and GAGGIOTTI (2008), implemented in the BayeScan software. This method calculates the locus-population-specific F_{ST} coefficients (which are different from observed F_{ST} values in FDIST2) and the posterior probability that a locus is subject to selection as measured by the decimal logarithm of the Bayes factor. The Bayes factor provides a scale of evidence in favor of selection model versus neutral model. To calculate these values I used in total 600 non-singleton SNP loci, both non-synonymous and synonymous, from three R-genes and eight reference genes. The obtained F_{ST} distribution and the results from FDIST2 allowed me to adjust a threshold posterior p-value, which determines a set of candidate loci subject to directional and balancing selection.

4. Protein sequence analyses

I used two methods developed to identify coevolving residues between protein domains to determine which residues in Pto or Fen were likely to be coevolving with Prf. The first method called CAPS (coevolution analysis using protein sequences) is based on a correlation coefficient and measures the correlation between two sites in the pairwise amino acid variability, relative to the mean pairwise variability per site (FARES and TRAVERS 2006). This method can be used to detect sites in which radical changes in one position are matched with radical changes in a second position. The significance of the correlation values was determined by randomization of pairs of sites in the alignment, calculation of their correlation values and comparison of the distribution of 10,000 randomly sampled values with the real values. To correct for multiple tests and for non-independence of data, the step-down permutation procedure was applied (WESTFALL and YOUNG 1993). I implemented this method using the program CAPS (FARES and MCNALLY 2006).

The second method, called ELSC (explicit likelihood of subset covariation), is based on alignment perturbation and also evaluates correlation between sites (DEKKER *et al.* 2004). Here however, the full joint alignment of the two proteins is broken into subalignments based on a per-site inspection. For example, a given site polymorphic in Pto (denoted here as site A) is chosen and the alignment is broken into two sub-alignments – the one subalignment containing all haplotypes linked to the major allele (the most prevalent amino acid polymorphism) at site A and the other subalignment containing the haplotypes associated with the minor allele or alleles at this site. Then the distribution of amino acids at a polymorphic site in Prf (denoted here as site B) in the subalignment containing the major allele of Pto at site A is compared to the distribution of all amino acids at site B. A normalized statistic that gives the probability of drawing at random the composition observed in the subalignment relative to the probability of drawing the most likely composition is then calculated. The final score is the negative natural log of this ratio of likelihoods. High values (>3) are indicative of sites that show correlated evolution. The algorithm was executed in the package provided from www.afodor.net. For both analyses, a multiple sequence alignments of Pto, Fen and Prf from the three populations of *S. peruvianum* were used. The two pseudogene sequences of Pto and Fen were excluded. Gametic phase between Pto and Prf or Fen and Prf was inferred using the ELB algorithm implemented in Arlequin (EXCOFFIER *et al* 2003, 2005).

Tertiary structures of Pto, Fen and Prf

The Pto crystal structure was determined by XING *et al.* (2007; PDB 2qkw), but the native tertiary structures of Fen and Prf have not yet been experimentally determined. Therefore I used I-TASSER (iterative threading assembly refinement algorithm; ZHANG 2008) to predict the structures of Fen and Prf. This method first searches for template proteins of similar folds from the PDB (protein database) library. Then the continuous fragments from the PDB templates are reassembled into full-length models by replica-exchange Monte Carlo simulations and the unaligned regions (mainly loops) are built by *ab initio* modeling. When no appropriate template is identified, I-TASSER builds the entire structure by *ab initio* modeling. Subsequently, fragment assembly simulation is performed to refine the global topology of the protein structure. Final full-atomic models are obtained by optimization of the hydrogen-bonding network. Due its high sequence similarity and evolutionary relatedness, Fen (GenBank accession AAF76307) was modeled by threading onto the crystal structure of Pto. For Prf (GenBank accession AAF76312), only the first 1500 residues were analyzed, including the region, which has been shown to interact with Pto and Fen. In the modeling process several parent proteins with functions essential in disease resistance were used: (1) a protein phosphatase – scaffold protein from human (PDB 1b3u:A), (2) oxidoreductase from *Neurospora crassa* involved in response to oxidative stress (PDB 1sy7:A), (3) clathrin adaptor protein core from mouse (PDB 1w63:A) involved in binding and intracellular protein transport, (4) importin β subunit from human (PDB 1qgr:A), which transfers proteins into nucleus, (5) β-catenin from human (PDB 1jdh:A) that functions in transcription process, (6) TIP20 protein from human (TATA binding protein that enhances transcription, part of multisubunit cullin-dependent ubiquitin ligase), which is involved in protein ubiquitination, negative regulation of catalytic activity and positive regulation of transcriptional complex assembly (PDB 1u6g:C), (7) apoptosis regulatory complex ced-4/ced-9 from nematode (PDB 2a5y) and (8) apoptotic protease activating factor from human (PDB 1z6t:A). Interesting residues identified as either coevolving between Prf and Pto/Fen or under natural selection are highlighted on these protein structures using the program PyMOL (DELANO 2008). Amino acid positions are numbered according to a reference protein sequence from *S. pimpinellifolium* (Pto, GenBank accession AAF76306) and *S. lycopersicum* (Fen, GenBank accession AAF76314; Prf, GenBank accession AAF76312).

CHAPTER 3

RESULTS

1. Nucleotide diversity in five genes from the Pto signaling pathway

I describe here sequence variation and level of evolutionary constraint for three R-genes in the *Pto* resistance gene cluster (*Pto*, *Fen* and *Prf*) and two candidate genes involved in the Pto signaling pathway (*Pfi* and *RIN4*). The Tarapaca population belongs to a TGRC core collection, which was carefully selected using multiple criteria to represent as much *S. peruvianum* diversity as possible (tgrc.ucdavis.edu; GORDILLO *et al.* 2008). Therefore this population was used to study level of polymorphism in the five genes of Pto signaling pathway. For three R-genes and two downstream candidate genes, 20 alleles were amplified and sequenced from this population.

Total polymorphism in the Tarapaca population in these five genes, as quantified by average pairwise differences across all sites π, ranged from 0.6% (*Prf*) to 1.6% (*Pfi*) (Table 2). For comparison, the mean π across the set of 14 reference genes for this same population is 1.3%. *Pto* and *Pfi* showed the highest polymorphism at synonymous sites (2.0% and 2.2%, respectively), as well as at non-synonymous sites (1.3% at both loci). The ratio of π_{non} to π_{syn} was 0.63 and 0.57 for *Pto* and *Pfi*, respectively, while this ratio was consistently much lower at the 14 reference loci (mean π_{non} to π_{syn} = 0.10). I used neutral coalescent simulations to test if the value of π observed at non-synonymous and synonymous sites fell within the 95% confidence interval of simulations in which θ was estimated from the average π across 14 reference genes from these same individuals (HUDSON 1990). These coalescent simulations indicated that both *Pto* and *Pfi* show excess variation, specifically at non-synonymous sites, while at synonymous sites the observed level of variation at *Pto* and *Pfi* is within the 95% confidence interval based on θ across these 14 other genes (Table 3). A significant departure from neutrality at *Pfi* is also captured in the MK test (Table 4). According to this test, *Pfi* displays significantly more variation at non-synonymous positions than expected under neutrality. A closer inspection of the distribution of variation across this large gene reveals that the NLS and the hydrolase-like region harbor substantial amounts of non-synonymous variation (π_{non} = 2.16%). In contrast, non-synonymous variation for the remainder of the gene is 0.423% (Figure 2).

Fen and *Prf* show lower levels of polymorphism among these five loci and intermediate values for the ratio of π_{non} versus π_{syn}. *Fen*, like *Pto*, is a small gene (963 nucleotides) and encodes a 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). 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 5' portion of the protein, which binds to other proteins including Pto and Fen.

The gene showing the greatest level of evolutionary constraint is *RIN4* homolog. This gene has the lowest level of non-synonymous polymorphism and the lowest ratio of π_{non} to π_{syn} of these five genes (Table 2). 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 very strong at this locus (Figure 4). 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 (Figure 5). Collectively, these groups containing identical sequence types show multiple fixed differences with respect to the other alleles. In particular the group of alleles: T7232A1, T7233A1 and T7240A1, shows nine fixed differences relative to the other alleles. Considering all non-singleton polymorphisms, these nine positions are in significant LD and, relative to the allele from the outgroup species *S. habrochaites*, these nine sites are derived in alleles T7232A1, T7233A1 and

T7240A1. Seven of nine of these changes are derived relative to the more distantly related outgroup*,* *S. lycopersicoides* (Figure 6). These nine fixed differences are distributed throughout the *RIN4* coding sequence. Two of these fixed, derived differences are non-synonymous, 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 LD involving derived changes, two of which are non-synonymous, and the low to moderate frequency of this sequence type, is consistent with the presence of partial or ongoing sweep at *RIN4*-like gene in *S. peruvianum*.

The three R-genes *Pto*, *Fen* and *Prf* are well known molecules in the Pto signaling pathway and physical interaction between Pto/Fen kinase and Prf was extensively studied (MUCYN *et al.* 2006, 2009; CHEN *et al.* 2008). To analyze the coevolutionary relationship between these molecules, I sequenced in total 44 alleles of *Pto*, *Fen* and *Prf* from 22 individuals across three populations of *S. peruvianum*. One allele of *Pto* from individual 7232 and one allele of *Fen* from individual 7236 appeared to be pseudogenes, based on the presence of frameshift mutations and pre-mature stop codons. Thus, these alleles were excluded from further analysis.

Average sequence polymorphism across three populations at synonymous sites across three R-genes from the *Pto* cluster is half that observed at the eight reference loci from these same individuals (1.56% at *Pto, Fen* and *Prf* versus 2.95% at the reference loci; Table 5). In contrast, non-synonymous polymorphism is more than three and half times higher at the resistance gene loci as compared to the eight reference loci (1.04% at *Pto*, *Fen* and *Prf* versus 0.29% at the reference loci). As a result, the ratio of non-synonymous to synonymous polymorphism is more than six times higher for the resistance genes compared to the reference loci. The sequence variation of R-gene *Pto* and the functional consequences of this variation within and between populations of seven tomato species were previously characterized (ROSE *et al.* 2005, 2007). Those studies also reported a significantly higher level of non-synonymous variation at *Pto* in *S. peruvianum* compared to a set of reference genes. Evidence for elevated levels of amino acid polymorphism is consistent with balancing selection at this locus. Here I observe similar patterns at two additional genes found in the same resistance gene cluster as *Pto*.

TABLE 2. Summary of polymorphism π (NEI 1987) across five genes from the Pto signaling pathway in the Tarapaca population of *S. peruvianum.*

Gene	Total sites	π_{total}	π_{syn}	π_{non}	$\pi_{\text{non}} / \pi_{\text{syn}}$
Pto	960	0.01450	0.02038	0.01278	0.63
Fen	963	0.00871	0.01560	0.00676	0.43
Prf	5541	0.00667	0.01386	0.00448	0.32
Pfi	5556	0.01662	0.02233	0.01277	0.57
RIN4	1176	0.00924	0.01984	0.00320	0.16

TABLE 3. Results of coalescent simulations for *Pto* and *Pfi.*

	π at reference genes	π at Pto	p-value (π exp > π obs)[c]	π at Pfi	p-value (π exp > π obs)[d]
syn[a]	0.023	0.02	0.39	0.022	0.413
non[b]	0.0024	0.013	0.00**	0.013	0.00**

a – Arithmetic mean of π at synonymous sites from 14 reference genes

b – Arithmetic mean of π at non-synonymous sites from 14 reference genes.

c – Probability of observing a value of π greater than that observed at *Pto* in 1,000 coalescent simulations, conditioned on the π values of the reference gene set.

d – Probability of observing a value of π greater than that observed at *Pfi* in 1,000 coalescent simulations, conditioned on the π values of the reference gene set.

TABLE 4. MK test (MCDONALD and KREITMAN 1991) using nucleotide variation at *Pfi.*

	Fixed differences	Polymorphisms
Silent	131	252
Replacement	18	90

p-value = 0.00026, based on a G-test of independence, *Solanum lycopersicoides* as outgroup.

TABLE 5. Summary statistics for R-genes *Pto*, *Fen, Prf* and eight reference genes within and across three populations of *S. peruvianum.*

Locus	Population	Length[a]	n[b]	S[c]	Hd[d]	π_{syn}[e]	π_{non}[f]	$\frac{\pi_{non}}{\pi_{syn}}$	*D*[g]	Z_{nS}[h]	ρ[i]
Pto	Total	960	43	68	0.982	1.808	1.437	0.79	-0.262	0.097	0.070
	Tarapaca		19	43	0.912	2.038	1.278	0.63	0.350	0.235	0.027
	Nazca		12	50	1.000	1.971	1.784	0.91	0.257	0.255	0.035
	Canta		12	34	0.985	1.021	1.124	1.10	-0.293	0.195	0.046
Fen	Total	957	43	76	0.984	2.018	0.919	0.46	-1.320	0.081	0.060
	Tarapaca		19	34	0.942	1.560	0.676	0.43	-0.476	0.173	0.031
	Nazca		12	25	1.000	1.558	0.506	0.32	-0.631	0.178	0.047
	Canta		12	53	0.955	2.277	1.545	0.68	-0.327	0.345	0.027
Prf	Total	1701	44	75	0.979	0.842	0.764	0.91	-0.818	0.062	0.017
	Tarapaca		20	49	0.947	1.114	0.705	0.63	-0.071	0.132	0.011
	Nazca		12	32	1.000	0.421	0.702	1.67	0.114	0.194	0.021
	Canta		12	26	0.848	0.331	0.487	1.47	-0.481	0.410	0.002
CT066	Total	1346	34	66	0.966	3.392	0.217	0.06	-0.616	0.091	0.024
	Tarapaca		10	40	0.933	3.369	0.204	0.06	-0.307	0.281	0.010
	Nazca		12	25	0.773	2.431	0.141	0.06	0.650	0.453	0.001
	Canta		12	43	0.985	2.880	0.177	0.06	-0.930	0.145	0.043
CT093	Total	1389	34	60	0.991	1.473	0.195	0.13	-1.708	0.105	0.012
	Tarapaca		10	23	0.956	1.763	0.105	0.06	-0.141	0.242	0.013
	Nazca		12	24	0.955	1.208	0.169	0.14	-0.598	0.371	0.004
	Canta		12	31	1.000	1.111	0.241	0.22	-1.001	0.187	0.023
CT166	Total	1265	32	114	0.986	0.894	0.069	0.08	-1.622	0.097	0.013
	Tarapaca		8	42	0.893	0.548	0.000	0.00	-0.514	0.475	0.002
	Nazca		12	45	0.970	1.019	0.000	0.00	-1.067	0.238	0.014
	Canta		12	75	0.970	1.131	0.164	0.15	-0.753	0.264	0.005
CT179	Total	899	34	91	0.991	4.355	0.082	0.02	-1.112	0.057	0.097
	Tarapaca		10	29	0.911	3.456	0.000	0.00	-0.003	0.284	0.011
	Nazca		12	49	1.000	3.751	0.117	0.03	-0.368	0.137	0.102
	Canta		12	56	0.985	4.604	0.117	0.03	-0.400	0.156	0.055
CT198	Total	693	30	101	0.986	5.439	0.364	0.07	-0.732	0.088	0.060
	Tarapaca		10	62	0.911	5.648	0.182	0.03	0.070	0.342	0.010
	Nazca		10	50	0.978	4.312	0.364	0.08	-0.050	0.223	0.041
	Canta		10	57	0.978	5.891	0.519	0.09	-0.050	0.294	0.015
CT208	Total	1069	32	83	0.938	1.720	0.018	0.01	-0.951	0.161	0.002
	Tarapaca		8	41	0.893	0.993	0.074	0.07	-0.227	0.745	0.000
	Nazca		12	47	0.803	1.419	0.000	0.00	-0.540	0.408	0.000
	Canta		12	43	0.773	1.784	0.000	0.00	-0.509	0.552	0.000
CT251	Total	1672	32	127	0.990	2.978	0.811	0.27	-0.877	0.092	0.031
	Tarapaca		10	70	0.933	3.443	0.721	0.21	-0.140	0.383	0.005
	Nazca		12	55	0.970	2.198	0.719	0.33	0.280	0.227	0.015
	Canta		10	70	1.000	2.598	0.643	0.25	-0.350	0.197	0.029
CT268	Total	1881	34	128	1.000	3.360	0.569	0.17	-1.019	0.054	0.080
	Tarapaca		10	56	1.000	2.586	0.446	0.17	-0.510	0.246	0.031
	Nazca		12	70	1.000	3.451	0.615	0.18	0.150	0.173	0.042
	Canta		12	68	1.000	3.061	0.516	0.17	-0.350	0.119	0.077

Total = pooled sample, treated as a single population; a – excluding indels; b – number of alleles analyzed; c – segregating sites; d – haplotype diversity; e – percent, nucleotide diversity per synonymous site; f – percent, nucleotide diversity per non-synonymous site; g – Tajima's *D* for all sites (TAJIMA 1989); h – intralocus linkage disequilibrium, average of R^2 across all pairwise comparisons of polymorphic sites (KELLY 1997); i – population recombination rate per site (HUDSON 2001).

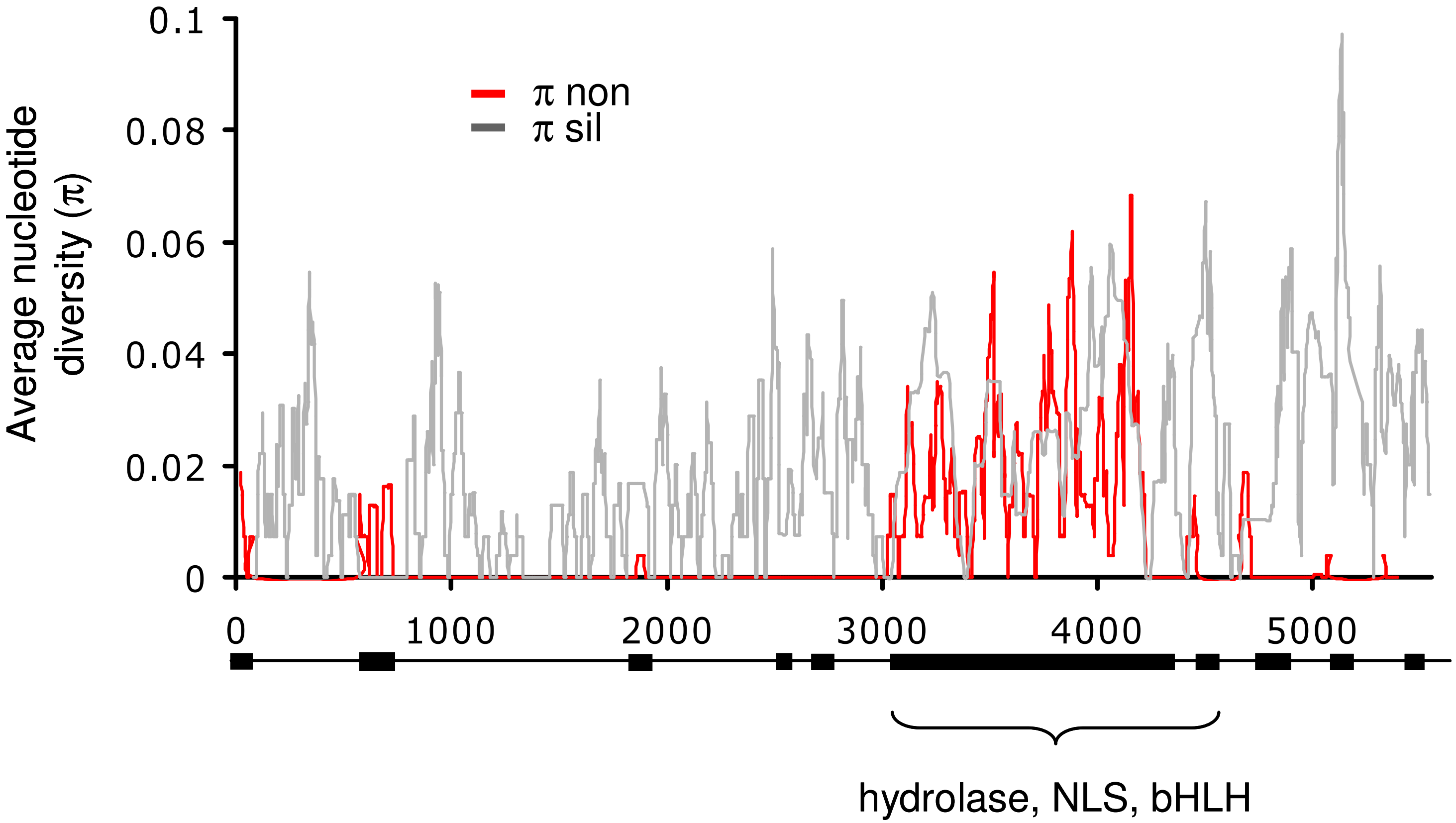


FIGURE 2. Sliding window plot of nucleotide diversity (π) for *Pfi* in the Tarapaca population of *S. peruvianum*. Values are midpoints of 30 bp windows. The gene structure is located below the graph. Boxes indicate exons, solid lines indicate introns. The important regions are indicated below the appropriate exons: a putative hydrolase motif, NLS – nuclear localization signal, bHLH – basic helix-loop-helix-like DNA binding domain.

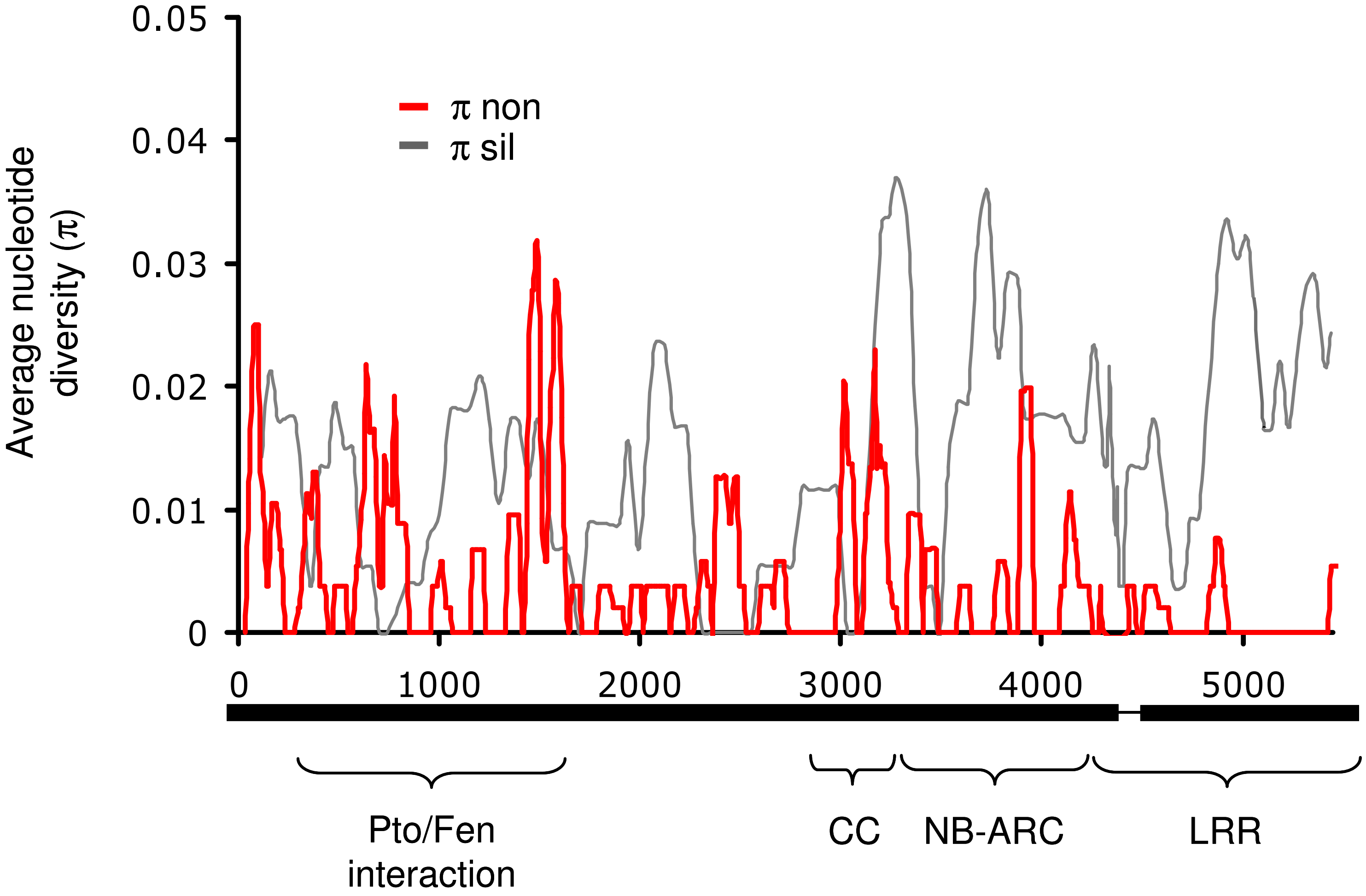


FIGURE 3. Sliding window plot of nucleotide diversity for *Prf* in the Tarapaca population of *S. peruvianum*. Values are midpoints of 50 bp windows. The gene structure is located below the graph. Boxes indicate exons, solid line indicates an intron. The functional regions are indicated below the appropriate exons (see also Figure 14).

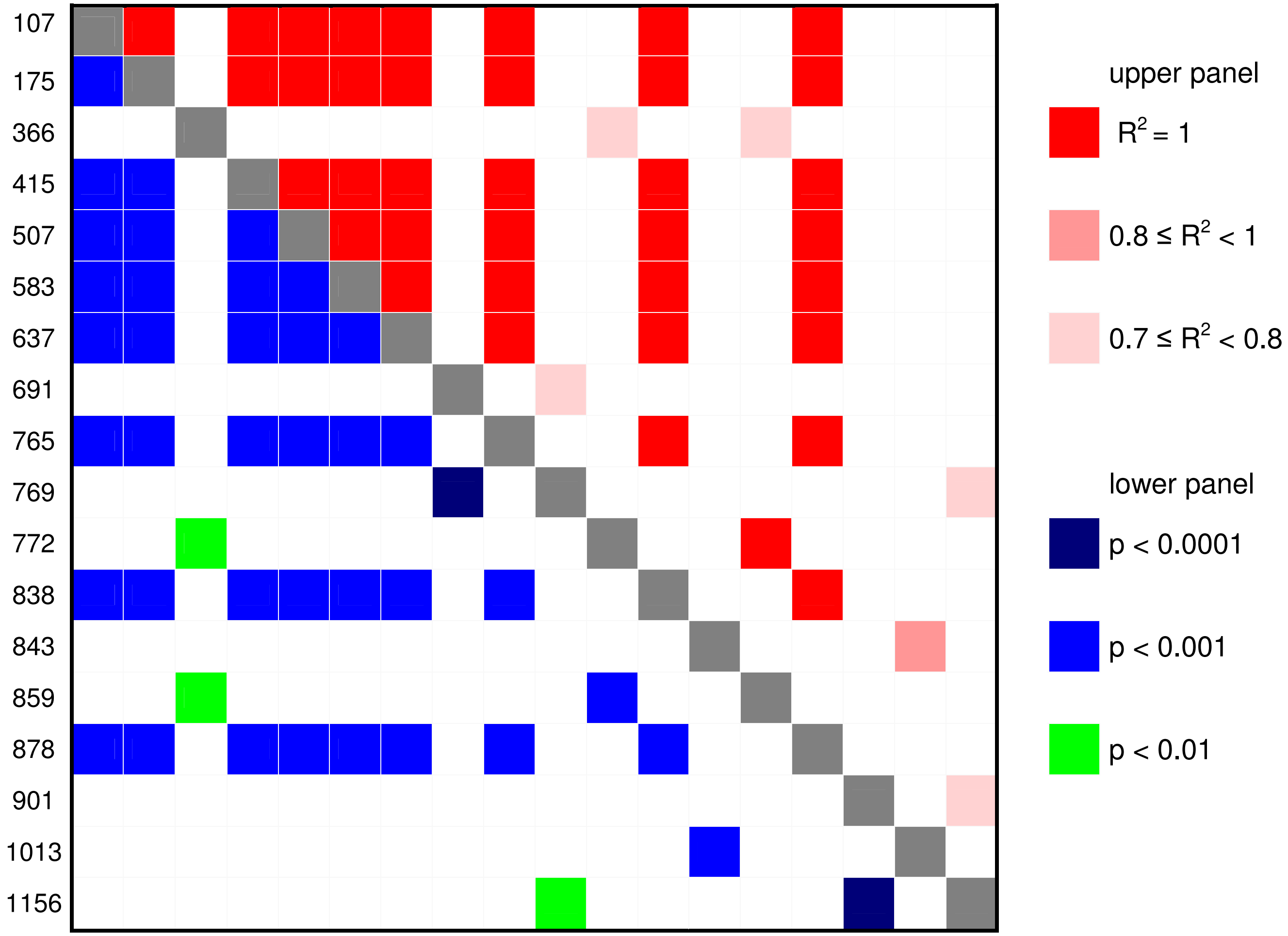


FIGURE 4. Significant linkage disequilibrium between polymorphic sites in *RIN4*-like gene. Above the diagonal: R^2 (measure of LD), below the diagonal: p-values after multiple test correction.

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					
T7232A1	A	C	T	T	G	G	A	G	A	G	T	A	T	G	T	T	C	C	A	A	T	C	C	C	A	A	C	T	T
T7233A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
T7240A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
T7234A1	C	.	G	.	.	T	G	.	G	A	G	G	C	.	A	C	T	.	.	G	C	.	.	A	T	.	A	C	.
T7236A1	C	.	G	.	.	T	G	.	G	A	G	G	C	.	A	C	T	.	.	G	C	.	.	A	T	.	A	C	.
T7238A1	C	.	G	.	.	T	G	.	G	A	G	G	C	.	A	C	T	.	.	G	C	.	.	A	T	.	A	C	.
T7236A2	C	.	G	.	.	T	G	A	G	A	G	G	C	.	.	C	.	A	.	G	.	.	T	A	.	.	.	.	.
T7237A1	C	.	G	.	.	T	G	A	G	A	G	G	C	.	.	C	.	A	.	G	.	.	T	A	.	.	.	.	.
T7241A1	C	.	G	.	.	T	G	A	G	A	G	G	C	.	.	C	.	A	.	G	.	.	T	A	.	.	.	.	.
T7232A2	.	T	G	A	.	T	G	.	G	.	G	G	C	T	.	C	.	.	G	G	C	.	.	A	.	.	A	.	.
T7233A2	.	.	G	.	.	T	.	.	G	A	G	G	C	.	.	C	.	.	.	G	C	.	.	A	.	.	A	.	.
T7234A2	.	.	G	.	.	T	.	A	G	A	G	G	C	.	.	C	.	.	.	G	C	T	.	A	G	G	A	.	.
T7235A1	.	.	G	.	A	T	.	.	G	.	G	G	C	.	.	C	.	.	.	G	C	.	.	A	.	.	A	.	C
T7235A2	.	.	G	A	.	T	G	.	G	.	G	G	C	T	.	C	.	.	G	G	C	.	.	A	.	.	A	.	.
T7237A2	.	T	G	A	.	T	G	.	G	.	G	G	C	.	.	C	.	.	.	G	C	T	.	A	.	.	A	.	.
T7238A2	C	.	G	.	A	T	.	.	G	.	G	G	C	.	.	C	.	.	.	G	C	.	.	A	.	.	A	.	C
T7239A1	.	.	G	.	.	T	.	.	G	.	G	G	C	T	.	C	.	.	.	G	C	.	.	A	.	G	A	.	.
T7239A2	.	.	G	.	.	T	G	.	G	A	G	G	C	.	C	C	.	.	.	G	C	.	.	A	.	.	A	.	.
T7240A2	.	.	G	.	.	T	G	.	G	.	G	G	C	.	.	C	.	.	.	G	.	.	.	A	G	.	A	C	C
T7241A2	C	.	G	.	.	T	.	.	G	A	G	G	C	.	.	C	.	.	.	G	C	.	.	A	.	.	A	.	.
S. habrochaites	.	.	G	.	.	T	.	.	G	A	G	G	C	.	.	C	.	.	.	G	.	.	.	A	.	.	A	.	.
S. lycopersicoides	.	.	G	.	.	T	.	.	G	A	G	G	.	.	.	.	.	.	.	G	C	.	.	A	.	.	A	.	.

FIGURE 5. Segregating sites across *RIN4*-like gene for the Tarapaca population of *S. peruvianum*. Dots indicate positions matching the reference allele T7232A1. Positions showing statistically significant LD (see Figure 4) are indicated along the top row with an asterisk. The type of mutation (i.e., synonymous or non-synonymous) of these positions is indicated in the third row. The lower rows contain the nucleotide states of outgroups of *S. peruvianum* at these same positions.

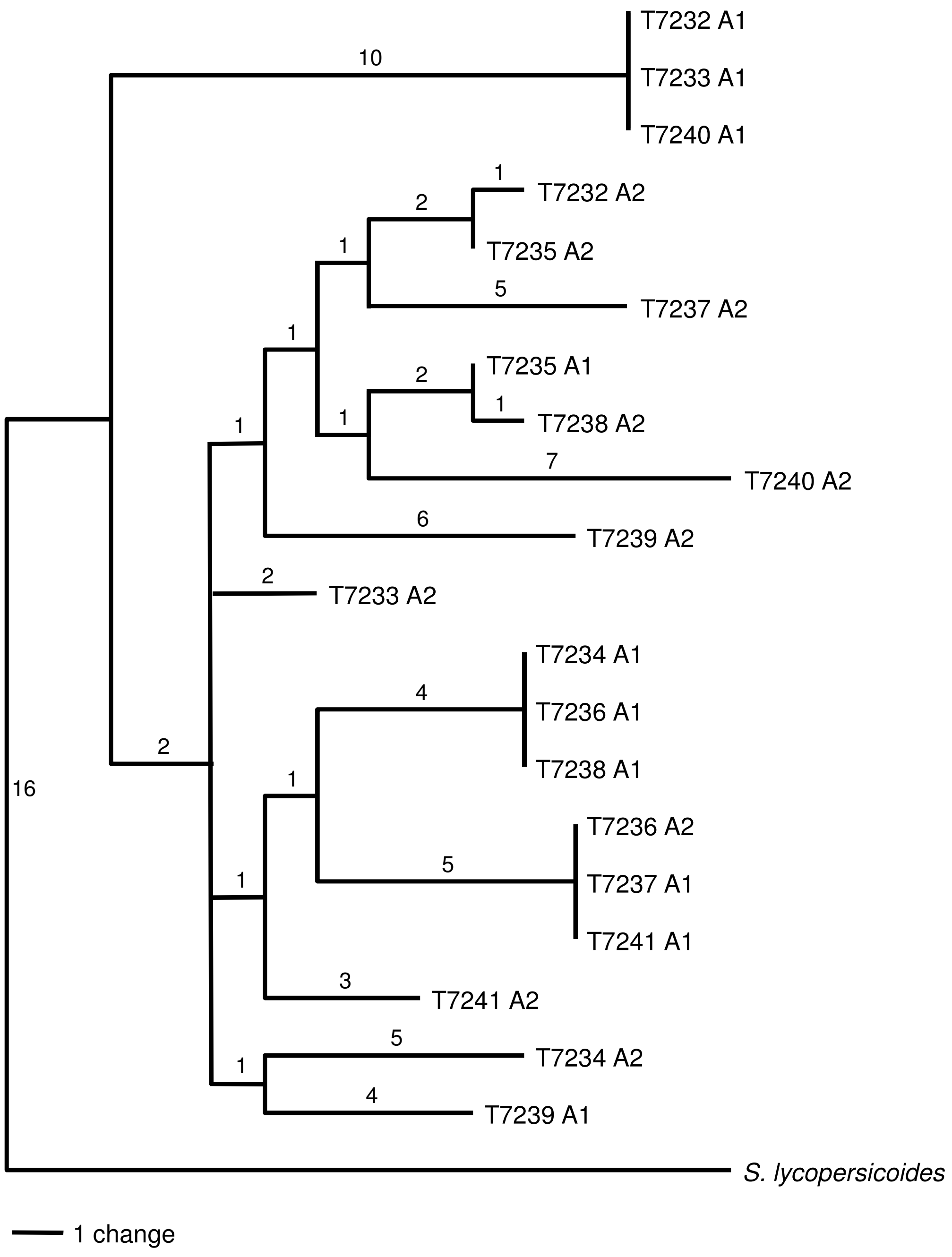


FIGURE 6. One of 10 equally most parsimonious trees of the *RIN4* nucleotide sequences in the Tarapaca population of *S. peruvianum*. The tree was rooted with the allele of *RIN4* from the outgroup species *S. lycopersicoides*. Branch lengths (number of steps) are indicated above the branches.

2. Population differentiation in Pto, Fen and Prf

The level of genetic differentiation between populations can be influenced by both demographic history and natural selection. Large differences in the amount of population differentiation between loci can point to individual loci that have been the targets of selection. I compared the levels of genetic differentiation between these three resistance genes and eight reference genes. F_{ST} ranged from 0.08 at *Pto* to 0.22 at *Prf*. These values were within the range of variation I observed at other loci from these same individuals (Table 6). Therefore, on an individual gene basis, I did not detect deviations among these genes in the degree of population differentiation.

Recent methods have been developed to evaluate whether individual nucleotide positions within a gene show greater or lesser differentiation than expected based on population differentiation at an independent set of reference loci. Using these methods, I identified candidates for either balancing selection or directional selection in the R-genes (Figures 7 and 8). In general, given the amount of population differentiation estimated from the reference loci and used for generating the 95% confidence intervals for FDIST2 test of BEAUMONT and NICHOLS (1996) (mean "neutral" F_{ST} = 0.156), I had limited power to detect selection at these positions. Also, using the Bayesian method of FOLL and GAGGIOTTI (2008) with my data set, I obtained low Bayes factor values. Thus, I considered results as significant for all BF > 1.8, which is in agreement with results from the FDIST2 test. This corresponds to the posterior p-values between 0.64 and 0.8. According to the scale of evidence described by JEFFREYS (1961), it represents weak to substantial evidence for the model assuming that the SNP loci are subject to selection. Collectively, using both methods, I detected two replacement sites experiencing directional selection and six candidates for balancing selection. What is known about these candidates in terms of protein biochemistry and function is discussed below.

These three resistance genes behave similarly to one another based on their patterns of nucleotide diversity and population differentiation. I was particularly interested how physical linkage and/or epistatic selection may have affected the evolutionary history of these genes. Pto and Fen both form complexes with Prf and are physically linked to Prf. Genome sequencing of this region in *S. pimpinellifolium* indicates that the *Fen* gene is only 2 kb from the coding region of *Prf*, while *Pto* is over 25 kb away. Recombination rates, which determine how quickly linkage associations break down, vary substantially across the tomato genome (see Table 5). The weighted average estimate of ρ across the set of reference loci in *S. peruvianum* is 0.0234 and LD decays rapidly in this outcrossing species (ARUNYAWAT *et al.* 2007). *Pto* and *Fen* show relatively high levels of recombination (ρ = 0.07 and ρ = 0.06, respectively), while *Prf* shows more moderate amount of recombination (ρ = 0.017). LD decays quite rapidly in these genes and the expectation of R^2 drops below 0.05 in less than 0.4 kb (Figure 9).

TABLE 6. Population differentiation F_{ST} (HUDSON *et al.* 1992) across *Pto*, *Fen*, *Prf* and eight reference loci in *S. peruvianum.*

Locus	F_{ST}	Locus	F_{ST}
Pto	0.08088	CT066	0.21720
Fen	0.11376	CT093	0.11138
Prf	0.21729	CT166	0.06221
		CT179	0.16311
		CT198	0.08318
		CT208	0.21463
		CT251	0.13948
		CT268	0.11416

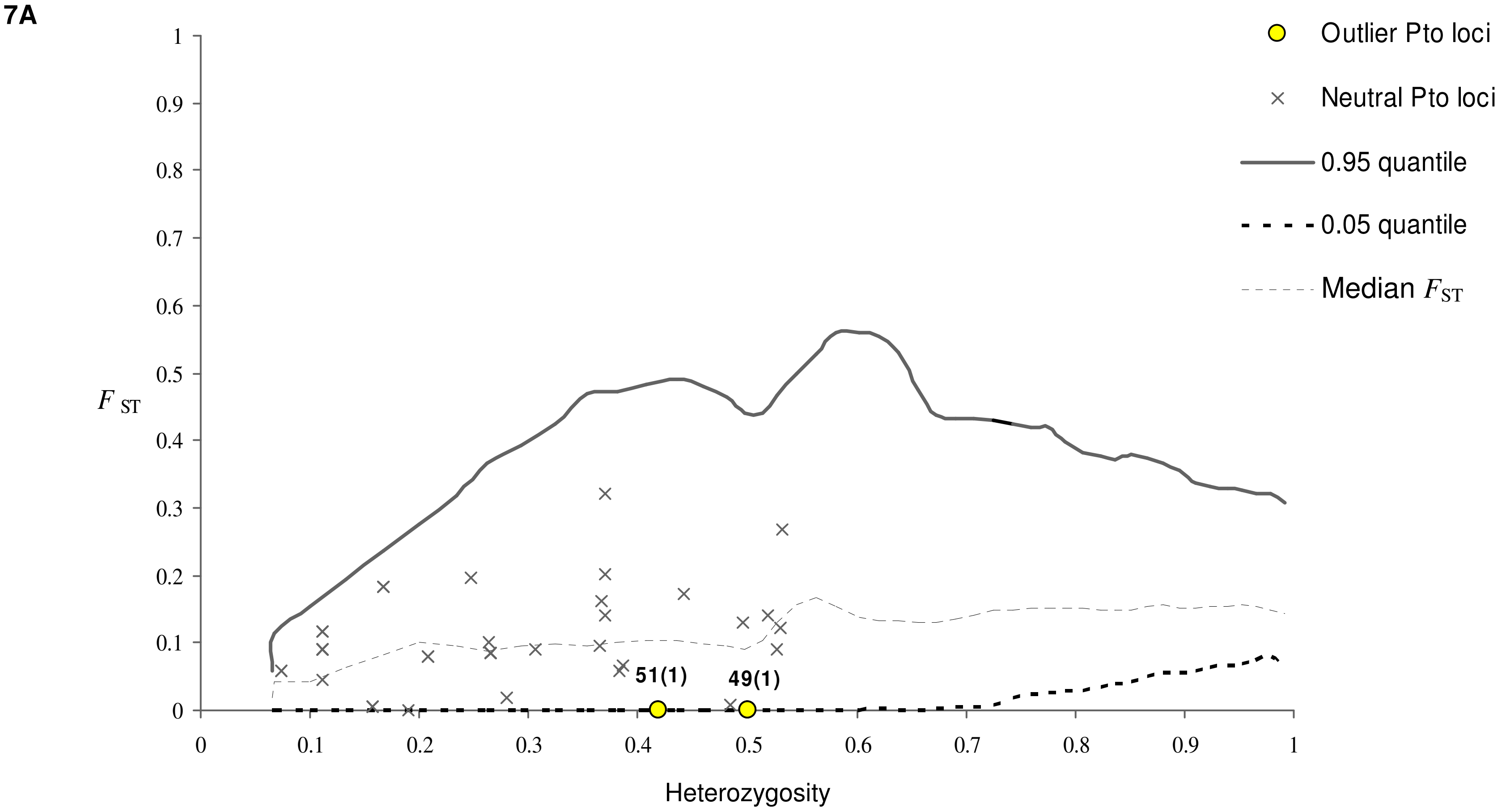
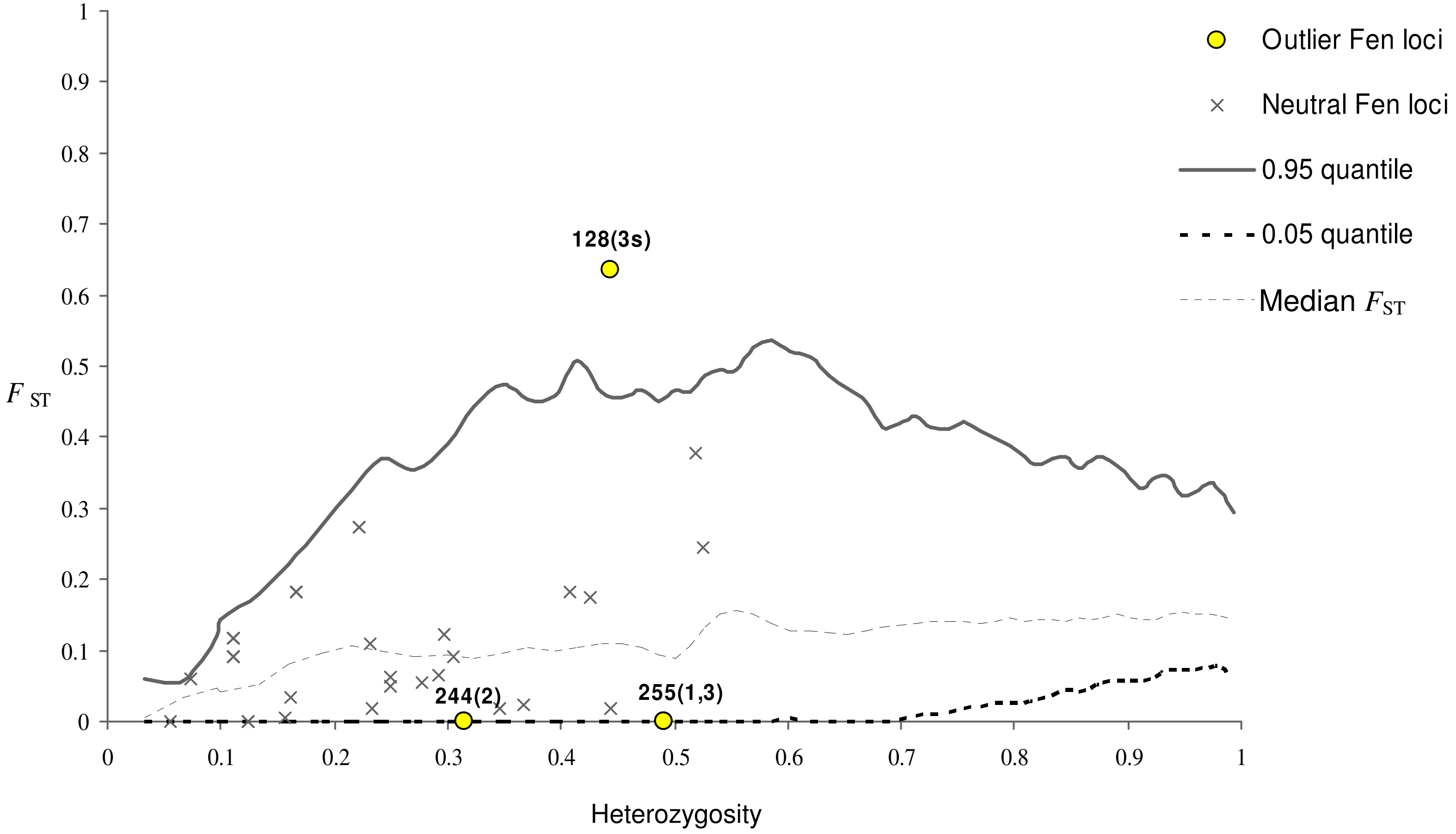


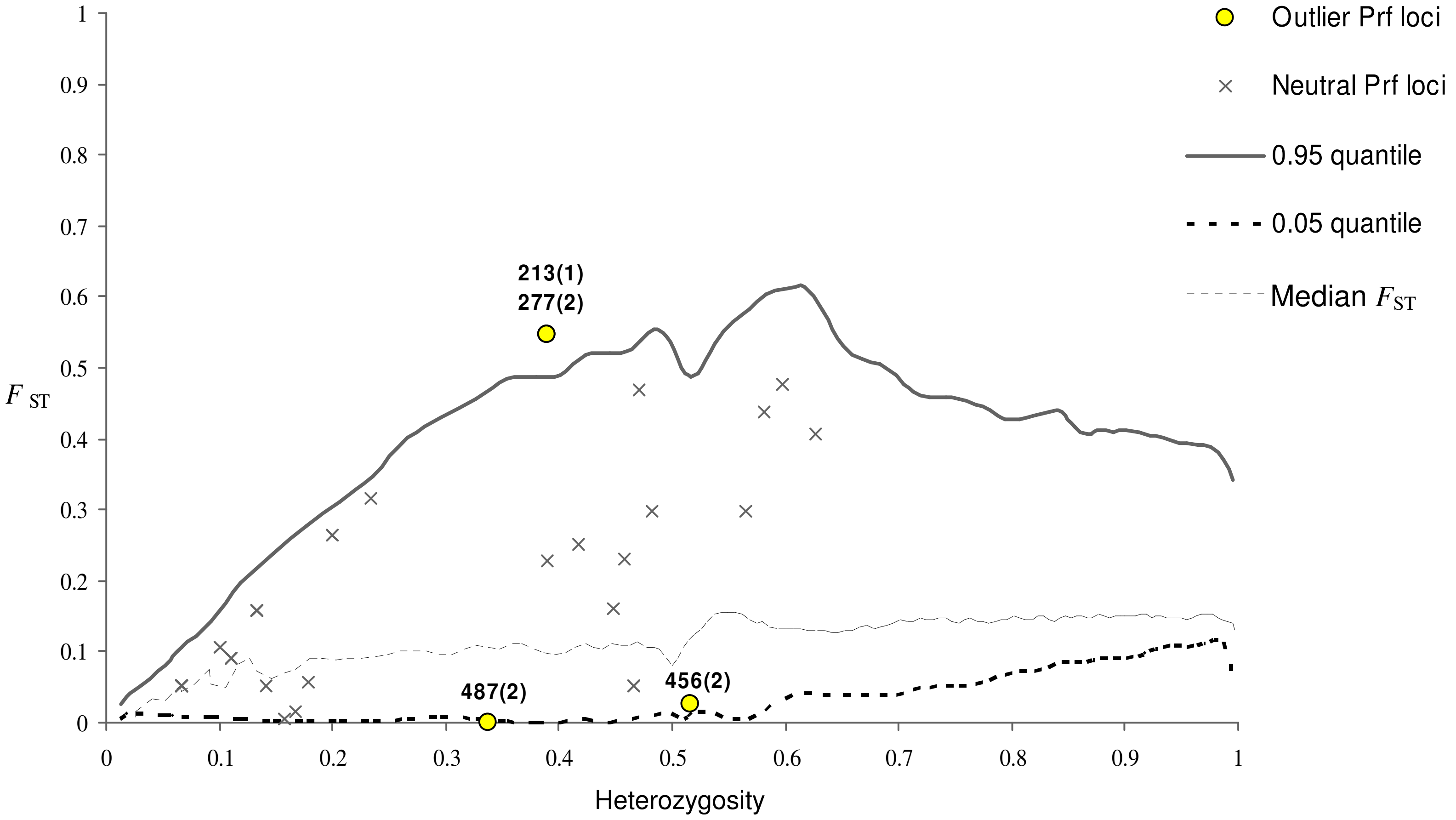
FIGURE 7. Outlier SNP loci in (A) Pto, (B) Fen, (C) Prf, based on the method of BEAUMONT and NICHOLS (1996). Each data point is a SNP locus. Loci with an F_{ST} value in 95% confidence interval were considered to be outlier loci (below 0.05 quantile – candidates for balancing selection, above 0.95 quantile – candidates for directional selection/local adaptation). Numbers indicate encoded amino acid and base position in codon (in parentheses; s – synonymous site).

7B



Outlier Fen loci
Neutral Fen loci
0.95 quantile
0.05 quantile
Median F_{ST}
F_{ST}
Heterozygosity
128(3s)
244(2)
255(1,3)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

7C



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
F_{ST}
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Heterozygosity
213(1)
277(2)
487(2)
456(2)
● Outlier Prf loci
× Neutral Prf loci
0.95 quantile
0.05 quantile
Median F_{ST}

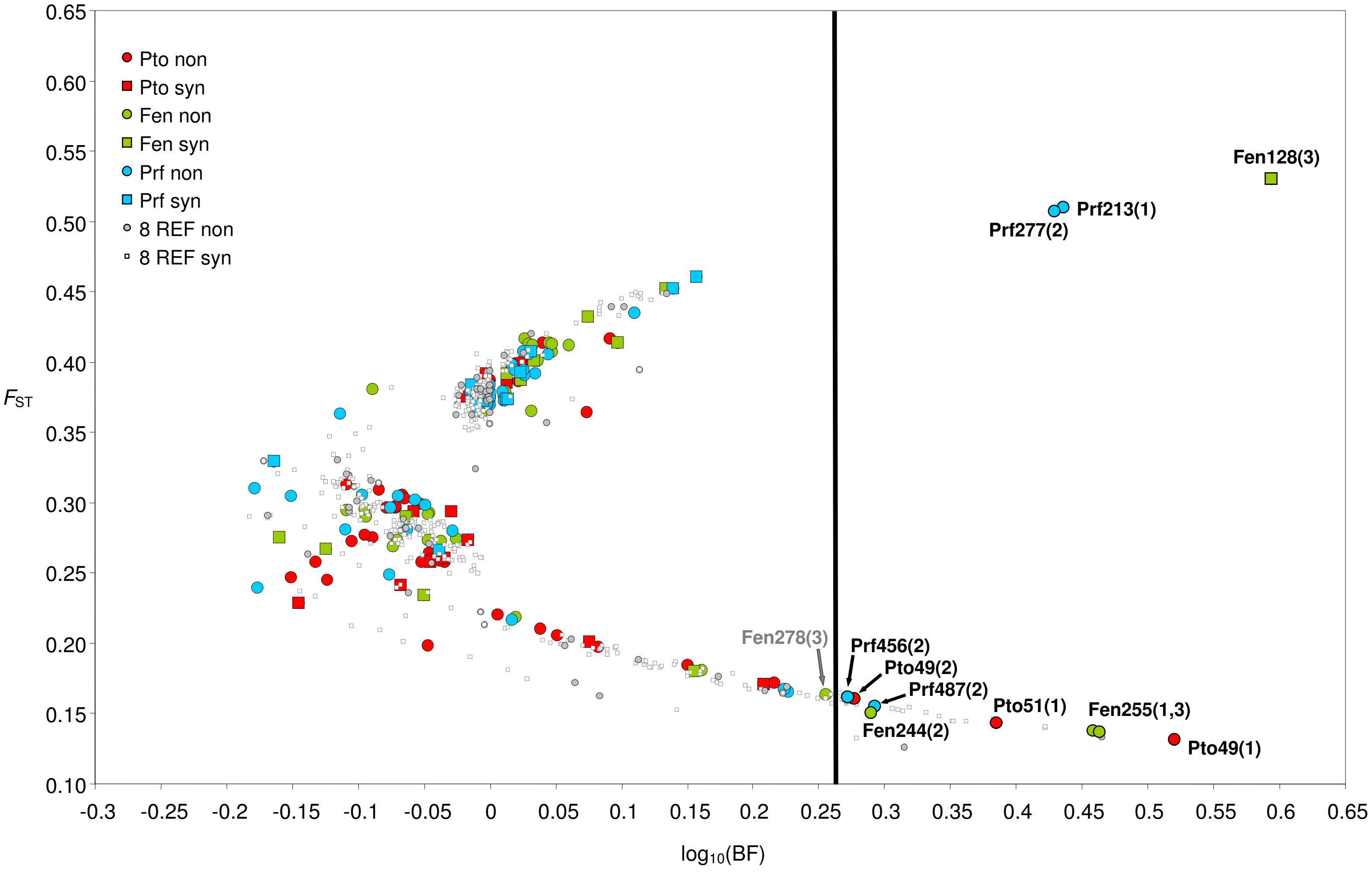


FIGURE 8. Outlier SNP loci from *Pto*, *Fen* and *Prf* based on method of FOLL and GAGGIOTTI (2008); syn – synonymous sites, non – non-synonymous sites, 8 REF – eight reference genes. Numbers denote encoded amino acid and base position in codon (in parentheses). The line is a threshold indicating candidate sites consistent with results of method by BEAUMONT and NICHOLS (1996).

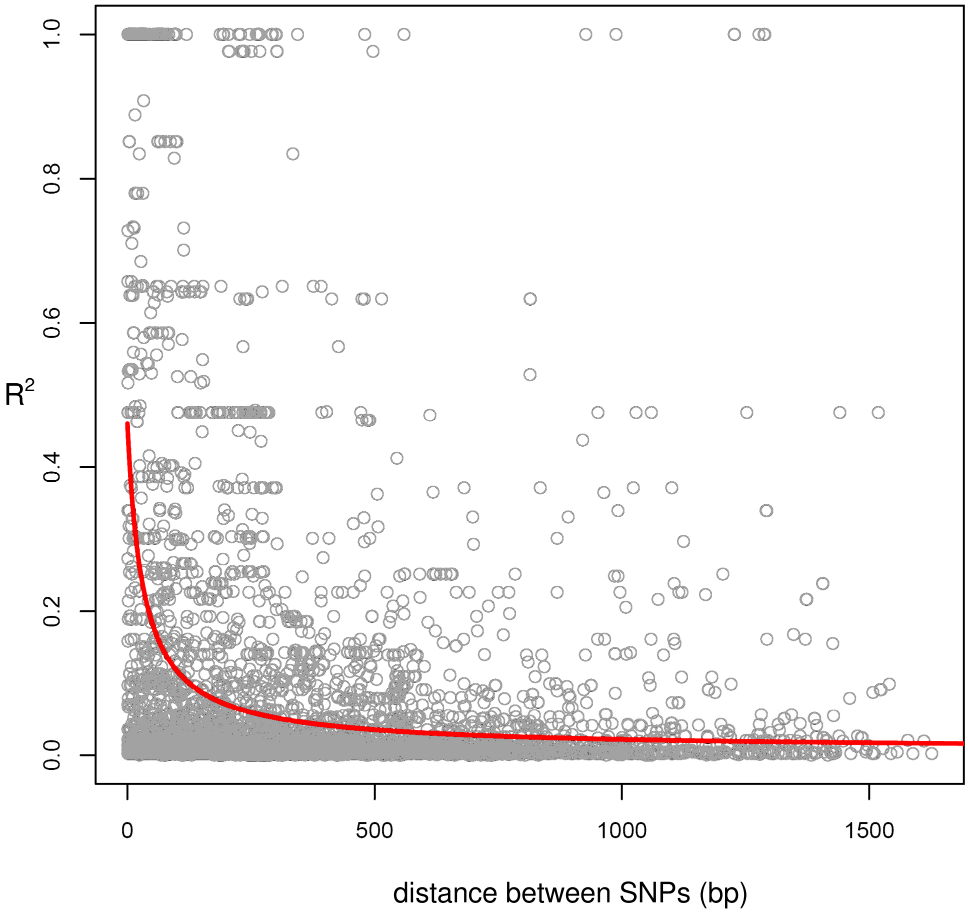


FIGURE 9. Decay of linkage disequilibrium R^2 as a function of distance between pairs of polymorphic sites in *Pto*, *Fen* and *Prf*. The red line depicts the expected decline of LD against distance based on the equation given by HILL and WEIR (1988).

3. Linkage disequilibrium between Pto/Fen and Prf

Since LD decays on average relatively rapidly within these three R-genes, associations through chromosomal linkage may only play a minor role in correlated evolutionary patterns between these genes. However, if natural selection favors particular combinations of alleles at these loci, epistatic selection may still contribute to correlated evolutionary histories. To test for coevolution between these genes, I estimated LD at pairs of polymorphic non-synonymous positions between genes. This estimate of LD is based on observed genotypes and does not require the data to be phased. I analyzed associations between loci for each population separately, since pooling alleles across populations could lead to spurious associations. In the Tarapaca population, I discovered 29 pairs of sites that were in LD between Pto and Prf (Table 7). For Fen, I discovered 14 pairs of sites in LD with Prf (Table 8). For this population, since I sequenced the entire *Prf* coding region, I could detect LD not only with the N-terminal region known to bind Pto and Fen, but also with other regions of Prf. Statistical significance of these associations was evaluated based on two kinds of p-values. Approximate p-values are derived from the composite disequilibrium coefficient using a chi-square approximation, while permutation based p-values correspond to the proportion of times the R^2 test statistic computed from randomly sampled data was found to be as extreme or more extreme than the statistical value of the original data. For the Tarapaca population, I report all pairs of sites for which these two p-values fell below the 0.05 level. For the other two populations, since fewer alleles were sampled, I had less power to detect associations. For these two populations, I report the LD between Pto/Fen and Prf that had approximate p-values lower than 0.05. For none of these pairs of sites, however, did the permutation based p-values fall below 0.05. Consequently, in Nazca I detected 18 pairs of sites that showed LD between Pto and Prf and 3 pairs of sites that showed LD between Fen and Prf (Table 9). In turn, in Canta, I detected 7 pairs of sites that showed LD between Pto and Prf and 17 pairs of sites that showed LD between Fen and Prf (Table 10).

TABLE 7. LD between non-synonymous polymorphisms in Pto and Prf within the Tarapaca population of *S. peruvianum* ($p < 0.05$). LD values are arranged by position in PTO. Numbers PTO and PRF indicate encoded amino acid and base position in codon (in parentheses).

	PTO	PRF	R^2	Approx. p-value	Permut. p-value
1.	49(2)	62(2)	0.725476	0.029523	0.04774
2.	49(2)	491(3)	0.725476	0.029523	0.04774
3.	49(2)	492(1)	0.725476	0.029523	0.04774
4.	49(2)	1002(2)	0.725476	0.029523	0.04774
5.	49(2)	1149(2)	0.725476	0.029523	0.04774
6.	88(2)	821(2)	0.811107	0.014961	0.01344
7.	135(1)	803(1)	0.944911	0.004586	0.01264
8.	135(2)	803(1)	0.944911	0.004586	0.01264
9.	154(2)	397(1)	0.737043	0.027027	0.02868
10.	154(2)	1013(1)	0.763158	0.022052	0.03242
11.	154(2)	1047(1)	0.763158	0.022052	0.03242
12.	154(2)	1066(2)	0.763158	0.022052	0.03242
13.	154(2)	1121(1)	0.763158	0.022052	0.03242
14.	168(1)	1013(1)	0.802955	0.016002	0.03498
15.	168(1)	1047(1)	0.802955	0.016002	0.03498
16.	168(1)	1066(2)	0.802955	0.016002	0.03498
17.	168(1)	1121(1)	0.802955	0.016002	0.03498
18.	168(2)	1013(1)	0.894737	0.007270	0.00984
19.	168(2)	1047(1)	0.894737	0.007270	0.00984
20.	168(2)	1066(2)	0.894737	0.007270	0.00984
21.	168(2)	1121(1)	0.894737	0.007270	0.00984
22.	200(1)	1013(1)	0.894737	0.007270	0.00984
23.	200(1)	1047(1)	0.894737	0.007270	0.00984
24.	200(1)	1066(2)	0.894737	0.007270	0.00984
25.	200(1)	1121(1)	0.894737	0.007270	0.00984
26.	232(3)	1013(1)	0.635851	0.026285	0.04706
27.	232(3)	1047(1)	0.635851	0.026285	0.04706
28.	232(3)	1066(2)	0.635851	0.026285	0.04706
29.	232(3)	1121(1)	0.635851	0.026285	0.04706

Table 8. LD between non-synonymous polymorphisms in Fen and Prf within the Tarapaca population of *S. peruvianum* ($p < 0.05$). LD values are arranged by position in FEN. Numbers FEN and PRF indicate encoded amino acid and base position in codon (in parentheses).

	FEN	PRF	R^2	Approx. p-value	Permut. p-value
1.	151(2)	62(2)	0.892218	0.007436	0.02778
2.	151(2)	491(3)	0.892218	0.007436	0.02778
3.	151(2)	492(1)	0.892218	0.007436	0.02778
4.	151(2)	1002(2)	0.892218	0.007436	0.02778
5.	151(2)	1149(2)	0.892218	0.007436	0.02778
6.	241(1)	803(1)	0.866025	0.009375	0.02400
7.	244(2)	803(1)	0.866025	0.009375	0.02400
8.	247(2)	803(1)	0.866025	0.009375	0.02400
9.	255(2)	803(1)	0.866025	0.009375	0.02400
10.	278(3)	62(2)	0.731727	0.028151	0.04818
11.	278(3)	491(3)	0.731727	0.028151	0.04818
12.	278(3)	492(1)	0.731727	0.028151	0.04818
13.	278(3)	1002(2)	0.731727	0.028151	0.04818
14.	278(3)	1149(2)	0.731727	0.028151	0.04818

TABLE 9. LD between non-synonymous polymorphisms in (A) Pto and Prf, (B) Fen and Prf within the Nazca population of *S. peruvianum* (approx. $p < 0.05$). LD values are arranged by position in PTO and FEN. Numbers PTO, FEN and PRF indicate encoded amino acid and base position in codon (in parentheses).

(A)

	PTO	PRF	R^2	Approx. p-value	Permut. p-value
1.	87(2)	252(2)	1	0.0143059	0.16390
2.	87(2)	385(2)	1	0.0143059	0.16390
3.	87(2)	451(1)	1	0.0143059	0.16390
4.	87(2)	453(2)	1	0.0143059	0.16390
5.	87(2)	203(1)	0.87831	0.0314437	0.16390
6.	127(2)	107(1)	1	0.0143059	0.16526
7.	154(2)	156(3)	1	0.0143059	0.06722
8.	154(2)	159(1)	1	0.0143059	0.06722
9.	154(2)	487(2)	1	0.0143059	0.06722
10.	158(3)	107(1)	1	0.0143059	0.16722
11.	197(1)	525(2)	1	0.0143059	0.06678
12.	197(1)	525(3)	1	0.0143059	0.06678
13.	205(1)	252(2)	1	0.0143059	0.16390
14.	205(1)	385(2)	1	0.0143059	0.16390
15.	205(1)	451(1)	1	0.0143059	0.16390
16.	205(1)	453(2)	1	0.0143059	0.16390
17.	205(1)	203(1)	0.87831	0.0314437	0.16390
18.	295(2)	344(1)	0.92582	0.0233422	0.06688

(B)

	FEN	PRF	R^2	Approx. p-value	Permut. p-value
1.	76(1)	165(3)	1	0.0143059	0.1651
2.	103(2)	120(2)	1	0.0143059	0.1661
3.	116(2)	165(3)	1	0.0143059	0.1651

TABLE 10. LD between non-synonymous polymorphisms in (A) Pto and Prf, (B) Fen and Prf within the Canta population of *S. peruvianum* (approx. $p < 0.05$). LD values are arranged by position in PTO and FEN. Numbers PTO, FEN and PRF indicate encoded amino acid and base position in codon (in parentheses).

(A)

	PTO	PRF	R^2	Approx. p-value	Permut. p-value
1.	51(1)	212(2)	0.774597	0.0273237	0.06700
2.	51(1)	487(2)	1	0.0143059	0.06700
3.	87(2)	368(1)	1	0.0143059	0.06686
4.	87(2)	456(2)	1	0.0143059	0.06686
5.	205(1)	233(1)	1	0.0143059	0.16738
6.	205(1)	397(2)	1	0.0143059	0.16738
7.	232(3)	451(1)	1	0.0024787	0.16528

(B)

	FEN	PRF	R^2	Approx. p-value	Permut. p-value
1.	72(3)	233(1)	0.87831	0.031444	0.16954
2.	72(3)	397(2)	0.87831	0.031444	0.16954
3.	73(2)	233(1)	0.87831	0.031444	0.16954
4.	73(2)	397(2)	0.87831	0.031444	0.16954
5.	74(3)	233(1)	0.87831	0.031444	0.16954
6.	74(3)	397(2)	0.87831	0.031444	0.16954
7.	76(1)	233(1)	0.87831	0.031444	0.16954
8.	76(2)	233(1)	0.87831	0.031444	0.16954
9.	76(1)	397(2)	0.87831	0.031444	0.16954
10.	76(2)	397(2)	0.87831	0.031444	0.16954
11.	78(1)	233(1)	0.87831	0.031444	0.16954
12.	78(1)	397(2)	0.87831	0.031444	0.16954
13.	103(2)	233(1)	0.87831	0.031444	0.16954
14.	103(2)	397(2)	0.87831	0.031444	0.16954
15.	153(3)	233(1)	0.87831	0.031444	0.16954
16.	153(3)	397(2)	0.87831	0.031444	0.16954
17.	215(1)	451(1)	1	0.014306	0.16706

4. Partitioning of LD variance components

I used a method developed by OHTA (1982a, b) to determine the relative contribution of epistatic selection to overall LD observed between these genes. I found that a significant portion of the LD observed between these genes could be attributed to neutral causes (genetic drift, population subdivision and limited migration; Table 11). Only a small fraction of the sites had a signature of what is considered unequal systematic disequilibrium. Unequal systematic disequilibrium can arise when epistasis is present in some, but not all subpopulations. Between Pto and Prf, 26 pairs of sites were identified for which LD was considered to be unequal, systematic (Table 12). Between Fen and Prf, 42 pairs of sites were identified for which LD was considered to be unequal, systematic (Table 13). Six Fen-Prf SNP pairs found to be candidates for epistasis/ systematic disequilibrium between Fen and Prf included synonymous sites or doubletons, therefore were not considered further.

TABLE 11. Summary of Ohta's LD coefficients between: (A) 3420 SNP pairs of Pto and Prf, (B) 2640 SNP pairs of Fen and Prf

(A)

Dual relationship	Number of Pto-Prf SNP pairs	Average values of Ohta's LD coefficients across SNP pairs				
		D_{IS}	D'_{IS}	D_{ST}	D'_{ST}	D_{IT}
1. $D_{IS} < D_{ST}$ and $D'_{IS} > D'_{ST}$	3309 (96,75%)	0.005081	0.222157	0.056112	0.004625	0.226781
2. $D_{IS} > D_{ST}$ and $D'_{IS} < D'_{ST}$	0	–	–	–	–	–
3. $D_{IS} > D_{ST}$ and $D'_{IS} > D'_{ST}$	111 (3,25%)	0.030078	0.073305	0.022024	0.015795	0.089097

(B)

Dual relationship	Number of Fen-Prf SNP pairs	Average values of Ohta's LD coefficients across SNP pairs				
		D_{IS}	D'_{IS}	D_{ST}	D'_{ST}	D_{IT}
1. $D_{IS} < D_{ST}$ and $D'_{IS} > D'_{ST}$	2543 (96,32%)	0.004655	0.232894	0.059046	0.00409	0.236982
2. $D_{IS} > D_{ST}$ and $D'_{IS} < D'_{ST}$	6 (0.23%)	0.066890	0.026460	0.012030	0.03128	0.057740
3. $D_{IS} > D_{ST}$ and $D'_{IS} > D'_{ST}$	91 (3,45%)	0.043789	0.107362	0.024747	0.017258	0.124619

Interpretation:
1. nonsystematic disequilibrium: restricted migration, genetic drift
2. systematic disequilibrium: epistatic selection
3. unequal systematic disequilibrium: partial epistatic selection

TABLE 12. Ohta's LD coefficients partially consistent with epistasis for pairs of Pto-Prf amino acid polymorphisms. LD values are arranged by position in PTO. Numbers PTO and PRF indicate encoded amino acid and base position in codon (in parentheses).

	PTO	PRF	D_{IS}	D'_{IS}	D_{ST}	D'_{ST}	D_{IT}
1.	43(1)	487(2)	0.03851	0.08881	0.03232	0.03049	0.11930
2.	46(3)	487(2)	0.03851	0.08881	0.03232	0.03049	0.11930
3.	49(1)	23(1)	0.04960	0.20635	0.04490	0.02666	0.23300
4.	49(1)	34(1)	0.04916	0.20862	0.04596	0.02629	0.23491
5.	49(1)	62(2)	0.02116	0.08088	0.01460	0.00742	0.08830
6.	49(1)	252(2)	0.07231	0.17271	0.03734	0.07281	0.24553
7.	49(1)	487(2)	0.01146	0.03855	0.00871	0.00527	0.04381
8.	49(2)	487(2)	0.04644	0.10469	0.02257	0.01041	0.11510
9.	49(3)	487(2)	0.03851	0.08881	0.03232	0.03049	0.11930
10.	49(1)	491(3)	0.02116	0.08088	0.01460	0.00742	0.08830
11.	49(1)	492(1)	0.02116	0.08088	0.01460	0.00742	0.08830
12.	51(1)	487(2)	0.02440	0.05858	0.00941	0.00116	0.05974
13.	51(2)	487(2)	0.03851	0.08881	0.03232	0.03049	0.11930
14.	51(3)	487(2)	0.03851	0.08881	0.03232	0.03049	0.11930
15.	70(3)	487(2)	0.03851	0.08881	0.03232	0.03049	0.11930
16.	71(1)	487(2)	0.03851	0.08881	0.03232	0.03049	0.11930
17.	72(1)	487(2)	0.04556	0.09448	0.03460	0.03982	0.13430
18.	88(2)	397(2)	0.05879	0.16251	0.05353	0.02970	0.19221
19.	88(2)	487(2)	0.08554	0.15835	0.04664	0.04073	0.19908
20.	124(3)	397(2)	0.01302	0.04460	0.01294	0.00249	0.04708
21.	135(1)	536(3)	0.02116	0.03855	0.01465	0.01041	0.04896
22.	154(2)	397(2)	0.08102	0.34392	0.05538	0.02378	0.36769
23.	154(2)	487(2)	0.05526	0.19766	0.04337	0.02740	0.22506
24.	205(1)	487(2)	0.08907	0.45540	0.08735	0.02892	0.48433
25.	273(1)	456(2)	0.05614	0.15684	0.04462	0.04447	0.20131
26.	295(2)	456(2)	0.07143	0.24263	0.05205	0.02240	0.26503

Table 13. Ohta's LD coefficients partially consistent with epistasis for pairs of Fen-Prf amino acid polymorphisms. LD values are arranged by position in FEN. Numbers FEN and PRF indicate encoded amino acid and base position in codon (in parentheses).

	FEN	PRF	D_{IS}	D'_{IS}	D_{ST}	D'_{ST}	D_{IT}
1.	76(1)	233(1)	0.09268	0.40514	0.06785	0.02531	0.43045
2.	76(2)	233(1)	0.08475	0.34845	0.07032	0.02865	0.37710
3.	76(1)	397(2)	0.06907	0.07710	0.03413	0.03971	0.11681
4.	76(2)	397(2)	0.06907	0.09751	0.04277	0.04218	0.13968
5.	76(1)	487(2)	0.04909	0.07067	0.02444	0.03067	0.10134
6.	76(2)	487(2)	0.04556	0.07634	0.02897	0.02575	0.10209
7.	76(1)	536(3)	0.05121	0.09977	0.03602	0.02452	0.12429
8.	76(2)	536(3)	0.05121	0.12018	0.04349	0.02575	0.14593
9.	151(2)	62(2)	0.07113	0.08768	0.04219	0.04447	0.13215
10.	151(2)	456(2)	0.03762	0.18745	0.03089	0.00907	0.19652
11.	151(2)	491(3)	0.07113	0.08768	0.04219	0.04447	0.13215
12.	151(2)	492(1)	0.07113	0.08768	0.04219	0.04447	0.13215
13.	153(3)	456(2)	0.03175	0.15797	0.03100	0.00227	0.16024
14.	153(3)	487(2)	0.02205	0.09297	0.01851	0.00594	0.09891
15.	244(2)	456(2)	0.03233	0.09448	0.01908	0.00907	0.10355
16.	244(2)	487(2)	0.01146	0.03704	0.01006	0.00742	0.04446
17.	255(1)	23(1)	0.05556	0.21202	0.04988	0.01671	0.22872
18.	255(3)	23(1)	0.05556	0.21202	0.04988	0.01671	0.22872
19.	255(1)	34(1)	0.05379	0.21429	0.05026	0.01763	0.23192
20.	255(3)	34(1)	0.05379	0.21429	0.05026	0.01763	0.23192
21.	255(1)	62(2)	0.04762	0.07332	0.02147	0.02892	0.10224
22.	255(3)	62(2)	0.04762	0.07332	0.02147	0.02892	0.10224
23.	255(1)	220(2)	0.04762	0.13379	0.03614	0.03374	0.16752
24.	255(3)	220(2)	0.04762	0.13379	0.03614	0.03374	0.16752
25.	255(1)	252(2)	0.05203	0.21542	0.03333	0.00296	0.21838
26.	255(3)	252(2)	0.05203	0.21542	0.03333	0.00296	0.21838
27.	255(1)	397(2)	0.03197	0.13076	0.01873	0.00019	0.13095
28.	255(3)	397(2)	0.03197	0.13076	0.01873	0.00019	0.13095
29.	255(1)	456(2)	0.08289	0.15949	0.01840	0	0.15949
30.	255(3)	456(2)	0.08289	0.15949	0.01840	0	0.15949
31.	255(1)	487(2)	0.08907	0.16856	0.01358	0.01499	0.18355
32.	255(3)	487(2)	0.08907	0.16856	0.01358	0.01499	0.18355
33.	255(1)	491(3)	0.04762	0.07332	0.02147	0.02892	0.10224
34.	255(3)	491(3)	0.04762	0.07332	0.02147	0.02892	0.10224
35.	255(1)	492(1)	0.04762	0.07332	0.02147	0.02892	0.10224
36.	255(3)	492(1)	0.04762	0.07332	0.02147	0.02892	0.10224
37.	278(3)	456(2)	0.05876	0.16440	0.02783	0.00227	0.16667
38.	278(3)	487(2)	0.04203	0.08579	0.02763	0.01235	0.09814
39.	283(2)	397(2)	0.05143	0.06122	0.02521	0.02448	0.08571
40.	283(2)	487(2)	0.03851	0.04006	0.01659	0.01499	0.05505
41.	291(1)	397(2)	0.05143	0.06122	0.02521	0.02448	0.08571
42.	291(1)	487(2)	0.03851	0.04006	0.01659	0.01499	0.05505

5. Correlated substitutions in proteins

I applied two methods that do not rely explicitly on LD to determine if sites between Pto/Fen and Prf proteins are coevolving. The first method was CAPS (FARES and TRAVERS 2006). This method is designed to discover coevolving pairs of sites based on their correlations in the underlying matrices of pairwise biochemical divergence. I identified 12 pairs of coevolving sites between Pto and Prf and five pairs between Fen and Prf (Table 14). It is interesting to note that only two sites in Prf were implicated as coevolving with residues of Pto and Fen. The functional significance of these sites is discussed below.

TABLE 14. Putative coevolving amino acid residues between (A) Pto and Prf, (B) Fen and Prf, inferred by the CAPS method ($p < 0.001$). Coevolving sites are arranged by position in PTO and FEN.

(A)

	PTO	PRF	Correlation
1.	49	34	0.6974
2.	49	120	0.6457
3.	51	34	0.4784
4.	51	120	0.4905
5.	72	34	0.3102
6.	72	120	0.3042
7.	115	34	0.2324
8.	115	120	0.2868
9.	168	34	0.6170
10.	168	120	0.7020
11.	178	34	0.1691
12.	178	120	0.1881

(B)

	FEN	PRF	Correlation
1.	76	34	0.1535
2.	76	120	0.1582
3.	116	34	0.3875
4.	255	34	0.7020
5.	255	120	0.6863

The ELSC method identifies putatively coevolving sites by evaluating how the distribution of amino acid residues at one site is dependent on the distribution of amino acid residues at a second site. This method does not take into account biochemical characteristics of the residues as the CAPS method does, but considers how the distribution of amino acid residues at different sites in a protein changes in sub-alignments, conditioned on a single site of the protein. Using this method, I identified eight pairs of sites between Pto and Prf and 14 pairs of sites between Fen and Prf that were putatively coevolving (Table 15).

TABLE 15. Putative coevolving amino acid residues between (A) Pto and Prf, (B) Fen and Prf, inferred by the ELSC method. Coevolving sites with score > 3 are arranged by position in PTO and FEN.

(A)

	PTO	PRF	ELSC score
1.	88	212	3.32
2.	88	233	3.13
3.	115	120	3.73
4.	132	212	4.49
5.	135	212	4.49
6.	168	120	3.99
7.	200	120	3.89
8.	273	203	4.35

(B)

	FEN	PRF	ELSC score
1.	46	156	3.61
2.	46	159	3.61
3.	46	212	3.19
4.	76	397	3.15
5.	136	120	7.01
6.	136	135	9.04
7.	136	203	4.65
8.	136	213	9.12
9.	136	277	9.12
10.	136	510	4.28
11.	151	62	3.61
12.	151	491	3.61
13.	151	492	3.61
14.	241	536	3.61

6. Candidate sites in Pto

Previous molecular and biochemical studies have identified many residues in Pto that are important for interaction with AvrPto and AvrPtoB and downstream signaling. Here I describe the functional context of the 19 sites in Pto that were recognized as candidates for natural selection and coevolution with Prf, using the methods described above. Ten of these sites were identified based on two or more methods (Figures 10 and 11).

Domain I

The protein polymorphism in Pto between sites 43 and 88 is structured into two major haplotypes and many of the variable sites in this region show a pattern of LD that is consistent with unequal systematic epistasis (Table 12). Sites 43 and 46 are associated with each other and are polymorphic in the Tarapaca and Nazca populations of *S. peruvianum*. This region of Pto was also identified in a DNA shuffling study as important for AvrPto and AvrPtoB binding (BERNAL *et al.* 2005). Sites 49 and 51 in Pto form hydrophobic contacts with AvrPto molecule and are described as one interface with AvrPto (XING *et al.* 2007). F_{ST} analyses identified these sites as candidates of balancing selection. Site 49, in particular, was identified as a candidate in four independent analyses. LD-based analyses pinpointed this site as associated with Prf. This site was also identified using CAPS. Three alleles (H, E, A) segregate at site 49 in these populations. These segregating amino acid residues have very different biochemical properties (i.e. H is polar, basic and large, E is polar, acidic and small, while A is nonpolar and small). Three alleles (V, L and G) also segregate at site 51 in these populations and these amino acid differences are conservative. Site directed mutagenesis at sites 49 and 51 in Pto showed that the joint replacement of the H49E and V51G/D resulted in significantly reduced interactions with AvrPto, but not AvrPtoB (XING *et al.* 2007; DONG *et al.* 2009). Many of my alleles also have the combination of amino acid E49 with G51, as does the paralog Fen. Functional studies of other Pto alleles that contained E49/G51 from wild tomato species were able to activate an AvrPto specific resistance response (ROSE *et al.* 2005). However, these alleles (parv94, chm115, peru567) differ not only at these two positions and therefore variation at other amino acid positions may have contributed to AvrPto binding and activation of disease resistance. Since functional versions of Pto and Fen have the combination of E49 and G51 and the proteins successfully signal through Prf, these substitutions do not appear to compromise Prf signaling. The observed correlation of above-mentioned Pto substitutions with Prf may be driven in part because this portion of Pto forms an exposed interface, perhaps not only for the pathogen ligand AvrPto, but for other interacting molecular partners such as Prf.

Domain II

The next sites in Pto strongly correlated with positions in Prf are sites 70, 71, and 72. Variation at these sites is structured into two distinct protein haplotypes (RRQ and SCK). These positions are variable in the Tarapaca and Nazca populations and all are non-conservative. Close to these sites is site K69 which is invariant in protein kinases and is required for ATP binding. Mutations at K69 abolish Pto kinase activity and the ability of Pto to interact with AvrPto (SCOFIELD *et al.* 1996; TANG *et al.* 1996). Along with sites P73, E74, S76, G78, this region is necessary for binding of AvrPtoB, but not AvrPto (BERNAL *et al.* 2005).

Domain III

Site 88, in kinase domain III, is also associated with Prf and was identified using the ELSC and the LD-based method. The T/I polymorphism in populations of Tarapaca and Nazca is rather conservative. This region is involved in anchoring and orienting the ATP molecule and is generally a strongly conserved in protein kinases (HANKS and HUNTER 1995).

Domain V

Sites 115 and 124 in domain V were also identified using these analytical methods. Four amino acids segregate at site 115 in the populations of *S. peruvianum* studied here and this site was identified using the CAPS and ELSC methods. The K allele is found in all three populations, and the minor alleles Q, D and E are found in Tarapaca, Nazca and Canta, respectively. Although these substitutions are radical relative to one another, site-directed substitutions of K115E, and K115D in Pto did not affect the ability to bind to AvrPto and AvrPtoB (BERNAL *et al.* 2005), indicating that even radical changes at this position may not negatively affect downstream signaling through Prf. Site 124 shows a S/R polymorphism in each population and was identified as a candidate for partial epistasis with Prf using Ohta's LD partitioning method.

Domain VIa

Sites 132 and 135 occur at the junction between domains V and VI. These sites are polymorphic in Nazca and form two haplotypes: P132/S135 and L132/F135. The major allele P132/S135 is conserved across most alleles of Pto in other wild tomato species, as well as in Pth2, Pth3, Pth5 and Fen in *S. pimpinellifolium* (ROSE *et al.* 2005). The P to L substitution at 132 is conservative (both amino acids are small and nonpolar), while the S to F substitution at 135 is non-conservative (S is polar, neutral and small, while F is nonpolar and large). Domain VIa normally forms an extensive hydrophobic α-helix that stretches through the large lobe of the protein kinase. A polymorphism at site 154, towards the end of the α-loop, is correlated with variation in the Prf gene. This site is polymorphic in all three *S. peruvianum* populations and the two amino acid residues, F and Y, are at nearly equal frequency in these populations. The functional effect of this substitution has not been explicitly tested and this site was not polymorphic among the chimeras tested for AvrPto and AvrPtoB recognition by BERNAL and colleagues (2005).

Domain VIb

This domain contains two β-strands with an intervening loop. The loop is known as the catalytic loop because it helps mediate phosphoryl transfer. In protein kinases, this loop is formed by the sequence HRD(L/V)KxxN. Across my alleles, no polymorphism is present in this loop except at position 168. This corresponds to the first small “x” of the consensus sequence. This site is polymorphic in all three populations and is associated with Prf. Three different amino acid segregate (S, I and T), with S being the minor allele. Substitutions of I and T are conservative, while a substitution of S is non-conservative. This residue, along with 169, is predicted to be surface-exposed and control Pto signaling. However, variation at 168 did not directly affect AvrPto and AvrPtoB recognition (WU *et al.* 2004).

Domain VII

This domain forms part of the activation segment of protein kinases. A polymorphism at site 178 was identified using the CAPS method as coevolving with Prf. This site is polymorphic in Nazca and Canta. Alanine is the major allele at this locus and found in all Tarapaca individuals sampled. Some Nazca individuals are heterozygous at this position for A and P. Pto alleles with either A or P at site 178 were functional in recognizing AvrPto (ROSE *et al.* 2005). In contrast, some Canta individuals are heterozygous for T and A. The T substitution was found in another *S. peruvianum* population and versions of Pto with this substitution were unable to recognize AvrPto and/or activate a disease resistance response.

Domain VIII

Domain VIII comprises the P+1 loop and plays a major role in ligand recognition. Two sites emerged as interesting candidates in this domain. Site 200 is polymorphic for I and V and both variants are segregating in all three populations. This site was identified through the ELSC method as coevolving with Prf. Residue 205 was identified as a candidate for partial epistasis with Prf. Site 205 is polymorphic in all three populations and the L/F polymorphism is rather conservative (both are nonpolar, small → large). Together with sites T204, I208, F213, site 205 forms the second interface for binding AvrPto (XING *et al.* 2007) and with residues F213, V242, V250, N251, the first interface for binding AvrPtoB. Site directed mutagenesis at sites 205 and 213 in Pto showed that the joint replacement of the L205A and F213A disrupted the interaction of Pto with AvrPtoB, but not AvrPto (DONG *et al.* 2009). Furthermore, this residue along with I214 and N251 form a negative regulatory patch (NRP), which controls many aspects of signaling, including a negative regulation of signaling through Prf (WU *et al.* 2004).

Domain X

One site is found to be candidate for coevolution in domain X. Site 273 emerged as a candidate from both Ohta's LD analysis and the ELSC method. This site is polymorphic for I/L in all populations. Little is known about the potential functional effects of a variation at this position; however residues in the domain X are required for interaction with the pathogen effectors and downstream signaling (BERNAL *et al.* 2005).

Domain XI

One site in this domain was identified as a candidate for partial epistasis with Prf. Site 295 is polymorphic in Nazca (L/S), but not in the other two populations. This polymorphism results in a non-conservative change. Mutational analysis of this site showed that the non-conservative substitution of L295D behaved as wild type and was able to induce AvrPto-dependent cell death (MUCYN *et al.* 2009). This may indicate that functional differences between the L and S alleles may not be evident in AvrPto-based detection assays.

7. Candidate sites in Fen

Functional information is also available on the Fen protein kinase. Here I describe 12 sites identified in Fen as candidates for natural selection and coevolution with Prf. One half of these sites were recognized using two or more independent methods (Figures 12 and 13)

Domain I

Site 46 of Fen is polymorphic in only a single population of *S. peruvianum*, namely Nazca. The major allele at this locus encodes a phenylalanine, however most individuals in Nazca are heterozygous for F and L. ELSC identified this as potentially coevolving with sites in Prf.

Domain II

Site 76 was identified in three methods as coevolving with sites in Prf. A number of sites were identified as coevolutionary partners in Prf, however site Prf397 was discovered in two of these methods. The homologous position in Pto has been shown to be critical for AvrPtoB binding and this site is monomorphic for serine in our collection of 54 Pto alleles across seven tomato species. In contrast, at Fen four different amino acid residues are present: R, K, S, and G, and the major allele is R. Alleles of Fen from

S. lycopersicum and *S. pimpinellifolium* bind AvrPtoB and encode S at this site, as does Pto. This raises the question whether the Fen alleles from *S. peruvianum* are also able to bind AvrPtoB in a similar way and what role this site may play in protein interactions with Prf.

Domain V

This region connects the two lobes of the protein kinase and is important for anchoring the ATP molecule. Site 116 is polymorphic in all three populations, but the major allele encodes a tyrosine. This site was identified as coevolving with Prf by the CAPS method.

Domain VIa

The region typically forms a large α-helix away from the active site of the protein and may serve as a structural support of the kinase (HANKS and HUNTER 1995). Site 136 within this region was polymorphic in Nazca and Canta for I and M, but fixed for I in the Tarapaca population. This site was identified as coevolving with five sites in Prf. Site 151, polymorphic only in the Tarapaca population, also was identified as coevolving with sites in Prf. Three coevolving sites in Prf were consistently identified across three methods (LD, Ohta's LD partitioning and ELSC) and these Prf sites were different from those identified as coevolving with site Fen136. Site 153 in this domain was polymorphic in the Nazca and Canta populations and was identified as coevolving with Prf using Ohta's LD partitioning method.

Domain X

Three sites in domain X were identified as coevolving with Prf or under balancing selection. Site 241 was polymorphic in Tarapaca and identified as coevolving with a site in Prf. Sites 244 and 255 were polymorphic in all three populations and were identified as both coevolving with Prf and experiencing balancing selection. There is an overlap in the coevolving partners identified in Prf for these two Fen polymorphisms. It was shown previously that the region between residues 243 and 258 in Pto is important either for correct protein folding or binding to the Avr proteins and downstream components (BERNAL *et al*. 2005).

Domain XI

Three sites in domain XI were identified as coevolving with Prf. Site 278 is polymorphic in all three populations and identified as experiencing balancing selection by one of the F_{ST}-based methods, while site 283 and site 291 are polymorphic in Tarapaca and Canta, but not in Nazca. The coevolving sites in Prf identified for these three sites in Fen are located towards the distal region of Prf N-terminus.

8. Candidate sites in Prf

Prf is a large protein with 5 domains (Figure 14). The N-terminal domain of Prf physically interacts with Pto and Fen (MUCYN *et al.* 2006, 2009; CHEN *et al.* 2008) and shows an excess of non-synonymous variation, compared with other domains in this protein (Figure 3). Twenty one amino acid positions were identified as candidates for natural selection and coevolution with Pto and Fen. Three regions in the N-terminal domain of Prf can be recognized: (1) proximal, amino acid sites 23–120, (2) middle, 135–277 and (3) distal 397–536 (Figures 14 and 15).

The proximal region of Prf N-terminus

Sites 23 and 34 are polymorphic in Tarapaca and Nazca and form two haplotypes: R23/H34 and W23/Y34. The haplotype W23/Y34 is fixed in Canta, R23/H34 is present in Prf allele from *S. lycopersicum* (GenBank AAF76312) and a combination R23/Y34 in the *S. habrochaites* outgroup allele. Both substitutions are rather conservative (R23W – basic, large, polar → nonpolar and H34Y – polar, large, basic → neutral). Both sites are associated with candidates for balancing selection (Pto49 and Fen255) and this is consistent with partial epistatic selection as showed by partitioning of LD. In addition site 34 was identified in the CAPS method as coevolving with multiple sites of Pto, namely 49/51/72/115/168/178 and sites 76, 116 and 255 of Fen.

Site 62 is polymorphic only in Tarapaca and a substitution at this site from F to Y is rather conservative. This site showed significant associations with putative balanced polymorphisms at residues Pto49, Fen255 and Fen278. In addition, the ELSC method indicated that this site is coevolving with Fen151, which was supported by LD analysis.

Site 120 is polymorphic in all three populations with the minor allele L only in Tarapaca and the major allele changing from R in Tarapaca to Q in Canta with a transitory state R/Q in Nazca. The R allele is present in Prf from *S. lycopersicum* and the Q allele in *S. habrochaites*. A displacement of the polar, basic, large R to polar, neutral, small Q and nonpolar, small L is a radical change. The CAPS method detects this site as coevolving with Fen76 and Fen255 and Pto sites 49/51/72/115/168/178. Furthermore, the ELSC method corroborates associations with similar region in Pto (sites 115, 168, 200) and site 136 in Fen.

The middle region of Prf N-terminus

Sites in this region are identified by different methods as coevolving with Fen only.

Site 135 is polymorphic in Nazca and Canta with a conservative change from V to L. The transition from the fixed allele V in Tarapaca to the predominating allele L in Canta is similar to that observed at site 120. This position is associated with Fen136 as predicted by the ELSC method.

Sites 156 and 159 in Prf are putatively coevolving with amino acid Fen46 as was shown by the ELSC analysis. Both sites are polymorphic and represented as two distinct haplotypes in these populations (S156/P159, R156/S159). The replacement S156R is a radical change (serine is neutral and small, arginine is basic and large), while the substitution P159S is rather conservative (both are small, nonpolar → polar neutral). The combination S156/P159 is the major allele, present also in *S. habrochaites* and *S. lycopersicum*. The allele R156/S159 is segregating only in Nazca with one case observed in Canta.

Sites 213 and 277 are polymorphic only in the Canta population with 213D/277T as the major allele. Sequences from *S. habrochaites* and *S. lycopersicum* have 213H/277I allele, which is fixed in Nazca and Tarapaca. The change H213D is a radical change (histidine is basic and large, whereas asparagine is acidic and small). In contrast, the replacement I277T is rather conservative (both isoleucine and threonine are small, nonpolar → polar neutral). Moreover, residues Prf213 and Prf277 are shown by F_{ST}-

based methods as significant candidates for directional selection. The method ELSC indicated both these loci as coevolving with Fen136.

Site 220 is a candidate for experiencing epistatic selection together with Fen255. Prf220 in the Tarapaca population is segregating for K and I. The I allele is the major allele, present also in Prf from *S. habrochaites*, while K appears in Prf from *S. lycopersicum*. The substitution I220K is a radical change – isoleucine is nonpolar and small, whereas lysine is polar, basic and large.

Other sites in this region (203, 212, 233 and 252) were candidates for coadaptation with both Pto and Fen.

Site 203 is polymorphic across three populations of *S. peruvianum*. In the Tarapaca population, A is the major allele while T allele is the major in Nazca and Canta. Replacement from A to T is rather a conservative change. The residue 203 was predicted only by the ELSC method as a candidate for correlated evolution with Pto273 and Fen136.

Site 212 is segregating in three populations with C as the major allele in Tarapaca and Canta, present also in *S. lycopersicum*. In Nazca, F predominates and appears also in *S. habrochaites*. A third allele Y is present in Nazca and Canta. This site was identified by the ELSC method as coevolving with Pto sites 88/132/135 and site 46 in Fen.

Site 233 is polymorphic for L and M in these three populations. The allele L is the major in Tarapaca and Canta, whereas in Nazca the two alleles are present in equal frequency. The methionine is found in Prf sequence from *S. habrochaites* and the substitution L233M is conservative. The ELSC method indicates this site as putatively coevolving with Pto88, while the partitioning of LD suggests that this site is in epistatic relationship with Fen76.

Site 252 in Prf, along with site 220 in this region, is another candidate for partial epistatic selection not only with Fen255, but also Pto49. This site segregates in three populations of *S. peruvianum* with alleles K and T. The K allele is the major allele in Tarapaca, while T predominates in Nazca and Canta. The change from K to T is radical – lysine is polar and large and threonine is neutral and small.

The distal region of Prf N-terminus

In this region many polymorphic residues show partial epistasis with Fen and Pto.

Site 397 is polymorphic in the Tarapaca population for Q and L. The L allele is the minor allele in this population and appears also in Prf from *S. habrochaites*. This residue is in LD with Pto154 in Tarapaca and is a candidate for epistatic selection with many sites in Pto and Fen, namely Pto49, 124, 154 and Fen76, 255, 283, 291. In addition, site Fen76 was pointed out as coevolving with Prf397 by the ELSC method.

Site 456 is polymorphic in all three populations, with C as the major allele in Tarapaca, Y predominating in Canta, and with both alleles in equal frequency in Nazca (as in position Prf220). This site shows unequal systematic disequilibrium with residues Pto273 and Pto295, as well as sites 151, 153, 244, 255 and 278 in Fen. Furthermore, this locus is a candidate for balancing selection detected in F_{ST}-based methods.

Site 487 is also segregating in all three populations (S/F) with the major allele S. The substitution S487F is a radical change (polar, small → nonpolar, large) and F allele is present in *S. habrochaites*. This site is not only a candidate for epistatic relationship with multiple sites in Pto and Fen (namely Pto43, 46, 49, 51, 70, 71, 88, 154, 205 and Fen76, 153, 244, 255, 278, 283, 291), but also indicated as experiencing balancing selection in F_{ST}-based tests. Of these Pto and Fen residues, Pto49, 51 and Fen244, 255, 278 are also candidates for balancing selection.

Sites 491 and 492 are polymorphic only in Tarapaca, where they form haplotypes K491/A492 and N491/S492. The former is the major allele and the replacement K491N is a radical change (basic, large → neutral, small), whereas the replacement A492S is rather conservative. The residues Prf491 and Prf492 in the Tarapaca population are in LD with sites Pto49 and Fen151, and are candidates for epistatic selection with these loci. The site Fen151 was identified by the ELSC method as coevolving with these Prf loci. Two additional sites in Fen (Fen255 and Fen278) showed significant LD with these Prf sites.

The next two residues were identified in the ELSC method as coevolving partners of Fen only.

Prf site 510 is putatively coevolving with Fen136. This locus is segregating for S and T in Nazca and Canta, and is fixed for T in Tarapaca. This replacement is conservative, however, the S allele is a major allele in Canta and both alleles are present in a nearly equal frequency in Nazca.

Site 536 is polymorphic only in Tarapaca for I and M with I as the major allele. This locus was identified by the ELSC method as putatively coevolving with Fen241.

In summary, for the most part the proximal region of Prf N-terminus contains candidates indentified by the CAPS method as coevolving with Pto or Fen. Polymorphisms in the middle region were identified as coevolving with Pto or Fen based on the ELSC method. Some of these same sites were identified by F_{ST}-based methods as candidates for directional selection (i.e. local adaptation). Several sites in the distal region of Prf N-term domain were identified as candidates for unequal systematic disequilibrium (i.e. partial epistasis) with Pto or Fen using Ohta's partitioning of LD method. Some of these same sites were identified by F_{ST}-based methods as candidates for balancing selection. The partner sites in Pto or Fen were also identified as candidates for balancing selection (Figures 16 and 17).

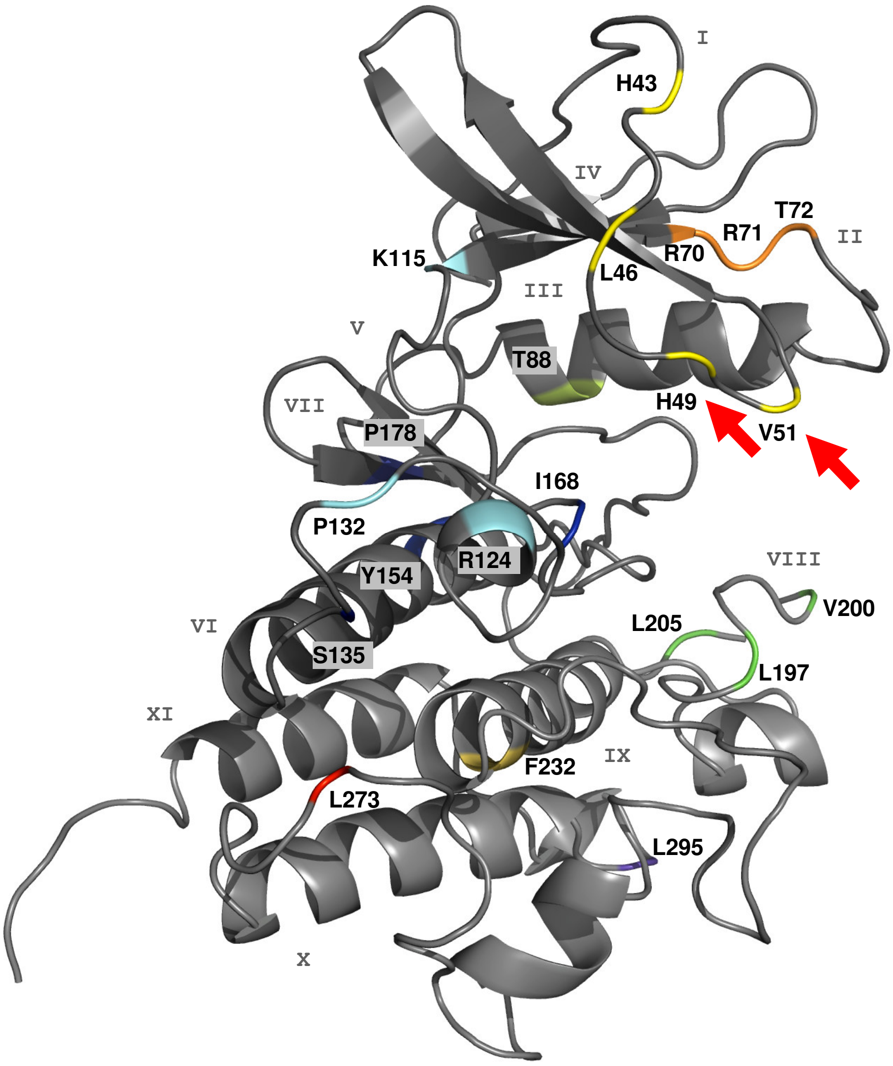


FIGURE 10. Ribbon diagram of the Pto crystal structure (*S. pimpinellifolium*, PDB 2qkw). Red arrows indicate the two polymorphic residues that are candidates for balancing selection. Other important polymorphisms are indicated by boldface letter and position. Colors correspond to individual kinase domains marked with Roman numerals.

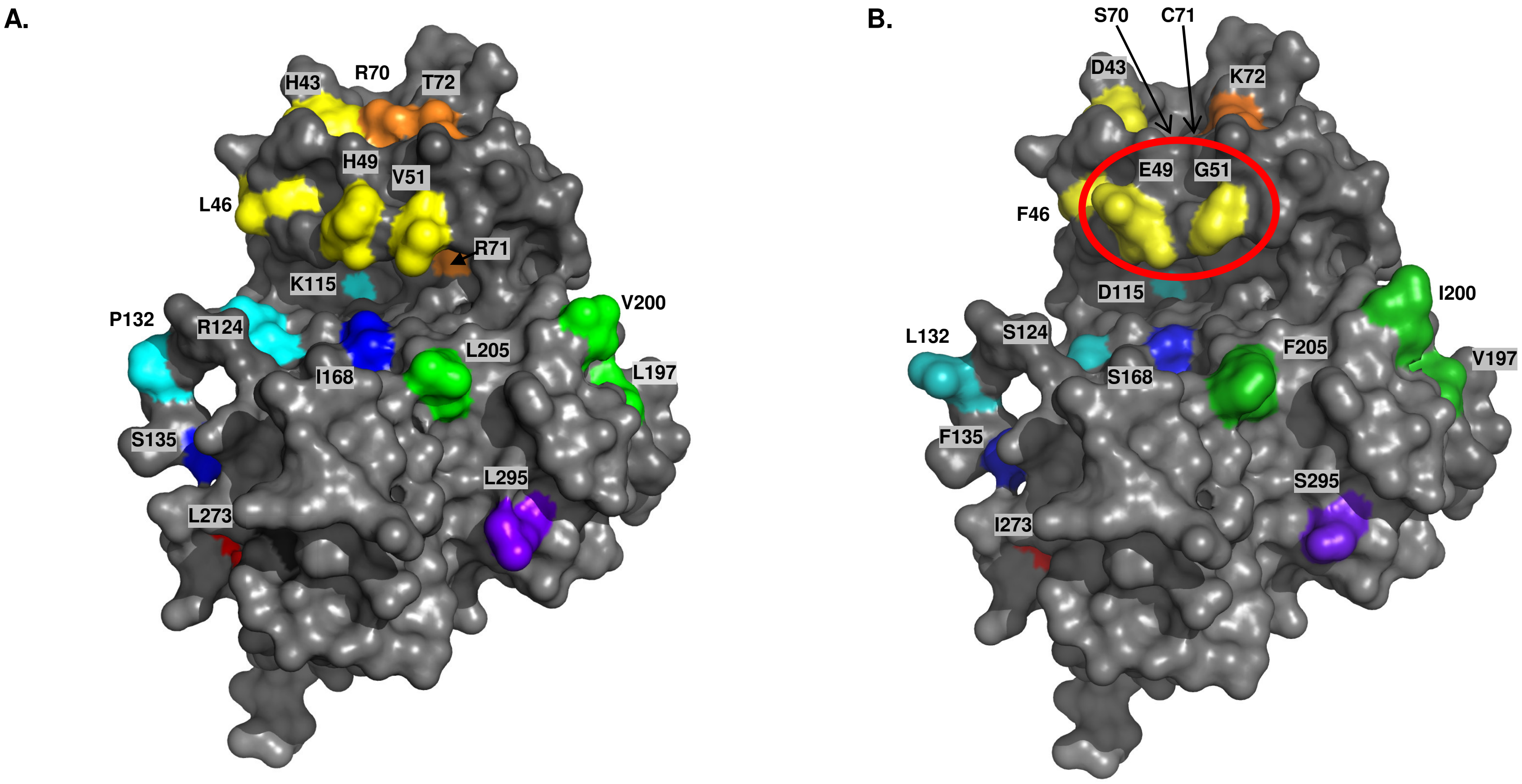


FIGURE 11. Molecular surface representation of Pto molecules highlighting exposed polymorphic amino acids. (A) Allele from *S. pimpinellifolium*. (B) Model showing alternative allele states found in *S. peruvianum*. Colors correspond to individual kinase domains (yellow – domain I, orange – II, light blue – V, blue – VI, green – VIII, purple – X, red – XI). Red circle indicates residues predicted to experience balancing selection in *S. peruvianum*.

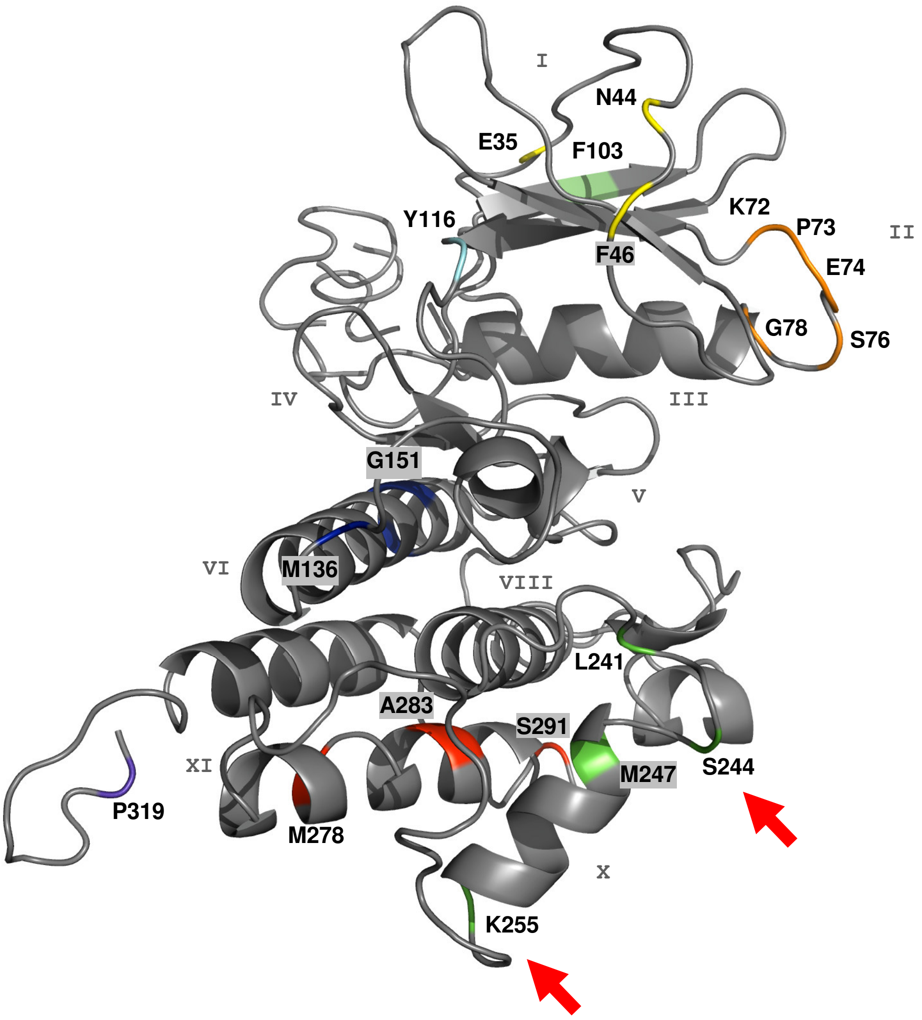


FIGURE 12. Ribbon diagram of the Fen protein model, based on allele from *S. pimpinellifolium*. Red arrows indicate the two polymorphic residues that are candidates for balancing selection. Other important polymorphisms are indicated by boldface letter and position. Colors correspond to individual kinase domains marked with Roman numerals.

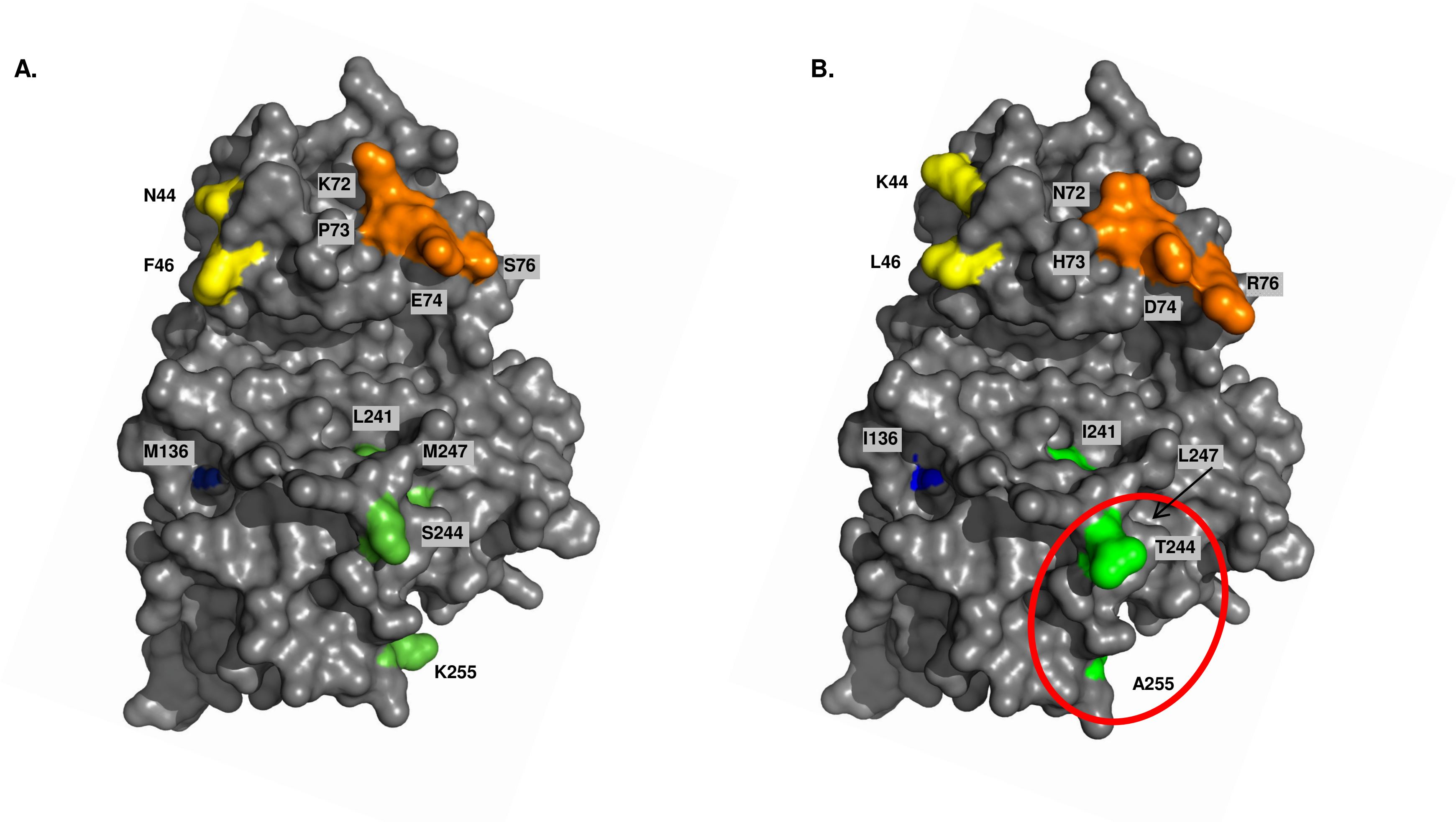


FIGURE 13. Molecular surface representation of Fen molecules highlighting exposed polymorphic amino acids. (A) Allele from *S. pimpinellifolium*. (B) Model showing alternative allele states found in *S. peruvianum.* Colors correspond to individual kinase domains (yellow – domain I, orange – II, blue – VI, green – X). Red circle indicates residues predicted to experience balancing selection in *S. peruvianum*.

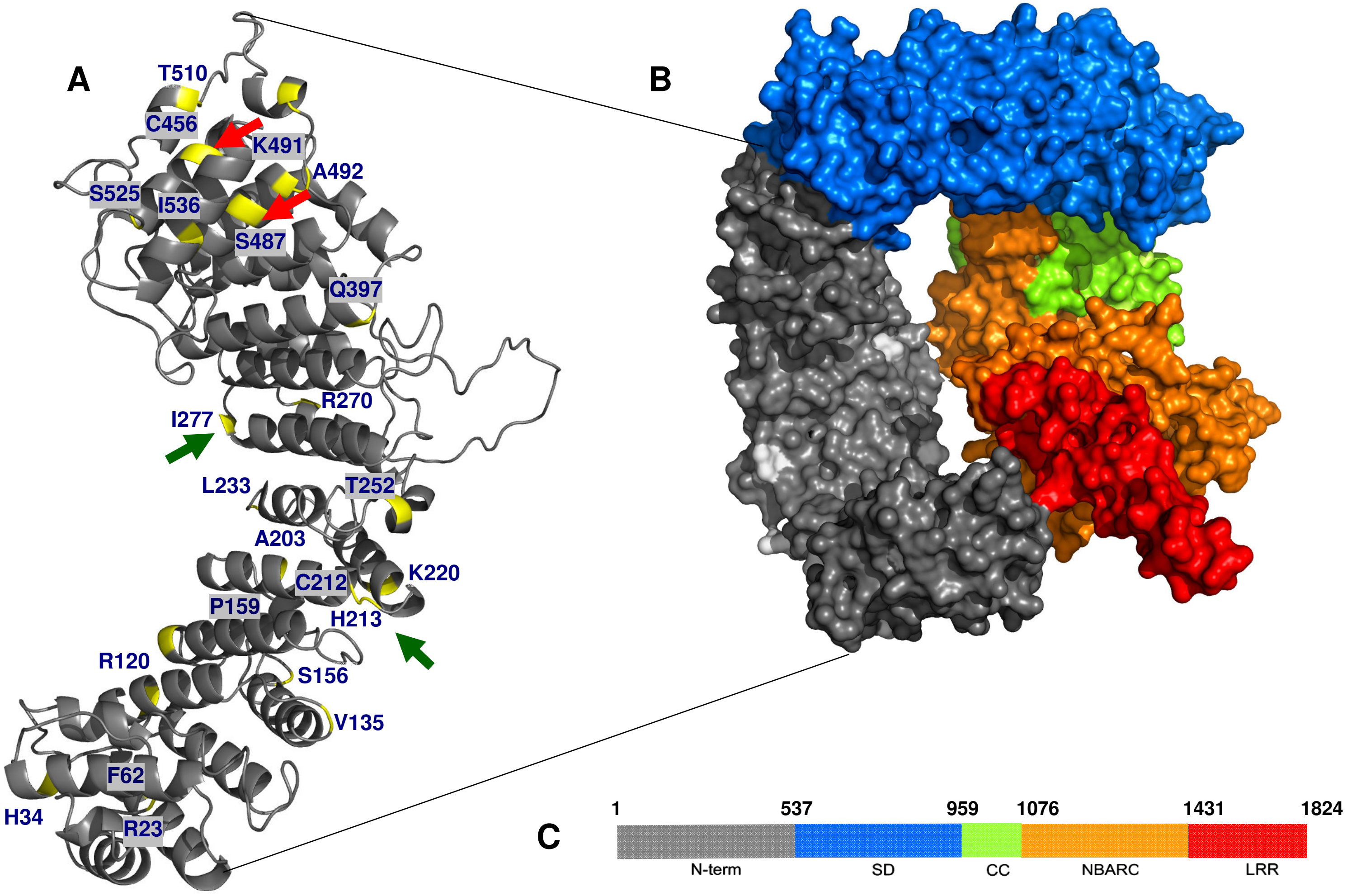


FIGURE 14. Model of the Prf protein: (A) N-terminal domain shown as a ribbon diagram with polymorphic sites in yellow; arrows indicate candidate sites for balancing selection (red) and directional selection (green). (B) Molecular surface representation of Prf. (C) Domain diagram of Prf, colored as in B; numbers indicate amino acid positions.

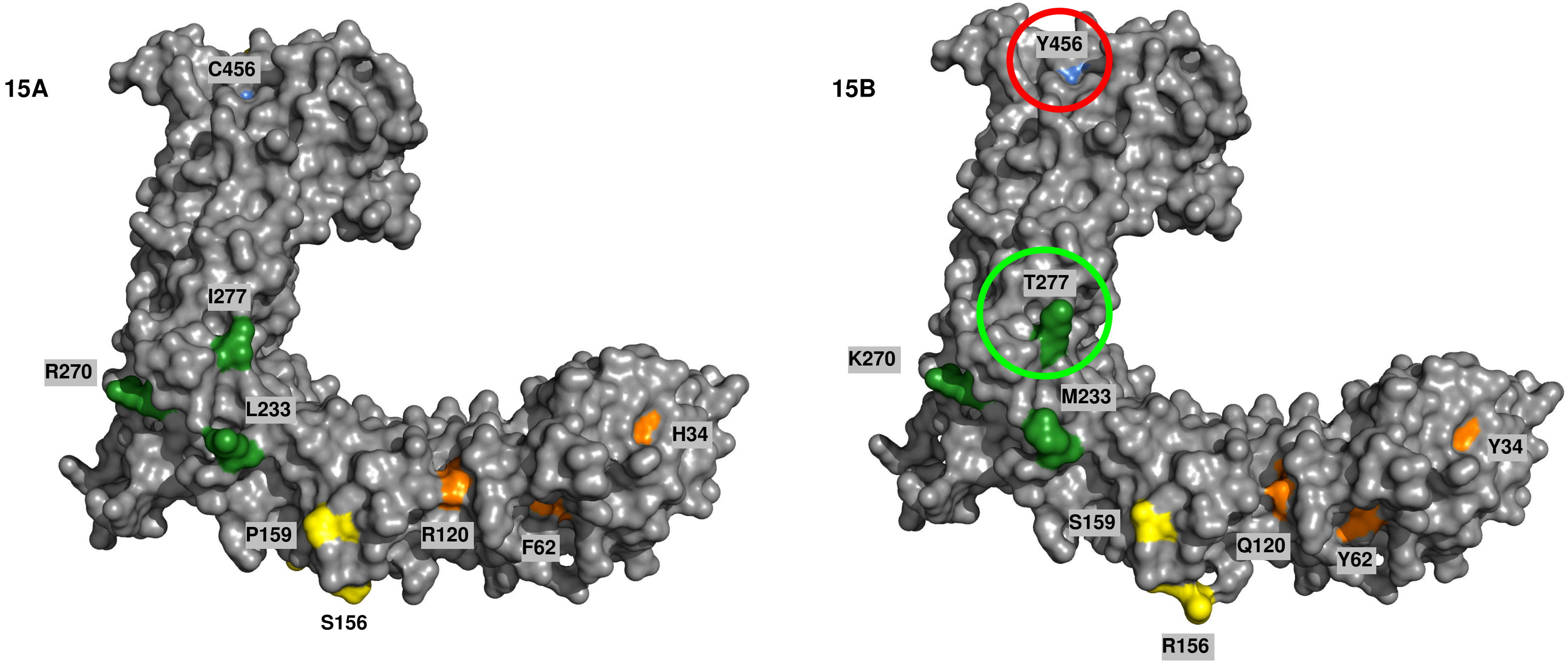
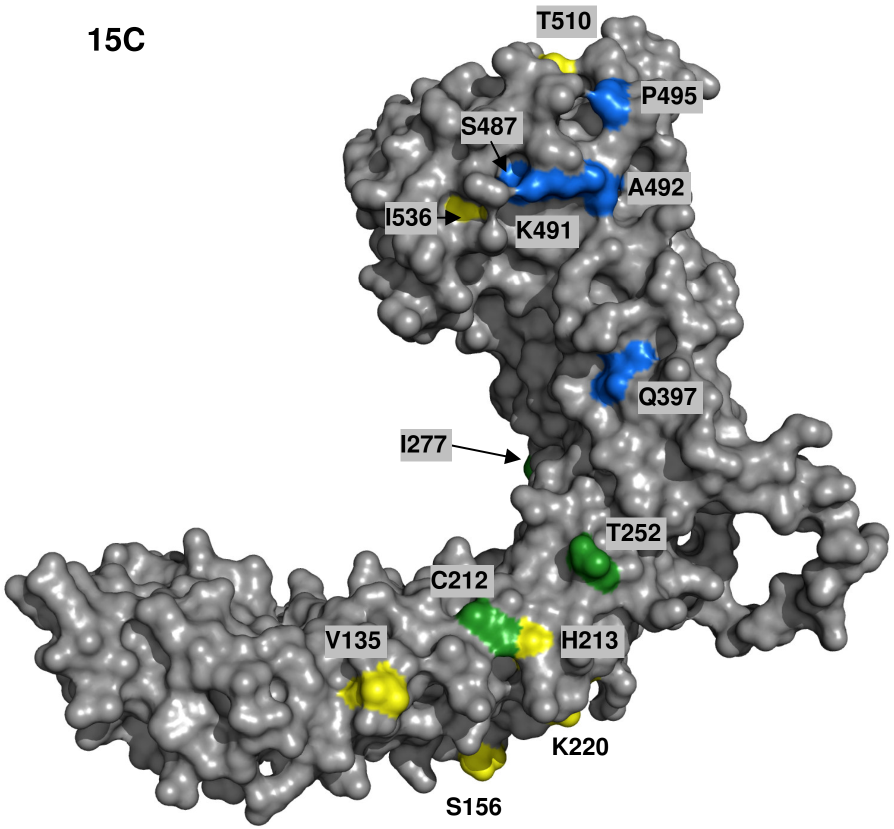


FIGURE 15. Molecular surface representation of the Prf N-terminal domain highlighting three regions (in different orientations): proximal (orange), middle (green), distal (blue). (A) and (C) Allele from *S. lycopersicum*; (B) and (D) Model showing alternative allele states found in *S. peruvianum*. Red circle indicates candidate sites for balancing selection; green circle indicates candidate sites for directional selection; sites predicted as coevolving with Fen only are shown in yellow.

15C


T510
P495
S487
A492
I536
K491
Q397
I277
T252
C212
H213
V135
K220
S156

15D

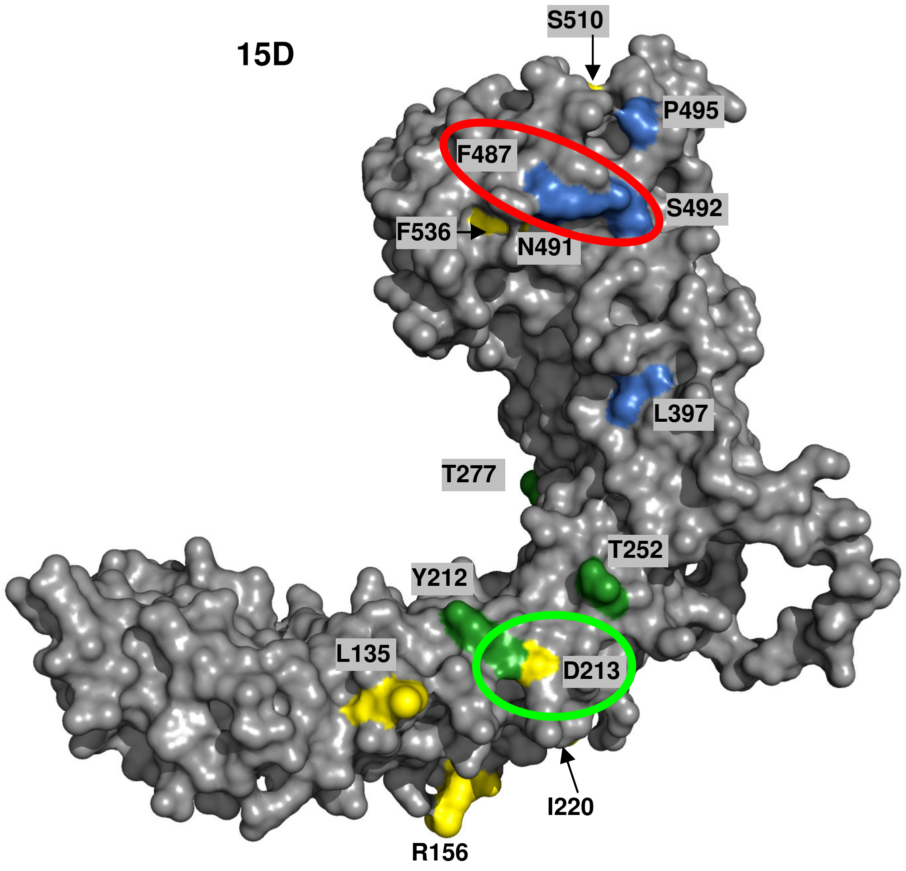

S510
P495
F487
S492
F536
N491
L397
T277
T252
Y212
D213
L135
I220
R156

FIGURE 16. Combined picture of regions coevolving between Pto and Prf detected by three independent methods. Associated residue pairs are depicted on model structures of Pto and Prf and connected by colored lines: the Ohta's partitioning of LD (blue), ELSC (green), CAPS (orange).

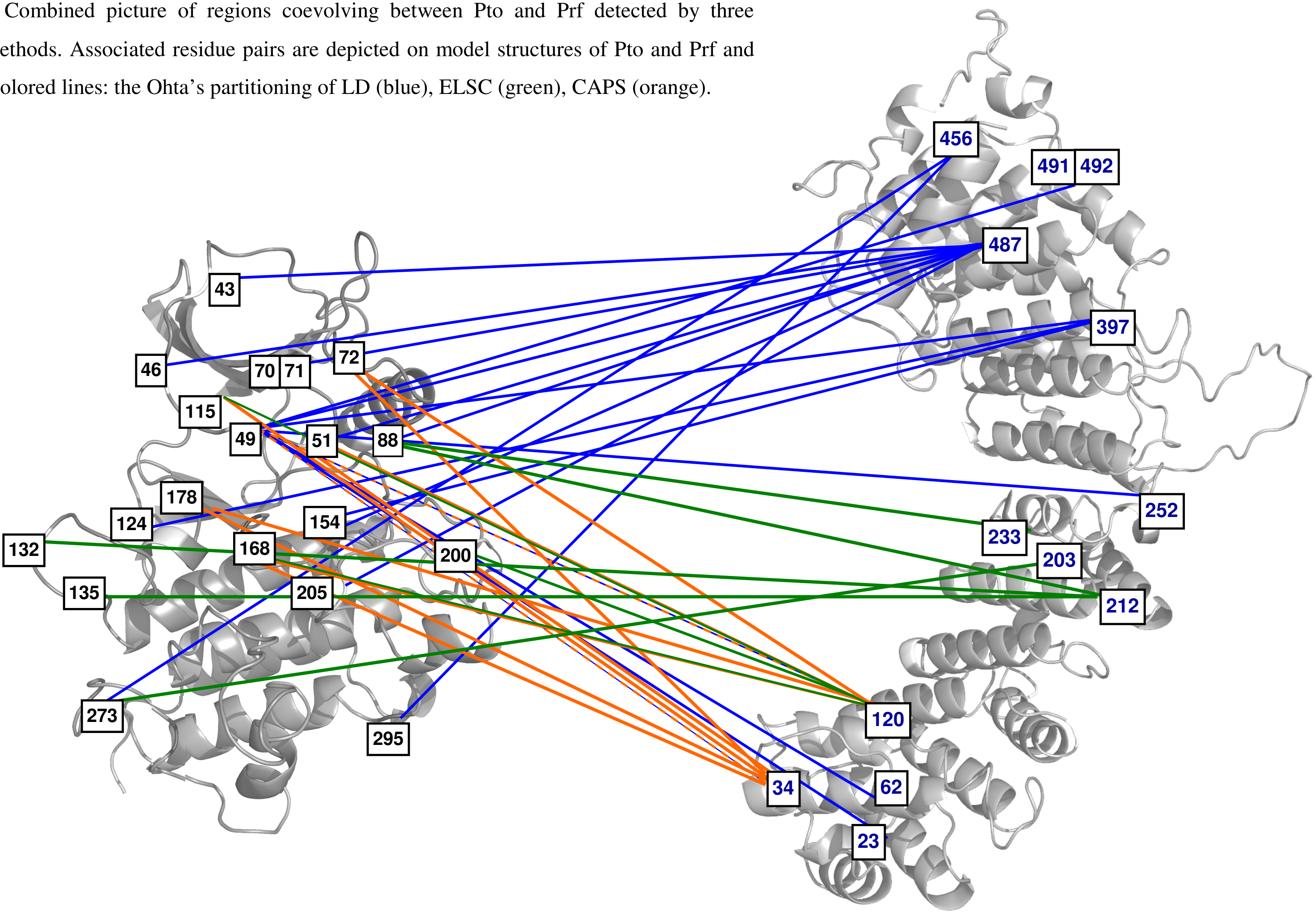
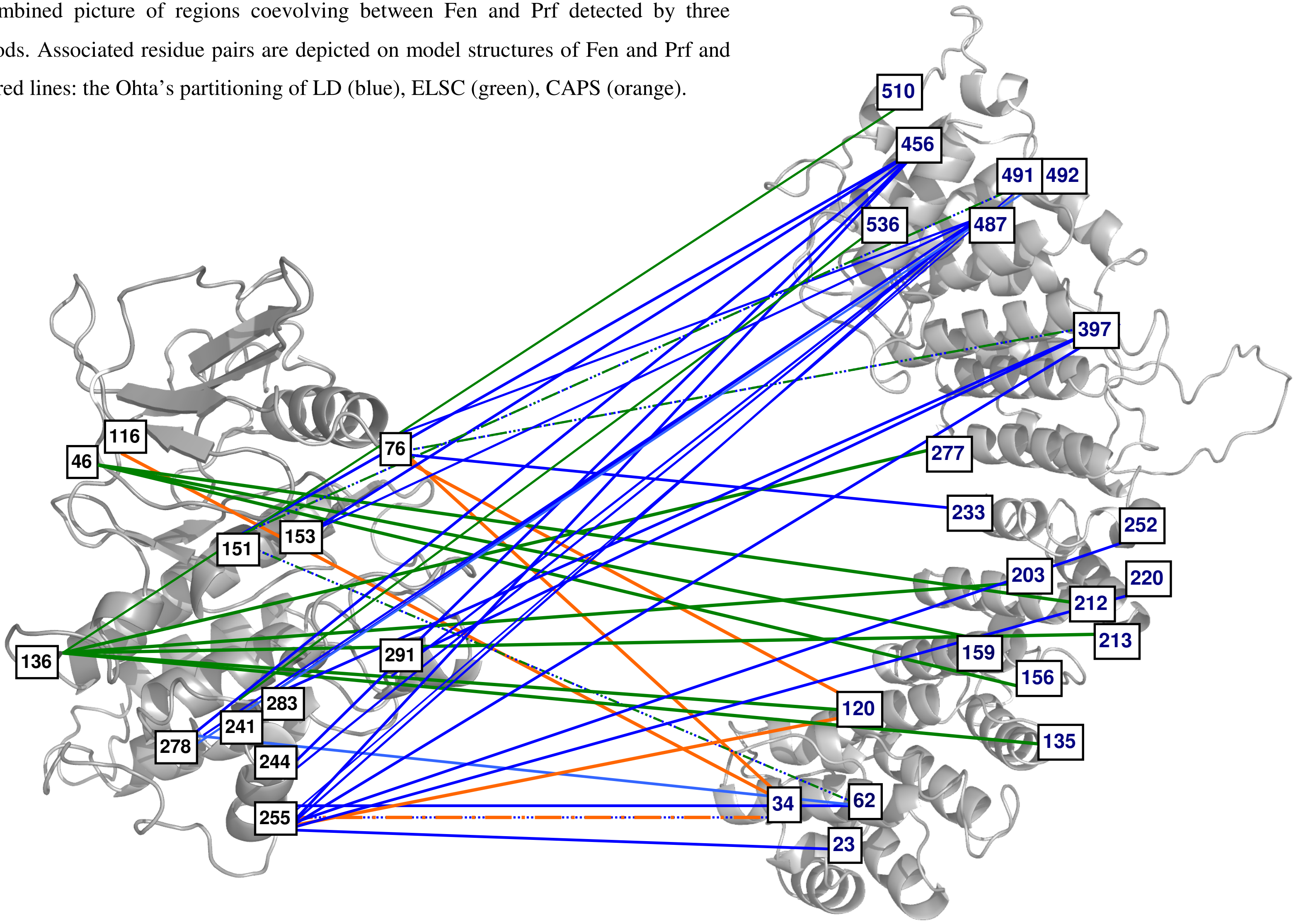


FIGURE 17. Combined picture of regions coevolving between Fen and Prf detected by three independent methods. Associated residue pairs are depicted on model structures of Fen and Prf and connected by colored lines: the Ohta's partitioning of LD (blue), ELSC (green), CAPS (orange).



CHAPTER 4

DISCUSSION

Nothing in evolution makes sense except in the light of population genetics

M. Lynch (2007)

1. Evolution of genes at different points in a signaling pathway in tomatoes

In the study of five genes involved in the Pto signaling pathway I found two loci with elevated amino acid polymorphism consistent with balancing selection, namely in *Pto* and *Pfi*. A third gene, *Prf*, showed signatures of both balancing selection and purifying selection, while two other genes, namely *Fen* and *RIN4*, showed predominantly purifying selection. Previous studies had reported that *Pto* is subject to balancing selection within different wild tomato species and, given the substantial functional information available for *Pto*, a scenario of balancing selection is not surprising (ROSE *et al*. 2005, 2007). Pto binds and recognizes two different pathogen ligands and triggers a defense response in wild tomato. The maintenance of different host resistance proteins in natural populations is consistent with an ongoing coadaptation between host and pathogen.

The second gene that showed elevated amino acid polymorphism relative to neutral expectations was *Pfi*. This gene is further down in the signaling pathway and acts as a negative regulator of defense (TAI 2004). The protein product of *Pfi* physically interacts with Prf, has a putative nuclear localization signal and is predicted to encode a transcription factor. As such, it may respond to an activated form of Prf by moving into the nucleus. There, it may mediate the downstream resistance responses including the hypersensitive response. As a component of the signaling pathway, rather than a known pathogen target, it is 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, although enzymatic assays to confirm hydrolytic activity have yet to be conducted. Provided that 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 defense. Alternatively, this molecule could serve as a direct target by other tomato pathogens. Recent studies reveal that all points in immune signaling pathways can be vulnerable to pathogen manipulation. Pathogens may specifically secrete proteins (i.e. effector molecules) to target downstream points in the pathway and suppress host resistance. Since Pfi is a negative regulator of defense, alteration of protein stability could result in suppression of the hypersensitive response. In this case, balancing selection may not be specifically operating on enzymatic function, but rather on pathogen evasion. Alternative forms of *Pfi* may vary in their “resistance” to manipulation by pathogen molecules.

Prf, one of the central molecules of this pathway, showed two distinctive signals of natural selection. The region known to physically interact with Pto and Fen showed elevated amino acid polymorphism, providing the first hints that balancing selection at *Pto*, may be carrying over to its interacting partner, *Prf*. These sorts of correlated selective histories open the door to more complex forms of selection, such as epistatic selection between molecules. Evaluation of the evidence for epistatic selection is discussed below. In comparison to the first half of *Prf*, the second half of this gene shows greater evolutionary constraint, consistent with its presumed role in downstream signaling.

The *Fen* and *RIN4* genes showed 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 yet. It is possible that this is a “defeated” resistance gene, i.e. it no longer recognizes contemporary pathogens of tomatoes or, alternatively, Fen does not operate in the same isolate specific manner that Pto does. If *Fen* is involved in basal defense and not in isolate-specific defense, it would be subject to different evolutionary forces than molecules known to be involved in isolate-specific defense, such as *Pto*. One such molecule that is known to contribute to basal defense and is involved in different resistance pathways (at least in *A. thaliana*) is *RIN4*. Strong protein conservation at *RIN4* is observed in the analyzed tomato population. However, the frequency spectrum of mutations and pattern of LD among *RIN4* alleles expose additional aspects of the history of *RIN4*, including the presence of a young, but divergent *RIN4* allelic type, carrying several derived mutations. One possible explanation of this pattern is that this divergent *RIN4* allele is passaging through the population as an advantageous allele. However, capturing a selective sweep in progress is quite unlikely because the sojourn times of advantageous alleles are generally too fast. 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 shows 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 and STEPHAN (1996). Here two or more sites experience positive selection, but are found on different haplotypes. The fixation process is paused until recombination can bring the adaptive mutations together into one haplotype with higher fitness. 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*, one may expect 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 there is an elevated substitution rate at non-synonymous sites between *S. peruvianum* and outgroup species, in combination with a reduction of variation at *RIN4* within *S. peruvianum*. However, I found no evidence for recurrent selective events at *RIN4* in the history of this tomato population. 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 metabolic pathways showed the greatest protein conservation due to selective constraint, as compared to downstream genes (RAUSHER *et al.* 1999; LU and RAUSHER 2003; RAUSHER *et al.* 2008). 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, signaling pathways and metabolic pathways may operate under a similar rules involving pleiotropy, the pleiotropy gradient in signaling pathways may be "inverted" relative to what is observed in metabolic pathways (i.e. the genes with the greatest pleiotropy may be located further downstream rather than upstream, if they serve as convergence points for different host signals).

In previous studies of plant metabolic pathways the possibility of bouts of positive selection or adaptive evolution could be excluded as the reason why downstream genes showed "relaxed constraint" relative to upstream genes (RAUSHER *et al.* 2008; RAMSAY *et al.* 2009). In contrast, genes at proximal points of signaling pathways for pathogen defense may well be expected to experience adaptive evolution. A few recent studies in *A. thaliana* have evaluated a number of defense genes, some of which are known to operate together in specific signaling pathways (BAKKER *et al.* 2006, 2008; CALDWELL and MICHELMORE 2009). 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 defense genes in *A. thaliana* revealed that, while some R-genes showed histories of transient balancing selection or partial selective sweeps, genes further downstream experienced almost exclusively purifying selection (BAKKER *et al.* 2006, 2008). At a broad scale, these results are consistent with expectations that genes downstream in defense pathways experience greater evolutionary constraint and upstream genes 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 and MICHELMORE 2009). In a study of 10 downstream defense 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 gene. Interestingly, a fourth gene in the CALDWELL and MICHELMORE (2009) study overlapped with one in our 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* harbors substantial silent polymorphism within *A. thaliana*, displaying more genetic variation than found at 93.5% in a set of 355 reference loci (CALDWELL and MICHELMORE 2009). To what degree this elevation in genetic diversity reflects past selective events, has not been investigated.

Compared to these other studies, I do not find a strong correlation of selective constraint and pathway position. This may be a result of pathway length, since longer pathways usually result in stronger correlations with functional constraint (RAMSAY *et al*. 2009). Perhaps, the present choice of genes captures only the very proximal part of the signaling pathway and therefore does not include genes analogous to those reported in previous studies. As more genes downstream in the *Pto* signaling 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 defense pathways can be targets of adaptive evolution. Consequently, population genetic studies such as the one here, can uncover very interesting candidates for future functional studies. For example, the consequences of RIN4 protein polymorphism on *Pto*-specific resistance and possibly basal defense 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 defense repertoire of plants, since although this gene displays a signature of balancing polymorphism similar to R-genes in plants, this gene does not share the motifs of most R-genes.

2. Detecting epistatic selection between interacting proteins

I implemented population genetics and bioinformatics methods to infer associations between proteins interacting in the tomato disease resistance pathway. In general, the results based on partitioning of LD variance and two correlated substitutions methods did not overlap greatly. The differences are not surprising, since although these methods were developed to infer molecular coevolution, their underlying assumptions are quite different. This may be why attempts to use coevolution signals to predict sequence regions involved in protein-protein interactions report different levels of success (PAZOS and VALENCIA 2002; HALPERIN *et al.* 2006).

One of the approaches used was Ohta's method to partition the total variance of linkage disequilibrium into within and between population components (OHTA 1982a,b). This method was developed to discriminate between epistatic natural selection and stochastic processes as the main cause of the observed LD. Systematic associations among alleles in isolated populations of a species may be taken as evidence of the direct action of natural selection on the loci involved (LEWONTIN 1974). For systematic associations, there is a relatively large within-population component and a relatively small between-population component, because LD is in the same direction in each population. In contrast, a large between-population component of LD is most readily attributable to nonselective effects of population subdivision or founder effects (BROWN and FELDMAN 1981; OHTA 1982a, b). Interestingly, although epistatic selection plays an important role in population genetics, this method has not been widely employed to identify natural selection operating within or between molecules. In one case, however, epistatic selection was detected using Ohta's method (e.g. in the alcohol dehydrogenase gene of *Drosophila*). Significant LD between sites in two introns of *Adh* within populations was detected, despite high levels of recombination (SCHAEFFER and MILLER 1993). Follow-up studies suggested that epistatic selection at *Adh* maintains the pre-mRNA structure necessary for stem-loop formation (KIRBY *et al.* 1995). Functional experiments confirmed predicted long-range interactions (PARSCH *et al.* 1997; BAINES *et al.* 2004) and demonstrated the role of an intronic hairpin structure in the splicing process (CHEN and STEPHAN 2003).

In a different study, WHITTAM *et al.* (1983) detected systematic associations between allozymes in three geographically isolated natural populations of *E. coli*. Several of these enzymes are functionally interrelated, occurring in the same metabolic pathway (e.g., ACO and IDH in the TCA cycle). Thus, epistatic selection could increase favorable allozyme combinations and maintain stable disequilibria in all populations. It is also possible that physiological differences between allozymes of these enzymes are expressed as selective differences between genotypes in certain genetic and/or environmental backgrounds.

Recently, DA SILVA (2009) used Ohta's method to test if amino acid covariation of the highly polymorphic HIV-1 exterior envelope glycoprotein V3 region is due to fitness epistasis between residues. In this case, fitness interactions among V3 amino acids could be hypothesized, since several sites appear to be involved in determining coreceptor usage. Furthermore, structural analyses have suggested interactions between some V3 sites that may affect V3 structural conformation, but none of these interactions has been demonstrated through functional analyses or fitness assays. In fact, positive selection in DA SILVA (2009) might explain differences in allele frequencies among subpopulations, indicating that these differences are adaptive rather than due to genetic drift. However, the substantial LD, or amino acid covariation, reported from previous analyses of one or a few V3 sequences from each of many patients (KORBER *et al.*1993; BICKEL *et al.* 1996; GILBERT *et al.* 2005; POON *et al.* 2007; TRAVERS *et al.* 2007) can be explained by population subdivision. The absence of a correlation between LD and coreceptor usage phenotype suggests that fitness epistasis is an unlikely cause of LD.

In this study, results of LD partitioning also suggest that restricted migration and genetic drift are the main causes of observed associations between genes from the *Pto* cluster. In addition, since these genes are tightly linked, linkage and population subdivision might enhance the effects of each other.

Approximately 3% of the pairs of polymorphic sites of *Pto* and *Prf* or *Fen* and *Prf* met the criteria for unequal systematic disequilibrium. Unequal systematic disequilibrium is an intermediate between systematic and non-systematic disequilibrium and is equivalent to partial epistasis. This means that epistatic selection occurs in only a few subpopulations or might be also interpreted as interaction of genetic drift and epistatic selection. A scenario of natural selection favoring particular combinations of alleles is also supported by F_{ST}-outlier tests. These methods show that protein polymorphisms within interacting protein pairs are experiencing balancing selection. These same sites identified based on F_{ST}-outlier methods show epistatic associations between proteins. Balancing selection maintains alternative alleles in a population for much longer periods of time than neutral alleles persist under random genetic drift (GILLESPIE 1991; TAKAHATA 1992). Together with linkage, balancing selection elevates the amount of variation within the region above that expected from the balance between mutation and random drift. One example is the polymorphism in the *Adh* gene of *D.*

melanogaster, where balancing selection seems to maintain strong LD. With many selected sites, linked polymorphisms may show strong LD, with only two common segregating haplotypes (KELLY and WADE 2000) or random fluctuations reduce variation below the predictions with stable genotype frequencies (NAVARRO and BARTON 2002). In the present study, this interpretation could explain variation at many sites in protein domains consisting of candidates for balancing selection (domain I in Pto, domain X in Fen and distal region of the Prf N-terminus). Here the results imply that balancing selection could maintain not only strong LD in close polymorphic sites within gene, but also associations between particular sites between genes via epistatic selection. This is consistent with linked coadapted loci, such as the components of the *Brassica* self-incompatibility system (SATO *et al.* 2002). Likewise strong LD, both within and between MHC genes (TAKAHATA and SATTA 1998; SANCHEZ-MAZAS *et al.* 2000; MEYER and THOMSON 2001), predisposes MHC loci to epistatic interactions and genetic hitchhiking (NAVARRO and BARTON 2002; VAN OOSTERHOUT 2009). Epistasis is evident from the differences in disease phenotype caused by distinct combinations of alleles at multiple loci (GREGERSEN *et al.* 2006 – see also Case studies of epistatic selection in Introduction). Furthermore, the MHC genes are surrounded by linked genetic variation that is associated with more diseases than any other part of human genome (VAN OOSTERHOUT 2009). These patterns of variability within the MHC genes suggest that different loci may be involved in different kinds of interactions. However, Ohta consistently interpreted the observed large variance of LD in human and mouse MHC as a result of population subdivision and limited migration, but not epistatic selection (OHTA 1982a, b). One alternative explanation is that local adaptation to different parasite communities may be responsible for the unexpectedly large differentiation of the MHC (e.g. BERNATCHEZ and LANDRY 2003).

Therefore, the interpretation that stochastic processes are the main source of observed LD in the present research could be too conservative. Linkage disequilibrium may have been caused, in part, by sampling of spatially isolated populations or sampling during expansion of successful alleles. If epistatic selection could be a general explanation for the occurrence of LD, then systematic disequilibria would be observed for those allele combinations that are favored in all localities and nonsystematic disequilibrium would represent combinations that are locally or temporally adaptive

(WHITTAM *et al.* 1983). Thus, if subpopulations are not identical, because they occupy different environments, then nonsystematic disequilibrium may indicate either genetic drift or “epistatic adaptation” to local environments as the cause of LD (DA SILVA 2009).

This interpretation is supported by another method used in this study, called ELSC (DEKKER *et al.* 2004). ELSC is the alignment perturbation method, which does not assume any mutational model, since it was not developed specifically for population genetics analyses, but to study orthologous sequences of interacting proteins from different species. In this analysis, the pooled multiple alignment of the R-proteins from three populations was used. The ELSC method introduces structure in the total alignment by creating the subalignment based on allelic state of a site. Using this method, I found that most intermolecular associations involved sites from the middle region of Prf N-terminus. However, according to Ohta’s method, the source of linkage associations of many of these sites could be attributed to stochastic processes. Results of F_{ST}-based outlier detection methods suggest that some of sites in this Prf region are candidates for experiencing directional selection. In addition, the interacting sites from Pto and Fen are polymorphic in a few populations only. Taken together the set of significant Pto-Prf and Fen-Prf pairs detected by the ELSC method could reflect sites coevolving due to local adaptation to particular environments (i.e. epistatic adaptation).

The third method used in this study to detect coevolving sites between molecules is CAPS (FARES and TRAVERS 2006). This method is based on correlation of variability between protein residues and uses not only sequence information, but also the fact that each of the 20 standard protein amino acids has its own unique properties. This means that the likelihood of the substitution of each particular residue by another residue during evolution could be different. Briefly, the more similar the physico-chemical properties of two residues, the greater the chance that the substitution will not have a detrimental effect on the protein function and hence on the organism’s fitness. In this method, a generalized measure of the likelihood of amino acid substitutions is used, so that each substitution is given an appropriate score (weight) in sequence comparisons. The Blocks Substitution Matrix (BLOSUM) (HENIKOFF and HENIKOFF 1992) is used to compare the transition scores at pairs of sites. Therefore it can distinguish between sites experiencing correlated radical or conservative replacements in proteins. This is, of course, an oversimplification because the effect of a substitution depends on the structural and functional background where it occurs.

Previously the CAPS method was used to evaluate coevolving sites, for example, within the complete *env* gene of HIV-1 (TRAVERS *et al.* 2007), within prokaryotic membrane proteins (FUCHS *et. al.* 2007) and between chaperones (TRAVERS and FARES 2007). It is worth noting that in the study of FUCHS *et al.* (2007) the number of significantly correlated residues obtained with this method was smaller than with the other prediction algorithms (including ELSC). Also, the results from the study of the *env* HIV-1 gene (TRAVERS *et al.* 2007) could not be confirmed by DA SILVA (2009; see above). His results of the Ohta's partitioning of LD method within the *env* V3 loop of HIV-1 suggest that fitness epistasis is not the cause of observed covariation in this region. Collectively, lack of agreement of the results in the present study from the CAPS method with either the Ohta's method or the ELSC method and the fact that the coevolving sites inferred by the CAPS method as significant in TRAVERS *et al.* (2007), could later be attributed to population subdivision as the main cause of covariation (DA SILVA 2009), may caution against interpreting results of prediction methods such as CAPS. On the other hand, these results could be due to different assumptions of these methods and may still provide useful insights. However, comparative studies such as the one here, highlight the need to evaluate these methods on other datasets to confirm their usefulness to detect true coevolutionary histories.

3. Distribution of natural selection across genes in the Prf complex

Many of the sites in Pto associated with Prf are known to form contact interfaces with the bacterial effectors. Based on this study, these same sites are likely to be experiencing balancing selection. Likewise, residues in domain VIII of Pto form a negative regulatory patch (NRP), and when these sites are mutated, Pto fails to interact with its pathogen ligands (WU *et al.* 2004; XING *et al.* 2007; MUCYN *et al.* 2009; DONG *et al.* 2009). Therefore, this overlap between NRP and AvrPto/AvrPtoB interaction sites in Pto, together with the results in present study, imply that effector binding interferes with inhibitory residues of Pto and disrupts negative regulation to trigger Prf-dependent immune response.

A crystal structure of Fen is not solved, but due its homology to Pto, the functional importance of some of these same sites may be extrapolated. In contrast to Pto, residues in domain X of Fen seem to experience balancing selection. Domain X is highly variable among kinases and seems to be more conserved in subfamilies that share similar functions. The homologous region in Pto is conserved in *S. peruvianum* and previously it was observed that replacements of sites from 243 to 258 disrupted all phenotypes, suggesting the importance of this region for pathogen ligand biding, downstream signaling or correct protein folding (BERNAL *et al.* 2005).

Lack of a crystal structure of Prf makes it difficult to ascertain functionally important amino acids within this protein. Therefore the structure of the Prf protein was modeled in this study. According to the protein model and putative coevolving sites within Prf, the N-terminal domain forms a large molecular arm jutting out from the one side of the protein (Figure 14). The distal region of Prf N-term shows most of the associations with Pto and Fen, consistent with partial epistatic selection. In addition some sites in this region were identified as candidates for balancing selection. Therefore this region is likely involved in contact with Pto and Fen. In contrast, the middle region of Prf N-term domain shows coevolving candidate sites detected by the ELSC method and a few sites consistent with partial epistatic selection between Prf and Pto/Fen. Many of these residues are predicted to coevolve with Fen, but not with Pto and are not polymorphic in all subpopulations. Moreover, some sites of this region are candidates for positive directional selection. This may suggest that these amino acids experience epistatic adaptation with Pto and Fen in some subpopulations. The proximal region shows candidates for coevolution with Pto and Fen, as detected by Ohta's LD partitioning method and the protein sequence based method CAPS. These methods, together, could indicate that this region is dependent on Pto and Fen kinases, but not necessarily due to physical contact.

A key question is what role the N-term domain of Prf plays in interaction with Fen and Pto. Although this domain is a novel sequence of unknown function, it seems likely that the Prf N-term–kinase complex could provide a regulatory node, in addition to the NBARC-LRR portion of Prf. The Prf complex controls immune signaling and Fen/Pto kinase requires Prf for function, although how Prf contributes to Pto or Fen-mediated resistance is unknown, but probably includes control of kinase activity. It is proposed that Fen/Pto kinase operates as a regulatory subunit of Prf and that Pto and Fen are important to Prf stability. Prf contributes to the recognition specificity of the kinase and activates downstream signaling. Thus, the Prf protein complex can be seen as a molecular switch that is targeted by the bacterial effectors (MUCYN *et al.* 2006, 2009).

Mutational analyses suggest that the NBARC-LRR moiety of Prf acts downstream of the N-term/kinase switch, but these nodes do not act independently during signaling. Hence, Fen/Pto and Prf may associate to form a recognition complex characterized by multiple regulatory molecular interactions. This suggests that radical replacements within the interacting interfaces of either Prf protein or Pto/Fen kinase could lead to inactivation or incorrect activation of signaling, which may be detrimental to the host. This can explain the genomic collocation and tightly coadaptation of Prf with the Pto gene family. Another example of tightly linked genes involved in the same physiological processes is the S locus in the *Brassica* species. There two genes that control self-incompatibility, *SRK* and *SLG*, are separated by a maximum distance of 220 kb (BOYES and NASRALLAH 1993). The *SRK* gene encodes a receptor kinase that determines specificity of the stigma in self-incompatibility recognition reactions and *SLG* encodes a glycoprotein that can enhance this process. By analogy to the Prf protein complex, SRK and SLG are proposed to interact (TAKASAKI *et al.* 2000) and like in the plant defense response, this involves a growth restriction of an invading organism (in this case the pollen tube). The components of these recognition complexes may be somewhat unique in that they are dependent on each other for a specific function and they can have a very big influence on the fitness of the progeny, at least in certain environments.

The high degree of *Prf* sequence conservation suggests its ancient origin. *Pto* and *Fen* share their common ancestor between 27.9 and 34.0 mya (ROSE 2002). Since both Fen and Pto require Prf, this suggests that Prf evolved to function with a progenitor of the Pto family (ROSEBROCK *et al.* 2007). How the structurally unrelated *Prf* gene became clustered in the *Pto* gene family is an interesting question. The structural differences between components of these multifunctional loci exclude the possibility that they arose by duplication and divergence of a single ancestral gene. Another pathway that uses both a protein kinase and an LRR–containing protein is the pathway involved in resistance of rice to bacterial blight. There the kinase and the LRR domain are both encoded by a single gene (SONG *et al.* 1995). It is common that functionally interacting proteins that are encoded by separate genes in some organisms are fused in a single polypeptide chain in others. Thus, one evolutionary scenario is that the Prf and Pto family members are derived from an ancestral tomato resistance gene in which these domains were fused (SALMERON *et al.* 1996). On the other hand, some type of transposition or rearrangement brought the two types of genes close to each other and selection favored this system because of the correlation of the genes for the resistance phenotype. In this case, proximity to *Prf* and possible simultaneous expression of proteins with distinct kinases, could be more flexible than fusion with only one kinase (e.g. the closely located *Fen*). This may be required by the host to counteract ongoing pathogen evolution. For instance, Pto homologs may be able to confer resistance to different pathogen isolates (CHANG *et al.* 2002).

Interaction between Prf and Pto/Fen kinase may suggest that Prf residues could contribute to pathogen detection via Avr proteins binding (MUCYN *et al.* 2006; BALMUTH and RATHJEN 2007). In addition to its role in signaling, Prf might serve as a targeting subunit, which acts as organizing platform that recruits both the kinase and the effector (or kinase substrate) to the same complex. The role as a scaffolding adaptor protein is suggested by function of parental proteins used in this study for Prf modeling (see Materials and methods). Furthermore, Prf was proposed as an indirect target of bacterial effectors (NTOUKAKIS *et al.* 2009), thus some of the associations detected in this study may emerge from an indirect link with Avr proteins.

Another explanation for this complicated pattern of Prf interactions could reside in a possibility that Prf molecules form together multimeric structures. Indeed, it was proposed that Prf may form homomultimers and mediates indirect self-association of Pto to build Pto–Prf heterodimers (GUTIERREZ-PULGAR, MUCYN and RATHJEN 2007).

4. Future directions

What is the pattern of linkage disequilibrium in the close vicinity of *Pto* and *Fen*?

Results of this study show that some polymorphic sites at *Pto*, *Fen* and *Prf* experience balancing selection and this is partially consistent with epistatic selection between

Pto/*Fen* and *Prf*, which could maintain variation at these loci. This supports the hypothesis that *Pto*, *Fen* and *Prf* may share similar evolutionary histories (Figure 1B). The extent of coevolution between these genes can be further investigated by studying linkage disequilibrium in the neighborhood of *Pto* and *Fen* towards *Prf*. The rate of LD decay may indicate whether epistatic selection has played an important role in the evolution of the *Pto* cluster. If LD remains high across a few kb flanking *Pto* and *Fen* in the direction of *Prf*, this will support strong epistatic selection as the important force maintaining associations of particular alleles of *Pto*/*Fen* and *Prf*. If instead LD decays within a few kb towards *Prf*, it would be an equally interesting observation. That might indicate that while the potential for epistatic selection between *Pto*/*Fen* and *Prf* exists through the physical and functional linkage required for disease resistance, these genes could have separate evolutionary histories.

What are the functional consequences of amino acid variation in Pto and Prf on recognition specificity and signaling?

Agrobacterium-mediated plant transformation of the *Pto* and *Prf* alleles into plants lacking (or not expressing) functional copies of these genes can be used to link the observed sequence variation at the *Pto* and *Prf* genes to the phenotypic variation in disease resistance. This method allows us to identify those polymorphisms that disrupt protein function and those polymorphisms that are selectively neutral or weakly selected relative to the defined protein function (e.g. avirulence protein mediated pathogen recognition or activation of disease resistance). Most importantly, these transformations will compensate the population genetics studies and determine the mode of epistatic selection between *Pto* and *Prf*.

Interactions between alleles of *Pto* and *Prf* can be functionally tested by co-infiltration of *Prf* alleles with *Pto* alleles derived from these same *S. peruvianum* individuals. The number of different pairwise comparisons could be determined by the range of amino acid variation observed among the alleles of *Pto* and *Prf* from different populations. If only subtle differences among pairwise combinations of the alleles are observed, a quantitative assay for the activation of the disease resistance response, such us measure of electrolyte leakage, can be used. This method provides a more sensitive measure of the activation of the resistance response compared to the more conventional method of simply assaying leaves for macroscopic cell death. Comparisons among the amino acid sequences of functional *Pto* and *Prf* alleles will allow us to determine which sites in the proteins can tolerate amino acid polymorphisms without affecting protein function. The co-infiltration studies of *Pto* and *Prf* alleles would specifically determine which amino acid positions affect the epistatic interactions between these two proteins and will give further insight into how epistatic selection has shaped the evolution of this adaptive trait.

Disease resistance is a complex trait, but one of the most desirable traits, since plant disease remains one of the major restricting factors in plant growth and food production. As more signaling pathway components are identified, progress has been made in understanding the characteristics of immune genes in plants. However, there are still many questions. For example, how do hosts coordinate immune responses when attacked by different pathogens simultaneously? Do these responses share the same signaling pathways? Using population genetics approach to study selected genotypes allows pathway function to be tested in the context of particular genetic backgrounds. More work is needed to define resistance-mediating variants in regions of linkage disequilibrium, not to mention contributions from intergenic regions. As such studies will certainly add more complexity and we need to consider how we can use evolutionary analyses to interpret and understand the function of these variants in disease resistance. It will be necessary to integrate this information from genetic associations with protein–protein interactions to carry out modeling and simulation studies of pathways that are implicated in disease development. Other pathways will become the focus of future functional studies, but it will be challenging to create models of disease in which protein expression levels are important and affect multiple pathways. Finally, the biggest challenge will be to use genetic information to ask questions about the environmental factors that interact with gene products and contribute to disease development and resistance. Hence, further paraphrasing T. DOBZHANSKY (1973), it can be said that nothing in population genetics makes sense except in the light of systems biology.

Appendix

Polymorphic amino acid sites at Pto, Fen and Prf

Sites include candidates for natural selection and coevolution in three populations of *S. peruvianum* (without singletons and doubletons across populations). Samples are ordered from the north to the south of the *S. peruvianum* geographic range (Canta “C”, Nazca “N”, Tarapaca “T”).

Pto	43	46	49	51	70	71	72	88	115	124	132	135	154	168	178	197	200	205	232	273	295
C261A1	H	L	A	V	R	R	Q	T	K	R	P	S	Y	T	A	L	I	L	F	I	L
C261A2	.	.	H	L	.	.	.	.	.	.	.	.	F	I	.	.	V	F	.	L	.
C262A1	.	.	.	.	.	.	.	.	.	.	.	.	F	I	.	.	V	F	L	L	.
C262A2	.	.	H	L	.	.	.	.	E	.	.	.	F	I	.	.	V	.	.	.	.
C263A1	.	.	.	.	.	.	.	.	.	.	.	.	F	I	.	.	V	F	.	L	.
C263A2	.	.	H	L	.	.	.	.	D	S	.	.	.	.	.	.	.	.	.	L	.
C264A1	.	.	H	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
C264A2	.	.	.	.	.	.	.	.	.	.	.	.	F	.	.	.	.	F	.	.	.
C265A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C265A2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C266A1	.	.	H	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C266A2	.	.	H	L	.	.	.	.	.	.	.	.	.	.	T	.	.	F	.	.	.
N251A1	D	F	E	G	S	C	K	I	.	.	L	F	.	I	.	V	V	.	.	.	.
N251A2	.	.	.	.	.	.	K	.	D	.	.	.	.	.	P	.	.	.	.	.	.
N252A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	L	.
N252A2	.	.	H	L	.	.	.	.	.	S	L	F	.	I	.	V	V	.	.	.	.
N253A1	.	.	H	L	.	.	.	.	D	.	.	.	.	I	.	V	V	.	.	.	.
N253A2	.	.	H	.	.	.	.	.	.	.	.	.	F	I	.	V	.	F	.	.	.
N254A1	.	.	.	.	.	.	.	.	.	.	.	.	F	I	.	V	V	.	.	L	.
N254A2	.	.	.	.	.	.	.	.	D	S	.	.	.	.	P	.	.	.	.	.	S
N255A1	.	.	H	L	.	.	.	I	.	.	.	.	F	I	.	.	V	.	.	.	S
N255A2	D	F	E	G	S	C	K	I	.	.	L	F	.	I	.	V	V	.	.	.	S
N256A1	.	.	H	L	.	.	.	I	.	.	.	.	F	I	P	.	.	.	.	L	.
N256A2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	L	.
T7232A1	.	.	H	.	.	.	.	.	Q	.	.	.	.	S	.	.	.	.	.	.	.
T7233A1	.	.	H	.	.	.	.	.	.	.	.	.	F	I	.	.	V	F	.	.	.
T7233A2	.	.	H	.	.	.	.	.	Q	.	.	.	.	S	.	.	.	.	.	.	.
T7234A1	D	F	E	G	S	C	K	I	.	.	.	.	F	I	.	.	V	F	L	.	.
T7234A2	.	.	H	.	.	.	.	.	.	.	.	.	F	I	.	.	V	F	L	.	.
T7235A1	.	.	.	.	.	.	.	.	.	.	.	.	F	I	.	.	V	F	.	.	.
T7235A2	.	.	H	.	.	.	.	.	.	.	.	.	.	S	.	.	.	.	.	.	.
T7236A1	.	.	.	.	.	.	.	.	.	.	.	.	F	I	.	.	V	F	.	.	.
T7236A1	.	.	H	L	.	.	.	.	.	.	.	.	F	I	.	.	V	F	L	.	.
T7237A1	.	.	H	L	.	.	.	I	.	S	.	.	.	.	.	.	.	.	.	.	.
T7237A2	.	.	H	.	.	.	.	.	.	.	.	.	.	S	.	.	.	.	.	.	.
T7238A1	D	F	E	G	S	C	K	I	.	.	.	.	F	I	.	.	V	F	L	.	.
T7238A2	.	.	H	L	.	.	.	.	.	.	.	.	F	I	.	V	V	.	.	.	.
T7239A1	D	F	E	G	S	C	K	I	.	.	.	.	F	I	.	.	V	F	L	.	.
T7239A2	D	F	E	G	S	C	K	I	.	.	.	.	.	I	.	V	V	.	.	L	.
T7240A1	.	.	.	.	.	.	.	.	.	.	.	.	F	I	.	.	V	F	.	.	.
T7240A2	D	F	E	G	S	C	K	I	.	.	.	.	F	I	.	.	V	F	L	.	.
T7241A1	.	.	.	.	.	.	.	.	.	.	.	.	F	I	.	.	V	F	.	.	.
T7241A2	.	.	H	L	.	.	.	.	.	.	.	.	F	I	.	.	V	F	L	.	.

FIGURE A1. Polymorphic amino acid sites at Pto.

Fen	35	44	46	72	73	74	76	78	103	116	136	151	153	241	244	247	255	278	283	291	319
C261A1	D	K	F	N	H	D	R	S	F	Y	M	G	Q	L	S	M	A	M	A	S	P
C261A2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C262A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	L	T	K	I	.	.	T
C262A2	.	.	.	.	·	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.
C263A1	E	N	.	K	P	E	S	G	Y	H	.	.	H	.	L	T	K	I	.	.	T
C263A2	.	.	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	.	V	G	.
C264A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C264A2	E	N	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.
C265A1	E	N	.	K	P	E	S	G	Y	.	.	.	H	.	.	.	T	.	.	.	.
C265A2	E	.	.	K	P	E	S	G	Y	.	.	.	H	.	.	.	K	I	.	.	.
C266A1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C266A2	.	.	.	.	.	.	.	.	.	.	I	.	.	.	L	T	K	I	.	.	T
N251A1	.	.	L	.	.	.	.	.	.	.	.	.	H	.	.	.	T	.	.	.	.
N251A2	.	.	.	.	.	.	.	.	.	.	.	.	H	.	.	.	.	.	.	.	.
N252A1	.	.	L	.	.	.	G	.	.	.	I	.	.	.	.	.	T	.	.	.	.
N252A2	.	.	.	.	.	.	.	.	.	F	I	.	.	.	.	.	.	.	.	.	.
N253A1	.	.	L	.	.	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.
N253A2	.	.	.	.	.	.	.	.	.	.	I	.	.	.	.	.	T	.	.	.	T
N254A1	.	.	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.
N254A2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	L	.	.	I	.	.	T
N255A1	E	.	.	.	.	.	.	.	.	.	I	.	.	.	.	.	T	.	.	.	.
N255A2	.	.	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.
N256A1	.	.	L	.	.	.	.	.	Y	.	I	.	.	.	.	.	.	.	.	.	.
N256A2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	L	.	.	I	.	.	T
T7232A1	E	N	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.
T7232A2	.	.	.	.	.	.	.	.	.	.	I	.	.	.	.	.	T	I	.	.	.
T7233A1	.	.	.	.	.	.	.	.	.	.	I	.	.	I	L	T	K	I	.	.	.
T7233A2	E	N	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.
T7234A1	.	.	.	.	.	.	.	.	.	.	I	.	.	I	L	T	K	I	.	.	.
T7234A2	.	.	.	.	Y	.	.	.	Y	.	I	.	.	.	.	.	.	.	.	.	.
T7235A1	.	.	.	.	.	.	.	.	.	.	I	.	.	.	.	.	T	I	.	.	.
T7235A2	.	.	.	.	.	.	.	.	.	.	I	A	.	.	.	.	T	.	.	.	.
T7236A1	E	N	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.
T7237A1	E	N	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	I	V	G	.
T7237A2	.	.	.	.	.	.	.	.	.	F	I	.	.	.	.	.	.	I	V	G	.
T7238A1	E	N	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.
T7238A2	.	.	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	I	.	.	.
T7239A1	.	.	.	.	.	.	K	.	.	.	I	.	.	.	.	.	.	.	.	.	.
T7239A2	.	.	.	.	.	.	K	.	.	.	I	.	.	I	L	T	K	I	.	.	.
T7240A1	.	.	.	.	.	.	.	.	.	.	I	A	.	.	.	.	T	.	.	.	.
T7240A2	.	.	.	.	.	.	.	.	.	.	I	.	.	.	.	.	T	.	.	.	.
T7241A1	E	N	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.
T7241A2	.	.	.	.	.	.	.	.	.	.	I	A	.	.	.	.	T	.	.	.	.

FIGURE A2. Polymorphic amino acid sites at Fen.

Prf	23	34	62	120	135	156	159	203	212	213	220	233	252	270	277	397	456	487	491	492	510	525	536
C261A1	W	Y	F	Q	L	S	P	T	C	D	I	L	T	R	T	Q	Y	S	K	A	S	S	I
C261A2	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C262A1	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C262A2	.	.	.	R	.	.	.	A	.	H	.	.	K	.	I	.	C	.	.	.	T	F	.
C263A1	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C263A2	.	.	.	.	.	.	.	.	.	H	.	.	.	.	I	.	.	.	.	.	.	.	.
C264A1	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C264A2	.	.	.	.	.	R	S	.	Y	H	.	.	.	.	I	.	C	F	.	.	T	.	.
C265A1	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C265A2	.	.	.	.	.	.	.	.	F	H	.	M	.	.	I	L	C	F	.	.	.	.	.
C266A1	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C266A2	.	.	.	R	.	.	.	A	.	H	.	.	K	.	I	.	C	.	.	.	T	F	.
N251A1	.	.	.	.	.	R	S	.	Y	H	.	.	.	.	I	.	C	F	.	.	T	.	.
N251A2	.	.	.	R	.	.	.	A	.	H	.	.	.	.	I	.	C	.	.	.	T	F	.
N252A1	.	.	.	.	.	R	S	.	Y	H	.	.	.	.	I	.	C	F	.	.	.	.	.
N252A2	.	N	.	R	.	.	.	.	Y	H	.	.	.	.	I	.	C	.	.	.	.	F	.
N253A1	.	.	.	.	.	.	.	A	.	H	.	.	K	.	I	.	C	.	.	.	T	.	.
N253A2	.	.	.	R	.	.	.	A	.	H	.	.	K	.	I	.	C	.	.	.	.	.	.
N254A1	.	.	.	R	.	.	.	.	F	H	.	M	.	.	I	.	.	.	.	.	.	.	.
N254A2	R	H	.	.	L	.	.	.	F	H	.	M	.	.	I	.	.	.	.	.	.	F	.
N255A1	.	.	.	R	.	.	.	.	F	H	.	M	.	.	I	.	.	.	.	.	.	.	.
N255A2	R	H	.	.	L	.	.	.	F	H	.	M	.	.	I	.	.	.	.	.	.	F	.
N256A1	R	H	.	.	L	.	.	.	F	H	.	M	.	.	I	.	.	.	.	.	T	F	.
N256A2	.	.	.	.	L	.	.	.	F	H	.	M	.	.	I	.	.	.	.	.	T	F	.
T7232A1	.	.	.	L	V	.	.	A	.	H	K	.	.	.	I	.	.	.	.	.	T	F	.
T7232A2	R	H	.	R	.	.	.	A	.	H	.	.	K	K	I	.	C	.	.	.	T	.	M
T7233A1	.	.	.	.	.	.	.	.	.	H	.	.	.	.	I	.	.	.	.	.	T	.	M
T7233A2	R	H	.	.	.	.	.	A	.	H	K	.	.	.	I	.	.	.	.	.	T	.	.
T7234A1	.	.	.	R	.	.	.	A	.	H	.	.	K	.	I	.	C	.	.	.	T	F	.
T7234A2	.	.	.	R	.	.	.	A	.	H	.	.	K	.	I	.	C	.	.	.	T	F	.
T7235A1	.	.	.	R	.	.	.	A	.	H	.	.	K	K	I	.	C	.	.	.	T	F	.
T7235A2	R	H	Y	R	.	.	.	A	.	H	K	.	.	.	I	.	.	.	N	S	T	.	.
T7236A1	.	.	.	R	.	.	.	A	.	H	.	.	K	.	I	.	C	.	.	.	T	F	.
T7236A1	R	H	.	R	.	.	.	A	.	H	.	.	K	K	I	.	C	.	.	.	T	F	.
T7237A1	.	.	.	R	.	.	.	A	.	H	.	.	K	.	I	L	C	F	.	.	T	.	.
T7237A2	.	.	.	R	.	.	.	A	.	H	.	.	K	.	I	L	C	F	.	.	T	.	.
T7238A1	.	.	.	R	.	.	.	A	.	H	.	.	K	K	I	.	C	F	.	.	T	.	.
T7238A2	.	.	.	R	.	.	.	A	.	H	.	.	K	K	I	.	C	F	.	.	T	.	.
T7239A1	R	H	.	R	.	.	.	.	F	H	.	M	.	.	I	L	C	F	.	.	T	.	M
T7239A2	R	H	.	R	.	.	.	.	F	H	.	M	.	.	I	L	C	F	.	.	T	.	M
T7240A1	R	H	Y	R	.	.	.	A	.	H	K	.	.	.	I	.	.	.	N	S	T	.	.
T7240A2	R	H	Y	R	.	.	.	A	.	H	K	.	.	.	I	.	.	.	N	S	T	.	.
T7241A1	R	H	Y	R	.	.	.	A	.	H	K	.	.	.	I	.	.	.	N	S	T	.	.
T7241A2	R	H	.	R	.	.	.	A	.	H	.	.	K	K	I	.	C	.	.	.	T	F	.

FIGURE A3. Polymorphic amino acid sites at Prf.

Abbreviations

ACO:	aconitase
Adh:	alcohol dehydrogenase
ATP:	adenosine triphosphate
Avr:	avirulence
BAC:	bacterial artificial chromosome
BF:	Bayes factor
bHLH:	basic helix-loop-helix
BLAST:	basic local alignment search tool
bp:	base pair
CAPS:	coevolution analysis using protein sequences
CC:	coiled-coil
cDNA:	complementary DNA
CTAB:	cetyl trimethyl ammonium bromide
DNA:	deoxyribonucleic acid
EDS1:	enhanced disease susceptibility 1
ELB:	Excoffier-Laval-Balding
ELSC:	explicit likelihood of subset covariation
env:	envelope
EST:	expressed sequence tag
ETI:	effector-triggered immunity
exp:	expected
Fen:	sensitivity to fenthion
FLC:	flowering locus c
FRI:	frigida
HKA:	Hudson-Kreitman-Aguade
HIV-1:	human immunodeficiency virus type 1
HR:	hypersensitive response
hrp:	hypersensitive response and pathogenicity
IDH:	isocitrate dehydrogenase
kb:	kilobasepair
kDa:	kilo Dalton
LD:	linkage disequilibrium
LRR:	leucine reach repeat
MADS:	MCM1-agamous-deficiens-serum response factor
MAMP:	microbe-associated molecular pattern
MAP:	mitogen-activated protein (kinase)
Mbp:	megabasepair
MHC:	major histocompatibility complex
MK:	McDonald-Kreitman
MTI:	MAMP-triggered immunity
mya:	million years ago
NADP:	nicotinamide adenine dinucleotide phosphate
NBARC:	nucleotide binding domain shared by Apaf-1, certain R-gene products, and CED-4 fused to C-terminal leucine-rich repeats
NBS:	nucleotide binding site

NLR:	NOD-like receptor
NLS:	nuclear localization signal
non:	non-synonymous
NPC:	nuclear pore complex
NPR1:	nonexpressor of PR genes
NRP:	negative regulatory patch
obs:	observed
ORF:	open reading frame
PAD4:	phytoalexin deficient 4
PCD:	programmed cell death
PCR:	polymerase chain reaction
PDB:	protein data bank
Pfi:	Prf interactor
Prf:	*Pseudomonas* resistance and fenthion sensitivity
PRR:	pattern recognition receptor
Pst:	*Pseudomonas syringae* pv. tomato
Pth:	Pto homolog
Pti:	Pto interactor
Pto:	resistance to *Pseudomonas syringae* pv. tomato
pv:	pathovar
R:	resistance
RIN4:	RPM1-interacting protein 4
RNA:	ribonucleic acid
ROS:	reactive oxygen species
SAR:	systemic acquired resistance
SD:	solanaceae domain
sil:	silent
SNP:	single nucleotide polymorphism
syn:	synonymous
T3SS:	type three secretion system
TCA:	tricarboxylic acid
TE:	Tris-EDTA
TGRC:	Tomato Genetics Research Center
var:	variety
VIGS:	virus induced gene silencing

Nucleic acid bases:

A: adenine
C: cytosine
G: guanine
T: thymine

Amino acids:

A = Ala:	Alanine	M = Met:	Methionine
C = Cys:	Cysteine	N = Asn:	Asparagine
D = Asp:	Aspartic acid	P = Pro:	Proline
E = Glu:	Glutamic acid	Q = Gln:	Glutamine
F = Phe:	Phenylalanine	R = Arg:	Arginine
G = Gly:	Glycine	S = Ser:	Serine
H = His:	Histidine	T = Thr:	Threonine
I = Ile:	Isoleucine	V = Val:	Valine
K = Lys:	Lysine	W = Trp:	Tryptophan
L = Leu:	Leucine	Y = Tyr:	Tyrosine

References

ABOU JAMRA R, FUERST R, KANEVA R, OROZCO DIAZ G, RIVAS F, *et al.*, 2007. The first genomewide interaction and locus-heterogeneity linkage scan in bipolar affective disorder: Strong evidence of epistatic effects between loci on chromosomes 2q and 6q. Am J Hum Genet. 81: 974-986.

ABRAMOVITCH RB, JANJUSEVIC R, STEBBINS CE and MARTIN GB, 2006. Type III effector AvrPtoB requires intrinsic E3 ubiquitin ligase activity to suppress plant cell death and immunity. PNAS 103: 2851-2856.

ARIE T, TAKAHASHI H, KODAMA M and TERAOKA T, 2007. Tomato as a model plant for plant-pathogen interactions. Plant Biotechnology 24: 135-147.

ARUNYAWAT U, STEPHAN W and STAEDLER T, 2007. Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. Mol Biol Evol. 24: 2310-2322.

ASAI T, STONE JM, HEARD JE, KOVTUN Y, YORGEY P, *et al.*, 2000. Fumonisin B1–induced cell death in *Arabidopsis* protoplasts requires jasmonate-, ethylene-, and salicylate-dependent signaling pathways. Plant Cell 12: 1823-1836.

AUSUBEL FM, 2005. Are innate immune signaling pathways in plants and animals conserved? Nature Immunol. 6: 973-979.

AVERY PJ and HILL WG, 1979. Distribution of linkage disequilibrium with selection and finite population size. Genet. Res. 33: 29-48.

BAINES JF, PARSCH J and STEPHAN W, 2004. Pleiotropic effect of disrupting a conserved sequence involved in a long-range compensatory interaction in the *Drosophila Adh* gene. Genetics 166: 237-242.

BAKKER EG, TOOMAJIAN C, KREITMAN M and BERGELSON J, 2006. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. Plant Cell 18:1803-1818.

BAKKER EG, TRAW MB, TOOMAJIAN C, KREITMAN M and BERGELSON J, 2008. Low levels of polymorphism in genes that control the activation of defense response in *Arabidopsis thaliana*. Genetics 178: 2031-2043.

BALMUTH A and RATHJEN JP, 2007. Genetic and molecular requirements for function of the Pto/Prf effector recognition complex in tomato and *Nicotiana benthamiana*. Plant J. 51: 978-990.

BATESON W, 1909. Heredity and variation in modern lights. In: Darwin and Modern Science (Seward AC, ed.), pp. 85-101, Cambridge University Press.

BATISTA R and OLIVEIRA MM, 2009. Facts and fiction of genetically engineered food. Trends Biotechnol: doi:10.1016/j.tibtech.2009.01.005.

BAUDRY E, KERDELHUE C, INNAN H and STEPHAN W, 2001. Species and recombination effects on DNA variability in the tomato genus. Genetics 158: 1725-1735.

BEAUMONT MA and NICHOLS RA, 1996. Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond B. 263: 1619-1626.

BENDER CL, STONE HE, SIMS JJ and COOKSEY DA, 1987. Reduced pathogen fitness of *Pseudomonas syringae* pv. *tomato* Tn5 mutants defective in coronatine production. Physiol Mol Plant Pathol. 30: 273-283.

BERGELSON J and PURRINGTON CB, 1996. Surveying patterns in the cost of resistance in plants. Am Nat. 148: 536-558.

BERNAL AJ, PAN QL, POLLACK J, ROSE L, KOZIK A, *et al.*, 2005. Functional analysis of the plant disease resistance gene *Pto* using DNA shuffling. J Biol Chem. 280: 23073-23083.

BERNATCHEZ L and LANDRY C, 2003. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J Evol Biol. 16: 363-377.

BHARDWAJ N and LU H, 2005. Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics 21: 2730-2738.

BICKEL PJ, COSMAN PC, OLSHEN RA, SPECTOR PC, RODRIGO AG, *et al.*, 1996. Covariability of V3 loop amino acids. AIDS Res. Hum. Retroviruses 12: 1401-1411

BLACK WC and KRAFSUR ES, 1985. A FORTRAN program for the calculation and analysis of two-locus linkage disequilibrium coefficients. Theor Appl Genet. 70: 491-496.

BLOOM AJ, ZWIENIECKI MA, PASSIOURA JB, RANDALL LB, HOLBROOK NM and ST CLAIR DA, 2004. Water relations under root chilling in a sensitive and tolerant tomato species. Plant Cell Environment 27: 971-979.

BOGDANOVE AJ, 2002. Pto update: recent progress on an ancient plant defence response signalling pathway. Mol Plant Pathol.3: 283-288.

BOMBLIES K, LEMPE J, EPPLE P, WARTHMANN N, LANZ C, *et al.*, 2007. Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. PLoS Biol. 5: e236.

BORODOVSKY M and MCININCH J, 1993. GeneMark: parallel gene recognition for both DNA strands. Comput Chem. 17: 123-133.

BOYES DC and NASRALLAH JB, 1993. Physical linkage of the *SLG* and *SRK* genes at the self-incompatibility locus of *Brassica oleracea.* Mol Gen Genet. 236: 369-373.

BRAUN DM and WALKER JC, 1996. Plant transmembrane receptors: new pieces in the signalling puzzle. Trends Biol Sci. 21: 70-73.

BROWN AHD and FELDMAN MW, 1981. Population structure of multilocus associations. PNAS 78: 5913-5916.

CAICEDO AL, STINCHCOMBE JR, OLSEN KM, SCHMITT J and PURUGGANAN MD, 2004. Epistatic interaction between Arabidopsis *FRI* and *FLC* flowering time genes generates a latitudinal cline in a life history trait. PNAS 101: 15670-15675.

CALDWELL KS and MICHELMORE RW, 2009. *Arabidopsis thaliana* genes encoding defense signaling and recognition proteins exhibit contrasting evolutionary dynamics. Genetics 181: 671-684.

CAO H, BALDINI RL and RAHME LG, 2001. Common mechanisms for pathogens of plants and animals. Annu Rev Phytopathol. 39: 259–284.

CAO ZO, HENZEL WJ and GAO XO, 1996. IRAK: A kinase associated with the interleukin-1 receptor. Science 271: 1128-1131.

CARANTA C, LEFEBVRE V and PALLOIX A, 1997a. Polygenic resistance of pepper to potyviruses consists of a combination of isolate-specific and broad-spectrum quantitative trait loci. Mol Plant Microbe Interact. 10: 872-878.

CARANTA C, PALLOIX A, LEFEBVRE V and DAUBEZE AM, 1997b. QTLs for a component of partial resistance to cucumber mosaic virus in pepper: Restriction of virus installation in host-cells. Theor Appl Genet. 94: 431-438.

CHANG JH, TAI Y-S, BERNAL AJ, LAVELLE DT, STASKAWICZ BJ and MICHELMORE RW, 2002. Functional analysis of the *Pto* resistance gene family in tomato and the identification of a minor resistance determinant in a susceptible haplotype. Mol Plant Microbe Interact. 15: 281-291.

CHARLESWORTH B and CHARLESWORTH D, 1973. Study of linkage disequilibrium in populations of *Drosophila melanogaster*. Genetics 73: 351-359.

CHARLESWORTH B, 1990. Mutation-selection balance and the evolutionary advantage of sex and recombination. Genet Res. 55: 199-221.

CHARLESWORTH D and CHARLESWORTH B, 1975. Theoretical genetics of Batesian mimicry II. Evolution of supergenes. J Theor Biol. 55: 305-324.

CHARLESWORTH D, 2002. Self-incompatibility: How to stay incompatible. Curr Biol. 12: 424-426.

CHEN H, ZOU Y, SHANG Y, LIN H, WANG Y, *et al.*, 2008. Firefly luciferase complementation imaging assay for protein-protein interactions in plants. Plant Physiol. 146: 368-376.

CHEN Y and DOKHOLYAN NV, 2006. The coordinated evolution of yeast proteins is constrained by functional modularity, Trends Genet. 22(8): 416-419.

CHEN Y and STEPHAN W, 2003. Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster Adh* gene. PNAS 100: 11499-11504.

CHETELAT RT and JI Y, 2007. Cytogenetics and evolution. In: Genetic Improvement of Solanaceous Crops (Razdan MK and Mattoo AK, eds.), pp. 77-112, Science Publishers, Enfield, NH .

CHETELAT RT, PERTUZE RA, FAUNDEZ L, GRAHAM EB and JONES CM, 2009. Distribution, ecology and reproductive biology of wild tomatoes and related nightshades from the Atacama Desert region of northern Chile. Euphytica 167: 77-93.

CHEVERUD JM, 2000. Detecting epistasis among Quantitative Trait Loci. In: Epistasis and the Evolutionary Process (Wolf JB, Brodie ED III and Wade MJ, eds.), pp.58-81, Oxford University Press.

CHISHOLM ST, COAKER G, DAY B and STASKAWICZ BJ, 2006. Host-microbe interactions. Shaping the evolution of the plant immune response. Cell 124: 803-814.

COCKERHAM CC and WEIR BS, 1977. Digenic descent measures for finite populations. Genet Res. 30: 121-147.

CORDELL HJ, 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet. 11: 2463-2468.

COUTINHO AM, SOUSA I, MARTINS M, CORREIA C, MORGADINHO T, *et al.*, 2007. Evidence for epistasis between *SLC6A4* and *ITGB3* in autism etiology and in the determination of platelet serotonin levels. Hum Genet. 121: 243-256.

COYNE JA and ORR, 2004. Speciation, Sinauer Associates, Sunderland, MA.

DA CUNHA L, MCFALL AJ and MACKEY D, 2006. Innate immunity in plants, a continuum of layered defenses. Microb Infect. 8: 1372-1381.

DA SILVA J, 2009. Amino acid covariation in a functionally important Human Immunodeficiency Virus type 1 protein region is associated with population subdivision. Genetics 182: 265-275.

DANGL JL and JONES JD, 2001. Plant pathogens and integrated defence responses to infection. Nature 411: 826-833.

DE LAAT W and GROSVELD F, 2003. Spatial organization of gene expression: the active chromatin hub. Chromosome Res. 11: 447-459.

DE TORRES M, MANSFIELD JW, GRABOV N, BROWN IR, AMMOUNEH H, *et al.*, 2006. *Pseudomonas syringae* effector AvrPtoB suppresses basal defence in *Arabidopsis*. Plant J. 47: 368-82.

DEKKER JP, FODOR A, ALDRICH RW and YELLEN G, 2004. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. Bioinformatics 20: 1565-1572.

DELANO WL, 2008. The PyMOL Molecular Graphics System. DeLano Scientific LLC, Palo Alto, CA.

DOBZHANSKY T. 1973. Nothing in biology makes sense except in the light of evolution. Am Biol Teacher 35: 125-129.

DONG J, XIAO F, FAN F, GU L, CANG H, *et al.*, 2009. Crystal structure of the complex between *Pseudomonas* effector AvrPtoB and the tomato Pto kinase reveals both a shared and a unique interface compared with AvrPto-Pto. Plant Cell 21: 1846-1859.

DOYLE JJ and DOYLE JL, 1987. A rapid DNA isolation procedure from small quantities of fresh leaf tissues. Phytochem Bull. 19: 11-15.

Drosophila 12 Genomes Consortium, 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450: 203-218.

DUNHAM RA, 2009. Transgenic fish resistant to infectious diseases, their risk and prevention of escape into the environment and future candidate genes for disease transgene manipulation, Comparative Immunology, Microbiology and Infectious Diseases 32: Genetically modified animals: 139-161.

DURRANT WE and DONG X, 2004. Systemic acquired resistance. Annu Rev Phytopathol. 42: 185-209.

DYKHUIZEN D and HARTL DL, 1980. Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. Genetics 96: 801-817.

EHRENREICH IM and PURUGGANAN MD, 2006. The molecular genetic basis of plant adaptation. Am J Bot. 93: 953-962.

ENDLER JA, 1986. Natural Selection in the Wild, Princeton University Press, NJ.

EXCOFFIER L, LAVAL G and BALDING D, 2003. Gametic phase estimation over large genomic regions using an adaptive window approach. Hum. Genomics 1: 7-19.

EXCOFFIER LG, LAVAL and SCHNEIDER S, 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evol. Bioinformatics Online 1: 47-50.

FARES MA and MCNALLY D, 2006. CAPS: coevolution analysis using protein sequences. Bioinformatics 22: 2821-2822.

FARES MA and TRAVERS SA, 2006. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. Genetics 173: 9-23.

FELDMAN MW, FRANKLIN I and THOMSON GJ, 1974. Selection in complex genetic systems I. The symmetric equilibria of the three-locus symmetric viability model. Genetics 76: 135-162.

FLINT J, BOND J, REES DC, BOYCE AJ, ROBERTS-THOMPSON JM, *et al.*, 1999. Minisatellite mutational processes reduce F_{ST} estimates. Hum Genetics 105: 567-576.

FLUHR R and KAPLAN-LEVY RN, 2002. Plant disease resistance: commonality and novelty in multicellular innate immunity. Curr Top Microbiol Immunol. 270: 23-46.

FOLEY DL, CRAIG JM, MORLEY R, OLSSON CJ, TERENCE DWYER, *et al.*, 2009. Prospects for epigenetic epidemiology. Am J Epidemiol. 169: 389-400.

FOLL M and GAGGIOTTI O, 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. Genetics 180: 977-993.

FOOLAD MR, 2004. Recent advances in genetics of salt tolerance in tomato. Plant Cell Tissue and Organ Culture 76:101-119.

FRANKLIN I and LEWONTIN RC, 1970. Is the gene the unit of selection? Genetics 65: 707-734.

FRASER HB, HIRSH AE, STEINMETZ LM, SCHARFE C and FELDMAN MW, 2002. Evolutionary rate in the protein interaction network. Science 296: 750-752.

FRASER HB, HIRSH AE, WALL DP and EISEN MB, 2004. Coevolution of gene expression among interacting proteins. PNAS 101: 9033-9038.

FREDERICK RD, THILMONY RL, SESSA G and MARTIN GB, 1998. Recognition specificity for the bacterial avirulence protein AvrPto is determined by Thr-204 in the activation loop of the tomato Pto kinase. Mol Cell 2: 241-245.

FRIEDMAN AR and BAKER BJ, 2007. The evolution of resistance genes in multi-protein plant resistance systems. Curr Opin Genet Dev. 17: 1-7.

FUCHS A, MARTIN-GALIANO AJ , KALMAN M, FLEISHMAN S, BEN-TAL N and FRISHMAN D, 2007. Co-evolving residues in membrane proteins. Bioinformatics 23: 3312-3319.

GARNIER-GERE P and DILLMANN C, 1992. A computer program for testing pairwise linkage disequilibria in subdivided populations. J Hered. 83: 239.

GEHRIG H, SCHUSSLER A and KLUGE M, 1996. *Geosiphon pyriforme*, a fungus forming endocytobiosis with *Nostoc* (cyanobacteria) is an ancestral member of the *Glomales*: evidence by SSU rRNA analysis. J Mol Evol. 43: 71-81.

GILBERT PB, NOVITSKY V and ESSEX M, 2005. Covariability of selected amino acid positions for HIV type 1 subtypes C and B. AIDS Res. Hum. Retroviruses 21: 1016-1030.

GILLESPIE JH, 1991. The Causes of Molecular Evolution. Oxford University Press.

GIMENEZ-IBANEZ S, HANN DR, NTOUKAKIS V, PETUTSCHNIG E, LIPKA V and RATHJEN JP, 2009. AvrPtoB targets the LysM receptor kinase CERK1 to promote bacterial virulence on plants. Curr Biol. 19: 423-429.

GLOOR GB, MARTIN LC, WAHL LM and DUNN SD, 2005. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 44: 7156-7165.

GOEHRE V, SPALLEK T, HAWEKER H, MERSMANN S, MENTZEL T, *et al.*, 2008. Plant pattern recognition receptor FLS2 is directed for degradation by the bacterial ubiquitin ligase AvrPtoB. Curr Biol. 18: 1824-1832.

GOH C-S, BOGAN AA, JOACHIMIAK M, WALTHER D and COHEN FE, 2000. Co-evolution of proteins with their interaction partners. J Mol Biol. 299: 283-293.

GOODNIGHT C J, 2000. Modeling gene interaction in structured populations. In: Epistasis and the Evolutionary Process (Wolf JB, Brodie ED III and Wade MJ, eds.), pp.129-145, Oxford University Press.

GORDILLO LF, STEVENS MR, MILLARD MA and GEARY BD, 2008. Screening two *Lycopersicon peruvianum* collections for resistance to tomato spotted wilt virus. Plant Disease 92: 694-704.

GREENBERG JT and AUSUBEL FM, 1993. *Arabidopsis* mutants compromised for the control of cellular damage during pathogenesis and aging. Plant J. 4: 327-341.

GREGERSEN JW, KRANC KR, KE X, SVENDSEN P, MADSEN LS, *et al.*, 2006. Functional epistasis on a common MHC haplotype associated with multiple sclerosis. Nature 443: 574-577.

GURR SJ and RUSHTON PJ, 2005. Engineering plants with increased disease resistance: what are we going to express? Trends Biotechnol. 23: 275-282.

GUTIERREZ-PULGAR JR, MUCYN T and RATHJEN JP, 2007. Functional cooperativity between Prf-Pto heterodimers in host resistance of tomato against *Pseudomonas syringae*. In: XIII International Congress on Molecular Plant-Microbe Interactions. July 21-27, 2007, Sorrento, Italy. Book of Abstracts : PS 1-84.

HAGENBLAD J and NORDBORG M, 2002. Sequence variation and haplotype structure surrounding the flowering time locus *FRI* in *Arabidopsis thaliana*. Genetics 161: 289-298.

HAKES L, LOVELL S, OLIVER SG and ROBERTSON DL, 2007. Specificity in protein interactions and its relationship with sequence diversity and coevolution. PNAS 104: 7999-8004.

HALPERIN I, WOLFSON H and NUSSINOV R, 2006. Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins 63: 832-845.

HALTERMAN DA, 1999. Dissertation: Characterization of the Fenthion response in tomato and the identification of genes that encode Fen-interacting proteins. Purdue University.

HAMMOND-KOSACK KE and JONES JD, 1997. Plant disease resistance genes. Annu Rev Plant Physiol Plant Mol Biol. 48: 575-607.

HANKS SK and HUNTER T, 1995. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. FASEB J. 9: 576-596.

HEDRICK PW, 1987. Gametic disequilibrium measures: proceed with caution. Genetics 117: 331-341.

HEDRICK PW, 1999. Balancing selection and MHC. Genetica 104: 207-214.

HENIKOFF S and HENIKOFF JG, 1992. Amino acid substitution matrices from protein blocks. PNAS 89: 10915-10919.

HERITAGE J, 2005. Transgenes for tea? Trends Biotechnol. 23: 17-21.

HILL WG and ROBERTSON A, 1968. Linkage disequilibrium in finite populations. Theor Appl Genet. 38: 226-231.

HILL WG and WEIR BS, 1988. Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol. 33: 54-78.

HILL WG, 1975. Linkage disequilibrium among multiple neutral alleles produced by mutation in a finite population. Theor Popul Biol. 8: 117-126.

HILL WG, 1976. Non-random association of neutral linked genes in finite populations, pp. 339–376. In: Population Genetics and Ecology (Karlin S and Nevo E, eds.). Academic Press, NY.

HUDSON RR, 1990. Gene genealogies and the coalescent process. In: Oxford Surveys in Evolutionary Biology, vol. 7 (Futujma D and Antonovics J, eds), pp. 1-44. Oxford University Press.

HUDSON RR, 2001. Two-locus sampling distributions and their application. Genetics 159: 1805-1817.

HUDSON RR, SLATKIN M and MADDISON WP, 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583-589.

HUGHES AL and NEI M, 1989. Nucleotide substitution at major histocompatibility complex class II loci: Evidence for overdominant selection. PNAS 86: 958-962.

HURST LD, WILLIAMS EJ and PAL C, 2002. Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet. 18: 604-606.

INNAN H and STEPHAN W, 2001. Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. Genetics 159: 389-399.

IZAGUIRRE MM, SCOPEL AL, BALDWIN IT and BALLARE CL, 2003. Convergent responses to stress: solar ultraviolet-B radiation and *Manduca sexta* herbivory elicit overlapping transcriptional responses in field-grown plants of *Nicotiana longiflora*. Plant Physiol. 132: 1755-1767.

JANJUSEVIC R, ABRAMOVITCH RB, MARTIN GB and STEBBINS CE, 2006. A bacterial inhibitor of host programmed cell death defenses is an E3 ubiquitin ligase. Science 311: 222-226.

JEFFREYS H, 1961. Theory of Probability, 3rd ed. Oxford University Press

JIA Y, LOH YT, ZHOU J and MARTIN GB, 1997. Alleles of *Pto* and *Fen* occur in bacterial speck-susceptible and fenthion-insensitive tomato cultivars and encode active protein kinases. Plant Cell 9: 61-73.

JIN T, BOKAREWA M, FOSTER T, MITCHELL J, HIGGINS J and TARKOWSKI A, 2004. *Staphylococcus aureus* resists human defensins by production of staphylokinase, a novel bacterial evasion mechanism. J Immunol. 172: 1169-1176.

JOHANSON U, WEST J, LISTER C, MICHAELS S, AMASINO R and DEAN C, 2000. Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. Science 290: 344-347.

JONES JB, 1991. Bacterial speck. In: Compendium of Tomato Diseases (Jones JB, Jones JP, Stall RE and Zitter TA, eds.), pp. 26-27, APS Press, St. Paul, MN.

JONES JDG and DANGL JF, 2006. The plant immune system. Nature 444: 323-329.

JORON M, PAPA R, BELTRAN M, CHAMBERLAIN N, MAVAREZ J, *et al.*, 2006. A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. PLoS Biol. 4: e303.

KARLIN S and FELDMAN MW, 1970. Linkage and selection: two locus symmetric viability model. Theor Popul Biol. 1: 39-71.

KARLIN S and FELDMAN MW. 1978. Simultaneous stability of D=0 and D≠0 for multiplicative viabilities at two loci. Genetics 90: 813-825.

KELLY JK and WADE MJ, 2000. Molecular evolution near a two-locus balanced polymorphism. J. Theor. Biol. 204: 83-101.

KELLY JK, 1997. A test of neutrality based on interlocus associations. Genetics 146: 1197-1206.

KELLY JK, 2000. Epistasis, linkage, and balancing selection. In: Epistasis and the Evolutionary Process (Wolf JB, Brodie ED III and Wade MJ, eds.), pp.146-157, Oxford University Press.

KIM MG, DA CUNHA L, MCFALL AJ, BELKHADIR Y, DEBROY S, *et al.*, 2005. Two *Pseudomonas syringae* type III effectors inhibit RIN4-regulated basal defense in *Arabidopsis*. Cell 121: 749-759.

KIMURA M, 1985. The role of compensatory neutral mutations in molecular evolution. J Genet. 64: 7-19.

KIRBY DA and STEPHAN W, 1996. Multi-locus selection and the structure of variation at the white gene of *Drosophila melanogaster*. Genetics 144: 635-645.

KIRBY DA, MUSE SV and STEPHAN W, 1995. Maintenance of pre-mRNA secondary structure by epistatic selection. PNAS 92: 9047-9051.

KONDRASHOV AS, 1994. Muller's ratchet under epistatic selection. Genetics 136: 1469-1473.

KORBER BT, FARBER RM, WOLPERT DH and LAPEDES AS, 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. PNAS 90: 7176-7180.

KOVER PX and CAICEDO AL, 2001. The genetic architecture of disease resistance in plants and the maintenance of recombination by parasites. Mol Ecol. 10: 1-17.

KRUEGER J, THOMAS CM, GOLSTEIN C, DIXON MS, SMOKER M, *et al.*, 2002. A tomato cysteine protease required for *Cf-2*-dependent disease resistance and suppression of autonecrosis. Science 296: 744-747.

LANGLEY CH, TOBARI YN and KOJIMA KI, 1974. Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Genetics 78: 921-936.

LAZZARO BP, SACKTON TB and CLARK AG, 2006. Genetic variation in *Drosophila melanogaster* resistance to infection: A comparison across bacteria. Genetics 174: 1539-1554.

LAZZARO BP, SCEURMAN BK and CLARK AG, 2004. Genetic basis of natural variation in *D. melanogaster* antibacterial immunity. Science 303: 1873-1876.

LE CORRE V, ROUX F and REBOUD X, 2002. DNA polymorphism at the *FRIGIDA* gene in *Arabidopsis thaliana*: extensive nonsynonymous variation is consistent with local selection for flowering time Mol Biol Evol. 19: 1261-1271.

LEE JM and SONNHAMMER ELL, 2003. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res 13: 875-882.

LEGNANI R, GOGNALONS P, SELASSIE KG, MARCHOUX G, MORETTI A and LATERROT H, 1996. Identification and characterization of resistance to tobacco etch virus in *Lycopersicon* species. Plant Disease 80: 306-309.

LEMPE J, BALASUBRAMANIAN S, SURESHKUMAR S, SINGH A, SCHMID M, *et al.*, 2005. Diversity of flowering responses in wild *Arabidopsis thaliana* strains. PLoS Genet. 1: 109-118.

LERCHER MJ, BLUMENTHAL T and HURST LD, 2003. Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. Genome Res. 13: 238-243.

LEWONTIN R, 1974. The Genetic Basis of Evolutionary Change. Columbia University Press, NY.

LEWONTIN RC and KOJIMA K, 1960. The evolutionary dynamics of complex polymorphisms. Evolution 14: 458-472.

LI W and NEI M, 1974. Stable linkage disequilibrium without epistasis in subdivided populations. Theor Popul Biol. 6: 173-183.

Li WH, 1997. Molecular Evolution. Sinauer, Sunderland.

LIBRADO P and ROZAS J, 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452.

LINCOLN MR, MONTPETIT A, CADER MZ, SAARELA J, DYMENT DA, *et al.*, 2005. A predominant role for the HLA class II region in the association of the MHC region with multiple sclerosis. Nature Genet. 37: 1108-1112.

LIU J, ELMORE JM, FUGLSANG AT, PALMGREN MG, STASKAWICZ BJ and COAKER G, 2009. RIN4 functions with plasma membrane H^+-ATPases to regulate stomatal apertures during pathogen attack. PLoS Biol. 7: e1000139.

LOH YT and MARTIN GB, 1995. The *Pto* bacterial resistance gene and the *Fen* insecticide sensitivity gene encode functional protein kinases with serine/threonine specificity. Plant Physiol. 108: 1735-1739.

LU Y and RAUSHER MD, 2003. Evolutionary rate variation in anthocyanin pathway genes. Mol Biol Evol. 20: 1844-1853.

LUO Y, CALDWELL KS, WROBLEWSKI T, WRIGHT ME and MICHELMORE RW, 2009. Proteolysis of a negative regulator of innate immunity is dependent on resistance genes in tomato and *Nicotiana benthamiana* an induced by multiple bacterial effectors. Plant Cell 21: 2458-2472.

LYNCH M, 2007. The Origins of Genome Architecture. Sinauer Associates, Sunderland, MA.

MACKEY D, HOLT BF, WIIG A and DANGL JL, 2002. RIN4 interacts with *Pseudomonas syringae* type III effector molecules and is required for RPM1-mediated resistance in *Arabidopsis*. Cell 108: 743-754.

MARTIN GB, BROMMONSCHENKEL SH, CHUNWONGSE J, FRARY A, GANAL MW, *et al.*, 1993. Map-based cloning of a protein kinase gene conferring disease resistance in tomato. Science 262: 1432-1436.

MARTIN GB, FRARY A, WU TY, BROMMONSCHENKEL S, CHUNWONGSE J, *et al.*, 1994. A member of the tomato *Pto* gene family confers sensitivity to fenthion resulting in rapid cell death. Plant Cell 6: 1543-1552.

MARTINON F and TSCHOPP J, 2005. NLRs join TLRs as innate sensors of pathogens. Trends Immunol. 26: 447-454.

MCCARTER SM, JONES JB, GITAITIS RD and SMITELY DR, 1983. Survival of *Pseudomonas syringae* pv. *tomato* in association with the tomato seed soil host tissue and epiphytic weed host in Georgia. Phytopathology 73: 1393-1398.

MCDONALD JH and KREITMAN M, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351: 652-654.

MCDOWELL JM and SIMON SA, 2006. Recent insights into R gene evolution. Mol Plant Path. 7: 437-448.

McVean G, Awadalla P and Fearnhead P, 2002. A coalescent-based method for detecting and estimating recombination rates from gene sequences. Genetics 160: 1231-1241.

Meyer D and Thomson G, 2001. How selection shapes variation of the human major histocompatibility complex: a review. Ann Hum Genet. 2001 65: 1-26.

Meyers BC, Kaushik S and Nandety RS, 2005. Evolving disease resistance genes. Curr Opin Plant Biol. 8: 129-134

Mezey JG, Nuzhdin SV, Fangfei Y and Jones CD, 2008. Coordinated evolution of co-expressed gene clusters in the *Drosophila* transcriptome. BMC Evol Biol. 8: 2.

Michaels SD, Bezerra IC and Amasino RM, 2004. *FRIGIDA*-related genes are required for the winter-annual habit in *Arabidopsis*. PNAS 101: 3281-3285.

Michelmore RW and Meyers BC, 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. 8: 1113-1130.

Miller JC and Tanksley SD, 1990. RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. Theor Appl Genet. 80: 437-448.

Moore JH and Williams SM, 2005. Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. Bioessays 27: 637-646.

Mucyn TS, Clemente A, Andriotis VME, Balmuth AL, Oldroyd GED, *et al.*, 2006. The tomato NBARC-LRR protein Prf interacts with Pto kinase *in vivo* to regulate specific plant immunity. Plant Cell 18: 2792-2806.

Mucyn TS, Wu A-J, Balmuth AL, Arasteh JM and Rathjen JP, 2009. Regulation of tomato Prf by Pto-like protein kinases. Mol Plant Microbe Interact. 22: 391-401.

Mygind PH, Fischer RL, Schnorr KM, Hansen MT, Soenksen CP, *et al.*, 2005. Plectasin is a peptide antibiotic with therapeutic potential from a saprophytic fungus, Nature 437: 975-980.

Nakazato T, Bogonovich M and Moyle LC, 2008. Environmental factors predict adaptive phenotypic differentiation within and between two wild Andean tomatoes. Evolution 62: 744-792.

Navarro A and Barton NH, 2002.The effects of multilocus balancing selection on neutral variability. Genetics 161: 849-863.

Nei M, 1987. Molecular Evolutionary Genetics. Columbia University Press, NY.

Nordborg M, 2000. Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. Genetics 154: 923-929.

Ntoukakis V, Mucyn TS, Gimenez-Ibanez S, Chapman HC, Gutierrez JR, *et al.*, 2009. Host inhibition of a bacterial virulence effector triggers immunity to infection. Science 324: 784-787.

Nuernberger T and Brunner F, 2002. Innate immunity in plants and animals: emerging parallels between the recognition of general elicitors and pathogen-associated molecular patterns. Curr Opin Plant Biol. 5: 318-324.

Nuernberger T and Lipka V, 2005. Non-host resistance in plants: new insights into an old phenomenon. Mol Plant Pathol. 6: 335-345.

NUERNBERGER T, BRUNNER F, KEMMERLING B and PIATER L, 2004. Innate immunity in plants and animals: striking similarities and obvious differences. Immunol Rev. 198: 249-266.

OHTA T and KIMURA M, 1969a. Linkage disequilibrium due to random genetic drift. Genet Res. 13: 47-55.

OHTA T and KIMURA M, 1969b. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics 63: 229-238.

OHTA T, 1982a. Linkage disequilibrium due to random genetic drift in finite subdivided populations. PNAS 79: 1940-1944.

OHTA T, 1982b. Linkage disequilibrium with the island model. Genetics 101: 139-155.

O'NEIL P, 1999. Selection on flowering time: an adaptive fitness surface for nonexistent character combinations. Ecology 80: 806-820.

PARSCH J, TANDA S and STEPHAN W, 1997. Site-directed mutations reveal long-range compensatory interactions in the *Adh* gene of *Drosophila melanogaster.* PNAS 94: 928-933.

PAZOS F and VALENCIA A, 2002. *In silico* two-hybrid system for the selection of physically interacting protein pairs. Proteins 47: 219-227.

PELLEGRINI M, MARCOTTE EM, THOMPSON MJ, EISENBERG D and YEATES TO, 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. PNAS 96: 4285-4288.

PERALTA IE and SPOONER DM, 2001. Granule-bound starch synthase (GBSSI) gene phylogeny of wild tomatoes (*Solanum* L. section *Lycopersicon* Mill. Wettst. subsection *Lycopersicon*). Am J Bot. 88: 1888-1902.

PERALTA IE, KNAPP SK and SPOONER DM, 2005. New species of wild tomatoes (*Solanum* section *Lycopersicon*: *Solanaceae*) from northern Peru. Syst Bot. 30: 424-434.

PERALTA IE, SPOONER DM and KNAPP S, 2008. Taxonomy of wild tomatoes and their relatives (*Solanum* sect. *Lycopersicoides*, sect. *Juglandifolia*, sect. *Lycopersicon*; *Solanaceae*). Syst Bot Monogr. 84: 1-186.

PETERS AD and LIVELY CM, 2000. Epistasis and the maintenance of sex. In: Epistasis and the Evolutionary Process (Wolf JB, Brodie ED III and Wade MJ, eds.), pp.99-112, Oxford University Press.

PHILLIPS PC, 1996. Waiting time for a compensatory mutation: phase zero of the shifting-balance process. Genet Res. 67: 271-283.

PICO B, HERRAIZ J and NUEZ F, 2000. *Lycopersicon chilense*-derived bridge lines for introgressing *L. peruvianum* traits into the *esculentum* genome. Report of the Tomato Genetics Cooperative 50: 30-32.

PICO B, HERRAIZ J, RUIZ JJ and NUEZ F, 2002. Widening the genetic basis of virus resistance in tomato, Scientia Hortic. 94: 73-89.

POON AF, LEWIS FI, POND SL and FROST SD, 2007. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. PLoS Comput. Biol. 3: e231.

POYSA V, 1990. The development of bridge lines for interspecific gene transfer between *Lycopersicon esculentum* and *Lycopersicon peruvianum*. Theor Appl Genet. 79: 187-192.

PRESGRAVES DC and STEPHAN W, 2007. Pervasive adaptive evolution among interactors of the *Drosophila* hybrid inviability gene, *Nup96*. Mol Biol Evol. 24: 306-314.

PRESGRAVES DC, 2003. A fine-scale genetic analysis of hybrid incompatibilities in *Drosophila*. Genetics 163: 955-972.

PRESTON GM, 2000. *Pseudomonas syringae* pv. *tomato*: the right pathogen, of the right plant, at the right time. Mol Plant Pathol. 1: 263-275.

PRITHIVIRAJ B, WEIR T, BAIS HP, SCHWEIZER HP and VIVANCO JM, 2005. Plant models for animal pathogenesis. Cell Microbiol. 7: 315-324.

RAMSAY H, RIESEBERG LH and RITLAND K, 2009. The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. Mol Biol Evol. 26: 1045-1053.

RATHJEN JP, CHANG JH, STASKAWICZ BJ and MICHELMORE RW, 1999. Constitutively active Pto induces a Prf-dependent hypersensitive response in the absence of avrPto. EMBO 18: 3232-3240.

RAUSHER M, LU Y and MEYER K, 2008. Variation in constraint versus positive selection as an explanation for evolutionary rate variation among anthocyanin genes. J Mol Evol. 67: 137-144.

RAUSHER MD, MILLER RE and TIFFIN P, 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol Biol Evol. 16**:** 266-274.

RAWSON PD and BURTON RS, 2002. Functional coadaptation between cytochrome *c* and cytochrome *c* oxidase within allopatric populations of a marine copepod. PNAS 99: 12955-12958.

RICE SH, 2000. The evolution of developmental interactions: epistasis, canalization and integration. In: Epistasis and the Evolutionary Process (Wolf JB, Brodie ED III and Wade MJ, eds.), pp. 82-98, Oxford University Press.

RICK CM and CHETELAT RT, 1995. Utilization of related wild species for tomato improvement. Acta Hortic. 412: 21-38.

RICK CM and FOBES JF, 1975. Allozyme variation in the cultivated tomato and closely related species. Bulletin of the Torrey Botanical Club 102: 376-384.

RICK CM, 1963. Barriers to inbreeding in *Lycopersicon peruvianum*. Evolution 17: 216-232.

RICK CM, 1973. Potential genetic resources in tomato species: clues from observations in native habitats. In: Genes, Enzymes and Populations (Srb AM, ed), pp. 255-269, Plenum Press, NY.

RICK CM, 1976a. Natural variability in wild species of *Lycopersicon* and its bearing on tomato breeding. Genet Agraria 30: 249-259.

RICK CM, 1976b. Tomato, *Lycopersicon esculentum* (Solanaceae). In: Evolution of Crop Plants (Simmonds NW, ed.), pp. 268-273, Longman, London.

RICK CM, 1979a. Biosystematic studies in *Lycopersicon* and closely related species of *Solanum.* Biol Taxon Solanaceae 7: 667-679.

RICK CM, 1979b. Potential improvement of tomatoes by controlled introgression of genes from wild species. In: Proceedings of the Conference on Broadening the Genetic Base of Crops, pp. 167-173, Pudoc, Wageningen, The Netherlands.

RICK CM, 1986. Reproductive isolation in the *Lycopersicon peruvianum* complex. In: *Solanaceae*: Biology and Systematics (D'Arcy WG, ed.), pp. 477-495. Columbia University Press, NY.

RICK CM, 1991. Tomato paste: a concentrated review of genetic highlights from the beginnings to the advent of molecular genetics. Genetics 128: 1-5.

RIELY BK and MARTIN GB, 2001. Ancient origin of pathogen recognition specificity conferred by the tomato disease resistance gene *Pto*. PNAS 98: 2059-2064.

ROBATZEK S, BITTEL P, CHINCHILLA D, KOECHNER P, FELIX G, *et al.*, 2007. Molecular identification and characterization of the tomato flagellin receptor LeFLS2, an orthologue of *Arabidopsis* FLS2 exhibiting characteristically different perception specificities. Plant Mol Biol. 64: 539-547.

ROONEY HCE, VAN'T KLOOSTER JW, VAN DER HOORN RAL, JOOSTEN M, JONES JDG and DE WIT PJGM, 2005. Cladosporium Avr2 inhibits tomato *Rcr3* protease required for *Cf-2*-dependent disease resistance. Science 308: 1783-1786.

ROSE LE, 2002. Dissertation: The population genetics and functional analysis of the *Pto* disease resistance gene in *Lycopersicon spp*. and the *RPP13* gene in *Arabidopsis thaliana*. University of California, Davis.

ROSE LE, BITTNER-EDDY PD, LANGLEY CH, HOLUB EB, MICHELMORE RW and BEYNON JL, 2004. The maintenance of extreme amino acid diversity at the disease resistance gene, *RPP13*, in *Arabidopsis thaliana*. Genetics 166: 1517-1527.

ROSE LE, LANGLEY CH, BERNAL AJ and MICHELMORE RW, 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 LE, MICHELMORE RW and LANGLEY CH, 2007. Natural variation in the *Pto* disease resistance gene within species of wild tomato (*Lycopersicon*). II. Population genetics of *Pto*. Genetics 175: 1307-1319.

ROSEBROCK TC, ZENG L, BRADY JJ, ABRAMOVITCH RB, XIAO F and MARTIN GB, 2007. A bacterial E3 ubiquitin ligase targets a host protein kinase to disrupt plant immunity. Nature 448: 370-375.

RUBIN GM, YANDELL MD, WORTMAN JR, GABOR MIKLOS GL, NELSON CR, *et al.*, 2000. Comparative genomics of the eukaryotes. Science 287: 2204-2215.

RUDICH J and LUCHINSKY U, 1986. Water economy. In: The Tomato Crop: A Scientific Basis for Improvement (Atherton JG and Rudich J, eds.), pp. 335-368. Chapman and Hall, NY and London.

SACRISTAN S and GARCIA-ARENAL F, 2008. The evolution of virulence and pathogenicity in plant pathogen populations. Mol Plant Path. 9: 369-384.

SALMERON J, OLDROYD G, ROMMENS C, SCOFIELD S, KIM H-S, *et al.*, 1996. Tomato *Prf* is a member of the leucine-rich repeat class of plant disease resistance genes and lies embedded within the *Pto* kinase gene cluster. Cell 86: 123-133.

SANCHEZ-DONAIRE A, ENCINA CL, CUARTERO J and GUERRA-SANZ JM, 2000. Increased efficiency of interspecific hybrids by embryo rescue in crosses between *L. esculentum* and *L. peruvianum*. Report of the Tomato Genetics Cooperative 50: 35-37.

SANCHEZ-MAZAS A, DJOULAH S, BUSSON M, LE MONNIER DE GOUVILLE I, POIRIER JC, *et al.*, 2000. A linkage disequilibrium map of the MHC region based on the analysis of 14 loci haplotypes in 50 French families. Eur J Hum Genet. 8: 33-41.

SATO T, NISHIO T, KIMURA R, KUSABA M, SUZUKI G, *et al.*, 2002. Coevolution of the S-locus genes SRK, SLG and SP11/SCR in *Brassica oleracea and B. rapa*. Genetics 162: 931-940.

SCHAEFFER SW and MILLER EL, 1993. Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics 135: 541-552.

SCHLENKE TA and BEGUN DJ, 2005. Linkage disequilibrium and recent selection at three immunity receptor loci in *Drosophila simulans*. Genetics 169: 2013-2022.

SCHLOSSER G and WAGNER P, 2008. A simple model of co-evolutionary dynamics caused by epistatic selection. J Theor Biol. 250: 48-65.

SCOFIELD SR, TOBIAS CM, RATHJEN JP, CHANG JH, LAVELLE DT, *et al.*, 1996. Molecular basis of gene-for-gene specificity in bacterial speck disease of tomato. Science 274: 2063-2065.

SCOTT SJ and JONES RA, 1982. Low temperature seed germination of *Lycopersicon* species evaluated by survival analysis. Euphytica 31: 869-883.

SESSA G and MARTIN GB, 2000. Signal recognition and transduction mediated by the tomato Pto kinase: A paradigm of innate immunity in plants. Microbes and Infection 2: 1591-1597.

SHAN L, HE P, LI J, HEESE A, PECK SC, *et al.*, 2008. Bacterial effectors target the common signaling partner BAK1 to disrupt multiple MAMP receptor-signaling complexes and impede plant immunity. Cell Host Microbe 4: 17-27.

SHELDON CC, BURN JE, PEREZ PP, METZGER J, EDWARDS JA, *et al.*, 1999. The FLF MADS Box gene: a repressor of flowering in *Arabidopsis* regulated by vernalization and methylation. Plant Cell 11: 445-458.

SHELTON CA and WASSERMAN SA, 1993. *Pelle* encodes a protein kinase required to establish dorsoventral polarity in the *Drosophila* embryo. Cell 72: 515-525.

SHEN QH and SCHULZE-LEFERT P, 2007. Rumble in the nuclear jungle: Compartmentalization, trafficking, and nuclear action of plant immune receptors. EMBO J. 26: 4293-4301.

SINGER GA, LLOYD AT, HUMINIECKI LB and WOLFE KH, 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol Biol Evol. 22: 767-775.

SINGH KB, FOLEY RC and ONATE-SANCHEZ L. 2002. Transcription factors in plant defense and stress responses. Curr Opin Plant Biol. 5: 430-436.

SLATKIN M, 1975. Gene flow and selection in a 2-locus system. Genetics 81: 787-802.

SOCOLICH M, LOCKLESS SW, RUSS WP, LEE H, GARDNER KH and RANGANATHAN R, 2005. Evolutionary information for specifying a protein fold. Nature 437: 512-518.

SONG WY, WANG GL, CHEN LL, KIM HS, PI LY, *et al.*, 1995. A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. Science 270: 1804-1806.

STEIN JC, DIXIT R, NASRALLAH ME and NASRALLAH JB, 1996. SRK, the stigma-specific S locus receptor kinase of *Brassica*, is targeted to the plasma membrane in transgenic tobacco. Plant Cell 8: 429-445.

STEPHAN W and LANGLEY CH, 1998. DNA polymorphism in *Lycopersicon* and crossing-over per physical length. Genetics 150: 1585-1603.

STEPHAN W, 1996. The rate of compensatory evolution. Genetics 144: 419-426.

STEVENS MA and RICK CM, 1986. Genetics and breeding. In: The Tomato Crop: A Scientific Basis for Improvement (Atherton JG and Rudich J, eds.), pp. 35-109. Chapman and Hall, NY and London.

STINCHCOMBE JR, WEINIG C, UNGERER M, OLSEN KM, MAYS C, *et al.*, 2004. A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene *FRIGIDA*. PNAS 101: 4712-4717.

STOLC V, GAUHAR Z, MASON C, HALASZ G, VAN BATENBURG MF, *et al.*, 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science 306: 655-660.

SUEL GM, LOCKLESS SW, WALL MA and RANGANATHAN R, 2003. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nature Struct Biol. 10: 59-69.

SVEJGAARD A, 2008. The immunogenetics of multiple sclerosis. Immunogenetics 60: 275-286.

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

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

TAKAHASI KR and TAJIMA F, 2005. Evolution of coadaptation in a two-locus epistatic system. Evolution 59: 2324-2332.

TAKAHASI KR, 2007. Evolution of coadaptation in a subdivided population. Genetics 176: 501-11.

TAKAHATA N and SATTA Y, 1998. Footprints of intragenic recombination at HLA loci. Immunogenetics 47: 430-441.

TAKAHATA N, SATTA Y and KLEIN J, 1992. Polymorphism and balancing selection at major histocompatibility complex loci. Genetics 130: 925-938.

TAKASAKI T, HATAKEYAMA K, SUZUKI G, WATANABE M, ISOGAI A and HINATA K, 2000. The S receptor kinase determines self-incompatibility in *Brassica stigma*. Nature 403: 913-916.

TANG S and PRESGRAVES DC, 2009. Evolution of the *Drosophila* nuclear pore complex results in multiple hybrid incompatibilities. Science 323: 779-782.

TANG X, FREDERICK RD, ZHOU J, HALTERMAN DA, JIA Y and MARTIN GB, 1996. Initiation of plant disease resistance by physical interaction of AvrPto and Pto kinase. Science 274: 2060-2063.

TANKSLEY SD, GANAL MW, PRINCE JP, DE VICENTE MC, BONIERBALE MW, *et al.*, 1992. High density molecular linkage maps of the tomato and potato genomes. Genetics 132: 1141-1160.

TAYLOR IB, 1986. Biosystematics of the tomato. In: The Tomato Crop: A Scientific Basis for Improvement (Atherton JG and Rudich J, eds.), pp. 1-34. Chapman and Hall, NY and London.

TEMPLETON AR, 2000. Epistasis and complex traits. In: Epistasis and the Evolutionary Process (Wolf JB, Brodie ED III and Wade MJ, eds.), pp. 41-57, Oxford University Press.

THALER JS and BOSTOCK RM, 2004. Interactions between abscisic-acid-mediated responses and plant resistance to pathogens and insects. Ecology 85: 48-58.

THOMMA BP, CAMMUE BP and THEVISSEN K, 2002. Plant defensins. Planta 216: 193-202.

THOMPSON JN, 1994. The Coevolutionary Process. The University of Chicago Press.

TIGCHELAAR EC, 1986. Tomato breeding. In: Breeding Vegetable Crops (Bassett MJ, ed.), pp. 135-171, AVI, Westport, CT.

TOOMAJIAN C, HU TT, ARANZANA MJ, LISTER C, TANG C, *et al.*, 2006. A nonparametric test reveals selection for rapid flowering in the *Arabidopsis* genome. PLoS Biol. 4: e137.

TRAVERS SAA and FARES MA, 2007. Functional coevolutionary networks of the Hsp70–Hop–Hsp90 system revealed through computational analyses. Mol Biol Evol. 24: 1032-1044.

TRAVERS SAA, TULLY DC, MCCORMACK GP and FARES MA, 2007. A study of the coevolutionary patterns operating within the *env* gene of the HIV-1 group M subtypes. Mol Biol Evol. 24: 2787-2801.

TROWSDALE J, 2006. Multiple sclerosis: putting two and two together. Nature Med. 12: 1119-1121.

TSAI C-T, HWANG J-J, RITCHIE MD, MOORE JH, CHIANG F-T, *et al.*, 2007. Renin-angiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: detection of high order gene–gene interaction. Atherosclerosis 195: 172-180.

VALLEJOS CE, 1979. Genetic diversity of plants for response to low temperatures and its potential use in crop plants. In: Low Temperature Stress in Crop Plants: The Role of The Membrane (Lyons JM, Graham D and Raison JK, eds.), p. 565, Academic Press, NY.

VAN BAARLEN P, VAN BELKUM A and THOMMA BP, 2007. Disease induction by human microbial pathogens in plant-model systems: potential, problems and prospects. Drug Discovery Today 12: 167-173.

VAN BAARLEN P, VAN ESSE HP, SIEZEN RJ, THOMMA BP, 2008. Challenges in plant cellular pathway reconstruction based on gene expression profiling. Trends Plant Sci. 13: 44-50.

VAN OOIJEN G, VAN DEN BURG HA, CORNELISSEN BJC and TAKKEN FLW, 2007. Structure and function of resistance proteins in solanaceous plants. Annu Rev Phytopathol. 45: 43-72.

VAN OOSTERHOUT C, 2009. A new theory of MHC evolution: beyond selection on the immune genes. Proc R Soc B 276: 657-665.

VIVIER E and MALISSEN B, 2005. Innate and adaptive immunity: specificities and signaling hierarchies revisited. Nature Immunol. 6: 17-21.

WAGNER A, 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol. 18: 1283-1292.

WAGNER GP and ALTENBERG L, 1996. Complex adaptations and the evolution of evolvability. Evolution 50: 967-976.

WANG DY, KUMAR S and HEDGES SB, 1999. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. Proc R Soc Lond B. 266: 163-171.

WANG ZO and POLLOCK DD, 2007. Coevolutionary patterns in cytochrome c oxidase subunit I depend on structural and functional context. J Mol Evol. 65: 485-495.

WAXMAN D and PECK J, 1998. Pleiotropy and the preservation of perfection. Science 279: 1210-1213.

WEGNER KM, 2008. Clustering of *Drosophila melanogaster* immune genes in interplay with recombination rate. PLoS ONE 3: e2835.

WEIR BS and COCKERHAM CC, 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38: 1358-1370.

WEIR BS, 1979. Inferences about linkage disequilibrium. Biometrics 35: 235-254.

WEIR BS, 1996. Genetic Data Analysis II. Sinauer Associates, Sunderland, MA.

WESTFALL PH and YOUNG SS, 1993. Resampling-based multiple testing. Wiley and Sons, NY.

WHITTAM TS, OCHMAN H and SELANDER RK, 1983. Geographic components of linkage disequilibrium in natural populations of *Escherichia coli*. Mol Biol Evol. 1: 67-83.

WILFERT L and SCHMID-HEMPEL P, 2008. The genetic architecture of susceptibility to parasites. BMC Evol Biol. 8: 187.

WILLIAMS EJB and BOWLES DJ, 2004. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Res. 14: 1060-1067.

WILLIAMS N, 2009. The growing GM challenge. Curr Biol. 19: R268-R269.

WILLIAMS SG and LOVELL SC, 2009. The effect of sequence evolution on protein structural divergence. Mol Biol Evol. 26: 1055-1065.

WILTSHIRE S, BELL JT, GROVES CJ, DINA C, HATTERSLEY AT, *et al.*, 2006. Epistasis between type 2 diabetes susceptibility loci on chromosomes 1q21–25 and 10q23–26 in northern Europeans. Ann Hum Genet. 70: 726-737.

WOLF JK, 2000. Indirect genetic effects and gene interactions. In: Epistasis and the Evolutionary Process (Wolf JB, Brodie ED III and Wade MJ, eds.), pp 158-176, Oxford University Press.

WRIGHT S, 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc. 6th Intl. Congr. Genet. 1: 356-366.

WRIGHT S, 1940. Breeding structure of populations in relation to speciation. Am. Nat. 74: 232-248.

WU A-J, ANDRIOTIS VME, DURRANT MC and RATHJEN JP, 2004. A patch of surface-exposed residues mediates negative regulation of immune signaling by tomato Pto kinase. Plant Cell 16: 2809-2821.

WULFF BB, KRUIJT M, COLLINS PL, THOMAS CM, LUDWIG AA, *et al.*, 2004. Gene shuffling-generated and natural variants of the tomato resistance gene *Cf-9* exhibit different auto-necrosis-inducing activities in *Nicotiana* species. Plant J. 40: 942-956.

XIANG T, ZONG N, ZOU Y, WU Y, ZHANG J, *et al.*, 2008. *Pseudomonas syringae* effector AvrPto blocks innate immunity by targeting receptor kinases. Curr Biol. 18: 4-80.

XING W, ZOU Y, LIU Q, LIU J, LUO X, *et al.*, 2007. The structural basis for activation of plant immunity by bacterial effector protein AvrPto. Nature 449: 243-247.

YI G, SZE SH and THON MR. 2007. Identifying clusters of functionally related genes in genomes. Bioinformatics 23: 1053-1060.

YUNIS H, BASHAN Y, OKON Y and HENIS Y, 1980. Weather dependence, yield losses, and control of bacterial speck of tomato caused by *Pseudomonas tomato.* Plant Disease 64: 937-939.

ZAYKIN DV, PUDOVKIN A and WEIR BS, 2008. Correlation-based inference for linkage disequilibrium with multiple alleles. Genetics 180: 533-545.

ZENG L-R, VEGA-SANCHEZ ME, ZHU T and WANG G-L, 2006. Ubiquitination-mediated protein degradation and modification: an emerging theme in plant-microbe interactions. Cell Res. 16: 413-426.

ZHANG J and ROSENBERG HF, 2002. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. PNAS 99: 5486-5491.

ZHANG Y, 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9: 40.

ZIPFEL C, KUNZE G, CHINCHILLA D, CANIARD A, JONES JDG, *et al.*, 2006. Perception of the bacterial PAMP EF-Tu by the receptor EFR restricts *Agrobacterium* mediated transformation. Cell 125: 749-760.

ZIPFEL C, ROBATZEK S, NAVARRO L, OAKELEY E, JONES JDG, *et al.*, 2004. Bacterial disease resistance in *Arabidopsis* through flagellin perception. Nature 428: 764-767.

Acknowledgements

My thank-you goes to:

Wolfgang Stephan and Laura Rose for guidance, shared discussions, considerable reserves of patience and introducing me to the field of theoretical population genetics. Laura for optimism and trust, Wolfgang for encouragement in his own way. I would like to add that people are happy if they have one brilliant advisor and I have been lucky to have two.

Faculty and committee members, including: Susanne Renner, John Parsch, Martin Parniske, Dirk Metzler, Dario Leister and John Baines for reviewing this thesis, reading my proposals and providing invaluable feedback and advice in the course of my research.

Simone Lange for a great help and an unforgettable atmosphere during my stay in the lab from our first day on the job.

Thomas Staedler, Ann Arunyawat and Traudl Feldmaier-Fuchs, who kindly shared not only plant material, but also expertise and enthusiasm in the study of wild tomatoes.

Students from the Rose lab, who were working on projects related to my dissertation research: Martina Jugl, Anja Hoerger and Florian Clemente.

Katrin Kuempfbeck for helping me to take my first steps in the real world of German administration.

Members of the Munich Evolutionary Biology group for comfortable working conditions, good intentions and random distractions after hours.

My family and my friends for their support and interest in my work.

If you reached this page, make sure that you read the content of this publication.
I wish you a fascinating and enlightening reading.

Ehrenwörtliche Versicherung

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

Erklärung

Hiermit erkläre ich, dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich anderweitig einer Doktorprüfung ohne Erfolg nicht unterzogen habe.

München, den 01.Oktober 2009.

Lukasz Grzeskowiak