# Host–parasite co-evolution and its genomic signature

*Dieter Ebert* [1,2]✉ *and Peter D. Fields* [1]

Abstract | Studies in diverse biological systems have indicated that host–parasite co-evolution is responsible for the extraordinary genetic diversity seen in some genomic regions, such as major histocompatibility (MHC) genes in jawed vertebrates and resistance genes in plants. This diversity is believed to evolve under balancing selection on hosts by parasites. However, the mechanisms that link the genomic signatures in these regions to the underlying co-evolutionary process are only slowly emerging. We still lack a clear picture of the co-evolutionary concepts and of the genetic basis of the co-evolving phenotypic traits in the interacting antagonists. Emerging genomic tools that provide new options for identifying underlying genes will contribute to a fuller understanding of the co-evolutionary process.

Parasites
Organisms, including pathogens, that take advantage of other organisms (hosts), thereby instigating a process of selection by the host to defend against the parasite.

Genomic signatures
Characteristic patterns of genetic variation, observed at a genomic region in a sample of genomes.

Selective sweep
The spread of a beneficial mutant and the hitch-hiking of genetic variants close to it in the genome. Beneficial mutants may have arisen de novo or were segregating in the population before the sweep and become beneficial after a change in conditions.

Host–parasite co-evolution occurs when selection by parasites triggers diverse host adaptations that reduce the costs of infection, which in turn prompts parasites to adapt anew to their hosts. This process may be among the most important generators of biological diversity over the past 3.5 billion years[1,2], including the generation and maintenance of genetic diversity within populations and species, and the sharing of certain variants across species boundaries[3–9]. As such, the genomes of organisms would be expected to show signatures of this host–parasite co-evolution. Recognizing and characterizing these genomic signatures would be expected to lead to a better understanding of how diversity has evolved, and is still evolving, across the tree of life. Furthermore, identifying genomic signatures of co-evolution can help to narrow down the loci under selection and provide mechanistic insights into host–parasite interactions.

Antagonistic co-evolution has been defined as reciprocal selection between two closely interacting species[10,11]. This definition focuses on the phenotypic traits of the co-evolving antagonists that negatively influence each other. It specifies that the traits are responsible for the interaction and that their underlying genes co-evolve — not the species[12,13]. It also sets antagonistic co-evolution apart from scenarios of host–parasite co-association and the resulting long-term patterns of co-speciation or co-cladogenesis, which do not require adaptive processes[14].

The process of antagonistic co-evolution can be described by two types of model: models of specific co-evolution, in which one host and one parasite species interact (sometimes called pairwise co-evolution); and models of unspecific co-evolution, better known as diffuse co-evolution[10], in which multiple hosts and/or parasite species contribute to the process. Models of specific co-evolution can be further grouped into selective sweep models, in which novel variants are selected for and rise to high frequencies, and balancing selection models, in which alternative variants at specific loci fluctuate in frequency over time[11,15]. The different modes of action underlying these models create different signatures in the genomes of the antagonists. Thus, observation of a particular signature in a genome allows (within limits) conclusions to be drawn about the evolutionary mechanism producing them (TABLE 1).

A genomic perspective of co-evolution considers the entire region around the co-evolving genes, because making use of the additional information present in the flanking sites strongly increases the power of the analysis to detect genomic signatures. Despite being functionally independent, sites physically tightly linked to the co-evolving genes are influenced by linked selection, which causes genetic hitch-hiking, and, unlike unlinked sites, their fate is determined by the dynamics of the selected genes and the rate of recombination among them[16,17]. Thus, genomic signatures of co-evolution can be detected in populations by comparing patterns of genetic variation in these regions with the patterns in the genomic background that presumably evolved by neutral evolutionary processes. The genomic signatures of models of specific co-evolution have received much attention, resulting in a good picture of the expected patterns in population samples of genomes. By contrast, the genomic signatures expected under diffuse co-evolution have not yet been determined; nevertheless, some heuristics exist.

In this Review, we address conceptual issues concerning host–parasite co-evolution and how they manifest in the genomes of the antagonists. We highlight the differences between specific and diffuse co-evolution and point out the role of spatial population structure in

[1]Department of Environmental Sciences, Zoology, University of Basel, Basel, Switzerland.

[2]Wissenschaftskolleg zu Berlin, Berlin, Germany.

✉e-mail: dieter.ebert@unibas.ch

https://doi.org/10.1038/s41576-020-0269-1

Balancing selection
Selection that occurs because the alleles involved have, on average, a selective advantage that correlates negatively with their frequency within a population or species. This term does not include loci that are at a balance between gaining and losing variants, such as mutation-drift and mutation-selection balance.

Linked selection
The evolution of (nearly) neutral SNPs influenced by selection on loci physically linked to them.

Genetic hitch-hiking
The process by which a genetic variant changes in frequency because it is physically linked to another variant that is changing in frequency due to selection.

Co-genomic
The simultaneous analysis of the function, structure and evolution of pairs of associated genomes in closely interacting organisms, such as host and parasites.

Overdominance
The scenario in which heterozygotes have a more extreme phenotypic trait value than all homozygotes. Overdominance for fitness results in balancing selection for the alleles causing the advantage for the heterozygote genotypes.

Local adaptation
If the effect of an allele is habitat-specific, such that it is beneficial in one habitat and detrimental in another, and vice versa for the alternative allele, local adaptation may evolve with directional selection within each population. Local adaption is a powerful mechanism to maintain genetic diversity within species.

Direct negative frequency-dependent selection
The selective benefit of an allele depends directly and negatively on its frequency, for example at sex-determining loci and plant self-incompatibility loci.

Indirect negative frequency-dependent selection
The selective benefit of an allele depends on the frequency of an allele in a co-evolving species.

Table 1 | **Comparison of features of specific co-evolution by selective sweep and balancing selection**

| Feature | Selective sweep co-evolution | Balancing selection co-evolution |
|---|---|---|
| Form of selection | Positive selection drives sweeps; selection is directional | Negative frequency-dependent selection gives common alleles a disadvantage; selection results in a balance of the frequencies of genetic variants |
| Functional polymorphisms | Visible only during selective sweeps | Maintained constantly and potentially for very long time periods |
| Underlying genetic system | Beneficial mutation in the host and parasite at any locus in the nuclear or cytoplasmic genome may sweep | Frequencies of alternative alleles at a few selected loci are balanced |
| Role of mutations | Mutations define the onset of new selective sweeps (hard sweeps) | Mutations are not necessary but do create rare variants, which may be selected and contribute to balancing selection or even replace a previous variant |
| Temporal continuity | Process can be highly stochastic and does not need to be continuous; long periods without sweeps are possible | Process must operate continuously because genetic variants may otherwise be lost. In a spatial setting, previously lost alleles may be reintroduced from other populations |
| Timescale of phenotypic change | Relatively slow because new mutations take a long time to reach a high enough frequency to be recognized. Sweeps starting from standing genetic variation progress more quickly | Fast because genetic variants are always at intermediate frequencies where selection results in fast changes |
| Population divergence | Sweeps drive population and species divergence | Population divergence is prevented in the long term, although it may occur in the short term |
| Evolutionary outcome | Creates macroevolutionary patterns (lineage divergence) | Explains high levels of genetic diversity within populations and species |
| Introgression among species | May introduce beneficial new alleles that can sweep | May introduce new functional variants that can contribute to balancing selection, but may create a fake picture of trans-species polymorphism |

shaping genomic signatures. Using these concepts, we discuss the evolution of trans-species polymorphisms (TSPs). Finally, we address new co-genomic methods that allow co-evolving loci in hosts and parasites to be directly pinpointed and what we can learn from these recent developments.

## Models of specific co-evolution

The two models of specific host–parasite co-evolution, selective sweep selection and balancing selection, have different effects on genetic variation. Selective sweep co-evolution is arguably one of the most important generators of macroevolutionary patterns, and explains differences in immune systems among lineages, drives speciation and is implicated in some of the major transitions in evolution[2,18–20]. However, it fails to explain the extraordinary genetic diversity observed at some host resistance loci. By contrast, co-evolution by balancing selection is best known for its potential to maintain high levels of genetic diversity within populations and species. The theory states that hosts and parasites undergo continuous antagonistic co-evolution, often referred to as Red Queen dynamics or trench warfare, which results in a balance of different variants at loci related to host defence and parasite offence[21–23]. Recent studies have revealed that there are more genomic regions in plants and animals that undergo balancing selection than previously thought[8,24–29].

The study of co-evolution by analysing patterns of genetic variation has a long tradition, but has numerous limitations that make it hard to reach strong conclusions about underlying evolutionary processes. For example, genomic studies on co-evolution are mostly performed separately in host and parasite genomes (but see Joint analysis of genomes). Thus, a given finding may be difficult to attribute to host–parasite co-evolution or to other processes, because the bioinformatic and population genetic approaches that are used to identify genomic regions with characteristics of selective sweeps or balancing selection do not identify the cause of selection — antagonistic interactions are only one of many possibilities. Additional efforts are needed to differentiate between these options. This issue is particularly pronounced for balancing selection, as different mechanisms (such as overdominance, local adaptation, direct negative frequency-dependent selection and indirect negative frequency-dependent selection) can all produce genomic signatures of balancing selection. However, only indirect negative frequency-dependent selection (NFDS) is associated with antagonistic co-evolution, and so distinguishing among these mechanisms is crucial.

***Selective sweep co-evolution.*** A selective sweep describes the rapid increase in frequency of a beneficial variant, such as a novel mutation. In the extreme case, the variant will reach fixation, replacing alternative variants. Adaptive evolution in this form is considered to be a dominant driver of protein adaptation and species divergence[30,31]. During host–parasite co-evolution, both antagonists may experience sweeps at loci that play a functional role in their interaction[11,15,32–35] (FIG. 1). For co-evolution to occur, sweeps do not need to alternate

Selective interference
In clonal, but not in sexual, populations, beneficial mutations interfere with each other, such that at a given moment the fittest mutation will outcompete weaker beneficial mutations. Interference can also affect the spread of mutants in genomic regions with low recombination rates.

Linkage disequilibrium
A statistical measure of the distribution of combinations of alleles at different loci, which is zero if this distribution follows the expectation based only on allele frequencies. Non-zero values of linkage disequilibrium can arise due to hitch-hiking, selection on allele combinations and stochastic processes, and may occur among loci without physical linkage.

Selection coefficient
A measure of fitness of genotypes or alleles relative to a reference, such as the ancestral form.

between hosts and parasites. Variants at multiple sites may even spread at the same time, so long as they are decoupled by recombination, because otherwise selective interference takes place[33,34,36].

A selective sweep leaves a characteristic valley of locally reduced genetic variation in population samples of genomes. These diversity valleys are formed because variants that are in proximity to the sweeping variant hitch-hike along and replace alternative variants, which results in a corresponding pattern of increased linkage disequilibrium[37]. Such valleys allow us to detect the approximate position of the beneficial mutant, but the genomic signatures are rarely clear and lose their distinct features over time as recombination breaks down associations and as new mutations arise. Thus, ancient selective sweeps are no longer detectable, although they leave their traces in signatures of positive selection that result in patterns of divergence among lineages[20]. Likewise, sweeping mutants in the initial phase of a sweep are hard to detect, as the reduction in local genetic diversity is not yet pronounced (FIG. 2). Additionally, each sweep event has a different history: each one is associated with a different selection coefficient and may occur at a different location in the genome, and possibly experience different local recombination rates[38]. Finally, signatures of sweeps strongly depend on the initial variation on which selection acted[37].

Variants spreading to fixation in co-evolving hosts and parasites may be functionally independent, other than that they each provide a benefit to their carrier and are disadvantageous to the antagonist[23]. However, when particular host or parasite proteins interact, selective sweep co-evolution can occur through alternating changes to functionally important segments of these proteins. For example, the restriction factor *TRIM5*, an antiviral protein in Old World primates, seems to have co-evolved in tight interaction with capsid proteins in lentiviruses[39].

Due to their high rate of evolution, parasites are among the fastest changing selecting agents that host organisms experience. Therefore, the speed with which a host population can respond to a new variant of a parasite is important. However, selective sweep evolution from a novel mutation to fixation (that is, a hard sweep) is a rather slow process. In particular, the initial increase in allele frequencies is slow: in a large population, a novel mutation with a 10% fitness advantage takes about 200 generations to reach a frequency of 10%, but only about 15 generations more to reach 50%[40,41]. Thus, sweeps starting from standing genetic variation (known as soft sweeps) can occur faster than hard sweeps, as the initial allele frequencies are higher. Indeed, evidence is accumulating that many sweeps observed in natural populations are soft sweeps[41–43]. Interestingly, standing

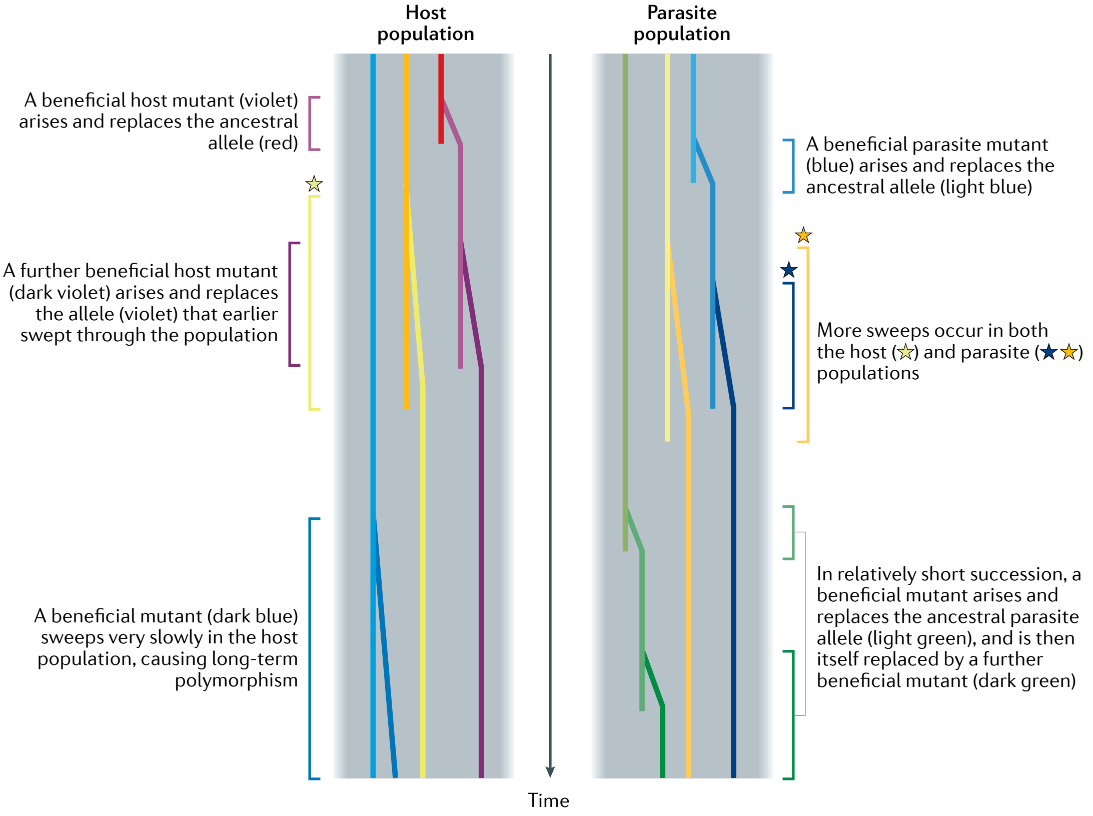

Host population
Parasite population
A beneficial host mutant (violet) arises and replaces the ancestral allele (red)
☆
A further beneficial host mutant (dark violet) arises and replaces the allele (violet) that earlier swept through the population
A beneficial mutant (dark blue) sweeps very slowly in the host population, causing long-term polymorphism
A beneficial parasite mutant (blue) arises and replaces the ancestral allele (light blue)
★
☆
More sweeps occur in both the host (☆) and parasite (★☆) populations
In relatively short succession, a beneficial mutant arises and replaces the ancestral parasite allele (light green), and is then itself replaced by a further beneficial mutant (dark green)
Time

Fig. 1 | **Schematic representation of selective sweep co-evolution in the gene pools of a host and parasite population.** For each antagonist, timelines of alleles at different loci are shown in different colours. Over the shown time period, each of the alleles is replaced at least once by a mutant, indicating selective sweeps. Genetic variation at loci — that is, the presence of multiple alleles in the gene pool at a given moment in time — is only visible during the sweep. Host and parasite sweeps are not linked. Loci can be present anywhere in the genome. Other loci may undergo sweeps at the same time for reasons unrelated to the co-evolutionary interactions.

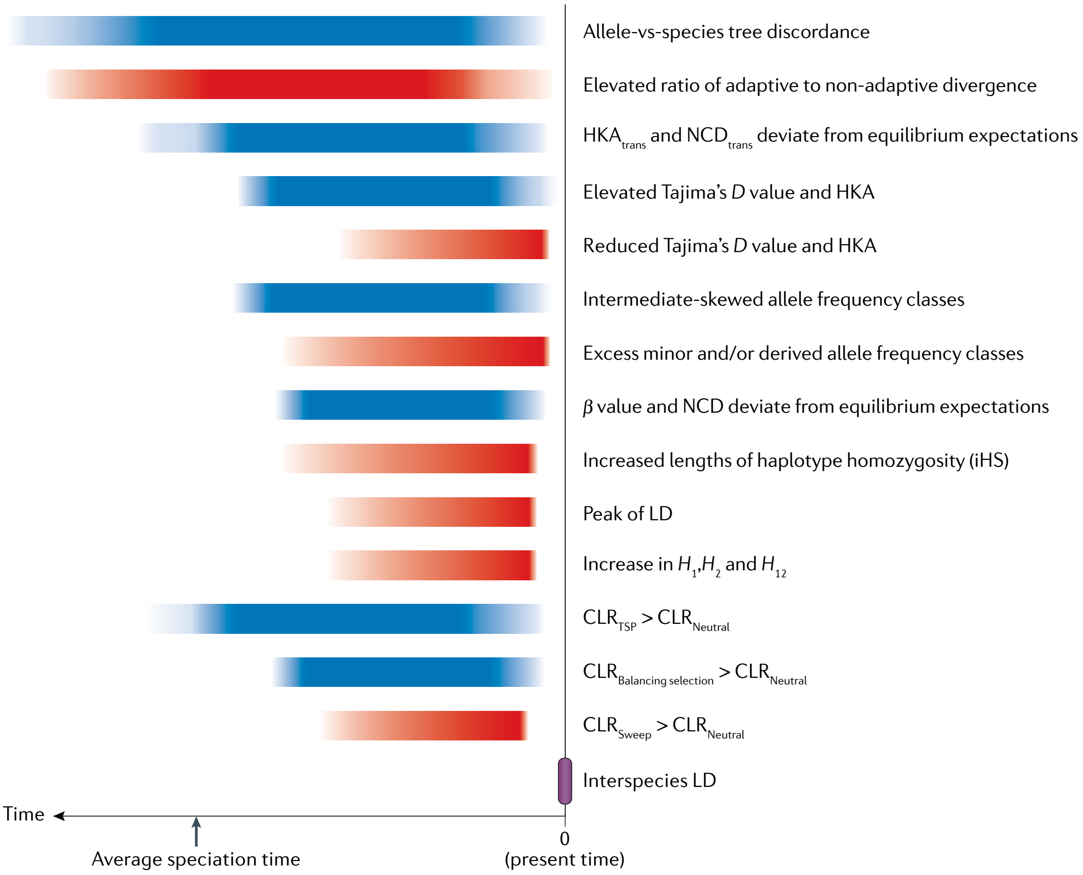


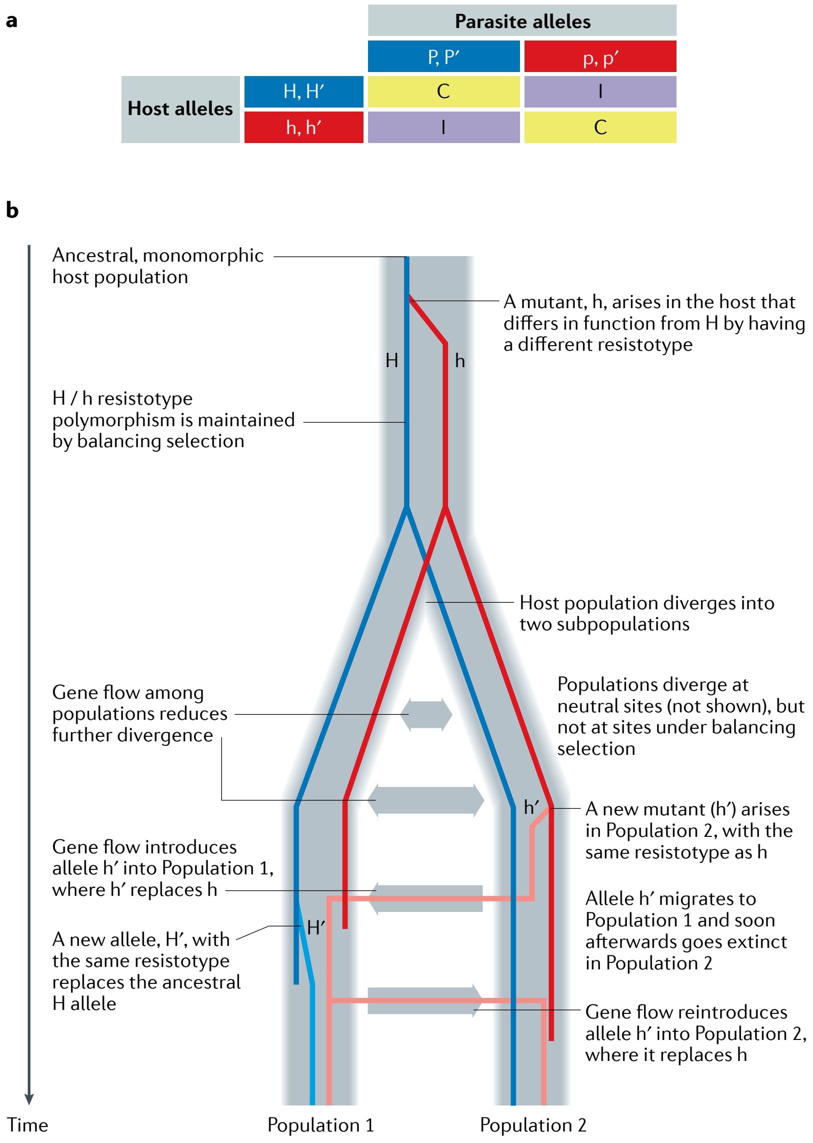
Fig. 2 | **Temporal dynamics of genomic signatures.** Approximate timescales during which genomic signatures for balancing selection (blue), positive selection and/or selective sweep (red) and interspecies linkage disequilibrium (purple) are detectable with common (although not exhaustive) population genetic tests (TABLE 2). Although signatures of balancing selection (especially those that are associated with trans-species polymorphism (TSP) or approximations thereof; see REF.[29]) and positive selection may extend quite deep into evolutionary time, patterns of very rapid evolutionary change associated with positive selection or selective sweeps will show a much more limited time depth. Concomitantly, the number of generations a given selective regime has to have in place before inference of its signature is possible may be different between balancing selection and positive selection, and, on the shortest times scales and with a sample from a single population, the signatures of balancing selection and positive selection may be indistinguishable. As a selective sweep grows older, its signature will disappear and become more similar to the signature of positive selection, which is visible only as elevated ratios of adaptive to synonymous divergence. Patterns of interspecies linkage disequilibrium provide no historical record of the co-evolutionary process, and so additional methods (such as those described in this figure) are required to determine what (if any) selective regime has been at play at the statistically associated loci. It should be noted that the suggested time frame of inference for individual tests may differ substantially for individual data sets as a result of experimental factors, such as sample size, and as the result of species-specific factors, such as demography. Allele frequency classes may be a component of summary statistics (such as Tajima's $D$ value) or may be the specific metric of interest (for example, site frequency spectrum). $\beta$ value, measure of allele frequency correlation[183]; CLR, composite likelihood ratio test[187]; $H_1$, $H_2$ and $H_{12}$, frequencies of the most frequent haplotype, of the second most frequent haplotype and of the first and second most frequent haplotypes, respectively; HKA, Hudson, Kreitman and Aguadé; $HKA_{trans}$, modified version of the HKA test to accommodate genomic data from multiple species in order to detect TSP; iHS, integrated haplotype score[184]; LD, linkage disequilibrium[182]; NCD, non-central deviation summary to determine deviation from neutral allele frequency expectations; $NCD_{trans}$, version of the NCD summary statistic extended to multispecies data sets in order to detect TSP[28,29] (see TABLE 2).

genetic variation is often high for loci related to host–parasite interactions and, thus, can drive faster sweeps. Although it is not fully known why this is the case, the two-speed genome model[44–46] attempts to explain this paradox by suggesting that specific regions in host and parasite genomes have strongly elevated mutation rates, constantly refuelling standing genetic variation, whereas housekeeping genes have lower mutation rates and evolve by purifying selection.

***Co-evolution by balancing selection.*** The Red Queen model of balancing selection was introduced by Clarke[47] to explain unexpectedly high genetic diversity within populations. It is based on the assumption of strong host–genotype by parasite–genotype interactions[48–50] (FIG. 3a), which has been finding increasing empirical support[51–53]. As opposed to the selective sweep model, in which mutant alleles have a general advantage over the wild-type alleles, genetic variants under balancing selection provide an advantage only in specific relation to a corresponding genetic variant in the antagonist. These host and parasite genes are functionally coupled, forming the driving force for reciprocal selection[22,54–56]. According to the Red Queen model, a parasite allele will increase in frequency when the host allele that allows it to infect is common; as hosts carrying this allele succumb to increased parasitism, their numbers will decline and — with a time lag — so too will parasite genotypes that depend on these particular host genotypes, that is, the loci in the host and the parasite are under NFDS. As parasites with particular infectivity alleles track corresponding host alleles, cyclical dynamics of host and parasite alleles might arise, with parasite cycles lagging behind the corresponding host allele cycles[57,58]. NFDS reduces the likelihood that alleles go extinct by chance because rare alleles gain an advantage; thus, genetic polymorphisms in hosts and parasites can be maintained, and the frequencies of the functional variants at the involved loci are balanced over time (FIG. 3b). In this form of balancing selection, an allele's selection coefficient does not depend directly on its own frequency, but rather depends on the frequencies of specific alleles in the antagonist. Therefore, it is called indirect NFDS[21].

Relative to the rest of the genome, genomic regions containing genes under balancing selection are expected to show high genetic diversity and a more even frequency spectrum of variants, resulting in positive Tajima's *D* values[16,26,59,60] (TABLE 2). Scans of host genomes have revealed ample evidence of regions under balancing selection, with the most well-known being the major histocompatibility (MHC) class 1 and 2 genes of jawed vertebrates and the ABO blood group system of higher primates, but also other regions of vertebrate genomes related to immune function[25–28,61–64]. Scans in plants and invertebrates have also revealed balancing selection in diverse genomic regions, again often regions associated with immune function[8,65–71]. However, current bioinformatic and population genetic methods may still miss many regions of interest. To produce a recognizable genomic signature, selection should be strong and should have acted for a sufficiently long time period (FIG. 2). Regions around sites under selection usually show strong local linkage disequilibrium, and thus the local recombination rate will also affect the ability to detect signatures. Ideally, the recombination rate around the selected loci should be low to allow linked SNPs to hitch-hike; high recombination rates would reduce the region to the very site under balancing selection itself[16,21]. However, some level of recombination is helpful for the detection of the sites under selection, as otherwise the entire region will be inherited as one linkage block.

Less attention has been given to balancing selection in parasites than in hosts. Mapping parasite loci is more difficult because parasites often cannot be cultured and phenotyped outside the host. Furthermore, they often have no or infrequent genetic recombination, have extreme population structures and demography, carry extra chromosomal genetic elements (such as plasmids) and have substantial divergence in gene content as a result of horizontal gene transfer: each of these features may produce genomic signatures that make



a

| | | Parasite alleles | |
|---|---|---|---|
| | | P, P′ | p, p′ |
| Host alleles | H, H′ | C | I |
| | h, h′ | I | C |

b
Ancestral, monomorphic host population
A mutant, h, arises in the host that differs in function from H by having a different resistotype
H
h
H / h resistotype polymorphism is maintained by balancing selection
Host population diverges into two subpopulations
Gene flow among populations reduces further divergence
Populations diverge at neutral sites (not shown), but not at sites under balancing selection
h′
A new mutant (h′) arises in Population 2, with the same resistotype as h
Gene flow introduces allele h′ into Population 1, where h′ replaces h
Allele h′ migrates to Population 1 and soon afterwards goes extinct in Population 2
H′
A new allele, H′, with the same resistotype replaces the ancestral H allele
Gene flow reintroduces allele h′ into Population 2, where it replaces h
Time
Population 1
Population 2

Fig. 3 | **Host resistotype and parasite infectotype interaction matrix and balancing selection.** A simple matching allele model of two host resistotypes and two parasite infectotypes is sufficient to create long-term balancing selection within populations. **a** | Haploid hosts with the H allele have resistotype C (for compatible interactions, that is, the host is susceptible) for parasites with the P allele and resistotype I (for incompatible interactions, that is, the host is resistant) for parasites with the p allele; hosts with the h allele have resistotype I for parasites with the P allele and resistotype C for parasites with the p allele. The haploid parasite with the P allele has the infectotype C for hosts with the H allele and infectotype I for hosts with the h allele, whereas parasites with the p allele have infectotype I for hosts with the H allele and infectotype C for hosts with the h allele. A mutant of an allele may produce a new allele (H′, h′, P′ and p′), but if the function of the allele is not affected with regards to its interaction with the parasite, the resistotypes and infectotypes will stay the same. **b** | Schematic representation of balancing selection in a host population that diverges into two populations. Allele colours correspond to functional types of the host. Initially, the H allele (blue) gives rise to the h allele (red), which has a different resistotype. Afterwards, the two allele types are maintained by balancing selection. Their infection profile might resemble the matching allele model shown in part **a**. Further mutations in the H and h alleles change the genotype of these alleles, but not their resistotype. The H allele persists until it is replaced by the H′ allele (light blue) in Population 1. The h allele gives rise to the h′ allele (pink) in Population 2. H and H′, as well as h and h′, change in relative frequency independent of selection by the parasite. Gene flow between populations can introduce new alleles (h′ into Population 1) and reintroduce extinct alleles (h′ into Population 2), which can replace the resident allele.

Table 2 | **Genomic summaries used to determine the evolutionary process underlying a genomic signature**

| Name of summary | Type of summary | Description | Signature type[a] (relative to genomic background) | Local genomic scale of analysis | Other comments |
|---|---|---|---|---|---|
| Allele vs species-tree discordance[179] | Summary statistic | Compares phylogenetic relationships of genes with those of species | Discordant trees indicate balancing selection and/or TSP | Distinct genes and/or windows across the genome | – |
| Elevated ratio of adaptive to non-adaptive divergence[180,181] | Summary statistic and likelihood method | Measures the ratio of adaptive to synonymous divergence | High values indicate positive selection | Gene classes can be compared | Explicit tests arise from the MK test or its derivatives |
| $HKA_{trans}$[29] | Summary statistic | Adaptation of the standard HKA test to better accommodate genomic data from multiple species | Positive values of chi-squared test statistic are suggestive of balancing selection | Windows across the genome | – |
| $NCD_{trans}$[28,29] | Summary statistic | Extension of the NCD summary statistic to accommodate multispecies data | Smaller values will generally be indicative of balancing selection | Windows across the genome | Describes the NCD summary, which includes an outgroup and fixed differences as part of the test |
| Tajima's *D* value and HKA test[182] | Summary statistic | Excess of low-frequency class polymorphisms | Decreased values indicate positive selection/selective sweep | Windows across the genome | Summary can be confounded with demography due to population expansion causing similar measures of the summary statistic |
| | | Increased values indicate balancing selection | Increased values indicate balancing selection | Windows across the genome | Can be confounded with demography due to population bottlenecks causing similar measures of the summary statistic |
| Perturbations from equilibrium SFS[182] | Summary statistic and likelihood method | Intermediate-skewed allele frequency classes | Excess derived allele frequency classes as compared with neutral expectation is indicative of balancing selection | Windows across the genome | Summary can be confounded with demography due to population bottlenecks causing similar measures of the summary statistic |
| | | Excess minor and/or derived allele frequency classes | Excess minor or derived allele frequency classes as compared with neutral expectation is indicative of positive selection and selective sweeps | Windows across the genome | Summary can be confounded with demography due to population expansion causing similar measures of the summary statistic |
| $\beta$ value[183] | Summary statistic | Measure of allele frequency correlation and overall mutation rate | Values greater than zero are indicative of balancing selection | Windows across the genome | – |
| NCD[28] | Summary statistic | Measure of deviation of the minor allele frequency from the expectations under a scenario of balancing selection | Smaller values will generally be indicative of balancing selection | Windows across the genome | – |
| iHS[184] | Summary statistic | Lengths of haplotype homozygosity or the decelerated decay of LD | Increased lengths of homozygosity and the slower decay of LD compared with neutral expectations is indicative of a selective sweep | Windows across the genome | – |
| LD[182] | Summary statistic | Distinct increase in LD among a subset of adjacent loci | Elevated values above the genome-wide background indicate a selective sweep | Windows across the genome | – |
| $H_1$, $H_{12}$, $H_2$ (REFS[42,185]) | Summary statistic | Frequencies of ranked haplotypes, that is, haplotype spectra | Increased frequencies of distinct, ranked haplotypes (and their relative frequencies) is indicative of a sweep (hard vs soft) | Windows (SNPs) across the genome | – |
| $T_{trans}$[29] | Likelihood method | Adaptation of the *T* statistic from[186] to specifically detect TSP | Fit of the observed data to a model of long-term balancing selection or TSP is greater than that of a model of neutral evolution ($CLR_{TSP} > CLR_{Neutral}$) | Windows across the genome | – |

Table 2 (cont.) | **Genomic summaries used to determine the evolutionary process underlying a genomic signature**

| Name of summary | Type of summary | Description | Signature type[a] (relative to genomic background) | Local genomic scale of analysis | Other comments |
|---|---|---|---|---|---|
| *T* statistic[186] | Likelihood method | Aggregate of summary statistics of genomic diversity parameterizes a maximum likelihood-based comparison of the fit of a neutral vs balanced polymorphism model | Fit of the observed data to a model of balancing selection is greater than that of a model of neutral evolution ($\mathrm{CLR}_{\mathrm{Balancing\ selection}} > \mathrm{CLR}_{\mathrm{Neutral}}$) | Windows across the genome | – |
| CLR[187] | Likelihood method | Aggregate of summary statistics of genomic diversity parameterizes a maximum likelihood-based comparison of the fit of a neutral vs selective sweep model | Fit of the observed data to a model of balancing selection is greater than that of a model of neutral evolution ($\mathrm{CLR}_{\mathrm{Sweep}} > \mathrm{CLR}_{\mathrm{Neutral}}$) | Windows across the genome | Refers specifically to the model described in[187] and its derivatives[188–191]; although CLR is used sometimes to refer to this specific summary, composite likelihood ratio tests are a general statistical procedure for model comparison |

$\beta$ value, measure of allele frequency correlation; CLR, composite likelihood ratio test; $H_1$, $H_2$, $H_{12}$, frequencies of the most frequent haplotype, the second most frequent haplotype and the first and second most frequent haplotypes, respectively; HKA, Hudson, Kreitman and Aguadé; $\mathrm{HKA}_{\mathrm{trans}}$, modified version of the HKA test to accommodate genomic data from multiple species in order to detect TSP; iHS, integrated haplotype score; LD, linkage disequilibrium; MK, McDonald–Kreitman; NCD, non-central deviation; $\mathrm{NCD}_{\mathrm{trans}}$, version of the NCD summary statistic extended to multispecies data sets in order to detect TSP; SFS, site frequency spectrum; TSP, trans-species polymorphism; $T_{\mathrm{trans}}$, likelihood ratio test statistics for detecting TSP. [a]Refers to the two models of specific co-evolution that may be supported with this summary.

it more difficult to identify the signature of balancing selection[72,73]. New bioinformatic tools for genome analysis are now being developed that address some of these challenges (see Table 3 in REF.[73]), and theoretical analyses suggest that balancing selection might show stronger genomic signatures in parasites than hosts[21]. Examples of loci likely under balancing selection have been described for a range of parasites[74–79]. For West Nile virus in mosquitoes and deformed wing virus in honeybees, it has been suggested that genetic diversity is maintained by selection from RNA interference (RNAi) in the host[78,80,81]. Interestingly, the genes for RNAi in *Drosophila* were suggested to evolve by very high rates of positive selection with evidence for recent selective sweeps[82]. Two closely related human pathogenic bacteria, *Staphylococcus aureus* and *Staphylococcus epidermidis*[74,75], show evidence for balanced polymorphisms, but for different genes. The *hrpA* gene of the plant pathogen *Pseudomonas syringae*, which encodes part of the type III secretion system, was found to be under balancing selection[76]. In addition, a rare example of a balanced polymorphism involving structural variation, which determines the presence or absence of alternative pathogenicity islands, was described in *Pseudomonas viridiflava*, a plant pathogen often found on *Arabidopsis*[83]. Interestingly, several of these bacteria are known for their wide host range, hinting that diffuse co-evolution may contribute to the observed patterns (see Diffuse co-evolution section). Sites with high genetic diversity and polymorphisms have also been described in human malaria parasites[84,85], although the signature of balancing selection here may be confounded with processes resulting from acquired immunity. Finally, in the planktonic crustacean *Daphnia magna* and its highly specific bacterial parasite *Pasteuria ramosa*, evidence for balancing selection comes from both the infectivity loci in the bacteria[79] and from the resistance loci in the host[53].

The genomic signatures of co-evolution in hosts and parasites cannot be expected to resemble each other, as the two antagonists experience selection in very different ways. For parasites, the fitness difference between infecting and failing to infect a host is very large, whereas the lifetime fitness difference between hosts that resist a parasite attack or not is likely much smaller[86]. In addition, some hosts may not encounter the parasite during their lifetime, reducing selection for resistance even further. Hosts and parasites also may have different generation times, recombination rates, effective population sizes, ploidy levels, mutation rates, population structures and demography, to name just a few dissimilarities. Therefore, the footprint of co-evolution and the genomic signatures of the two co-evolving antagonists may look very different and may sometimes only be detectable in one of the antagonists[21].

Initially, population genetic modelling of co-evolution by NFDS was performed for infinite population sizes, which strongly reduces the chance of extinction of host and parasite variants. Real populations, however, are of finite size and subject to stochastic effects, so functional variants may be lost due to genetic drift, bringing — in the most extreme case — co-evolution to a halt[21,87–89]. These effects can substantially blur the genomic signatures of co-evolution in natural populations, making their interpretation more difficult[21]. However, spatial structures of host populations can strongly reduce these stochastic events, as long as subpopulations are not evolving in synchrony[90–92].

***Specific co-evolution and spatial structure.*** Simple models of host–parasite co-evolution generally focus on evolutionary dynamics in a single, large, panmictic population. However, populations are almost always spatially distributed, with an extended geographic structure and gene flow among populations. This spatial structuring profoundly influences co-evolution and

Tajima's *D*
A population genetic summary statistic describing the frequency distribution of polymorphisms in a population, with *D* being zero under neutral evolution and positive under balancing selection.

Genetic drift
A neutral evolutionary process that influences allele frequencies based on the random sampling of genetic variants during reproduction.

Panmictic
Random mating within a population.

Fixation index
A measure of genetic differentiation of spatially structured populations, usually estimated from SNP or microsatellite data.

Directional selection
A mode of natural selection by which a genetic variant is predicted to spread to fixation (also known as positive selection).

Supergene
A group of tightly linked genes on a chromosome that are inherited together as a haplotype and often have reduced recombination.

divergence[58,90,91,93–97], and accounting for it can strongly refine population genomic analyses[33,73,98].

Research on subdivided populations has examined how co-evolution influences the spread of variants related to host–parasite interactions, thereby contributing to patterns of genetic variation on a species-level scale (rather than a population-level scale)[96,99,100]. In a selective sweep scenario, strong gene flow allows globally beneficial mutations to spread quickly from population to population, whereas weak gene flow leads to population divergence[37,101]. More complex patterns of divergence may arise, for example, when gene flow is rare and unbalanced among subpopulations, when it differs for the two antagonists and when populations adapt locally to other environmental factors[102]. Genomic patterns of such processes are hard to study, especially if the evolving genes are not known beforehand[37,101].

By contrast, under balancing selection, gene flow enables alleles to persist much longer in the overall gene pool, as it can bring locally extinct alleles back into the population[60,93,100,103,104] (FIG. 3b). In addition, as immigrating alleles are likely to be rare upon their arrival, they are expected to have an advantage and experience a lower rate of extinction and a higher likelihood of spreading locally[104,105], thereby increasing their effective migration rate compared with neutral genetic variants. The number of alleles maintained on a species level thus rises accordingly, and the loci under NFDS are expected to show less differentiation among populations than neutral loci that do not benefit from NFDS[79,106]. This difference is seen when comparing isolation-by-distance patterns for these two groups of loci: whereas neutral loci show increasing differentiation (visible, for example, as an increased fixation index) with increasing distance among populations, loci involved in NFDS show no or a much weaker pattern of isolation by distance[107], which is the opposite of what local positive selection would produce[104,108,109]. However, although the immigrant advantage reduces differentiation at the loci under NFDS among populations over the long term, it may drive the divergence of neighbouring populations in the short term, for example if strong selection occurs at a given gene but for different variants. Such patterns have been observed in some population genetic studies of resistance genes[100,106,110]. Thus, populations may show more variable pairwise fixation indices at sites under NFDS than at neutral sites.

Although limited gene flow causes populations to show patterns of isolation by distance[111,112], the perspective described above does not include common ecological and biological circumstances that may influence gene flow, such as spatially divergent selection, population size and genetic drift, metapopulation and source–sink population structure, and historic events. Parasites may differ in local abundance and may not be present in every host population. Thus, the propensity for host–parasite co-evolution may vary in space and time. Host genes involved in host–parasite interactions may become neutral in the absence of the parasite, or even detrimental if there are costs of resistance. The combined dynamics of these and other evolutionary and ecological processes that influence spatial and temporal variation have been described as the geographic mosaic of co-evolution, with hot spots showing strong and rapid co-evolutionary dynamics, and cold spots marked by slow or no co-evolution[113,114]. For any given system, the greater the proportion of co-evolutionary cold spots, the weaker the overall signature of balancing selection. However, even at cold spots, one may find elevated genetic diversity at loci that are under balancing selection in hot spots, because gene flow from hot spots may prevent extinction of variants in cold spots. Species wide, the overall genomic signature of balancing selection may, therefore, be relatively insensitive to the geographic mosaic of co-evolution. However, this will certainly depend on the interplay of migration, genetic drift and selection.

## Trans-species polymorphism

Because alleles under balancing selection are less likely to be lost from a gene pool, they segregate for longer time periods than genes undergoing neutral evolution or directional selection[16]. This theory can be tested using coalescence approaches that compare regions under balancing selection with the genomic background[38]. In extreme cases, ancient identical-by-descent genetic variants at polymorphic sites may have pre-existed even before the last speciation event, so that these variants are shared across closely related species (FIG. 4a). This TSP is indicated when haplotypes from the two species cluster by allele and not by species[16,115] (FIG. 4b), as expected for neutrally evolving alleles. In the context of host–parasite interactions, the signal of TSP is very clear in MHC genes, in the ABO blood group system of higher primates and in several other vertebrate loci[25,29,64,115–117]. It has also been observed in plants[8,118,119], but rarely in invertebrates[68]. Genomic regions with evidence of TSP are typically enriched with immune function genes[6,8,26,97,105,119,120], suggesting that host–parasite co-evolution might be the driving force behind a genomic signature of TSP[20,120,121]. Experimental data further support a link between TSP and parasitism[105,122].

Given the large timescale involved in TSP, the impact of the co-evolving genes on the immediate genomic neighbourhood of the selected sites will be reduced, because recombination decouples selected sites from linked neutral sites[16,123]. Therefore, long-term polymorphisms of single SNPs are difficult to detect by analysing genomic signatures alone. This can be seen with the SNP in the human haemoglobin gene responsible for the balanced sickle cell anaemia–malaria resistance polymorphism in large parts of Africa. Long-term balancing selection can best be detected when the balanced haplotypes do not recombine (for example, if they are located in inverted regions of the genome or are part of a supergene) or if multiple polymorphisms are together under selection, so that they maintain linkage disequilibrium across polymorphic sites[25,53,123].

## Diffuse co-evolution

The two models of specific co-evolution described above assume that the relevant interactions are between one host and one parasite species. However, interactions may include multiple parasites and/or multiple hosts. This scenario is known as diffuse co-evolution[10].

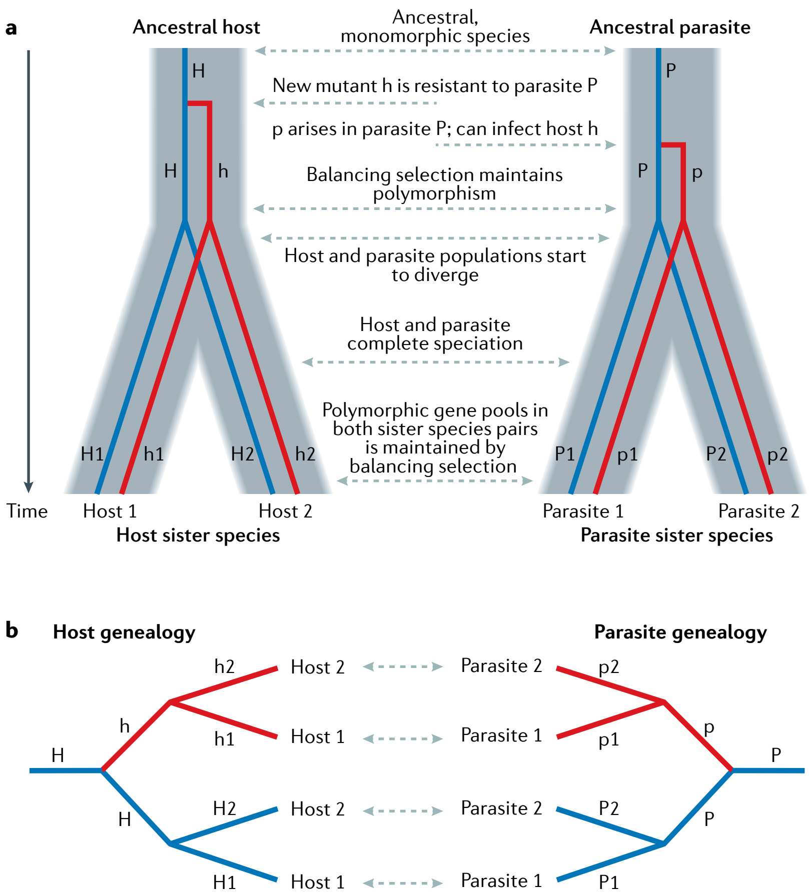

a
Ancestral host
Ancestral, monomorphic species
Ancestral parasite
H
P
New mutant h is resistant to parasite P
p arises in parasite P; can infect host h
H
h
P
p
Balancing selection maintains polymorphism
Host and parasite populations start to diverge
Host and parasite complete speciation
Polymorphic gene pools in both sister species pairs is maintained by balancing selection
H1
h1
H2
h2
P1
p1
P2
p2
Time
Host 1
Host 2
Host sister species
Parasite 1
Parasite 2
Parasite sister species
b
Host genealogy
Parasite genealogy
H
h
H
h2
h1
H2
H1
Host 2
Host 1
Host 2
Host 1
Parasite 2
Parasite 1
Parasite 2
Parasite 1
p2
p1
P2
P1
p
P
P

Fig. 4 | **Idealized scenario for host–parasite co-evolution by long-term balancing selection leading to trans-species polymorphisms. a** | When balancing selection maintains allelic variants over long time periods, trans-species polymorphism (TSP) may be visible, with the polymorphisms existing prior to the split into two species. In contrast to the scenario in FIG. 3b, here, speciation completely blocks gene flow and hybridization. If TSP results from strict long-term co-evolution, similar evolutionary histories are expected for the functionally linked variants in hosts and parasites (FIG. 3a). **b** | For the scenario outlined in part **a**, the genes undergoing long-term balancing selection are expected to cluster by function in a gene tree (red and blue colour), not by species. Host and parasite genealogies should, however, be congruent (that is, have the same topology), indicating that speciation events occurred in parallel.

Functional guilds
Groups of organisms with similar lifestyle characteristics that perform the same ecological function, such as gut parasites, pollinators and filter-feeders.

R-gene
Resistance genes of plants that convey resistance against diseases by producing R proteins.

Diffuse co-evolution refers particularly to the interactions of functional guilds, such as diverse species of herbivores and their plant hosts or parasites and hosts. In its simplest form, co-evolution is diffuse when a host trait and the underlying genes in the genome evolve in response to at least two parasites, or a parasite trait evolves in response to selection caused by more than one host. In such cases, co-evolution is influenced by a combination of frequency dependence (according to the genetic composition of the parasite populations) and density dependence (according to the abundance of the parasite species). Furthermore, the parasite composition may differ in different subpopulations of the host and may change over time (owing to diversifying selection and/or spatio-temporal dynamics)[124–126]. The more hosts or parasites that participate, the more complex the evolution of the participants will be and the harder it will be to predict the genomic signature of the diffusely evolving regions[13].

The need for consideration of diffuse co-evolution is underlined by examples of host loci that interact with multiple parasites. Human variants of CCR5, TRIM5α and APOBECG3 interact with HIV-1, but were suggested to have been under positive selection by other viruses with which they interacted in the past[127–129]. The tomato R-gene *cf2* confers resistance to the parasitic nematode *Globodera rostochiensis* and to the fungus *Cladosporium fulvum*[130]. Parasites from three kingdoms interact in part with the same proteins in *Arabidopsis*[131]. The MHC regions of diverse vertebrates are well known for their interactions with many parasites[125,126,132,133].

The converse, a parasite interacting with more than one host, seems also to be widespread. Sympatric species of sticklebacks are infected by the same parasites, which interact with the host's MHC[116]. The pathogenic bacteria *S. aureus*, *P. viridiflava* and *P. syringae* are known to infect various host species with overlapping genetic mechanisms[74,76,134]. Finally, hosts may even exchange immune genes to fight parasites: for example, related *Arabidopsis* species have been shown to exchange resistance genes through hybridization and introgression[135]. Likewise, parasites may exchange co-evolving genes via horizontal gene transfer to overcome host defences[136].

Co-evolution may become even more diffuse when the interactions of various hosts and parasite species show spatio-temporal dynamics. Co-evolution in such multispecies scenarios is based on variable interactions between changing communities of hosts and parasites. Thus, the signatures observed in the host and parasite genomes cannot be attributed to one ecological setting but result from a history with diverse settings with different interactions in host and parasite communities. Each setting may result in a different evolutionary trajectory and may well include temporal phases of more narrowly defined co-evolution by NFDS or selective sweeps between pairs of antagonists[137]. The signatures of these phases of specific co-evolution will not be recognizable in the genome, unless the ecological setting in which they occur is sufficiently stable for a given period of time.

Models of single hosts co-evolving with multiple parasites support the idea that long-term maintenance of genetic polymorphisms is possible[7,48,59,124,138,139]. Empirical data from systems likely under diffuse co-evolution are consistent with this idea, leading to the widely held belief that the complexity of multispecies interactions maintains genetic diversity, including TSP[92,116,125–127,140–142]. Thus, diffuse co-evolution offers an explanation for the 'missing antagonist problem' that typifies TSP studies, which typically lack knowledge of the parasite that co-evolved with the host in the past[120]. If TSP is caused by diffuse co-evolution, there is no single co-evolving parasite, but an association with a changing pool of parasite species over time and space. Indeed, most of the associations of the human MHC with parasites (such as HIV, West Nile virus, dengue, hepatitis B, hepatitis C, tuberculosis and leprosy[133]) are believed to be rather young — much younger than 5 million years, the approximate date of the last common human–chimpanzee ancestor for which TSP at the MHC was observed[25,29,116]. Thus, the parasites we see interacting with humans now may not be the same as those that interacted with us some million years ago. Malaria, which is a parasite specific to humans, may be an exception. Recent work suggests that human–malaria co-evolution could be as old as the human split from our closest living relatives[143–146]. However, the hypothesis that malaria is a missing antagonist, explaining TSP in humans, requires further investigation.

To further our understanding of TSP it will be necessary to distinguish between cases of specific long-term co-evolution and diffuse co-evolution. Although demonstrating that TSP is the consequence of ancient specific co-evolution is challenging, we can make testable predictions. In the strictest case of specific co-evolution by balancing selection, the aim would be to demonstrate that two closely related host species with evidence of TSP are parasitized by two closely related species of co-evolving parasites (FIG. 4a) and that the genes in question play a functional role in the interaction of the two pairs of antagonists (FIG. 3a). In such a scenario, an overlay of long-term Red Queen dynamics with the co-speciation of the host and the parasite might be observed[14] (FIG. 4b). Such a strict set of conditions may not be very likely, but the requirements may be relaxed to accommodate ecological, historical, epidemiological and biogeographical features[116,138].

## Joint analysis of genomes

Unlike the approaches described above, emerging co-genomic methods jointly analyse the genomes of hosts and those of their parasites. These methods focus on the genes responsible for the phenotypic interactions of the antagonists and allow subsequent analysis of their genomic signatures[72,147–149].

***Identifying the genes involved in host–parasite interactions.*** A co-genomics approach to identify the genes that directly interact with each other in hosts and parasites is best illustrated with a host–parasite matrix that shows the functional specificity of variants of the antagonists that interact with each other to produce phenotypes — disease (compatible) and resistance (incompatible) (FIG. 3a). Although many different matrices have been proposed[150–155], we still do not know much about them in natural host–parasite systems. Studies of interacting host–parasite genes have a long tradition in plant–pathogen systems, starting with H. H. Flor's gene-for-gene system for flax and one of its rust pathogens[156] (for reviews see REFS[22,91]), but only a few examples have so far been confirmed for animal systems[52,53]. It is widely believed that specific interaction matrices like these underlie co-evolution by NFDS and, thus, are responsible for maintaining genetic diversity. Finding the genes underlying such matrices is still cumbersome and time consuming, but this may change with the development of co-genomics approaches that detect interspecies (or intergenomic) linkage disequilibrium (iLD). If the only polymorphism in a population that explains variation in infection success is the interaction between the host's A locus and the parasite's B locus, then infected individuals could only ever carry the host–parasite allele combinations H/P and h/p (FIG. 3a). Other combinations (H/p and h/P) would not produce infections (FIG. 3a). This non-random association between host and parasite alleles produces iLD because the phenotype (infection) depends on the combination of genetic variants in different species. The statistical signal of iLD is only seen at the interacting loci, not at sites in the genetic background, as long as recombination efficiently decouples selected loci from non-selected loci (FIG. 5a). Using iLD is a powerful tool to detect interacting genes for numerous reasons. First, its presence indicates that host and parasite individuals carry alleles related to each other in a fitness context, that is, for a trait (infection) likely under selection. Second, it allows the interacting loci of both antagonists to be pinpointed. Third, it can be used even when only samples of infected hosts (such as infected patients) are available. Fourth, it is free from assumptions about the shape of the interaction matrix, which is important as we currently have little understanding of how the matrix is structured. Last, the method is not limited to model species, as it does not require previous knowledge about the system.

However, statistical limitations are apparent. The method works best when the alleles in question have intermediate frequencies and when the number of tests to be conducted (every host polymorphism is tested in relation to every parasite polymorphism) is not too large relative to the sample size: as the number of tests, and therefore false positives, increases with the product of the numbers of polymorphisms in both antagonists, larger sample sizes are required for antagonists with large genomes. In addition to these limitations, other factors can reduce the power to detect significant iLD. For example, if genes involved in iLD have epistatic (that is, non-additive) interactions with other loci in the same genome, the statistical power to detect them is reduced. Furthermore, simulations of different co-evolutionary models showed that the inferential power of iLD will vary over time due to the underlying dynamics of a host–parasite interaction matrix, and relevant loci will be more identifiable if co-evolution occurs by NFDS than if by selective sweep co-evolution[157]. Finally, multiple infections (that is, more than one parasite genotype being involved in the infection of individual hosts) may cause problems in the analysis. Multiple infections of humans and animals are believed to be very common[158].

Numerous pioneering co-genomics studies have been performed with human patients infected with different parasites. In genome-to-genome analyses, iLD was used to test for associations between SNPs in humans and SNPs in HIV-1 (REF.[147]), hepatitis C[159] and *Streptococcus pneumoniae*[160]. These studies were able to identify previously unknown interacting genes in the host and the parasite, with the virus studies being more powerful because they have smaller genomes than *Streptococcus* and require fewer tests. This method is currently being further developed to increase its sensitivity and to take population stratification into account[161,162]. A modified version of this method was used to test for associations of SNPs in human candidate genes with phylogenetic lineages of *Mycobacterium tuberculosis*, revealing a host SNP–parasite lineage association[163].

Another co-genomics method, Analysis with a Two-Organism Mixed Model (ATOMM), also aims to

Approximate Bayesian computation
A Bayesian statistical approach wherein parameter inference and model selection are conducted in the absence of likelihood functions. Instead, approximate Bayesian computation relies on summary statistics and simulations to infer posterior distributions of parameters and/or models of interest.

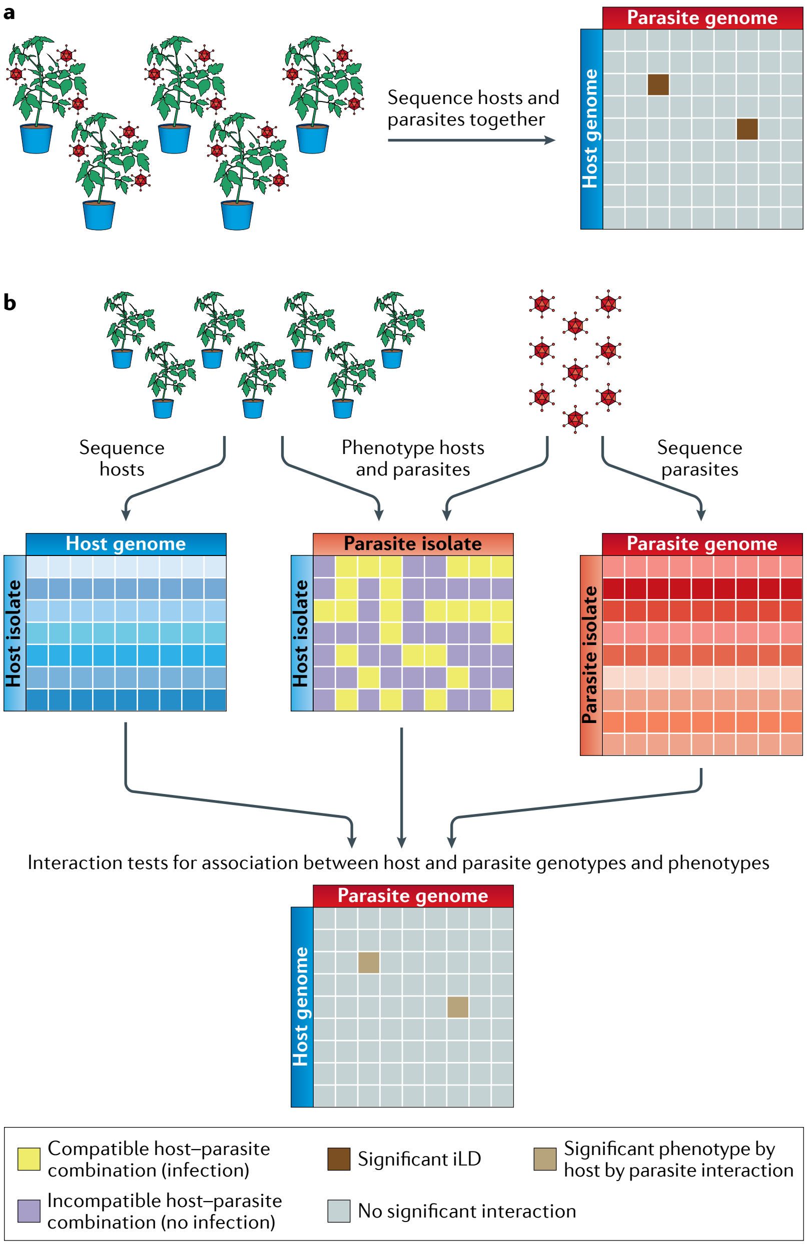

a
Sequence hosts and parasites together
Parasite genome
Host genome
b
Sequence hosts
Phenotype hosts and parasites
Sequence parasites
Host genome
Host isolate
Parasite isolate
Host isolate
Parasite genome
Parasite isolate
Interaction tests for association between host and parasite genotypes and phenotypes
Parasite genome
Host genome
Compatible host–parasite combination (infection)
Significant iLD
Significant phenotype by host by parasite interaction
Incompatible host–parasite combination (no infection)
No significant interaction

Fig. 5 | **Co-genomic approaches to find genes involved in host–parasite interactions.** A co-genomic analysis tests for associations between polymorphisms in host and parasite genomes. **a** | The genome-to-genome method analyses samples of hosts naturally infected with a parasite to find host variants in strong interspecies linkage disequilibrium (iLD) with variants in the parasites. Phenotypic data are not necessary. The figure shows two pairs of sites in the host and parasite genomes that significantly associate with each other, that is, they show strong iLD. **b** | Analysis with a Two-Organism Mixed Model (ATOMM) includes phenotypic data from a host–parasite infection matrix, which are analysed together with polymorphism data for both the host and parasite isolates[164]. The method allows interacting loci in hosts and parasites to be detected. Host and parasite genomes are represented as a series of squares, where each square indicates the position of a polymorphism in the sampled genomes. Shades of blue and red indicate different host and parasite genomes, respectively. Unequal numbers of host and parasite isolates can be used.

find interacting loci in hosts and parasites[164] (FIG. 5b). This method requires experimental data for phenotypes, such as infection and disease symptoms, from all possible combinations of host and parasite genotypes. It allows researchers to map these phenotypes to host and parasite genomes, while accounting for confounding factors such as the phylogenetic or spatial structure.

Although not yet routinely used, co-genomic methods offer exciting possibilities for future studies of host–parasite interactions. The new pairs of loci identified with these methods can then be analysed for their genomic signatures, allowing the picture of the signature of co-evolution in host and parasite genomes to be fine-tuned. The use of the genome-to-genome method with non-model organisms will allow the study of systems previously out of the reach of co-evolutionary research.

***Using co-genomics to infer signatures of the co-evolutionary process.*** As yet, we do not have access to analytical derivations that determine exactly how different co-evolutionary processes affect allele frequencies at loci across host–parasite genomes. However, we can simulate much of the biological complexity that arises as a result of the co-evolutionary process and explore how genomic signatures might differ under particular scenarios[157] (BOX 1). For example, one approach simulates co-evolutionary dynamics to generate allele frequency expectations, which can then be used to look back in time to determine the coalescent properties that might generate these expected allele frequencies at loci in host and parasite genomes[21]. This general idea has been adapted into an inferential framework using approximate Bayesian computation[149] for population genetic simulations of different co-evolutionary scenarios. The results of these simulations can be summarized by both traditional population genetic summaries (TABLE 2) as well as novel summaries of the host and parasite allele frequency spectra concurrently, which are then statistically compared with polymorphism data obtained from host and parasite genomes to determine their similarity. This approach enables the identification of loci that underlie the co-evolutionary process in both host and parasite genomes simultaneously, as well as identifying important eco-evolutionary parameters such as the cost of infectivity. The method can be applied to genome-wide polymorphism data gathered from controlled laboratory experiments or from natural populations. It may be used either in concert with the iLD approach or entirely independently, as the false positive rate for detecting genomic regions undergoing co-evolution is rather low.

## Conclusions and future perspectives

In a world of rapidly increasing transportation and migration, accelerating climate change, deforestation and reforestation, altered agricultural and food handling practises and increasing human and livestock densities, host–parasite contact rates are already very high and are only expected to increase, resulting in more intense reciprocal selection. These changes affect many naturally co-evolving systems in which humans have a vested interest, such as malaria–mosquito, virus (including dengue virus, zika virus and West Nile virus)–mosquito, *Borrelia*–ticks and *Mycoplasma*–house finch, as well as parasites believed to co-evolve with humans[163,165–171]. Although we still have a limited understanding of how

Box 1 | **Emerging co-genomics methodologies**

**Likelihood-free inference**

The biological complexity inherent to the co-evolutionary process and the distinct limitations of population genetic summary statistics, including the lack of a suitable and/or tractable likelihood function, for dealing with certain aspects of this complexity, have driven researchers to seek new solutions[149]. A very successful approach for bypassing some limitations of individual summary statistics and likelihood-based inference has been approximate Bayesian computation. In approximate Bayesian computation, the requirement of a likelihood function is bypassed by the use of simulated data sets. Summarization of these simulations via summary statistics followed by a statistical evaluation, or rejection procedure, identifies which simulations most closely approximate the set of summary statistic values obtained from an observed data set, thereby allowing inference of the posterior probabilities of model parameters of interest[192]. Another exciting approach that allows the aggregate use of simulation and summary statistics to evaluate complex biological dynamics is supervised machine learning. This works on the principle of using training sets to generate a predictive model, for example a simulated data set of the relevant evolutionary scenarios to a given problem of interest. This predictive model is then used to determine how a set of input variables predicts a given response[193]. However, these different approaches should not be considered mutually exclusive[194].

**Ancestral recombination graphs**

To identify a signature of selective sweep or balancing selection using the methods described above, a clear understanding of how a neutral coalescent differs from a non-neutral one is crucial. Given that so much useful information can be gleaned from the flanking genomic regions as well as the locus experiencing selection, traditional summaries of the coalescent that rely on treating loci or genomic windows as (semi-)independent will throw away important, if not crucial, information needed to infer the dynamics of the co-evolutionary process. The ancestral recombination graph (ARG) provides a representation of the relationships among genomic segments, mediated through recombination, as a network[195]. Historically, it has been computationally prohibitive to reconstruct these ARGs in even small genomic data sets. However, recent advances in the data structures required to encode the information of an ARG[196], and in the simulation of complicated, non-neutral perturbations of the coalescent in the context of whole genomes[197], have begun to allow the inference of both selective sweeps and balancing selection from reconstructed ARGs[195,198]. It is perhaps not inconceivable that future advances may well allow for the direct reconstruction of the pairwise relationships of genomic segments in the host and the parasite, which is mediated though the reproductive processes of the host and the parasite, and the infection process that connects them.

co-evolution works in natural populations, the knowledge derived from the state-of-the-art population genomic methodology described here can guide and inform co-evolutionary research and experimentation. We are now in a position to approach new questions and gain a new perspective on old ones.

With few exceptions, we know little about the role of structural variation, such as copy number variation and inversions, in the co-evolutionary process. Copy number variation, which is well known in the MHC, can be maintained by balancing selection (reviewed in[125]). Supergenes have also been shown to play an important part in balanced polymorphisms in various traits[172] and have recently been suggested to have a decisive role in host–parasite co-evolution[53]. Although it is still cumbersome to identify structural polymorphisms in large samples, improved methods will certainly make this an important aspect of co-evolutionary research.

A broad survey of the literature on co-evolution would suggest a taxon-specific division between plant systems, centred on gene-for-gene infection models, and animal systems, focused on matching allele models without invoking costs. It is currently not clear whether this division reflects a bias in our research efforts or has a biological basis. Now, co-genomic methods will enable us to move beyond a few model systems and rapidly collect more data on the type of infection matrix that predominates in natural populations. This information is particularly required for animal systems and non-agricultural plant systems, for which there are only a few examples of interaction matrices. Understanding the costs of resistance will need fitness assays in the presence and absence of the parasite.

Another challenge is to understand how genes within the same genome interact to modify the dynamics with genes in the antagonist. Epistasis has long been thought to be important in host–parasite interactions[48], and epistasis between host resistance loci has been proposed for numerous systems[22,52,53,133,173]. However, the generality and importance of these observations for host–parasite co-evolution is not yet clear. Epistasis is central to the theory of co-evolution and of the evolution and maintenance of genetic recombination. Genetic recombination works to speed up evolutionary responses under antagonistic co-evolution only if the recombining loci show epistatic interactions[48,174,175] because epistasis may create negative linkage disequilibrium among resistance alleles.

Knowing which genes are functionally linked and what epistatic relationships exist would allow us to predict how reciprocal selection acts, whether specific or diffuse co-evolution occurs and what traces we might expect at the genome level. After decades of co-evolution research, we still do not have examples of temporal dynamics allele frequencies and their associated phenotypes in either hosts or parasites, or an understanding of how the corresponding genomic sites in the antagonists are functionally linked. However, recent progress in the field promises change in the near future. Finally, our understanding of the genomic signatures of co-evolution is still largely correlational. Rigorous experimental work, including experimental evolution studies, can help us scrutinize the evidence for co-evolution, link it to genomic signatures and test specific model predictions[56,176–178].

Published online: 28 August 2020

Likelihood function
The analytical formulation of a set of parameters that can be used to assess the fit of a given observed data set to a predetermined model.

Supervised machine learning
Machine learning is a statistical methodology that uses artificial intelligence to automate inferential processes with minimal explicit instruction. Supervised machine learning is a type of machine learning that uses (labelled) training sets to generate a target function when the correspondence between the function of interest and the response variable is known. This target function can then be applied to unclassified (unlabelled) data to make statistical inferences.

Ancestral recombination graph
(ARG). A genealogical or phylogenetic representation of the network of coalescence and recombination events in a collection of orthologous DNA sequences.

1. Majerus, M., Amos, W. & Hurst, G. *Evolution: The Four Billion Year War* (Addison, Wesley Longman, 1996).
2. Jack, R. & Du Pasquier, L. *Evolutionary Concepts in Immunology* (Springer Nature Switzerland, 2019).
3. Fumagalli, M. et al. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* **7**, e1002355 (2011).
4. Karasov, T. L., Horton, M. W. & Bergelson, J. Genomic variability as a driver of plant–pathogen coevolution? *Curr. Opin. Plant. Biol.* **18**, 24–30 (2014).
5. Apanius, V., Penn, D., Slev, P. R., Ruff, L. R. & Potts, W. K. The nature of selection on the major histocompatibility complex. *Crit. Rev. Immunol.* **37**, 75–120 (2017).
6. Lenz, T. L., Hafer, N., Samonte, I. E., Yeates, S. E. & Milinski, M. Cryptic haplotype-specific gamete

selection yields offspring with optimal MHC immune genes. *Evolution* **72**, 2478–2490 (2018).

7. Penman, B. S. & Gupta, S. Detecting signatures of past pathogen selection on human HLA loci: are there needles in the haystack? *Parasitology* **145**, 731–739 (2018).
8. Koenig, D. et al. Long-term balancing selection drives evolution of immunity genes in *Capsella*. *eLife* **8**, e43606 (2019).
9. Guoy, A. & Excoffier, L. Polygenic patterns of adaptive introgression in modern humans are mainly shaped by response to pathogens. *Mol. Biol. Evol.* **37**, 1420–1433 (2020).
10. Janzen, D. H. When is it coevolution. *Evolution* **34**, 611–612 (1980).
11. Woolhouse, M. E. J., Webster, J. P., Domingo, E., Charlesworth, B. & Levin, B. R. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat. Genet.* **32**, 569–577 (2002).
12. Kiester, A. R., Lande, R. & Schemske, D. W. Models of coevolution and speciation in plants and their pollinators. *Am. Nat.* **124**, 220–243 (1984).
13. Wade, M. J. The co-evolutionary genetics of ecological communities. *Nat. Rev. Genet.* **8**, 185–195 (2007).
14. de Vienne, D. M. et al. Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. *N. Phytol.* **198**, 347–385 (2013).
15. Ebert, D. Host–parasite coevolution: insights from the *Daphnia*–parasite model system. *Curr. Opin. Microbiol.* **11**, 290–301 (2008).
16. Charlesworth, D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* **2**, 379–384 (2006).
**This paper is an authoritative review on the population genetics of balancing selection**.
17. Zivkovic, D., John, S., Verin, M., Stephan, W. & Tellier, A. Neutral genomic signatures of host–parasite coevolution. *BMC Evol. Biol.* **19**, 230 (2019).
18. Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (Oxford Univ. Press, 1995).
19. Dodds, P. N. & Rathjen, J. P. Plant immunity: towards an integrated view of plant–pathogen interactions. *Nat. Rev. Genet.* **11**, 539–548 (2010).
20. Sironi, M., Cagliani, R., Forni, D. & Clerici, M. Evolutionary insights into host–pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* **16**, 224–236 (2015).
21. Tellier, A., Moreno-Gamez, S. & Stephan, W. Speed of adaptation and genomic footprints of host–parasite coevolution under arms race and trench warfare dynamics. *Evolution* **68**, 2211–2224 (2014).
22. Thrall, P. H., Barrett, L. G., Dodds, P. N. & Burdon, J. J. Epidemiological and evolutionary outcomes in gene-for-gene and matching allele models. *Front. Plant Sci.* **6**, 1084 (2016).
23. Ebert, D. Open questions: what are the genes underlying antagonistic coevolution? *BMC Biol.* **16**, 114 (2018).
24. Fischer, M. C., Foll, M., Heckel, G. & Excoffier, L. Continental-scale footprint of balancing and positive selection in a small rodent (*Microtus arvalis*). *PLoS ONE* **9**, e112332 (2014).
25. Leffler, E. M. et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**, 1578–1582 (2013).
**This study reports on many sites with TSPs and ancient balancing selection in genomes of humans and chimpanzees**.
26. Key, F. M., Teixeira, J. C., de Filippo, C. & Andres, A. M. Advantageous diversity maintained by balancing selection in humans. *Curr. Opin. Genet. Dev.* **29**, 45–51 (2014).
27. Schweizer, R. M. et al. Natural selection and origin of a melanistic allele in North American Gray Wolves. *Mol. Biol. Evol.* **35**, 1190–120 (2018).
28. Bitarello, B. D. et al. Signatures of long-term balancing selection in human genomes. *Genome Biol. Evol.* **10**, 939–955 (2018).
29. Cheng, X. & DeGiorgio, M. Detection of shared balancing selection in the absence of trans-species polymorphism. *Mol. Biol. Evol.* **36**, 177–199 (2019).
30. Enard, D., Cai, L., Gwennap, C. & Petrov, D. A. Viruses are a dominant driver of protein adaptation in mammals. *eLife* **5**, e124699 (2016).
31. Schirrmann, M. K. et al. Genomewide signatures of selection in *Epichloe* reveal candidate genes for host specialization. *Mol. Ecol.* **27**, 3070–3086 (2018).
32. Persoons, A. et al. The escalatory Red Queen: population extinction and replacement following arms race dynamics in poplar rust. *Mol. Ecol.* **26**, 1902–1918 (2017).
33. Mohd-Assaad, N., McDonald, B. A. & Croll, D. Genome-wide detection of genes under positive selection in worldwide populations of the barley scald pathogen. *Genome Biol. Evol.* **10**, 1315–1332 (2018).
34. Badouin, H. et al. Widespread selective sweeps throughout the genome of model plant pathogenic fungi and identification of effector candidates. *Mol. Ecol.* **26**, 2041–2062 (2017).
35. Obbard, D. J., Gordon, K. H. J., Buck, A. H. & Jiggins, F. M. The evolution of RNAi as a defence against viruses and transposable elements. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 99–115 (2009).
36. Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat. Rev. Genet.* **14**, 827–839 (2013).
37. Hermisson, J. & Pennings, P. S. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol. Evol.* **8**, 700–716 (2017).
38. Hahn, M. W. *Molecular Population Genetics* (Sinauer Associates, 2018).
39. McCarthy, K. R., Kirmaier, A., Autissier, P. & Johnson, W. E. Evolutionary and functional analysis of old world primate TRIM5 reveals the ancient emergence of primate lentiviruses and convergent evolution targeting a conserved capsid interface. *PLoS Pathog.* **11**, e1005085 (2015).
40. Elena, S. F., Cooper, V. S. & Lenski, R. E. Punctuated evolution caused by selection of rare beneficial mutations. *Science* **272**, 1802–1804 (1996).
41. Anderson, T. J. C. et al. Population parameters underlying an ongoing soft sweep in southeast asian malaria parasites. *Mol. Biol. Evol.* **34**, 131–144 (2017).
42. Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet.* **11**, e1005004 (2015).
43. Sanchez-Vallet, A. et al. The genome biology of effector gene evolution in filamentous plant pathogens. *Ann. Rev. Phytopathol.* **56**, 21–40 (2018).
44. Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4**, 24–33 (1994).
45. Raffaele, S. et al. Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science* **330**, 1540–1543 (2010).
46. Croll, D. & McDonald, B. A. The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathog.* **8**, e1002608 (2012).
47. Clarke, B. C. in *Genetic Aspects of Host–Parasite Relationships* (eds A. E. R. Taylor & R. M. Muller) 87–104 (Blackwell, 1976).
48. Hamilton, W. D., Axelrod, R. & Tanese, R. Sexual reproduction as an adaptation to resist parasites. *Proc. Natl Acad. Sci. USA* **87**, 3566–3573 (1990).
49. Fenton, A., Antonovics, J. & Brockhurst, M. A. Inverse-gene-for-gene infection genetics and coevolutionary dynamics. *Am. Nat.* **174**, E230–E242 (2009).
50. Schmid-Hempel, P. Evolutionary Parasitology: The Integrated Study of Infections, Immunology, Ecology, and Genetics (Oxford Univ. Press, 2011).
**This textbook comprehensively summarizes the field of host–parasite evolution and co-evolution**.
51. Ben Khalifa, M., Simon, V., Fakhfakh, H. & Moury, B. Tunisian potato virus Y isolates with unnecessary pathogenicity towards pepper: support for the matching allele model in eIF4E resistance–potyvirus interactions. *Plant Pathol.* **61**, 441–447 (2012).
52. Luijckx, P., Fienberg, H., Duneau, D. & Ebert, D. A matching-allele model explains host resistance to parasites. *Curr. Biol.* **23**, 1085–1088 (2013).
**This study provides an early example of a well worked out, matching allele host–parasite interaction matrix**.
53. Bento, G. et al. The genetic basis of resistance and matching-allele interactions of a host–parasite system: the *Daphnia magna–Pasteuria ramosa* model. *PLoS Genet.* **13**, e1006596 (2017).
54. King, K. C., Jokela, J. & Lively, C. M. Parasites, sex, and clonal diversity in natural snail populations. *Evolution* **65**, 1474–1481 (2011).
55. Ashby, B. & Boots, M. Multi-mode fluctuating selection in host–parasite coevolution. *Ecol. Lett.* **20**, 357–365 (2017).
56. Papkou, A. et al. The genomic basis of Red Queen dynamics during rapid reciprocal host–pathogen coevolution. *Proc. Natl Acad. Sci. USA* **116**, 923–928 (2019).
**This study of experimental evolution with nematodes and a bacterial pathogen demonstrates the complexity of co-evolutionary interactions emerging in seemingly simple systems**.
57. Koskella, B. & Lively, C. M. Evidence for negative frequency-dependent selection during experimental coevolution of a freshwater snail and a sterilizing trematode. *Evolution* **63**, 2213–2221 (2009).
58. Lively, C. M. Habitat heterogeneity, host population structure, and parasite local adaptation. *J. Hered.* **109**, 29–37 (2018).
59. Ejsmond, M. J., Babik, W. & Radwan, J. MHC allele frequency distributions under parasite-driven selection: a simulation model. *BMC Evol. Biol.* **10**, 332 (2010).
60. Fijarczyk, A. & Babik, W. Detecting balancing selection in genomes: limits and prospects. *Mol. Ecol.* **24**, 3529–3545 (2015).
61. Bubb, K. L. et al. Scan of human genome reveals no new loci under ancient balancing selection. *Genetics* **173**, 2165–2177 (2006).
62. Cagliani, R. et al. The signature of long-standing balancing selection at the human defensin β-1 promoter. *Genome Biol.* **9**, R143 (2008).
63. Fumagalli, M. et al. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res.* **19**, 199–212 (2009).
64. Segurel, L. et al. The ABO blood group is a trans-species polymorphism in primates. *Proc. Natl Acad. Sci. USA* **109**, 18493–18498 (2012).
65. Bergelson, J., Kreitman, M., Stahl, E. A. & Tian, D. C. Evolutionary dynamics of plant R-genes. *Science* **292**, 2281–2285 (2001).
66. Hoerger, A. C. et al. Balancing selection at the tomato *RCR3* guardee gene family maintains variation in strength of pathogen defense. *PLoS Genet.* **8**, e1002813 (2012).
67. Llaurens, V., Whibley, A. & Joron, M. Genetic architecture and balancing selection: the life and death of differentiated variants. *Mol. Ecol.* **26**, 2430–2448 (2017).
68. Croze, M. et al. A genome-wide scan for genes under balancing selection in *Drosophila melanogaster*. *BMC Evol. Biol.* **17**, 15 (2017).
69. Buckley, J., Holub, E. B., Koch, M. A., Vergeer, P. & Mable, B. K. Restriction associated DNA-genotyping at multiple spatial scales in *Arabidopsis lyrata* reveals signatures of pathogen-mediated selection. *BMC Genomics* **19**, 496 (2018).
70. Wu, Q. et al. Long-term balancing selection contributes to adaptation in *Arabidopsis* and its relatives. *Genome Biol.* **18**, 217 (2017).
71. Unckless, R. L., Howick, V. M. & Lazzaro, B. P. Convergent balancing selection on an antimicrobial peptide in *Drosophila*. *Curr. Biol.* **26**, 257–262 (2016).
72. Bartoli, C. & Roux, F. Genome-wide association studies in plant pathosystems: toward an ecological genomics approach. *Front. Plant. Sci.* **8**, 763 (2017).
73. Power, R. A., Parkhill, J. & de Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.* **18**, 41–50 (2017).
74. Thomas, J. C., Godfrey, P. A., Feldgarden, M. & Robinson, A. Candidate targets of balancing selection in the genome of *Staphylococcus aureus*. *Mol. Biol. Evol.* **29**, 1175–1186 (2012).
75. Zhang, L. F., Thomas, J. C., Didelot, X. & Robinson, D. A. Molecular signatures identify a candidate target of balancing selection in an arcD-like gene of *Staphylococcus epidermidis*. *J. Mol. Evol.* **75**, 43–54 (2012).
76. Guttman, D. S., Gropp, S. J., Morgan, R. L. & Wang, P. W. Diversifying selection drives the evolution of the type III secretion system pilus of *Pseudomonas syringae*. *Mol. Biol. Evol.* **23**, 2342–2354 (2006).
77. Castillo, J. A. & Agathos, S. N. A genome-wide scan for genes under balancing selection in the plant pathogen *Ralstonia solanacearum*. *BMC Evol. Biol.* **19**, 123 (2019).
78. Ryabov, E. V. et al. Dynamic evolution in the key honey bee pathogen deformed wing virus: novel insights into virulence and competition using reverse genetics. *PLoS Biol.* **17**, e3000502 (2019).
79. Andras, J. P., Fields, P. D., Du Pasquier, L., Fredericksen, M. & Ebert, D. Genome-wide association analysis identifies a genetic basis of infectivity in a model bacterial pathogen. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msaa173 (2020).

80. Brackney, D. E., Beane, J. E. & Ebel, G. D. RNAi targeting of West Nile virus in mosquito midguts promotes virus diversification. *PLoS Pathog.* **5**, e1000502 (2009).

81. Brackney, D. E., Schirtzinger, E. E., Harrison, T. D., Ebel, G. D. & Hanley, K. A. Modulation of flavivirus population diversity by RNA interference. *J. Virol.* **89**, 4035–4039 (2015).

82. Obbard, D. J., Jiggins, F. M., Halligan, D. L. & Little, T. J. Natural selection drives extremely rapid evolution in antiviral RNAi genes. *Curr. Biol.* **16**, 580–585 (2006).

83. Araki, H. et al. Presence/absence polymorphism for alternative pathogenicity islands in *Pseudomonas viridiflava*, a pathogen of *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **103**, 5887–5892 (2006).

84. Nygaard, S. et al. Long- and short-term selective forces on malaria parasite genomes. *PLoS Genet.* **6**, e1001099 (2010).

85. Ochola, L. I. et al. Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in plasmodium falciparum. *Mol. Biol. Evol.* **27**, 2344–2351 (2010).

86. Salathe, M., Kouyos, R. D., Regoes, R. R. & Bonhoeffer, S. Rapid parasite adaptation drives selection for high recombination rates. *Evolution* **62**, 295–300 (2008).

87. Gokhale, C. S., Papkou, A., Traulsen, A. & Schulenburg, H. Lotka–Volterra dynamics kills the Red Queen: population size fluctuations and associated stochasticity dramatically change host–parasite coevolution. *BMC Evol. Biol.* **13**, 254 (2013).

88. Schenk, H., Schulenburg, H. & Traulsen, A. How long do Red Queen dynamics survive under genetic drift? A comparative analysis of evolutionary and eco-evolutionary models. *BMC Evol. Biol.* **20**, 8 (2020).

89. MacPherson, A., Keeling, M. J. & Otto, S. P. Coevolution does not slow the rate of loss of heterozygosity in a stochastic host-parasite model with constant population size. *bioRxiv* https://doi.org/10.1101/2020.04.07.024661 (2020).

90. Thrall, P. H. & Burdon, J. J. Effect of resistance variation in a natural plant host–pathogen metapopulation on disease dynamics. *Plant. Pathol.* **49**, 767–773 (2000).

91. Brown, J. K. M. & Tellier, A. Plant–parasite coevolution: bridging the gap between genetics and ecology. *Annu. Rev. Phytopathol.* **49**, 345–367 (2011).

92. Radwan, J., Babik, W., Kaufman, J., Lenz, T. L. & Winternitz, J. Advances in the evolutionary understanding of MHC polymorphism. *Trends Genet.* **36**, 298–311 (2020).

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

94. Eizaguirre, C., Lenz, T. L., Kalbe, M. & Milinski, M. Divergent selection on locally adapted major histocompatibility complex immune genes experimentally proven in the field. *Ecol. Lett.* **15**, 723–731 (2012).

95. Rico, Y. et al. Spatial patterns of immunogenetic and neutral variation underscore the conservation value of small, isolated American badger populations. *Evol. Appl.* **9**, 1271–1284 (2016).

96. Jousimo, J. et al. Ecological and evolutionary effects of fragmentation on infectious disease dynamics. *Science* **344**, 1289–1293 (2014).

97. Crispo, E. et al. The evolution of the major histocompatibility complex in upstream versus downstream river populations of the longnose dace. *Ecol. Evol.* **7**, 3297–3311 (2017).

98. Keller, M. F. et al. Trans-ethnic meta-analysis of white blood cell phenotypes. *Hum. Mol. Genet.* **23**, 6944–6960 (2014).

99. Morgan, A. D., Gandon, S. & Buckling, A. The effect of migration on local adaptation in a coevolving host–parasite system. *Nature* **437**, 253–256 (2005).

100. Thrall, P. H. et al. Rapid genetic change underpins antagonistic coevolution in a natural host–pathogen metapopulation. *Ecol. Lett.* **15**, 425–435 (2012).

101. Kawecki, T. J. & Ebert, D. Conceptual issues in local adaptation. *Ecol. Lett.* **7**, 1225–1241 (2004).

102. Croll, D. & McDonald, B. A. The genetic basis of local adaptation for pathogenic fungi in agricultural ecosystems. *Mol. Ecol.* **26**, 2027–2040 (2017).

103. Laine, A. L., Burdon, J. J., Dodds, P. N. & Thrall, P. H. Spatial variation in disease resistance: from molecules to metapopulations. *J. Ecol.* **99**, 96–112 (2011).

104. Bolnick, D. I. & Stutz, W. E. Frequency dependence limits divergent evolution by favouring rare immigrants over residents. *Nature* **546**, 285–288 (2017).
**This experimental study with fish shows that rare immigrants have an advantage over resident genotypes and demonstrates elegantly that resistance genes have higher effective migration rates.**

105. Phillips, K. P. et al. Immunogenetic novelty confers a selective advantage in host–pathogen coevolution. *Proc. Natl Acad. Sci. USA* **115**, 1552–1557 (2018).

106. Rico, Y., Morris-Pocock, J., Zigouris, J., Nocera, J. J. & Kyle, C. J. Lack of spatial immunogenetic structure among wolverine (*Gulo gulo*) populations suggestive of broad scale balancing selection. *PLoS ONE* **10**, e0140170 (2015).

107. Leducq, J. B. et al. Effect of balancing selection on spatial genetic structure within populations: theoretical investigations on the self-incompatibility locus and empirical studies in *Arabidopsis halleri*. *Heredity* **106**, 319–329 (2011).

108. Castric, V., Bechsgaard, J., Schierup, M. H. & Vekemans, X. Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genet.* **4**, e1000168 (2008).

109. Hoban, S. et al. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am. Nat.* **188**, 379–397 (2016).

110. Borg, A. A., Pedersen, S. A., Jensen, H. & Westerdahl, H. Variation in MHC genotypes in two populations of house sparrow (*Passer domesticus*) with different population histories. *Ecol. Evol.* **1**, 145–159 (2011).

111. Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).

112. Fields, P. D., Reisser, C., Dukic, M., Haag, C. R. & Ebert, D. Genes mirror geography in *Daphnia magna*. *Mol. Ecol.* **24**, 4521–4536 (2015).

113. Thompson, J. N. *The Geographic Mosaic of Coevolution* (Univ. of Chicago Press, 2005).

114. Laine, A. L., Barres, B., Numminen, E. & Siren, J. P. Variable opportunities for outcrossing result in hotspots of novel genetic variation in a pathogen metapopulation. *eLife* **8**, e47091 (2019).

115. Klein, J. *Immunology* (Blackwell, 1990).

116. Lenz, T. L., Eizaguirre, C., Kalbe, M. & Milinski, M. Evaluating patterns of convergent evolution and trans-species polymorphism at MHC immunogenes in two sympatric stickleback species. *Evolution* **67**, 2400–2412 (2013).
**This study demonstrates TSP in two sympatric stickleback fish sharing the same parasites. The authors were able to rule out convergent evolution as an alternative explanation for TSP.**

117. Tesicky, M. & Vinkler, M. Trans-species polymorphism in immune genes: general pattern or MHC-restricted phenomenon? *J. Immunol. Res.* https://doi.org/10.1155/2015/838035 (2015).

118. Mboup, M., Fischer, I., Lainer, H. & Stephan, W. Trans-species polymorphism and allele-specific expression in the CBF gene family of wild tomatoes. *Mol. Biol. Evol.* **29**, 3641–3652 (2012).

119. Novikova, P. Y. et al. Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).

120. Azevedo, L., Serrano, C., Amorim, A. & Cooper, D. N. Trans-species polymorphism in humans and the great apes is generally maintained by balancing selection that modulates the host immune response. *Hum Genomics* **9**, 21 (2015).

121. Lenz, T. L. Computational prediction of MHC II–antigen binding supports divergent allele advantage and explains trans-species polymorphism. *Evolution* **65**, 2380–2390 (2011).

122. Eizaguirre, C., Lenz, T. L., Kalbe, M. & Milinski, M. Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. *Nat. Commun.* **3**, 621 (2012).

123. Gao, Z. Y., Przeworski, M. & Sella, G. Footprints of ancient-balanced polymorphisms in genetic variation data from closely related species. *Evolution* **69**, 431–446 (2015).

124. Hedrick, P. W. Pathogen resistance and genetic variation at MHC loci. *Evolution* **56**, 1902–1908 (2002).

125. Eizaguirre, C. & Lenz, T. L. Major histocompatibility complex polymorphism: dynamics and consequences of parasite-mediated local adaptation in fishes. *J. Fish. Biol.* **77**, 2023–2047 (2010).

126. Osborne, M. J., Pilger, T. J., Lusk, J. D. & Turner, T. F. Spatio-temporal variation in parasite communities maintains diversity at the major histocompatibility complex class II in the endangered Rio Grande silvery minnow. *Mol. Ecol.* **26**, 471–489 (2017).

127. Daugherty, M. D. & Malik, H. S. Rules of engagement: molecular insights from host–virus arms races. *Annu. Rev. Genet.* **46**, 677–700 (2012).

128. Cagliani, R. et al. A positively selected APOBEC3H haplotype is associated with natural resistance to HIV-1 infection. *Evolution* **65**, 3311–3322 (2011).

129. Davis, Z. H. et al. Global mapping of herpesvirus–host protein complexes reveals a transcription strategy for late genes. *Mol. Cell* **57**, 349–360 (2015).

130. Lozano-Torres, J. L. et al. Dual disease resistance mediated by the immune receptor Cf-2 in tomato requires a common virulence target of a fungus and a nematode. *Proc. Natl Acad. Sci. USA* **109**, 10119–10124 (2012).

131. Wessling, R. et al. Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. *Cell Host Microbe* **16**, 364–375 (2014).

132. Wegner, K. M., Kalbe, M., Kurtz, J., Reusch, T. B. H. & Milinski, M. Parasite selection for immunogenetic optimality. *Science* **301**, 1343–1343 (2003).

133. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18**, 76 (2017).

134. Karasov, T. L., Barrett, L., Hershberg, R. & Bergelson, J. Similar levels of gene content variation observed for *Pseudomonas syringae* populations extracted from single and multiple host species. *Plos ONE* **12**, e0184195 (2017).

135. Bechsgaard, J., Jorgensen, T. H. & Schierup, M. H. Evidence for adaptive introgression of disease resistance genes among closely related arabidopsis species. *Genes Genomes Genet.* **7**, 2677–2683 (2017).

136. Gluck-Thaler, E. & Slot, J. C. Dimensions of horizontal gene transfer in eukaryotic microbial pathogens. *PLoS Pathog.* **11**, e1005156 (2015).

137. Campbell, M. C., Ashong, B., Teng, S. L., Harvey, J. & Cross, C. N. Multiple selective sweeps of ancient polymorphisms in and around LT alpha located in the MHC class III region on chromosome 6. *BMC Evol. Biol.* **19**, 218 (2019).

138. Karasov, T. L. et al. The long-term maintenance of a resistance polymorphism through diffuse interactions. *Nature* **512**, 436–440 (2014).

139. Rabajante, J. F. et al. Red Queen dynamics in multi-host and multi-parasite interaction system. *Sci. Rep.* **5**, 10004 (2015).

140. Kamath, P. L., Turner, W. C., Kusters, M. & Getz, W. M. Parasite-mediated selection drives an immunogenetic trade-off in plains zebras (Equus quagga). *Proc. Biol. Sci.* **281**, 20140077 (2014).

141. Nadeem, A. & Wahl, L. M. Prophage as a genetic reservoir: promoting diversity and driving innovation in the host community. *Evolution* **71**, 2080–2089 (2017).

142. Fortuna, M. A. et al. Coevolutionary dynamics shape the structure of bacteria–phage infection networks. *Evolution* **73**, 1001–1011 (2019).

143. Silva, J. C. et al. Genome sequences reveal divergence times of malaria parasite lineages. *Parasitology* **138**, 1737–1749 (2011).

144. Galen, S. C. et al. The polyphyly of *Plasmodium*: comprehensive phylogenetic analyses of the malaria parasites (order Haemosporida) reveal widespread taxonomic conflict. *Roy. Soc. Open Sci.* **5**, 171780 (2018).

145. Otto, T. D. et al. Genomes of all known members of a *Plasmodium* subgenus reveal paths to virulent human malaria. *Nat. Microbiol.* **3**, 687–697 (2018).

146. Pacheco, M. A. et al. Mode and rate of evolution of haemosporidian mitochondrial genomes: timing the radiation of avian parasites. *Mol. Biol. Evol.* **35**, 383–403 (2018).

147. Bartha, I. et al. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *eLife* **2**, e01123 (2013).
**This study describes a pioneering method in co-genomics, applied to interacting genomic sites in hosts and parasites.**

148. Lees, J. A., Tonkin-Hill, G. & Bentley, S. D. GENOME WATCH stronger together. *Nat. Rev. Microbiol.* **15**, 516–516 (2017).

149. Märkle, H. & Tellier, A. Inference of coevolutionary dynamics and parameters from host and parasite polymorphism data of repeated experiments. *PLoS Comput. Biol.* **16**, e1007668 (2020).

150. Otto, S. P. & Nuismer, S. L. Species interactions and the evolution of sex. *Science* **304**, 1018–1020 (2004).
151. Tellier, A. & Brown, J. K. M. Polymorphism in multilocus host–parasite coevolutionary interactions. *Genetics* **177**, 1777–1790 (2007).
152. Engelstadter, J. & Bonhoeffer, S. Red Queen dynamics with non-standard fitness interactions. *PLoS Comput. Biol.* **5**, e1000469 (2009).
153. Best, A. et al. The evolution of host–parasite range. *Am. Nat.* **176**, 63–71 (2010).
154. Fenton, A., Antonovics, J. & Brockhurst, M. A. Two-step infection processes can lead to coevolution between functionally independent infection and resistance pathways. *Evolution* **66**, 2030–2041 (2012).
155. Kwiatkowski, M., Engelstadter, J. & Vorburger, C. On genetic specificity in symbiont-mediated host–parasite coevolution. *PLoS Comput. Biol.* **8**, e1002633 (2012).
156. Flor, H. H. Host–parasite interaction in flax rust — its genetics and other implications. *Phytopathology* **45**, 680–685 (1955).
157. Märkle, H., Tellier, A. & John, S. Cross-species association statistics for genome-wide studies of host and parasite polymorphism data. Preprint at *bioRxiv* https://doi.org/10.1101/726166 (2019).
158. Balmer, O. & Tanner, M. Prevalence and implications of multiple-strain infections. *Lancet Infect. Dis.* **11**, 868–878 (2011).
159. Ansari, M. A. et al. Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nat. Genet.* **49**, 666–673 (2017).
**This study describes a strong example of the application of co-genomics to find interacting loci in humans infected with hepatitis C virus**.
160. Lees, J. A. et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat. Commun.* **10**, 2176 (2019).
161. Naret, O. et al. Correcting for population stratification reduces false positive and false negative results in joint analyses of host and pathogen genomes. *Front. Genet.* **9**, 266 (2018).
162. Ansari, M. A. et al. Interferon λ4 impacts the genetic diversity of hepatitis C virus. *eLife* **8**, e42463 (2019).
163. McHenry, M. L. et al. Interaction between host genes and mycobacterium tuberculosis lineage can affect tuberculosis severity: evidence for coevolution? *PLoS Genet.* **16**, e1008728 (2020).
164. Wang, M. Y. et al. Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. *Proc. Natl Acad. Sci. USA* **115**, E5440–E5449 (2018).
**This study describes the development of a powerful co-genomics method that utilizes data from an interaction matrix of all combinations of host and parasite genotypes to find the genomic sites that underlie the interaction**.
165. Hill, A. V. S., Jepson, A., Plebanski, M. & Gilbert, S. C. Genetic analysis of host–parasite coevolution in human malaria. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **352**, 1317–1325 (1997).
166. Lacroix, R., Mukabana, W. R., Gouagna, L. C. & Koella, J. C. Malaria infection increases attractiveness of humans to mosquitoes. *PLoS Biol.* **3**, e298 (2005).
167. Bonneaud, C. et al. Rapid evolution of disease resistance is accompanied by functional changes in gene expression in a wild bird. *Proc. Natl Acad. Sci. USA* **108**, 7866–7871 (2011).
168. Bonneaud, C. et al. Rapid antagonistic coevolution in an emerging pathogen and its vertebrate host. *Curr. Biol.* **28**, 2978–2983 (2018).
169. Tschirren, B. et al. Polymorphisms at the innate immune receptor TLR2 are associated with *Borrelia* infection in a wild rodent population. *Proc. R. Soc. Lond B Biol Sci.* **280**, 20130364 (2013).
170. Heeney, J. L., Dalgleish, A. G. & Weiss, R. A. Origins of HIV and the evolution of resistance to AIDS. *Science* **313**, 462–466 (2006).
171. Hertz, T. et al. Mapping the landscape of host–pathogen coevolution: HLA class I binding and its relationship with evolutionary conservation in human and viral proteins. *J. Virol.* **85**, 1310–1321 (2011).
172. Schwander, T., Libbrecht, R. & Keller, L. Supergenes and complex phenotypes. *Curr. Biol.* **24**, R288–R294 (2014).
173. Lenz, T. L. et al. Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. *Nat. Genet.* **47**, 1085–1090 (2015).
174. Salathe, M., Kouyos, R. D. & Bonhoeffer, S. The state of affairs in the kingdom of the Red Queen. *Trends Ecol. Evol.* **23**, 439–445 (2008).
175. da Silva, J. & Galbraith, J. D. Hill–Robertson interference maintained by Red Queen dynamics favours the evolution of sex. *J. evol. Biol.* **30**, 994–1010 (2017).
176. Kubinak, J. L. et al. Experimental viral evolution reveals major histocompatibility complex polymorphisms as the primary host factors controlling pathogen adaptation and virulence. *Genes Immun.* **14**, 365–372 (2013).
177. Brockhurst, M. A. & Koskella, B. Experimental coevolution of species interactions. *Trends Ecol. Evol.* **28**, 367–375 (2013).
178. Retel, C. et al. The feedback between selection and demography shapes genomic diversity during coevolution. *Sci. Adv.* **5**, eaax0530 (2019).
179. Figueroa, F., Günther, E. & Klein, J. MHC polymorphism pre-dating speciation. *Nature* **335**, 265–267 (1988).
180. Mcdonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
181. Eyre-Walker, A. & Keightley, P. D. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* **26**, 2097–2108 (2009).
182. Nielsen, R. Molecular signatures of natural selection. *Ann. Rev. Genet.* **39**, 197–218 (2005).
183. Siewert, K. M. & Voight, B. F. Detecting long-term balancing selection using allele frequency correlation. *Mol. Biol. Evol.* **34**, 2996–3005 (2017).
184. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
185. Messer, P. W. & Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.* **28**, 659–669 (2013).
186. DeGiorgio, M., Lohmueller, K. E. & Nielsen, R. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet.* **10**, e1004561 (2014).
187. Kim, Y. & Stephan, W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765–777 (2002).
188. Kim, Y. & Nielsen, R. Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**, 1513 (2004).
189. DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I. & Nielsen, R. SweepFinder 2: increased sensitivity, robustness and flexibility. *Bioinformatics* **32**, 1895–1897 (2016).
190. Pavlidis, P., Živković, D., Stamatakis, A. & Alachiotis, N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol. Biol. Evol.* **30**, 2224–2234 (2013).
191. Alachiotis, N., Stamatakis, A. & Pavlidis, P. OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics* **28**, 2274–2275 (2012).
192. Csilléry, K., Blum, M. G. B., Gaggiotti, O. E. & François, O. Approximate Bayesian computation (ABC) in practice. *Trends Ecol. Evol.* **25**, 410–418 (2010).
193. Schrider, D. R. & Kern, A. D. Supervised machine learning for population genetics: a new paradigm. *Trends Genet.* **34**, 301–312 (2018).
**This paper offers an accessible description of both present applications and possible future developments of supervised machine learning for understanding signatures of selection in genomic-scale data**.
194. Raynal, L. et al. ABC random forests for Bayesian parameter inference. *Bioinformatics* **35**, 1720–1728 (2018).
195. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* **10**, e1004342 (2014).
196. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Comput. Biol.* **12**, e1004842 (2016).
197. Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W. & Ralph, P. L. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Mol. Ecol. Resour.* **19**, 552–566 (2019).
**This paper describes the implementation of tree-sequence recording into the already multifaceted and powerful SLiM simulation framework, and provides one of the most important schemes needed to model neutral and non-neutral dynamics on genome-scale data**.
198. Hejase, H. A., Dukler, N. & Siepel, A. From summary statistics to gene trees: methods for inferring positive selection. *Trends Genet.* **36**, 243–258 (2020).
**This paper is an exceptionally comprehensive review of both historical and present approaches for detecting forms of positive selection. Although the focus is on positive selection, many of the focal methodologies would, with some modification, be applicable for detecting the many signatures of host–parasite co-evolution**.

## Acknowledgements

The authors thank the Ebert laboratory for fruitful discussion and S. Zweizig for comments on the language of the manuscript. This work is supported by a grant from the Swiss National Science Foundation.

## Author contributions

The authors contributed equally to all aspects of the article.

## Competing interests

The authors declare no competing interests.

## Peer review information

*Nature Reviews Genetics* thanks A. Tellier and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

## Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.