Molecular Ecology (2005) **14**, 3623–3642 doi: 10.1111/j.1365-294X.2005.02690.x

INVITED REVIEW

Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies

A. VASEMÄGI and C. R. PRIMMER
Department of Biology, 20014, University of Turku, Finland

Abstract

Strategies for the identification of functional genetic variation underlying phenotypic traits of ecological and evolutionary importance have received considerable attention in the literature recently. This paper aims to bring together and compare the relative strengths and limitations of various potentially useful research strategies for dissecting functionally important genetic variation in a wide range of organisms. We briefly explore the relative strengths and limitations of traditional and emerging approaches and evaluate their potential use in free-living populations. While it is likely that much of the progress in functional genetic analyses will rely on progress in traditional model species, it is clear that with prudent choices of methods and appropriate sampling designs, much headway can be also made in a diverse range of species. We suggest that combining research approaches targeting different functional and biological levels can potentially increase understanding the genetic basis of ecological and evolutionary processes both in model and non-model organisms.

Keywords: admixture mapping, allele-specific expression, association analysis, gene expression, neutrality test, QTL mapping

Received 30 April 2005; revision accepted 30 June 2005

Introduction

Identification of functional genetic variation underlying phenotypic traits of evolutionary, agronomic or medical importance is one of the most intriguing tasks in biology. Until recently, extensive cross-field research directed at understanding the role of functional genetic variation in evolution has been primarily limited to the traditional genetic model organisms, i.e. species with short generation times and well characterized genomes (e.g. Krebs & Feder 1998; Cullum *et al*. 2001). However, recent technological and analytical advances enable similar cross-field studies in species that have been the focus of considerable ecological research, but are rather poorly characterized genetically and physiologically (Feder & Mitchell-Olds 2003). Such non-model organisms are currently extensively screened to describe phylogeographical relationships and population genetic structure using assumedly neutral nonfunctional molecular genetic variation. However, as stated by Lewontin (1974), 'To concentrate only on genetic change, without attempting to relate it to the kinds of physiological, morphogenetic, and behavioural evolution that are manifest in the fossil record and in the diversity of extant organisms and communities, is to forget entirely what we are trying to explain in the first place.' This statement characterizes the impetus behind the currently emerging research field of ecological and evolutionary functional genomics, which aims to apply cross-field approaches for identifying functional genetic polymorphisms that affect the phenotypic traits and fitness in natural environment (e.g. Jackson *et al*. 2002; Feder & Mitchell-Olds 2003).

Rapid progress in genomics and related fields has resulted in a wave of recent reviews advocating the application of genomic approaches in various research areas, including ecology and evolution (Boake *et al*. 2002; Heckel 2003; Stearns & Magwene 2003; Delwiche 2004; Thomas & Klaper 2004; Hofmann *et al*. 2005; Mauricio 2005), describing the development of gene-targeted or 'functional' genetic markers (van Tienderen *et al*. 2002; Andersen & Lübberstedt

Correspondence: Craig Primmer, Fax: +358 2333 6680; E-mail: craig.primmer@utu.fi

© 2005 Blackwell Publishing Ltd

2003; Rudd 2003; Varshney *et al*. 2005), and providing detailed overviews on specific strategies for identification functionally important variation (Table 1). However, most reviews concentrate on small number of closely related methodologies, often without judicious consideration for other potentially useful alternatives. In addition, many reports assume that extensive genomic resources, such as large-insert genomic libraries, comprehensive expressed sequence tag collections, microarray analysis platforms or full genome sequence are already available for the species of interest. Hence, their relevance to genetically less-studied species is reduced.

The aim of this paper is to bring together and compare various potentially useful research strategies for dissecting functionally important genetic variation. We do not intend to give a comprehensive overview of all available methods and technical advances potentially useful for identifying functional DNA polymorphisms, but rather we explore the relative strengths and limitations of the currently most widely used strategies and promising recent developments. Finally, we emphasize the power of integration of complementary research strategies and suggest that combining research approaches targeting different functional and biological levels can significantly increase understanding the genetic basis of various ecological and evolutionary processes in wide range of organisms.

Research strategies for dissecting functional DNA variation at different biological levels

For the purposes of this review, research strategies have been classified into nine categories that utilize DNA, messenger RNA (mRNA), protein, phenotypic and environmental variation and can be applied at the individual, family (pedigree), population or species level (Table 1; Information Box). The outlined methods can be categorized on a scale from 'bottom-up' to 'top-down' approaches depending on the focus of the research along the genotype–phenotype pathway. 'Bottom-up' based approaches utilize DNA polymorphisms and traditionally assess if a particular locus has been affected by directional or balancing selection either directly or via linkage as opposed to neutral evolutionary forces. When the neutral hypothesis is rejected, it can be supposed that the particular locus or nearby region of the

Information Box — Outline of nine research strategies for dissecting functional DNA variation

Single-locus and sequence-based 'neutrality' tests represent a diverse group of statistical methods to test for the null hypothesis of neutral evolution at different timescales extending from current generation to hundreds of thousands of generations.

Multiple-marker-based 'neutrality' tests utilize information from numerous loci or genome regions to construct the neutral null distribution based on the variability characteristics of the markers in the analysed samples. The combination of multiple marker information enables distinguishing locus-/region-specific effects, such as selection from the genome-wide effects due to, e.g. random genetic drift and gene flow.

Quantitative trait locus mapping of messenger RNA (mRNA) expression variation (also known as eQTL mapping) uses the mRNA transcript abundance as a molecular 'phenotype' and utilizes a linkage mapping strategy (see below) to identify particular regions of the genome that are associated with variation in gene expression levels within a pedigree (family).

Allele-specific mRNA expression analysis enables estimation of the expression levels of alternative alleles within heterozygous individuals based on polymorphism in the transcribed region of the gene.

Quantitative trait locus mapping of protein expression variation combines linkage mapping (see below) and protein expression analysis to identify particular regions of the genome that are associated with variation in quantitative and qualitative protein expression levels within a pedigree (family).

Environmental association analysis evaluates if a particular genotype/allele frequency is significantly associated with a certain environmental variable over time or space. Commonly, a significant association between an environmental variable and an allele (or haplotype) frequency is taken as evidence for directional selection affecting the particular locus.

Quantitative trait locus (QTL) analyses (also known as linkage mapping) identify particular regions of the genome that are associated with the trait being assessed or measured within a pedigree (family). Three basic requirements for QTL mapping include genetic linkage map information based on polymorphic markers, pedigree (family) material to trace the segregation of the markers and phenotypic data of individuals in the pedigree.

Admixture mapping enables identification of genomic regions that are nonrandomly distributed among naturally occurring interpopulation or interspecific advanced generation hybrids and associated with phenotypic or fitness related traits.

Association analysis (linkage disequilibrium mapping) tests if a certain genotype (or haplotype) associates nonrandomly with the phenotypic trait of interest within families or populations.

Table 1 Comparison of various strategies for identifying functional DNA polymorphisms

Functional level	Strategy	Species limitations	Gene ascertainment bias	Sampling unit	Cost	Genome coverage (1 = low; 10 = high)	Directness* (1 = low; 10 = high)	Reviews and recent developments
DNA	Single-locus and sequence-based 'neutrality' tests	yes/no	yes/no	cohort/ population/ species	low	1–3	1–2; 3–4†	Kreitman 2000; Yang & Bielawski 2000; Nielsen 2001; Ford 2002; Wright & Gaut 2004
DNA	Multiple-marker-based 'neutrality' tests	no	yes/no	cohort/ population/ species	low/ medium	3–6	3–5	Sabeti *et al*. 2002; Luikart *et al*. 2003; Guinand *et al*. 2004; Storz *et al*. 2004; Beaumont 2005
mRNA-DNA	Quantitative trait locus mapping of mRNA expression variation	yes	yes	pedigree	high	10	1–2; 4–6‡	Jansen & Nap 2001; Pastinen & Hudson 2004; Stamatoyannopoulos 2004
mRNA-DNA	Allele-specific mRNA expression analysis	yes	yes	heterozy-gous individual`	medium /high	1–2	9	Buckland 2004; Pastinen & Hudson 2004; Knight 2004; Stamatoyannopoulos 2004
Protein-DNA	Quantitative trait locus mapping of protein expression variation	yes	yes	pedigree	high	10	1–2; 4–6‡	de Vienne *et al*. 2001; Consoli *et al*. 2002; Thiellement *et al*. 2002
DNA–environment	Environmental association	yes	yes/no	cohort/ population	low/ medium	1–3	1–2; 3–5§	Eanes 1999; Watt & Dean 2000
Phenotype — DNA	Quantitative trait locus (QTL) analysis	yes	no	pedigree	medium /high	10	1–4	Mackay 2001; Erickson *et al*. 2004; Slate 2005
Phenotype — DNA	Admixture mapping	yes	no	population/ species of hybrid origin	medium /high	10	4–6	Rieseberg *et al*. 2000; Rieseberg & Buerkle 2002; Lexer *et al*. 2004; McKeigue 2005
Phenotype — DNA	Association analysis	yes	yes/no	family/ population	low-high	1–10¶	5–9**	Neale & Savolainen 2004; Hirschhorn & Daly 2005

*Strength of evidence that particular polymorphism affects phenotype (incl. mRNA and protein expression 'phenotype').
†Tests based on comparisons between classes of mutations (e.g. d_N/d_S ratio) are expected to provide less ambiguous evidence for selection (*directness*: 3–4) than methods based on allelic distributions or levels of variability (*directness*: 1–2) (e.g. Nielsen 2001; Schlötterer *et al*. 2004).
‡When the expression level of a gene is affected by loci in physically unlinked genomic regions, known as *trans*-acting factors, the strength of evidence that particular polymorphism affects mRNA or protein expression 'phenotype' is relatively low (*directness*: 1–2). When the expression level variation of a gene maps to the same position as the gene itself, it provides more direct support that transcription of particular gene might be controlled by polymorphisms tightly linked to that gene, known as *cis*-acting factors (*directness*: 4–6).
§The strength of evidence that particular polymorphism affects phenotype is expected to be higher for temporal and microhabitat association studies (*directness*: 3–5) than for large-scale (regional) environmental association analyses (*directness*: 1–2) (see text).
¶Association analysis can be performed at various genomic scales starting from a single locus polymorphism analysis (*genome coverage*: 1) to a high density whole genome screen (*genome coverage*: 10).
**The strength of evidence that particular polymorphism affects phenotype depends on the extent of linkage disequilibrium (LD), which varies widely between species, populations and genomic regions (see text). The lower the level of LD, the stronger the evidence that particular polymorphism or adjacent variant affects phenotype.

genome may contain genetic variants that affect phenotypic variation and hence, fitness of the organism. At the opposite end of the scale, 'top-down' approaches, such as quantitative trait loci (QTL) and association mapping, begin with the phenotypic variation and aim to identify the genetic basis of the phenotypic trait by describing the chromosomal locations, number of loci, and the strengths and directions of their effects. The main characteristics of the strategies, including potential limitations, and the directness (in terms of the strength of evidence that a particular polymorphism affects the phenotype) of the approaches are summarized in Table 1 and described in more detail below. Clearly, there are substantial trade-offs between genome coverage, directness and the cost of each approach which indicates that generally no single strategy exists that is optimal for every certain case or species. However, each strategy has its strengths, and it is hoped that this review will assist researchers in choosing the most appropriate strategy, or combination of strategies for their particular study system.

Single-locus and sequence-based 'neutrality' tests

Both single marker and sequence data have been extensively used to test the neutral null hypothesis to infer the evidence of selection using a wide range of statistical methods. Several excellent reviews on this topic exist (see Table 1 for references) and hence, we provide only a brief overview of the main characteristics related to single-locus and sequence-based 'neutrality' tests. In its simplest form, deviations from the expected Hardy–Weinberg genotypic proportions within a population can be used to infer the potential effect of strong selection on a particular marker (Watt & Dean 2000). Unfortunately, this procedure has very low statistical power (Lewontin & Cockerham 1959) and many other confounding factors like null alleles, genotyping errors, population substructure and migration can cause false-positive results. Another single-locus method that evaluates the functional significance of particular polymorphism in the current generation is the test for random mating in natural or experimental conditions, which has been used for example to evaluate the importance of major histocompatibility complex (MHC) gene sequence variation in mate choice (reviewed in Bernatchez & Landry 2003; Garrigan & Hedrick 2003).

The next category of methods are based on allele frequency distribution and/or level of variability and include Ewens–Watterson test, HKA test, Tajima's *D*-test and their subsequent developments (Watterson 1977; Hudson *et al*. 1987; Tajima 1989; Fu 1996). Both the Ewens–Watterson test, Tajima's *D*-test and its derivates have been extensively used to test the hypothesis of neutral equilibrium model which includes strong assumptions regarding population demographics, such as no population structure and constant population size. In addition, recent simulations have showed that Ewens–Watterson test for stepwise-mutation model (Kimura & Ohta 1975; Cornuet & Luikart 1996) is not a valid method to infer signatures of selection at microsatellite loci (Schlötterer *et al*. 2004). As a result, it has been increasingly realized that statistical methods based on single locus allelic distribution and/or levels of variability cannot provide robust inferences of selection as they strongly depend on demographic model (Nielsen 2001). For example, Wright & Gaut (2005) suggested that the proportion of genes that have reported to be under balancing or positive selection (> 20%) in maize (*Zea mays*) and *Arabidopsis* sp. represents most likely considerable overestimate. Excellent depiction of the challenges related to neutral equilibrium model–based tests is provided by Akey *et al*. (2004), who analysed 132 human gene sequences and applied extensive coalescent simulations to evaluate the effect of various demographic scenarios to the performance of four common 'neutrality' tests.

The third group of methods are based on comparisons of between different classes of mutations within a locus and include tests based on the nonsynonymous and synonymous substitution ratio (d_N/d_S or K_A/K_S) and McDonald–Kreitman type test (Hughes & Nei 1988; McDonald & Kreitman 1991). Unlike the methods described in the previous paragraph, this category of tests are not explicitly based on neutral equilibrium model and therefore provide stronger evidence of selection at the amino acid sequence level (reviewed in Yang & Bielawski 2000; Nielsen 2001; Yang 2002). In McDonald–Kreitman test, nonsynonymous and synonymous polymorphism ratio within and between species is compared in a 2×2 contingency table as d_N/d_S ratio within and between species is expected to be the same under neutrality. However, interpretation of the significant outcome of the McDonald–Kreitman test can be complicated because changes in population size together with weak selection against deleterious mutations can either increase or decrease the number of amino acid-changing polymorphisms (Eyre-Walker 2002).

Estimation of the average nonsynonymous (d_N) and synonymous (d_S) substitution ratio between two sequences has been traditionally used to infer whether the particular gene has been under negative selection ($d_N < d_S$), neutral evolution ($d_N = d_S$) or positive selection ($d_N > d_S$) (Nei & Gojobori 1986; Li 1993; Ina 1995). As the majority of amino acid sites are highly conserved and only few substitutions are expected to enhance the function of the protein, $d_N > d_S$ calculated over the whole gene provides an extremely stringent criterion for inferring the presence of positive selection (Endo *et al*. 1996; Tiffin & Hahn 2002). Nevertheless, there is an increasing list of genes that have been found to be under positive selection, especially those related to reproduction and host–parasite interactions indicating that the co-evolutionary arms race between sexes or host and pathogen/parasite can lead to rapid protein evolution

(reviewed in Ford 2002; Swanson & Vacquier 2002). More elaborate approaches have been used to infer lineage-specific episodes of positive selection from multiple sequences in a phylogenetic framework (e.g. Messler & Stewart 1997) and various methods have been developed to identify specific regions or even single sites under positive selection (Fitch *et al*. 1997; Nielsen & Yang 1998; Suzuki & Gojobori 1999; Fares *et al*. 2002; Suzuki 2004; Massingham & Goldman 2005). Application of these methods represents very powerful tool to obtain evidence of positive selection and study function of the protein (Clark *et al*. 2003; Sawyer *et al*. 2005). Nevertheless, such tests can produce misleading results if the underlying assumptions (i.e. no recombination), ancestral sequence reconstruction or nucleotide substitution model is violated or incorrect (e.g. Kosakovsky Pond & Frost 2005). There has been active debate recently over the usefulness of two currently most widely used approaches, maximum-likelihood test (Nielsen & Yang 1998) and parsimony test (Suzuki & Gojobori 1999) to correctly identify positively selected genes and specific sites. It has been suggested that the likelihood methods can provide frequently false-positive results while parsimony method has a low type I error rate (Suzuki & Nei 2001, 2002, 2004; Zhang 2004). Others stress that likelihood method does not lead to an excess of false positives when the underlying models are applied correctly and problems with optimization are ruled out, while the parsimony method has little power to detect positively selected sites unless the data set is large (Wong *et al*. 2004).

In brief, single-locus and sequence-based 'neutrality' tests represent a diverse group of methods based on various assumptions to test the neutral hypothesis at different evolutionary timescales from current generation to hundreds of thousands of generations. Obviously, methods that estimate the role of selection over the history of species (e.g. d_N/d_S) do not provide information whether or not the observed substitution is currently under positive selection (Garrigan & Hedrick 2003). Also, neutral equilibrium model–based tests that infer selection over the history of populations provide unambiguous evidence of selection only in very limited circumstances, as described above. In addition, even if the signal of selection at a particular gene has been detected, often the selective agent remains unknown. In summary, single-locus and sequence-based 'neutrality' tests generally provide only limited evidence that a certain DNA variant or tightly linked polymorphism currently has direct phenotypic and fitness consequences.

Multiple-marker-based 'neutrality' test

Multiple-marker-based 'neutrality' tests (also termed as a 'population genomics' approach, Black *et al*. 2001; or 'genome scan', e.g. Storz 2005) are based on the concept of genetic hitch-hiking (Maynard Smith & Haigh 1974) that predicts that when the frequency of non-neutral polymorphism changes as a result of selection, it also affects the flanking neutral variation. By genotyping large numbers of markers in sets of individuals taken from one or more populations or species, it is possible to identify genomic regions or 'outlier loci' that exhibit patterns of variation that deviate from the rest of the genome due to the effects of selection (reviewed in Black *et al*. 2001; Luikart *et al*. 2003; Schlötterer 2003; Storz 2005). Hence, a significant decrease in variability, elevated divergence and/or an increase of linkage disequilibrium is consistent with directional (positive) selection, while more similar allele frequencies between populations than expected can be caused by balancing selection. In practice, however, it might be very difficult to detect loci under balancing selection when the divergence between samples is moderate or low (Beaumont & Balding 2004; Beaumont 2005). On the other hand, there are an increasing number of studies that have applied multiple-marker-based 'neutrality' tests to identify genes and genomic regions potentially affected by positive selection related to colonization of a novel environment (Schlötterer 2002a; Storz *et al*. 2004), drug and pesticide resistance (Kohn *et al*. 2000; Wootton *et al*. 2002; Nair *et al*. 2003; Catania *et al*. 2004), domestication (Vigouroux *et al*. 2002, 2005; Goldringer & Bataillon 2004; but see also Wiener *et al*. 2003), reproductive isolation (Wilding *et al*. 2001; Campbell & Bernatchez 2004; Emelianov *et al*. 2004; Rogers & Bernatchez 2005), local adaptation (Kayser *et al*. 2003; Storz & Dubach 2004; Vasemägi *et al*. 2005) and interspecific differentiation (Scotti-Saintagne *et al*. 2004). Similarly to single-locus and sequence-based methods, the multiple-marker-based approach represents 'bottom-up' strategy and as it does not require a priori definition of phenotype or candidate loci (Kauer *et al*. 2002; Schlötterer 2002b; Vasemägi *et al*. 2005).

Multilocus 'neutrality' tests for inferring positive selection can be broadly divided into three categories based on: (i) reduction in variability within population, (ii) elevated levels of divergence between samples, and (iii) increased linkage disequilibrium along the chromosome. The first empirical approach commonly utilizes highly variable microsatellite loci to compare genetic diversity between two samples by calculating the variance (V) in allele size (ln RV test; Schlötterer 2002a) or expected heterozygosity (H) (ln RH test; Kauer *et al*. 2002). It has been demonstrated that ln-transformed ratio of V and H are roughly normally distributed under various demographic scenarios (Kauer *et al*. 2002; Schlötterer 2002a). Therefore, loci that have been affected by recent strong directional selection are expected to lie in the tails of empirical genome-wide distribution of ln RV and ln RH. Simulations indicate that ln RV test has lower power than the ln RH test, as the former suffers from higher variance and is sensitive to non-stepwise mutations (Kauer *et al*. 2002). Nevertheless, the ln RV-based test is expected to be useful over longer time interval than the

ln *RH* test, as *V* is supposed to return to its equilibrium value more slowly than *H* after the selective sweep (Kimmel *et al*. 1998). One of the critical prerequisite of both ln *RV* and ln *RH* test is the assumption of constant within-locus mutation rate across populations. Given that there is strong correlation between microsatellite repeat number and mutation rate (Neff & Gross 2001) such an expectation is probably unrealistic and different allele-specific mutation rates can potentially have a significant bias to the outcome ln *RV* and ln *RH* test.

The second group of multilocus neutrality tests compare the level of divergence between two or more samples to detect outlier loci that exhibit higher (or lower) than expected intergeneration/population/species differentiation (commonly measured as F_{ST} or analogous statistic). The approach was originally proposed by Lewontin & Krakauer (1973) and has been further developed, e.g. to take account for population structure (Beaumont & Nichols 1996; Porter 2003; Beaumont & Balding 2004), to improve accuracy by comparing only two, rather than multiple populations (Tsakas & Krimbas 1976; Vitalis *et al*. 2001), to enable use of the test with data for dominant markers (Wilding *et al*. 2001) and to allow comparison of temporal samples within a single population (Goldringer & Bataillon 2004). Commonly, the significance levels of these tests are derived via simulations (Vitalis *et al*. 2001; Porter 2003; Goldringer & Bataillon 2004; Beaumont 2005) but occasionally, when the number of analysed markers is very large, potential outlier loci have been directly identified from the empirical data (Akey *et al*. 2002). To overcome the uncertainties related to various test assumptions in model-based simulations, it has been advocated to employ simultaneously two or more tests which are based on different models and parameter estimation, and select outlier loci for further analysis that are supported by multiple methods (e.g. Storz *et al*. 2004; Vasemägi *et al*. 2005). Also, comparison of different classes of markers (e.g. coding vs. noncoding loci) can be used to infer the presence of selection in distinct groups of polymorphisms (Scotti-Saintagne *et al*. 2004).

An efficient way to increase the reliability of the genome scans is to utilize genetic linkage map information (Emelianov *et al*. 2004; Scotti-Saintagne *et al*. 2004) or available genome sequence (Harr *et al*. 2002; Schlötterer 2003) to identify specific chromosomal regions, rather than individual markers, that are affected by directional selection. Such a strategy, called 'hitch-hiking mapping', has been successfully employed at various genomic scales in different organisms, including rat (Kohn *et al*. 2000), *Drosophila* (Harr *et al*. 2002; Kauer *et al*. 2002; Catania *et al*. 2004), larch budmoth (*Zeiraphera diniana*) (Emelianov *et al*. 2004), malaria parasite (*Plasmodium falciparum*) (Wootton *et al*. 2002; Nair *et al*. 2003) and closely related oak species (Scotti-Saintagne *et al*. 2004). Several examples of how this approach has been applied are outlined in Fig. 1. An alternative strategy,

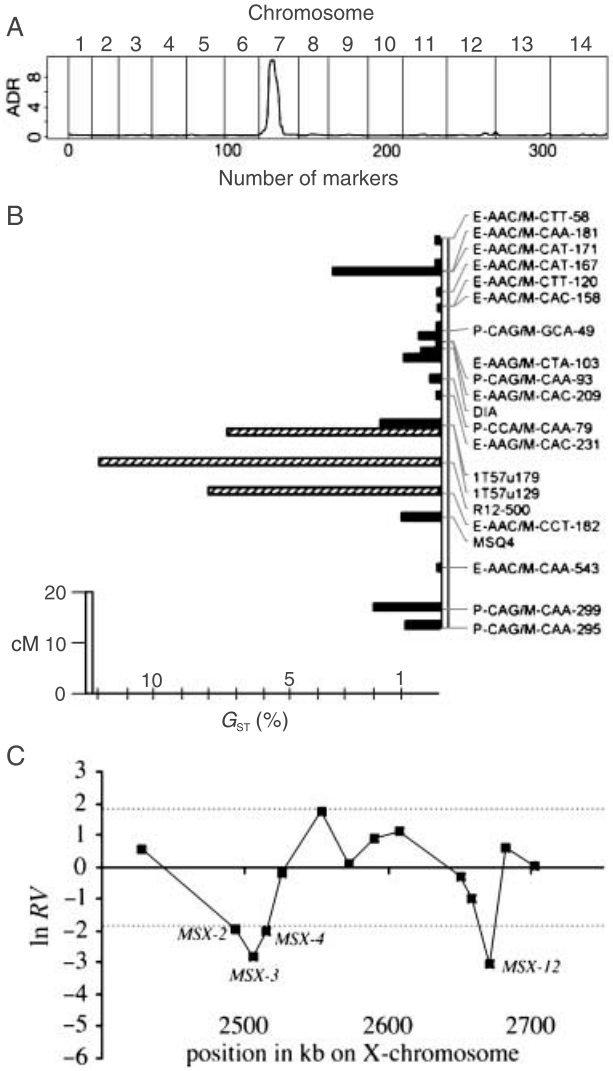

Fig. 1 Examples of hitch-hiking mapping at whole genome, chromosomal and fine-scale levels. (A) Comparison of allelic diversity ratio (ADR) along the genome of chloroquine-resistant and susceptible malaria parasite, *Plasmodium falciparum*, populations from Africa measured using 342 microsatellite markers. Highly significant reduction in diversity on chromosome 7 is consistent with a strong selective sweep driven by drug resistance (modified from Wootton *et al*. 2002). (B) Interspecific differentiation (G_{ST} values) between two closely related oak species (*Quercus petraea* and *Quercus robur*) for 21 loci in a single linkage group. Hatched bars indicate outlier loci which differ significantly from the rest of the genome, most likely caused by hitchhiking effects due to divergent selection (modified from Scotti-Saintagne *et al*. 2004). (C) Relative variability estimates expressed as ln *RV* values for 14 microsatellites covering an X chromosome region of 274.4 kb affected by putative selective sweeps associated with the out-of-Africa colonization in *Drosophila melanogaster*. The 95% confidence intervals of the ln *RV* statistic are indicated by dashed lines (modified from Harr *et al*. 2002).

which does not depend on the information of the genomic location of the markers, is to exploit polymorphisms tightly linked to the coding sequences, such as expressed sequence tag (EST) linked microsatellites (Vigouroux *et al*. 2002; Vasemägi *et al*. 2005). This approach is expected to be more efficient than utilization of unmapped anonymous markers [e.g. signs of selection were identified in 12–21% of gene-linked loci (Scotti-Saintagne *et al*. 2004; Vasemägi *et al*. 2005) compared to 0–9% of anonymous loci (Wilding *et al*. 2001; Campbell & Bernatchez 2004; Scotti-Saintagne *et al*. 2004; Vasemägi *et al*. 2005)] and provides an efficient means for further confirmatory sequence analysis of the closest candidate gene when the full genome sequence of the species of interest is not available. Other confirmatory evidence of positive selection include searching for parallel footprints of selection from multiple samples or populations (Wilding *et al*. 2001; Campbell & Bernatchez 2004), unusually long high-frequency haplotypes (Sabeti *et al*. 2002; Bersaglieri *et al*. 2004; Kim & Nielsen 2004), association with certain environmental variables (Storz & Dubach 2004) and co-localization with the known QTL region (Rogers & Bernatchez 2005). Such complementary validation steps do not only reduce the number of false positives from the initial genome scan, but can also generate useful hypothesis about the underlying phenotypic trait and potential selective agents for further functional analysis. For example, recent genome scan in Atlantic salmon (*Salmo salar*) identified microsatellite locus *Ssa14* as a highly significant outlier potentially affected by divergent selection (Fig. 2A; Vasemägi *et al*. 2005). Notably, the same microsatellite locus has been found to linked to the upper thermal tolerance QTL in two other salmonid fish species, rainbow trout (*Oncorhynchus mykiss*) (Fig. 2B) and Arctic charr (*Salvelinus alpinus*) (Fig. 2C) (Somorjai *et al*. 2003), which suggests that *Ssa14* may be linked to loci affected by similar temperature-related selection in Atlantic salmon.

The most significant advantage of multiple-marker approach over the single-locus 'neutrality' tests is that the null distribution is constructed using the variability characteristics of the observed samples. Hence, multilocus 'neutrality' tests should not be biased by population history and demographic effects and can provide more direct evidence of functional significance of particular gene (Table 1). In addition, without a priori requirement to define fitness-related phenotype, genome scans are expected to be particularly suitable for analysis of adaptation and speciation processes in natural populations of non-model organisms.

cDNA microarrays and quantitative trait locus mapping of mRNA expression variation

It has been generally recognized that DNA polymorphisms in protein-coding regions account for only a fraction of the total phenotypic diversity observed both between and within species (Mackay 2001; Rodriguez-Trelles *et al*. 2003). Accordingly, it was proposed already three decades ago that variation in gene expression, rather than coding sequences, might explain a substantial proportion of the differences between closely related species, such as humans and chimpanzees (King & Wilson 1975; but see also Glazko *et al*. 2005). Recent large-scale gene expression studies using

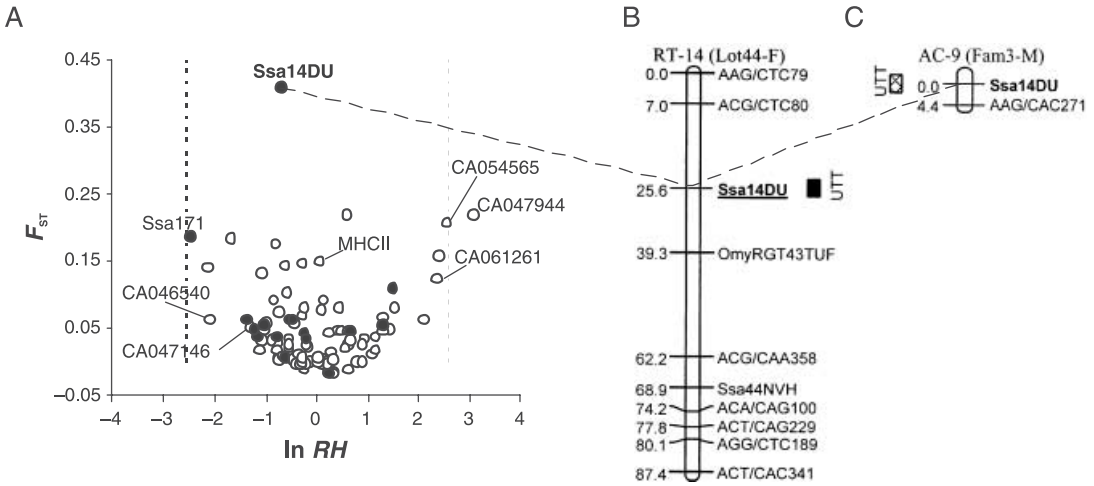

Fig. 2 Multiple evidence of the presence of functionally important variation linked to the genomic microsatellite *Ssa14* in (A) Atlantic salmon (*Salmo salar*) as revealed by genome-scan using 78 EST-linked (empty bullets) and 17 genomic (black bullets) tandem repeat markers between two populations from the Baltic Sea. Dashed lines indicate the 99% confidence interval of standardized ln *RH* estimates. Accession numbers or locus names of putative candidate loci potentially affected by directional selection are indicated (modified from Vasemägi *et al*. 2005). Genetic linkage maps in (B) rainbow trout (*Oncorhynchus mykiss*) and (C) Arctic charr (*Salvelinus alpinus*) indicating the position of the upper thermal tolerance (UTT) QTL as a solid and cross-hatched box (modified from Somorjai *et al*. 2003).

cDNA microarrays have found considerable variation in gene expression levels across individuals, populations and species (e.g. Oleksiak *et al*. 2002), which have invoked heated debate over the biological significance of such variation (Gibson 2003; Rodriguez-Trelles 2004; Feder & Walser 2005). Perhaps not surprisingly, several studies have suggested that a large part of the transcriptome differences within and between species can be explained by the stabilizing (Hsieh *et al*. 2003; Rifkin *et al*. 2003; Lemos *et al*. 2005) or neutral models of transcriptome evolution (Oleksiak *et al*. 2002; Khaitovich *et al*. 2004). On the other hand, it is also evident that quantitative mRNA expression variation of certain genes has probable functional significance and plays significant role in adaptive evolution (Ferea *et al*. 1999; Enard *et al*. 2002; Oleksiak *et al*. 2002).

Identification of the causative genetic polymorphisms from gene expression data alone, however, is difficult (Bochdanovits *et al*. 2003; Abzhanov *et al*. 2004; Drnevich *et al*. 2004). This is extremely relevant because correlation between transcript level and trait does not directly infer that allelic variation of a particular gene directly affects the phenotype as variation upstream in a particular pathway can alter downstream gene expression. Given that it has been shown that polymorphism in one gene can often affect expression of many physically unlinked genes (*trans*-acting factors) both in yeast, mice and humans (Brem *et al*. 2002; Schadt *et al*. 2003a; Morley *et al*. 2004), the assumption that a differentially expressed gene will contain a polymorphism directly responsible for the variation in transcript abundance is often unlikely to be realistic. Therefore, identification of differentially expressed genes alone can be regarded as rather 'short' step towards identification of genetic polymorphisms underlying phenotypic diversity (Feder & Walser 2005; but see also Gracey & Cossins 2003; Nuzhdin *et al*. 2004).

Recently, it has been proposed that mRNA transcript abundance can be considered as a quantitative trait amenable for genetic linkage (QTL) mapping (Jansen & Nap 2001; Doerge 2002). The rationale behind this strategy is straightforward: by treating transcript abundance as a 'phenotype' and assessing whether or not a particular marker cosegregates with the mRNA expression variation within a family (or pedigree) it is possible to map the genomic regions responsible for expression regulation of particular gene (see Information box and Jansen & Nap 2004). When the expression level variation of a gene maps to the same position as the gene itself, it provides evidence that transcription of particular gene is controlled by polymorphisms tightly linked to that gene (known as *cis*-acting factors; Fig. 3). Alternatively, the expression level of a particular gene can be controlled by loci in other physically unlinked genomic regions, known as *trans*-acting factors. Consequently, linkage mapping of mRNA expression level variation [coined as 'genetical genomics' (Jansen & Nap 2001) or eQTL mapping (Schadt *et al*. 2003a)] is particularly suitable for estimation of the proportion of variation controlled by *cis*- vs. *trans*-acting factors. Studies using eQTL approach have already provided valuable insights into transcriptional regulation in yeast, mice, maize and humans

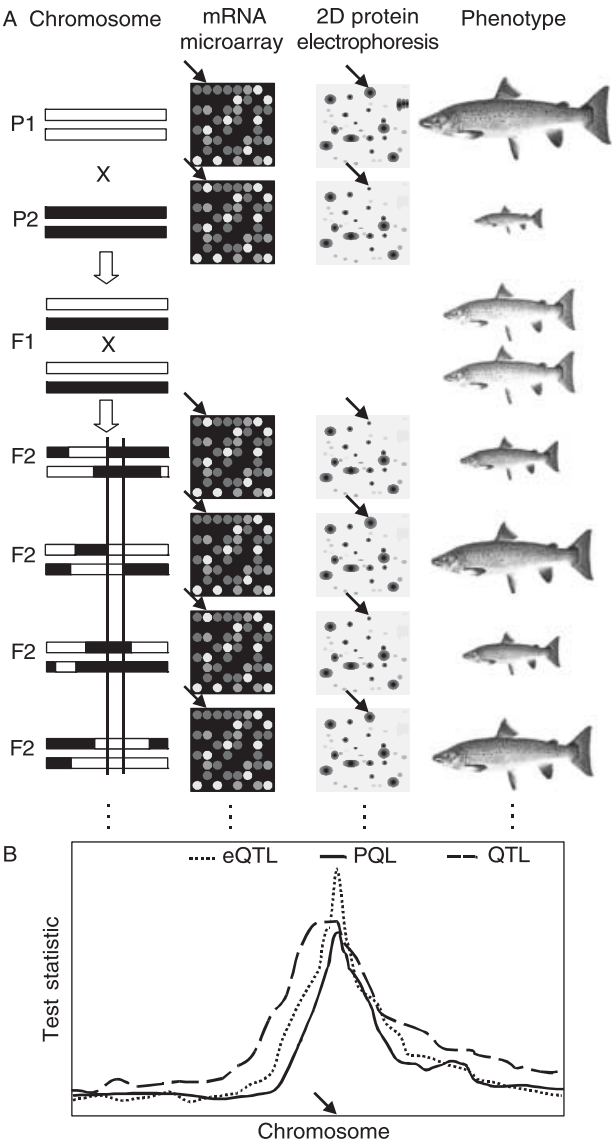

Fig. 3 Schematic illustration of (A) genetic linkage mapping combined with messenger RNA and protein transcript abundance analysis to map the genomic regions (indicated between the two vertical lines) in a segregating cross (F_2) of parental chromosomes (P1 and P2) responsible for variation in mRNA expression (eQTL), protein abundance (PQL) and a phenotypic trait (QTL for size). A differentially expressed hypothetical gene is denoted with an arrow. (B) Physical co-localization of the hypothetical gene, its mRNA expression variation, protein expression variation and phenotypic QTL for size. Such a result would indicate that *cis*-acting polymorphism(s) which affect the particular gene's expression level could be the causative agent resulting in variation of the phenotypic trait.

suggesting that a substantial part of gene expression is controlled by one or several loci (Brem *et al*. 2002; Schadt *et al*. 2003a). As a result, the integration of genetic linkage mapping with gene transcription, protein expression and phenotypic analysis approaches has been recognized as very promising strategy for identifying polymorphisms underlying complex traits in model organisms (Schadt *et al*. 2003b; Jansen & Nap 2004; Kirst *et al*. 2004; Hubner *et al*. 2005).

All the above-mentioned studies have used microarray platforms for mRNA expression profiling and extensive genetic map information to search for the regulatory polymorphisms of gene expression variation in model organisms. The obvious limitations of the eQTL strategy in non-model organisms are therefore following: (i) microarrays are developed only in a limited number of species (ii) high density genetic maps and pedigree information are often not available (iii) the significant costs of simultaneous expression analysis and genetic linkage mapping. Nevertheless, we suggest that eQTL approach can be successfully performed in non-model organisms, especially for identification of variation in *cis*-acting regulatory elements. First, it has been demonstrated that cross-species use of microarrays is possible, with some extreme examples being the use of a human microarray for analysis of pig (Moody *et al*. 2002) and salmon gene expression (Tsoi *et al*. 2003). While it is still too early to determine if there are any general guidelines regarding the expected error rate and bias towards highly conserved genes involved with cross-species microarray use, it is clear that transfer between closely related species will generally be feasible (e.g. Rise *et al*. 2004). Second, a number of alternative mRNA transcript profiling methods, such as differential display (Venkatesh *et al*. 2005) and others (e.g. Breyne *et al*. 2003) may in fact be applicable to an eQTL approach in non-model organisms. Third, it is relatively straightforward to develop hundreds of polymorphic markers and construct low-resolution genetic maps in previously genetically uncharacterized species in a relatively short period of time using molecular markers such as amplified fragment length polymorphisms (AFLP) and related methods (Vos *et al*. 1995; Myburg *et al*. 2001). However, as with any form of lineage mapping, the resolution, or directness, of eQTL mapping is limited by the number of recombination events that have occurred during the crossing experiment. Therefore, dissecting the genetic basis of *trans*-acting regulatory elements remains a formidable task unless a high-density genetic map or whole genome sequence information is available, Nevertheless, less ambitious aims, such as identification of the proportion of *cis*- vs. *trans*-acting factors, could be a realistic goal in many cases. Obtaining such basic information in non-model organisms would be of great interest as this has been studied in only a handful of species (reviewed in Jansen & Nap 2004), and determining the proportion of genes regulated by *cis*-acting factors will be important for evaluating the prospects of further fine-scale characterization of functional genetic variation. In addition, it would further our understanding of the general mechanisms of gene regulation.

Allele-specific mRNA expression analysis

A more direct approach for identification of regulatory polymorphisms tightly linked to the transcribed gene is to evaluate the evidence for allele-specific expression differences (reviewed in Buckland 2004; Knight 2004; Yan & Zhou 2004). The key advantage of this approach is that relative expression levels of alternative alleles are measured within heterozygous individual by designing allele-specific primers based on known polymorphism in the transcribed region of the gene (Singer-Sam *et al*. 1992). As a result, allele-specific expression analysis circumvents the potential biases of several confounding factors, such as environmental variation and mRNA quality. Previously, allele-specific differences in expression levels have shown to be associated with the epigenetic phenomena and genomic imprinting (e.g. Li *et al*. 1993). However, recent findings have shown that a considerable proportion of genes in mammals (Cowles *et al*. 2002; Yan *et al*. 2002; Bray *et al*. 2003) and plants (Schaart *et al*. 2005) show highly tissue- and context-specific allelic variation in gene expression. For example, a recent study in maize demonstrated that alleles of two stress-related genes, lipid transfer protein (*LPT*) and pro-rich protein (*PRP*), respond differentially to stress treatments (high density and drought) compared to the control group (Guo *et al*. 2004). The biggest limitation of PCR-based allele-specific expression profiling methods is the need of polymorphism within the transcribed sequence and relatively low throughput: identification of the polymorphisms themselves and subsequent analyses can be time consuming. Several sophisticated approaches have been recently proposed to overcome these limitations (Ding & Cantor 2003; Lo *et al*. 2003; Ronald *et al*. 2005) but the robustness and the usefulness of these methods in non-model organisms remains to be evaluated. Therefore, allele-specific expression analyses of previously uncharacterized genes has a low probability of revealing functional polymorphisms that underlie phenotypic traits in non-model organisms. However, allele-specific expression profiling holds promise for exploring the potential presence of *cis*-acting regulatory DNA polymorphisms in previously identified candidate genes for which there is some prior evidence of its potential effect to particular phenotype.

Quantitative trait locus mapping of protein expression variation

As numerous pre-, co- and post-translational modifications take place prior to the formation of a functional protein,

mapping of genomic regions associated with the variation in protein expression lies a step closer to the phenotype compared to mRNA expression studies (Feder & Walser 2005). As for eQTL analysis (see above), quantitative trait locus mapping of protein expression variation utilizes linkage mapping to identify the genomic regions responsible for protein expression variation within a known pedigree (Fig. 3). For the past 30 years, two-dimensional (2D) gel electrophoresis has been the most widely used technique to separate and estimate the expression level of hundreds of proteins simultaneously according to isoelectric point, molecular mass and solubility (reviewed in Gorg *et al*. 2004). The first application of a QTL approach to study quantitative and qualitative protein variation using 2D gel electrophoresis was performed in maize (Damerval *et al*. 1994). Altogether 70 genomic regions were identified that affected the expression of 42 polypeptides, one to five protein quantity loci (PQL) per gene (Damerval *et al*. 1994). Commonly, two different types of protein polymorphisms can be distinguished: variation in protein quantity and electrophoretic position (position shift). The position shift loci have been mostly mapped to the structural genes themselves while protein abundance is often regulated by numerous physically unlinked *trans*-acting loci (de Vienne *et al*. 2001; Klose *et al*. 2002). Recently, the same strategy (termed 'quantitative proteomics') was used to analyse over 8700 proteins in a backcross of two mouse species (*Mus musculus* and *Mus spretus*). Among 1324 proteins that showed consistent variation, Klose and colleagues (2002) mapped 665 proteins and identified 466 of them using mass spectrometry demonstrating that combining large-gel 2D electrophoresis and linkage mapping can identify regulatory regions of hundreds of proteins. Obviously, the protein quantity is not the only component that determines the physiological effect of the protein. Linkage mapping studies have been also used to search for genomic regions that underlie enzyme activities, a trait that depends both on the abundance and efficiency of the protein (e.g. Prioul *et al*. 1999).

Generally, PQL suffers of similar limitations as the eQTL approach such as a requirement of large family (pedigree) material and reasonable amount of fresh tissue. Also, similar to eQTL analysis, PQL mapping does not directly reveal whether the regulatory polymorphism that affects the expression of the particular protein has any phenotypic consequences. Several studies have attempted to assess the functional link between gene expression polymorphism and phenotype by evaluating whether certain PQLs (or eQTLs) co-localize with phenotypic QTLs (de Vienne *et al*. 2001; Hubner *et al*. 2005). For example, co-localization of three genomic regions responsible for certain drought-induced protein abundance (PQL) and growth under stress (QTL) has been reported in maize (de Vienne *et al*. 2001). As triple co-localization is unlikely to be due to chance, the authors suggested that regulatory polymorphisms that affect the abundance of particular protein could be directly linked for the growth under stress. The biggest advantage of the PQL strategy is currently related to the use of 2D gel electrophoresis that enables the separation and abundance estimation of hundreds of proteins simultaneously without any prior information of the polypeptide. However, compared to mRNA microarray platforms 2D gel electrophoresis can be still considered a relatively low throughput technique, as generally only hundreds, rather than thousands of genes, can be simultaneously analysed.

Environmental association

Environmental association studies generally assess whether a particular genotype correlates with a certain environmental variable. Such studies can be divided into two broad categories: (i) temporal approach, where allele frequency change together with the environmental factors are monitored over time within a single cohort or population, or where the survival of alternate genotypes in certain environments are monitored over generations; (ii) spatial approach, where the association between allele frequencies and environmental variables from spatially distinct samples/populations is measured. The first strategy has been widely used in artificial selection studies in microorganisms to study the genetic changes that underlie adaptation to various environmental factors (e.g. Elena & Lenski 2003). For example, in one of the longest running evolutionary experiments, Lenski (2004) reviews over a 20 000 generation long study in the prokaryote *Escherichia coli*. Occasionally, temporal association studies have also been applied in free-living natural populations. For example, Planes & Romans (2004) followed a single cohort of marine fish *Diplodus sargus* over 4 months from pelagic larval stage to settlement and reported gradual allele frequency change in phosphoglycomutase enzyme (*Pgm*) which together with common-garden aquarium experiments provided convincing evidence that variation in *Pgm* or closely linked loci affects the growth and survival of this species.

Spatial association studies can be divided into small- (microhabitat) and large-scale (regional) approaches. The distinction between these two is not directly related to the physical distance per se but rather depends on the spatial scale of dispersal and steepness of the environmental gradient. Thus, gene flow is expected to strongly counterbalance selection in the microhabitat environmental gradient and only strong selection pressures are expected to produce associations (a clinal relationship) between allele or haplotype frequency and environmental variables. On a small spatial scale, historical processes such as founder effects and admixture of previously isolated populations are expected to have no confounding effects as the genetic

consequences of these processes should be common to all samples. Small-scale environmental association studies have been carried out mainly to evaluate the functional significance of allozyme polymorphisms and several studies have demonstrated that differential selection gradients can exist over very short distances (Johannesson *et al*. 1995). For example, in the northern acorn barnacle (*Semibalanus balanoides*) Schmidt & Rand (1999) reported a significant and temporally consistent association between mannose-6-phosphate isomerase (*Mpi*) polymorphism and intertidal thermal microhabitats, suggesting that *Mpi* or a closely linked locus plays important role in adaptation to environmental heterogeneity. Similarly, in the polychaete *Alvinella pompejana* that lives in deep-sea hydrothermal vents, phosphoglycomutase enzyme (*Pgm*) polymorphism has been shown to correlate strongly with the mean temperature of the habitat (Piccino *et al*. 2004). It is interesting to note that the presence of temperature-related selection on *Pgm* has been reported in a wide range of organisms (e.g. Eanes 1999; Verrelli & Eanes 2001).

By far the most commonly reported marker–environment associations in natural populations are large-scale latitudinal analyses where the presence of statistically significant associations between latitude and allele (or haplotype) frequency have been frequently taken as support for the hypothesis that positive natural selection maintains single-locus clinal variation (e.g. Eanes 1999; Baines *et al*. 2004; Sezgin *et al*. 2004). The major difficulty is, however, that no population-specific null expectation exists regarding how frequently neutral polymorphisms are also expected to show clinal variation. Consequently, it is often overlooked that large-scale clinality of molecular variation can be easily explained by historical processes, such as founder effects or hybridization of previously isolated populations and migrational patterns, such as spatially restricted gene flow (isolation by distance). Furthermore, environmental association studies traditionally utilize a very limited number of loci (but see also Storz & Dubach 2004) which often makes difficult, if not impossible, to evaluate the alternative neutral explanation. Therefore, as emphasized by Clarke (1975), selection should only be invoked as a reasonable explanation when additional functional evidence, independent parallel clines within or between species is found. Finally, it is important not to underestimate the difficulties of correctly determining and measuring the biologically relevant environmental factors to avoid spurious or misleading associations. For example, a significant single-locus latitudinal cline can suggest, but not identify, what type of biotic or abiotic factors have operated as selective agents. Clearly, it is not a trivial task even in relatively well-studied systems where reasonable environmental variable candidate may exist (Schmidt & Rand 1999; Veliz *et al*. 2004). To conclude, empirical environmental association studies, and especially temporal analysis, can provide an invaluable link between the genetic variation and selective processes in natural environments, but the successful execution of such an approach requires thorough evaluation of the alternative neutral scenarios together with a shift from single-locus towards multilocus analyses.

Quantitative trait locus analysis

Quantitative trait locus (QTL) analysis (also referred as linkage mapping) has been extensively used in model organisms (Mackay 2001), plants (Mauricio 2001) and domestic animals (Andersson & Georges 2004) to identify certain regions of the genome that are associated with the phenotypic traits of interest. Several recent reviews have comprehensively addressed the application of QTL analysis for studying a wide range of ecological and evolution questions in non-model species (Erickson *et al*. 2004; Slate 2005) and hence details of different mapping strategies will not be covered here. In general, the effectiveness of a QTL mapping strategy for identifying genes affecting the trait of interest will be most efficient when (i) detailed multigenerational pedigree information is available and (ii) a dense genetic linkage map is available for the species. Clearly the number of free-living species for which these criteria are fulfilled is quite limited. However, the recent reviews listed above describe a number of strategies through which significant progress can still be made even when all criteria are not fulfilled. Nevertheless, it is important to remember that while the knowledge of the number and effect size of the genomic region(s) controlling the trait of interest is very valuable, the procedure to identifying the gene(s) and subsequently the quantitative trait nucleotide(s) (QTN) of the QTL is still very time consuming (Slate 2005). Therefore, in this review, we focus on applications of QTL mapping which are integrated with other approaches (Fig. 3; see above) and several additional mapping strategies suitable for application in free-living populations and discuss their potential integration with other approaches (see below).

Admixture mapping

Hybridization and introgression is common both in plants and animals (Dowling & Secor 1997; Allendorf *et al*. 2001) and offers a unique opportunity to study the genetic basis underlying the processes of speciation and adaptation (Arnold 1997; Mallet 2005). The use of naturally occurring advanced generation hybrids for linkage mapping studies can essentially be viewed as an extension of conventional linkage mapping in an artificial experimental cross resembling an advanced intercross line (Darvasi & Soller 1995). The basic principle of admixture mapping is straightforward: if individuals from two genetically distinct populations that

differ in allele frequencies affecting the trait of interest hybridize, it is expected that the ancestry of the chromosomal region linked to that locus is non-randomly associated with the underlying trait (Fig. 4A). Because recombination reduces the size of the chromosomal blocks inherited from each of the contributing population in every subsequent generation, admixture mapping is expected to have substantially higher resolution than traditional QTL analysis that utilizes second-generation hybrids as the regions in which significant linkage is detected will be smaller due to the increased number of recombination events (Table 1). On the other hand, the resolution of admixture mapping is still lower than of association analysis that is based on ancestral linkage disequilibrium (LD) pattern generated over hundreds and thousands of generations (see below). Several recent reviews have emphasized the potential value of advanced backcrosses for genetic mapping in naturally occurring hybrid zones in plants (e.g. Rieseberg *et al*. 2000; Rieseberg & Buerkle 2002) but the approach can be successfully applied in many other organisms of mixed genetic origin (Rogers *et al*. 2001; Darvasi & Shifman 2005; McKeigue 2005).

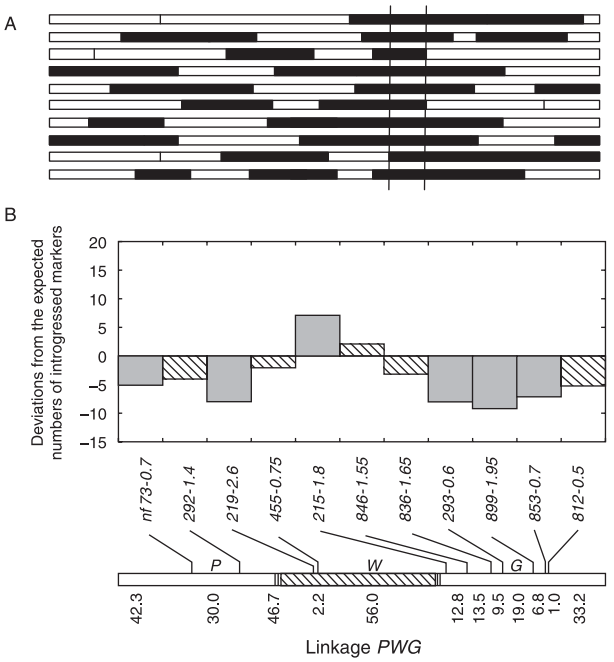

Fig. 4 Admixture mapping. (A) Schematic illustration of the nonrandom ancestry of chromosomal regions in a hybrid population originating from two genetically distinct populations/species (marked as black and white segments). Nonrandom genomic composition of a particular chromosomal region (indicated between two vertical lines) is suggestive of natural selection, rather than genetic drift affecting the genomic composition of the region. (B) Deviation from random introgression along the *PWG* linkage group in a natural hybrid zone between two sunflower species (*Helianthus annuus* and *Helianthus petiolaris*). Gray and hatched bars correspond to significant and nonsignificant deviations from random genome introgression, respectively, measured at 11 species-specific markers. The introgression pattern was highly consistent in three separate hybrid zones (modified from Rieseberg *et al*. 1999).

Compared to QTL analysis in advanced inbred lines, admixture mapping in natural populations poses several practical and theoretical challenges (reviewed in Rieseberg & Buerkle 2002; McKeigue 2005). First, the admixture level among individuals varies in natural populations, which can generate false associations if not taken into account. Second, the ancestral populations that contributed to the admixed population may not be available or known. Third, there may be no fixed differences in allele frequencies which prohibits the development of fully diagnostic markers for separating ancestral populations. However, recently developed statistical tools (Hoggart *et al*. 2004; Montana & Pritchard 2004) have provided a means to overcome such shortcomings and demonstrated that admixture mapping can be efficient and cost-effective way to identify genetic regions responsible for variation in complex phenotypic traits (Zhu *et al*. 2005). Importantly, map-based analysis of highly recombinant natural hybrids allows an evaluation of the fitness of parental chromosome segments even without direct information about the underlying phenotype. For example, linkage analysis of hybrid zones in two sunflower species (*Helianthus petiolaris*, *Helianthus annuus*) revealed large numbers of chromosomal segments that appeared to be under the same direction of selection in three hybrid populations indicating that natural selection, rather than genetic drift, has been the primary evolutionary force affecting the genomic composition of the ancient hybrid populations (Fig. 4B; Rieseberg *et al*. 1999). Nevertheless, admixture mapping together with phenotypic information can provide further information about the underlying traits and the evolutionary mechanisms that contribute to isolation. For, example, in sunflowers, association between pollen sterility and negatively selected genomic regions provided a reasonable explanation why these segments have been disadvantageous in hybrids (Rieseberg *et al*. 1999). In addition, in a follow-up study (Rieseberg *et al*. 2003) compared the genetic composition of ancient hybrids with artificially made backcrosses and demonstrated that the same chromosomal combinations are required to generate similar phenotypes in both synthetic and ancient hybrids. The authors suggest that the results corroborate the importance of the role that hybridization can play in adaptive evolution and suggest that hybridization offers a means for rapid adaptive evolutionary transitions.

Association analysis

Association studies (also referred as linkage disequilibrium mapping) test if a certain genotype (or haplotype) is

associated more frequently with a phenotypic trait than expected by chance alone by utilizing the non-random occurrence of alleles at linked loci, known as linkage disequilibrium (LD). The extent of LD depends on many biological and demographic factors, including recombination rate, population history, selection and mating system characteristics. As a result, the extent of linkage disequilibrium varies widely across species, populations and genomic regions (Jorde 2000). For example, LD extends over hundreds of kilobases in predominantly selfing species like *Arabidopsis* and rice (Nordborg *et al*. 2002; Garris *et al*. 2003) whereas linkage disequilibrium in outcrossing species can rapidly decay to distances of less than 1 kb (Langley *et al*. 2000; Remington *et al*. 2001). In addition, both the rate and pattern of LD can differ considerably between genetically closely related species (Remington *et al*. 2001; Hamblin *et al*. 2004; Ptak *et al*. 2005). Linkage disequilibrium has crucial implications to association mapping as it affects both the power and directness of the approach. Extended LD enables to screen for associations with the limited number of markers, but the precise location of the QTL cannot be inferred. On the contrary, reduced linkage disequilibrium allows accurate fine-scale mapping but the number of markers required for the whole-genome scan can be extremely large. In humans, one million random single nucleotide polymorphisms (SNPs) has been suggested to provide reasonable whole–genome coverage for association studies (Hirschhorn & Daly 2005). As a result, association studies that screen comprehensively the entire genome are generally not feasible for non-model organisms in the near future. As an alternative to the whole-genome screens, several recent reviews have advocated the utility of regional or candidate–gene association approach where only a small part of the entire genome is covered using random or gene-targeted markers (Cardon & Bell 2001; Tabor *et al*. 2002; Simko 2004). Such a candidate–region/gene association approach has been recognized as a promising means of finding functionally important polymorphisms in naturally occurring long-lived species with the large genome size (such as in conifers) where traditional linkage mapping using experimental crossing is impossible or prohibitively time-consuming (Neale & Savolainen 2004).

Two general classes of association tests can be distinguished: population based case-control studies and pedigree-based analyses. Case-control design compares the marker frequencies among individuals that differ in the phenotypic trait of interest within a single panmictic population, while pedigree-based analyses estimate whether heterozygous parents transmit a particular allele more or less frequently to the progeny with certain phenotype than expected. Both types of tests have been extensively developed in medical genetics (Hirschhorn & Daly 2005) but occasionally, LD mapping has also been applied to other organisms (e.g. Hernandez-Sanchez *et al*. 2003; Streelman *et al*. 2003). The biggest advance of case-control design is its relative simplicity to obtain individuals that differ in particular trait. At the same time, case-control tests can have several potential confounding effects related to poorly matched control group and population stratification (Cardon & Bell 2001; Hirschhorn & Daly 2005). Recently developed statistical methods that account for the population stratification have enabled construction of valid case-control tests even in the presence of population structure (Pritchard *et al*. 2000; Pritchard & Donnelly 2001). Pedigree–based association studies (such as the transmission-disequilibrium test, TDT) are not affected by population stratification but are susceptible to false-positive results from genotyping errors and missing data (Mitchell *et al*. 2003; Hirschhorn & Daly 2005) and require family based samples that can be difficult to obtain in non-model organisms (but see Moen *et al*. 2004).

To date, most LD mapping studies conducted in non-model organisms have tested the association between a particular phenotype and small number of previously identified candidate genes. The diverse range of organisms in which associations between major histocompatibility complex (MHC) genes and immune response have been identified (reviewed in Bernatchez & Landry 2003; Garrigan & Hedrick 2003) are a classic example of utilizing such an approach. In addition, there is increasing evidence that some genes can control similar phenotypic traits even in distantly related species (Fitzpatrick *et al*. 2005) which will aide the identification of candidate genes study in non-model organisms. For example, mutations in the Melanocortin-1 receptor gene (*MC1R*) were first implicated in controlling coat colour variation in a range of model organisms including mice (Robbins *et al*. 1993), humans (Valverde *et al*. 1995), chickens (Takeuchi *et al*. 1996) and pigs (Kijas *et al*. 1998). Subsequent studies have identified strong associations between mutations in the same gene and colour phenotypes in a range of free-living species including several bird species (Theron *et al*. 2001; Mundy *et al*. 2004), pocket mice (Nachman *et al*. 2003) and lizards (Rosenblum *et al*. 2004). However, studies of the same gene also provide a reminder that such a candidate gene approach will not always prove successful in identifying ecologically important mutations. For example, MacDougall-Shackleton *et al*. (2003) did not identify any mutations in *MC1R* which corresponded to variation in the plumage patterns of old world leaf warblers. Such studies provide examples how different genetic solutions can be obtained for the same evolutionary problem, and highlight the importance of the reporting of negative results.

Commonly, associations are tested separately for every single polymorphism within the gene even when there is extensive linkage disequilibrium between markers. However, it has been recognized that utilization of haplotypes, rather than single locus polymorphisms, can provide a

more powerful strategy to test for phenotypic associations (e.g. Akey *et al*. 2001; Zaykin *et al*. 2002; Zhao *et al*. 2003; Simko 2004; but see also Cuppen 2005 for more critical view). Recently, a simple method was introduced that partitions the haplotype tree into two or more clades and estimates the phenotypic association related to the specific substitution(s) in the haplotype tree (Templeton *et al*. 2005). Such analyses provide a promising evolutionary framework enabling interpretation of the associations of multiple linked loci and are expected to be especially useful in candidate gene association studies of non-model organisms in the future.

Limitations

Identification of the causative polymorphisms underlying phenotypic traits is an extremely challenging task in organisms which lack extensive sequence information and genomic resources. Moreover, even in well-studied species the discovery of functional genetic variation that affects complex (multigenic) traits can be often more time consuming and complex than expected (Flint & Mott 2001; Mackay 2001; Caicedo *et al*. 2004). To illustrate the difficulties related to gene discovery starting from QTL mapping, Nadeau & Frankel (2000) described the procedure as a 'long and bumpy road'. We argue that such a view may actually under-estimate the challenge of identifying genetic variation of functional importance in many non-model organisms, as the 'roads' are yet to be constructed for many species. In addition, formal evidence, or burden of proof, that a certain genetic variant affects particular trait can be very difficult to obtain and has been demonstrated rather limited number of cases (Glazier *et al*. 2002; Abiola *et al*. 2003). For example, even when perfect association between a polymorphic site and phenotypic trait is found, other tightly linked polymorphisms can be responsible for the variation of the trait. As a result, most of the evidence on the functional role of a particular polymorphism in non-model species can be considered as inconclusive as the possibility that other closely linked variants affect the phenotypic trait of interest have been formally rejected in limited number of cases (but see also Schulte *et al*. 2000; Stolz *et al*. 2003; Colosimo *et al*. 2005). Therefore, it is unrealistic for studies of functional variation in non-model organisms to reach the same levels of confirmation achieved in research on model organisms. Nevertheless, this does not make such research efforts futile (see below).

Determination of the generality of a particular finding is another critical factor often neglected as analysis of functional genetic variation is frequently performed within just a single family or population. For example, Symonds *et al*. (2005) analysed four crosses derived from different natural populations in *Arabidopsis thaliana* and identified nine QTL with significant effects on trichome density. However, only two of the nine QTL mapped to the same linkage positions in all four populations, while seven QTL were population specific. Similarly, Nachman *et al*. (2003) reported that while *MC1R* mutations were associated with coat colour polymorphism in one pocket mouse population, while no similar association was observed in three additional populations situated 700 km away. Therefore, it is important to recognize that utilizing multiple, independent samples (individuals, families, populations or species) is expected to provide better understanding of the functional genetic variation than any single sample approach alone.

Future directions: the promise of integrative approaches

Despite the fact that in many cases it will not be possible for studies of functional variation in non-model organisms to reach the same levels of confirmation achieved in research on model organisms, valuable information can still be gained by taking smaller 'steps'. Clearly, combining more than one of the research approaches described above (Table 1) will increase the reliability of the findings (e.g. Rieseberg *et al*. 2003; Rogers & Bernatchez 2005). Several recent studies, together with the examples described earlier, illustrate this point by successfully combining several complementary approaches at multiple levels of biological complexity (Liu *et al*. 2001). For example, using quantitative trait locus mapping, Wayne & McIntyre (2002) initially identified 5286 genes that fell within the QTL region responsible for the variation of ovariole number in *Drosophila melanogaster*. Deficiency (deletion) mapping reduced the number of positional candidate genes to 548 and expression analysis identified 34 of these to be differentially expressed. These 34 loci are strong candidates to be associated with the trait but in order to identify particular functional polymorphisms causing the observed phenotypic variation, additional approaches such as allele-specific expression profiling, eQTL mapping, association analysis or transformation are needed. At present, we are not aware of any single study that have utilized linkage mapping to simultaneously identify genomic regions responsible for the variation in gene transcription, protein expression and phenotypic trait. Nevertheless, it has been increasingly recognized that a 'systems biology approach', i.e. integration of genotypic, transcriptomic, proteomic and phenotypic information provides a promising means of searching for causative polymorphisms underlying complex trait of interest (Schadt *et al*. 2003b).

In conclusion, combination of independent research approaches at various functional (i.e. DNA, mRNA, protein) and biological (i.e. individual, population, species) levels can provide a better understanding of the role of molecular variation in genetic, ecological and evolutionary processes in a wide range of species. It is likely that much of the

future progress in functional genetic analyses will rely on progress in model species, as has been the case in the past, but, as evidenced by the examples outlined in this review, with prudent application of methodologies and appropriate sampling designs, much headway will continue to be made in an increasingly diverse range of organisms.

Acknowledgements

We thank Louis Bernatchez, Martin Feder and anonymous reviewer for their useful comments. We appreciate the support of funding from the Academy of Finland and the Finnish Ministry for Agriculture and Forestry.

References

Abiola O, Angel JM, Avner P *et al.* (2003) The nature and identification of quantitative trait loci: a community's view. *Nature Reviews Genetics*, **4**, 911–916.

Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ (2004) *Bmp4* and morphological variation of beaks in Darwin's finches. *Science*, **305**, 1462–1465.

Akey J, Jin L, Xiong MM (2001) Haplotypes vs. single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics*, **9**, 291–300.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805–1814.

Akey JM, Eberle MA, Rieder MJ *et al.* (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology*, **2**, 1591–1599.

Allendorf FW, Leary RF, Spruell P, Wenburg JK (2001) The problems with hybrids: setting conservation guidelines. *Trends in Ecology & Evolution*, **16**, 613–622.

Andersson L, Georges M (2004) Domestic animal genomics: deciphering the genetics of complex traits. *Nature Reviews Genetics*, **5**, 202–212.

Andersen JR, Lübberstedt T (2003) Functional markers in plants. *Trends in Plant Science*, **8**, 554–560.

Arnold ML (1997) *Natural Hybridization and Evolution*. Oxford University Press, New York.

Baines JF, Das A, Mousset S, Stephan W (2004) The role of natural selection in genetic differentiation of worldwide populations of *Drosophila ananassae*. *Genetics*, **168**, 1987–1998.

Beaumont MA (2005) Adaptation and speciation: what can F_{ST} tell us? *Trends in Ecology & Evolution*, **20**, 435–440.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **263**, 1619–1626.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, **16**, 363–377.

Bersaglieri T, Sabeti PC, Patterson N *et al.* (2004) Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics*, **74**, 1111–1120.

Black WC, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: genome-wide sampling of insect populations. *Annual Review of Entomology*, **46**, 441–469.

Boake CRB, Arnold SJ, Breden F *et al.* (2002) Genetic tools for studying adaptation and the evolution of behavior. *American Naturalist*, **160**, S143–S159.

Bochdanovits Z, van der Klis H, de Jong G (2003) Covariation of larval gene expression and adult body size in natural populations of *Drosophila melanogaster*. *Molecular Biology and Evolution*, **20**, 1760–1766.

Bray NJ, Buckland PR, Owen MJ, O'Donovan MC (2003) Cis-acting variation in the expression of a high proportion of genes in human brain. *Human Genetics*, **113**, 149–153.

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Breyne P, Dreesen R, Cannoot B *et al.* (2003) Quantitative cDNA-AFLP analysis for genome-wide expression studies. *Molecular Genetics and Genomics*, **269**, 173–179.

Buckland PR (2004) Allele-specific gene expression differences in humans. *Human Molecular Genetics*, **13**, R255–R260.

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD (2004) Epistatic interaction between *Arabidopsis* FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proceedings of the National Academy of Sciences, USA*, **101**, 15670–15675.

Campbell D, Bernatchez L (2004) Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. *Molecular Biology and Evolution*, **21**, 945–956.

Cardon LR, Bell JI (2001) Association study designs for complex diseases. *Nature Reviews Genetics*, **2**, 91–99.

Catania F, Kauer MO, Daborn PJ *et al.* (2004) World-wide survey of an Accord insertion and its association with DDT resistance in *Drosophila melanogaster*. *Molecular Ecology*, **13**, 2491–2504.

Clark AG, Glanowski S, Nielsen R *et al.* (2003) Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. *Science*, **302**, 1960–1963.

Clarke B (1975) Contribution of ecological genetics to evolutionary theory — detecting direct effects of natural-selection on particular polymorphic loci. *Genetics*, **79**, 101–113.

Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928–1933.

Consoli L, Lefevre A, Zivy M, de Vienne D, Damerval C (2002) QTL analysis of proteome and transcriptome variations for dissecting the genetic architecture of complex traits in maize. *Plant Molecular Biology*, **48**, 575–581.

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, **144**, 2001–2014.

Cowles CR, Hirschhorn JN, Altshuler D, Lander ES (2002) Detection of regulatory variation in mouse genes. *Nature Genetics*, **32**, 432–437.

Cullum AJ, Bennett AF, Lenski RE (2001) Evolutionary adaptation to temperature. IX. Preadaptation to novel stressful environments of *Escherichia coli* adapted to high temperature. *Evolution*, **55**, 2194–2202.

Cuppen E (2005) Haplotype-based genetics in mice and rats. *Trends in Genetics*, **21**, 318–322.

Damerval C, Maurice A, Josse JM, Devienne D (1994) Quantitative trait loci underlying gene-product variation — a novel perspective for analyzing regulation of genome expression. *Genetics*, **137**, 289–301.

Darvasi A, Shifman S (2005) The beauty of admixture. *Nature Genetics*, **37**, 118–119.

Darvasi A, Soller M (1995) Advanced intercross lines, an experimental population for fine genetic-mapping. *Genetics*, **141**, 1199–1207.

Delwiche CF (2004) The genomic palimpsest: genomics in evolution and ecology. *Bioscience*, **54**, 991–1001.

Ding CM, Cantor CR (2003) A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. *Proceedings of the National Academy of Sciences, USA*, **100**, 3059–3064.

Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics*, **3**, 43–52.

Dowling TE, Secor CL (1997) The role of hybridization and introgression in the diversification of animals. *Annual Review of Ecology and Systematics*, **28**, 593–619.

Drnevich JM, Reedy MM, Ruedi EA, Rodriguez-Zas S, Hughes KA (2004) Quantitative evolutionary genomics: differential gene expression and male reproductive success in *Drosophila melanogaster*. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **271**, 2267–2273.

Eanes WF (1999) Analysis of selection on enzyme polymorphisms. *Annual Review of Ecology and Systematics*, **30**, 301–326.

Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Reviews Genetics*, **4**, 457–469.

Emelianov I, Marec F, Mallet J (2004) Genomic evidence for divergence with gene flow in host races of the larch budmoth. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **271**, 97–105.

Enard W, Khaitovich P, Klose J *et al.* (2002) Intra- and interspecific variation in primate gene expression patterns. *Science*, **296**, 340–343.

Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. *Molecular Biology and Evolution*, **13**, 685–690.

Erickson DL, Fenster CB, Stenøien HK, Price D (2004) Quantitative trait locus analyses and the study of evolutionary process. *Molecular Ecology*, **13**, 2505–2522.

Eyre-Walker A (2002) Changing effective population size and the McDonald–Kreitman test. *Genetics*, **162**, 2017–2024.

Fares MA, Elena SF, Ortiz J, Moya A, Barrio E (2002) A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. *Journal of Molecular Evolution*, **55**, 509–521.

Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. *Nature Reviews Genetics*, **4**, 651–657.

Feder ME, Walser J-C (2005) The biological limitations of transcriptomics in elucidating stress and stress responses. *Journal of Evolutionary Biology*, **18**, 901–910.

Ferea TL, Botstein D, Brown PO, Rosenzweig RF (1999) Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences, USA*, **96**, 9721–9726.

Fitch WM, Bush RM, Bender CA, Cox NJ (1997) Long term trends in the evolution of H (3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences, USA*, **94**, 7712–7718.

Fitzpatrick MJ, Ben-Shahar Y, Smid HM *et al.* (2005) Candidate genes for behavioural ecology. *Trends in Ecology & Evolution*, **20**, 96–104.

Flint J, Mott R (2001) Finding the molecular basis of quantitative traits: successes and pitfalls. *Nature Reviews Genetics*, **2**, 437–445.

Ford MJ (2002) Applications of selective neutrality tests to molecular ecology. *Molecular Ecology*, **11**, 1245–1262.

Fu YX (1996) New statistical tests of neutrality for DNA samples from a population. *Genetics*, **143**, 557–570.

Garrigan D, Hedrick PW (2003) Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. *Evolution*, **57**, 1707–1722.

Garris AJ, McCouch SR, Kresovich S (2003) Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa* L.). *Genetics*, **165**, 759–769.

Gibson G (2003) Population genomics: celebrating individual expression. *Heredity*, **90**, 1–2.

Glazier AM, Nadeau JH, Aitman TJ (2002) Finding genes that underlie complex traits. *Science*, **298**, 2345–2349.

Glazko G, Veeramachaneni V, Nei M, Makalowski W (2005) Eighty percent of proteins are different between humans and chimpanzees. *Gene*, **346**, 215–219.

Goldringer I, Bataillon T (2004) On the distribution of temporal variations in allele frequency: consequences for the estimation of effective population size and the detection of loci undergoing selection. *Genetics*, **168**, 563–568.

Gorg A, Weiss W, Dunn MJ (2004) Current two-dimensional electrophoresis technology for proteomics. *Proteomics*, **4**, 3665–3685.

Gracey AY, Cossins AR (2003) Application of microarray technology in environmental and comparative physiology. *Annual Review of Physiology*, **65**, 231–259.

Guinand B, Lemaire C, Bonhomme F (2004) How to detect polymorphisms undergoing selection in marine fishes? A review of methods and case studies, including flatfishes. *Journal of Sea Research*, **51**, 167–182.

Guo M, Rupe MA, Zinselmeier C *et al.* (2004) Allelic variation of gene expression in maize hybrids. *Plant Cell*, **16**, 1707–1716.

Hamblin MT, Mitchell SE, White GM *et al.* (2004) Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. *Genetics*, **167**, 471–483.

Harr B, Kauer M, Schlötterer C (2002) Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences, USA*, **99**, 12949–12954.

Heckel DG (2003) Genomics in pure and applied entomology. *Annual Review of Entomology*, **48**, 235–260.

Hernandez-Sanchez J, Visscher P, Plastow G, Haley C (2003) Candidate gene analysis for quantitative traits using the transmission disequilibrium test: the example of the melanocortin 4-receptor in pigs. *Genetics*, **164**, 637–644.

Hirschhorn JN, Daly MJ (2005) Genome–wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, **6**, 95–108.

Hofmann GE, Burnaford JL, Fielman KT (2005) Genomics-fueled approaches to current challenges in marine ecology. *Trends in Ecology & Evolution*, **20**, 305–311.

Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM (2004) Design and analysis of admixture mapping studies. *American Journal of Human Genetics*, **74**, 965–978.

Hsieh WP, Chu TM, Wolfinger RD, Gibson G (2003) Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics*, **165**, 747–757.

Hubner N, Wallace CA, Zimdahl H *et al.* (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*, **37**, 243–253.

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class-I loci reveals overdominant selection. *Nature*, **335**, 167–170.

Ina Y (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Journal of Molecular Evolution*, **40**, 190–226.

Jackson RB, Linder CR, Lynch M *et al.* (2002) Linking molecular insight and ecological research. *Trends in Ecology & Evolution*, **17**, 409–414.

Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. *Trends in Genetics*, **17**, 388–391.

Jansen RC, Nap JP (2004) Regulating gene expression: surprises still in store. *Trends in Genetics*, **20**, 223–225.

Johannesson K, Johannesson B, Lundgren U (1995) Strong natural-selection causes microscale allozyme variation in a marine snail. *Proceedings of the National Academy of Sciences, USA*, **92**, 2602–2606.

Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. *Genome Research*, **10**, 1435–1444.

Kauer M, Zangerl B, Dieringer D, Schlötterer C (2002) Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. *Genetics*, **160**, 247–256.

Kayser M, Brauer S, Stoneking M (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. *Molecular Biology and Evolution*, **20**, 893–900.

Khaitovich P, Weiss G, Lachmann M *et al.* (2004) A neutral model of transcriptome evolution. *PLoS Biology*, **2**, 682–689.

Kijas JMH, Wales R, Tornsten A, Chardon P, Moller M, Andersson L (1998) Melanocortin receptor 1 (MC1R) mutations and coat color in pigs. *Genetics*, **150**, 1177–1185.

Kim Y, Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. *Genetics*, **167**, 1513–1524.

Kimmel M, Chakraborty R, King JP *et al.* (1998) Signatures of population expansion in microsatellite repeat data. *Genetics*, **148**, 1921–1930.

Kimura M, Ohta T (1975) Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proceedings of the National Academy of Sciences, USA*, **72**, 2761–2764.

King MC, Wilson AC (1975) Evolution at 2 levels in humans and chimpanzees. *Science*, **188**, 107–116.

Kirst M, Myburg AA, De Leon JPG *et al.* (2004) Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. *Plant Physiology*, **135**, 2368–2378.

Klose J, Nock C, Herrmann M *et al.* (2002) Genetic analysis of the mouse brain proteome. *Nature Genetics*, **30**, 385–393.

Knight JC (2004) Allele-specific gene expression uncovered. *Trends in Genetics*, **20**, 113–116.

Kohn MH, Pelz HJ, Wayne RK (2000) Natural selection mapping of the warfarin-resistance gene. *Proceedings of the National Academy of Sciences, USA*, **97**, 7911–7915.

Kosakovsky Pond SL, Frost SDW (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, **22**, 1208–1222.

Krebs RA, Feder ME (1998) Experimental manipulation of the cost of thermal acclimation in *Drosophila melanogaster*. *Biological Journal of the Linnean Society*, **63**, 593–601.

Kreitman M (2000) Methods to detect selection in populations with applications to the human. *Annual Review of Genomics & Human Genetics*, **1**, 539–559.

Langley CH, Lazzaro BP, Phillips W, Heikkinen E, Braverman JM (2000) Linkage disequilibria and the site frequency spectra in the su (s) and su (w) (a) regions of the *Drosophila melanogaster* X chromosome. *Genetics*, **156**, 1837–1852.

Lemos B, Meiklejohn CD, Caceres M, Hartl DL (2005) Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. *Evolution*, **59**, 126–137.

Lenski RE (2004) Phenotypic and genomic evolution during a 20 000-generation experiment with the bacterium *Escherichia coli*. *Plant Breeding Reviews*, **24**, 225–265.

Lewontin RC (1974) *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York and London.

Lewontin RC, Cockerham CC (1959) The goodness of fit test for detecting natural-selection in random mating populations. *Evolution*, **13**, 561–564.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Lexer C, Heinze B, Alia R, Rieseberg LH (2004) Hybrid zones as a tool for identifying adaptive genetic variation in outbreeding forest trees: lessons from wild annual sunflowers (*Helianthus* spp.). *Forest Ecology and Management*, **197**, 49–64.

Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution*, **36**, 96–99.

Li E, Beard C, Jaenisch R (1993) Role for DNA methylation in genomic imprinting. *Nature*, **366**, 362–365.

Liu H-C, Cheng HH, Tirunagaru V, Sofer L, Burnside J (2001) A strategy to identify positional candidate genes conferring Marek's disease resistance by integrating DNA microarrays and genetic mapping. *Animal Genetics*, **32**, 351–359.

Lo HS, Wang ZN, Hu Y *et al.* (2003) Allelic variation in gene expression is common in the human genome. *Genome Research*, **13**, 1855–1862.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

MacDougall-Shackleton EA, Blanchard L, Gibbs HL (2003) Unmelanized plumage patterns in old world leaf warblers do not correspond to sequence variation at the melanocortin-1 receptor locus (*MCIR*). *Molecular Biology and Evolution*, **20**, 1946–1946.

Mackay TFC (2001) The genetic architecture of quantitative traits. *Annual Review of Genetics*, **35**, 303–339.

Mallet J (2005) Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, **20**, 229–237.

Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–1762.

Mauricio R (2001) Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. *Nature Reviews Genetics*, **2**, 370–381.

Mauricio R (2005) Can ecology help genomics: the genome as ecosystem? *Genetica*, **123**, 205–209.

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favorable gene. *Genetical Research*, **23**, 23–35.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, **351**, 652–654.

McKeigue PM (2005) Prospects for admixture mapping of complex traits. *American Journal of Human Genetics*, **76**, 1–7.

Messler W, Stewart CB (1997) Episodic adaptive evolution of primate lysozymes. *Nature*, **385**, 151–154.

Mitchell AA, Cutler DJ, Chakravarti A (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *American Journal of Human Genetics*, **72**, 598–610.

Moen T, Fjalestad KT, Munck H, Gomez-Raya L (2004) A multi-stage testing strategy for detection of quantitative trait loci affecting disease resistance in Atlantic salmon. *Genetics*, **167**, 851–858.

Montana G, Pritchard JK (2004) Statistical tests for admixture mapping with case-control and cases-only data. *American Journal of Human Genetics*, **75**, 771–789.

Moody DE, Zou Z, McIntyre L (2002) Cross-species hybridisation of pig RNA to human nylon microarrays. *BMC Genomics*, **3**.

Morley M, Molony CM, Weber TM *et al.* (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.

Mundy NI, Badcock NS, Hart T, Scribner K, Janssen K, Nadeau NJ (2004) Conserved genetic basis of a quantitative plumage trait involved in mate choice. *Science*, **303**, 1870–1873.

Myburg AA, Remington DL, O'Malley DM, Sederoff RR, Whetten RW (2001) High-throughput AFLP analysis using infrared dye-labeled primers and an automated DNA sequencer. *BioTechniques*, **30**, 348.

Nachman MW, Hoekstra HE, D'Agostino SL (2003) The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences, USA*, **100**, 5268–5273.

Nadeau JH, Frankel WN (2000) The roads from phenotypic variation to gene discovery: mutagenesis versus QTLs. *Nature Genetics*, **25**, 381–384.

Nair S, Williams JT, Brockman A *et al.* (2003) A selective sweep driven by pyrimethamine treatment in Southeast Asian malaria parasites. *Molecular Biology and Evolution*, **20**, 1526–1536.

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. *Trends in Plant Science*, **9**, 325–330.

Neff BD, Gross MR (2001) Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats. *Evolution*, **55**, 1717–1733.

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, **3**, 418–426.

Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity*, **86**, 641–647.

Nielsen R, Yang ZH (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929–936.

Nordborg M, Borevitz JO, Bergelson J *et al.* (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*, **30**, 190–193.

Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM (2004) Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Molecular Biology and Evolution*, **21**, 1308–1317.

Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. *Nature Genetics*, **32**, 261–266.

Pastinen T, Hudson TJ (2004) Cis-acting regulatory variation in the human genome. *Science*, **306**, 647–650.

Piccino P, Viard F, Sarradin PM *et al.* (2004) Thermal selection of *PGM* allozymes in newly founded populations of the thermotolerant vent polychaete *Alvinella pompejana*. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **271**, 2351–2359.

Planes S, Romans P (2004) Evidence of genetic selection for growth in new recruits of a marine fish. *Molecular Ecology*, **13**, 2049–2060.

Porter AH (2003) A test for deviation from island-model population structure. *Molecular Ecology*, **12**, 903–915.

Prioul JL, Pelleschi S, Sene M *et al.* (1999) From QTLs for enzyme activity to candidate genes in maize. *Journal of Experimental Botany*, **50**, 1281–1288.

Pritchard JK, Donnelly P (2001) Case–control studies of association in structured or admixed populations. *Theoretical Population Biology*, **60**, 227–237.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. *American Journal of Human Genetics*, **67**, 170–181.

Ptak SE, Hinds DA, Koehler K *et al.* (2005) Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genetics*, **37**, 429–434.

Remington DL, Thornsberry JM, Matsuoka Y *et al.* (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences, USA*, **98**, 11479–11484.

Rieseberg LH, Buerkle CA (2002) Genetic mapping in hybrid zones. *American Naturalist*, **159**, S36–S50.

Rieseberg LH, Whitton J, Gardner K (1999) Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, **152**, 713–727.

Rieseberg LH, Baird SJE, Gardner KA (2000) Hybridization, introgression, and linkage evolution. *Plant Molecular Biology*, **42**, 205–224.

Rieseberg LH, Raymond O, Rosenthal DM *et al.* (2003) Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, **301**, 1211–1216.

Rifkin SA, Kim J, White KP (2003) Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nature Genetics*, **33**, 138–144.

Rise ML, von Schalburg KR, Brown GD *et al.* (2004) Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Research*, **14**, 478–490.

Robbins LS, Nadeau JH, Johnson KR *et al.* (1993) Pigmentation phenotypes of variant extension locus alleles result from point mutations that alter *MSH* receptor function. *Cell*, **72**, 827–834.

Rodriguez-Trelles F (2004) Evolutionary genetics — Transcriptome evolution — much ado about nothing? *Heredity*, **93**, 405–406.

Rodriguez-Trelles F, Tarrio R, Ayala FJ (2003) Evolution of cis-regulatory regions versus codifying regions. *International Journal of Developmental Biology*, **47**, 665–673.

Rogers SM, Bernatchez L (2005) Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). *Molecular Ecology*, **14**, 351–361.

Rogers SM, Campbell D, Baird SJE, Danzmann RG, Bernatchez L (2001) Combining the analyses of introgressive hybridisation and linkage mapping to investigate the genetic architecture of population divergence in the lake whitefish (*Coregonus clupeaformis*, Mitchill). *Genetica*, **111**, 25–41.

Ronald J, Akey JM, Whittle J *et al.* (2005) Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Research*, **15**, 284–291.

Rosenblum EB, Hoekstra HE, Nachmanc MW (2004) Adaptive reptile color variation and the evolution of the *MC1R* gene. *Evolution*, **58**, 1794–1808.

Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequences? *Trends in Plant Science*, **8**, 321–329.

Sabeti PC, Reich DE, Higgins JM *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Sawyer SL, Wu LI, Emerman M, Malik HS (2005) Positive selection of primate *TRIM5* alpha identifies a critical species-specific retroviral restriction domain. *Proceedings of the National Academy of Sciences, USA*, **102**, 2832–2837.

Schaart JG, Mehli L, Schouten HJ (2005) Quantification of allele-specific expression of a gene encoding strawberry polygalacturonase-inhibiting protein (PGIP) using pyrosequencing (TM). *Plant Journal*, **41**, 493–500.

Schadt EE, Monks SA, Drake TA *et al.* (2003a) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.

Schadt EE, Monks SA, Friend SH (2003b) A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets. *Biochemical Society Transactions*, **31**, 437–443.

Schlötterer C (2002a) A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics*, **160**, 753–763.

Schlötterer C (2002b) Towards a molecular characterization of adaptation in local populations. *Current Opinion in Genetics and Development*, **12**, 683–687.

Schlötterer C (2003) Hitchhiking mapping — functional genomics from the population genetics perspective. *Trends in Genetics*, **19**, 32–38.

Schlötterer C, Kauer M, Dieringer D (2004) Allele excess at neutrally evolving microsatellites and the implications for tests of neutrality. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **271**, 869–874.

Schmidt PS, Rand DM (1999) Intertidal microhabitat and selection at *Mpi*: interlocus contrasts in the northern acorn barnacle, *Semibalanus balanoides*. *Evolution*, **53**, 135–146.

Schulte PM, Glemet HC, Fiebig AA, Powers DA (2000) Adaptive variation in lactate dehydrogenase-B gene expression: role of a stress-responsive regulatory element. *Proceedings of the National Academy of Sciences, USA*, **97**, 6597–6602.

Scotti-Saintagne C, Mariette S, Porth I *et al.* (2004) Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. & *Q. petraea* (Matt.) Liebl.]. *Genetics*, **168**, 1615–1626.

Sezgin E, Duvernell DD, Matzkin LM *et al.* (2004) Single-locus latitudinal clines and their relationship to temperate adaptation in metabolic genes and derived alleles in *Drosophila melanogaster*. *Genetics*, **168**, 923–931.

Simko I (2004) One potato, two potato: haplotype association mapping in autotetraploids. *Trends in Plant Science*, **9**, 441–448.

Singer-Sam J, Chapman V, Lebon JM, Riggs AD (1992) Parental imprinting studied by allele-specific primer extension after PCR — paternal-X chromosome-linked genes are transcribed prior to preferential paternal X-chromosome inactivation. *Proceedings of the National Academy of Sciences, USA*, **89**, 10469–10473.

Slate J (2005) Quantitative trait locus mapping in natural populations: progress, caveats and future directions. *Molecular Ecology*, **2**, 363–379.

Somorjai IML, Danzmann RG, Ferguson MM (2003) Distribution of temperature tolerance quantitative trait loci in Arctic charr (*Salvelinus alpinus*) and inferred homologies in rainbow trout (*Oncorhynchus mykiss*). *Genetics*, **165**, 1443–1456.

Stamatoyannopoulos JA (2004) The genomics of gene expression. *Genomics*, **84**, 449–457.

Stearns SC, Magwene P (2003) The naturalist in a world of genomics. *American Naturalist*, **161**, 171–180.

Stolz U, Velez S, Wood KV, Wood M, Feder JL (2003) Darwinian natural selection for orange bioluminescent color in a Jamaican click beetle. *Proceedings of the National Academy of Sciences, USA*, **100**, 14955–14959.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

Storz JF, Dubach JM (2004) Natural selection drives altitudinal divergence at the albumin locus in deer mice, *Peromyscus maniculatus*. *Evolution*, **58**, 1342–1352.

Storz JF, Payseur BA, Nachman MW (2004) Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Molecular Biology and Evolution*, **21**, 1800–1811.

Streelman JT, Albertson RC, Kocher TD (2003) Genome mapping of the orange blotch colour pattern in cichlid fishes. *Molecular Ecology*, **12**, 2465–2471.

Suzuki Y (2004) New methods for detecting positive selection at single amino acid sites. *Journal of Molecular Evolution*, **59**, 11–19.

Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, **16**, 1315–1328.

Suzuki Y, Nei M (2001) Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, **18**, 2179–2185.

Suzuki Y, Nei M (2002) Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, **19**, 1865–1869.

Suzuki Y, Nei M (2004) False-positive selection identified by ML-based methods: examples from the *Sig1* gene of the diatom *Thalassiosira weissflogii* and the tax gene of a human T-cell lymphotropic virus. *Molecular Biology and Evolution*, **21**, 914–921.

Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. *Nature Reviews Genetics*, **3**, 137–144.

Symonds VV, Godoy AV, Alconada T *et al.* (2005) Mapping quantitative trait loci in multiple populations of *Arabidopsis thaliana* identifies natural allelic variation for trichome density. *Genetics*, **169**, 1649–1658.

Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics*, **3**, 391–A396.

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

Takeuchi S, Suzuki H, Yabuuchi M, Takahashi S (1996) A possible involvement of melanocortin 1-receptor in regulating feather color pigmentation in the chicken. *Biochimica et Biophysica Acta — Gene Strucure and Expression*, **1308**, 164–168.

Templeton AR, Maxwell T, Posada D *et al.* (2005) Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. *Genetics*, **169**, 441–453.

Theron E, Hawkins K, Bermingham E, Ricklefs RE, Mundy NI (2001) The molecular basis of an avian plumage polymorphism in the wild: a melanocortin-1-receptor point mutation is perfectly associated with the melanic plumage morph of the bananaquit, *Coereba flaveola*. *Current Biology*, **11**, 550–557.

Thiellement H, Zivy M, Plomion C (2002) Combining proteomic and genetic studies in plants. *Journal of Chromatography B—Analytical Technologies in the Biomedical & Life Sciences*, **782**, 137–149.

Thomas MA, Klaper R (2004) Genomics for the ecological toolbox. *Trends in Ecology & Evolution*, **19**, 439–445.

Tiffin P, Hahn MW (2002) Coding sequence divergence between two closely related plant species: *Arabidopsis thaliana* and *Brassica rapa* ssp. *pekinensis*. *Journal of Molecular Evolution*, **54**, 746–753.

Tsakas S, Krimbas CB (1976) Testing heterogeneity of *F*-values — suggestion and a correction. *Genetics*, **84**, 399–401.

Tsoi SCM, Cale JM, Bird IM *et al.* (2003) Use of human cDNA microarrays for identification of differentially expressed genes in Atlantic salmon liver during *Aeromonas salmonicida* infection. *Marine Biotechnology*, **5**, 545–554.

Valverde P, Healy E, Jackson I, Rees JL, Thody AJ (1995) Variants of the melanocyte stimulating hormone receptor are associated with red hair and fair skin in humans. *Nature Genetics*, **11**, 328–330.

van Tienderen PH, de Haan AA, van der Linden CG, Vosman B (2002) Biodiversity assessment using markers for ecologically important traits. *Trends in Ecology & Evolution*, **17**, 577–582.

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology*, **23**, 48–55.

Vasemägi A, Nilsson J, Primmer CR (2005) Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar* L.). *Molecular Biology and Evolution*, **22**, 1067–1076.

Veliz D, Bourget E, Bernatchez L (2004) Regional variation in the spatial scale of selection at *MPI** and *GPI** in the acorn barnacle *Semibalanus balanoides* (Crustacea). *Journal of Evolutionary Biology*, **17**, 953–966.

Venkatesh B, Hettwer U, Koopmann B, Karlovsky P (2005) Conversion of cDNA differential display results (DDRT-PCR) into quantitative transcription profiles. *BMC Genomics*, **51**.

Verrelli BC, Eanes WF (2001) Clinal variation for amino acid polymorphisms at the Pgm locus in *Drosophila melanogaster*. *Genetics*, **157**, 1649–1663.

de Vienne D, Bost B, Fievet J, Zivy M, Dillmann C (2001) Genetic variability of proteome expression and metabolic control. *Plant Physiology and Biochemistry*, **39**, 271–283.

Vigouroux Y, McMullen M, Hittinger CT *et al.* (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proceedings of the National Academy of Sciences, USA*, **99**, 9650–9655.

Vigouroux Y, Mitchell S, Matsuoka Y *et al.* (2005) An analysis of genetic diversity across the maize genome using microsatellites. *Genetics*, **169**, 1617–1630.

Watt WB, Dean AM (2000) Molecular-fuctional studies of adaptive genetic variation in procaryotes and eucaryotes. *Annual Reviews in Genetics*, **34**, 593–622.

Vitalis R, Dawson K, Boursot P (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics*, **158**, 1811–1823.

Vos P, Hogers R, Bleeker M *et al.* (1995) AFLP — a new technique for DNA-fingerprinting. *Nucleic Acids Research*, **23**, 4407–4414.

Watterson GA (1977) Heterosis or neutrality. *Genetics*, **85**, 789–814.

Wayne ML, McIntyre LM (2002) Combining mapping and arraying: An approach to candidate gene identification. *Proceedings of the National Academy of Sciences, USA*, **99**, 14903–14906.

Wiener P, Burton D, Ajmone-Marsan P *et al.* (2003) Signatures of selection? Patterns of microsatellite diversity on a chromosome containing a selected locus. *Heredity*, **90**, 350–358.

Wilding CS, Butlin RK, Grahame J (2001) Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *Journal of Evolutionary Biology*, **14**, 611–619.

Wong WSW, Yang ZH, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, **168**, 1041–1051.

Wootton JC, Feng XR, Ferdig MT *et al.* (2002) Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature*, **418**, 320–323.

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology and Evolution*, **22**, 506–519.

Yan H, Zhou W (2004) Allelic variations in gene expression. *Current Opinion in Oncology*, **16**, 39–43.

Yan H, Yuan WS, Velculescu VE, Vogelstein B, Kinzler KW (2002) Allelic variation in human gene expression. *Science*, **297**, 1143–1143.

Yang ZH (2002) Inference of selection from multiple species alignments. *Current Opinion in Genetics and Development*, **12**, 688–694.

Yang ZH, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, **15**, 496–503.

Zaykin DV, Westfall PH, Young SS *et al.* (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity*, **53**, 79–91.

Zhang J (2004) Frequent false detection of positive selection by the likelihood method with branch-site models. *Molecular Biology and Evolution*, **21**, 1332–1339.

Zhao HY, Pfeiffer R, Gail MH (2003) Haplotype analysis in population genetics and association studies. *Pharmacogenomics*, **4**, 171–178.

Zhu XF, Luke A, Cooper RS *et al.* (2005) Admixture mapping for hypertension loci with genome-scan markers. *Nature Genetics*, **37**, 177–181.

Anti Vasemägi is a postdoc in Craig Primmer's research group. The group is one of three units in the recently formed Finnish Centre of Excellence in Evolutionary Genetics and Physiology. Future research aims in the group include application of some of the approaches outlined in this review to identify functionally important genetic variation in non-mammalian vertebrates.
