ELSEVIER

International Journal of Mass Spectrometry 226 (2003) 37–54

International Journal of Mass Spectrometry

www.elsevier.com/locate/ijms

Review

Single nucleotide polymorphism analyses by MALDI-TOF MS

Kai Tang*, David Opalsky, Ken Abel, Dirk van den Boom, Ping Yip, Guy Del Mistro, Andreas Braun, Charles R. Cantor

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

Received 31 August 2001; accepted 20 July 2002

Abstract

Matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) mass spectrometry (MS) has been used to detect single nucleotide polymorphisms (SNPs). Methods of generating allelic-specific DNA fragments to facilitate MALDI analysis have been employed. Interactive data acquisition has been developed to ensure high data quality for multiplexed assays, making possible ultra-high throughput, real-time genotyping. DNA pooling strategies combined with MALDI-MS have been used for SNP validation and to estimate allelic frequencies. The impact of these strategies on genome-wide association studies and phamacogenomics are discussed.

Keywords: Phamacogenomics; Genotyping; Time-of-flight

1. Introduction

MALDI-TOF mass spectrometry is now being used for analysis of nucleic acids [1–3], including genetic variations such as microsatellites [4–6], insertion/deletions [7,8], and, especially, single nucleotide polymorphisms (SNPs) [9–13]. The output data is a measure of an intrinsic characteristic of the DNA products being studied (molecular weight in Daltons); no indirect measurement of the products is involved, as with fluorescent or radiolabel tagging. The ability to resolve oligonucleotides varying in mass by less than a single nucleotide makes MALDI-TOF mass spectrometry an excellent platform for SNP and mutation analysis.

A highly automated processing platform incorporating MALDI-TOF mass spectrometry, designated DNA MassARRAY™, has been developed [14–17]. DNA MassARRAY™ uses samples in chip-based, high-density arrays. This system accurately calls SNPs in individual DNA samples, or alternatively determines SNP allele frequencies in DNA pools. Assay design for MassARRAY™ is simple, flexible and has been automated to allow designing vast numbers of assays, all of which can be run using a universal set of reaction conditions.

DNA MassARRAY™ analysis of SNPs incorporates region-specific amplification of genomic DNA spanning the genetic variation, selection of one PCR product strand to serve as a template for a variant-specific primer extension (usually 1–2 bases), dispensing of nanoliter quantities of extension products

* Corresponding author.
E-mail address: ktang@sequenom.com (K. Tang).

1387-3806/02/$ – see front matter
PII S1387-3806(02)00980-6

onto a matrix-loaded chip array, and product analysis by MALDI-TOF mass spectrometry. Spectra are interpreted automatically according to input assay design parameters to identify the variant DNAs (alleles or mutants).

DNA MassARRAY™ offers several advantages for analysis of DNA. First, MALDI-TOF mass spectrometry allows the separation and detection of a mixture of large biomolecules in seconds without the need for labels or internal standards. The resultant mass signals provide absolute information on an inherent molecular property (molecular weight), thus differentiating MALDI-TOF from other detection methods that are indirect. Second, a test strategy (MassEXTEND™) has been designed to reveal only the essential information relevant to the particular assay being performed. Third, the use of miniaturized chip-based sample preparations conserves samples and also enables automated scanning of numerous samples on one array in the mass spectrometer. This combination of high throughput with high definition signals is a dramatic improvement over other SNP genotyping technologies, which have such drawbacks as prohibitive speed and cost (gel-based methods), requirements for specific temperatures or reagents for specific hybridization, need for expensive labels or dyes which provide only indirect measurements of diagnostic DNAs, problems with accuracy, and low flexibility for multiplexing (detecting multiple SNPs in a single assay) or for developing new SNP assays.

2. SNP genotyping methods using MALDI-MS detection

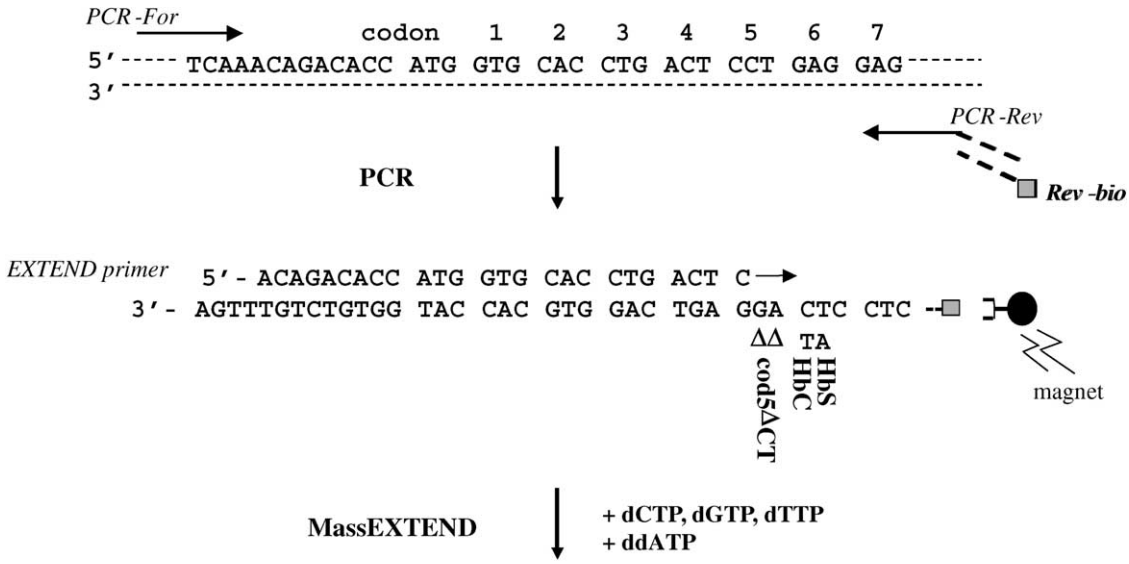
2.1. Standard (solid-phase) MassEXTEND™ (sME) assay

Reactions employed in DNA MassARRAY™ are performed as illustrated in Fig. 1. The particular assay shown is capable of distinguishing between the normal allele and several codon 5 and 6 mutations in the human β-globin gene. Initially, genomic sequences containing polymorphisms or mutations are PCR-amplified in a reaction including two sequence-specific primers (*PCR-For* and *-Rev*), plus a biotinylated and artificial sequence primer (*Rev-bio*). This primer, whose sequence is also present at the 5′ end of one PCR primer, is included in large molar excess so that virtually all of the amplified products contain the biotin tag. Since the same *Rev-bio* primer can be used for all assays, the cost of the biotinylated primer per individual assay is negligible. PCR products are captured using streptavidin-coated magnetic beads in the presence of a magnet. Following DNA denaturation, the captured strand serves as the template for an optimized primer extension reaction (MassEXTEND™), generating allele-specific oligonucleotide products [18]. This reaction includes a primer that binds adjacent to the polymorphism, a DNA polymerase, and a mixture of deoxy- and dideoxy-NTPs. When a ddNTP is present in the mixture, the same dNTP is absent. Chain termination results from incorporation of a ddNTP as dictated by the sequence surrounding the polymorphism, generating products of different lengths and masses. Depending on the local sequence context, the primer extension assays are typically designed so the products terminate after extension by either one or two nucleotides. The diagnostic primer-extension products are washed to eliminate enzymes and buffer salts and then denatured to remove them from the immobilized template. The resulting supernatant is used for MALDI-TOF analysis.

The advantage of the solid-phase assay is that the primer-extension products can be purified easily by washing the magnetic beads. However, the amount of the diagnostic products is limited by the amount of the immobilized template since any non-hybridized products are washed away with buffer salts. This assay also has the disadvantage of the added cost of streptavidin beads.

2.2. Homogeneous MassEXTEND™ (hME) assay

The principle design of the hME assay is the same as that of the solid-phase assay. However, it is a single-tube reaction carried out in solution without using any immobilization step [17]. Therefore, no additional biotinylated primer is required during the



			bases	mass
	primer	ACAGACACCATGGTGCACCTGACTC	25	7596
EXTEND products:	HbA	ACAGACACCATGGTGCACCTGACTC *CTGA*	29	8816
	HbS	ACAGACACCATGGTGCACCTGACTC *CTGTGGA*	32	9778
	HbC	ACAGACACCATGGTGCACCTGACTC *CTA*	28	8487
	cod5ΔCT	ACAGACACCATGGTGCACCTGACTC *GA*	27	8222

Fig. 1. DNA MassARRAY^{TM} assay for distinguishing the normal β-globin allele and mutant alleles in several hemoglobinopathies including HbS (sickle cell) and HbC (both in codon 6), and a −CT deletion in codon 5. The first ATG codon is not counted because the initiator methionine is processed from the mature polypeptide. All MassEXTEND products are terminated by differential incorporation of dideoxy-ATP. Extend product masses are shown in Daltons.

PCR. After the PCR, a dephosphorylation step with shrimp alkaline phosphatase (SAP) is used to destroy the remaining free dNTPs in the solution. Primer extension reaction (hME) is then performed in 40 cycles of denaturation, primer annealing and primer extension, by varying the temperature. The products are desalted by adding a suspension of cation exchange beads (NH_4^+ form) [19]. After a brief centrifugation, the supernatant is used for MALDI-TOF analysis.

Since hME requires only liquid addition steps throughout the whole procedure, it is easily compatible with automated liquid handling stations and thermal cyclers currently available. The amount of products can also exceed the amount of template since the hME cycling amplifies the products linearly and they remain in the solution. Since hME eliminates the use of expensive streptavidin-coated magnetic beads, the cost is lower. However, the desalting step now becomes critical. The purity of the products directly affects the success of MALDI-MS.

2.3. Multiplexing

The typical mass range of primer-extension products is between 5000 and 10,000 Da, corresponding to 17–33 nt in length. This provides a wide window for multiplexing since MassEXTEND^{TM} primers targeting different SNPs can be chosen so that all extended products and primers do not overlap in the mass spectrum. The mass separation among peaks of uniplex

reactions is at least one base (∼300 Da). However, due to the high accuracy of TOF mass spectrometers (∼0.1% in linear mode), smaller mass difference can be unambiguously discriminated without running into the risk of peak misidentification. Normally neighboring peaks with mass difference of 50 Da are well separated in linear TOF instruments (requiring resolution of only 100–200) and require a mass accuracy of only 0.5–1% for identification. Fifty Dalton is therefore used as the minimum requirement for peak separations in multiplex reactions.

Multiplexing can also start at the PCR level, which requires more careful primer design. Automatic assay design software has been created to address the need for reliability and optimization simultaneously. This will be discussed in Section 3.1. The need for real-time quality control for all assays in the mass spectrum of a multiplexed genotyping reaction has also prompted the development of real-time data acquisition software, which will be discussed in Section 3.4.

Compared to gel electrophoresis, which lacks in resolution, and fluorescent detection, which has limited available wavelengths, MALDI mass spectrometry provides high resolution, high accuracy and wide mass range for designing highly multiplexed genotyping assays. Normally, five-fold multiplexing can be routinely carried out with assay optimization and 10-fold multiplexing is possible.

2.4. Allele frequency analysis

As described above, DNA MassARRAY™ has great utility for highly accurate uniplex or multiplex genotyping of individual DNA samples. Pilot studies have suggested this approach can also be used to estimate SNP allele frequencies in pooled DNA samples [20,21]. This derives from the quantitative nature of MALDI mass spectrometry, such that allele frequencies are proportional to mass spectral peak areas.

2.4.1. Quantitative MALDI

It has been known that with conventional dried-droplet sample preparations, the distribution of 3-HPA matrix crystals throughout the sample spot area is extremely uneven. This makes it necessary to search for a hot spot to obtain a satisfactory mass spectrum. Mainly for this reason the signal intensity does not always represent absolute sample quantity, making quantitative comparisons between spectra difficult. However, based on the observation that the mixing of the same kind of analytes in a given sample is homogeneous, the ratio of peak areas within a mass spectrum is representative of the ratio of analyte quantities [22]. This provides an approach for relative quantitation that can be applied to allele frequency determination in MALDI mass spectrometry.

2.4.2. Pooling strategy

While sample processing is essentially the same as for genotyping, an additional level of information is possible through interpretation of mass spectra obtained from pooled samples. The general concept for estimating allele frequencies based on the pooled samples is relatively simple. Following automated peak recognition as for genotyping, an evaluation of the areas under those peaks is performed. For each spectrum acquired, peak areas corresponding to expected MassEXTEND products are automatically normalized relative to baseline and integrated. Within a given spectrum, allele frequencies are then estimated as the proportion of either peak area relative to the total area for both expected peaks. Frequency estimates, as well as estimation of error associated with these frequencies, are now done in an automated manner using SpectroTYPER™ software (see Sections 3.4.2 and 3.4.3). Fig. 2 presents example spectra, showing clear allele frequency differences for a SNP in the ICAM-1 gene among Californians generally of African, Asian, Caucasian, and Hispanic descent. Knowledge of such differences may be key to planning ethnicity-based case and control association studies, to minimize the chance of false positive associations that can result from analyzing unstratified datasets.

Preliminary studies have shown that allele frequency estimates obtained from pooled DNAs can be very accurate, comparable to the actual frequencies determined by genotyping individual DNAs. An example is shown in Fig. 3, using data for a codon 353

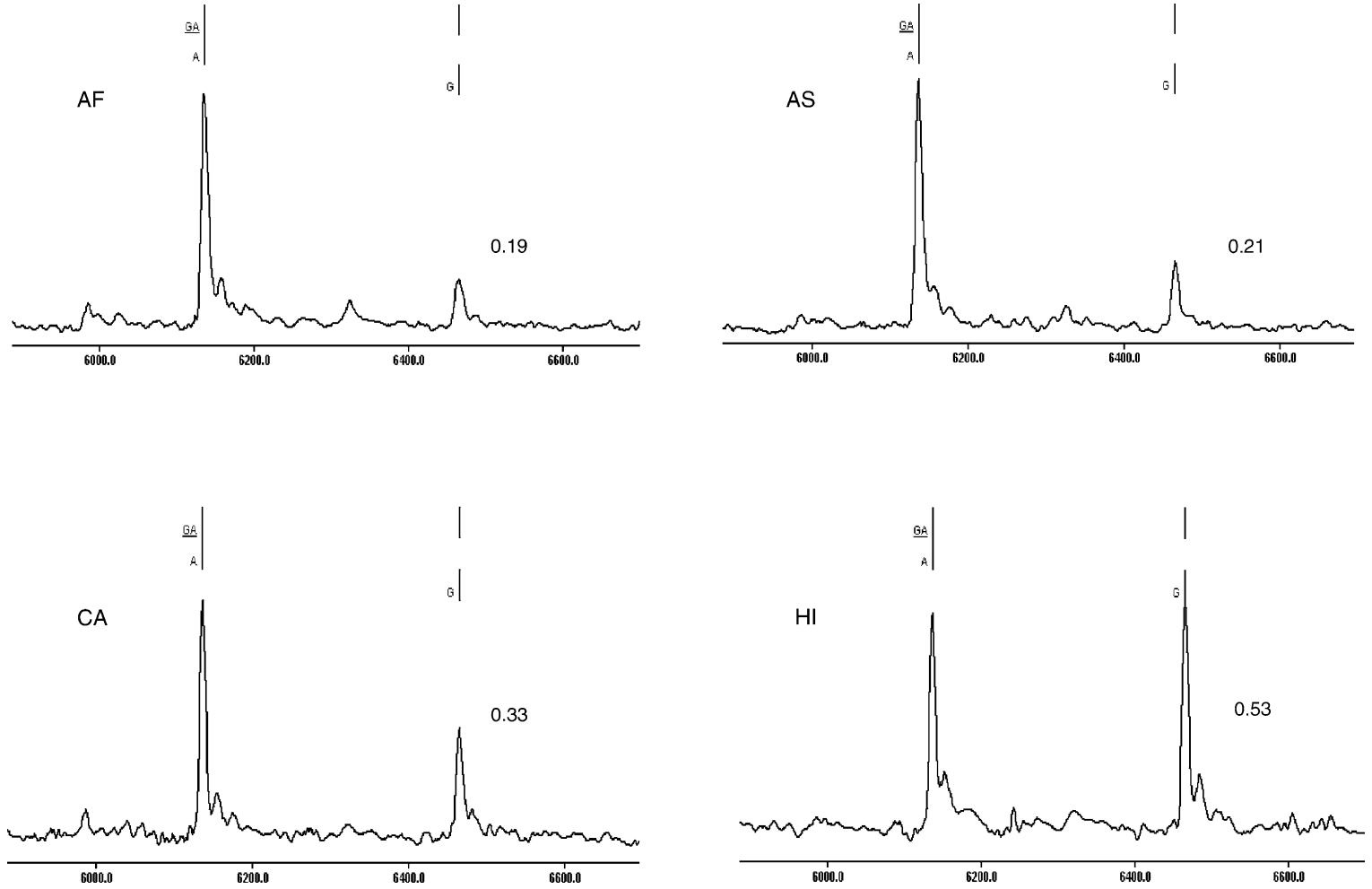


Fig. 2. Comparison of allele frequencies of an ICAM-1 SNP, obtained from DNA pools representing several broad, California-based ethnic groups. (AF, African-American; AS, Asian-American; CA, Caucasian-American; HI, Hispanic-American.)

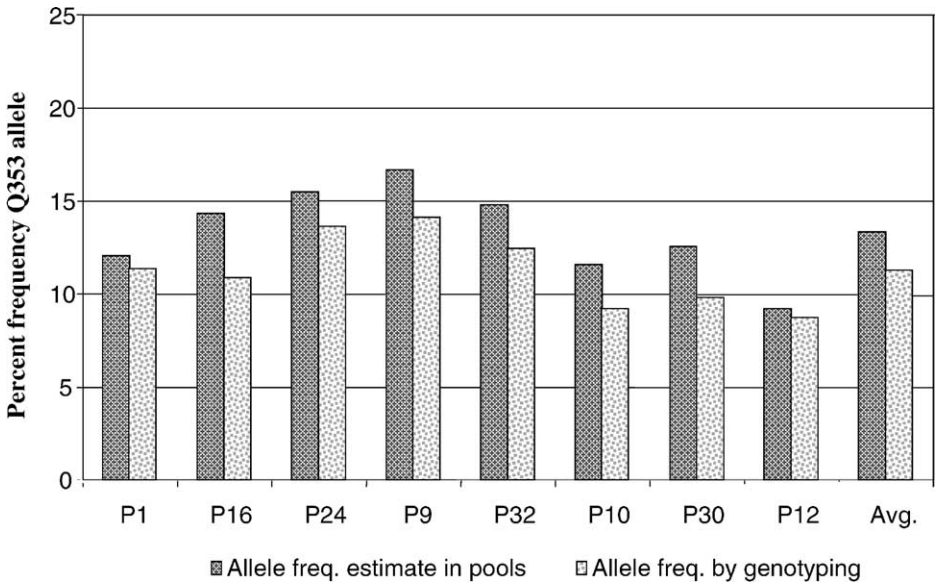


Fig. 3. Comparison of allele frequencies for the Factor VII R353Q SNP among eight different pools, each containing DNAs from nearly 100 Caucasian individuals, by pooled sample analysis and genotyping.

SNP in the clotting Factor VII gene (Arg353Gln). Frequencies for the minor allele (Q353) were estimated in eight different pools comprised of nearly 100 Caucasian individuals, and compared with frequencies determined by genotyping the individual DNAs represented in those pools. The results were extremely similar. Actual frequencies (genotyping) for Q353 in each pool averaged 11.3% (range 9–14%). By comparison, the estimated frequencies (pooling) averaged 13.3% (range 9–16%). With regards to accuracy, the differences within a pool between actual and estimated frequencies averaged only 2%, with no difference greater than 3.5%. It should be noted the differences in estimated frequencies within a pool were in fact much less than the actual differences seen between pools, presumably attributable to random sampling.

3. Automation for high throughput analysis

3.1. Assay design

The process of assay design, as rendered by the SpectroDESIGNER^{TM} software, presents two challenges. First, the design of PCR and extend primers will reliably amplify and interrogate a specific SNP site and the optimal multiplexing of assays is possible such that no single assay will fail due to competing kinetics or cross-hybridization reactions. Second, the products (analytes and by-products) are well resolved in the resulting mass spectra.

Primers are designed with respect to target sequences of a given SNP strand such that the length must be between user-specified limits (e.g., 17–24 bases) and must not contain any bases that are uncertain in the target sequence. The hybridization strength is gauged by calculating the sequence-dependent melting (or hybridization/dissociation) temperature, T_m. A particular primer choice may be disallowed, or penalized relative to other choices of primers, because of its hairpin potential, false priming potential, primer–dimer potential, and problematic subsequences such as GGGG.

For extend primer design there are only two choices of primer sequence, adjacent to the SNP site on either side. These primers may be of various lengths, and each of these primer choices is scored using general

primer design concerns and additional factors which are dependent on the set of extension products that result from employing each of the terminator mixes available. Generally, it is rare for SNP targets to fail at extend primer design on both sides of the SNP. The scores are mainly used to choose between alternative extend primer choices and preferred terminator mixes. However, particular primer lengths for extend primers may not be suitable because of mass conflicts with contaminants (e.g., biotin-tags) or by-products (e.g., depurination products or, possibly, secondary extend-pausing products that are terminated by a dNTP instead of ddNTP). For example, an extension primer prematurely terminated with dA would have exactly the same mass as if normally terminated with ddG, and therefore these products are indistinguishable. To avoid miscalls in genotypes, such designs should not be chosen.

For PCR primer design many more primers need to be evaluated but there is a greater chance of finding pairs of primers that satisfy optimal design parameters. Each potential primer is scored with respect to an optimal length (20 bases), an optimal T_m (60 °C) and an optimal G–C content (50%). Only primer pairs that would produce an amplicon length that satisfies the user-supplied minimum and maximum amplicon length bounds are suitable. The normalized combination of these scoring components is referred to as the *uniplex PCR confidence score* and is recorded in the output for successful assay designs.

For multiplexed assays, the same general guideline for primer design applies such as avoiding false priming and primer dimers, only now more primers are involved. In addition, the analyte peaks in the mass spectra for one assay must be sufficiently well resolved from any product of any assay it is multiplexed with, including pausing peaks and any other user-specified by-product peaks. In addition, analyte peaks must fall within the user-specified mass window. Efforts are also made to ensure that strong assays are not multiplexed with weak assays to avoid the respective analyte peaks appearing too unbalanced in the mass spectra.

3.2. Sample preparation and transfer

All samples are processed in microtiter plates (MTPs, 96- or 384-well). The PCRs are performed in either a PTC-225 DNA Engine Tetrad™ thermo cycler (MJ Research, Watertown, MA) or a GeneAmp PCR System 9700 (Applied Biosystems, Foster City, CA). Primer extension assays (either sME or hME) are set-up using an automated 96-channel pipette station (Multimek, Beckman Coulter, Fullerton, CA; also available through SEQUENOM as SpectroPREP™ with customized add-ons and programs).

The primer extension products in each MTP can then be transferred to a silicon chip using different types of nano-dispensing devices. A four channel piezoelectric pipette (SpectroJET™, SEQUENOM) is often used for both 96- and 384-well MTP [14]. These piezoelectric tips are capable of dispensing sub-nanoliter volume per drop and are calibrated individually to dispense the same amount of analyte. The SpectroJet™ is programmed to aspirate in parallel 1 μL of analyte from the MTP and dispense 14 nL in serial onto the corresponding pads on the silicon chip. The silicon chip (SpectroCHIP™ from SEQUENOM) has a highly hydrophobic surface with 96 or 384 hydrophilic pads in an array format preloaded with 3-hydroxypicolinic acid (3-HPA) matrix [23]. When dispensed, the aqueous analyte solution partially re-dissolves the matrix and the droplet quickly shrinks toward the hydrophilic pad and re-crystallize with the matrix within the 200 μm × 200 μm pad. We have found that the re-crystallized samples consistently yield better MALDI performance. The advantage of the piezoelectric pipette is the nature of non-contact dispensing which does not disturb the matrix crystals on the chip. However, the drawbacks of being sensitive to liquid viscosity (different piezo voltages are needed when the viscosity changes) and the relatively low throughput (∼70 min for 384-well dispensing) make the SpectroJet™ inappropriate for ultra-high throughput genotyping.

Nanoliter sample transfer can also be performed using pintool devices (SpectroPOINT™, SEQUENOM). In order to avoid destruction of matrix crystals

upon contact dispensing, slot pins with openings bigger than 200 μm are used so that when the pins touch the chip surface, the matrix spot at center of the slot can be spared. The volume of liquid delivery has been found to be linear with the pin velocity upon contact, therefore can be calibrated and precisely controlled. Since the pins can be mounted in an array format precisely with 4500 ± 20 μm spacing, the pintool can be used to dispense analyte in parallel. By using 24 pins in a 4 × 6 array, the SpectroPOINT™ is able to transfer 15 nL from a 384-well MTP to a 384-well chip in 12 min.

The advantage of using nanoliter sample preparation for MALDI mass spectrometry is that the miniaturized sample spot (200 μm × 200 μm) provides more homogeneous sample distribution relative to the laser spot ($d \approx$ 50–100 μm). Searching for hot spots during MALDI-MS of such samples is usually not necessary.

3.3. MALDI-TOF-MS

Most commercial TOF mass spectrometers can be modified to analyze SpectroCHIPs™. This can be achieved by modifying the existing sample target to hold at least one chip, having a precise XY-stage, a customized geometry file that maps the array positions, and a sample vision system to monitor and align the array of spots. Data acquisition is usually controlled by a customized software developed by SEQUENOM and is discussed in Section 3.4. The instrument parameters are usually set on the local mass spectrometer using the instrument control software provided. These parameters are usually the same as those used for analysis of oligonucleotides with a typical dried-droplet sample preparation using 3-HPA matrix. Typically, parameters used on the Bruker Biflex III TOF-MS (linear mode) are; accelerating voltage of +20 kV, P2 lens, +18.9 kV, focusing lens voltage of 9.4 kV, and long delay >600 ns. The Biflex III is equipped with a LeCroy Waverunner digitizer; 6000 spectral points are acquired with sample bin set at 5 ns. The bandwidth-limiting filter (BWL) is on, corresponding to an input bandwidth of ∼250 MHz in order to provide hardware smoothing of the mass spectra. The detector is gated so that ions below mass 2500 Da do not saturate the detector. The focal position of the nitrogen laser is adjusted to the surface of the sample. For high throughput analysis, a 20 Hz or faster laser and a high speed XY-table (e.g., 25 mm/s top speed) are necessary. The reliability and reproducibility of the XY-table is essential.

3.4. Interactive data acquisition

The data acquisition and analysis unit, SpectroTYPER RT™ (SEQUENOM, Inc., San Diego, CA), is a MALDI-based genotyping system that executes in real-time the signal processing and genotyping algorithms first used in its predecessor, SpectroTYPER™. The SpectroTYPER RT™ system uses biology-based results to control data acquisition in the mass spectrometer, thereby significantly improving call efficiency and increasing the instrument throughput.

Commercial MALDI mass spectrometers typically are capable of performing automated measurements on a series of samples. Standard software packages that enable automation include integrated algorithms that are used to judge the quality of the spectra. Such algorithms assess parameters such as the signal-to-noise ratio, peak resolution, and/or signal intensity within a specified mass range. If an acquired spectrum is determined to be of low quality, the instrument parameters may be adjusted and/or the stage may be moved (rastered) to another section of the sample for re-acquisition of the spectrum. The cycle of evaluation and re-acquisition is repeated until either a spectrum of sufficient quality is acquired or a pre-specified number of acquisition attempts have been made. The spectrum is then saved and the system moves on to the next sample. In these systems, the integrated judging algorithms make their determination based on qualities of the spectra that are independent of the underlying assay or biological information contained in the spectra. During an automated run, a spectrum for each sample is stored. Then special purpose algorithms are employed that automatically determine the sample genotype. SpectroTYPER™ is an automated

data processing system that determines one or more genotypes in each sample, depending on the assay definition for that sample, and assigns each a quality that is, from best to worst—conservative, moderate, aggressive, low probability, or bad spectrum.

In the system just described, a combination of automated data collection routines and automated data processing routines, two different sets of criteria are used to judge the spectra; one set of criteria is used to control the data acquisition process, and a separate set is used to determine the biological significance of the acquired spectrum. Using such two-step acquisition and analysis routines, however, may result in missed calls and unnecessarily long acquisition times. This is because the spectral features that define a clean acquisition are not necessarily the same features required for accurate genotyping. For example, the presence of large primer peaks due to incomplete extension may render a spectrum acceptable in terms of signal-to-noise criteria in a predefined mass window, but the resulting spectrum might not be of sufficient quality to allow determination of an unambiguous genotype. It is also possible that a spectrum that is of high quality for genotyping has a signal-to-noise ratio that causes repeated sampling by the data collection algorithm. In this case, unneeded data would be collected with a corresponding decrease in throughput. When different criteria are used for data collection and for data analysis it will always be possible that either the data collected does not give a suitable biological result or that extra data is collected resulting in lower throughput. Furthermore, the mismatches between the two judging methods become more common as the spectra from a sample become more complex, as with highly multiplexed samples. Integration of the data analysis and data collection algorithms should therefore result in faster, more accurate MALDI genotyping. A potential problem with a system that runs biology-based signal processing in real-time is throughput. The assay-based algorithms can take a significant amount of time to run. In order to have acceptable performance and actually realize the possible throughput advantages from biology-based instrument control, it is necessary to optimize the algorithms and to optimize the hardware platform that the algorithms run on.

The SpectroTYPER RT™ system is a modification of the SpectroTYPER™ system that includes highly optimized versions of the calling algorithms with a streamlined interface to a database to store the genotyping results. As part of the optimization, a well-defined programming interface was developed that controls the dialogue between the data acquisition component and the biological-calling component of spectra analysis. The interface is flexible and modular to allow modification of the calling algorithms.

3.4.1. Instrumentation

A Bruker Biflex instrument was modified to include the real-time genotype calling capabilities of SpectroTYPER RT™. The modifications include the installation of a SpectroTYPER RT™ workstation. The SpectroTYPER RT™ workstation is a personal computer (PC) equipped with a Signatec PDA500 500 MHz 8-bit digitizer and a National Instruments IMAQ-PCI 1411 frame grabber. Four signals are disconnected from the Biflex and routed to the SpectroTYPER RT™. These signals are the output of the microchannel plate (MCP) detector, the trigger for the laser and high voltage electronics, the output of a photo-diode detector used to trigger the data acquisition, and the video signal from the sample visualization camera. The output of the detector in the Biflex passes through a gain of five pre-amplifier (Stanford Research Systems SR445) and a passive low pass filter with a cut-off frequency of 90 MHz (Mini Circuits BLP-90). In addition, there is a TCP/IP network connection between the PC workstation and the controlling computer on the Biflex (Sun workstation).

The software on the Sun workstation is modified to accept commands over the TCP/IP interface to move the stage from sample-to-sample and to different raster positions within a sample. The SpectroTYPER RT™ workstation is equipped with software that triggers the mass spectrometer (laser and high voltage pulsing) and acquires the spectrum. The software also controls stage position.

The SpectroTYPER RT™ system has the capability to control the mass spectrometer and acquire spectra and process these spectra in real-time. The assay-based results are used to decide whether or not to raster. The software is implemented in such a way as to allow plug-in algorithms; the following algorithm was used to collect the data presented here. A series of samples is automatically run. For each sample, a set number of shots determined by operator input are averaged to create a spectrum. The system collects the shots, averages them, and then each assay defined for that sample is judged independently. If the score for an assay is less than moderate the system will collect more data for that assay. Another set of shots is averaged, and the result of this data collection is added to the first. Again, a result for each assay is determined and if the score for that assay improves by adding the new shots the new sum is kept. The cycle of data collection and judging continues until a set number of attempts expires, or a score of moderate or better is achieved for each assay in the sample.

3.4.2. *SpectroTYPER RT™ results*

In a typical experiment, 48 new non-established (the first time run for each of these assays) 4-plexes were performed on eight different DNAs (384 reactions) and spotted on a 384-well chip. The same 384-well chip was measured consecutively on three different Biflex instruments. The first run was performed using the standard Biflex Autoxecute software. The standard acquisition available with the Biflex uses fuzzy logic to control rastering based on resolution and signal-to-noise ratio over a fixed mass range. The next two runs were performed on Biflex instruments equipped with SpectroTYPER RT™. Normally the quality of the data would decrease in consecutive runs because the sample is depleted by successive laser shots. The results are presented in Table 1.

We can examine some specific spectra from this experiment to better see how using assay-based judging helps increase call efficiency. Fig. 4a shows a spectrum acquired by the Biflex using its standard judging algorithms. This spectrum contains the results from four assays. In this case, the data produced two conservative calls (AC for assay I and C for assay III) and two low-probability calls (unable to determine genotypes for assays II and IV). Possible peaks at 7129 and 8355 Da significantly reduced the confidence level of genotype calls. As can be seen, the spectrum is complex and the simple judging employed by the Biflex found at least one peak of sufficient quality to save this spectrum and move on even though post-processing by SpectroTYPER™ showed that the data was not sufficient to provide the desired four high quality genotyping results.

Table 1
Comparison of results between Biflex Autoxecute and SpectroTYPER RT™

Call quality	Run 1: Bruker standard acquisition and control	Run 2: Real-time genocalling and control using SpectroTYPER RT	Run 3: Real-time genocalling and control using SpectroTYPER RT
Total possible calls (384 spots × 4 genotypes)	1536	1536	1536
Conservative[a]	1062	1310	1199
Moderate[a]	121	86	167
Aggressive[a]	90	5	23
Low probability[a]	140	58	98
Bad spectrum[a]	123	77	98
Total “good” calls = conservative plus moderate	1183	1396	1366
Improvement in efficiency over standard acquisition (%)	NA	18	15.5

[a] Automatically assigned definitions of assay quality, according to algorithms which assess parameters such as signal-to-noise ratio, peak resolution and mass accuracy within a specified mass range.

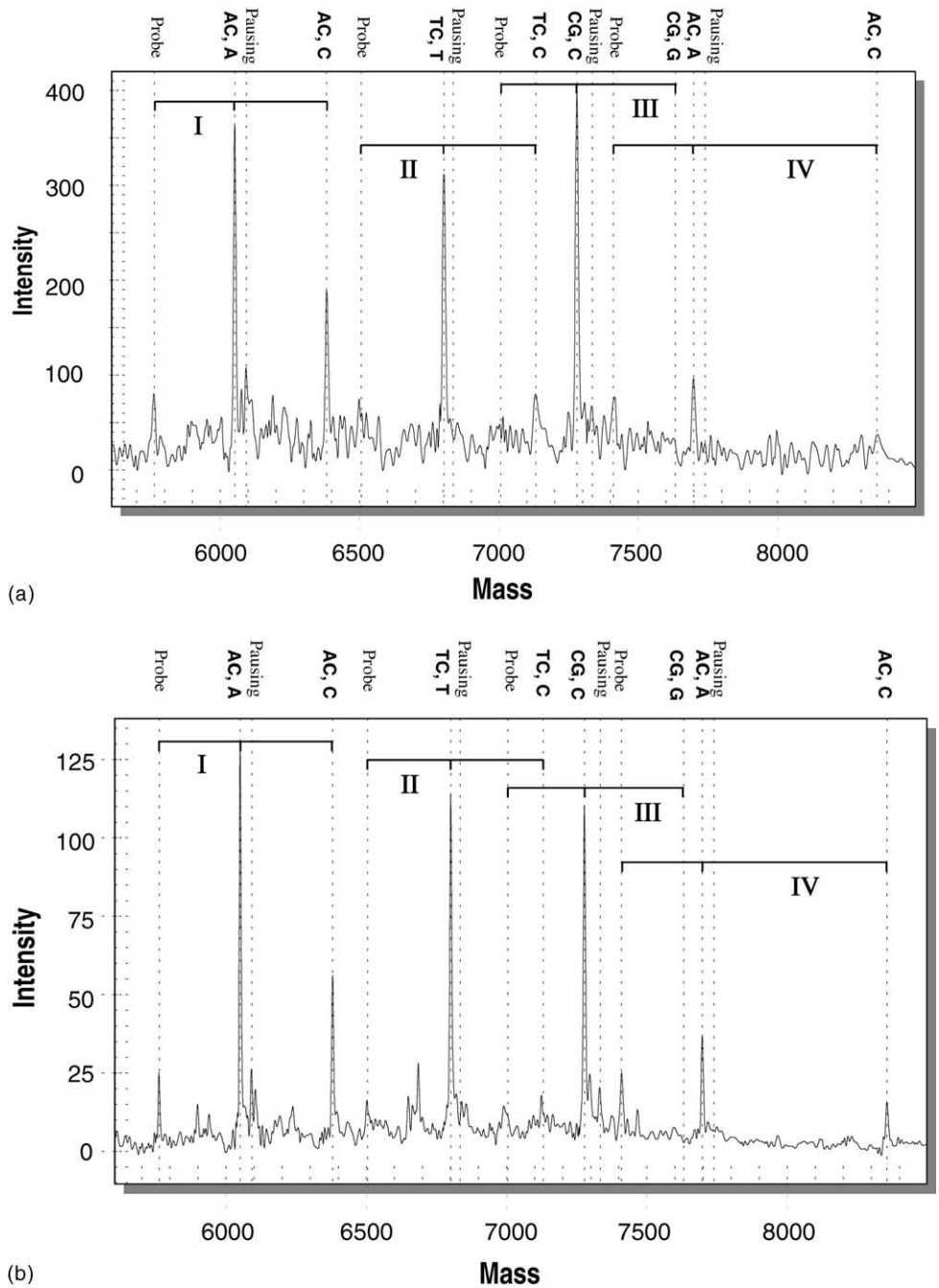


Fig. 4. Spectrum of multiplexed genotyping reaction (hME) acquired by using (a) standard judging algorithms on Bruker Biflex; (b) assay-based judging on SpectroTYPER RT™. Peaks from the same assay are grouped together. Probe: primer position; Pausing: possible peak position produced by termination with a dNTP.

Fig. 4b shows the spectra acquired using assay-based judging. In this case, the data resulted in three conservative calls and one moderate call. The assay-based judging will add shots from different raster positions until all assays give acceptable results. The assay-based judging is not fooled by large primer peaks or large peaks from assays other than the one of interest. The result can be seen by looking at the peaks at mass 7129 and 8355 Da in both figures. The 7129 Da peak represents the C allele of assay II in this sample. From the spectrum acquired with the SpectroTYPER RT™ system it is clear that this peak is not there and assay II should be called T and it was automatically called T with a conservative score. In the spectrum acquired by the Biflex acquisition software, it is less clear what the result should be. The 8355 Da peak represents the C allele of assay IV in the sample. From the spectrum acquired with the SpectroTYPER RT™ system, the peak was unambiguously identified and AC genotype called for assay IV, although a moderate score was given for the reduced signal-to-noise ratio of the 8355 Da peak. However, in the spectrum acquired by the Biflex acquisition software, the peak at mass 8355 Da is especially noisy and could not be identified. In this case averaging in more shots from a different section of the sample would clearly help, but the judging algorithms used by the Biflex do not know to expect a possible peak at this mass, and there are enough peaks elsewhere in the spectrum with good signal-to-noise ratio to judge the spectrum sufficient and move on.

As can be seen in Table 1, the call efficiency was improved by using assay-based judging to control data acquisition. The overall call efficiency went from 77% (1183 good calls/1536 possible) for the standard acquisition to 90.9% (1396/1536) in the first run by using assay-based judging.

The assay-based judging and mass spectrometer control described here and incorporated into SpectroTYPER RT™ has been used to run thousands of 384- and 96-position chips using a wide range of assays and DNAs. The improvement in calling efficiency has been observed to range from 0 to over 50%. The improvement strongly depends on the quality of the assay and the level of multiplexing. The improvement for strong uniplexes is typically not significant. Weak assays and multiplexed assays benefit more from the assay-based judging, and a multiplex with strong and weak assays together or other complex spectra would see the greatest benefit. The use of assay-based spectrum processing has been shown to provide significant improvement in call efficiency over standard methods.

In summary, the major advantage of the RT system over the traditional mass spectrometer-based genotyping systems is that the calling efficiency of multiplex assays has improved dramatically. In the traditional off-line analysis, a spectrum would be saved if some peaks pass a signal-to-noise threshold. In multiplex assays, some assays are invariably weaker than the rest. Typically within the same mass window, the stronger assays would produce intense peaks that meet the signal-to-noise and resolution criteria while the weaker assays would yield peaks that do not meet those criteria. Thus, off-line acquisition has a tendency of passing spectra that yield unambiguous results only for the stronger assays, while the weaker ones will not be called. In the real-time system, each assay is evaluated individually. Thus, even though all the strong assays have been unambiguously called, the acquisition will still continue (usually at different raster spots on the same sample pad) until the weaker assays are called as well. With this added awareness of the biological content of the spectrum, the real-time system can easily improve on the efficiency of calls over traditional systems.

3.4.3. Customized plug-in algorithms

We have implemented SpectroTYPER RT™ with a great degree of flexibility. A well-defined programming interface has been developed that controls the dialogue between the data acquisition component and the biological calling component. Each component is distinct and can be enhanced or modified without impacting the basic communication protocol as long as the programming interface is adhered to. One can thus easily develop very complex and elaborate algorithms for calling biological assays in the SpectroTYPER

RT™ system. For example, another area of innovation afforded by the real-time system is the ability to do on-the-fly statistics on peak heights, resolutions, and areas. This gives the system the control over the progress of acquisition and enhances the quality of resultant data. There are applications that use quantitative information such as determining allele frequencies. Because of the imperfect crystallization of the sample and matrix, it is imperative for quantitative applications to obtain a good sampling from many raster positions on the crystal. On-the-fly statistics allows one to gauge the progress of data acquisition and enhances the quality of the resulting data accumulated.

4. Other MALDI-MS-based genotyping assays

The use of MALDI mass spectrometry provides unparallel accuracy for unambiguous detection. Based on this method, many other strategies for assay design have been developed to generate base-specific information for genotyping and/or SNP discovery. One class of assays that shows great promise involves the use of base-specific cleavage of the amplicon of interest. Each complete base-specific cleavage of a given amplicon will generate a fragmentation pattern indicative of the sequence. A comparison between theoretical and experimental cleavage patterns allows identification of sequence changes. Sequence changes will either lead to addition or loss of cleavage fragments and their corresponding mass signals, or will shift an existing fragment by the mass of the substituted nucleotide. In a MALDI mass spectrum, such complete digest fragments usually distribute in the range from 1000 to 7000 Da, and are usually concentrated in the 1000–3000 Da region. The use of a high-resolution analyzer, such as a reflectron TOF or a QqTOF could facilitate more accurate peak identification.

4.1. DNA-based UDG fragmentation

In this approach, the DNA region of interest is amplified using PCR in the presence of dUTP instead of dTTP. The PCR product is then immobilized onto streptavidin-coated magnetic beads and denatured. The immobilized strand is treated with uracil-DNA-glycosylase (UDG) to generate nucleotide-U-specific abasic sites and then eluted and fragmented to completion by alkaline treatment. The fragments are then analyzed by MALDI-MS. This approach has been used for rapid bacterial identification [24] and can be easily adapted for scanning for SNPs. An example is shown in Fig. 5. A T/C polymorphism in the UCP-2 gene is screened by UDG fragmentation. The corresponding fragments are listed in Table 2. At least one peak is unique in each allele, which can be used as a marker. A heterozygote is represented by both markers, as shown in Fig. 5.

4.2. Base-specific RNA fragmentation

An alternative approach is to transcribe the PCR amplicon into RNA and use different RNA endonucleases to generate base-specific fragments. The PCR primer can be tagged with either a T7 or SP6 promoter. The in vitro transcript can be completely digested with RNase T1 at every G-position, or RNase U2 at every A-position. RNase PhyM (A and U specific) and RNase A (C and U specific) can also be used to generate corresponding base-specific fragments. The specificity of these ribonucleases appears to be good [25].

5. Advances made in SNP analysis by MALDI-MS

5.1. SNP validation

The ability to work with pooled DNA samples offers the opportunity to characterize large numbers of candidate SNPs whose polymorphism status has not yet been verified using population-based samples. Often, candidate SNPs are identified through alignments of multiple homologous DNA sequences. However, DNA sequencing quality can be variable, sometimes resulting in apparent sequence variation that is not actually present in a population. Therefore, there exists a need to confirm the existence of alternate alleles

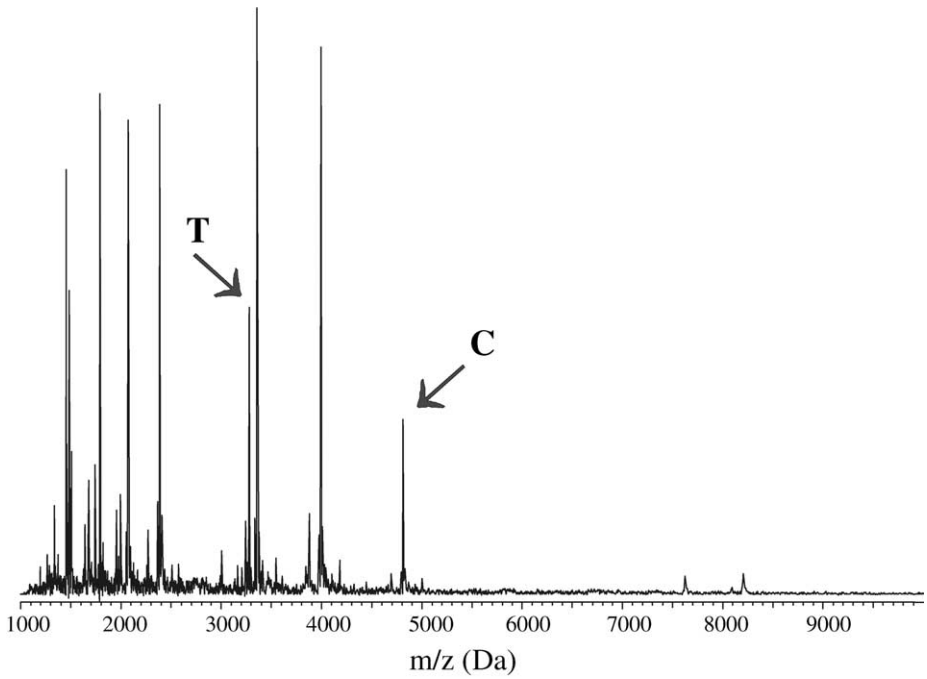


Fig. 5. A heterozygous C/T polymorphism is detected by using UDG fragmentation. The peaks indicated are unique to the alleles they represent, respectively, and can be used as markers.

suggested by candidate SNPs. Ideally such SNP validation efforts should also provide allele frequencies for one or more populations. Given the very large numbers of SNPs being identified through various genome-related efforts, the tasks of SNP validation and allele frequency determination would appear to be daunting. Certainly, validating hundreds of thousands of candidate SNPs throughout the genome, using an approach based on genotyping individual DNAs, is cost-prohibitive and not feasible in a reasonable time frame. However, pooling DNAs from large numbers of individuals representative of a population permits undertaking these tasks.

We recently demonstrated the validity of this approach in a collaborative study with the NCI [21]. The objective of this study was to apply a high-throughput, automated approach to validate more than 10,000 candidate SNPs identified in silico through analysis of aligned expressed sequence tags (ESTs) [26], using pooled DNAs. Besides confirming the polymorphic status of a candidate SNP, another objective was to simultaneously estimate the allelic frequencies within a pool comprised of nearly 100 individuals of European ancestry from the Centre d'Etude du Polymorphisme Humain (CEPH) panel [27]. Applying advances in automated assay design and sample processing, more than 9100 SNP assays were developed and tested using the CEPH DNA pool. To maximize throughput, all assays were designed to perform under a universal set of conditions. No optimization of assays was done prior to testing on the pooled DNAs.

Data was generated by MALDI mass spectrometric analysis of the MassEXTEND^{TM} products dispensed in high-density arrays onto chips. Products from each MassEXTEND^{TM} reaction were spotted in quadruplicate, as it was found that multiple data points improved the precision of the frequency estimates obtained. Each spot was scanned, and all four spectra were evaluated for data quality. Allele frequencies were averaged, and standard deviations determined from the multiple data. The variability measures

Table 2
Fragments generated by UDG-based fragmentation in two different alleles

T allele		C allele	
CCCAGTCACGACGTTGTAAAACGTCTTGGCCT TGCAGATCCAAGGAGAAAGTCAGGGGCCAGT GCGCGCTACAG**T**CAGCGCCCAGTACCGCGGT GTGATGGCCTGTGTGAAATTGTTATCCGCT		CCCAGTCACGACGTTGTAAAACGTCTTGGCCT TGCAGATCCAAGGAGAAAGTCAGGGGCCAGT GCGCGCTACAG**C**CAGCGCCCAGTACCGCGGT GTGATGGCCTGTGTGAAATTGTTATCCGCT	
	Mass		**Mass**
Primer cleavage fragments		Primer cleavage fragments	
bioCCCAGTCACGACGTTGTAAAACG	7453.6	bioCCCAGTCACGACGTTGTAAAACG	7453.6
Fragments from the sequence displayed above		Fragments from the sequence displayed above	
pCpR	483.2	pCpR	483.2
pApR	507.2	pApR	507.2
pGpR	523.2	pGpR	523.2
pGpR	523.2	pGpR	523.2
pGpR	523.2	pGpR	523.2
pGpR	523.2	pGpR	523.2
pGApR	836.4	pGApR	836.4
pCCGCpR	1390.8	pCCGCpR	1390.8
pGGCCpR	1430.8	pGGCCpR	1430.8
pGGCCpR	1430.8	pGGCCpR	1430.8
pACAGpR	1438.8	pGAAApR	1462.8
pGAAApR	1462.8	pGCAGApR	1768
pGCAGApR	1768	pGCGCGCpR	2049.2
pGCGCGCpR	2049.2	pACCGCGGpR	2362.4
pACCGCGGpR	2362.4	pCAGGGGCCAGpR	3334
pCAGCGCCCAGpR	3254	pCCAAGGAGAAAGpR	3968.4
pCAGGGGCCAGpR	3334	pACAGCCAGCGCCCAGpR	4788
pCCAAGGAGAAAGpR	3968.4	bioCCCAGTCACGACGTTGTAAAACGpR	7630.6
bioCCCAGTCACGACGTTGTAAAACGpR	7630.6		

provided useful guidance for interpreting spectra from assays suggesting very low allele frequencies. For most assays indicating minor allele frequencies $<10\%$ and having small standard deviations, follow-up genotyping of individual DNAs represented in the pool confirmed the presence of rare second alleles, thereby also demonstrating the sensitivity of DNA MassARRAY^{TM}.

Summarizing the results of this collaborative study, nearly 7700 of the designed assays (84%) performed successfully under universal reaction conditions. Of these, more than 6400 produced outcomes that could

be unequivocally scored as either polymorphic (minor allele frequency > 0.1) or non-polymorphic (no detected peak corresponding to the second allele in the four independent measurements). Among these, 3646 (57%) were classified as SNPs as described above, and data from an additional 690 candidates were indicative of rare minor alleles with frequencies in the range of 0.01–0.1. Once the designed reagents were in hand, these validation experiments for the more than 9100 SNP candidates were performed in less than 4 weeks.

Several technical issues were brought to our attention during this study. First, unexpected peaks were observed in approximately 1% of experimentally proven assays. Additional products could indicate tri-allelic SNPs, errors in the sequence upon which the assays were designed, or simply non-specificity of the amplification reaction. Co-amplification of homologous gene sequences or pseudogenes, or the presence of highly repetitive sequences (e.g., Alu) may account for many of these additional products. Second, it was observed that for same assays the PCR reaction appeared to amplify paired alleles non-uniformly, favoring amplification of one allele (skewing). For many of these assays the imbalance between the paired alleles was highly reproducible, and in these situations a correction factor could be applied based on characterizing peak ratios in genotype data from heterozygotes. However, for other assays this skewing was not reproducible for the same assay/template combination in multiple experiments, indicating the likely need for assay redesign. Last, it was shown that assay quality, and thus reliable frequency estimates, were highly dependent on the quality of synthetic oligonucleotides from commercial suppliers. The percentage of successful assays was observed to vary considerably between vendors, and even between batches from the same vendor (Fig. 6). Therefore, it is important to identify suppliers capable of providing high quality oligonucleotides routinely. In this light, it can be noted DNA MassARRAY™ also represents an ideal tool for quality control monitoring of oligonucleotides prior to their use in pooled sample analysis or any other applications whether based on mass spectrometry or not.

The NCI study demonstrated it is currently feasible to perform large-scale validation studies involving thousands of SNPs cost-efficiently, using analytical platforms currently available. The study made one of the largest collections of validated gene-based SNPs available to the genetics community, and demonstrated DNA MassARRAY™ as a breakthrough technology in large-scale SNP studies. Since the NCI study, improvements continue to be made in all aspects of the process, from the up-front in silico work, to sample processing, to data analysis. In particular, successful

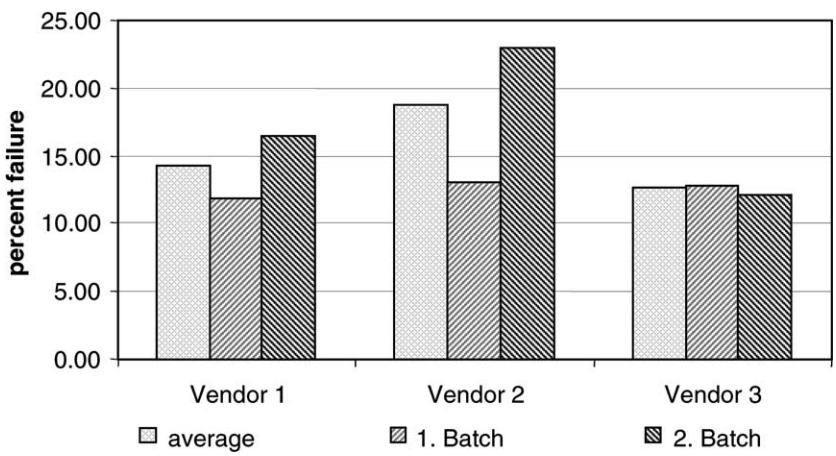


Fig. 6. Comparisons between assay performance (percentage failed assays) among three different vendors supplying oligonucleotides for the collaborative SNP validation study with the NCI. Each vendor provided oligonucleotides in two large batches.

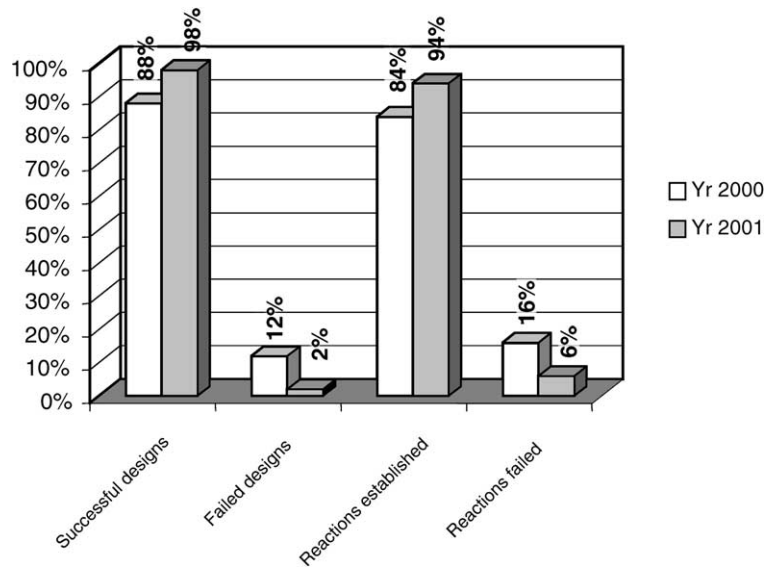


Fig. 7. Evolution of assay processing efficiency.

automated assay design and establishment of working assays have both improved by 10% during the year following the completion of this collaborative study (Fig. 7).

5.2. Genome-wide association studies using MALDI-MS

The throughput advantage offered by pooled DNA analysis allows genetic variability to be approached in a brute force way to rapidly identify medically relevant markers and therapeutic targets. With DNA MassARRAY™, much larger numbers of SNPs, either gene-based or genome-wide, can now be characterized in carefully designed study groups involving large numbers of individuals. It is now feasible to undertake SNP-phenotype association studies involving tens of thousands of assays on a genome-wide scale. These could be case-control studies to identify alleles associated with disease susceptibility, indicated by significant allele frequency differences between matched groups of affected vs. unaffected individuals, or pharmacogenetics/genomics studies involving groups organized according to varied therapeutic responses.

It is not possible today to predict allele frequency differences relevant to polygenic/multifactorial disorders compared to the healthy population, or accounting for highly variable clinical responses to drug therapies. There is simply not enough known about the genetic underpinnings of these processes to allow scientifically correct assumptions. Also the number of samples required is thought to be critical, but often difficult to predict. It is dependent upon a variety of factors, including the overall genetic heterogeneity of a given disease, how many alleles are required in concert to produce a particular phenotype, and the penetrance of particular alleles. To obtain maximum success in association studies it is expected that the sample size must be as large as practically possible. A straightforward genotyping approach would present practical limitations to either the number of SNPs that can be assayed in such studies, or the number of samples genotyped with a consequent loss of statistical power.

However, consider the possible throughput advantage of pooled sample analysis as illustrated in the following example. A realistic scenario for a disease gene association study could involve comparing

allele frequencies for 10,000 SNPs among 400 individuals (200 cases vs. 200 controls). Genotyping individual DNA samples would require the processing of 4,000,000 uniplex reactions at a substantial cost and time commitment. Even five-plex processing would still require 800,000 reactions. In contrast, combining all cases and controls in their respective pools offers a potential 40–200-fold first-pass advantage in throughput ($10{,}000 \text{ assays} \times 2 \text{ pools} = 20,000$ reactions processed). Even if efforts were then undertaken to verify the most promising assays through repeated pool analyses and/or individual genotyping, this strategy can still offer impressive reductions of up to 90% in both time and project cost. Pooled sample analysis using MALDI-TOF mass spectrometry (DNA MassARRAY^{TM}) therefore represents the best opportunity for completing genome-wide association studies to identify novel alleles underlying disease susceptibility, and for demonstrating the medical utility of assays for these alleles in diagnosis and improving therapy.

Acknowledgements

Our special thanks goes to Dr. Charlie Rodi for providing us an unpublished manuscript for the hME assay. We would also like to thank Niels Storm, Jhobe Steadman, Xian-Wei Yao, Mona Shahgholi, Chao Lin, Paul Heaney, Carolyn Wrightson and others of the SEQUENOM team for their contributions to the various projects involved. Finally, we would like to thank Prof. Franz Hillenkamp for his vision and numerous advice during each stage of the developments.

References

[1] M.C. Fitzgerald, L.M. Smith, Annu. Rev. Biophys. Biomol. Struct. 24 (1995) 117.
[2] E. Nordhoff, F. Kirpekar, P. Roepstorff, Mass Spectrom. Rev. 15 (1996) 67.
[3] M.T. Roskey, P. Juhasz, I.P. Smirnov, E.J. Takach, S.A. Martin, L.A. Haff, Proc. Natl. Acad. Sci. U.S.A. 93 (1996) 4724.
[4] P.L. Ross, P. Belgrader, Anal. Chem. 69 (1997) 3966.
[5] P.L. Ross, P.A. Davis, P. Belgrader, Anal. Chem. 70 (1998) 2067.
[6] N.I. Taranenko, V.V. Golovlev, S.L. Allman, N.V. Taranenko, C.H. Chen, J. Hong, L.Y. Chang, Rapid Commun. Mass Spectrom. 12 (1998) 413.
[7] L.Y. Ch'ang, K. Tang, M. Schell, C. Ringleberg, K.J. Matteson, S.L. Allman, C.H. Chen, Rapid Commun. Mass Spectrom. 9 (1995) 772.
[8] J. Bai, Y. Liu, X. Liang, Y. Zhu, D.M. Lubman, Rapid Commun. Mass Spectrom. 9 (1995) 1172.
[9] A. Braun, D.P. Little, D. Reuter, B. Muller-Mysok, H. Koster, Genomics 46 (1997) 18.
[10] D.P. Little, A. Braun, M.J. O'Donnell, H. Koster, Nat. Med. 3 (1997) 1413.
[11] L. Haff, I.P. Smirnov, Genome Res. 7 (1997) 378.
[12] P. Ross, L. Hall, I.P. Smirnov, L. Haff, Nat. Biotechnol. 16 (1998) 1347.
[13] K. Tang, D. Fu, D. Julien, A. Braun, C.R. Cantor, H. Köster, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 10016.
[14] D.P. Little, T.J. Cornish, M.J. O'Donnell, A. Braun, R.J. Cotter, H. Köster, Anal. Chem. 69 (1997) 4540.
[15] J. Leushner, N. Chiu, Mol. Diag. 5 (2000) 341.
[16] C. Jurinke, D. van den Boom, C.R. Cantor, H. Köster, Methods Mol. Biol. 170 (2001) 103.
[17] N. Storm, B. Darnhofer-Patel, D. van den Boom, C.P. Rodi, Methods in Molecular Biology, vol. 212, 2003, in press.
[18] A. Braun, D.P. Little, H. Koster, Clin. Chem. 43 (1997) 1151.
[19] E. Nordhoff, R. Cramer, M. Karas, F. Hillenkamp, F. Kirpekar, K. Kristiansen, P. Roepstoff, Nucleic Acids Res. 21 (1993) 3347.
[20] P. Ross, L. Hall, L.A. Haff, Biotechniques 29 (2000) 620.
[21] K.H. Buetow, M. Edmonson, R. Macdonald, R. Clifford, P. Yip, J. Kelley, D.P. Little, R. Strausberg, H. Koster, C.R. Cantor, A. Braun, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 581.
[22] R.W. Nelson, M.A. McLean, T.W. Hutchens, Anal. Chem. 66 (1994) 1408.
[23] K.J. Wu, A. Steding, C.H. Becker, Rapid Commun. Mass Spectrom. 7 (1993) 142.
[24] F. von Wintzingerode, S. Böcker, C. Schlötelburg, N.H.L. Chiu, N. Storm, C. Jurinke, U.B. Göbel, D. van den Boom, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 7039.
[25] S. Hahner, H. Lüdemann, F. Kirpekar, E. Nordhoff, P. Roepstoff, H. Galla, F. Hillenkamp, Nucleic Acid Res. 25 (1997) 1957.
[26] K.H. Buetow, M.N. Edmonson, A.B. Cassidy, Nat. Genet. 21 (1999) 323.
[27] J. Dausset, H. Cann, D. Cohen, M. Lathrop, J.M. Lalouel, R. White, Genomics 6 (1990) 575.