

Gene expression profiling in evolutionary genetics

DANIEL L. HARTL, COLIN D. MEIKLEJOHN,
CRISTIAN I. CASTILLO-DAVIS

Department of Organismic and Evolutionary Biology Harvard University

DUCCIO CAVALIERI

Harvard Center for Genomics Research Harvard University

JOSÉ MARIA RANZ, JEFFREY P. TOWNSEND

Department of Organismic and Evolutionary Biology Harvard University

4.1 Introduction

Lewontin (1974) has characterized much of the history of population genetics as “the struggle to measure variation,” especially genetic variation at the molecular level. His characterization portrays a time when population geneticists were severely limited in the techniques that could be applied to organisms in natural populations. Fortunately, during the past 25 years molecular biology has supported a steady stream of innovative approaches and techniques that are widely applicable to natural populations. Chief among these have been chain-termination methods of DNA sequencing (Sanger *et al.* 1977) and the polymerase chain reaction (Saiki *et al.* 1985). From these have emerged high-throughput DNA sequencing strategies resulting in the complete sequences of the genomes of innumerable organelles, viruses, prokaryotes, and agents of infectious disease, as well as the genomes of most of the key model organisms used in molecular genetics and, of course, the human genome. The availability of genomic sequences has already resulted in the new field of comparative genomics (Koonin *et al.* 2000).

By contrast, in population genetics the struggle to measure variation was largely a struggle to detect differences between genotypes of organisms within a single species. Variation within populations is important because it is essential to Darwinism to understand how genetic differences within species become transformed into differences between species over evolutionary time. Here, too, molecular techniques have eased the struggle. Although complete sequencing of multiple genomes from a single species has so far been restricted to a few organelles, viruses, and bacteria, the great interest in human single-nucleotide polymorphisms (Sachidanandam *et al.* 2001) for

¹ The last five authors contributed equally to this work.

disease-association studies has provided impetus for genome-wide studies of polymorphism in other organisms using either low-redundancy shotgun sequencing (e.g., random three-fold redundancy provides about 95% genome coverage) or high-density oligonucleotide arrays (Winzeler *et al.* 1998).

The upshot of the last 25 years is that population geneticists are no longer caught up in a struggle to measure variation. Population geneticists are rather awash in variation. The new struggle is not to measure variation, but to interpret and understand genetic variation in the context of population history, geographical structure, patterns of migration, and the internal constraints and external evolutionary forces that affect the level and distribution of genetic variation within and between genomes.

It remains a key challenge for evolutionary biology to learn how variation in genotype within populations of organisms is ultimately associated with adaptive variation in phenotype. Several new molecular approaches offer promising opportunities to meet this challenge. Among these are DNA microarrays that enable expression profiling (genome-wide analysis of the relative abundances of gene transcripts). In principle, this approach can form a bridge connecting genotype with phenotype, because specific, reproducible patterns of transcription associated with particular genotypes may also be associated with particular phenotypes and affect Darwinian fitness. Application of expression profiling to natural populations is still in its earliest stages. Hence, in this chapter we focus on some of the issues that are raised in these applications, and also give some examples.

4.2 Déjà vu all over again?

Although it is an admirable tradition of evolutionary biologists to put techniques developed by others to their own uses, the application to evolutionary issues often challenges the techniques and interpretations beyond the range of their original intent. Protein electrophoresis provides an example. A major advance in the study of genetic variation came with the use of starch-gel electrophoresis to study protein variation in natural populations of *Drosophila* (Hubby and Lewontin 1966, Lewontin and Hubby 1966). Starch-gel electrophoresis had been invented much earlier by Smithies (1954), who employed it to detect human genetic variation. However, when electrophoresis was used for a systematic analysis of genetic variation in natural populations, it became important to know what fraction of amino acid replacements present in a population are detected as changes in electrophoretic mobility, how many amino acid replacements differ between two electrophoretically distinct proteins encoded by alleles of a single gene, and whether polymorphic amino acid replacements affect fitness (Lewontin 1991). These issues were largely irrelevant in the original use of the technique to separate heterogeneous serum proteins (Smithies 1995). Likewise, the application of DNA sequencing to natural populations poses unique problems, because

differences in sequence between alleles do not necessarily reflect differences in fitness between genotypes. The challenge of inferring population structure, population history, and evolutionary forces based on DNA sequence has stimulated important advances in theoretical population genetics (Yang 1998, Fu and Li 1999, Wakeley 1999, Nielsen and Wakeley 2001, Bustamante *et al.* 2002).

The application of expression profiling to natural populations also raises special problems not encountered in typical experiments with laboratory populations. For example, one standard kind of experiment with laboratory populations examines global gene expression in a knockout mutant compared against an isogenic wild-type strain grown under the same conditions (Hughes *et al.* 2000). Hence there is one variable only – the gene that has been knocked out. Another standard kind of experiment examines gene expression of a single strain under two or more conditions, such as yeast cells grown in the presence or absence of ethanol (Alexandre *et al.* 2001). Here again there is a single variable, this time environmental in origin. In both examples, the genotypes are under the experimenter's control.

With natural populations, the situation is different. Most natural populations are highly heterozygous (Lewontin 1974), hence genotypes differ within populations as well as between populations. How much of the sequence variation between organisms is reflected in differences in gene expression that can be detected experimentally? In studying natural populations, a major issue to be faced in dealing with expression profiling is that, because of heterozygosity within populations, differential expression of any given gene may result from differences in DNA sequence within regulatory domains of the gene itself (cis-acting regulatory elements), or from epistatic effects of differences in one or more other genes (trans-acting regulatory elements). Correlating sequence variation with expression-profile variation therefore requires much more than a spreadsheet of genes, sequences, and expression levels.

In addition to issues arising from epistasis, another comes from potential nonlinearity of the transcriptional response, because a relatively small change in the expression of a trans-acting regulatory gene may produce a large change in the transcriptional level of the genes that it affects. While these and other problems should not be minimized, there is nevertheless great potential for expression profiling to reveal which differences in gene expression are associated with differences in fitness, and to learn how these differences relate to environmental variables in the habitat. Understanding the contributions of sequence variation and epistasis to differences in gene expression is vital to an incorporation of expression profiling into evolutionary genetics.

4.3 DNA microarrays and their applications

Two types of DNA microarrays (“DNA chips”) are commonly available. The first consists of 200 000–400 000 synthetic oligonucleotides (Chee *et al.* 1996,

Lipshutz *et al.* 1999), typically 25-mers, which are complementary to regions of genomic sequence and can be used for genotyping (Gingeras *et al.* 1998, Winzeler *et al.* 1998) or for assaying relative levels of gene-expression (Lockhart *et al.* 1996, Lockhart and Winzeler 2000). The second consists of up to 20 000 DNA fragments, typically cDNA clones amplified by the polymerase chain reaction, which are used primarily for competitive RNA hybridization to assess relative levels of gene expression (DeRisi *et al.* 1997, Eisen and Brown 1999). Hybridization of mRNA samples is equivalent to a massively parallel set of Northern blots, and hybridization of a genomic DNA sample is equivalent to a whole-genome Southern blot. Hence the fundamental principle of nucleic acid hybridization with labeled probes is nothing new: microarrays are merely the application of microscale high-throughput technology to leverage these techniques.

Whether used for clinical diagnosis or studies in evolutionary genetics, DNA microarrays generate a blizzard of data concerning the relative abundance (in two or more samples of mRNA, cDNA, or genomic DNA) of molecules that have high sequence similarity to the DNA at a particular position in a microarray. The ability to distinguish between signal and background noise is determined by the hybridization methods, the number of replicates of a given experiment, and the nature of the statistical analysis. In principle, if there is adequate replication, even a small difference in abundance of an mRNA sequence between two samples can be determined with a high degree of statistical confidence. However, once a difference in relative abundance has been found, microarrays alone are limited in their ability to identify the source of the difference. This is because a significant difference in level of an mRNA species between two samples could be due to a difference at any point in a hierarchy of regulatory processes that affect transcription, RNA processing, or mRNA stability. These include differences in both cis-regulatory and trans-regulatory elements, and in some cases differential mRNA abundance reflects changes in gene copy number (Lucito *et al.* 2000). One potential pitfall in evolutionary applications comes from the ability of individual members of conserved multi-gene families to cross-hybridize, which confounds gene expression with gene copy number. This caveat is particularly relevant to transposable elements, whose copy number can differ dramatically between individuals and between species. Consequently, for any particular coding sequence, tracking down the source of an observed difference requires more detailed study of the gene and its regulatory elements.

To date, most applications of microarray technology have dealt with clinical diagnosis (DeRisi *et al.* 1996, Kononen *et al.* 1998, Scherf *et al.* 2000), identifying interactions among metabolic pathways that control metabolic flux (DeRisi *et al.* 1997, Wodicka *et al.* 1997), investigating control mechanisms of the cell cycle (Cho *et al.* 1998, Chu *et al.* 1998), comparing patterns of gene expression of organisms with the same genotype grown under different conditions (Hardwick *et al.* 1999, Jelinsky and Samson 1999, Hughes *et al.* 2000,

Roberts *et al.* 2000), or describing the patterns in which genes are progressively deployed during development (White *et al.* 1999, Hill *et al.* 2000). Among the first applications of DNA microarrays to evolutionary studies were those of yeast strains that had undergone adaptive evolution in laboratory culture (Ferea *et al.* 1999).

With regard to natural populations, few studies have been published that deal with the extent of variation in either levels or patterns of global gene expression among organisms isolated from natural environments. Several key questions set the agenda for molecular evolutionary biology as it enters the post-genomics era. Is there significant variation in gene expression from one organism to the next? How many genes are differentially regulated, and to what extent? What are the molecular mechanisms behind the regulatory variation? Are there particular sets of related genes whose expression tends to vary together as a result of pleiotropic effects?

4.4 Statistical analysis

Considering the large number of pairwise and higher-order interactions that are possible, the statistical analysis of genome-wide expression data is a multivariate problem of extremely high dimensionality. Several types of clustering algorithms, such as hierarchical clustering or self-organizing maps, have been suggested as exploratory tools to identify genes with correlated expression profiles in order to discover networks of coordinately expressed genes and to assign open reading frames of unknown function to regulatory clusters (Bittner *et al.* 1999, Tavazoie *et al.* 1999, Alter *et al.* 2000, Holter *et al.* 2000, Jensen and Knudson 2000, Kim *et al.* 2000). More recent approaches have exploited analysis of variance to decompose the variance in gene expression into components due to genotype, sex, environment, and other factors, as well as due to their interactions (Kerr *et al.* 2000, Gibson 2002).

For application to natural populations, the first issue for expression profiling is level of resolution. What magnitude of difference in relative intensity of signal (for example, signal from mRNA isolated from two strains of *Drosophila*) can be regarded as significantly different from background noise? Much of the earlier literature uses the rule of thumb that a difference in relative intensity is considered significantly above background if the magnitude of the larger intensity is at least twofold greater than that of the smaller intensity. One problem with this rule is that it fails to take replications into account, because a difference that is smaller than twofold but consistent across multiple independent experiments is likely to be biologically meaningful. Another problem with the twofold rule is that false positives will often occur when the weaker signal is near the background level, because the signal intensity is usually corrected by subtracting the background, which in the twofold rule results in division by a number close to zero. This is, of course, a problem with any method of analysis that uses ratios.

Another approach to assessing the significance of a difference would be a conventional Student's t test in which the difference between means for the same gene across replicates is compared with the average standard deviation across replicates. This approach has the drawback that, unless the number of replicates is relatively large, the variance across replicates tends to be underestimated (Baldi and Long 2001). To overcome this limitation, Baldi and Long (2001) have developed a hierarchical Bayesian approach to the t test that regards the logarithm of the ratio of expression levels for each gene as a normal distribution whose mean, given the variance, is itself assumed to be normal and whose variance is assumed to be inverse gamma. From this analysis they implement a t test by adjusting the empirical variance of each log-expression value according to a local background variance associated with neighboring genes in the microarray. A type of nonparametric t test has also been suggested, in which the significance level for a given gene is assessed by comparing the test statistic with those obtained from random permutations of the data (Dudoit *et al.* 2000).

For evolutionary genetics, it is important to be able to compare levels of gene expression across experiments with different genotypes, because typically genotype A will be compared in one experiment with genotype B , genotype B in another experiment with genotype C , genotype C in still another experiment with genotype D , and so forth. This design is complicated further by the fact that each experiment may be replicated a different number of times, and there may also be some ad hoc comparisons, for example, genotype A with genotype C . Any real experiment is likely to be unbalanced in its comparisons and asymmetrical in its number of replicates, if for no other reason than that not all microarray hybridization experiments yield satisfactory data, and also because repetition of failed experiments is avoided if possible because of the time and expense.

One example of such a set of comparisons is shown in Figure 4.1, taken from unpublished work of J. Townsend and D. Cavalieri. The organisms in question are a parental strain of homothallic diploid yeast (M28) and four diploid progeny obtained from germination of the spores in a single tetrad (S1, S2, F1, and F2). Each arrow indicates a single microarray hybridization, and the arrowhead points to the strain whose mRNA was labeled with cyanine 3 fluorochrome ("green") versus cyanine 5 fluorochrome ("red").

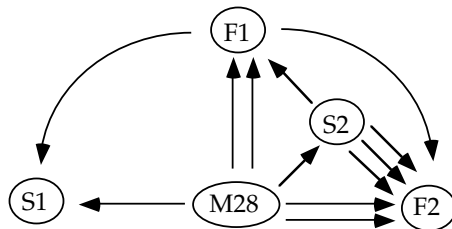


Figure 4.1. A typical unbalanced experimental design of gene-expression comparisons. M28 is a homothallic diploid natural isolate of *S. cerevisiae*, and S1, S2, F1, and F2 are four diploid progeny obtained from a single ascus.

diagram is not shown to suggest this as an ideal experimental design, but rather to show the unpredictable result of simultaneous exploratory research and technology development. The initial exploratory research indicated that, in addition to the parental strain, the progeny S2 and F2 were of greatest interest, hence these strains warranted the greatest number of additional experiments (Cavalieri *et al.* 2000). However, an optimal method of analysis would allow the multiple strains, unbalanced design, and unequal numbers of replicates to be taken into account to obtain, for each gene, an estimate of the relative expression level in each strain and a confidence interval around this estimate.

To enable the analysis of such complex experimental designs, Townsend (unpublished) has implemented a Bayesian analysis known as BAGEL (the acronym stands for Bayesian Analysis of Gene Expression Levels). In this approach, the level of each signal (green or red) in each experiment is regarded as coming from a normal distribution with some unknown mean and variance. Each gene is treated as independent. This assumption is not literally true for genes that are coordinately regulated, but it serves as a useful first approximation for evaluating gene-expression levels on a gene-by-gene basis. The algorithm implemented in BAGEL is based on the hierarchical assumption that the mean and variance of the normal distribution of signal intensity at any position in the microarray are random variables with uniform distributions on the positive real numbers. These distributions are vague, and they are also “improper” in the sense that they do not integrate to 1. Such a choice gives primacy to the observations, rather than the prior distributions, in estimating the parameters of interest.

The conditional distribution of the parameters in this model, given the data, are based on computer simulations defining a Markov chain that converges to the correct stationary distribution. The mean, mode, credible interval, and other characteristics of the parameter distributions then approximated by sampling from a long trajectory of this Markov chain. In Bayesian analysis, the 95% credible interval is the analog of the 95% confidence interval in conventional frequentist statistics, although the credible interval has the straightforward interpretation that 95% of the simulated realizations of the sample mean fall within the credible interval.

In the analysis of expression profiles of yeast isolates from natural populations discussed below, the mean levels of gene expression have been estimated using BAGEL. The mean expression levels are relative values, not absolutes. This is because the mean expression level for any gene in any treatment (e.g., genotype) is expressed as a ratio relative to the smallest mean across all treatments. Hence the expression levels are ranked with the smallest value set arbitrarily to 1. As a conservative criterion for a significant difference between any pair of estimates, we require that the 95% credible intervals be nonoverlapping.

4.5 DNA microarrays in evolutionary genetics: some specific examples

Microarray technology has only recently been used for evolutionary studies. *Saccharomyces cerevisiae* was the first eukaryotic organism for which DNA microarrays became available, and partly for this reason budding yeast is being used increasingly as a model system to study evolution in laboratory populations (Ferea *et al.* 1999, Zeyl 2000). Among other approaches, Hartwell *et al.* (1999) have advocated examining evolutionary constraints on functional modules in order to dissect the modularity of living systems, and Murray (2000) has stressed the issue of the evolvability of organisms.

Use of expression profiling to study the population genetics and molecular evolution of natural populations is only just beginning, but already it is clear that the possibilities are virtually unlimited. In this section we give some examples. Necessarily they have emerged from genomic sequencing and other research resources in model organisms (*Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Drosophila melanogaster*), because these are the organisms in which DNA microarrays first became available. However, as the technology has improved and disseminated, comparable approaches are rapidly being developed in many other organisms of evolutionary interest.

4.5.1 Molecular evolution of genes expressed early and late in nematode development

Microarray technology offers a substantial bridge to connect organismic with molecular evolution. For example, global gene-expression studies on the developmental timing and tissue-specific expression of genes (White *et al.* 1999, Hill *et al.* 2000) opens a large new field for evolutionary investigations. Among the issues that can be addressed are the role of genomic and developmental complexity in influencing evolutionary change, the relationship between variation in gene sequences versus variation in gene expression, and the molecular mechanisms of intraspecific and interspecific morphological divergence.

Using gene-expression data to inform molecular evolutionary studies has already begun to yield interesting results concerning the relationships between organismic and molecular evolution. For example, it has been shown that rates of protein evolution in mammals are markedly affected by tissue-specific patterns of expression; in particular, rates of nonsynonymous substitution are negatively correlated with breadth of tissue expression, which has been interpreted as resulting from selection against mutations that would result in strongly pleiotropic effects (Duret and Mouchiroud 2001). In a comparison of genes between *C. elegans* and *C. briggsae*, Castillo-Davis and Hartl (2002) found that genes expressed during embryogenesis have significantly fewer paralogs than do genes expressed after embryogenesis is completed. Among

early-expressed genes, 5–15% have paralogous copies, whereas among late-expressed genes 35–40% have paralogous copies in both genomes. It is not yet known whether the greater number of paralogs results from positive selection for tissue-specific expression or function (e.g., olfactory receptors) or from a shorter average persistence of duplicated copies of genes that are expressed early in embryogenesis due to deleterious dosage effect.

Whole-genome expression data from *C. elegans* along with comparative genomic sequence from *C. briggsae* has also afforded a test of the hypothesis that genes expressed early in embryogenesis are more conserved in sequence through evolutionary time than genes expressed later in life. For the number of nonsynonymous substitutions per nonsynonymous site, the 95% confidence intervals are 0.034–0.058 for genes expressed early and 0.040–0.067 for genes expressed late. There is clearly no significant difference, although this analysis deals with the average protein-wide rates of evolution rather than with the rates for particular functional sequence motifs. In contrast to the results for nonsynonymous substitutions, the rates of synonymous substitutions are very different between genes expressed early and late. For the number of synonymous substitutions per synonymous site, the 95% confidence intervals are 1.09–1.63 for the early genes and 0.76–0.98 for the late genes (Castillo-Davis and Hartl 2002). Genome-wide there is also a highly significant positive correlation between expression level and codon usage bias (Castillo-Davis and Hartl 2002). Because late-expressed genes have, on the average, higher expression levels than early-expressed genes (Castillo-Davis and Hartl 2002, from data in Hill *et al.* 2000), and hence more codon usage bias, a smaller rate of synonymous substitution is expected.

4.5.2 Expression profiling of natural isolates of vineyard yeast

For many years it was a mystery where yeast could be found in a natural environment, because yeast could not easily be isolated from grapes in vineyards (Mortimer 1999). Yet the organism must exist in vineyards, or in the wineries themselves, because crushed grapes begin to ferment owing to the presence of naturally occurring yeast. The mystery was resolved by the discovery that, whereas wine yeast is rarely found on grapes with an unbroken skin, viable cells are found in about one-third of damaged berries, inside of which they establish a little fermentation chamber (Mortimer 1999). Furthermore, most vineyard isolates are diploid and about 70% are homothallic (Mortimer 2000), which means that, after sporulation and the germination of a haploid spore, the mother cell changes mating type and mates with the daughter cell, forming a diploid that is completely homozygous except for the mating type locus. The diploid phase must persist long enough without sporulation to allow a significant number of mutations to occur, because there is known to be a great deal of functional heterozygosity among vineyard isolates. For example, when vineyard isolates are sporulated and their progeny tested for growth

on the sugars sucrose, maltose, and galactose, approximately 67% of the isolates segregate for the inability to utilize at least one of these sugars (Cavaliere *et al.* 1998). When a homothallic diploid undergoes sporulation, each haploid spore undergoes germination, mating-type switching, and mating to form a homozygous diploid once again. This process of rendering the genome completely homozygous has been called “genome renewal” (Mortimer *et al.* 1994).

Among natural isolates of vineyard yeast isolated from around the Tuscan wine capital of Montalcino, any randomly chosen pair of strains shows a significant difference in expression level for 1–2% of their genes (Townsend, unpublished). Although the percentage of differentially expressed genes is relatively small, the absolute number is large (60–120 genes), and the magnitude of the differential expression is rather large (e.g., twofold or more).

One interesting difference is found in the gene *SSU1*, which encodes a sulfite exporter. Mean expression levels in four Tuscan isolates are shown in Figure 4.2. The comparisons were carried out with a “circular” design in which each isolate is compared with the next in line, completing the circle by comparing the last with the first; each comparison was replicated twice by interchanging the flours, and the mean relative expression levels and their 95% credible intervals were estimated using BAGEL. Each of the Tuscan isolates has an *SSU1* expression level that is different from the others, and the

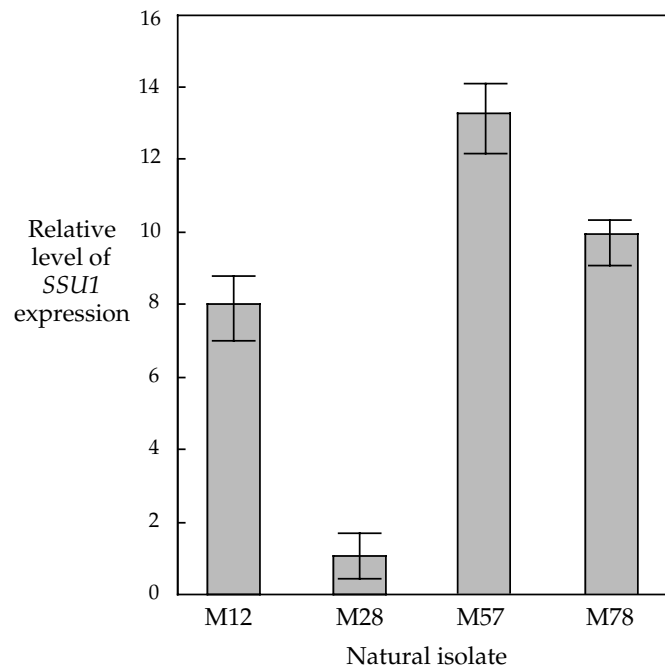


Figure 4.2. Mean expression levels of the sulfite exporter *SSU1* in four natural isolates of *S. cerevisiae* and their 95% credible intervals.

differences are all statistically significant. M28 has low expression of this sulfite exporter, and its progeny show segregation of a gene strongly affecting the closely related sulfur assimilation and methionine biosynthesis pathways (see below). It is interesting to speculate that the differences in *SSUI* expression may result from natural selection for resistance to sulfites in the environment. For as long as 200 years, Tuscan vintners have been treating vineyards with copper sulfate to inhibit the growth of molds on the grapes, and sodium sulfite, potassium metabisulfite, and sulfur dioxide are widely used during and after fermentation to stabilize the wine and kill bacteria. In this case the expression profiling may have revealed an evolutionary consequence of changes in the chemical ecology of vineyards.

The isolate M28 in Figures 4.1 and 4.2 proved to be particularly interesting. The original isolate attracted attention because it exhibited a slightly rough, delicately filigreed colony morphology. The progeny colonies showed 2:2 segregation for a smooth morphology (spores S1 and S2) or a more extremely filigreed morphology (spores F1 and F2). These are the organisms depicted in Figure 4.1.

For the full BAGEL analysis of the 12 arrays depicted in Figure 4.1, some 145 genes (~2% of the genome) had estimated levels of expression in F1 and S2 whose 95% credible intervals were nonoverlapping; 101 of these showed greater expression in F1, and 44 showed greater expression in S2. In contrast, the full BAGEL analysis of F1 and F2 showed only three genes (<0.1% of the genome) whose 95% credible intervals were nonoverlapping. The discrepancy of 145:3 is far too large to be explained by random segregation in two pairs of ascospores, hence it appeared likely that most of the differences between the F and S segregants resulted from the epistatic effects of one factor, or perhaps a small number of factors. From the standpoint of evolutionary genetics, it is worth noting that there is no detectable difference in growth rate between the filigreed and the smooth segregants under laboratory conditions, in spite of the dramatic difference in expression profile. Both types of segregants are extremely vigorous; in fact, in comparison with a standard laboratory strain, they both have a fitness advantage of 33–50%.

Detailed analysis showed that the most overexpressed genes in F1 relative to S2 encode metabolic enzymes, particularly those associated with amino acid biosynthesis. The most highly overexpressed genes include 12 in the methionine pathway, two in the serine pathway, two in the histidine pathway, and one in the arginine pathway. Entire metabolic pathways are upregulated, including the pathway from pyruvate to valine and leucine, that from phosphoribosyl pyrophosphate to histidine, and that leading from extracellular surface to methionine (Cavaliere *et al.* 2000). On the other hand, among the genes that are underexpressed in F1, relative to S2, are many amino acid permeases and transporters (Cavaliere *et al.* 2000). These results may be compared to those of Wodicka *et al.* (1997), who examined the expression profiles of a laboratory strain grown in minimal medium versus rich medium. Among 51 genes

expressed more abundantly in minimal medium, 8 encoded proteins involved in arginine and methionine biosynthesis, and among 84 genes expressed much more abundantly in rich medium, 7 encoded amino acid permeases.

The gene *PHD1* is a good example of the additional power gained from taking all experiments and replicates into account, even with a design as complex as that in Figure 4.1. This gene is a principal transcriptional regulator of filamentous growth in the morphogenetic pathway induced by ammonia starvation (Gimeno and Fink 1994). Since the *MEP2* ammonium permease regulates pseudohyphal differentiation in *S. cerevisiae* (Lorenz and Heitman 1998), and *MEP2* is overexpressed sixfold in the filamentous segregants, we were somewhat surprised to find that the upregulation of *PHD1* was less than twofold (Cavaliere *et al.* 2000). However, BAGEL analysis of the results of the experiments in Figure 4.1 yield the mean levels of *PHD1* expression and the 95% credible intervals shown in Figure 4.3. The mean expression of each strain is shown immediately above the bar. Both of the filigreed segregants have a level of *PHD1* expression that is significantly greater than either of the smooth segregants, even though the difference between F2 and S1 is a factor

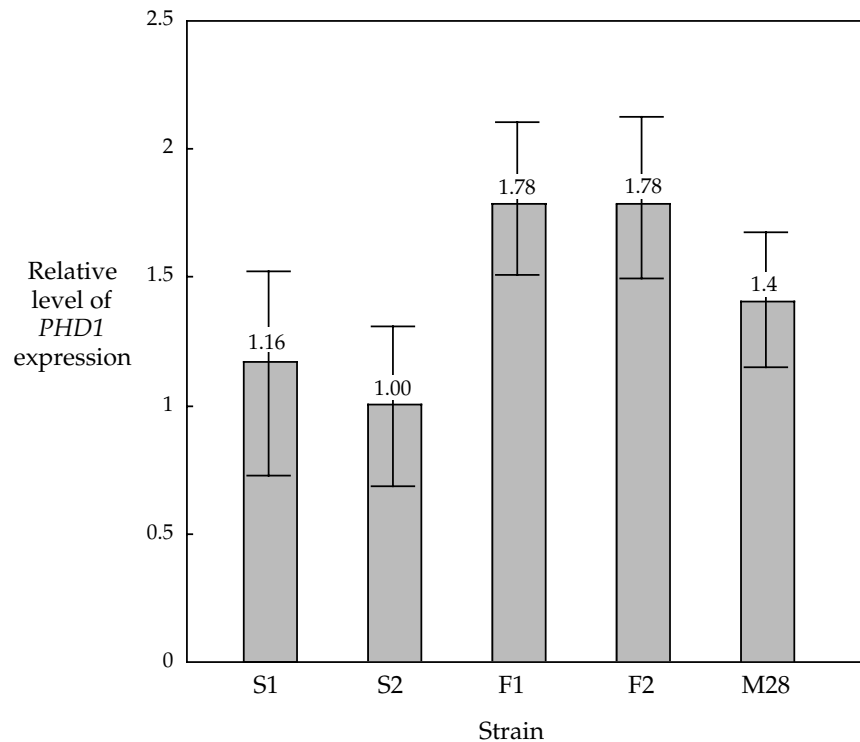


Figure 4.3. Mean expression levels of *PHD1*, a transcriptional regulator of filamentous growth, and their 95% credible intervals, estimated from a Bayesian analysis of the comparisons in Figure 4.1.

of 1.53. Hence, with sufficient replication, even relatively small differences can become significant. Although *PHD1* expression in M28 is not significantly different from either type of segregant, it is interesting that its level is intermediate, which is consistent with its less pronounced filigreed colony phenotype.

4.5.3 Variation in gene expression in *Drosophila*

Unpublished observations of C. Meiklejohn and J. Ranz also indicate that variation in gene-expression profiles exists between wild-type populations of *Drosophila melanogaster*. Using the standard twofold cutoff for a significant difference in replicated hybridizations, they found that 0.5–2% of approximately 4500 genes are differentially expressed between different strains, either strains collected from different parts of the world or different wild-type laboratory strains. This is in roughly the same range as found in yeast, and it will be interesting to learn whether the similarity holds as more extensive surveys of natural populations are carried out.

Thus far, the observed pattern of genome-wide expression variation follows the degree of genetic differentiation of the strains inferred from the sequences of a relatively small number of individual genes. In particular, strains from Africa (Zimbabwe) appear to harbor more sequence diversity than strains from elsewhere in the world (Begun and Aquadro 1993), and some African strains may be partially reproductively isolated from the rest of the species (Alipaz *et al.* 2001). In regard to gene-expression profiles, there is a greater number of differentially expressed genes (twofold cutoff) between African strains ($32/4489 = 0.71\%$) and between African and laboratory strains ($44/3793 = 1.16\%$ and $69/4250 = 1.53\%$) than between any two laboratory strains ($20/4427 = 0.45\%$). Eight genes that were highly variable (more than twofold in three comparisons) were analyzed by BAGEL. These eight genes showed varying patterns of variation, with some differences specific to a given strain, and some shared between Zimbabwe and laboratory strains. One gene appears to be differentially expressed at four levels between the strains analyzed.

Large differences in expression profile are associated with sex. Up to 6% of genes examined are more than five-fold upregulated or downregulated in males as compared with females. This is far larger than the differential expression found between males or between females, irrespective of their geographical origin. However, such a large number of differentially regulated genes between the sexes may be more apparent than real, because some unknown fraction of these differences undoubtedly results from allometric relationships of anatomy and cell types between the sexes.

Among the male-specific genes, a few are also differentially expressed in males from different populations. These sex-specific and geographically divergent genes are candidates for sexual selection or fast evolving genes involved in reproduction (Civetta and Singh 1995, Wu *et al.* 1996, Swanson *et al.* 2001).

Interestingly, one gene that is highly female specific (upregulated an average of 55-fold in females) is downregulated in Zimbabwe males relative to laboratory strains. Why a female specific gene would be downregulated in Zimbabwe males (or upregulated in laboratory males) is unclear. One possibility is that high expression of the gene has deleterious pleiotropic effects in males, but another is that the divergence is selectively neutral and results either from random genetic drift or from the manifestation in males of differential regulation in females.

Hybridization of microarrays with genomic DNA can also be used to identify differences in copy number between genotypes. For example, two genes appear to be either greatly reduced in copy number or entirely missing in Canton-S relative to a Zimbabwe strain, and one gene appears to be missing in the Zimbabwe strain relative to Canton-S. Two of these sequences are transposable elements, so differences in copy number are to be expected, but the third is unique-sequence DNA missing in Canton-S.

4.5.4 Interspecific comparisons in *Drosophila*

DNA microarrays also allow whole-genome comparisons between species that otherwise must be done on a gene-by-gene basis. Using traditional Southern blots, Schmid and Tautz (1997) found that about one-third of the open reading frames had diverged enough between *D. melanogaster* and *D. virilis* to fail in cross-hybridization. A similar figure was inferred from the success rate of *in situ* hybridizations of probes from *D. melanogaster* onto polytene chromosomes of *D. repleta* (Ranz *et al.* 2001). These species pairs both have estimated divergence times of 40 million years (Russo *et al.* 1995).

An example of this approach using microarrays comes from unpublished work of J. Ranz and C. Meiklejohn. They carried out a comparison between *D. melanogaster* and *D. simulans* in order to assess to what extent the expression profile of adults of the species has been modified during the approximately 3 million years since their divergence from a common ancestor. For this purpose, cDNA microarrays from *D. melanogaster* were used for competitive hybridization with mRNA extracted from adults of each sex of each species. As shown in the top part of Table 4.1, 4.4% of approximately 4800 spots on the microarray showed differential expression between the species using a twofold cutoff. About 75% of these genes are apparently overexpressed in *D. melanogaster*, relative to *D. simulans*. A total of 170 genes are differentially expressed in the other species in one sex only, and among these genes more than 80% are differentially expressed in males. How these findings may relate to the rapid sequence evolution of male-specific genes (Civetta and Singh 1995, 1998, Wu *et al.* 1996) has yet to be determined.

A somewhat different picture emerges when genes differentially expressed in the sexes are compared between the species. About 15% of the spots on the microarray fall into this category, and the comparative data are shown

Table 4.1. *Comparison of expression profiles of D. melanogaster and D. simulans*

<i>Genes differentially expressed by species and overexpressed twofold or more in:</i>	
Both sexes of <i>D. melanogaster</i>	34
Both sexes of <i>D. simulans</i>	11
Males only of <i>D. melanogaster</i>	114
Males only of <i>D. simulans</i>	28
Females only of <i>D. melanogaster</i>	18
Females only of <i>D. simulans</i>	10
TOTAL	215 (4.4%)
<i>Genes differentially expressed by sex and overexpressed twofold or more in:</i>	
Males of both <i>D. melanogaster</i> and <i>D. simulans</i>	211
Males of <i>D. melanogaster</i> only	90
Males of <i>D. simulans</i> only	63
Females of both <i>D. melanogaster</i> and <i>D. simulans</i>	251
Females of <i>D. melanogaster</i> only	182
Females of <i>D. simulans</i> only	102
TOTAL	899 (19%)

in the bottom part of Table 4.1. Among these genes, the majority (60%) are overexpressed in females. Furthermore, among those that are overexpressed in only one of the species, the majority (65%) are apparently overexpressed in females only.

We stress the use of the term “apparently overexpressed” when comparing *D. melanogaster* with *D. simulans* because the efficiency of hybridization of a probe depends not only on the relative expression level but also on the degree of sequence divergence. To distinguish between these possibilities it is also necessary to carry out competitive hybridizations with genomic DNA from the species. Preliminary data by Ranz and Meiklejohn indicate that a relatively small fraction of the differences in Table 4.1 can be attributed to sequence divergence. Hence, each of the genes that shows differential expression between the species is a candidate for having undergone interesting and informative regulatory changes in its recent evolution.

4.6 Concluding remarks

In population genetics, much of the last third of the twentieth century was devoted to answering the question: How much genetic variation is present in natural populations and how is it maintained? The struggle to measure variation was eased first by the application of protein electrophoresis (Lewontin 1974) and then virtually eliminated by the development of automated DNA sequencing and the polymerase chain reaction. The “how much” part of the question has been answered (or at least is answerable) in any organism to which these techniques can be applied. Much more difficult are issues related

to the adaptive significance of genetic variation within species and of genetic differences between species. The neutral theory (Kimura 1983) was debated fervently when data were scant and analytical methods only beginning to be developed. Much has happened both technologically and analytically, making it possible to close the book on the neutral theory and to turn to other questions. For example, how does polymorphism and divergence of regulatory elements compare with polymorphism and divergence of protein-coding sequences? Do genes with high levels of amino acid polymorphism also have high levels of expression polymorphism? Do genes that show evidence of positive selection at the amino acid level also show evidence of positive selection of their regulatory sequences? Do genes that show evidence of relaxed selection at the amino acid level also have relaxed selection on regulatory variation? More generally, considering the recent advances in functional genomics and proteomics, including expression profiling, the opportunities are auspicious for evolutionary biologists to be able to identify the genetic, cellular, developmental, and organismic changes that drive adaptive evolution and speciation.

4.7 Acknowledgments

This work was supported by NIH grants GM60035, GM58423, and HG01250 to DLH, and by a fellowship from the National Research Council of Spain to JMR. We thank John Parsch, Yun Tao, and Justin Blumenstiel of the Hartl lab for their help and advice with *Drosophila* microarrays; the Harvard Center for Genomics Research, especially Rachel Erlich, Claire Bailey, and Andrew Murray for their support and encouragement; and all the other members of the “*Drosophila* chip” consortium including Neal Silverman, Inez Alvarez-Garcia, Eric Bernstein, Yong Dai, Rich Dearborne, Stephanie Mohr, Sam Kunes, Bill Gelbart, and Tom Maniatis.

REFERENCES

- Alexandre, H., Ansanay-Galeote, V., Dequin, S., and Blondin, B. (2001). Global gene expression during short-term ethanol stress in *Saccharomyces cerevisiae*. *FEBS Letters* 498:98–103.
- Alipaz, J. A., Wu, C. I., and Karr, T. L. (2001). Gametic incompatibilities between races of *Drosophila melanogaster*. *Proc. R. Soc. Lond. Ser. B* 268:789–95.
- Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* 97:10101–6.
- Baldi, P., and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* 17:509–19.
- Begun, D. J., and Aquadro, C. F. (1993). African and North American populations of *Drosophila melanogaster* are very different. *Nature* 365:548–50.

- Bittner, M., Meltzer, P., and Trent, J. (1999). Data analysis and integration: of steps and arrows. *Nature Genet.* 22:213–5.
- Bustamante, C., Nielsen, R., Sawyer, S. A., Olsen, K. M., Purugganan, M. D., and Hartl, D. L. (2002). The cost of inbreeding in *Arabidopsis*. *Nature* 416:531–4.
- Castillo-Davis, C. I., and Hartl, D. L. (2002). Genome evolution and developmental constraint in *Caenorhabditis elegans*. *Mol. Biol. Evol.* 19:728–35.
- Cavalieri, D., Barberio, C., Casalone, E., Pinzauti, F., Sebastiani, F., Mortimer, R. K., and Polsinelli, M. (1998). Genetic and molecular diversity in *S. cerevisiae* natural populations. *Food Technol. Biotechnol.* 36:45–50.
- Cavalieri, D., Townsend, J. P., and Hartl, D. L. (2000). Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by DNA microarray analysis. *Proc. Natl. Acad. Sci. USA* 97:12369–74.
- Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., and Fodor, S. P. A. (1996). Accessing genetic information with high-density DNA arrays. *Science* 274:610–14.
- Cho, R. J., Campbell, M. J., Winzler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molec. Cell* 2:65–73.
- Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science* 282:699–705.
- Civetta, A., and Singh, R. S. (1995). High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. *J. Mol. Evol.* 41:1085–95.
- Civetta, A., and Singh, R. S. (1998). Sex-related genes, directional sexual selection, and speciation. *Mol. Biol. Evol.* 15:901–9.
- DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y. D., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* 14:457–60.
- DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680–6.
- Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Dept. Statistics Technical Report No. 578*, Univ. Calif., Berkeley.
- Duret, L., and Mouchiroud, D. (2001). Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* 17:68–74.
- Eisen, M. B., and Brown, P. O. (1999). DNA arrays for analysis of gene expression. *Methods Enzymol.* 303:179–205.
- Ferea, T. L., Botstein, D., Brown, P. O., and Rosenzweig, R. F. (1999). Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl. Acad. Sci. USA* 96:9721–6.
- Fu, Y. X., and Li, W. H. (1999). Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theor. Popul. Biol.* 56:1–10.
- Gibson, G. (2002). Microarrays in ecology and evolution: a preview. *Molec. Ecol.* 11:17–24.

- Gimeno, C. J., and Fink, G. R. (1994). Induction of pseudohyphal growth by over-expression of *PHD1*, a *Saccharomyces cerevisiae* gene related to transcriptional regulators of fungal development. *Mol. Cell. Biol.* 14:2100–12.
- Gingeras, T. R., Ghandour, G., Wang, E. G., Berno, A., Small, P. M., Drobniewski, F., Alland, D., Desmond, E., Holodniy, M., and Drenkow, J. (1998). Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic Mycobacterium DNA arrays. *PCR Methods & Applications* 8:435–48.
- Hardwick, J. S., Kuruvilla, F. G., Tong, J. K., Shamji, A. F., and Schreiber, S. L. (1999). Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins. *Proc. Natl. Acad. Sci. USA* 96:14866–70.
- Hartwell, L. H., Hopfield, J. J., Leibler, S. S., and Murray, A. W. (1999). From molecular to modular biology. *Nature* 402:C47–52.
- Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G., and Brown, E. L. (2000). Genomic analysis of gene expression in *C. elegans*. *Science* 290:809–12.
- Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R., and Fedoroff, N. V. (2000). Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl. Acad. Sci. USA* 97:8409–14.
- Hubby, J. L., and Lewontin, R. C. (1966). A molecular approach to the study of genic heterozygosity in natural populations I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics* 54:577–94.
- Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburty, K., Simon, J., Bard, M., and Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell* 102:109–26.
- Jelinsky, S. A., and Samson, L. D. (1999). Global response of *Saccharomyces cerevisiae* to an alkylating agent. *Proc. Natl. Acad. Sci. USA* 96:1486–91.
- Jensen, L. J., and Knudson, S. (2000). Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics* 16:326–33.
- Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *J. Computational Biol.* 7:819–37.
- Kim, S., Dougherty, E. R., Chen, Y., Sivakumar, K., Meltzer, P., Trent, J. M., and Bittner, M. (2000). Multivariate measurement of gene expression relationships. *Genomics* 67:201–9.
- Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge, England: Cambridge University Press.
- Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M. J., Sauter, G., and Kallioniemi, O. P. (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine* 4:844–7.
- Koonin, E. V., Aravind, L., and Kondrashov, A. S. (2000). The impact of comparative genomics on our understanding of evolution. *Cell* 101:573–6.
- Lewontin, R. C. (1974). *The Genetic Basis of Evolutionary Change*. New York, NY: Columbia University Press.

- Lewontin, R. C. (1991). Electrophoresis in the development of evolutionary genetics: milestone or millstone? *Genetics* 128:657–62.
- Lewontin, R. C., and Hubby, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54:595–609.
- Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nature Genet.* 21(Suppl S): 20–4.
- Lockhart, D. J., Dong, H. L., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C. W., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* 14:1675–80.
- Lockhart, D. J., and Winzeler, E. W. (2000). Genomics, gene expression and DNA arrays. *Nature* 405:827–36.
- Lorenz, M. C., and Heitman, J. (1998). The MEP2 ammonium permease regulates pseudohyphal differentiation in *Saccharomyces cerevisiae*. *EMBO J.* 17:1236–47.
- Lucito, R., West, J., Reiner, A., Alexander, J., Esposito, D., Mishra, B., Powers, S., Norton, L., and Wigler, M. (2000). Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Res.* 10:1726–36.
- Mortimer, R. K. (1999). On the origins of wine yeast. *Res. Microbiol.* 150:199–204.
- Mortimer, R. K. (2000). Evolution and variation of the yeast (*Saccharomyces*) genome. *Genome Res.* 10:891–9.
- Mortimer, R. K., Romano, P., Suzzi, G., and Polsinelli, M. (1994). Genome renewal: a new phenomenon revealed from a genetic study of 43 strains of *Saccharomyces cerevisiae* derived from natural fermentation of grape musts. *Yeast* 10:1543–52.
- Murray, A. W. (2000). Whither genomics? *Genome Biol.* 1:1–6.
- Nielsen, R., and Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158:885–96.
- Ranz, J. M., Casals, F., and Ruiz, A. (2001). How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res.* 11:230–9.
- Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., Tyers, M., Boone, C., and Friend, S. H. (2000). Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287:873–80.
- Russo, C. A. M., Takezaki, N., and Nei, M. (1995). Molecular phylogeny and divergence times of Drosophilid species. *Mol. Biol. Evol.* 12:391–404.
- Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z. M., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., Altshuler, D. et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–33.
- Saiki, R. K., Scharf, S. J., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. (1985). Enzymatic amplification of β -globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230:1350–4.

- Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-termination inhibitors. *Proc. Natl. Acad. Sci. USA* 74:5463–7.
- Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., Brown, P. O., and Weinstein, J. N. (2000). A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* 24:236–44.
- Schmid, K. J., and Tautz, D. (1997). A screen for fast evolving genes from *Drosophila*. *Proc. Natl. Acad. Sci. USA* 94:9746–50.
- Smithies, O. (1954). Zone electrophoresis in starch gels: group variation in the serum proteins of normal human adults. *Biochem. J.* 61:629–41.
- Smithies, O. (1995). Early days of gel electrophoresis. *Genetics* 139:1–3.
- Swanson, W. J., Clark, A. G., Waldrip-Dail, H. M., Wolfner, M. F., and Aquadro, C. F. (2001). Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proc. Natl. Acad. Sci. USA* 98:7375–9.
- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genet.* 22:281–5.
- Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics* 153:1863–71.
- White, K. P., Rifkin, S. A., Hurban, P., and Hogness, D. S. (1999). Microarray analysis of *Drosophila* development during metamorphosis. *Science* 286:2179–84.
- Winzeler, E. A., Richards, D. R., Conway, A. R., Goldstein, A. L., Kalman, S., Mccullough, M. J., Mccusker, J. H., Stevens, D. A., Wodicka, L., Lockhart, D. J., and Davis, R. W. (1998). Direct allelic variation scanning of the yeast genome. *Science* 281:1194–7.
- Wodicka, L., Dong, H. L., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997). Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.* 15:1359–67.
- Wu, C.-I., Johnson, N. A., and Palopoli, M. F. (1996). Haldane's rule and its legacy: why are there so many sterile males? *Trends Ecol. Evol.* 11:281–4.
- Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15:568–73.
- Zeyl, C. (2000). Budding yeast as a model organism for population genetics. *Yeast* 16:773–84.