Available online at www.sciencedirect.com

SCIENCE DIRECT®

International Journal of Mass Spectrometry 238 (2004) 173–188

International Journal of Mass Spectrometry

www.elsevier.com/locate/ijms

MALDI-TOF MS: a platform technology for genetic discovery

Dirk van den Boom*, Martin Beaulieu, Paul Oeth, Rich Roth, Christiane Honisch, Matthew R. Nelson, Christian Jurinke, Charles Cantor

SEQUENOM, Inc., 3595 John Hopkins Court, San Diego, CA 92121, USA

Received 6 February 2004; accepted 17 March 2004
Available online 1 October 2004

Abstract

Matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) mass spectrometry (MS) has been applied as a high-throughput platform technology for qualitative and quantitative nucleic acid analysis in the genetic discovery of target genes and their biological validation. Mass spectrometric methods for the elucidation of genetic variability and for subsequent large-scale genotyping of genetic markers are exemplified. The use of quantitative MALDI-TOF MS is described for large-scale validation of SNP markers and their analysis in DNA sample pools. Initial results of genome-wide association studies employing this technology are provided exemplifying a genetics-driven approach to drug discovery.

Keywords: MALDI-TOF MS; Genetic discovery; Genotyping; Whole genome association study

1. Introduction

Sequencing of several eukaryotic and prokaryotic organisms has created new opportunities for a systematic study of genotype–phenotype correlation and thus the elucidation of the genetic pathways that contribute to the physiology of an organism. The availability of an almost complete human genome sequence enables us to study the genes involved in human health and disease etiology [1]. The three billion base pairs of the human genome are assumed to encode for 30,000–40,000 genes. Only 5% of the whole genome (excluding introns) codes for proteins and current approaches in functional and comparative genomics may help to explore the function of the remainder. The availability of so much genomic information and the growing availability of high-throughput genomic technologies are providing an increasing range of methods for identifying genes that are important for human health.

The knowledge of the reference human sequence enables predictions on the impact of sequence changes (in the form of polymorphisms or mutations). Non-synonymous sequence changes alter the amino acid sequence (the primary structure) of a protein, and often have a significant effect on protein structure and function. The impact of non-coding sequence changes are less obvious. They might alter structural features of DNA packaging in the nucleus, thereby influencing expression of genes or whole gene regions. They might reside in promotor regions or other regulatory regions of genes, influencing their expression in a developmental or tissue-specific manner or in response to environmental factors. They can influence RNA stability and translation efficiencies. Numerous scenarios can be constructed for the influence of non-coding sequence changes. The mechanistic contribution of genetic variability in non-coding regions to the phenotype in general and to disease in particular, is thus more difficult to assess.

An alternative approach to identifying genes involved in human disease is through the changes in gene expression observed under various healthy and diseased conditions. Gene expression can be monitored at two key levels. One is at the level of RNA expression, commonly carried out using purified messenger RNA (mRNA). Highly parallel hybridization-based technologies of mRNA profiling have made this approach fairly common in genomics research, and

* Corresponding author. Tel.: +1 858 202 9066; fax: +1 858 202 9084.
E-mail address: dvandenboom@sequenom.com (D. van den Boom).

1387-3806/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijms.2004.03.015

have been successfully applied to identify genes in diseases pathways [2–4]. Gene expression can also be studied at the protein level [5–7]. The study of proteins may be preferable over mRNA due to the closer connection between proteins and pathophysiology. However, protein expression levels cannot currently be easily measured on a genomic scale and therefore more targeted applications have been developed [8,9]. The lack of amplification methods similar to those available for the study of nucleic acids increases the difficulty in analysis of low-abundant proteins and their modifications.

Searching for genetic variations associated with disease susceptibility has long been used to identify genes influencing human health. To date, most human diseases that have been genetically characterized are monogenic disorders [10]. In such disorders, mutations in one gene have a major effect so that the mutation appeared to cause the disease. Very often positional cloning could be used to localize and identify the disease gene. However, it can be expected that most common diseases (such as diabetes or heart disease) are complex and multifactorial in nature: alleles of multiple genes interact with one another and with multiple environmental exposures over time resulting in the disease phenotype [11]. If we consider the current estimate of 30,000 genes and the millions of variations in them, the identification of those influencing human health and the elucidation of the interaction can be a daunting task. Genetics can be used to pinpoint potential targets of proven relevance in humans, which can then be further validated by techniques such as gene expression profiling and protein analysis [12].

The characterization of genetic factors contributing to disease requires the large-scale analysis of genomic DNA sequence in large numbers of individuals from various populations. Since complete genome sequencing of thousands of individuals is currently not feasible, researchers focus on genotyping of single nucleotide polymorphisms (SNPs) as surrogate markers, to characterize genetic variability. The emergence of SNPs as the most useful marker resulted from the development of several high-throughput technologies for SNP analysis. To date over four million SNPs have been deposited in public databases and the allele-frequency of hundred thousands of them has been determined in at least four ethnic groups (Caucasians, Afro-Americans, Hispanics and Asians). This expanding resource is now used to carry out whole genome scans for disease genes.

Within the last decade, mass spectrometry has become an invaluable tool for the analysis and characterization of biomolecules. Matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) mass spectrometry has developed into one of the leading technologies for accurate large-scale genotyping of SNPs. The use of MALDI-TOF MS has recently been expanded to include relative and absolute transcript quantification and re-sequencing of larger target regions for discovery of new polymorphisms or identification of causal variants in target genes.

SEQUENOM has integrated these applications of MALDI-TOF MS into a top–down approach for disease gene identification by genome-wide association studies. This article provides a review of MALDI-TOF MS applications in nucleic acid analysis and how they can be merged into a concept for large-scale genetic discovery of target genes for drug development.

2. SNP typing using MALDI-TOF MS

In April 2003, 50 years after the discovery of the DNA double helix and after a decade of sequencing efforts, the Human Genome Project (HGP) announced completion of the 3.2 billion bases sequence constituting the human genetic code. With this reference sequence established, it is possible to extract medical and biological value by studying inter-individual variations. However, genome-wide individual re-sequencing is cost and time prohibitive. One of the most important findings of the HGP was the discovery of inter-individual sequence variation; the most prevalent type is called single nucleotide polymorphism (SNPs). The availability of a SNP map now opens the possibility to conduct genome-wide genetic association studies at a fraction of the price. SNP genotyping promises to reveal why some people are more susceptible to particular diseases. In the future, the efficiency of drugs or adverse reactions to drugs for individuals may be predictable, opening the door to personalized medicine. The forensic potential is undeniable.

The analytical accuracy of MALDI-TOF permits unambiguous mass distinction of DNA fragments that differ by only one base and thus facilitated the development of several SNP genotyping methods [13–18]. Alternative analytical methods often require indirect detection via labeled probes, which can add substantially to genotyping errors. A predominant advantage of mass spectrometry is the detection of an intrinsic molecular property of the analyte—the molecular mass. A high-throughput automated SNP genotyping platform, such as the one developed by SEQUENOM, allows the analysis of thousands of SNPs a day [13,18–20].

The MassEXTEND^{TM} assay combines primer oligonucleotide base extension with MALDI-TOF mass spectrometry. The assay consists of a post-PCR primer extension reaction that is carried out in the presence of one or more dideoxynucleotides (ddNTPs) resulting in allele-specific terminated extension products. Primers are designed to anneal adjacent to the SNPs of interest. Depending on the SNP identity, extension products of different length and mass are produced. A homozygous genotype leads to one extension product of defined mass. In heterozygous samples, two products of distinguishable mass are generated. The DNA fragments are usually conditioned and analyzed by MALDI-TOF MS (see Fig. 1).

2.1. Solid-phase primer extension assay

The first generation of MassEXTEND^{TM} was based on a solid-phase scheme (sME). It was necessary to label the

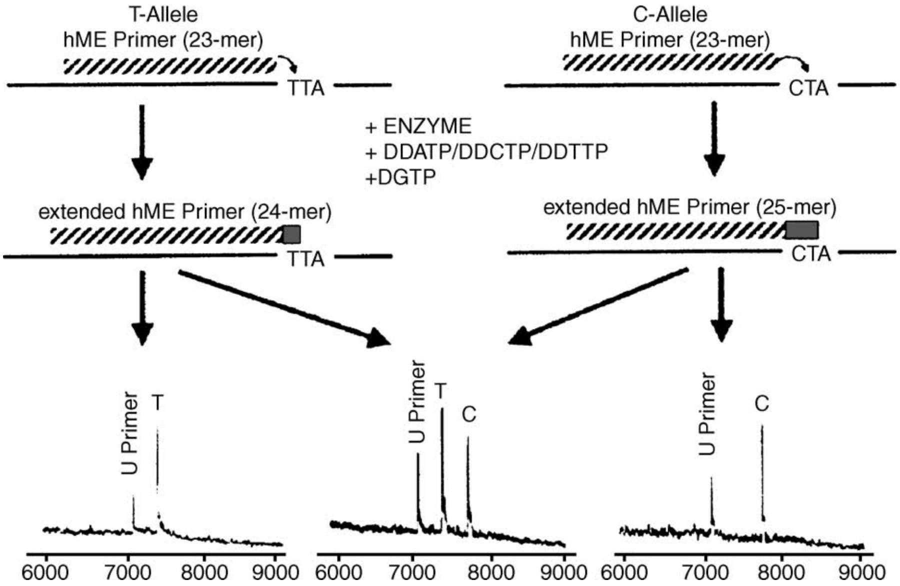


Fig. 1. MassEXTEND (hME) reaction: following PCR amplification of a locus of interest, a primer extension is performed using a hME primer that is designed to anneal next to the SNP. The key feature of the scheme, is the use of a terminator mixture that yields allele-specific extension products differing in length and mass by at least one nucleotide. In the depicted example, a dGTP is used along with terminators ddATP, ddCTP and ddTTP. For the T allele (T), ddATP is incorporated, extending the primer to a 24 mer. For the C allele (C), dGTP is incorporated prior to the termination of the extension by incorporation of a ddATP, and this yields a 25-mer. U primer marks unextended primer.

PCR products with a biotin tag during PCR. The PCR products were captured using streptavidin-coated magnetic beads. Magnetic separation, denaturation and wash steps were performed to generate a single-stranded template for a subsequent primer extension reaction. The primer extension reaction was allowed in the presence of specific deoxy/dideoxy nucleotide mixtures. Resulting allele-specific primer-extension products were washed to remove buffer components and denatured to remove the analyte from the immobilized template. The supernatant was analyzed by MALDI-TOF MS and detected mass signals for converted into genotype information [13].

The advantage of the solid-phase assay is that washing the immobilized products is an efficient means of purification of the primer-extension products. However, there are disadvantages associated with the use of solid-phase capture methods. Magnetic beads, such as those used initially, add significantly to the overall process cost. They increase the complexity of the process and its automation. Finally, solid phases usually have a limited binding capacity, which restricts attempts to higher multiplexing of assays.

2.2. Homogeneous primer extension assay

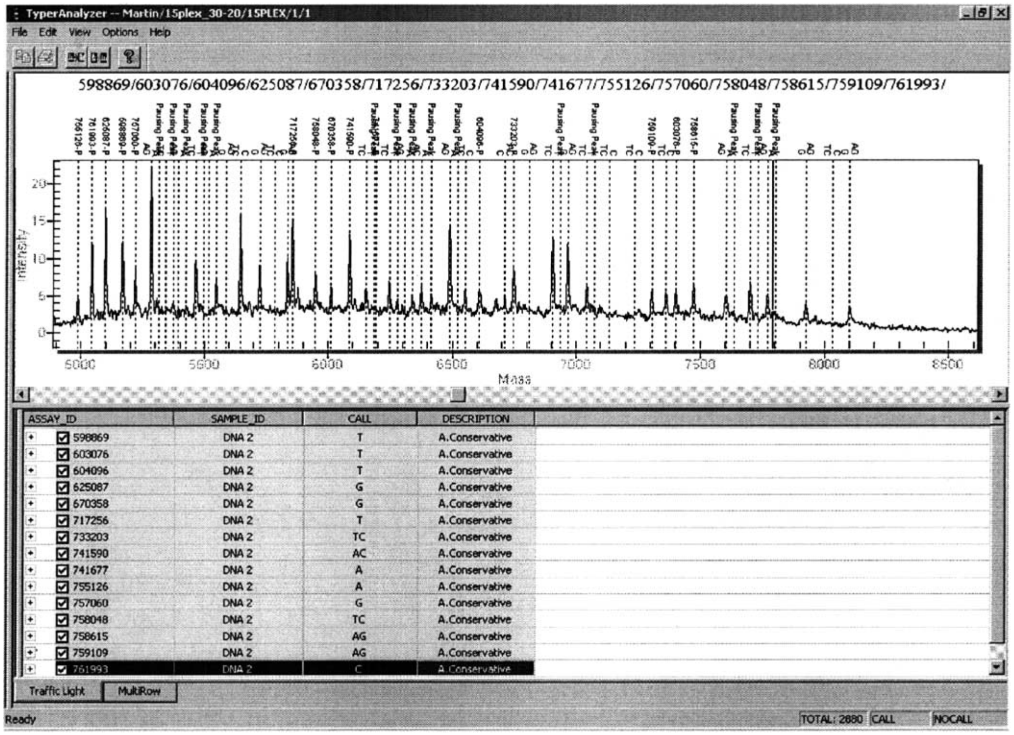
In 2001, a next generation of the MassEXTEND reaction was developed [20,21]. The homogeneous MassEXTEND reaction (hME) is a simplified primer extension method, in which the solid-phase purification step was replaced by a simple enzymatic and ion-exchange resin treatment. It is a single-tube reaction, which only requires addition of reagents required for subsequent reaction steps. Following PCR amplification the addition of shrimp alkaline phosphatase (SAP) facilitates deactivation of remaining unincorporated dNTPs from the PCR reaction. Allele-specific termination products are generated in a cycled primer extension step. Samples are then diluted with deionized water and cation exchange beads (NH_4^+ form) are added [22]. The supernatant is dispensed on miniaturized chip arrays with predefined matrix positions for MALDI-TOF MS analysis. This homogeneous assay format offers similar performances as solid-phase approaches. Since no fluids have to be removed in between steps, it is easier to automate and formed the basis for current high-throughput processes for MALDI-based nucleic acid analysis.

2.3. Multiplexed primer extension reactions

Performing multiple PCR/hME reactions in a single reaction vessel (multiplexing) is a way to increase analysis throughput and reduce the cost per genotype. Assays for SNPs or mutations at different genome locations can be combined to save time and consumables (see Fig. 2). The available mass window for analysis of primer-extension products usually ranges from 5000 to 9000 Da. This corresponds to oligonucleotide length between 17 and 30 Nts. If primers targeting different SNPs are designed such that the masses of each primer and their corresponding extension products do not overlap, each acquired mass spectrum can host multiple unambiguous genotyping results.

2.4. Assay design

In order to design compatible SNP assays for multiplexing, several considerations must be taken into account. Primer



ASSAY_ID	SAMPLE_ID	CALL	DESCRIPTION
☑ 598869	DNA 2	T	A.Conservative
☑ 603076	DNA 2	T	A.Conservative
☑ 604096	DNA 2	T	A.Conservative
☑ 625087	DNA 2	G	A.Conservative
☑ 670358	DNA 2	G	A.Conservative
☑ 717256	DNA 2	T	A.Conservative
☑ 733203	DNA 2	TC	A.Conservative
☑ 741590	DNA 2	AC	A.Conservative
☑ 741677	DNA 2	A	A.Conservative
☑ 755126	DNA 2	A	A.Conservative
☑ 757060	DNA 2	G	A.Conservative
☑ 758048	DNA 2	TC	A.Conservative
☑ 758615	DNA 2	AG	A.Conservative
☑ 759109	DNA 2	AG	A.Conservative
☑ 761993	DNA 2	C	A.Conservative

Fig. 2. Mass spectrum of a 15-plex and corresponding genotype analysis.

combinations for the multiplexed PCR amplification reactions must be designed such that no cross-loci amplification occurs. The design of primers for the extension reaction must yield a combination of expected peak masses for each assay that are resolvable for unambiguous genotyping. On a larger scale, multiplexed SNP assays are usually designed with a computer aided assay design program. MassEXTEND specific PCR primer designs usually adhere to the following: (1) balanced PCR primer length around 20 mer length, (2) balanced T_m around 60 °C and (3) balanced G–C content around 50% and (4) optimal amplicon length of around 100 bp. Potential primer sets should be searched against public databases to avoid single assay failure due to competing kinetics or cross-hybridization reactions with other loci.

For extend primer design there are only two choices of primer sequence, both adjacent to the SNP on either DNA strand. Almost always at least one side allows for successful assay design. Optimal hME primer design should consider the following rules: (1) optimal length of primers between 17 and 24 mer; (2) primers should not be designed in regions with degenerate or ambiguous bases; (3) primers with low T_m, hairpin formation potential, false priming potential, primer–dimer potential and problematic sequence repeats such as GGGG should be disregarded and (4) potential mass conflicts from by-products must be avoided: depurination products and so-called pausing products, which lack the terminal ddNTP due to an incomplete extension reaction, should not obscure unambiguous analysis. For example, an extension primer prematurely terminated with dA has the same mass as a ddG terminated product, and therefore these products are indistinguishable by mass. To avoid miscalls in genotypes these designs should be discarded.

In multiplexing, up to 15 or more primers are involved. The analyte peaks in the mass spectrum for one assay must be sufficiently well resolved from any product of any assay it is multiplexed with. This includes pausing peaks and potential salt adduct signals generated by sodium or potassium contamination. In addition, analyte peaks must fall within a mass window considered benign for automated acquisition with a defined instrument setting. With current linear MALDI-TOF MS, we usually acquire a mass window of 5000–9000 Da. Designs rules for intercalation of assays are based on a mass accuracy of 0.1% and a mass resolution of 500 throughout the specified mass range. These values can be easily achieved in high-throughput settings when miniaturized chip arrays are employed [23]. Based on these considerations, up to 15–20 assays can be combined successfully (see Fig. 2).

Another way to simplify assay multiplexing is to employ a pinpoint strategy [15,17,24]. In this scheme primer extension reactions are performed in the presence of all four ddNTPs. For both alleles of a SNP, the primer extension is terminated at the first non-primer nucleotide. Although this scheme offers greater simplicity and potentially higher plexing designs, it challenges the capabilities of currently available linear MALDI-TOF MS. The mass difference between ddATP and ddTTP is only 9 Da and the assignment of the "correct" extension product to a mass signal in a fully automated, high-throughput setting is not reliable enough. Additionally, some mass differences in pinpoint designs fall very close to either sodium- or potassium-adduct signals. Pinpoint designed assays, if used with automated generic conditions, can thus be prone to peak misinterpretation and genotyping errors. Allowing one allele to be extended through the SNP by using a suitable mixture of dNTP and ddNTPs creates more options in spreading mass signals appropriately in a mass spectrum.

2.5. PCR optimization

In order to perform successful high level multiplexed reactions, the biochemical conditions must be optimized for balanced amplification of all SNP-specific target regions. Commonly used PCR additives such as DMSO, BSA, detergents, urea, glycerol and the Q solution (Qiagen Inc.) cannot be utilized with MALDI-TOF MS readout because they disturb the analyte-matrix crystallization needed for MALDI-MS. To increase multiplexing success, the PCR amplification conditions were optimized using a statistical experiment design, the Taguchi method [25]. Orthogonal arrays L9 (3(4)) were used to estimate interactions and effects of different concentrations and ratios of reagents such as *Taq* enzyme, nucleotides, magnesium, primers, PCR buffer (mono-ionic salts) and genomic DNA. Initial results were showing that an increased concentration of PCR buffer to two-fold, improved overall amplification performances in multiplexing (Table 1). However, after analysis of primer extension reactions and MALDI-TOF MS analysis, the new conditions revealed detrimental effects on the genotyping performances (data not shown). Albeit the PCR improvements, the increased salt concentration (in particular mono-valent ions) had an inhibitory effect on the ThermoSequenase (TS) enzyme used for MassEXTEND. Furthermore, the increase in salt concentration imposed a greater challenge on the analyte conditioning by ion-exchange resin.

Using primer extension ratios and MALDI-readout as measurement points, PCR buffer concentrations were optimized to 1.25×. Robust PCR amplification can be achieved at this value without significant influence on TS activity, but at much higher genotyping success rates (see Fig. 3). The Taguchi optimization of PCR conditions further included the following changes: increased concentration of each of the nucleotides to 500 μM, MgCl$_2$ to 3.5 mM, HotStar Taq PCR enzyme to 0.15 U, PCR primers to 100 nM each and a slight reduction of sample DNA to 2 ng/5 μl reaction (see Table 1).

Table 1
Comparison of PCR conditions changes for high-level multiplexing MassEXTEND^{TM}

	Previous[*]	New A	New B
Genomic DNA (ng/rxt)	2.5	2.0	2.0
HotStarTaq PCR buffer	1×	2×	1.25×
dNTPs (μM)	200	500	500
MgCl$_2$ (mM)	2.5	3.5	3.5
PCR primers (nM)	50	100	100
HotStarTaq (U/rxt)	0.1	0.15	0.15

[*] Previous conditions as described in [20]. The new conditions A and B show only reagents that incurred a change. The cycling conditions were the same as previously described [20].

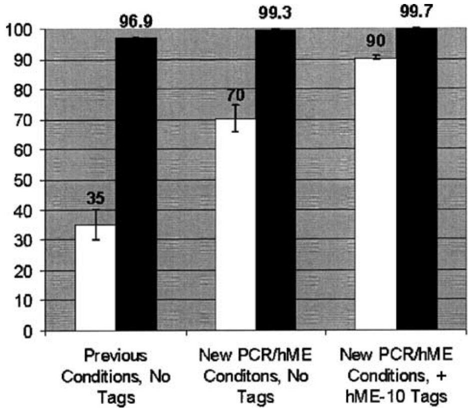


Fig. 3. Performance of 12-plexed hME under different reaction conditions. Three experimental conditions are compared. Results show averages of seven 12-plexed reactions (84 assays) performed on seven previously genotyped genomic DNA samples. The experimental conditions used are indicated at the bottom of the chart. Depicted are first pass results obtained from realtime analysis. The average percentage of successful real-time calls are indicated by open bars. The average percentage of accurate real-time calls are indicated by the closed bars. Standard deviations are indicated.

An important factor is the final concentration of "free" Mg^{2+} ions ([MgCl$_2$] − [total dNTPs]). Magnesium is an important co-factor to DNA polymerase enzymes. When altering nucleotide concentrations, the MgCl$_2$ concentration must be adjusted accordingly. We found the optimal free magnesium concentration to be between 1.3 and 1.7 mM. One of the most significant factor contributing to increased PCR amplification rates in multiplexing, was the addition of a 10 mer tag to the 5′ end of both, forward and reverse, PCR primers (5′-ACGTTGGATG-3′). The effect of the tags was additive and contributed to a further ∼30% improvement in genotyping rates (see Fig. 3).

2.6. Primer extension optimization

The increased amount of analytes in high-level multiplexing generates mass spectra of higher complexity and lower signal-to-noise ratios. Variations in analyte peak intensity may be caused by inconsistent oligonucleotide quality or quantity (concentration), or by differences in desorption/ionization behavior. To avoid decreased success rates, primer amounts can be normalized by the signal intensity in the mass spectrum, after aliquots of each multiplex primer mixture are analyzed with MALDI-TOF MS.

The Taguchi optimization of the primer extension reaction led to increased concentration of primers (from 0.5 to 1 μM each), an increased ThermoSequenase amount of 1.25 U/rxn, and as one of the most significant factors, an increased number of extension cycles (from 55 to 75).

2.7. Performance

Current MALDI-MS multiplexing (12-plex level) provides an average first-pass call rate of 90%, when the generic, Taguchi-optimized protocols are applied. The accuracy of the genotype data (automated calls, first pass) is averages to 99.7%. The overall call rate and accuracy obtained for 12-plex is in agreement with a previous study conducted by the Whitehead Institute Center for Genome Research initially performed at five plex-levels. Using the MassARRAY^{TM} platform, they estimated the accuracy rate to be at 99.6% [26]. Most importantly, the 10% no-call rate is not randomly distributed. No-calls are usually confined to specific assays, which perform weakly or not at all under generic, high-throughput conditions. Out of 84 assays used in a recent study, 10 exhibited significantly weaker extension rates as compared to the uniplex format. Those assays generally provide lower call rates and are more prone to wrong genotype calls. Biased amplification of alleles during PCR is another source of errors—allele bias can be observed in uniplexing but is enhanced in multiplexing. Additionally, assays exhibiting very low primer extension efficiencies are at risk for miscalls, especially when their analytes fall in a higher mass range (over 7000 Da). Because most of the miscalls are not random, problematic assays can be identified by automated analysis of primer extension ratios, cluster analysis of signal-to-noise ratios [27] and Hardy-Weinberg equilibrium analyses [26]. These statistical tools are useful in isolating problematic assays. Individual genotypes can be corrected or rejected. Selected assays can be filtered out from analysis and further experiments. Usually, after non-functional assays and weak performers are removed, assays can be called with an analytical efficiency over 97% and an accuracy of 99.8%. As with a variety of other genotyping technologies, the biggest contributor to the remaining assay failure is random PCR drop-out.

2.8. Throughput of MALDI-TOF MS-based genotyping

Current linear MALDI-TOF mass spectrometry instrumentation provides sufficient resolution, mass accuracy and mass range to allow the design of multiplexed genotyping assays. Nowadays, 12–15-fold multiplexing can be routinely carried out using the generic experimental conditions described here and on a principal basis, 20–25-fold seems feasible with further assay optimization. Routine 12-fold multiplexing with ∼30 min acquisition/real-time analysis per 384 MALDI sample spots, translates into an analytical speed of ∼150 genotypes per minute (9000 genotypes/h). This throughput coupled with a cost/genotype of currently below 10 renders whole-genome scan studies feasible and affordable.

3. Quantitative MALDI-TOF MS of nucleic acids

As described in the previous section, the resolution and mass accuracy of MALDI-TOF MS allows genotyping of genetic variations using primer extension products. Relative quantification of primer extension products can also be achieved with this platform. On a per sample basis, the ratio of analyte, co-crystallized with matrix and subsequently ionized after laser excitation, is proportional to the peak area observed in MALDI-TOF MS. This characteristic was observed for sets of protein analyte peaks present in the same reaction [28] and is also true for nucleic acids. This facilitates quantitation of DNA analytes relative to each other in the same sample. Comparison of single analyte products between reactions is much more difficult however. This difficulty arises from multiple factors such as the uneven distribution of 3-HPA matrix/analyte crystals and differential ionization of analytes from preparation to preparation. This difficulty is overcome by measuring at least two analyte products per reaction. This allows for the comparison of analyte ratios between samples.

Many interesting quantitative nucleic acid applications are now possible using MALDI-TOF MS for their analysis. Table 2 lists several of the applications currently being conducted. As illustrated in this table, two primary approaches to quantitation are used. The first involves the measurement of two unknown nucleic acid analytes with known mass but unknown concentration relative to each other. Allele frequency estimation in pooled populations of nucleic acids is a good example that is discussed below. In this case, the ratio of the two measured alleles is compared between samples to quantitatively determine differences. Relative quantitation is achieved by normalizing the baseline of each spectrum. Identified peaks at specified masses (corresponding to expected alleles) are then given a Gaussian fit and the area under the curve is integrated. Within each spectrum, allele frequencies of the primer extension products are estimated as the ratio of the area of one allelic peak to the total area of all of the expected allelic peaks. Since one is dealing with populations of molecules in this model the sum of the frequencies for the alleles under investigation always add up to 1 (100%).

The second approach involves the addition of an internal standard oligonucleotide of known sequence, mass and concentration to all reactions. Unknown analyte concentrations are determined by titrating the internal standard over a range of concentrations such that the point at which the unknown analyte(s) and the internal standard oligonucleotide are at approximately a 1:1 ratio is identified (based on peak areas of primer extension products). Exact molar concentrations can be assigned to the unknown analyte via regression analysis of the plotted peak area ratios for each sample and concentrations can be compared between samples. The approach is successfully used for the measurement of gene expression levels as described below.

3.1. Analysis of heterogeneous nucleic acid mixtures (allele frequency estimation)

The first description of relative quantitation between two specific alleles present in a mixed DNA population was described by Ross et al. [29]. In this study, DNA mixtures were prepared from genotyped homozygote and heterozygote

Table 2
Assays developed for relative and absolute quantitation of nucleic acids by MALDI-TOF MS and their corresponding applications

Quantitative method	Assay	Application
Ratio of two analytes of unknown concentration	Allele-specific quantitation	Disease association studies, allele-specific expression
	Allele determination in polyploidy genomes	Agricultural genetics
	Gene copy number	Monosomy, trisomy, transgenic animals
	Loss of heterozygosity	Cancer diagnostics
	Loss of imprinting	Cancer diagnostics
	Viral typing	Vaccine QC
Ratio of unknown analyte to internal standard of known concentration	Gene expression analysis	Quantitative gene expression analysis
	Quantitative PCR	Multiple applications
	Gene transfer estimations	Gene therapy
	Gene duplication/multiplication	Genetic diagnostics
	Viral/bacterial titering	Pathogen quantitation/treatment control monitoring

samples to create pooled samples with known allele frequencies for three primer extension assays. The study demonstrated that allele frequencies in complex mixtures of DNA had a limit of detection (LOD) to 2% and a limit of quantitation (LOQ) of 5–10% for minor allele frequencies using MALDI-TOF MS. These findings and those by our group conducting similar experiments at the time (see Figs. 4 and 5) have held up after extensive validation and are used as defining parameters for large-scale association studies and genome-wide scans. Fig. 4 shows a scatter plot of estimated allele frequencies versus allele frequencies derived from individual genotyping for 36 primer extension assays distributed randomly throughout the human genome. Ninety-six individual genomic DNAs were first genotyped for all assays to establish the exact allele frequencies in this population. The DNAs were then pooled at an equimolar ratio to test the accuracy of estimating allele frequencies relative to genotyped frequencies using the homogeneous primer extension assay [20]. Each assay was conducted in quadruplicate and standard deviations of 2% or less were achieved for the majority of assays as shown. The coefficient of determination (R^2) in Fig. 4 exceeds 0.90 indicating good, but not perfect, correlation between the allele frequencies calculated from the pooled template versus the genotyped ones. These differences arise from multiple issues such as uneven amplification of alleles during PCR and primer extension reactions that occur for all technologies relying on these processes [30,31]. Other MALDI-TOF MS specific issues such as uneven analyte/matrix cocrystallization and preferential analyte desorption/ionization between alleles can also contribute to this phenomenon. These effects, however, appear to have a minimal contribution (<2% inaccuracy) based on oligonucleotide titration's simulating primer extension products from 0 to 100% frequency for two alleles (data not shown). The summed effect of all

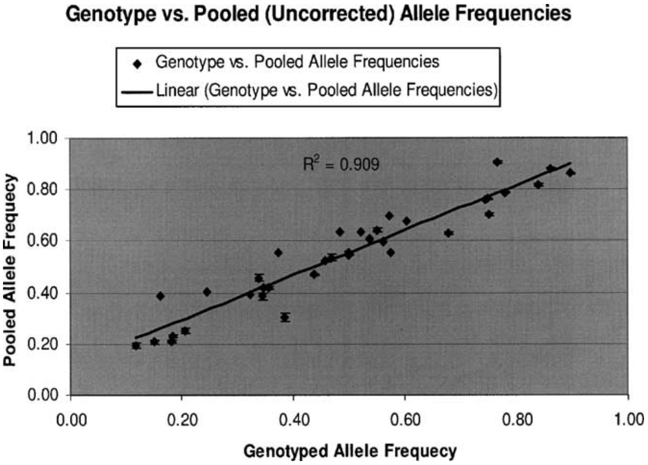


Fig. 4. Scatter plot of genotyped population allele frequencies (x-axis) vs. allele frequencies calculated using pooled population DNAs (y-axis). Thirty-six unique assays are depicted. The DNA population consisted if 96 individual DNAs at equimolar concentrations (260 pg per individual DNA/μl = 25 ng/μl). Frequencies were calculated using MassARRAY TYPER RT software. The calculated allele frequency with standard deviation for each assay represents the average of four replicate reactions each dispensed in replicates of four onto silicon chip arrays loaded with matrix. For genotyped frequencies, each of the 96 individual DNAs was genotyped for each of the 36 assays using the MassARRAY system. Best-fit line and coefficient of determination (R^2) were calculated using Excel 2000 (Microsoft).

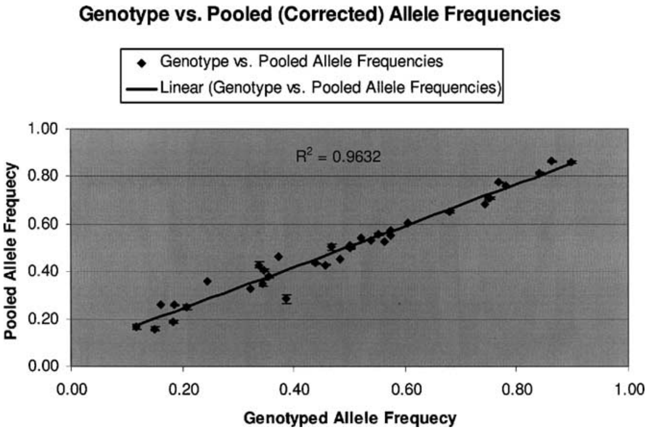


Fig. 5. Scatter plot of genotyped population allele frequencies (x-axis) vs. allele frequencies calculated using pooled population DNAs (y-axis) as described in Fig. 4. The pooled allele frequencies have been corrected for allelic imbalances of the 36 assays using technique described in the text. Note the improvement in the coefficient of determination (R^2) after correction with individual heterozygote allele ratios.

inaccuracies can be measured using individual heterozygous samples for each assay under investigation. In theory, heterozygotes have an allele ratio of 1:1 (one allele on each chromosome) and should therefore exhibit a 1:1 ratio of allele frequencies (0.5:0.5). Any deviation from this expected ratio can be quantified in the same manner as described for pooled templates. The average frequency for each allele from multiple heterozygotes in a population for a particular assay therefore represents the summed effect of any inaccuracies for that assay. The calculated deviation can then be applied to the pooled result as a correction factor [32]. Fig. 5 shows the 36 primer extension assays of Fig. 4 with corrected allele frequencies for the pooled DNA results. The coefficient of determination (R^2) improves to 0.96 as a result of this correction. Multiple studies have compared the quantitative abilities of various platforms used for estimating allele frequencies in populations of nucleic acid molecules [33,34]. MALDI-TOF MS based measurement of primer extension reactions have been shown to be as accurate, sensitive and reproducible as all available technologies according to these studies.

The ability of a MALDI-TOF MS based platform to determine allele frequencies in pooled nucleic acid populations for the purpose of conducting complex disease association studies has been investigated by several groups including us [35–37] and is indeed a viable solution to streamlining large studies which would otherwise require many thousands to millions of genotypes per association study which is cost and time ineffective. Our studies to date and an overview of their results are presented in the next section of this review.

Several other interesting applications have been carried out using semi-quantitative primer extension based MALDI-TOF MS analysis with unknown analyte concentrations. Knight et al. measured allele-specific differences between regulatory polymorphisms associated with the ability of RNA polymerase II to bind and assemble its transcription complex at the start site of transcription for several eukaryotic promoters [38]. This technique provides a powerful tool for identifying important regulatory SNPs and haplotypes in vivo.

Amexis et al. used semi-quantitative MALDI-TOF MS for vaccine QC of the mumps virus [39]. Here, ratios of mumps viral quasispecies were measured in Jeryl Lynn substrains for the live, attenuated mumps/measles vaccine. Ratios between the two substrains were calculated at five distinct nucleotide positions within the viral genome and the results were shown to match and improve upon the existing quality control methodology used by the FDA. The methodology can be used in vaccine QC of any RNA or DNA virus.

3.2. Gene expression analysis

The quantitative measurement of mRNA transcripts using primer extension products and MALDI-TOF MS was approached in a slightly different manner than that of allele frequency analysis in pooled nucleic acid mixtures. Transcripts usually exist in one allelic form and are therefore not useful for the comparison of two or more alleles. Ding and Cantor [40] established a solution for the quantitation of non-allelic templates by introducing a synthetic template that varies from the wild-type reverse transcribed mRNA (cDNA) by a single base. An oligonucleotide of 60–90 bases acts as a competitor in the PCR amplification step and the resulting product contains a differentiating allele, which can be used to distinguish the wild-type cDNA template from synthetic template. Since the synthetic template is of known concentration it can be titrated and used as an internal standard to quantitate the amount of a particular cDNA. This combination of competitive PCR coupled with primer extension has proven to be very effective. The detection of cDNA amounts is independent of PCR cycle numbers, because the two templates rely on the same primers and differ in sequence only by 1 bp. Primer extension reactions, which distinguish the cDNA from the internal standard, also rely on the same primer. Titrating the internal standard over a range of concentrations and conducting regression analysis of the plotted peak area ratios yields the point at which the cDNA and the internal standard alleles are at a 1:1 ratio. This point represents the concentration of cDNA present in a particular tissue or sample.

Fig. 6 shows results of a logarithmic titration of the internal standard versus a mix of cDNAs representing major tissue/organ systems of humans. Two "housekeeping" genes are shown, GAPDH and HMBS. Note that for these two genes the point at which the y-axis crosses 0.50 (point at which cDNA and standard are at 1:1 ratio) is approximately 3 logs apart. These results match those of a similar tissue set using 5′-nuclease digestion and fluorescence to quantitate the ex-

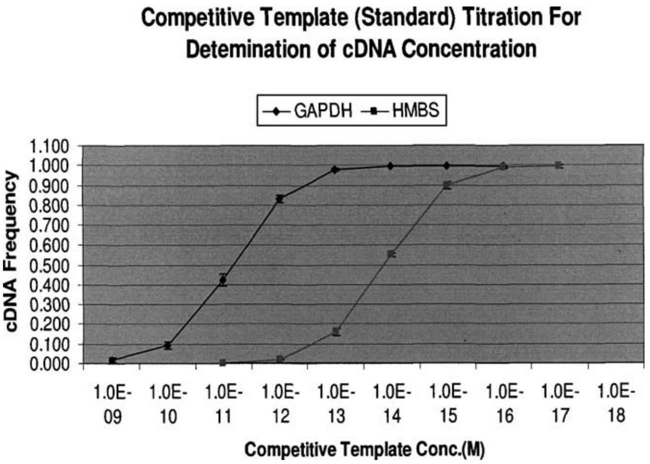


Fig. 6. Results of nine-log range titration of competitive template (internal standard) to determine relative gene expression levels for two "housekeeping" genes: glyceraldehyde 3 phosphate dehydrogenase (GAPDH) and hydroxymethylbilane (HMBS). Input cDNA amount was constant for all titration points. The point at which each titration curve crosses the y-axis at 0.50 represents the concentration at which a 1:1 ratio of cDNA to internal standard alleles has been observed and therefore represents the concentration of cDNA in the sample. Note that each gene shows an increase in cDNA frequency as the internal standard (competitor) concentration is decreased logarithmically until only cDNA template products are detected in MALDI-TOF MS. Each data point represents the average of replicate reactions with standard deviation indicated for each data point.

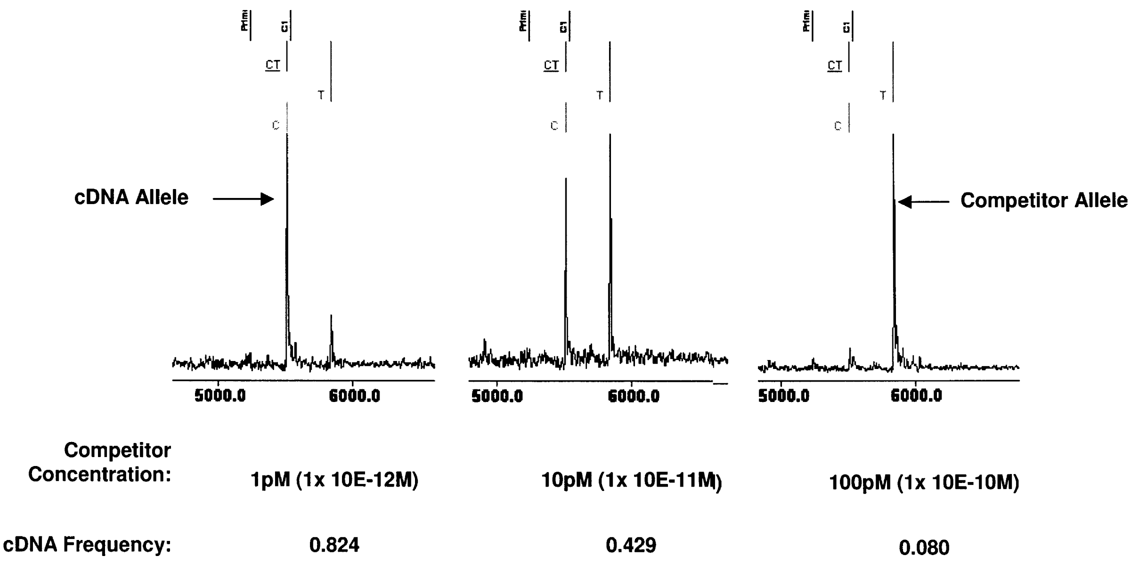


Fig. 7. Representative spectra for the GAPDH titration curve depicted in Fig. 6. The three spectra represent three unique reactions with one log differences in the concentration of the competitor molecule (internal standard) used in the PCR reaction as described in the text. The ratios of the cDNA and competitor alleles change as the amount of competitor included increases.

pression levels of these two genes [41]. Each reaction in Fig. 6 was conducted in triplicate and resulted in standard deviations of <2% as shown. Fig. 7 shows the representative spectra for the GAPDH gene in this titration. The low mass peak represents the wild-type cDNA. Notice how the ratio of the two allelic peaks change as competitor concentration increases. Multiplexed analysis up to a triplex reaction is possible without loss of accuracy (data not shown). The use of an internal standard for each reaction and the possibility to include a gene for sample normalization for each reaction via multiplexing make this methodology very attractive for expression studies on large numbers of samples in a time and cost efficient manner. In addition, allele-specific expression can be conducted in the presence or absence of the internal standard oligonucleotide using the methodologies described above. Large-scale identification of genetic variations and gene expression profiling offer a powerful combination, which should provide insight into the genetic components of complex diseases [42,43].

4. SNP and mutation discovery by MALDI-TOF MS

As described earlier, the exploration of SNPs is one of the first approaches to extract medical and biological value of the genome sequencing data and to elucidate inter- and intraspecies genetic variations. A sufficiently dense SNP map is a pre-requisite for genome-wide association studies and also forms the basis for building a haplotype map of the human genome. Although an almost complete human reference sequence is now available and over four million SNPs have been deposited in public database, the demand for de-novo DNA sequencing as well as differential sequencing is not diminishing. First, there is still a plethora of uncharacterized organisms, especially microbes. Second, several phases of genome scans and association studies continue to require DNA sequencing. Once chromosomal regions or gene regions are associated with a particular phenotypic trait, usually denser SNP panels are constructed. Public databases do not always reveal sufficient SNPs in the regions of interest and thus differential sequencing has to be performed. Third, once the gene(s) underlying the phenotypic trait are confirmed, causal variants have to be identified. This usually requires “re-sequencing” a multitude of individuals. Finally, as long as a disease cannot be attributed to specific recurring mutations at defined positions, the disease genes will have to be re-sequenced for patient-specific diagnostics. Recent years have seen significant developments to enable re-sequencing of longer target regions by MALDI-TOF MS and address some of the points mentioned above.

A novel high-throughput SNP discovery approach (the discovery of nucleotide substitutions, insertions and deletions) couples a homogeneous in vitro transcription/RNase cleavage system with MALDI-TOF analysis of the analyte intrinsic molecule property, the molecular mass, and unparalleled speed of signal acquisition as well as the potential for high degree automation thus facilitating industrial scale application. TOF instruments acquire data in microseconds as opposed to hours of analyte separation in conventional gel-electrophoresis.

Earlier attempts to apply MALDI-TOF MS to de novo sequencing and to improve speed and accuracy of Sanger sequencing, as an alternative method for separation and detection of Sanger sequencing ladders [44–46] were hindered by ion fragmentation, the exponential decay in sensitivity with increasing fragment mass, limiting mass resolution and accuracy of the high mass range in conventional axial-TOF instruments.

In analogy to dideoxy sequencing, mass differences between nested sets of truncated DNA sequence fragments originating from a primer were analysed by MALDI-TOF MS and the mass differences of the DNA fragments were used to calculate the nucleotide sequence. Despite several biochemical strategies generating DNA sequencing ladders of sufficient yield and purity to suit MALDI-TOF MS analysis and promising results for solid-phase based sequencing and cycle sequencing, routine read lengths exceeding the 100 bp barrier have never been achieved [46–48].

In contrast to the primer based extension reaction of Sanger sequencing the novel comparative re-sequencing scheme for SNP discovery is based on the generation of short base-specific cleavage products of a given nucleic acid amplificate in homogeneous single-tube assays. Parallel processing in 384 well formats is combined with miniaturized sample preparation on matrix pads on the surface of modified silicon chips carriers. The biochemistry comprises a T7-promotor mediated in vitro transcription and the analysis of cleaved RNA strands by mass spectrometry as shown in Fig. 8 [21,49,50].

The target region of interest is PCR amplified in two reactions utilizing different 5′-promoter tagged primer pairs. The first reaction introduces a T7-promotor tag in the forward strand of the amplification product, while the second reaction incorporates the T7-promotor tag in the complementary reverse strand. Following PCR and deactivation of the deoxy-NTPs by dephosphorylation, RNA polymerase and ribonucleotides are added to the reaction mixture. In vitro transcription generates a single-stranded RNA template facilitating further amplification. Under MALDI-TOF conditions RNA is more stable than DNA and it enables the use of base-specific RNases for base-specific template digestion. Four base-specific cleavage reactions are driven to completion reducing the original single copy of the target sequence to a specific set of RNA fragments in each of the reactions, which are desalted by ion-exchange, conditioned and readily separated and analysed by MALDI-TOF MS. A set of compomers can be unambiguously calculated and assigned to the signals in the resulting mass spectra. This experimental set of compomers is used to reconstruct the sequence by combining and cross-comparing the information of the four cleavage reactions to the in silico cleavage pattern of a known references sequence. Sequence changes have a profound impact on the mass signal pattern. A heterozygous sequence change can generate up to five discriminatory

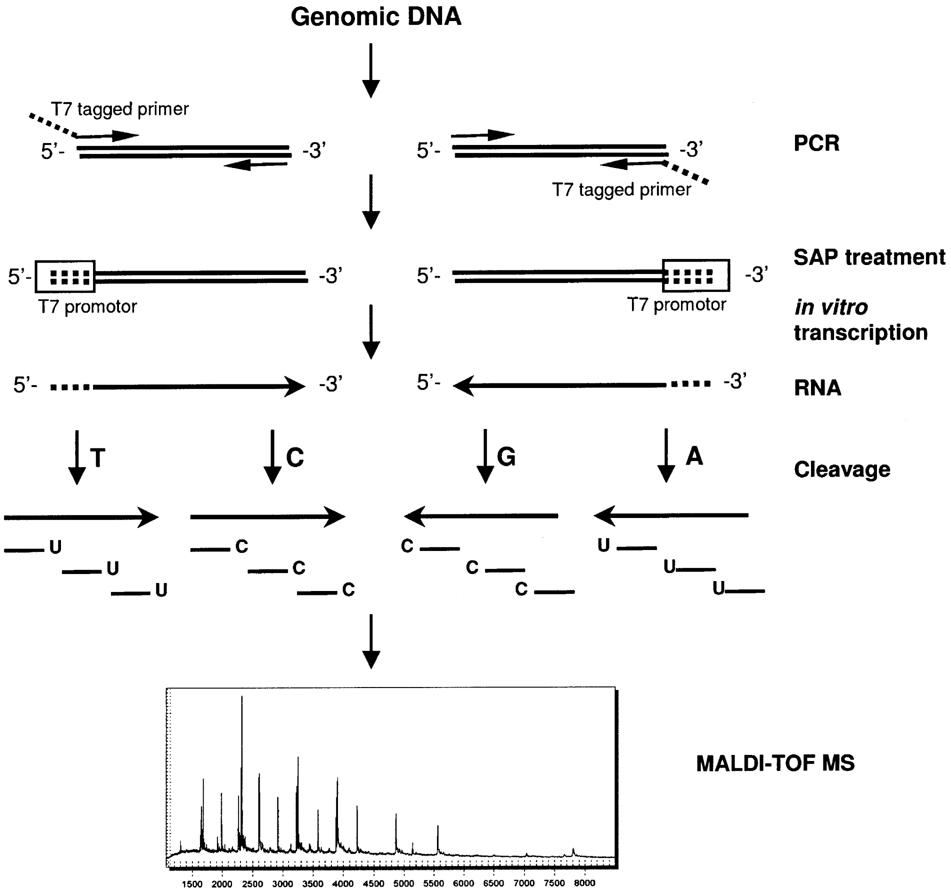


Fig. 8. MALDI-TOF based re-sequencing scheme. The PCR amplification of the sequence of interest is followed by T7-mediated in vitro transcription and base-specific cleavage. RNA of the forward strand is cleaved at U or C. An A and G-specific cleavage of the template sequence is facilitated by U or C cleavage of the reverse RNA strand. MALDI-TOF acquisition of spectra of each of the cleavage reactions is followed by comparison to reference sequence derived in silico cleavage pattern.

observations in the mass spectrum by adding or removing a cleavage site as well as shift the mass of single products by the mass difference of exchanged nucleotides. This compares to only one observation in conventional fluorescent Sanger sequencing: two colors with the same elapsed migration time. A homozygous sequence change might even result in up to 10 observations since not only additional but also missing signals can be used for SNP identification. In most cases, the combined observations of all reactions allow for the detection, identification and localization of the sequence change.

Fig. 9 illustrates the effects of a [C/T] sequence change in a 650-bp amplicon. Mass spectra were derived after a C- and a T-specific cleavage reaction of the forward and reverse strand—equivalent to four base-specific cleavage reactions.

For the wild type sample [C/C] signals of all cleavage reactions can easily be identified on the basis of the reference sequence derived from the in silico cleavage pattern. In this

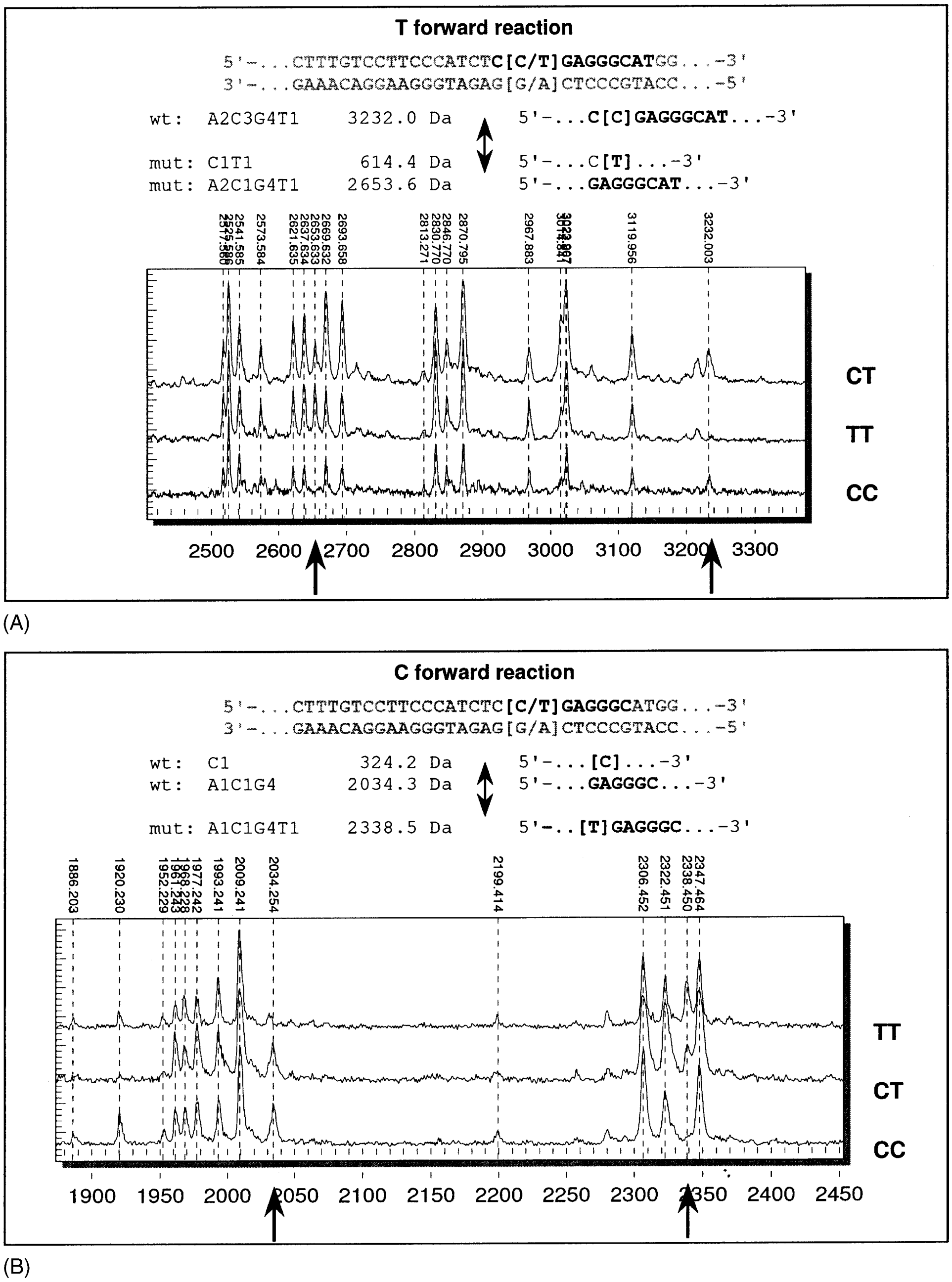


Fig. 9. Identification of a single nucleotide polymorphism by MALDI-TOF MS analysis. Panels A–D view the four base-specific cleavage reactions. Each spectrum reveals spectral changes resulting from the substitution of a C wt-allele for a T mutant-allele. Spectra of all three different genotypes are overlayed. Arrows indicate spectral changes. (A) T-specific cleavage of the forward transcript. (B) C-specific cleavage of the forward transcript. (C) T-specific cleavage of the reverse transcript. (D) C-specific cleavage of the reverse transcript. Spectra are shown covering all detected signals with an indication of the sequence change affected signals in a zoomed view.

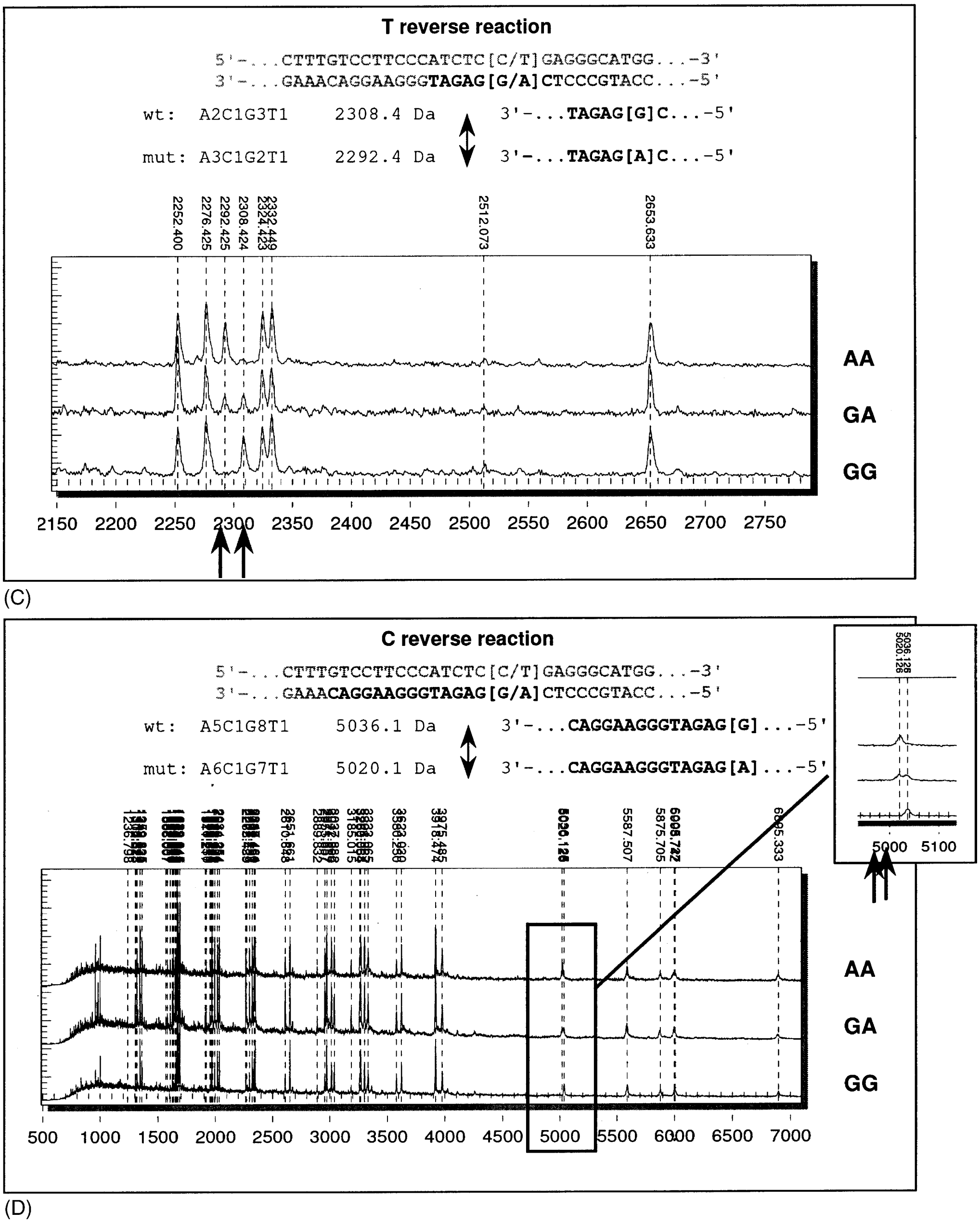


Fig. 9. (*Continued*).

case, the in silico pattern matches the detected mass peak pattern. Deviations from the in silico pattern lead to unambiguous identification of the sequence variation.

In the T-specific cleavage reaction of the forward transcript a sequence change C to T at position 391 of the amplicon sequence introduces a new cleavage side and thus splits the original 10-bp fragment into a 8-bp and a 2-bp fragment. For the mutant sample [T/T] this corresponds to a disappearance of a mass signal at 3232.0 Da and the appearance of mass signals at 2653.6 Da and 614.4 Da (signal not shown). Spectra containing all signals derive from samples with a heterozygous sequence change [C/T].

The C-specific reaction on the forward strand confirms the observation of a [C/T] substitution at position 391 of the amplicon. A cleavage site is removed and generates a 7-bp fragment of 2338.5 Da out of a 6-bp fragment of 2034.3 Da and a monomer C at the position of the substitution.

Additional confirmatory information is generated from the reverse strand of the amplicon. The T-specific cleavage reaction generates a fragment mass shift corresponding to an exchange of the base G on the 2308.4 Da fragment versus an A, shifting the mass of the fragment by −16 Da to 2292.4 Da. Peak intensities of the homo- and heterozygous samples correlate well with the related amount of cleavage product as the peak intensity for the signals of the homozygous samples decrease by half of their intensity in the heterozygous sample.

Again a mass shift of −16 Da applies in the C-specific reverse reaction for the 15-bp fragment at 5036.1 Da generating a fragment at 5020.1 Da.

Further experimental back-up is obtained when all supporting mass signals are correlated across a multitude of samples.

However, an analysis becomes increasingly challenging the more the region of interest deviates from the reference sequence.

Within the low mass range of the spectrum some cleavage information is lost due to overlapping mass signals, especially in the mass range of 200–1000 Da. Mono-, di-, tri- and even tetra-nucleotides are non-informative due to coinciding fragments with an additional diminished detection due to strong matrix signals within the low mass range.

The longer the target sequence the higher the likelihood of overlapping masses of cleavage products and the reduction of possible observations based on additional and/or missing signals. This limits the ability to exactly localize a sequence change in the region of interest to a mass window between about 1100 Da and 10,000 Da.

A simulation of arbitrarily chosen 500-bp amplicons in the human genome revealed that about 90% of all theoretical sequence changes could be detected, characterized and localized. Approximately 10% can still be detected and characterized with the fraction of non-detectable sequence changes below 1% [51]. An approach using signal intensity or peak area evaluation will support additional and missing signal observations, but is dependent on the reproducibility of signal-to-noise ratios between spectra and samples, a challenge related to homogeneous sample preparation, analyte homogeneity and spectra summing.

Recent software facilitates automated SNP discovery and mutation detection using a time-efficient algorithm to discover and pinpoint sequence variations based on the information content of the four cleavage reactions [52].

Considering four base-specific cleavage reactions, an amplicon length of approximately 500-bp and 5 s for data acquisition at a laser pulse repetition rate of 20 Hz, a single MALDI-TOF mass spectrometer can easily scan 2.5 million base pairs per 24 h. This is comparable to the throughput of a conventional capillary-based sequencing instrument.

Real-time quality control of the mass spectra has prompted the development and improvement of a real-time data-acquisition software [53]. Spectra judgment included baseline correction, peak identification as well as internal calibration.

As opposed to sequencing, the combination of mass spectrometry with base-specific cleavage offers the significant advantage of the redundancy of information from four cleavage reactions and the resulting enhancement of the reliability of results.

The complexity of spectra resulting from base-specific cleavage is starting to challenge the capabilities of current axial MALDI-TOF mass spectrometers. Mass accuracy, mass resolution, sensitivity and the dynamic range are becoming limiting factors, when one attempts to fully exploit these applications on a biochemical basis like for extended amplicon length or in the detection of rare genetic variants in DNA pools or mixtures such as tumor biopsies. The orthogonal ([O]-TOF mass spectrometer) TOF has a high potential to meet these increased requirements and initial experiments revealed promising results in the analysis of very high density nucleic acid fragment species [54].

The feasibility of the system exceeds SNP discovery to applications like pathogen identification and general marker detection, methods that can efficiently score large numbers of genetic markers in selected populations to determine genotypic as well as phenotypic correlations [55,56].

Additional fields of application for this large-scale comparative sequence analysis tool include methylation pattern analysis and mutation screening as well as large-scale characterization of cDNAs and their alternative splice variants—further steps in the attempt to elucidate the genetic code and its individual variations.

SNP-discovery by base-specific cleavage is an expansion of available molecular methods and a significant milestone in the portfolio of MALDI-TOF MS applications in the field of genomics.

5. Genome-wide associations studies using MALDI-TOF MS

Early in 2002 Sequenom initiated several large-scale single nucleotide polymorphism-based genetic association studies in multiple disease areas, including melanoma, breast cancer, lung cancer and prostate cancer, metabolic diseases such as type II diabetes and low/high-density lipoprotein cholesterol (HDL-C), musculoskeletal disorders such as osteoarthritis and osteoporosis, and central nervous system/brain disorders such as schizophrenia. In order to carry out these genetic association studies, Sequenom developed a gene-based SNP map consisting of over 100,000 SNP markers with experimentally validated allele frequencies (Braun et al., submitted for publication). Currently, genotyping large numbers of SNPs over sample numbers large enough to provide the power necessary to discern statistically significant frequency differences between case and control samples is not feasible, as both time and money are limiting. For example, genotyping 100,000 SNPs over 300 case samples and 300 control samples would involve 6×10^7 genotyping reactions, a staggering number at this time. To address this challenge, our group as well as others established strategies and methodologies to compare allele frequencies between case and control samples estimated from DNA sample pools [30,36,57]. By pooling DNA samples, the number of reactions performed is drastically reduced, providing a means to reduce the time and cost involved in the completion of high-throughput genome scans. As an example, the analysis

Table 3
Features of the 28,000 SNP marker set for genome scans

Gene coverage	15,272
Exonic	7,831
Intronic	12,630
10 kb upstream region	4,739
10 kb downstream region	5,550
Intergenic regions	2,765

A SNP is considered to be located in a gene region if it is within an exon, intron, or 10 kb up or downstream. Based on this characterization, 16% of SNPs mapped to two or more genes. The numbers of synonymous and non-synonymous SNPs were calculated using annotation of NCBI reference SNPs.

of 100,000 SNPs over two pools of 300 cases and 300 controls requires only 200,000 reactions, as compared to 6×10^7reactions required in an individual genotyping approach. We report here the application of this DNA pooling or 'allelotyping' approach in multiple genome scans for genes involved in complex traits using approximately 28,000 SNP markers in case-control DNA pools. The features of these SNP markers are summarized in Table 3.

The primary SNP set analyzed in this study included 25,488 SNPs previously confirmed polymorphic SNPs. These SNPs reside in genes or in close proximity to genes. More specifically, SNPs in the set are located in exons, introns, and within 10,000 base pairs upstream of a transcription start site of a gene. In addition, SNPs were selected according to the following localization: ESTs; Locuslink or Ensembl genes; and predicted promoter regions. SNPs in the set were also selected on the basis of even spacing across the genome. Another 3017 SNPs in the genome screen were derived from a various candidate genes and confirmed cSNPs. We mapped the positions of these SNPs against the NCBI's genome assembly build 30 from June 2002 (which is the same as Golden Path assembly version hg12). Out of the total 28,505 assays performed, 26,678 assays were mapped to the human genome map and 25,983 assays are uniquely mapped to the human genome only once as determined by proprietary algorithms of the eXTEND^{TM} software developed based on the ePCR program [58].

In order to carry out each disease-specific genome-scan a unique sample set of cases and matched controls were obtained from a clinical sample provider. DNA pools of these samples were constructed from individuals of uniform ethnicity, locality, age and gender. In addition to knowledge of the primary disease diagnosis of interest, additional and extensive phenotypic information on each case and control sample was obtained from each sample provider and stored in a database. This information served many purposes including a decision making tool as to which samples would be included or excluded in DNA pooling, as well as for secondary phenotype analysis. The number of samples obtained for each scan was sufficient to carry out our genetic association studies with enough power to detect differences in allele frequencies of common SNPs. Typically for each genome scan, between 200 and 400 individual case and control samples were utilized (see also Table 4).

We employed a multi-phased strategy to conduct each genome scan, which is summarized below. Equimolar quantities of DNA from each subject were pooled within each group. For diabetes and melanoma cohorts, four pools of DNA were constructed, separating females and males. For the other studies, only one case and one control DNA pool were constructed. In phase I, the allele frequency for each of the 28,000 SNPs in each DNA pool was estimated by allelotyping the DNA pools using a single PCR reaction and the MassEXTEND reaction as described in earlier sections of this article. SNPs were considered as being associated with diseases when allele frequency differences calculated between case and control pools, either male or female, were statistically significant ($p < 0.05$, Z-test). The significant associations from phase I were then subjected to another round of allelotyping, but this time triplicate PCRs were performed (phase II). SNPs maintaining statistically significant differences at this stage were then confirmed by genotyping (phase III) all subjects of each pool. Generally, this involved genotyping of approximately 100 SNPs per scan. This three-phased approach utilizing allelotyping (phases I and II) prior to genotyping (phase III) significantly reduced the number of reactions performed, which resulted in significant time and cost savings. Following a SNP's validation by genotyping in the discovery sample cohort, confirmation of a SNP's association was performed by genotyping in a second or replication cohort, which consisted of a new set of case and control samples that were not part of the discovery sample set. SNPs showing an association in this second sample set were considered replicated SNPs, providing further confidence that the particular SNP was indeed associated with the disease of interest.

As of today, 12 genome scans have been completed in the span of approximately 16 months, and many significant

Table 4
Select sample set statistics

	Sample source	Male control	Male case	Female control	Female case
Melanoma	German Caucasian	217 (47.3 ± 17.1)	236 (53.1 ± 14.8)	233 (47.6 ± 17.8)	266 (48.9 ± 16.5)
Breast cancer	German Caucasian			276 (56.2 ± 9.9)	272 (55.9 ± 13.7)
Type 2 diabetes	German Caucasian	254 (50.6 ± 7.1)	254 (50.0 ± 8.9)	244 (49.0 ± 6.3)	244 (52.5 ± 9.9)
HDL*	U.K. Caucasian			304 low HDL (46.55 ± 11.9)	295 high HDL (46.88 ± 11.9)
Osteoarthritis	U.K. Caucasian			335 (59.0 ± 4.8)	335 (58.4 ± 10.5)
Lung cancer	German Caucasian	287[a] (61.6 ± 13.2)	369 (64.6 ± 8.6)		

Depicted are the numbers of individuals used in each genome scan, with their respective race and ethnic backgrounds, and mean ages.

[a] Lung cancer control group consists of 196 males and 91 females.

associations have been found. Many of the confirmed associations identified from our genome scan analysis have been implicated directly or indirectly as candidate genes in their respective diseases/phenotypes. In the melanoma genome scan, a SNP present in intron 11 of the BRAF gene was significantly associated with melanoma in males ($p = 0.032$, submitted for publication). The BRAF gene encodes a serine/threonine kinase participating in the RAS/RAF MAP kinase signal transduction pathway. Davies et al. recently reported that somatic mutations in the BRAF gene occur in greater than 60% of human melanoma cases [59] while Polloack et al. reported that greater than 80% of nevi contain mutations in the BRAF gene [60]. In the breast cancer genome scan, one significant association identified was deleted in liver cancer 1 (DLC1) gene ($p = 0.018$, odds ratio = 0.6). The DLC1 gene is a rat Rho-GTPase-activating protein homologue and has been suggested as a candidate tumour suppressor gene for human liver cancer, as well as for prostate, lung, colorectal, and breast cancers [61]. Two SNPs on chromosome 22 in the seizure-related 6-like (SEZ6L) gene are significantly associated with lung cancer ($p = 0.10$ and 0.015). SEZ6L is a transmembrane protein functioning as an intracellular signal transducer via protein–protein interactions in a variety of human cells. Gene deletion at 22q12.1 near the SEZ6L gene has been reported in small cell lung cancer and non-small cell lung cancer, suggesting the presence of a tumor suppressor gene [62]. Our results point to the SNPs in SEZ6L being a predisposing factor of lung cancer. In the type II diabetes genome scan, we have identified significant associations with the peroxisome proliferator-activated receptor γ (PPARγ)($p = 0.057$ in males and $p = 0.056$ in females. PPARγ is a member of the nuclear hormone receptor subfamily of transcription factors involved in adipocyte differentiation and is a risk factor of type 2 diabetes confirmed by association studies [63]. For the high/low HDL genome scan study, we identified cholesteryl ester transfer protein (CETP, $p = 0.000009$). CETP gene polymorphisms are significantly associated with serum HDL levels and the progression of coronary atherosclerosis [57,64].

Genome-wide SNP analysis has been demonstrated as a powerful approach for mapping disease-associated genes [65–70]. It has been estimated that the number of SNP markers required for such studies will likely be between 30,000 and 600,000, depending on the linkage disequilibrium (LD). In the very near future the HapMap project will establish the LD across the entire human genome, providing a roadmap as to the locations and numbers of SNPs required to efficiently perform a whole genome scan. The results presented here demonstrate that a SNP-based genome-wide association study is feasible, particularly when allelotyping (DNA pooling) is applied to minimize the time and cost involved in performing this type of study. The discovery of candidate genes from each of the genome scans completed to date clearly validates the strategy employed for the mapping of multigenic disease loci. The genome scans described above have been completed in less than 1 and 1/2 year, with reagent costs of only a fraction of costs of individual genotyping. Genome scan association studies using allelotyping is a cost-effective high-throughput approach to discovering SNPs associated with complex diseases.

References

[1] S. Broder, J.C. Venter, Curr. Opin. Biotechnol. 11 (2000) 581.

[2] R. Bomprezzi, M. Ringner, S. Kim, M.L. Bittner, J. Khan, Y. Chen, A. Elkahloun, A. Yu, B. Bielekova, P.S. Meltzer, R. Martin, H.F. McFarland, J.M. Trent, Hum. Mol. Genet. 12 (2003) 2191.

[3] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, Nature 406 (2000) 536.

[4] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Science 286 (1999) 531.

[5] A. Pandey, M. Mann, Nature 405 (2000) 837.

[6] R. Aebersold, M. Mann, Nature 422 (2003) 198.

[7] B. Schweitzer, P. Predki, M. Snyder, Proteomics 3 (2003) 2190.

[8] S.L. Wu, H. Amato, R. Biringer, G. Choudhary, P. Shieh, W.S. Hancock, J. Proteome Res. 1 (2002) 459.

[9] S. Hanash, Nature 422 (2003) 226.

[10] D. Botstein, N. Risch, Nat. Genet. 33 (Suppl.) (2003) 228.

[11] C.F. Sing, J.H. Stengard, S.L. Kardia, Arterioscler. Thromb. Vasc. Biol. 23 (2003) 1190.

[12] L. Melton, Nature 422 (2003) 917.

[13] A. Braun, D.P. Little, H. Koster, Clin. Chem. 43 (1997) 1151.

[14] L.A. Haff, I.P. Smirnov, Biochem. Soc. Trans. 24 (1996) 901.

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

[16] D.P. Little, A. Braun, B. Darnhofer-Demar, H. Koster, Eur. J. Clin. Chem. Clin. Biochem. 35 (1997) 545.

[17] P. Ross, L. Hall, I. Smirnov, L. Haff, Nat. Biotechnol. 16 (1998) 1347.

[18] K. Tang, D.J. Fu, D. Julien, A. Braun, C.R. Cantor, H. Koster, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 10016.

[19] K. Tang, D. Opalsky, K. Abel, D. van den Boom, P. Yip, G. Mistro, A. Braun, C.R. Cantor, Int. J. Mass Spectrom. 226 (2002) 37.

[20] N. Storm, B. Darnhofer-Patel, D. van den Boom, C.P. Rodi, Methods Mol. Biol. 212 (2003) 241.

[21] C.P. Rodi, B. Darnhofer-Patel, P. Stanssens, M. Zabeau, D. van den Boom, Biotechniques Suppl. (2002) 62.

[22] E. Nordhoff, R. Cramer, M. Karas, F. Hillenkamp, F. Kirpekar, K. Kristiansen, P. Roepstorff, Nucleic Acids Res. 21 (1993) 3347.

[23] D.P. Little, A. Braun, M.J. O'Donnell, H. Koster, Nat. Med. 3 (1997) 1413.

[24] L.A. Haff, I.P. Smirnov, Nucleic Acids Res. 25 (1997) 3749.

[25] B.D. Cobb, J.M. Clarkson, Nucleic Acids Res. 22 (1994) 3801.

[26] S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, D. Altshuler, Science 296 (2002) 2225.

[27] M. Olivier, L.M. Chuang, M.S. Chang, Y.T. Chen, D. Pei, K. Ranade, A. de Witte, J. Allen, N. Tran, D. Curb, R. Pratt, H. Neefs, M. de Arruda Indig, S. Law, B. Neri, L. Wang, D.R. Cox, Nucleic Acids Res. 30 (2002) e53.

[28] R.W. Nelson, M.A. McLean, T.W. Hutchens, Anal. Chem. 66 (1994) 1408.

[29] P. Ross, L. Hall, L.A. Haff, Biotechniques 29 (2000) 620.

[30] P. Sham, J.S. Bader, I. Craig, M. O'Donovan, M. Owen, Nat. Rev. Genet. 3 (2002) 862.

[31] B.J. Barratt, F. Payne, H.E. Rance, S. Nutland, J.A. Todd, D.G. Clayton, Ann. Hum. Genet. 66 (2002) 393.

[32] C. Jurinke, P. Oeth, D. van den Boom, Mol. Biotechnol. (in press).

[33] S. Le Hellard, S.J. Ballereau, P.M. Visscher, H.S. Torrance, J. Pinson, S.W. Morris, M.L. Thomson, C.A. Semple, W.J. Muir, D.H. Blackwood, D.J. Porteous, K.L. Evans, Nucleic Acids Res. 30 (2002) 74.

[34] S. Shifman, A. Pisante-Shalom, B. Yakir, A. Darvasi, Mol. Cell Probes. 16 (2002) 429.

[35] K.H. Buetow, M. Edmonson, R. MacDonald, R. Clifford, P. Yip, J. Kelley, D.P. Little, R. Strausberg, H. Koester, C.R. Cantor, A. Braun, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 581.

[36] K.L. Mohlke, M.R. Erdos, L.J. Scott, T.E. Fingerlin, A.U. Jackson, K. Silander, P. Hollstein, M. Boehnke, F.S. Collins, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 16928.

[37] M. Werner, M. Sych, N. Herbon, T. Illig, I.R. Konig, M. Wjst, Hum. Mutat. 20 (2002) 57.

[38] J.C. Knight, B.J. Keating, K.A. Rockett, D.P. Kwiatkowski, Nat. Genet. 33 (2003) 469.

[39] G. Amexis, P. Oeth, K. Abel, A. Ivshina, F. Pelloquin, C.R. Cantor, A. Braun, K. Chumakov, A. Braun, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 12097.

[40] C. Ding, C.R. Cantor, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 3059.

[41] J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, F. Speleman, Genome Biol. 3 (2002) RESEARCH0034.

[42] A. Darvasi, Nature 422 (2003) 269.

[43] E.E. Schadt, S.A. Monks, T.A. Drake, A.J. Lusis, N. Che, V. Colinayo, T.G. Ruff, S.B. Milligan, J.R. Lamb, G. Cavet, P.S. Linsley, M. Mao, R.B. Stoughton, S.H. Friend, Nature 422 (2003) 297.

[44] L.M. Smith, Science 262 (1993) 530.

[45] E. Nordhoff, C. Luebbert, G. Thiele, V. Heiser, H. Lehrach, Nucleic Acids Res. 28 (2000) E86.

[46] H. Koster, K. Tang, D.J. Fu, A. Braun, D. van den Boom, C.L. Smith, R.J. Cotter, C.R. Cantor, Nat. Biotechnol. 14 (1996) 1123.

[47] F. Kirpekar, E. Nordhoff, L.K. Larsen, K. Kristiansen, P. Roepstorff, F. Hillenkamp, Nucleic Acids Res. 26 (1998) 2554.

[48] N.I. Taranenko, S.L. Allman, V.V. Golovlev, N.V. Taranenko, N.R. Isola, C.H. Chen, Nucleic Acids Res. 26 (1998) 2488.

[49] S. Krebs, I. Medugorac, D. Seichter, M. Forster, Nucleic Acids Res. 31 (2003) e37.

[50] R. Hartmer, N. Storm, S. Boecker, C.P. Rodi, F. Hillenkamp, C. Jurinke, D. van den Boom, Nucleic Acids Res. 31 (2003) e47.

[51] P. Stanssens, M. Zabeau, G. Meersseman, G. Remes, Y. Gansemans, N. Storm, R. Hartmer, C. Honisch, C.P. Rodi, S. Bocker, D. van den Boom, Genome Res. 14 (2004) 126.

[52] S. Bocker, Bioinformatics 19 (Suppl. 1) (2003) I44.

[53] K. Tang, D. Opalsky, K. Abel, D. van den Boom, P. Yip, G. Del Mistro, A. Braun, C.R. Cantor, Int. J. Mass Spectrom. 226 (2003) 37.

[54] S. Berkenkamp, F. Hillenkamp, D. van den Boom, ASMS, Montreal, Canada, 2003.

[55] F. von Wintzingerode, S. Bocker, C. Schlotelburg, N.H. Chiu, N. Storm, C. Jurinke, C.R. Cantor, U.B. Gobel, D. van den Boom, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 7039.

[56] M. Lefmann, C. Honisch, S. Bocker, N. Storm, F. von Wintzingerode, C. Schlotelburg, A. Moter, D. van den Boom, U.B. Gobel, J. Clin. Microbiol. 42 (2004) 339.

[57] A. Bansal, D. van den Boom, S. Kammerer, C. Honisch, G. Adam, C.R. Cantor, P. Kleyn, A. Braun, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 16871.

[58] G.D. Schuler, Genome Res. 7 (1997) 541.

[59] H. Davies, G.R. Bignell, C. Cox, P. Stephens, S. Edkins, S. Clegg, J. Teague, H. Woffendin, M.J. Garnett, W. Bottomley, N. Davis, E. Dicks, R. Ewing, Y. Floyd, K. Gray, S. Hall, R. Hawes, J. Hughes, V. Kosmidou, A. Menzies, C. Mould, A. Parker, C. Stevens, S. Watt, S. Hooper, R. Wilson, H. Jayatilake, B.A. Gusterson, C. Cooper, J. Shipley, D. Hargrave, K. Pritchard-Jones, N. Maitland, G. Chenevix-Trench, G.J. Riggins, D.D. Bigner, G. Palmieri, A. Cossu, A. Flanagan, A. Nicholson, J.W. Ho, S.Y. Leung, S.T. Yuen, B.L. Weber, H.F. Seigler, T.L. Darrow, H. Paterson, R. Marais, C.J. Marshall, R. Wooster, M.R. Stratton, P.A. Futreal, Nature 417 (2002) 949.

[60] P.M. Pollock, U.L. Harper, K.S. Hansen, L.M. Yudt, M. Stark, C.M. Robbins, T.Y. Moses, G. Hostetter, U. Wagner, J. Kakareka, G. Salem, T. Pohida, P. Heenan, P. Duray, O. Kallioniemi, N.K. Hayward, J.M. Trent, P.S. Meltzer, Nat. Genet. 33 (2003) 19.

[61] B.Z. Yuan, M.J. Miller, C.L. Keck, D.B. Zimonjic, S.S. Thorgeirsson, N.C. Popescu, Cancer Res. 58 (1998) 2196.

[62] M. Nishioka, T. Kohno, M. Takahashi, T. Niki, T. Yamada, S. Sone, J. Yokota, Oncogene 19 (2000) 6251.

[63] 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.

[64] J.A. Kuivenhoven, J.W. Jukema, A.H. Zwinderman, P. de Knijff, R. McPherson, A.V. Bruschke, K.I. Lie, J.J. Kastelein, N. Engl. J. Med. 338 (1998) 86.

[65] L. Kruglyak, Nat. Genet. 22 (1999) 139.

[66] J.C. Stephens, J.A. Schneider, D.A. Tanguay, J. Choi, T. Acharya, S.E. Stanley, R. Jiang, C.J. Messer, A. Chew, J.H. Han, J. Duan, J.L. Carr, M.S. Lee, B. Koshy, A.M. Kumar, G. Zhang, W.R. Newell, A. Windemuth, C. Xu, T.S. Kalbfleisch, S.L. Shaner, K. Arnold, V. Schulz, C.M. Drysdale, K. Nandabalan, R.S. Judson, G. Ruano, G.F. Vovis, Science 293 (2001) 489.

[67] N. Patil, A.J. Berno, D.A. Hinds, W.A. Barrett, J.M. Doshi, C.R. Hacker, C.R. Kautzer, D.H. Lee, C. Marjoribanks, D.P. McDonough, B.T. Nguyen, M.C. Norris, J.B. Sheehan, N. Shen, D. Stern, R.P. Stokowski, D.J. Thomas, M.O. Trulson, K.R. Vyas, K.A. Frazer, S.P. Fodor, D.R. Cox, Science 294 (2001) 1719.

[68] A.K. Daly, C.P. Day, Br. J. Clin. Pharmacol. 52 (2001) 489.

[69] R. Judson, B. Salisbury, J. Schneider, A. Windemuth, J.C. Stephens, Pharmacogenomics 3 (2002) 379.

[70] E. Lai, C. Bowman, A. Bansal, A. Hughes, M. Mosteller, A.D. Roses, Nat. Genet. 32 (2002) 353.