ELSEVIER

Journal of Chromatography B, 782 (2002) 3–25

JOURNAL OF CHROMATOGRAPHY B

www.elsevier.com/locate/chromb

Review

Sequence variation and the biological function of genes: methodological and biological considerations

Peter J. Oefner*

Genome Technology Center, Stanford University, 855 California Avenue, Palo Alto, CA 94304, USA

Abstract

Single nucleotide polymorphisms (SNPs) are expected to facilitate the chromosomal mapping and eventual cloning of genetic determinants of complex quantitative phenotypes. To date, more than 2.5 million non-redundant human SNPs have been reported in the public domain, of which approximately 100 000 have been validated by either independent investigators or by independent methods. Equally impressive is the myriad of methods developed for allelic discrimination. Nevertheless, reports of successful applications of SNPs to genome-wide linkage analysis of both mono- and polygenic traits are rare and limited to a few model organisms, that provide affordable platforms to test both novel methodological and biological concepts at a whole-genome scale under conditions that can be reasonably controlled. Progress in the analysis of SNPs needs to be complemented by methods that allow the systematic elucidation of both primary and secondary phenotypes of genes. Importantly, observations made in one species may very well be of immediate applicability to other species including human. This is particularly true for conserved biological processes such as mitochondrial respiration and DNA repair.

Keywords: Reviews; Denaturing HPLC; Single nucleotide polymorphism; Pyrosequencing; RNA

Contents

*Tel.: +1-650-812-1926; fax: +1-650-812-1975.
E-mail address: oefner@genome.stanford.edu (P.J. Oefner).

1570-0232/02/$ – see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S1570-0232(02)00699-2

1. Introduction

Single-nucleotide polymorphisms (SNPs) have been heralded as key to understanding the genetic factors that determine susceptibility and predisposition to such common diseases as diabetes, hypertension, and schizophrenia [1–3], or predict individual variability to drug response [4]. However, the number of instances in which certain alleles could be definitely associated with a certain phenotype remains small. Among the more well known examples are the association of the apolipoprotein E type 4 allele with late-onset familial Alzheimer disease [5], the protection against HIV-1 infection in individuals homozygous for the *CCR5-Δ32* allele [6], the association of the common Pro12Ala polymorphism in peroxisome proliferator-activated receptor-γ with type 2 diabetes [7], and the *APC* T3920A [8] and *CHEK2* 1100delC [9] variants, that are responsible for an increased susceptibility to colorectal and breast cancer, respectively. Reduced representation shotgun sequencing [10] and in silico comparison of overlapping genomic sequences generated by the human genome project [11] have resulted in the identification of millions of SNPs. This has been accompanied by the development of numerous methods for both individual and pooled genotyping of SNPs [12,13], many of which have been never assessed critically with regard to their success rate, accuracy, and cost effectiveness using, preferentially, a large common set of unbiased SNPs.

Here we review different techniques for both the discovery and genotyping of SNPs as well as for the functional characterization of genes that we have developed or, at least, tested over the years and applied to the cloning of the genetic determinants of simple and complex phenotypes. As such it is a personal account written with the intention of encouraging young analytical chemists to venture beyond the boundaries of their field of immediate interest.

2. Map-based cloning of induced mutations in *Arabidopsis thaliana*

In human, simple Mendelian recessive and dominant disorders have been mapped and cloned successfully by linkage analysis of extended families using microsatellites. In plants, in contrast, the object of study is usually not a naturally occurring monogenic trait, but rather that of a phenotype brought about by various mutagenesis procedures, such as insertional mutagenesis, gene silencing, and physical or chemical mutagenesis. This has proven quite effective in elucidating the function of unknown genes. While insertional mutagenesis offers the advantage that the gene affected can be easily identified, it lacks the versatility to create partial loss-of-function or gain-of-function alleles. The latter are particularly informative in dissecting all functional domains of a protein and the different phenotypes associated with a specific genetic locus. An interesting human example is the α_{1A}-voltage-dependent calcium channel gene *CACNL1A4*. Amino acid replacement and truncating mutations in this ion channel gene have been described to cause episodic disease where periods of well-being are interrupted by hemiplegic migraine or ataxia, i.e. an inability to coordinate voluntary muscular movements [14]. In contrast, expansion of a CAG repeat or polyglutamine tract in the same gene produces a permanent and progressive ataxia [15]. A major drawback in creating mutations by chemical or physical means is the laborious task of identifying the locus responsible for the mutant phenotype. Traditionally, this entails the tracking of DNA sequence variants that co-segregate with heritable traits. The majority of known DNA sequence variants in use over the past 20 years have been polymorphisms affecting the recognition sequence of a restriction endonuclease, commonly referred to as restriction fragment length polymorphisms (RFLPs) [16]. Their use in positional cloning has been hampered by the laborious process of digestion of genomic DNA with various restriction enzymes, gel electrophoretic separation and transfer of the cleaved fragments onto a nitrocellulose membrane, and detection of the DNA fragments by hybridization with radioactive probe sequences and autoradiography. This is an impractical process when hundreds of RFLPs are required for mapping a gene within a few centimorgans. With the development of the polymerase chain reaction (PCR), modified versions of RFLPs have been developed, such as cleaved amplified polymorphic sequences (CAPS) [17] and amplified fragment length polymorphisms

(AFLPs) [18]. The CAPS method uses amplified unique DNA fragments that are digested with a restriction endonuclease to display polymorphic sites that affect cleavage, and the products are analyzed by gel electrophoresis. Tens of CAPS of markers have been developed, that share an *AluI* or *Sau3A* restriction site and allow the mapping of a gene to one of the arms of the five *Arabidopsis* chromosomes. Plans of identifying hundreds of such markers have never materialized due to technical problems with the enrichment of *AluI* or *Sau3A* restriction site containing sequences by means of genomic subtraction. The AFLP method, in comparison, digests genomic DNA with restriction endonucleases, followed by linker addition and amplification with random sequence-tagged primers to yield fragment length polymorphisms. Its major advantage is that multiple markers can be analyzed in a single lane of a sequencing gel. Its major disadvantage is the inability to distinguish between homo- and heterozygotes. A general disadvantage of all methods relying on restriction endonucleases is that less than one third of polymorphisms affect the recognition sequence of a restriction enzyme. This prompted a still ongoing large-scale discovery effort for simple sequence polymorphisms by, initially, DHPLC and, subsequently, conventional sequencing of the two most commonly used ecotypes in genetic research of *Arabidopsis thaliana*, namely Columbia and Landsberg *erecta* [19].

2.1. High-density oligonucleotide variable detector array

In order to demonstrate the utility of SNPs for genome-wide linkage analysis in Arabidopsis, we applied a variable detector array (VDA) [20,21] to genotype 412 of the 487 initially discovered polymorphic loci between Columbia and Landsberg *erecta* [19]. The 75 polymorphisms, which had been excluded from genotyping, were embedded in sequence prone to indiscriminate or poor hybridization based on prior experience [20]. Indiscriminate hybridization is expected if a 20-mer probe has ≥8Cs, or if a window of 8 bases in a 20-mer probe has ≥4Cs. Poor hybridization can result if a 20-mer probe contains (1) >8 As, (2) >9 Ts, (3) T=0, (4) a run of As, Gs, or Ts >4, (5) a run of 2 bases >10, (6) a palindrome >6 bases, (7) A+T >13, and (8) A+G >14. The remaining 412 SNPs were interrogated by a total of 1648 variable detector arrays, each consisting of 44 25-mer probes (Fig. 1). Four VDAs are required to genotype a simple sequence polymorphism, two for each allele (one for the reverse strand and one for the forward). For each of the 11 positions examined, including the polymorphic site itself and five bases on each site, the VDA has a set of four 25-mer oligonucleotide probes. These probes are complementary to the reference sequence, except at the central, examined position, for which each of the four nucleotides is substituted in turn. Usually the reference sequence can be read from the hybridization pattern, because the perfectly matching probe yields a much stronger hybridization signal than do the three mismatching oligonucleotide probes in the same column. This scheme, which tests not only the polymorphic site but also a variety of contexts around the polymorphic site, with mismatches as controls, provides a much more sensitive and reliable assay for genotyping. Ideally, the forward and reverse strand hybridization patterns should corroborate each other. However, quite frequently only one of the two strands yields a hybridization pattern of sufficient quality for accurate allele calling. That is the reason for the redundant nature of the array design.

To enable the multiplex amplification of tens of polymorphic loci in a single PCR, we designed primers with similar calculated melting temperatures that flank the polymorphic site by a few bases on either side. Markers were then amplified individually and approximate PCR yield was determined on a standard ethidium bromide-stained agarose gel. Subsequently, 45–55 markers that had shown similar yields were pooled to amplify them in multiplex reactions. This multiplex strategy was shown to maintain discrimination for more than 95% of markers that had distinguished well between homozygotes in singleplex reactions [19]. The multiplex reactions were then pooled for a single chasing reaction to label all amplicons with biotin prior to visualization with a streptavidin R-phycoerythrin conjugate. For that purpose, 5′ biotinylated primers complementary to 23-mer T7 and T3 sequences incorporated at the 5′ ends of the primary forward and reverse amplimers, respectively, were used. The

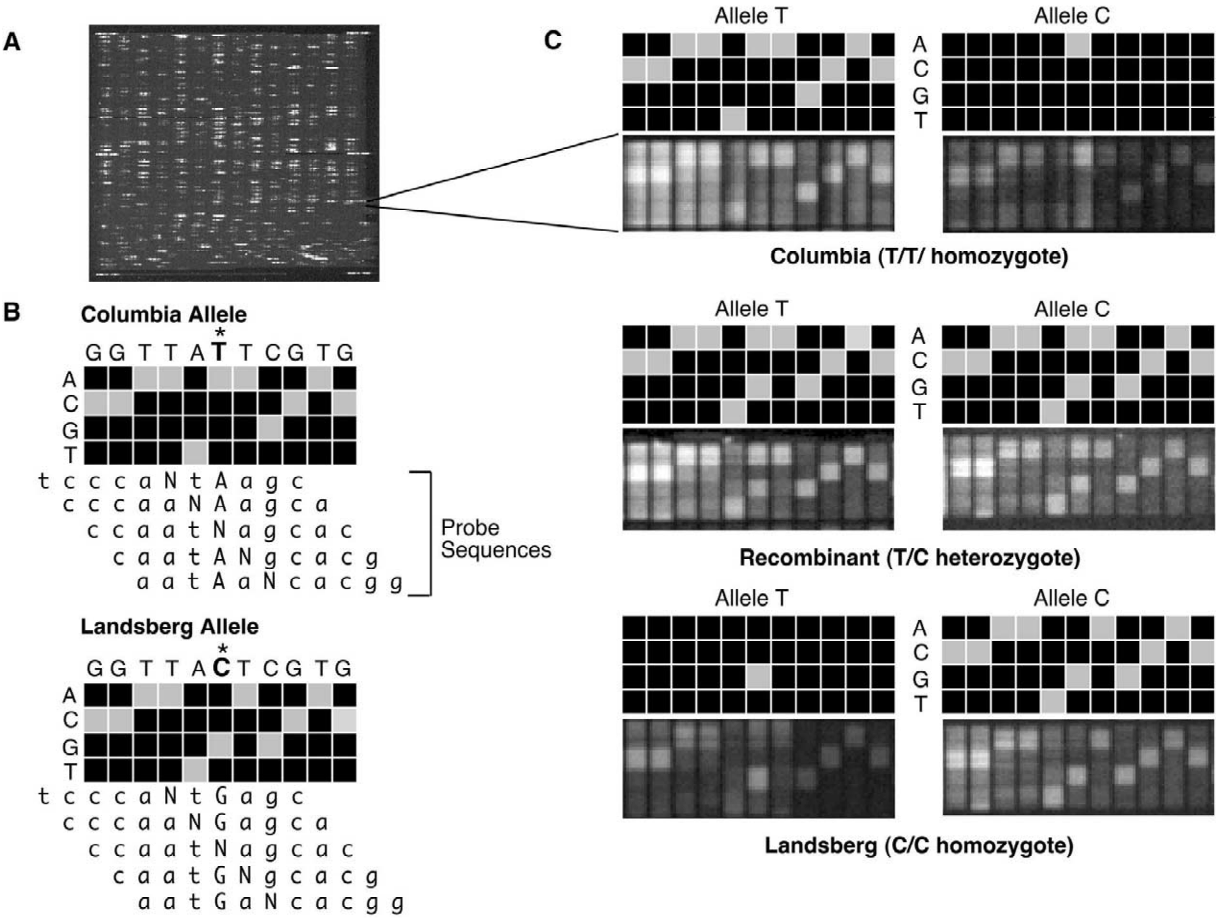

Fig. 1. Genotyping by differential array hybridization. (A) Fluorescence image of an entire oligonucleotide probe array following hybridization. (B) Scheme of the genotyping of the forward strand of SGCSNP4 in Columbia and Landsberg *erecta* ecotypes on a variant detector high-density oligonucleotide probe array. Variant detector arrays are designed to interrogate not only the polymorphic site (marked with an asterisk) using four 25-mer probes that have an A, C, G or T at the center position (N), but also the flanking five bases on either side. This design allows the determination of the sequence context in which the polymorphic site is embedded and adds to the robustness and accuracy of the genotyping assay. The target DNA hybridizes most strongly to the probe that complements its sequence most closely. Therefore, the probe with the correct base at each center position will produce the strongest hybridization signal. (C) The scans show the actual and schematic hybridization patterns for homozygous Columbia (top), homozygous Landsberg *erecta* (bottom) and a heterozygous recombinant (center). In this example, the variant bases are T and C. Hybridization of the T allele to the C allele variant detector array, and vice versa, will result only in one strong hybridization signal in the column that interrogates the polymorphic site itself. Interrogation of the flanking bases will yield no or only weak signals as the target sequence does not match perfectly the corresponding probe sequences due to the different allelic state at the polymorphic site. Reproduced from Ref. [22] with permission.

fluorescence intensity readings for every feature on the chip are recorded and processed sequentially using a package of algorithms publicly available at ftp://www.tairpub:tairpub@ftp.arabidopsis.org/home/tair/Software to compute the probability of observing a segregation pattern by chance at each marker [19].

Excluding 25 markers from a 150-kb region on chromosome 4, the current set of genetically and physically mapped simple sequence polymorphisms comprises 210 markers that can discriminate well between Columbia and Landsberg *erecta* homozygous plants and against the heterozygote. The average resolution of the linkage map is about 3.5 cM, and the largest gap between markers is approximately 15 cM. Hence, oligonucleotide array based genotyping of a limited number of 30–40 F2 segregating plants that carry the mutant phenotype will allow the mutation to be mapped onto one of the five Arabidopsis chromosomes to a region within a few centimorgans, corresponding to a few hundred thousand base pairs. The *rsf1* (reduced sensitivity to far-red light) mutation, for instance, was mapped recently unequivocally to a 500-kb interval on the top of chromosome 1 using 32 F2 plants [22].

In order to reduce the amount of comparative sequencing between the parent plant and its mutant offspring to identify the induced mutation, fine-mapping is performed to shorten the interval. This is accomplished by designing PCR primers approximately every 10 000 bp to yield amplicons with a size of about 500–600 bp. Screening inter- rather than intragenic regions increases the chances of finding polymorphisms. Denaturing high-performance liquid chromatography (DHPLC) is a particularly attractive tool for fine-mapping as it can be used both for the discovery and the genotyping of polymorphisms. For that purpose, mixtures of corresponding fragments amplified from genomic DNA of the Columbia and Landsberg *erecta* accessions are screened at temperatures recommended by the results of computer simulation of the DNA melting behavior of the amplicons [23]. Fragments found to be polymorphic between the accessions are then used as markers to shorten the interval in F2 plants. The closer one moves to the mutated gene the smaller the number of heterozygous F2 plants will become. Since one recombination event out of 100 products of meiosis is found about every 200 kb in *Arabidopsis* [24], it takes 100 meioses to map the mutant locus within 400 kb or 2 cM between nearest flanking crossovers. In practice, however, nearly twice as many meioses are required to achieve this level of resolution with 90% probability [25]. Consequently, in order to map a mutant to a chromosomal region of about 50 kb, it will take at least 800 meioses or F2 plants exhibiting the phenotype of interest. Indeed, in case of the *rsf1* mutant, fine structure genetic mapping by DHPLC in 692 F2 plants narrowed the chromosomal interval to 55 331 bp [22]. In the past, when sequencing was carried out manually with radioactively labeled reagents, usually thousands rather than hundreds of F2 plants were employed to define the location of the gene in order to minimize the amount of sequencing required. The excellent sensitivity and semi-automated and inexpensive nature of DHPLC, however, make it possible to screen larger intervals reducing effectively the time and labor to clone an induced mutation [26].

For mutational analysis by DHPLC, typically the entire interval between the two markers located closest to the nearest flanking proximal and distal crossover is amplified by PCR in fragments of 500–600 bp with a minimal overlap of 30 bp. Since chemical mutagenesis is carried out in a hemizygous background generated by repeated selfing, successful detection of the mutation requires that amplicons generated from mutant are mixed at equimolar ratio with corresponding amplicons from wild type. Typically, the appearance of one or more additional peaks in one or two chromatographic profiles, if the mutation happens to be located in the overlapping region between two adjacent amplicons, indicates the location of the mutation. Its exact position and chemical nature is eventually established by sequencing.

Successful and unequivocal mapping of a gene does not guarantee easy identification of the mutation causing the phenotype of interest. In one case, neither DHPLC analysis nor conventional sequencing of the entire map interval succeeded in detecting a mutation. Complementation of the mutant phenotype with clones from that interval representing wild type parent confirmed at the functional level that the mutant allele had been mapped correctly. Eventually, it was found that the presence of paralogs had led to faulty annotation of the sequence in that region and, as a consequence, several thousand base pairs had gone unscreened. In this context, it is important to remember that even the genomic sequence of *Saccharomyces cerevisiae*, which was the first eukaryotic organism to be sequenced completely in 1996, is still being annotated. A further possibility, though extremely rare, is the non-mutational imprinting of a gene by ethyl methanesulfonate [27].

Columbia and Landsberg *erecta* have been the preferred ecotypes since the 1960s for elucidating genetic influences on development and physiology. With the increasing interest in the study of naturally occurring quantitative genetic variation, for which

Arabidopsis provides a useful model system due to its widespread occurrence in the moderate temperature zones of the world [28,29], genetic markers will be needed that distinguish not only Columbia and Landsberg *erecta* but also the more than one hundred other known accessions [30]. To date, 221 simple sequence polymorphisms have been evaluated in almost 90 different ecotypes. On average, close to 95% of all accessions could be genotyped successfully for any given polymorphism. Only five of 221 markers distinguished solely Columbia and Landsberg *erecta*. Twenty-eight (12.7%) of the markers were rare with a minor allele frequency ≤10%, while 98 of 221 (44.3%) exhibited a minor allele frequency >20% (Fig. 2). Hence, even markers ascertained in Columbia and Landsberg *erecta* only will prove to be of general utility in marker-trait association studies of natural populations of *Arabidopsis* [24].

2.2. Allele-specific polymerase chain reaction

Variable detector arrays provide an elegant and swift approach to the genotyping of hundreds to thousands of markers in parallel. Most laboratories working with *Arabidopsis*, however, do not have excess to instrumentation necessary for handling DNA arrays. A relatively inexpensive alternative is the use of allele-specific PCR [31]. In its original form, the reaction comprises two allele-specific primers that differ in their 3′ terminal nucleotide and match either of the two alleles of a binary single nucleotide polymorphism [32]. Because mismatched 3′ termini are extended by DNA polymerases with much lower efficiency than correctly matched termini, the allele-specific primer amplifies preferentially the specific allele. However, in most cases a single base mismatch at the 3′ terminus is not sufficient to create reliable discrimination between

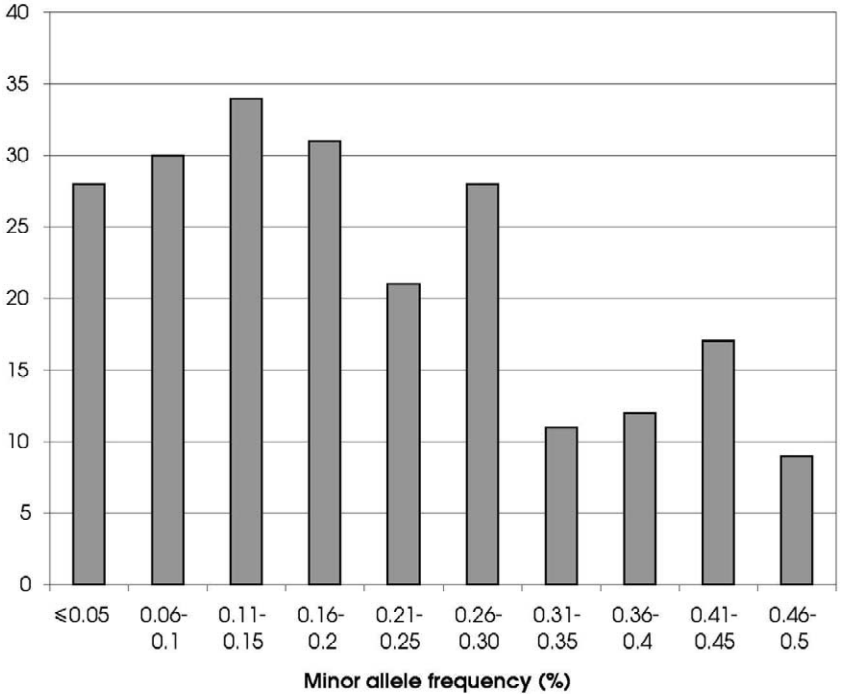

Fig. 2. Distribution of minor allele frequencies of 221 simple sequence polymorphisms ascertained in Columbia and Landsberg *erecta* ecotypes in 86 *Arabidopsis* accessions collected in Africa, Europe, Asia, and America, that had been kindly provided by Justin Berkovitz from the Plant Biology Laboratory at the Salk Institute for Biological Studies in La Jolla, CA, USA.

the two alleles [33,34]. In a modification of the original methodology, an additional base pair change is introduced within the last four bases of the primer [35]. This extra mismatch in addition to the one at the 3′ end produces a dramatic reduction in the PCR product yield of the nonspecific allele but has a relatively minor effect on the amplification of the specific allele. The challenge is to determine the nature and location of the additional mismatch that will yield the required primer specificity. To facilitate the process of primer design, a computer program has been written based on a set of empirical data that evaluates the effect of the addition of different mismatch alternatives on PCR amplification. The program, called SNAPER, generates a list of up to 32 possible primers per SNP site (16 alternatives for each allele) that contain an additional mismatch within the three bases closest to the 3′ end. The program provides information concerning the likelihood that the primer will be allele-specific, predicted by empirical data, and the position and type of base pair change introduced to generate the additional mismatch in the primer.

To validate SNAPER, a total of 331 primer pairs suggested by the program for 43 different single nucleotide polymorphisms were tested using 28 and 38 cycles of PCR amplification to ensure that the primers are specific over a 1000-fold range of template DNA concentrations. The program had an overall success rate of 53% in generating primers with the desired range of specificity, i.e. in 27 of 43 cases specific primer pairs for both alleles of a given SNP could be obtained [31]. For 14 of the 16 cases failed, a specific primer pair was obtained for only one of the alleles, while two SNPs failed to generate any allele-specific primers. Eventually, a second allele-specific primer could be obtained for six out of seven SNPs for which previously only one allele-specific primer had been found. Genotyping of 17 of the SNPs in 94 recombinant inbred lines of *Arabidopsis* produced only in 2.7% ambiguous data.

Eventually, 50 single nucleotide polymorphisms evenly distributed throughout the *Arabidopsis* genome were converted into markers amenable to allelic discrimination by allele-specific PCR. Since the reactions do not employ labeled primers and can be analyzed on an ethidium bromide stained agarose slab gel, allele-specific PCR constitutes—once established—an inexpensive mean to map an induced mutation to within a few million bases on one of the five chromosomes.

2.3. Genotyping of single nucleotide polymorphisms by liquid chromatography–electrospray ionization mass spectrometry

The majority of methods in use for interrogating single nucleotide polymorphisms rely on the ability to hybridize an oligonucleotide with high specificity to a target sequence that either contains or is located immediately adjacent or within a few base pairs of the variant site of interest [13]. Obviously, this cannot be always accomplished and it typically involves extra reagents in addition to those required for amplification of the polymorphic locus.

Liquid chromatography–electrospray ionization mass spectrometry enables the direct interrogation of single nucleotide polymorphisms without further manipulation of the amplicon containing the polymorphic locus [36,37]. The mass accuracy of modern quadrupole ion trap mass analyzers enables the discrimination of nucleic acid sequences that differ in molecular mass by as little as 3–6 Da in a total mass of 31 000 Da (~100 nt). Moreover, their resolving power is sufficient to differentiate two oligodeoxynucleotides having a mass difference of 9 Da (the mass difference between an adenine and a thymine) up to a length of ~75 nt. Liquid chromatography, in this process, serves a dual purpose. Firstly, by fractionating the amplicon from the excess of deoxynucleotides and primers contained in the polymerase chain reaction, it prevents a loss of sensitivity due to the preferential ionization of the latter over the higher molecular mass PCR product. Simultaneously, it suppresses the formation of complexes between DNA and mono- and divalent cations that adversely affect the quality of mass spectra by displacing the latter with triethylammonium ions contained in the mobile phase. Secondly, by carrying out liquid chromatography at elevated column temperatures ≥70 °C, PCR products are readily denatured into their single stranded components, which is essential to discriminate alleles, as in some cases mass changes in one strand of an intact DNA duplex (e.g. A>T) will be neutralized by mass changes in the complementary strand (e.g. T>A). However, changes in mass alone may lack the specificity to discriminate between two alleles at a given site, as an amplicon may harbor the same substitution at a different position or the first base substitution (e.g. A>G) may be neutralized by a second linked substitution (e.g. G>A). Depending on the size of the amplicon, this may be a rare event but clustering of polymorphisms has been observed in both *Arabidopsis* [26] and human [38]. Additional information may be gained from mass spectrometry by selecting a precursor ion from the series of multiply charged ions, which result from the variable number of protons that dissociate from the sugar-phosphate backbone of nucleic acids during electrospray ionization, and fragmenting it subsequently by gas phase collision-induced dissociation. The sequence of the resulting fragments ions is then deduced by means of a computer-based algorithm, which compares the measured spectrum with m/z values predicted from a reference sequence employing established fragmentation pathways [37,39].

2.4. Padlock probes and parallel genotyping directly in genomic DNA

All of the aforementioned approaches to interrogating SNPs require pre-amplification of the polymorphic loci. Although several studies have succeeded in amplifying tens of loci in a multiplex reaction [19,20,40], there are intrinsic limits to the number of loci that can be amplified in a single polymerase chain reaction as the probability of nonspecific amplifications grows approximately in proportion to the square of the number of primer pairs that are combined in a single reaction [41]. The use of padlock probes adds another layer of specificity to allelic discrimination enabling potentially the interrogation of thousands of SNPs without prior amplification of the polymorphic sites using a few nanograms of genomic DNA [42]. A padlock probe is designed to include target-complementary sequences at each end that are joined via a non-target complementary linking segment. The target-complementary sequences are selected such that the two ends are brought immediately next to each other upon hybridization to the target sequence. If perfectly hybridized, the probe ends can then be joined by enzymatic DNA ligation. The accuracy of allelic discrimination is highly dependent on the concentration of sodium chloride in the ligation reaction. At a concentration of 250 m*M* NaCl, ligation rates for T4 DNA ligase were shown to differ >1000-fold between a correct C/G match and a T/G mismatch, while ligation yield for the matched target was decreased insignificantly. Since discrimination is accomplished by allele-specific hybridization between the 3′ end of the linear padlock probe and the SNP of interest, two padlock probes are required to interrogate each locus. Depending on the scheme used to detect nick closure, padlock probes may be as long as 100 nucleotides, which generates significant cost related to their synthesis and purification. By designing the target-complementary sequences so that they hybridize adjacent to the polymorphic locus leaving a gap of one nucleotide to be filled by extension of the 3′ end using a DNA polymerase prior to ligation, only one padlock probe is required for successful interrogation [43]. This approach carries the additional advantage that a polymorphic locus is not called inadvertently homozygous due to degradation or failed addition of one of the padlock probes.

Circularized padlock probes have been traditionally detected by means of rolling-circle amplification [43,44]. For analysis of large sets of polymorphic loci, a highly parallel molecular bar-coding strategy is employed [42]. For that purpose, the linking segment of the padlock probe is designed to contain a forward priming sequence common to all probes, either of two reverse priming sequences specific for the two alleles, and a tag sequence unique for each locus [42]. In this case, allelic discrimination is accomplished in a single ligation reaction. In case of allelic discrimination by a gap-fill reaction, only one padlock probe containing forward and reverse priming sequences common to all probes is required, but filling of the single base gaps has to be carried out in four separate reactions to each of which one of the four deoxynucleotides is added. Following nick closure, only circularized probes serve as templates for PCR amplification. To reduce background from amplification of any unreacted probes, ligation reactions are digested prior to PCR with exonucleases to which circularized probes are insensitive [45]. In the presence of two padlock probes per biallelic locus, an invariant forward and two allele-specific reverse primers that are labeled with either a red or a green fluorescent dye are used to prime amplification. Allelic discrimination is accomplished by hybridizing the amplification products to a high-density oligonucleotide array that carries probes complementary to the unique tag sequences [46]. Homozygotes will yield either a red or a green fluorescence signal, while heterozygotes will appear yellow. Discrimination of alleles following DNA polymerase catalyzed gap-fill reactions is also accomplished by sorting of amplicons on a bar-code array. However, depending on whether a four- or a two-color scanner is available, priming is carried out with either four differently labeled common reverse primers, followed by pooling of the reactions prior to hybridization to a single array. Alternatively, only two differently labeled primers may be used. This allows pooling of only two of the four gap-fill reactions and, hence, genotyping of all four reactions requires hybridization to two bar-code arrays.

With only nanograms of genomic DNA required to genotype several hundred single nucleotide polymorphisms in parallel, genotyping by means of padlock probes and bar-code arrays is the method of choice whenever only small amounts of non-renewable DNA are available and knowledge of individual genotypes is required.

3. Targeted screening for induced mutations

The availability of high-throughput mutation screening methods makes it feasible to identify chemically induced mutations by a reverse genetic strategy in which the mutations are identified first before their affect on phenotype is evaluated. The first to apply this strategy successfully using DHPLC were McCallum et al. [47] who were interested in determining the biological functions of two novel chromomethylases in *Arabidopsis thaliana*. A total of 13 mutations, nine of which were redundant, were detected and confirmed by sequencing in approximately 2 Mb of DNA sequence screened. The same approach was also used to detect EMS-induced mutations in a *Drosophila* gene [48] and ethylnitrosurea (ENU)-induced mutations in four genes in mouse [49].

It is obvious that for this reverse genetic strategy to be applied on a whole genome scale, presently available single-column DHPLC instruments provide insufficient sample throughput. Significant increases in sample throughput can be accomplished by a combination of different means. Firstly, DNA samples from several individual mutants can be pooled [47]. Since heteroduplices absorb more UV light than their corresponding homoduplices, one mutant chromosome out of ten can be detected reliably without significantly compromising sensitivity [47,50,51]. Slightly more chromosomes may be pooled using proofreading DNA polymerases, which eliminate almost completely background heteroduplices in the chromatogram stemming from PCR artifacts [52,53]. Secondly, fragments amplified from different regions of the genome that share similar melting characteristics may be multiplexed by labeling them with different fluorophores [54]. Using an argon ion laser for excitation and measuring fluorescence emission at wavelengths of 505, 540, 575, and 595 nm, spectral resolution of the fluorescent dyes 5-carboxyfluorescein (FAM), hexachlorofluorescein (HEX), 7′-8′-benzo-5′-fluoro-2′-4,7-trichloro-5-carboxyfluorescein (NED), and 6-carboxy-X-rhodamine (ROX), was sufficient to reconstruct the individual chromatograms even when the differently labeled amplicons co-eluted [54,55]. In practice, however, only FAM, HEX and NED are useful tags, because ROX-labeled nucleic acids are retained significantly longer and gradient conditions have to be chosen carefully to avoid co-elution of ROX-labeled primer with PCR fragments labeled with any of the other three fluorescent dyes as the abundant primer signal would mask emission signals stemming from the latter. An important prerequisite for laser induced fluorescence detection was the replacement of 4.6 mm I.D. packed bed columns with styrene–divinylbenzene monoliths polymerized in situ in 0.2 mm I.D. fused-silica capillaries in order to achieve sufficiently high peak concentrations of analytes. The monoliths have been shown to yield separation efficiencies of homo- and heteroduplices identical to those obtained on a micropellicular poly-(styrene–divinylbenzene) stationary phase [56]. These capillaries also enabled the construction of arrays comprising eight [55] and 16 [57] separation channels, respectively. Delivery of the mobile phase is accomplished with a single low-pressure gradient pump that is operated at a fairly high primary flow-rate of 100–200 μl/min to ensure reproducible gradient formation. The actual flow in the individual capillaries, however, is only about 2 μl/min. Reduction of the flow is accomplished by splitting of the primary flow with most of the mobile phase currently going to waste. Hence, it is possible to operate at least 48 capillaries in parallel, but construction of such large HPLC arrays is presently hampered by the commercial unavailability of automated liquid handling and injection systems that allow the simultaneous loading of more than eight columns. Both column temperature and the concentration of acetonitrile required for elution of a DNA fragment control its degree of denaturation. Slight variation in temperature and the ratio of styrene to divinylbenzene during polymerization can result in differences in porosity and, consequently, surface area between different batches of both micropellicular and monolithic stationary phases. Since retention is proportional to the surface area of the adsorbent, the concentration of acetonitrile required to elute a DNA fragment is the greater the smaller the particle size or the pore diameter of the stationary phase. To compensate for such differences and still use only one uniform gradient of acetonitrile, one can adjust the temperature of the individual columns until they all yield similar elution profiles [55]. Generally, an increase in column temperature of 1 °C equals a decrease in acetonitrile concentration of approximately 0.8%. Alternatively, one can assemble only columns with very similar backpressure into the array, thereby, eliminating the need for individual column temperature control [58].

A potentially powerful alternative to DHPLC in the targeted screening of induced mutations is the use of the plant endonuclease CEL I that cuts DNA in two independent incision events, one in each strand, at the phosphodiester bond immediately on the 3′-side of mismatched bases resulting in truncated fragments that can be readily resolved and sized by denaturing gel electrophoresis, hence providing also information about the location of the mutation [59]. Although CEL I has a preference for certain mismatches in the order of C/C≥C/A~C/T≥G/G>A/C~A/A~T/C>T/G~G/T~G/A~A/G>T/T, at least one of the two alternate heteroduplices that are formed by two alleles should be a good substrate for the enzyme, thus, ensuring high mutation detection sensitivity [60]. However, in practice, detection of C to T transitions by CELI has proven far more reliable than that of A to T transversions, which are generated perdominantly in ENU-mutagenesis.

Based on typical *Arabidopsis* and mouse codon usage, only approximately 5 and 10%, respectively, of the mutations induced by EMS and ENU will introduce a stop codon, while approximately two thirds will be missense mutations resulting in the replacement of an amino acid, and the remainder will be silent changes. Hence, assays that detect translation-terminating mutations such as the protein truncation test [61] could be used to target this specific class of mutations. At present, such tests are too expensive and laborious for the screening of thousand of mutants. Moreover, null mutations often result in embryonic lethality, thus, precluding elucidation of secondary post-embryonic phenotypes [49]. Hence, generation of an allelic series that includes missense mutations, particularly, in evolutionary conserved regions offers a much greater chance of gaining insight into the pleiotropic effects of genes and in establishing a relationship between structure and function.

Although reverse genetics promises to accelerate the functional annotation of genomes, it does not come without logistical challenges. The overall rate with which mutations are found after chemical mutagenesis appears to vary, for reasons not understood at present, significantly between organisms. In *Arabidopsis* and *Drosophila*, mutations were recovered approximately every 200 kb [47,48]. In mouse, in contrast, the recovery rate was tenfold lower [49]. If this low recovery rate of mutations were to hold true, it could prove necessary to screen about 10 000 mice to detect a single nonsense or missense mutation with a probability of 95%.

4. Functional genomic analysis by RNA interference

A potent alternative to mutagenesis, at least in invertebrate systems such as the nematode *Caenorhabditis elegans*, is RNA-mediated interference

(RNAi) that enables the specific and transient inhibition of the activity of a gene by direct injection or ingestion of double stranded RNA (dsRNA) that interacts with complementary endogenous messenger RNA transcripts, which are subsequently degraded [62–64]. Feeding worms *Escherichia coli* expressing dsRNA rather than injecting it directly offers the advantage that the interference effect can be titrated to uncover a series of hypomorphic phenotypes informative about the typically multiple functions of a given gene. This is accomplished by varying the concentration of isopropylthiogalactoside in the nematode growth medium, which is necessary to induce the expression of dsRNA in the transformed bacteria contained in the medium [64]. Systematic functional analysis of the *C. elegans* genome by RNA interference has shed light on many biological processes and molecular pathways [65,66]. With approximately one third of the predicted *C. elegans* genes sharing a human ortholog [67], the systematic investigation of loss-of-function phenotypes in *C. elegans* also holds the promise of expanding our knowledge of the basic processes underlying human disease [68]. However, no technique is without its limitations. Of all known embryonic lethal genes on chromosome I of *C. elegans*, RNA silencing was sufficiently effective to detect 90% [65]. However, the success rate of assigning phenotypes to genes with known post-embryonic phenotype was only 45%. In some cases, the phenotype was simply overlooked, but in other instances RNA interference plainly failed to phenocopy the null phenotype, particularly, of genes involved in neuronal function and spermatogenesis.

Although specific dsRNA triggered silencing of gene activity has been reported for oocytes and early embryos of mice [69], zebrafish [70], and *Xenopus* [71], other studies have failed to observe gene-specific RNA interference in different vertebrate systems, demonstrating instead predominantly non-specific effects of dsRNA on gene expression [72,73]. The eventual observation that dsRNA is cleaved to RNA segments 21–23 nucleotides in length before it initiates the degradation of the targeted mRNA [74], led to the development of short interfering RNAs (siRNAs) that are 21–25 nucleotides in length and allow the specific inhibition of gene activity not only in invertebrate organisms but also in cultured somatic cells from both mouse and human [75,76]. The degree of gene silencing obtained with siRNAs can vary significantly from gene to gene and is rarely complete. Hence, the success of silencing has to be validated on an individual basis by using, for instance, a reporter system such as Green Fluorescent Protein (GFP) or Western immunoblotting of gel protein extracts.

5. Functional profiling of yeast genes and the identification of human disease genes

Cloning of single-gene Mendelian traits in human relies on the genotyping of variable number tandem repeats in extended pedigrees. Depending on the number and size of pedigrees available, mutant alleles can be mapped within a few hundred thousand to million base pairs. Intervals of that size are too large for mutation screening and, therefore, one would like to limit the number of candidate genes that may account for the phenotype of interest based on information on their function. A particularly powerful system for determining loss-of-function phenotypes is the nearly complete set (96% of all annotated open reading frames) of gene-disruption mutants in the yeast *Saccharomyces cerevisiae* [77,78]. This resource was constructed by directed gene replacement that led to the generation of almost 6000 individual yeast strains in each of which a different gene is deleted precisely from start to stop codon with a so-called deletion cassette [79]. The cassette contains the kanamycin resistance gene aminoglycoside phosphotransferase, which is flanked by a common priming site, two distinct 20-nucleotide sequences that serve as molecular bar codes to uniquely identify each deletion mutant by means of hybridization to a high-density oligonucleotide array, another common priming site, and finally 30 bases of sequence homologous to the yeast gene to be deleted. Replacement of each gene is accomplished by transformation of the cassette into a haploid yeast strain and homologous recombination that results in the specific replacement of the targeted open reading frame. Successful gene replacement is confirmed by plating transformed deletion strains onto an agar plate containing kanamycin. Information about the biological function of each gene is subsequently inferred by pooling of all yeast deletion strains and simultaneous monitoring of their fitness under a variety of selective growth conditions, such as media that lack essential nutrients [77], contain drugs [78,80], high salt [78], or non-fermentable substrates [81], or exposure to UV light [82]. The function of each gene is uncovered by monitoring the depletion of the corresponding deletion strain under a growth condition that affects its survival. For that purpose, at various points after inoculation, DNA is isolated from aliquots of the culture, followed by amplification of the unique tag sequences to reduce the complexity of the sample, and analysis on the array to determine the relative abundance of the different deletion strains in the pool.

Since proteins that are conserved throughout eukaryotes, including human, carry out most of the core biological functions, systematic analysis of the function of yeast genes can uncover genes that are involved in human disease [83]. Recently, we have demonstrated the use of the collection of single-gene yeast deletion mutants to the identification of nuclear-encoded mitochondrial genes. Genes encoding mitochondrial proteins were identified by pooling 4706 homozygous diploid single-gene deletion strains and monitoring in parallel their growth on both non-fermentable substrates, including glycerol, lactate and ethanol, and on fermentable sugar (glucose). Mutants with respiratory defects have impaired growth on non-fermentable substrates. Of 425 previously known genes encoding proteins involved in mitochondrial function and biogenesis, 353 were observed to grow on glucose. The remaining genes either failed to grow because of the lethal effect of their deletion, or they had failed deletion construction. Fifty-seven percent (201 of 353) of the viable mutants showed defects in growth on non-fermentable substrates, suggesting that about half of all mitochondrial-related proteins are essential for optimal respiratory activity and can therefore be identified by a quantitative growth selection screen. In addition to the 201 proteins with known mitochondrial localization or function in oxidative phosphorylation, the tricarboxylic acid cycle, mitochondrial protein synthesis and transport, ionic homeostasis, and the metabolism of vitamins, cofactors and prosthetic groups, 265 additional mutants showed more or less severe defects in fitness on non-fermentable substrates. Of the 265 genes, 104 encoded proteins that localized outside the mitochondria with functions in vacuolar and ion transport, transcription, and protein targeting, sorting and translocation. For the remaining 161 proteins, the subcellular localization and, in most cases, function was unknown. Fifty-one carried a putative mitochondrial import sequence and in vitro assessment of mitochondrial import using radiolabeled precursor proteins and isolated yeast mitochondria confirmed in five of the six proteins tested that they were indeed imported into the mitochondria in a membrane potential-dependent manner. Overall, the quantitative deletion screen led to a 6.1-fold enrichment of genes encoding known mitochondrial proteins, while independent gene-expression analysis of the diauxic shift from fermentation to respiration yielded an enrichment factor of only 1.2 [84]. This suggests that deletion phenotype is a more specific measure of gene function than is expression level. This was also observed in a recent study on the transcriptional response of *Saccharomyces cerevisiae* to DNA-damaging agents, which failed to identify most of the genes that had been found to be involved in the repair of double-strand breaks, pyrimidine dimers, single-strand breaks, base damage, and DNA cross-links, by means of a systematic screen of 4627 diploid yeast strains with homozygous deletions of nonessential genes [85]. A more informative utilization of genome-wide expression analysis is the generation of transcriptional signature profiles of deletion mutants [86]. Fourteen of 24 deletions of known mitochondrial proteins were clustered into the same group because they had a statistically similar expression profile. Quantitative growth selection of the yeast deletion pool on non-fermentable substrates, in contrast, identified the same 14, as well as a further three [83]. The obvious drawback of signature profiling, however, is that it requires one experiment per deletion strain and thus thousands of arrays to measure all strains, compared to only a single array at each time point for the functional screen.

For 255 of the 466 deletion mutants that had shown growth defects on non-fermentable substrates a human ortholog could be identified. Of these, 21 were genes already known to be involved in mitochondrial disease inherited in a Mendelian fashion. Additionally, eight orthologs were found associated with diseases for which a mitochondrial pathophysiology is plausible but has not been proven. In turn, for 33 of 102 known human genes associated with Mendelian mitochondrial disease, there was no corresponding yeast gene whose deletion was associated with growth defects, although in a few cases the deletion strains showed minor deficiencies. A further 15 of the 102 had not been measured at all, because they had been either lethal or not detectable, and 33 had yielded no yeast orthologs. Although the screen had been by no means comprehensive, the data showed that many human disease genes are associated with quantitative growth defects in yeast. Subsequently, seven mapped, putative mitochondrial disorders, for which affected individuals have either clinical symptoms or biochemical findings indicative of mitochondrial disease, were selected. Of the 255 human orthologs of yeast genes that had shown quantitative growth defects on non-fermentable substrates, 11 could be assigned as candidate genes to the reported chromosomal disease intervals. Ongoing mutational analysis of the candidate genes in index cases of the pedigrees that had been used for mapping has failed thus far to identify disease-causing mutations. There are several potential reasons for this failure. Firstly, due to the high degree of redundancy at the individual gene or pathway level that evolved most likely to buffer phenotypic consequences of genetic variation [87], quantitative growth selection of single-gene mutants will not always result in a detectable phenotype. A comparison of the frequency distribution of fitness for 1147 duplicate genes that had at least one homolog elsewhere in the yeast genome with that of 1275 singleton genes under the different growth conditions tested [81], showed among duplicate genes a significantly higher proportion of genes with a weak or no effect of gene deletion (64.3 vs. 39.5%) and a significantly lower proportion of genes with a lethal effect of deletion (12.4 vs. 29.0%). Furthermore, a high correlation was observed between the sequence similarity of duplicate genes and the likelihood that they will compensate each other's function [88]. Double mutants, which are generated by mating and meiotic recombination of single-gene deletion mutants, can often uncover redundant functions of two genes acting in a single biochemical pathway or within two distinct pathways that functionally compensate for defects in the other. Double mutants have been used successfully to identify genes involved in cell polarity, secretion, DNA repair, and cytoskeletal organization [89]. Secondly, the number of selective growth conditions tested has been by no means sufficient to identify all of the presumed 700–1000 proteins involved in mitochondrial function and biogenesis [90]. Independent of the eventual success of quantitative growth selection in identifying nuclear genes involved in the function and biogenesis of mitochondria, it will have to be complemented with analysis of the mitochondrial proteome to identify not only all proteins that localize to mitochondria but also to elucidate any quantitative changes in protein composition in mitochondrial disease. The latter is the more important as there is a growing body of evidence that changes in mRNA transcript levels may not reflect changes in protein expression due to differences in translation, protein modification or degradation [91]. A particularly powerful technique in this regard is Fourier transform ion cyclotron resonance mass spectrometry, the excellent mass accuracy of which has recently allowed the identification of 2762 (86.7%) of the 3187 proteins predicted from annotation of the *Deinococcus radiodurans* genome [92]. Thirdly, it cannot be excluded that, rather than a point mutation leading to premature termination of translation or aberrant splicing, a large genomic deletion that would have been missed by direct sequencing of PCR amplified exons is the actual cause of disease.

6. Cloning of quantitative trait loci in *Saccharomyces cerevisiae*

In nature, most phenotypic traits are quantitative and characterized by differences in degree rather than in kind. The wide range of phenotypes observed in quantitative traits stands in contrast to single-gene Mendelian traits that are characterized by few discrete phenotypic classes. While traditional genetic mapping using primarily highly polymorphic microsatellite loci has been applied repeatedly with success to the cloning of Mendelian disease alleles in humans that exhibit high penetrance, linkage studies in combination with transmission disequilibrium and case-control analyses have led only rarely to the identification of susceptibility gene variants in complex genetic disorders such as Crohn's disease [93] and asthma [94]. The difficulties in applying genome-wide approaches to quantitative traits are thought to be due to the different contributions of many underlying genes to the phenotype and the ability of different combinations of alleles with low penetrance and environmental factors to produce similar phenotypes [95]. Often disregarded, failure to classify these similar phenotypes correctly can lead to significant loss of power to detect a true linkage, particularly for loci with modest effects [96].

Only recently, 10 years after identification of the map interval, a gene contributing to fruit size in tomato has been cloned successfully [97]. This recent discovery represents the first quantitative trait gene directly identified solely by genetic methods, the only approach applicable to traits with no known developmental or physiological basis. This confirms that despite the genetic control achievable in model organisms, the identification of QTL genes remains difficult. However, it has been the promise of the human genome and other related projects that the availability of complete genome sequences and sophisticated tools to probe the genome will accelerate and increase the resolution of QTL mapping and the eventual cloning of the genetic determinants by marker-trait association [26].

Saccharomyces cerevisiae is a particularly suitable model organism to test the ability of new genome scale technologies and their impact on our knowledge of quantitative traits and gene function. Its genome was the first one of a eukaryote to be completely sequenced, comprising 12 Mb and approximately 6000 genes [98]. This allowed the production of the first high-density oligonucleotide probe array to simultaneously measure the expression of every gene in the entire yeast genome [99]. Subsequently, it was shown that hybridization of genomic DNA to the same array can be used effectively to map genes within a few thousand base pairs [100]. This is made possible by both the high frequency of allelic variation, on average every couple of hundred base pairs, between different yeast strains and the high degree of array coverage (21.8%) of the non-repetitive regions of the yeast genome, with every annotated open reading frame being represented by a minimum of 20 25-base oligonucleotide probes. All probes on the array are perfectly complementary to the genomic sequence of the laboratory strain S96. The presence of allelic variants in yeast strains other than S96 results in decreases in signal intensity at some of the probes relative to signal intensities observed after hybridization with DNA from the laboratory strain. The approach is comparable to mapping by means of differences in DHPLC elution profiles [22], as a change in signal intensity rather than knowledge about the precise location and chemical nature of the mismatch mimics a biallelic marker. The average physical marker spacing over the entire yeast genome is typically 3500 base pairs, corresponding to an average genetic distance of 1.2 centimorgans. The approach does not lend itself readily to the scoring of codominant markers. In yeast, however, diploid crosses are sporulated following meiotic recombination yielding a tetrad of four haploid segregants each of which carries only one of the two parental alleles.

The aforementioned approach of direct allelic variation scanning of the yeast genome was used to map genes that underlie the high temperature growth phenotype of clinical yeast isolates [101]. In infected mice, this phenotype is associated with virulence which makes it the more interesting to study. Crosses between strains that either have or lack the ability to grow at a temperature ≥ 41 °C have suggested that this phenotype is inherited in a non-Mendelian fashion. Typically, only one ninth of all segregants generated from a single cross of two haploid parent strains exhibit the high-temperature growth phenotype. This indicates multiple underlying genetic loci; that is, if each locus were essential for the trait, the high-temperature growth phenotype segregated as though it were conditioned by 3.2 unlinked Mendelian loci ($1/9=1/2^{3.2}$). In practice, however, the trait could be conditioned by a combination of alleles with different contributions that are each non-essential in isolation. This prevents the prediction of an exact number of loci and therefore the Mendelian estimate represents only a minimum number.

As mentioned above, the beauty of using DNA expression arrays for SNP genotyping is that no prior knowledge about the location and nature of sequence variation between two similar genomes is required. In yeast, partial shotgun sequencing of different strains has revealed on average one instance of allelic variation every 160 base pairs. Therefore, in combination with the almost 22% coverage of the non-repetitive regions of the yeast genome provided by the high-density oligonucleotide probe array, it is fairly easy to construct a genetic map containing thousands of closely spaced markers. These markers can then be used to map loci contributing to phenotypic differences between strains. In case of the high-temperature growth phenotype, 3444 biallelic markers were identified from probes with yielded reproducibly decreased signal strength in a clinical isolate relative to hybridization of DNA from a laboratory strain lacking the phenotype. Subsequent hybridization of 19 high-temperature growth segregants derived from a cross of the two strains identified two regions on chromosomes XIV and XVI, respectively, in which clearly more than half of the hybridized segregants inherited their DNA from the clinical isolate.

To confirm the heritability of both intervals and to precisely define the regions, DHPLC was used in the same manner as described for the fine mapping in *Arabidopsis thaliana* to identify 28 and 21 markers in the chromosome XIV and chromosome XVI interval, respectively. The markers were then genotyped in 104 high-temperature growth segregants. The chromosome XVI interval continued to have a low but significant level of association with the high temperature growth phenotype, with 66.7% of the segregants inheriting alleles derived from the clinical isolate. This translated into a relative risk of 2.1, i.e. the increased probability of displaying the high temperature growth phenotype if a strain carries alleles of the clinical rather than the laboratory strain.

A far greater association with the high temperature growth phenotype was identified for the chromosome XIV locus. Of the 104 segregants, 96.2% inherited alleles from the clinical isolate at this locus, yielding a relative risk of 30.6. Interestingly, fine-structure mapping had only succeeded in narrowing the interval from 51.6 to 32 Kb. This was larger than the expected 6 Kb or 2 cM calculated for a single-gene Mendelian trait locus defined with an equal number of meiotic products. This comparison is valid even when taking into consideration the incomplete inheritance of the QTL interval, because the near Mendelian segregation ratio (96%) makes the Mendelian map interval prediction a valid approximation. This was a first indication that this locus may contain more than one gene contributing to the high temperature growth phenotype.

Two lines of evidence further strengthened the association of the chromosome XIV interval with the high temperature growth phenotype. First, random segregants from the same cross displayed random segregation at the locus. Second, analysis of 64 high temperature growth segregants from a second cross between the same laboratory strain and an unrelated high temperature growth clinical isolate showed 87.5% association, confirming that the chromosome XIV interval is a major-effect quantitative trait locus. Therefore, it came as a big surprise that neither comparative expression analysis at 30 and 37 °C, respectively, nor sequence analysis of the six and seven yeast strains, respectively, that have or lack the ability to grow at high temperature, revealed any significant differences in expression or sequence that would have allowed to link one or more of the 15 genes located in the chromosome XIV locus to the high temperature growth phenotype. However, it lent support to the hypothesis that susceptibility alleles are likely to be common, with no allele being necessary or sufficient for expression of a particular phenotype.

To identify the phenotypically relevant allele(s) in the chromosome XIV interval, a new functional assay, called reciprocal-hemizygosity analysis, was developed [101]. Isogenic pairs of strains were constructed in the hybrid background of the clinical and the laboratory strain that differed genetically only in the alleles of one gene. In each strain one allele of one gene was deleted, producing a hemizygous diploid carrying the allele of either the laboratory strain or the clinical strain. Since the two alleles are replaced with different drug resistance markers, the isogenic pairs can be grown in competition before being plated on media containing either drug to count the number of colonies formed. A distinct advantage of reciprocal-hemizygosity analysis is that it also works for essential genes and is insensitive to potentially confounding gene dosage effects. Furthermore, as measurements are made in the hybrid strain background, segregating alleles from the genetic backgrounds of both the clinical and the laboratory strain can be detected in one assay. This turned out to be of particular significance in the dissection of the architecture of the chromosome XIV quantitative trait locus, as not only two alleles from the clinical strain, namely *MKT1* and *RHO2*, were found to confer a high temperature growth advantage, but also the laboratory strain-derived allele of *END3*. This observation explained why the hybrid containing the genetic material of both the clinical and laboratory strain had shown a significantly greater fitness at 41 °C than either strain on its own.

Using a backcrossing strategy to eliminate the genetic contribution of the chromosome XIV locus to high temperature growth has led in the meantime to the identification of three additional loci, all of which were found to contain at least two alleles conferring a high temperature growth advantage. This finding indicates that existing approaches to quantitative traits demands re-evaluation. If closely linked loci of both common and rare variants, as suggested by a recent theoretical study of the evolution of complex disease loci [102], are common, current single-gene-per-locus approaches might have intrinsic deficiencies. Although narrowing an interval in the hope of achieving a map interval that approaches a single point might serve to locate the major contributor, the effects of neighboring genetic factors could be missed. An additional interesting observation from the study of high temperature growth in yeast is that two of the genes at the chromosome XIV locus found to contribute to the phenotype, namely *END3* and *RHO2*, are both cytoskeleton proteins. Hence, one may argue that quantitative trait loci comprising several genes of similar function are more likely to contain more than just one allele conferring susceptibility to a disease [103].

7. Potato blight and the challenge of genotyping SNPs in a polyploid genome

One of the most destructive infectious diseases of potato is late blight and it represents a major threat to potato cultivation worldwide [104]. The cloning of resistance genes to the causative agent, the fungus *Phytophthora infestans*, is therefore of great importance to agriculture. Most crop plants, including potato, are polyploid. Hence, the study of association between polymorphic markers and quantitative resistance to late blight requires not only distinction between homozygous and heterozygous allelic state but also between different allele ratios. In case of tetraploid potato, five allelic ratios are distinguishable: 4:0, 3:1, 2:2, 1:3 and 0:4. This constituted an interesting case to test the quantitative accuracy of pyrosequencing and compare it to that of single nucleotide extension sequencing [105].

7.1. Allele quantitation by pyrosequencing

The principle of pyrosequencing is based on the detection of de novo incorporation of nucleotides [106]. Briefly, the four different nucleotides are added stepwise to the primed DNA template. Following polymerase mediated base incorporation, a proportional amount of pyrophosphate is released and converted to adenosine 5′-triphosphate by ATP-sulfurylase in the presence of adenosine 5′-phosphosulfate (APS). In turn, ATP is used in a luciferase reaction during which a luciferin molecule is oxidized. The ensuing light, which is proportional to the number of nucleotides incorporated, is detected by a charge couple device (CCD) camera. The iterative addition of nucleotides is possible as the excess of nucleoside triphosphates added to the reaction is continuously degraded between each cycle by apyrase into nucleoside diphosphates and, subsequently, nucleoside monophosphates. In contrast to single nucleotide extension sequencing, it is not necessary for the primer to anneal immediately adjacent to the polymorphic nucleotide site. This offers at least in theory more flexibility in primer design and should help to compensate for the low reaction temperature of 28 °C employed in pyrosequencing due to thermal instability of luciferase. Hence, the formation of dimers and hairpins in the sequencing primer as well as in the template itself has to be avoided. In our hands, pyrosequencing failed to generate interpretable results in 15 of 94 (16%) unselected polymorphic sites due to the aforementioned reasons, while the presence of paralogs caused two additional failures [105]. Overall, 77 of 94 (82%) polymorphic sites could be genotyped successfully. This compares favorably to other large-scale genotyping studies of SNPs in human [20,40] and *Arabidopsis thaliana* [19] using either hybridization to a high-density oligonucleotide tiling array [19,20] or single nucleotide extension in combination with hybridization to a generic tag array [40]. In those studies, robust scoring was obtained for 60–80% of the SNPs tested.

The mean relative standard deviation with which the percent distribution of two alleles could be reproduced was 3.6%, with a range of 1.5–9.1%. Observed allele frequencies for the three heterozygous states differed from the expected distribution as much as 10% in absolute values. Nevertheless, accuracy was still sufficient to call all five possible allelic states unambiguously using a single measurement only. For hexa- and octaploid genomes, it would be necessary to carry out triplicate measurements to determine allelic state with certainty. However, applicability of pyrosequencing to polyploid genomes is less determined by precision and accuracy of measurement but rather by sequence context downstream of the polymorphic site. In contrast to single nucleotide extension, which employs dideoxynucleotides that terminate elongation upon incorporation, pyrosequencing uses deoxynucleotides. Consequently, if a base of the same kind follows either allele of a polymorphic site, pyrosequencing will extend the primer not only by one but two bases. In case of tetraploid potato, for instance, in addition to the 0–4 bases at the polymorphic site, four additional bases of the same kind will be incorporated, bringing the totals to 4, 5, 6, 7, and 8 bases, respectively. Given the dynamic assay range of pyrosequencing, which extends approximately over one order of magnitude, it will be become impossible to genotype a single nucleotide polymorphism if two bases of the same kind follow the polymorphic site, as demonstrated recently [105]. It follows that, in a hexaploid genome, only polymorphic loci whose alleles are not followed by bases of the same kind can be genotyped correctly.

7.2. Allele quantitation by single nucleotide extension and completely denaturing HPLC

Single nucleotide extension is one of the most popular methods in use for SNP genotyping [13]. Since the basic method was first described [107], which is based on the annealing of an oligonucleotide primer immediately upstream or downstream from the polymorphic site and its extension by one or more bases in the presence of the appropriate deoxy- (dNTPs) and dideoxynucleoside 5′-triphosphates (ddNTPs), several techniques have been developed for detecting the extension products. These have included among others radiolabeling [108,109], luminous detection [110], colorimetric ELISA [111], fluorescence detection [40,112–115], mass spectrometry [116,117], and high-performance liquid chromatography in combination with UV absorbance [118] or fluorescence detection [57]. Although it is necessary, in contrast to pyrosequencing, for the 3′-terminal base of the extension primer to anneal immediately adjacent to the polymorphic site, sequence context appears to be in practice less of a limitation than in pyrosequencing or differential hybridization, as the higher temperature (60 °C) at which the extension reaction is typically carried out prevents in most instances the formation of hairpins and dimers [105].

A drawback of chromatographic analysis of extension products is the relatively low throughput on conventional single-column HPLC instruments. Throughput is further decreased by the fact, that it has proven necessary under completely denaturing conditions, with analysis temperature ranging from 70 to 80 °C, to wash the column with 25% (v/v) acetonitrile for 90 s between injections rather than only 20 s, as is sufficient under partially denaturing conditions, to maintain long-term performance of the column. Recent progress in the construction of capillary HPLC arrays will lead to significant improvements in throughput that can be further increased by tagging different reactions with different fluorophores [57].

What makes HPLC appealing in the analysis of single-nucleotide extension reactions is the high reproducibility of the quantitative measurements of allele ratios [57,105,119,120]. One study [119] reported mean standard deviations of 0.022–0.031 for seven repeated analyses of three different SNPs [119], while another study observed a mean standard deviation of 0.014 (90% confidence interval, 0.012–0.018) for 10 repeated measurements of reference samples with allele ratios ranging from 0.18 to 9.0 [120]. Consequently, and most importantly for association studies, differences in allele frequencies between pools of cases and controls can be determined with high accuracy. The mean error between observed differences between pools for a total of nine SNPs and true differences based on individual genotype data for every individual in the pools was 0.006, with a maximum error of 0.016 [119]. A more recent study employing single nucleotide extension with fluorescent dye-terminators followed by capillary electrophoresis observed a mean error of 0.01 between true and observed differences between pools of controls and cases for a total of 15 SNPs, with a maximum error of 0.022 [121]. However, estimates of absolute allele frequencies in the pools differed from individual genotyping results on average by 0.018 in absolute values, with a maximum of 0.036, and 0.024, with a maximum of 0.063, using HPLC [119] and capillary electrophoresis [121], respectively. Errors in estimating absolute allele frequencies in pools of potato recombinants were very similar between single nucleotide extension and pyrosequencing [105]. The smallest deviations from true absolute frequencies were observed with conventional dye-terminator sequencing, particularly, when the ratio of peak heights observed on forward and reverse strands were averaged. Since similar discrepancies were observed in determining the absolute allele ratios of 3:1 and 1:3 heterozygotes in individual potato plants, errors in preparing pools can be excluded as a major source of the observed discrepancies between true and measured allele frequencies.

The practical consequences of the magnitude of experimental error in determining differences in allele frequencies between pools of controls and cases on the power of marker-trait association tests are significant [120]. At a power level of 80% and with a significance level of 0.05, the comparison of two equal sized pools of 150 individuals can detect an increase in risk of 1.5 when the two alleles are equally frequent in the control pool. An experimental error of 0.013 in the allele quantification of the pools would necessitate a 20% increase in sample size to 180 individuals in both pools to reach significance. At an experimental error of 0.026, the size of each pool would have to be increased approximately 3-fold to 435 individuals. If one were then to consider additional loss of statistical power due to diagnostic misclassification [96], which would amount to 30% at a 10% rate of erroneous diagnosis, it becomes obvious how important minimization of the error of measuring differences in allele frequencies between pools is in keeping pool sizes and, consequently, cost of recruitment in association studies low.

Incorporation efficiencies of the different dideoxynucleotides are hardly ever equal. Using diploid heterozygotes as controls, observed allele ratios varied from 0.87 to 2.05 instead of the expected value of 1 [119]. Allele ratios even varied for identical SNPs embedded in different sequences. Fortunately, such differences are reproducible for a given SNP and, therefore, can be corrected. However, this requires the availability of a control with equal allele distribution. For diploid genomes, this is easily available, but for polyploid genomes this may be more difficult to obtain or require the synthesis of two synthetic templates that are identical to the genomic sequence in which the SNP is embedded. In pyrosequencing, unequal incorporation of nucleotides has been observed, but its degree is far less than that of single nucleotide extension. Therefore, pyrosequencing allows determination of allele frequencies with confidence, even when a heterozygous control with an allele ratio of 1 is unavailable. It has been suggested that differential PCR amplification of alleles rather than differences in efficiency with which DNA polymerases incorporate ddNTPs account for unequal allele representation [122]. However, this appears unlikely because for several SNPs tested the deviation in allele ratio was identical whether a heterozygous PCR product or two synthetic olignucleotides that represented the two alleles and had been mixed at an equimolar ratio were used as templates in the single nucleotide extension reaction [105].

7.3. Allelic discrimination using a co-spotted single nucleotide extension assay

Multiplexing of single-base extension reactions has proven even more difficult than that of polymerase chain reactions, with the maximum number of SNP genotyping reactions in a single tube typically not exceeding 30 SNPs [40,113,114]. Arrays offer a platform to carry out hundreds of single-base extensions in parallel, though physically separate from each other, at significantly lower cost due to the smaller reaction volume. Originally, oligonucleotides that corresponded to sequences immediately upstream or downstream from the polymorphic loci were covalently linked to an epoxide-activated glass surface via an amino group attached to their 5′ end. Subsequently, single-stranded DNA or RNA templates were hybridized to the arrayed oligonucleotides, each of which acted as a primer for a single nucleotide extension reaction with a DNA polymerase and radioactive or fluorescent ddNTPs [123–125]. The use of four differentially labeled ddNTPs carries the advantage that all possible extension reactions can be carried out on a single array. In an alternative approach offering increased detection efficiency and eliminating the need to generate single-stranded template, two primers per SNP were spotted onto the glass surface [113]. The primers differed at their 3′ end, which was complementary to either of the variant alleles. Following multiplex amplification of the polymorphic sequences of interest with one of the amplimers carrying a 5′ tail of T7 RNA polymerase promoter sequence, the PCR products were added directly to the primer array, along with the reaction mixture, which contained T7 RNA polymerase and rNTPs to generate RNA templates from the amplicons that would bind to the complementary probes on the array as well as reverse transcriptase and dNTPs labeled with the same fluorophore for the actual allele-specific genotyping reaction, which would only proceed if the 3′ end of the probe matched the template. The fluorescent signals from each primer pair were then compared to define genotype. However, aside from the higher cost per genotype, the power of discrimination between genotypes has been observed to be at least an order of magnitude lower using allele-specific hybridization than single-nucleotide extension [124].

High-density oligonucleotide arrays such as those prepared by photolithography [126] are not applicable to solid-phase primer extension, as the chemical synthesis proceeds in the 3′→5′ direction and does not leave the 3′ end free for extension. However, arrays of generic probe sequences [77], which are as different as possible to minimize cross-hybridization yet still retain similar hybridization properties to facilitate simultaneous analysis under standard conditions, have been used successfully to sort multiplex single nucleotide extension reactions carried out in solution [40,114]. Firstly, marker-specific primers are used in multiplex PCR amplifications of up to 30 genomic regions containing SNPs. Secondly, the amplification products are used as templates in single nucleotide extension reactions using bifunctional primers with 3′ complementarity to the specific SNP loci and 5′ complementarity to specific probes on the array. Following extension in the presence of labeled ddNTPs, using a different fluorophore for each of the two SNP alleles, the resulting products are hybridized to the array. Thirdly, genotypes are deduced from the fluorescence intensity ratios of the two colors. Using this approach, over 100 extension reactions obtained by pooling of the multiplexes were analyzed simultaneously with approximately 99% accuracy.

Recently, yet another approach to the use of arrays for SNP genotyping has been described [127]. Firstly, the genomic DNA region spanning the SNP of interest was amplified by PCR using a 5′-amino modified forward primer and 5′ biotin-modified reverse primer. Secondly, the forward strand of the PCR product was separated from the reverse strand using magnetic streptavidin beads. Thirdly, the forward strand was mixed with an extension primer whose 5′ end was amino modified and whose 3′ end was penultimate to the polymorphic site. Fourthly, following in-solution hybridization, which allows detection of smaller quantities of amplified target compared to on-chip hybridization, the primer/template pair was co-spotted in quintuplicate onto a functionalized glass surface, which covalently binds both the primer and the template, allowing very stringent washes to reduce background fluorescence. Fifthly, all samples were extended simultaneously with labeled ddNTPs. Finally, the slides were scanned. The fluorescence intensity signals of the five reactions per SNP were averaged and the genotype deduced. The approach carries the main advantage that co-spotting of primer and template at distinct physical locations eliminates potential cross-interactions between different pairs and ensures specificity of probe/target identity. Interestingly, the rate of successful genotyping reactions was determined primarily by the glass surface used. It varied from 86 to 99% for glass slides from Surmodics (Sunnyvale, CA, USA) and Zyomyx (Hayward, CA, USA), respectively. This was due to the fact that Surmodics slides often exhibited large areas of high background noise probably as a result of non-specific binding of fluorescent dideoxynucleotides. On average, signal-to-noise ratio for the Zyomyx slides was about 90:1, but in regions of high background it was as low as 5:1, which was still sufficient to call alleles reliably. The average signal-to-noise ratio for the Surmodics slides was 50:1, but in some regions of the slide the background exceeded the average signal by a factor of two, causing complete loss of data.

It is obvious that determination of individual genotypes causes significant cost and labor. Therefore, one might wonder whether individual SNP genotyping is really necessary given the high accuracy of determining differences in allele frequencies between pools of cases and controls. However, observations made on the *ADAM33* gene, a putative asthma susceptibility gene, suggest that the association of certain combinations of SNP alleles within a gene with increased susceptibility can be by several orders of magnitude greater than that of individual SNPs [94]. Consequently, one may use pools first to establish differences in allele frequencies for individual SNPs between cases and controls. In order to validate the case-control study, which can generate false positive results due to population admixture, one should then conduct a family-based transmission disequilibrium test using father–mother-affected child trios to evaluate the transmission of the associated marker allele from a heterozygous parent to an affected offspring [128]. In a final step, one should test whether any combinations of SNP alleles are associated more significantly with susceptibility or transmitted preferentially to affected offspring.

8. Conclusions

The release of draft sequences of the human genome and other eukaryotes has accelerated the development of new tools that enable the systematic genome-wide analysis of sequence variation, expression of genes at the transcriptional and translational level, and the elucidation of the primary and secondary functions of genes. Model organisms, such as yeast, worm, fruit fly, *Arabidopsis*, and mouse provide excellent platforms for testing new methods and biological concepts that also benefit our understanding of the molecular basis of human disease. None of the tools in use can be considered perfect, which provides analytical chemists with a continued opportunity to make a significant contribution to our conquest of understanding the molecular determinants of life.

References

[1] N. Risch, K. Merikangas, Science 273 (1996) 1516.
[2] E.S. Lander, Science 274 (1996) 536.
[3] F.S. Collins, M.S. Guyer, A. Chakravarti, Science 287 (1997) 1580.
[4] A.D. Roses, Hum. Mol. Genet. 10 (2001) 2261.
[5] W.J. Strittmatter, A.M. Saunders, D. Schmechel, M. Pericak-Vance, J. Enghild, G.S. Salvesen, A.D. Roses, Proc. Natl. Acad. Sci. USA 90 (1993) 1977.
[6] M. Carrington, M. Dean, M.P. Martin, S.J. O'Brien, Hum. Mol. Genet. 8 (1999) 1939.
[7] D. Altshuler, J.N. Hirschhorn, M. Klannemark, C.M. Lindgren, M.C. Vohl, J. Nemesh, C.R. Lane, S.F. Schaffner, S. Bolk, C. Brewer, T. Tuomi, D. Gaudet, T.J. Hudson, M. Daly, L. Groop, E.S. Lander, Nat. Genet. 26 (2000) 76.
[8] S.J. Laken, G.M. Petersen, S.B. Gruber, C. Oddoux, H. Ostrer, F.M. Giardiello, S.R. Hamilton, H. Hampel, A. Markowitz, D. Klimstra, S. Jhanwar, S. Winawer, K. Offit, M.C. Luce, K.W. Winzeler, B. Vogelstein, Nat. Genet. 17 (1997) 79.
[9] H. Meijers-Heijboer, A. van den Ouweland, J. Klijn, M. Wasielewski, A. de Snoo, R. Oldenburg, A. Hollestelle, M. Houben, E. Crepin, M. van Veghel-Plandsoen, F. Elstrodt, C. van Duijn, C. Bartels, C. Meijers, M. Schutte, L. McGuffog, D. Thompson, D. Easton, N. Sodha, S. Seal, R. Barfoot, J. Mangion, J. Chang-Claude, D. Eccles, R. Eeles, D.G. Evans, R. Houlston, V. Murday, S. Narod, T. Peretz, J. Peto, C. Phelan, H.X. Zhang, C. Szabo, P. Devilee, D. Goldgar, P.A. Futreal, K.L. Nathanson, B. Weber, N. Rahman, M.R. Stratton, Nat. Genet. 31 (2002) 55.
[10] D. Altshuler, V.J. Pollara, C.R. Cowles, W.J. Van Etten, J. Baldwin, L. Linton, E.S. Lander, Nature 407 (2000) 513.
[11] P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, P.Y. Kwok, Genome Res. 8 (1998) 748.
[12] V.N. Kristensen, D. Kelefiotis, T. Kristensen, A.L. Borresen-Dale, BioTechniques 30 (2001) 318.
[13] I.G. Gut, Hum. Mutat. 17 (2001) 475.
[14] R.A. Ophoff, G.M. Terwindt, M.N. Vergouwe, R. van Eijk, P.J. Oefner, S.M.G. Hoffmann, J.E. Lamerdin, H.W. Mohrenweiser, D.E. Bulman, M. Ferrari, J. Haan, D. Lindhout, G.J.B. van Ommen, M.H. Hofker, M.D. Ferrari, R.R. Frants, Cell 87 (1996) 543.
[15] O. Zhuchenko, J. Bailey, P. Bonnen, T. Ashizawa, D.W. Stockton, C. Amos, W.B. Dobyns, S.H. Subramony, H.Y. Zoghbi, C.C. Lee, Nat. Genet. 15 (1997) 62.
[16] D. Botstein, R.L. White, M. Skolnick, R.W. Davis, Am. J. Hum. Genet. 32 (1980) 314.
[17] S. Weining, P. Langridge, Theor. Appl. Genet. 82 (1991) 209.

[18] C.M. Thomas, P. Vos, M. Zabeau, D.A. Jones, K.A. Norcott, B.P. Chadwick, J.D.G. Jones, Plant J. 8 (1995) 785.

[19] R.J. Cho, M. Mindrinos, D.R. Richards, R.J. Sapolsky, M. Anderson, E. Drenkard, J. Dewdney, T.L. Reuber, M. Stammers, N. Federspiel, A. Theologis, W.H. Yang, E. Hubbell, M. Au, E.Y. Chung, D. Lashkari, B. Lemieux, C. Dean, R.J. Lipshutz, F.M. Ausubel, R.W. Davis, P.J. Oefner, Nat. Genet. 23 (1999) 203.

[20] D.G. Wang, J.B. Fan, C.J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M.S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T.J. Hudson, R. Lipshutz, M. Chee, E.S. Lander, Science 280 (1998) 1077.

[21] R.J. Sapolsky, L. Hsie, A. Berno, G. Ghandour, M. Mittmann, J.B. Fan, Genet. Anal. 14 (1999) 187.

[22] J.I. Spiegelman, M.N. Mindrinos, C. Fankhauser, D. Richards, J. Lutes, J. Chory, P.J. Oefner, Plant Cell 12 (2000) 2485.

[23] A.C. Jones, J. Austin, N. Hansen, B. Hoogendoorn, P.J. Oefner, J.P. Cheadle, M.C. O'Donovan, Clin. Chem. 45 (1999) 1133.

[24] M. Nordborg, J.O. Borevitz, J. Bergelson, C.C. Berry, J. Chory, J. Hagenblad, M. Kreitman, J.N. Maloof, T. Noyes, P.J. Oefner, E.A. Stahl, D. Weigel, Nat. Genet. 30 (2002) 190.

[25] M. Boehnke, Am. J. Hum. Genet. 55 (1994) 379.

[26] J.I. Spiegelman, M.N. Mindrinos, P.J. Oefner, Biotechniques 29 (2000) 1084.

[27] M. Katoh, N.L. Cacheiro, C.V. Cornett, K.T. Cain, J.C. Rutledge, W.M. Generoso, Mutat. Res. 210 (1989) 337.

[28] T. Mitchell-Olds, Trends Ecol. Evol. 10 (1995) 324.

[29] C. Alonso-Blanco, M. Koornneef, Trends Plant Sci. 5 (2000) 22.

[30] R.A. Price, I.A. Al-Shehbaz, J.D. Palmer, in: E. Meyerowitz, C. Somerville (Eds.), Arabidopsis, Cold Spring Harbor Press, Cold Spring Harbor, NY, 1994, p. 7.

[31] E. Drenkard, B.G. Richter, S. Rozen, L.M. Stutius, N.A. Angell, M. Mindrinos, R.J. Cho, P.J. Oefner, R.W. Davis, F.M. Ausubel, Plant Physiol. 124 (2000) 1483.

[32] L. Ugozzoli, R.B. Wallace, Allele-specific polymerase chain reaction, Methods Enzymol. 2 (1991) 42.

[33] R.S. Cha, H. Zarbl, P. Keohavong, W.G. Thilly, PCR Methods Appl. 2 (1992) 14.

[34] S. Kwok, S.Y. Chang, J.J. Sninsky, A. Wang, PCR Methods Appl. 3 (1994) S39.

[35] C.R. Newton, A. Graham, L.E. Heptinstall, S.J. Powell, C. Summers, N. Kalsheker, J.C. Smith, A.F. Markham, Nucleic Acids Res. 17 (1989) 2503.

[36] H. Oberacher, P.J. Oefner, W. Parson, C.G. Huber, Angew. Chem. Int. Ed. 40 (2001) 3828.

[37] H. Oberacher, P.J. Oefner, G. Hölzl, A. Premstaller, K. Davis, C.G. Huber, Nucleic Acids Res. 30 (2002) e67.

[38] L. Jin, P.A. Underhill, V. Doctor, R.W. Davis, P. Shen, L.L. Cavalli-Sforza, P.J. Oefner, Proc. Natl. Acad. Sci. USA 96 (1999) 3796.

[39] H. Oberacher, B. Wellenzohn, C.G. Huber, Anal. Chem. 74 (2001) 211.

[40] J.B. Fan, X. Chen, M.K. Halushka, A. Berno, X. Huang, T. Ryder, R.J. Lipshutz, D.J. Lockhart, A. Chakravarti, Genome Res. 10 (2000) 853.

[41] U. Landegren, M. Nilsson, Ann. Med. 29 (1997) 585.

[42] M. Nilsson, J. Banér, M. Mendel-Hartvig, F. Dahl, D.O. Antson, M. Gullberg, U. Landegren, Hum. Mutat. 19 (2002) 410.

[43] P.M. Lizardi, X. Huang, Z. Zhu, P. Bray-Ward, D.C. Thomas, D.C. Ward, Nat. Genet. 19 (1998) 225.

[44] J. Banér, M. Nilsson, M. Mendel-Hartvig, U. Landegren, Nucleic Acids Res. 26 (1998) 5073.

[45] M. Nilsson, H. Malmgren, M. Samiotaki, M. Kwiatkowski, B.P. Chowdhary, U. Landegren, Science 265 (1994) 2085.

[46] D.D. Shoemaker, D.A. Lashkari, D. Morris, M. Mittmann, R.W. Davis, Nat. Genet. 14 (1996) 450.

[47] C.M. McCallum, L. Comai, E.A. Greene, S. Henikoff, Nat. Biotechnol. 18 (2000) 455.

[48] A. Bentley, B. MacLennan, J. Calvo, C.R. Dearolf, Genetics 156 (2000) 1169.

[49] E.L. Coghill, A. Hugill, N. Parkinson, C. Davison, P. Glenister, S. Clements, J. Hunter, R.D. Cox, S.D. Brown, Nat. Genet. 30 (2002) 255.

[50] W. Liu, D.I. Smith, K.J. Rechtzigel, S.N. Thibodeau, C.D. James, Nucleic Acids Res. 26 (1998) 1396.

[51] A.C. Jones, J.R. Sampson, J.P. Cheadle, Hum. Mutat. 17 (2001) 233.

[52] D.T. Gjerde, C.P. Hanna, D. Hornby (Eds.), DNA Chromatography, Wiley–VCH, Weinheim, 2002, p. 91.

[53] D. Muhr, T. Wagner, P.J. Oefner, J. Chromatogr. B 782 (2002) 105.

[54] W. Xiao, D. Stern, M. Jain, C. Huber, P.J. Oefner, BioTechniques 30 (2001) 1332.

[55] A. Premstaller, W. Xiao, H. Oberacher, M. O'Keefe, D. Stern, T. Willis, C.G. Huber, P.J. Oefner, Genome Res. 11 (2001) 1944.

[56] C.G. Huber, A. Premstaller, H. Oberacher, W. Xiao, G.K. Bonn, P.J. Oefner, J. Biochem. Biophys. Methods 47 (2001) 5.

[57] A. Premstaller, H. Oberacher, A. Rickert, C.G. Huber, P.J. Oefner, Genomics 79 (2002) 793.

[58] A. Premstaller, H. Oberacher, C.G. Huber, P.J. Oefner, Anal. Chem. (2002) in press.

[59] C.M. McCallum, L. Comai, E.A. Greene, S. Henikoff, Plant Physiol. 123 (2000) 439.

[60] C.A. Oleykowski, C.R. Bronson Mullins, A.K. Godwin, A.T. Yeung, Nucleic Acids Res. 26 (1998) 4597.

[61] P.A.M. Roest, R.G. Roberts, S. Sugino, G.J.B. van Ommen, J.T. den Dunnen, Hum. Mol. Genet. 2 (1993) 1719.

[62] A. Fire, D. Albertson, S.W. Harrison, D.G. Moerman, Development 113 (1991) 503.

[63] A. Fire, S. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, C.C. Mello, Nature 391 (1998) 806.

[64] R.S. Kamath, M. Martinez-Campos, P. Zipperlen, A.G. Fraser, J. Ahringer, Genome Biol. 2 (2001) 0002.1.

[65] A.G. Fraser, R.S. Kamath, P. Zipperlen, M. Martinez-Campos, M. Sohrmann, J. Ahringer, Nature 408 (2000) 325.

[66] P. Gönczy, C. Echeverri, K. Oegema, A. Coulson, S.J. Jones, R.R. Copley, J. Duperon, J. Oegema, M. Brehm, E. Cassin,

E. Hannak, M. Kirkham, S. Pichler, K. Flohrs, A. Goessen, S. Leidel, A.M. Alleaume, C. Martin, N. Ozlu, P. Bork, A.A. Hyman, Nature 408 (2000) 331.

[67] G.M. Rubin, M.D. Yandell, J.R. Wortman, G.L. Gabor Miklos, C.R. Nelson, I.K. Hariharan, M.E. Fortini, P.W. Li, R. Apweiler, W. Fleischmann, J.M. Cherry, S. Henikoff, M.P. Skupski, S. Misra, M. Ashburner, E. Birney, M.S. Boguski, T. Brody, P. Brokstein, S.E. Celniker, S.A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R.F. Galle, W.M. Gelbart, R.A. George, L.S. Goldstein, F. Gong, P. Guan, N.L. Harris, B.A. Hay, R.A. Hoskins, J. Li, Z. Li, R.O. Hynes, S.J. Jones, P.M. Kuehl, B. Lemaitre, J.T. Littleton, D.K. Morrison, C. Mungall, P.H. O'Farrell, O.K. Pickeral, C. Shue, L.B. Vosshall, J. Zhang, Q. Zhao, X.H. Zheng, S. Lewis, Science 287 (2000) 2204.

[68] J. Ahringer, Curr. Opin. Genet. Dev. 7 (1997) 410.

[69] F. Wianny, M. Zernicka-Goetz, Nat. Cell Biol. 2 (2000) 70.

[70] A. Wargelius, S. Ellingsen, A. Fjose, Biochem. Biophys. Res. Commun. 263 (1999) 156.

[71] H. Nakano, S. Amemiya, K. Shiokawa, M. Taira, Biochem. Biophys. Res. Commun. 274 (2000) 434.

[72] A.C. Oates, A.E. Bruce, R.K. Ho, Dev. Biol. 224 (2000) 20.

[73] Z. Zhao, Y. Cao, M. Li, A. Meng, Dev. Biol. 229 (2001) 215.

[74] P.D. Zamore, T. Tuschl, P.A. Sharp, D.P. Bartel, Cell 101 (2000) 25.

[75] N.J. Caplen, S. Parrish, F. Imani, A. Fire, R.A. Morgan, Proc. Natl. Acad. Sci. USA 98 (2001) 9742.

[76] M. Abdelrahim, I. Samudio, R. Smith, R. Burghardt, S. Safe, J. Biol. Chem. (2002) in press.

[77] D.D. Shoemaker, D.A. Lashkari, D. Morris, M. Mittmann, R.W. Davis, Nat. Genet. 14 (1996) 450.

[78] G. Giaever, A.M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, A.P. Arkin, A. Astromoff, M. El Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.D. Entian, P. Flaherty, F. Foury, D.J. Garfinkel, M. Gerstein, D. Gotte, U. Guldener, J.H. Hegemann, S. Hempel, Z. Herman, D.F. Jaramillo, D.E. Kelly, S.L. Kelly, P. Kotter, D. LaBonte, D.C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S.L. Ooi, J.L. Revuelta, C.J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D.D. Shoemaker, S. Sookhai-Mahadeo, R.K. Storms, J.N. Strathern, G. Valle, M. Voet, G. Volckaert, C.Y. Wang, T.R. Ward, J. Wilhelmy, E.A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J.D. Boeke, M. Snyder, P. Philippsen, R.W. Davis, M. Johnston, Nature 418 (2002) 387.

[79] A. Wach, A. Brachat, R. Pohlmann, P. Philippsen, Yeast 10 (1994) 1793.

[80] G. Giaever, D.D. Shoemaker, T.W. Jones, H. Liang, E.A. Winzeler, A. Astromoff, R.W. Davis, Nat. Genet. 21 (1999) 278.

[81] L.M. Steinmetz, C. Scharfe, A.M. Deutschbauer, D. Mokranjac, Z.S. Herman, T. Jones, A.M. Chu, G. Giaever, H. Prokisch, P.J. Oefner, R.W. Davis, Nat. Genet. (2002) in press.

[82] G.W. Birrell, G. Giaever, A.M. Chu, R.W. Davis, J.M. Brown, Proc. Natl. Acad. Sci. USA 98 (2001) 12608.

[83] S.A. Cherwitz, L. Aravind, G. Sherlock, C.A. Ball, E.V. Koonin, S.S. Dwight, M.A. Harris, K. Dolinski, S. Mohr, T. Smith, S. Weng, J.M. Cherry, D. Botstein, Science 282 (1998) 2022.

[84] J.L. DeRisi, V.R. Iyer, P.O. Brown, Science 278 (1997) 680.

[85] G.W. Birrell, J.A. Brown, H.I. Wu, G. Giaever, A.M. Chu, R.W. Davis, J.M. Brown, Proc. Natl. Acad. Sci. USA 99 (2002) 8778.

[86] T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, S.H. Friend, Cell 102 (2000) 109.

[87] J.L. Hartman, B. Garvik, L. Hartwell, Science 291 (2001) 1001.

[88] Z. Gu, L.M. Steinmetz, X. Gu, C. Scharfe, R.W. Davis, W.H. Li, Nature (2002) submitted.

[89] A.H. Tong, M. Evangelista, A.B. Parsons, H. Xu, G.D. Bader, N. Page, M. Robinson, S. Raghibizadeh, C.W. Hogue, H. Bussey, B. Andrews, M. Tyers, C. Boone, Science 294 (2001) 2364.

[90] D.C. Wallace, Science 283 (1999) 1482.

[91] T. Ideker, V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, J.K. Eng, R. Bumgarner, D.R. Goodlett, R. Aebersold, L. Hood, Science 292 (2001) 929.

[92] M.S. Lipton, L. Pasa-Tolic, G.A. Anderson, D.J. Anderson, D.L. Auberry, J.R. Battista, M.J. Daly, J. Fredrickson, K.K. Hixson, H. Kostandarithes, T.P. Conrads, C.D. Masselon, L.M. Markille, R.J. Moore, M.F. Romine, Y. Shen, N. Tolic, H.R. Udseth, T.D. Veenstra, A. Venkateswaran, K.K. Wong, R. Zhao, R.D. Smith, Proc. Natl. Acad. Sci. USA 99 (2002) 11049.

[93] Y. Ogura, D.K. Bonen, N. Inohara, D.L. Nicolae, F.F. Chen, R. Ramos, H. Britton, T. Moran, R. Karaliuskas, R.H. Duerr, J.P. Achkar, S.R. Brant, T.M. Bayless, B.S. Kirschner, S.B. Hanauer, G. Nunez, J.H. Cho, Nature 411 (2001) 603.

[94] P. Van Eerdewegh, R.D. Little, J. Dupuis, R.G. Del Mastro, K. Falls, J. Simon, D. Torrey, S. Pandit, J. McKenny, K. Braunschweiger, A. Walsh, Z. Liu, B. Hayward, C. Folz, S.P. Manning, A. Bawa, L. Saracino, M. Thackston, Y. Benchekroun, N. Capparell, M. Wang, R. Adair, Y. Feng, J. Dubois, M.G. FitzGerald, H. Huang, R. Gibson, K.M. Allen, A. Pedan, M.R. Danzig, S.P. Umland, R.W. Egan, F.M. Cuss, S. Rorke, J.B. Clough, J.W. Holloway, S.T. Holgate, T.P. Keith, Nature 418 (2002) 426.

[95] E.S. Lander, N.J. Schork, Science 265 (1994) 2037.

[96] M.S. Silverberg, M.J. Daly, D.N. Moskovitz, J.D. Rioux, R.S. McLeod, Z. Cohen, G.R. Greenberg, T.J. Hudson, K.A. Siminovitch, A.H. Steinhart, Gut 49 (2001) 773.

[97] A. Frary, T.C. Nesbitt, S. Grandillo, E. Knaap, B. Cong, J. Liu, J. Meller, R. Elber, K.B. Alpert, S.D. Tanksley, Science 289 (2000) 85.

[98] A. Goffeau, B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston, E.J. Louis, H.W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, S.G. Oliver, Science 274 (1996) 546.

[99] L. Wodicka, H. Dong, M. Mittmann, M.H. Ho, D.J. Lockhart,

Nat. Biotechnol. 15 (1997) 1359.

[100] E.A. Winzeler, D.R. Richards, A.R. Conway, A.L. Goldstein, S. Kalman, M.J. McCullough, J.H. McCusker, D.A. Stevens, L. Wodicka, D.J. Lockhart, R.W. Davis, Science 281 (1998) 1194.

[101] L.M. Steinmetz, H. Sinha, D.R. Richards, J.I. Spiegelman, P.J. Oefner, J.H. McCusker, R.W. Davis, Nature 416 (2002) 326.

[102] J.K. Pritchard, Am. J. Hum. Genet. 69 (2001) 124.

[103] J.D. Rioux, M.J. Daly, M.S. Silverberg, K. Lindblad, H. Steinhart, Z. Cohen, T. Delmonte, K. Kocher, K. Miller, S. Guschwan, E.J. Kulbokas, S. O’Leary, E. Winchester, K. Dewar, T. Green, V. Stone, C. Chow, A. Cohen, D. Langelier, G. Lapointe, D. Gaudet, J. Faith, N. Branco, S.B. Bull, R.S. McLeod, A.M. Griffiths, A. Bitton, G.R. Greenberg, E.S. Lander, K.A. Siminovitch, T.J. Hudson, Nat. Genet. 29 (2001) 223.

[104] E.F. Fry, S.B. Goodwin, Plant Dis. 81 (1997) 1349.

[105] A.M. Rickert, A. Premstaller, C. Gebhardt, P.J. Oefner, BioTechniques 32 (2002) 592.

[106] M. Ronaghi, M. Uhlén, P. Nyrén, Science 281 (1998) 363.

[107] M.H. Kuppuswamy, J.W. Hoffman, C.K. Kasper, S.G. Spitzer, S.L. Groce, S.P. Bajaj, Proc. Natl. Acad. Sci. USA 88 (1991) 1143.

[108] A. Jalanko, J. Kere, E. Savilathi, M. Schwartz, A. Syvanen, M. Ranki, H. Soderlund, Clin. Chem. 38 (1992) 39.

[109] A. Syvänen, A. Sajantila, M. Lukka, Am. J. Hum. Genet. 52 (1993) 46.

[110] P. Nyrén, B. Pettersson, M. Uhlén, Anal. Biochem. 208 (1993) 171.

[111] T.T. Nikiforov, R.B. Rendle, P. Goelet, Y.H. Rogers, M.L. Kotewicz, S. Anderson, G.L. Trainor, M.R. Knapp, Nucleic Acids Res. 22 (1994) 4167.

[112] T. Pastinen, J. Partanen, A. Syvänen, Clin. Chem. 42 (1996) 1391.

[113] T. Pastinen, M. Raitio, K. Lindroos, P. Tainola, L. Peltonen, A.C. Syvänen, Genome Res. 20 (2000) 1031.

[114] J.N. Hirschhorn, P. Sklar, K. Lindblad-Toh, Y.M. Lim, M. Ruiz-Gutierrez, S. Bolk, B. Langhorst, S. Schaffner, E. Winchester, E.S. Lander, Proc. Natl. Acad. Sci. USA 97 (2000) 12164.

[115] X. Chen, L. Levine, P.Y. Kwok, Genome Res. 9 (1999) 492.

[116] D.P. Little, A. Braun, B. Darnhofer-Demar, A. Frilling, Y. Li, R.T. McIver, H. Köster, J. Mol. Med. 75 (1997) 745.

[117] L.A. Haff, I.P. Smirnov, Genome Res. 7 (1997) 378.

[118] B. Hoogendoorn, M.J. Owen, P.J. Oefner, N. Williams, J. Austin, M.C. O’Donovan, Hum. Genet. 104 (1999) 89.

[119] B. Hoogendoorn, N. Norton, G. Kirov, N. Williams, M.L. Hamshere, G. Spurlock, J. Austin, M.K. Stephens, P.R. Buckland, M.J. Owen, M.C. O’Donovan, Hum. Genet. 107 (2000) 488.

[120] M. Giordano, M. Mellai, B. Hoogendoorn, P. Momigliano-Richiardi, J. Biochem. Biophys. Methods 11 (2001) 101.

[121] N. Norton, N.M. Williams, H.J. Williams, G. Spurlock, G. Kirov, D.W. Morris, B. Hoogendoorn, M.J. Owen, M.C. O’Donovan, Hum. Genet. 110 (2002) 471.

[122] Q. Liu, E.C. Thorland, S.S. Sommer, BioTechniques 22 (1997) 292.

[123] J.M. Shumaker, A. Metspalu, C.T. Caskey, Hum. Mutat. 7 (1996) 346.

[124] T. Pastinen, A. Kurg, A. Metspalu, L. Peltonen, A.C. Syvänen, Genome Res. 7 (1997) 606.

[125] G. Tully, K.M. Sullivan, P. Nixon, R.E. Stones, P. Gill, Genomics 34 (1996) 107.

[126] S.P.A. Fodor, J.L. Read, M.C. Pirrung, L. Stryer, A.T. Lu, D. Solas, Science 251 (1991) 767.

[127] M. Jain, Y.R. Thorstenson, D. Faulkner, N. Pourmand, T. Jones, M. Au, P.J. Oefner, K.P. White, R.W. Davis, Hum. Mutat. (2002) submitted.

[128] R.S. Spielman, R.E. McGinnis, W.J. Ewens, Am. J. Hum. Genet. 52 (1993) 506.