An approach to high-throughput genotyping.

J M Hall, C A LeDuc, A R Watson, et al.

Genome Res. 1996 6: 781-790
Access the most recent version at doi:10.1101/gr.6.9.781

REVIEW

An Approach to High-throughput Genotyping

Jeff M. Hall,[1] Carrie A. LeDuc, Andy R. Watson, and Alan H. Roter

Sequana Therapeutics, La Jolla, California 92037

Since the advent of DNA-based genetic typing, much effort in the field of human genetics has gone into identifying greater numbers of markers of a highly informative nature. The generation of large numbers of such markers, particularly microsatellite repeats, consisting of di-, tri-, and tetranucleotide motifs, has resulted in dense, genome-wide marker maps (Cooperative Human Linkage Center 1994; Utah Marker Development Group 1995), the latest of which contains >5000 microsatellite markers (Dib et al. 1996). These mapping resources have increased greatly our ability to perform genetic analysis of human populations and has allowed the field to begin the search for genes involved in complex, multifactoral human diseases such as diabetes, asthma, schizophrenia, and other polygenic illnesses.

The result of these improved mapping tools has been the increased need for high-throughput methods for generating large numbers of human genotypes. For instance, a typical current human genetic disease mapping project might involve applying 300 markers to scan the genomes of a population of 500 individuals. This example would require 150,000 genotypes, an effort that until recently could have occupied a laboratory full-time for several years. High-throughput methods allow an individual worker to produce 5000 genotypes per week so that a 150,000 genotype genome scan can be completed in 10 weeks by a team of three people. This article addresses the current status of high throughput human genotyping using examples from the approaches our laboratory has taken in constructing a high-throughput facility capable of generating >2 million genotypes per year. This review also considers data management issues that arise when using high-throughput methods, common problem areas in high-throughput genotyping, and probable future developments and improvements in these methods.

[1]Corresponding author.
E-MAIL jhall@sequana.com; FAX (619) 452-6653.

Current Techniques in High-throughput Genotyping

Input DNA

DNA samples to be used in a genome scan must be uniform in nature and of high quality. Patient samples may be received in the form of fresh blood, frozen blood, cell lines, or previously extracted DNA. For samples where DNA extraction is required, the extraction may be performed either manually or using an automated instrument such as the Genepure 341 Nucleic Acid Purification System (Applied Biosystems Division, Perkin Elmer, Foster City, CA). DNA extraction can be performed using organic solvents such as phenol and chloroform, as in a recently described single-tube isolation procedure (McIndoe et al. 1995) or using nonorganic methods in commercially available kits such as the ones from Gentra (Minneapolis, MN), which rely on salt precipitations, or those sold by Qiagen (Chatsworth, CA), which utilize binding of the DNA to a resin contained within a spin column. Sequana has recently developed a semiautomated robotic system using the Gentra reagents in a 96-well format on a Beckman Biomek 2000 platform (Fullerton, CA) that has increased the consistency and efficiency of DNA extraction from whole blood and from cell lines.

Following DNA extraction, the DNA may be quantified using spectrophotometric or fluorometric methods (Ahn et al. 1996). We employ a sensitive double-stranded DNA-specific assay using the intercalating dye Picogreen (Molecular Probes, Eugene, OR), which is read in a 96-well fluorometer, the Cytofluor II (Perceptive Biosystems, Framingham, MA). Finally, before undertaking a genome scan a quality control step

should be implemented. A small number of the samples is subjected to genotyping with a panel of 10 test markers to be certain that the DNA will amplify well and be readable in the fluorescence-based genotyping system. These procedures for purification and quality control of the sample DNAs can greatly reduce problems of variable genotyping performance that would preclude high throughput efficiencies during a genome scan.

The most widely used sample format for high-throughput genotyping is the 96-well plate. The majority of wells on the plate contain the various study subjects' DNA, and several wells at the end of each plate can be reserved for use as positive or negative controls. Before making the plates for genotyping, it is best to have the stock DNAs adjusted to a common concentration, which facilitates delivery of an accurate volume by the dispensing robot. We use a Packard MultiPROBE robot (Meridian, CT) to bring all the DNAs to a standard concentration of 4 ng/μl. We dispense 5 μl of each sample from the 4 ng/μl stocks for a total of 20 ng of DNA to be used in an eventual PCR reaction volume of 20 μl per well. The 4 ng/μl stock DNAs can be dispensed using a 96-channel pipetting robot such as the Robbins Micro Dispenser (Sunnyvale, CA). Theoretically, a genome scan using 300 markers can be done using 6 μg of DNA. Our laboratory requires a minimum of 20 μg of DNA per sample, as following up presumptive linkages from a genome scan can require substantial additional genotyping to saturate the regions of interest. Once dispensed, the plates containing the study DNAs may be used in the wet state or dried down, which is a more convenient way to store large numbers of plates over time. It is convenient when possible to group the samples on a plate by family, as this can lead to easier allele calling later.

Sample Tracking

A robust system for the management of sample identity and history is required for any high-throughput laboratory. We have designed a system for sample labeling that uses preprinted, bar-code labels for all steps in the sample handling process. Blood vacutainers are affixed with bar-code labels in the field and duplicate labels are affixed to the accompanying data sheets. The bar code consists of a three-character study code (indicating the disease project) followed by a six-digit identifier. Once received at our facility, vacutainer bar codes are scanned into our LIMS system (Laboratory Information Management System; see Data Management section below). This process records the sample ID, date received, and any related information. During all subsequent manipulations the sample maintains its unique bar code, which can be traced back to the LIMS data base. After robotic distribution of DNAs into 96 well plates, the plates are also labeled with a bar code. Plate bar codes consist of a three-character study code, the letter P for plate, a two-digit template identifier (which specifies the plate template that defines which patient DNAs are in which wells), and a four-digit plate identifier. Every PCR plate therefore has a reference to a template that details the plate contents and a unique identifier that allows the tracking of specific genotypes back to the plate.

Microsatellite Markers

As originally proposed (Botstein et al. 1980), a set of 300 evenly spaced markers can be used to map the entire human genome at an average density of approximately 10 cM. The actual number of markers needed in a genome scan depends on the strength of genetic signal one anticipates for a given disease, the number of genes involved, and the number of affected individuals available for typing (Lander and Schork 1994; Weeks and Lathrop 1995). It has been argued that by using a larger population size, one may employ fewer markers (Elston et al. 1996); however, the expense of ascertaining additional subjects and acquiring and typing their DNAs must be considered. For example, we estimate that the cost for a single genotype is $1, so being able to decrease the total number of markers in a genome scan from 300 to 200 would save $100 per individual typed. However, the cost to obtain the additional patients required to compensate for the decreased marker density can easily cost $100 or more per patient, a cost that includes measuring phenotypes as well as bleeding, mailing, and then extracting the DNAs.

Sets of evenly spaced microsatellite markers covering the human genome can be chosen from publicly available markers, but not all published markers are found to perform equally in all laboratories. Some individual discrimination and choosing may be required before an acceptable set of genome scan markers are found. Preselected fluorescent microsatellite marker sets that span the human genome are available from sev-

eral commercial sources, including Research Genetics (Huntsville, AL), the ABI Divison of Perkin Elmer, and Genpak (Brighton, England). Often these have been optimized to run under similar PCR conditions, grouped to cover certain chromosomes, and/or grouped to give maximum coverage when multiplexing the gel loadings. Although these markers have been screened for robust performance, individual laboratories still experience some difficulties in getting all the markers to work, and some fine-tuning of the PCR conditions is usually required. At Sequana a routine optimization process is run that tests all new markers at annealing temperatures from 54° C to 64° C and Mg^{++} concentrations from 1 to 4 mM to determine optimal amplification conditions. For particularly difficult markers we often rely on a "touchdown" PCR protocol, which starts at an annealing temperature of 66° C and decreases at every other cycle by 1° until 52° C is reached and then continues for 10 cycles at 52° C annealing temperature.

Multiplexing Issues

While multiplexing several markers during the actual PCR step is possible, it is often difficult to achieve uniform amplification across all of the multiple markers within a single PCR reaction. An additional amount of primer design and reaction optimization is required to ensure that multiplexed primer sets actually work well together. Some advances incorporating the addition of common tails to the marker-specific primers may allow more efficient PCR multiplexing (Shuber et al. 1995; Lin et al. 1996), but these methods have not been applied successfully yet to high-throughput fluorescent genotyping. In our laboratory we proceed with the amplification of the individual markers, followed by pooling the products from the individual PCR reactions prior to electrophoresis.

Individual markers for PCR are set up in a 96-well format using a Packard MultiPROBE robot, which can perform both the preparation and aliquotting of the PCR master mix. PCR is performed using the PTC-100 thermal cycler (MJ Research, Watertown, MA), and the amount of PCR product from each marker is measured to calculate the optimal pooling amounts. A Hamilton Micro Lab 2200 robot (Reno, NV) is used to pool the reactions from the various individual 96-well plates onto one master plate. This plate is dried down and the samples on it are resuspended prior to electrophoresis in a loading buffer that contains a fluorescent size standard.

There are now enough markers available for a genome scan to choose between using dinucleotide or tetranucleotide repeat markers. Dinucleotide repeat markers, primarily CA repeats, allow for sets that can be very dense. In our own laboratory using dinucleotides we have successfully created multiplexing sets of up to 17 markers per lane, and others have reported multiplexing up to 24 markers in a lane (Reed et al. 1994). Tetranucleotide repeat markers cannot be as densely multiplexed, but they improve the discrimination of alleles because of their wider spacing. Our own current genome scanning sets incorporate a 50:50 mixture of di- and tetranucleotide repeat markers. For specific project applications other than genome scanning, such as the saturation of a small chromosomal region, it is necessary to generate new multiplexed marker sets specific to that task. We have written a software program that facilitates the creation of such templates by using data base information for each marker, such as the allele size range and fluorescent label, and then automatically creates optimal multiplexed marker sets.

Hardware Issues

Currently DNA electrophoresis hardware that incorporates software for the analysis of microsatellite markers is available from the ABI Division of Perkin Elmer, Li-Cor (Lincoln, NB), and Pharmacia (Uppsala, Sweden). We currently use ABI 373 and 377 instruments. The run times needed on such instruments for markers in the 100- to 500-bp size range are 3–6 hr each, so that multiple runs can be scheduled in a single day. Some groups have constructed their own instrumentation to accomplish high-throughput genotyping (J.L. Weber, D.A. Vaske, C. Blanchette, D.E. David, C.A. Christenson, and T.L. Rusch, pers. comm.). No commercial instrument is available yet that can run 96 samples simultaneously, so the products of a single 96-well plate are run on three sequencers. Although the majority of automated genotyping is done using direct fluorescent genotyping on DNA sequencing instruments, there are groups that employ multiplexed PCR and gel separation of markers followed by transfer of the nonlabeled products to a membrane and repeated probing of that membrane with multiplexed sets of labeled probes (L. Doucette-Stamm, M. Wang, R. Patel, D. Blakely, H.

Yu, J. Carulli, T. Keith, J. Harris, P. Richterich, and J. Mao, pers. comm.). This procedure involves scoring autoradiograms or the imaged output from the hybridized probes (Ginns et al. 1996), which can be facilitated using imaging instruments such as the Fluorimager (Molecular Dynamics, Sunnyvale, CA). It is also possible to employ the method of incorporating a radioactive primer in the PCR step, followed by gel electrophoresis and autoradiography in high-throughput genotyping. The disadvantage in this method is the loss of multiplexing capacity that is afforded with multicolor systems. However, in mapping efforts where allele sizes can be predetermined, such as the mouse genome project (Dietrich et al. 1996), the density of markers per lane may be optimized to increase greatly the multiplexing capacity.

Data Management

A number of data management issues are encountered in high-throughput genotyping for a large disease mapping project. The data management system employed must allow the import of raw data from the laboratory, the processing of that raw data to generate finished genotypes, error analysis and correction of the finished data (which requires the ability to go back to the raw data), the compilation of all data, and export in a finished form to suitable programs for genetic analysis. The example given earlier in this review of a typical human disease mapping project generating 150,000 genotypes over a 10-week period would correspond to data processing on approximately 5 gigabytes of electronic data. Managing the data for 10 such projects involves processing 1 gigabyte of new data per day. One of the challenges of high-throughput genotyping is to create an informatics environment that can support such a large data flow efficiently.

Network

The amount of new data per day generated in high-throughput genotyping does not flow evenly throughout a 24 hour period. The data flow is dependent on the periodic delivery of newly generated data from the genotyping instruments. This can correspond to surges of 100 to 350 megabytes of data during a 10-min period several times a day. During these periods the data surge of about 6 megabit/sec can be 60% of the capacity of a standard 10 megabit/sec Ethernet network. At this collision rate the network efficiency of data transfer is reduced dramatically, and the standard Ethernet connection can create a bottleneck. Other genome laboratories have installed standard Ethernet networks (Adams et al. 1994), as the faster Ethernet equipment has become available only recently. Sequana has addressed this problem by installing both standard 10 BaseT Ethernet cabling and fiber optic cabling as part of our network infrastructure. Such a network can support higher data transfer speeds in conjunction with upgraded computers to support these speeds.

Data Base

It is advisable to employ a central facility for storing all the information related to the research activities involved in the high-throughput effort. Such a system can store all relevant information from a study, including phenotype data, pedigree structures, DNA sample information, genotypes, and any information culled from any of the steps involved in the entire high-throughput process. A custom LIMS has been built to centralize data storage for all of our research activities. We store all genotype data in LIMS, including the raw electropherogram data. Data automation and analysis tools interact directly with the LIMS data base for their input (I) and output (O), which simplifies software application development by generalizing I/O mechanisms. The Sybase Database Management System has become a standard in the genome community for public data bases, such as the Genome Data Base, the Genome Sequence Data Base, and the National Center for Biotechnology Information, as well as for the support of high-throughput laboratories such as the Institute for Genome Research (Fleischmann et al. 1995) and the Cooperative Human Linkage Center (Murray et al. 1994). Our LIMS uses a Sybase data-base management system (Emeryville, CA) and runs on a two-processor Sun SparcStation 1000 (Mountain View, CA) with a 30-gigabyte RAID data storage system. This standard configuration allows incremental performance tuning and hardware upgrades as needed. We use both Unix and MacOS clients to access the LIMS data-base server. The Macintosh clients support in-house software that simplifies interacting with LIMS, and the Unix clients provide around-the-clock reliability for automated processes. Client-server configurations are quite common and provide a foundation for software development in some informatics groups (J.M. Gill II, pers.

comm). Genotype data is stored in LIMS with a number of boolean flags that indicate the status of each genotype as it moves through the automated system. At any point in time, a scientist can query LIMS to determine the status of a specific genotype or a genome scan.

Automated Data Processing

High-throughput genotyping requires that the genotyping instruments be available continuously for the next run as soon as the prior run is complete. Since the software that comes with most genotyping instruments is designed for analysis on the instrument, this creates the need to move the newly collected data off the instrument for analysis elsewhere. Different groups have solved this problem in different ways. One option is to dedicate two Macintosh computers to each automated sequencer (R. Gibbs, pers. comm.). This allows the data collection on the second computer while the first is being used for data analysis. Other laboratories have replaced the Applied Biosystems Division, Perkin Elmer data collection software completely (Golden et al. 1993) by designing their own data collection and analysis software. We have removed all of the analysis components from the ABI instrument and installed them on our dedicated computation servers. We have written a MacOS application that runs on the instrument computer along with the ABI data collection software and upon completion of the gel data collection assigns the gel a unique run identifier and then moves the data to a centralized file server where it waits for the computation server for processing.

Processing of the raw data from the genotyping instrument involves the identification of lanes, extraction of the signal from each lane, identification of the sample and size standard peaks, and calculation of the base-pair size of sample peaks. These functions have been integrated in most existing genotyping software into a semiautomated system. These systems allow the scientist to review the results of the automatic analysis. In some laboratories, the ABI/PE Genotype software package is used for semiautomated high-throughput allele calling (Reed et al. 1994). We use Genotyper as a semiautomated tool for the front-end steps of allele identification, specifically for the demultiplexing of each gel lane. The process of demultiplexing involves identifying the one or two peaks corresponding to alleles for each marker in each size range and color of a lane. To accomplish this, we set up templates corresponding to each multiplexed marker set. The template contains category definitions of the size range and color of each marker in the set. It also contains a macro that automates the identification of the correct peaks for each marker. For semiautomated demultiplexing, Genescan results are loaded into the appropriate worksheet template and the macro is run. This identifies the peaks and generates an output table in an appropriate format for import into LIMS. After review of the results, the output table is saved to disk and imported into the LIMS system.

Allele Calling and Mendel Checking

Allele calling is the process of assigning the same single character code to every microsatellite allele with the same base-pair size. Because experimental error leads to a ± 0.5-bp spread around the actual base-pair size, it is necessary to use statistical analysis to determine the allele categories. Once the correct size bins are determined the allele codes can be assigned automatically by a software program. Once allele codes have been assigned they can be checked for non-Mendelian behavior in the nuclear or extended families, and errors may be detected and corrected before the data is sent on for analysis. Mendel checking in the high-throughput setting requires a framework for evaluating and correcting erroneous allele calls. For example, we have written a software program that brings together on a single computer all of the data needed for Mendel checking and correction. An initial check is made of each marker for non-Mendelian inheritance. When a problem is identified in a family all the relevant genotypes are presented within the context of that family pedigree, facilitating the identification of the genotypes that caused the non-Mendelian inheritance problems. In cases where the allele calls may be in question, clicking on a pedigree member queries LIMS and displays the electropherogram trace that gave rise to that member's genotype. This allows visual examination of the peaks and their calls. The peak calls can be adjusted easily with this tool, and the allele in error can be recalled.

Common High-throughput Genotyping Problems

Stutters and Plus A Problems

Two problems that are experienced frequently

with microsatellite markers, particularly the dinucleotide repeat type, are stuttering or shadow bands (a leading ladder of minor products preceding the primary allele peak) and "plus A" (the addition of an extra A, at the 3′ end of the amplified product). Both of these artifacts can cause difficulties in allele calling, particularly when analyzing dinucleotide repeats on heterozygous individuals with two alleles of close sizes. It has been proposed that by characterizing the stutter bands and incorporating that information in the allele calling analysis, it may be possible to achieve more accurate discrimination and calling of closely spaced alleles (Perlin et al. 1995), but this technique is still experimental. Plus A can cause splitting of the electropherogram peak, particularly when the addition is only partial. This problem does not occur with all markers, and may be intermittent even for a marker that does experience this artifact. The addition of this extra base seems to be influenced by the sequence at the 3′ end of the PCR product and by the PCR conditions, particularly the post-PCR holding temperatures and time length, although neither of these variables accounts for the entire problem. It has been proposed that the post-PCR addition of T4 polymerase can correct this problem by removing the extra A but this treatment still leaves some plus A products uncorrected for certain markers (Ginot et al. 1996). It is most efficient to optimize the PCR conditions for each marker to push for either complete addition or lack of addition of the extra base, such as the recently described "PIG" tailing approach, in which a specific seven-base tail added to the reverse primer resulted in nearly 100% adenylation of the PCR products (Brownstein et al. 1996).

Quality Issues

When undertaking high-throughput genotyping it is important that high quality be maintained at the same time high quantity is achieved. It is possible to define the level of accuracy and reproducibility in a high-throughput genotyping environment. Reproducibility can be determined by running replicates of samples with known genotypes. In our laboratory a collection of 12 markers run in duplicate on 268 individuals yielded 35 allele discrepancies out of the 6432 alleles called, for an irreproducibility rate of 0.55%, or 1 in 180 alleles. Actual errors in the genotyping data can occur at various stages and are less easily identified. Mendel checking can be quite effective at revealing many errors but is dependent on the samples in question belonging to families with sufficient structure to detect such problems. Both null alleles (Tamaki et al. 1992; Callen et al. 1993) and new alleles (Smith et al. 1990) exist in human populations, and it is often difficult to distinguish a genotyping error from either of these. In a genome scan, null alleles will result in loss of information but not errors in the analysis, while patients exhibiting new alleles will be discarded from the analysis if the data are amenable to being checked for agreement with Mendelian inheritance patterns. Analytical approaches have been proposed that examine errors that are not obviously wrong using Mendelian inheritance (Ehm 1996), but these methods are dependent on having adequate family structure; as studies move to sib-pair analysis and TdT (Transmission disequilibrium Testing) there will be a certain level of errors resulting from null and new alleles that will go undetected. Some level of PCR failure is to be expected, and in our laboratory the average percent of missing genotypes is 2% for high-quality markers. These have been monitored over time in a variety of our studies and do not result from specific samples not amplifying consistently, but are truly random and probably represent the possibility of random failures at each step in the process.

PCR Contamination

The issue of PCR contamination must be dealt with in the high-throughput environment. We follow the procedures recommended by Kwok and Higuchi (1989). DNA to DNA and primer to primer cross-contamination during the manipulation of the genomic DNA samples and during PCR set up by the robotic pipettor are avoided by multiple water rinsing of the pipetting probes. Internal tests at Sequana (results not shown) have shown this to provide a sufficient template dilution effect to prevent production of detectable PCR product in control negative wells. A 10% bleach rinse is also used periodically to degrade remaining DNA (Hayatsu et al. 1971). The plates are capped following reaction setup and for thermal cycling. These caps are not removed until the plates are in a separate post-PCR processing area. Post-PCR sample pro-

cessing is carried out using robotic pipettors used only for pipetting PCR products, with multiple water rinsing between samples. By separating physically the areas for reaction set up and post-PCR handling and providing each area with dedicated pipetting systems, we have obviated the need for more complex contamination avoidance measures such as the use of UTP–uracil-*N*-glycosylase-containing systems (Longo et al. 1990).

A Case Study

An example of high-throughput genotyping applied to a specific project is a genome scan undertaken at Sequana on an Amish population in which diabetes was present as the polygenic disease of interest. This study consisted of 268 patient samples to be genotyped with 290 markers, for a total of 77,720 genotypes. The patient blood for this study was collected over the course of 6 months and arrived at our facility at the rate of 10–15 bloods a week. For efficiencies of scale, these samples were extracted in batches along with samples from other disease studies coming in at the same time. Upon completion of the DNA extraction of the entire population for this study, the samples were arranged into three 96 well plates, with each plate containing a positive control DNA [a Centre d'Etude du Polymorphisme Humain (CEPH) family member] and a negative well containing only water. An initial 180 replicates containing 20 ng per well of patient DNA were made of each of the three master plates and dried down for immediate use. A small number of the initial batch of replicate plates was tested with a panel of 10 microsatellite markers to be certain that the plates and DNA were of sufficient quality. The material passed that quality-control step, with all 10 markers amplifying on greater than 90% of the samples. Further replicates were made as required during the course of the scan, in batches of 100 at a time.

The genotyping began with a set of 290 markers arranged into 40 multiplexed sets, with an average density of 7.25 markers per set. These markers had been chosen by the following criteria: All had a polymorphism information content (PIC) greater than 0.7; the markers covered the human genome at an average density of 15 cM; and there were no gaps larger than 30 cM. The lab work for the initial amplification and gel analysis of all 290 markers on the replicates made from the three 96-well plates was done by three people and took 8 weeks to complete. The average time for PCR and gel analysis was two days per multiplex set per technician, but two concurrent sets could be staggered one day apart and run at the same time, which resulted in higher throughput. Upon completion of the first pass with the 40 sets of 290 markers each of the three master plates was examined for any markers among the 290 that had failed on that plate. We found an average of 1 in 10 markers randomly failed on a plate as a result of human error, poor PCR, or poor resolution on the sequencing gel. That failure was random across the three master plates, therefore plate-specific fill-in sets were designed to complete the 290 markers needed for each plate. The density of markers in these fill-in sets was 3.5 markers per set on average, a number that was constrained by the fact that these markers had already been optimized to fit in the initial sets and could not be packed as densely in the custom-designed fill-in sets. A further month was required to generate these fill-in data.

At this point letter alleles were assigned for each marker from the genotypes for all three combined plates. After the allele-calling process was completed for each marker, a program was run to check the Mendelian inheritance of each marker in the families in this study. Using the output from the Mendel checking program, the original gel files and genotyping data for all of the individuals from any pedigrees that showed Mendel errors were re-examined. Obvious genotyping errors that could be corrected, such as miscalled peaks, missed peaks, and lane tracking errors, were edited and those corrections were put into the data base. Individuals for whom no obvious errors in the genotyping could be identified were "turned off" in the data base—i.e., the data was left in the raw data files but was flagged to not be used in any subsequent genetic analysis. The process of allele calling took an average of 5 min per marker, and Mendel editing (correcting of any Mendel errors) took an additional 30 min per marker. Allele calling and Mendel checking/editing for all 290 markers took one technician a month to complete. In summary, the entire process of genotyping this population of 268 individuals with 290 markers took 10 months: 6 months to collect and process the sample DNAs, 3 months to generate all the raw genotypes, and 1 month to clean up and edit the final genotype data so that it was ready for linkage analysis.

Outline of the High-throughput Genotyping at Sequana Therapeutics

Bloods are received and each sample is labeled with a unique bar code.

▼

The DNA is extracted from the blood samples, and all of the DNAs for a study are assembled into 96-well master plates.

▼

The DNA concentrations are measured using a 96-well fluorometer and the DNA concentration of all the samples is adjusted to 4 ng/μl.

▼

Multiple replicate plates are made from each master plate, with 20 ng of DNA in each well. These replicated plates are dried down for storage. A small number of plates is tested for quality assurance with a panel of 10 markers before a genome scan is begun.

▼

PCR of individual markers is performed on the dried-down 96-well plates. These markers have been chosen previously and arranged to form multiplexable sets for analysis on a fluorescent sequencer.

▼

The PCR products are measured from each plate and the products from 10–12 plates are pooled together at the appropriate relative concentrations. Loading dye containing a size standard is added to each well.

▼

The pooled multiplexed markers from a single 96-well plate are run on three ABI 373 instruments (32 wells per instrument). The gel file output from each ABI is sent to the LIMS data base.

▼

The gels are checked for correct tracking and size standard calling. Genescan and Genotyper programs are run.

▼

When a given marker has been run on all plates, the letter alleles are defined and the data is checked for correct Mendelian inheritance patterns.

Future Developments

Improvements to current genotyping methods are likely to be made in the short term by incremental improvements to the microsatellite methods and in the long term by the use of single nucleotide polymorphism typing.

Short-term Developments

Higher density microplates (384 wells) for PCR should increase the throughput of thermal cycling and should also reduce reagent costs by allowing reaction volumes as low as 2 μl. Such microvolume PCR reactions may be performed in capillaries or on miniaturized sample handling formats. Multicapillary electrophoresis systems (Wang et al. 1995) will reduce dramatically the labor costs associated with pouring and loading slab gel machines as well eliminate the labor associated with the lane retracking necessary when using slab gels. Improvements in dye chemistries have been proposed recently. Problems with current dyes such as spectral overlap and different quantum yields are being addressed by newer dyes such as the energy transfer dyes (Ju et al. 1996), which also require less PCR product per lane, allowing use of smaller-volume PCR reactions. Increases in the numbers of dyes available and decreased spectral overlap will allow denser multiplexed sets of markers. Finally, robotic integration of 96-channel pipettors, high-density thermal cyclers, and, eventually, 96-channel capillary electrophoresis should allow 24-hr fully automated robotic operation with high data quality and minimal sample manipulation errors.

Long-term Improvements

Single nucleotide polymorphism (SNP) typing, also known as biallelic marker typing, has been proposed as a method that eventually may replace microsatellite markers. Biallelic markers are more easily automated, give less ambiguous results than microsatellites, and allow for easier data processing and analysis. All of these features come with the tradeoff of SNPs being substantially less informative than microsatellite markers. However, by increasing the density of SNP markers and by using SNPs with reasonable polymorphism content, this lack of information can be compensated for. The utility of typing known mutations using these methods for diagnostic use or for association studies has been demonstrated;

however, the power to detect linkage using a high-density SNP genome scan has not yet been shown. Use of physically close clusters of SNPs to form haplotypes may improve the informativeness of this method. With several thousand mapped SNP markers, genome scanning may be carried out on DNA array chips (Fodor et al. 1993) given the requisite assay development. Scaling up current microplate methods (Delahunty et al. 1995) offers an alternative approach.

Conclusion

High-throughput genotyping has become routine in a small number of laboratories, and this review has attempted to give an overview from one such location on the current status of the methods involved. By running one 96-well plate per day with 10 markers a single technician can perform ~1000 genotypes per day. Allowing for data management and analysis time, much of which can be accomplished in between wet lab processes, a single technician can produce 5000 genotypes per week. The use of robotics for many of the wet lab procedures in high-throughput genotyping allows the laboratory personnel to spend >50% of their time doing data processing and analysis. Skilled personnel who have expertise with robotics and LIMS type data analysis are needed for high-throughput genotyping, and such personnel are not abundant in the current work force. The cost per genotype as estimated by ours and other laboratories can be in the range of $1 per genotype, and if PCR volume can be reduced in the future that cost would be decreased. The typical cost to set up a high-throughput laboratory with a minimum of three genotyping instruments and incorporating robotics and informatics components is >$500,000. Such a cost, although prohibitive to the average academic laboratory, is within the budget of a core facility. Current high-throughput genotyping is semiautomated; a fully automated procedure could increase throughput, decrease cost, and minimize errors that inevitably result from human sample handling.

Many of the current human gene hunts involve complex phenotypes (diabetes, schizophrenia, cardiovascular diseases), and the numbers of genotypes required to map the genes involved will be much higher than efforts in the past on simpler genetic diseases. Only by employing high-throughput methods will it be possible to identify the genes involved in these types of diseases.

Note Concerning Software

Several types of software mentioned in this paper were written at Sequana. This software is being considered for commercial release, but is not currently available. Interested parties should contact the authors at Sequana for more details on the future release dates and availability of this software.

REFERENCES

Adams, M.D., A.R. Kerlavage, J.M. Kelley, J.D. Gocayne, C. Fields, C.M. Fraser, and J.C. Venter. 1994. A model for high throughput automated DNA sequencing and analysis core facilities. *Nature* **368:** 474–475.

Ahn, S.J., J. Costa, and J.R. Emanuel. 1996. PicoGreen quantitation of DNA: Effective evaluation of samples pre- or post-PCR. *Nuc. Acids Res.* **24:** 2623–2625.

Botstein, D., R. White, M. Skolnick, and R. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32:** 314–331.

Brownstein, M.J., J.D. Carpten, and J.R. Smith. 1996. Modulation of non-templated nucleotide addition by Taq DNA Polymerase: Primer modifications that facilitate genotyping. *BioTechniques* **20:** 1004–1010.

Callen, D.F., A.D. Thompson, Y. Shen, H.A. Phillips, R.I. Richards, J.C. Mulley, and G.R. Sutherland. 1993. Incidence and origin of null alleles in the (AC)n microsatellite markers. *Am. J. Hum. Genet.* **52:** 922–927.

Cooperative Human Linkage Center. 1994. A comprehensive human linkage map with centimorgan density. *Science* **265:** 2049–2054.

Delahunty, C.M., W. Ankener, S. Brainerd, D. Nickerson, and I. Mononen . 1995. Finnish-type aspartylglucosaminuria detected by oligonucleotide ligation assay. *Clin. Chem.* **41:** 59–61.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Dourot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette, and J. Weissenbach. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380:** 152–154.

Dietrich, W.F., J. Miller, R. Steen, M.A. Merchant, D. Damron-Boles, Z. Husain, R. Dredge, M.J. Daly, K.A. Ingalls, T.J. O'Connor, et al. 1996. A comprehensive genetic map of the mouse genome. *Nature* **380:** 149–152.

Ehm, M.G., M. Kimmel, and R .Cottingham. 1996. Error detection for genetic data, using likelihood methods. *Am. J. Hum. Genet.* **58:** 225–234.

Elston, R.C., X. Guo, and L. Williams. 1996. Two-stage global search designs for linkage analysis. *Genet. Epidemiol.* (in press).

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, W.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J.D. Gocayne, J. Scott, R. Shirley, L.I. Liu, A. Glodek, J.M. Kelley, J.F. Weidman, C.A. Phillips, T. Spriggs, E. Hedblom, M.D. Cotton, T.R. Utterback, M.C. Hanna, D.T. Nguyen, D.M. Saudek, R.C. Brandon, L.D. Fine, J.L. Fritchman, J.L. Fuhrmann, N.S.M. Geoghagen, C.L. Gnehm, L.A. McDonald, K.V. Small, L.M. Fraser, H.O. Smith, and J.C. Venter. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512.

Fodor, S.P., R.P. Rava, X.C. Huang, A.C. Pease, C.P. Holmes, C.L. Adams. 1993. Multiplexed biochemical assays with biological chips. *Nature* **364:** 555–556.

Ginns, E., J. Ott, J. Egeland, C. Allen, C. Fann, D. Pauls, J. Weissenbach, J. Carulli, K. Falls, T. Keith, and S. Paul. 1996. A genome-wide search for chromosomal loci linked to bipolar affective disorder in the old Order Amish. *Nature Genet.* **12:** 431–435.

Ginot, F., I. Bordelais, S. Nguyen, and G. Gyapay. 1996. Correction of some genotyping errors in automated fluorescent microsatellite analysis by enzymatic removal of one base overhangs. *Nuc. Acids Res.* **24:** 540–541.

Golden, J.B., III, D. Torgersen, and C. Tibbetts. 1993. Pattern recognition for automated DNA sequencing: I. On-line signal conditioning and feature extraction for basecalling. *Intelligent Sys. Molec. Biol.* **1:** 136–144.

Hayatsu, H., S.K. Pan, and T. Ukita. 1971. Reaction of sodium hypochlorite with nucleic acids and their constituents. *Chem. Pharm. Bull.* **19:** 2189–2192.

Ju, J., A.N. Glazer, and A Mathies. 1996. Energy transfer primers: A new fluorescence labeling paradigm for DNA sequencing and analysis. *Nature Med.* **2:** 246–249.

Kwok, S. and R. Higuchi. 1989. Avoiding false positives with PCR . *Nature* **339:** 237–238.

Lander, E.S. and N.J. Schork. 1994. Genetic dissection of complex traits. *Science* **265:** 2037–2048.

Lin, Z., C. Xiangfeng, and H. Li. 1996. Multiplex genotype determination at a large number of gene loci. *Proc. Natl. Acad. Sci.* **93:** 2582–2587.

Longo, M.C., M.S. Berninger, and J.L. Hartley. 1990. Use of uracil DNA glycosylase to control carry-over contamination in polymerase chain reactions. *Gene* **93:** 125–128.

McIndoe, R.A., M.S. Linhardt, and L. Hood. 1995. Single-tube genomic DNA isolation from whole blood without pre-isolating white blood cells. *BioTechniques* **19:** 30–32.

Murray, J.C., K.H. Buetow, J.L. Weber, S. Ludwigsen, T. Scherpbier-Heddema, F. Manion, J. Quillen, V.C. Sheffield, S. Sunden, and G.M. Duyk. 1994. A comprehensive human linkage map with centimorgan density. *Science* **265:** 2049–2054.

Perlin, M., G. Lancia, and S. Ng. 1995. Toward fully automated genotyping: Genotyping microsatellite markers by deconvolution. *Am. J. Hum. Genet.* **57:** 1199–1210.

Reed, P., J. Davies, J. Copman, S. Bennett, S. Palmer, L. Pritchard, S. Gough, Y. Kawaguchi, H. Cordell, K. Balfour, S. Jenkins, E. Powell, A. Vignal, and J. Todd. 1994. Chromosome-specific microsatellite sets for fluorescence-based semi-automated genome mapping. *Nature Genet.* **7:** 390–395.

Shuber, F., V. Grondin, and K. Klinger. 1995. A simplified procedure for developing multiplex PCRs. *Genome Res.* **5:** 488–493.

Smith, J.C., R. Anwar, J. Riley, D. Jenner, A.F. Markham, and A.J. Jeffreys. 1990. Highly polymorphic minisatellite sequences: Allele frequencies and mutation rates for five locus-specific probes in a Caucasian population. *J. Forensic Sci. Soc.* **30:** 19–24.

Tamaki, K., D.G. Monckton, A. MacLeod; D.L. Neil, M. Allen, and A.J. Jeffreys. 1992. Minisatellite variant repeat (MVR) mapping: Analysis of "null" repeat units at D1S8. *Hum. Mol. Genet.* **1:** 401–406.

Utah Marker Development Group. 1995. A collection of ordered tetranucleotide-repeat markers from the human genome. *Am. J. Hum. Genet.* **57:** 619–628.

Wang, Y., J. Ju, B. Carpenter, J. Atherton, G. Sensabaugh, and R. Mathies. 1995. Rapid sizing of short tandem repeat alleles using capillary array electrophoresis and energy-transfer fluorescent primers. *Anal. Chem.* **67:** 1197–1203.

Weeks, D.E. and G.M. Lathrop. 1995. Polygenic disease: Methods for mapping complex disease traits. *Trends Genet.* **11:** 513–519.