1
Overview of Genotyping

Mónica Bayés and Ivo Glynne Gut

Introduction

Several types of variants exist in the human genome: single nucleotide polymorphisms (SNPs), short tandem repeats (STRs) also called microsatellites, small insertions or deletions (InDels), copy number variants (CNVs) and other structural variants (SVs) (Figure 1.1). SNPs are changes in a single base at a specific position in the genome, in most cases with only two alleles (Brookes 1999). By definition the rarer allele should be more abundant than 1% in the general population otherwise referred to as mutations. SNPs are found at a frequency of about one every 100–300 bases in the human genome. Since the completion of the Human Genome Project (HGP) (International Human Genome Sequencing Consortium 2004), SNPs have been discovered at an unprecedented rate and currently there are more than 24 million human reference SNP (rs) entries in the most extensive SNP database (dbSNP Build 132, www.ncbi.nlm.nih.gov/projects/SNP/). SNPs, however, are not randomly distributed across the genome and occur much less frequently in coding sequences than in noncoding regions. SNPs located in regulatory or protein coding regions are more likely to alter the biological function of a gene than those in intergenic regions.

Genotyping is the process of assignment of different variants in an otherwise conserved DNA region. The relative simplicity of methods for SNP genotyping, the abundance of SNPs in the human genome and their low mutation rates have made them very popular in the past decade. SNP genotyping has currently many applications: disease gene localization and identification of disease-causing variants, quantitative trait loci (QTL) mapping, pharmacogenetics, identity testing based on
a. Single nucleotide polymorphisms (SNPs)

\[
\begin{align*}
\text{T} & \text{C} \text{G} \text{A} \\
\text{A} & \text{G} \text{C} \text{T} \\
\text{T} & \text{T} \text{G} \text{A} \\
\text{A} & \text{A} \text{G} \text{C} \text{T}
\end{align*}
\]

b. Short tandem repeats (STRs)

\[
\begin{align*}
\text{C} & \text{A} \text{C} \text{A} \text{C} \text{A} \text{C} \text{A} \text{C} \\
\text{G} & \text{T} \text{G} \text{T} \text{G} \text{T} \text{G} \text{T} \text{G} \text{T} \\
\text{C} & \text{A} \text{C} \text{A} \text{C} \text{A} \\
\text{G} & \text{T} \text{G} \text{T} \text{G} \text{T} \text{G} \text{T}
\end{align*}
\]

c. Small insertions or deletions (InDels)

\[
\begin{align*}
\text{T} & \text{C} \text{G} \text{A} \text{T} \text{A} \text{T} \\
\text{A} & \text{G} \text{C} \text{T} \text{A} \text{A} \text{T} \text{G} \\
\text{T} & \text{T} \text{A} \text{C} \\
\text{A} & \text{A} \text{T} \text{G}
\end{align*}
\]

d. Copy number variants (CNVs)

\[
\begin{align*}
\text{seg1} & \text{seg2} \text{seg3} \\
\text{seg1} & \text{seg2}
\end{align*}
\]

e. Other structural variants (inversion)

\[
\begin{align*}
\text{seg1} & \text{seg2} \\
\text{seg1} & \text{seg2}
\end{align*}
\]

**Figure 1.1** Types of genetic variants. Each arrow represents a DNA segment of more than 1 kb.

Genetic fingerprinting, just to mention the major ones. Genotyping applications extend beyond human genetics to animals and plants.

Although some SNP alleles confer susceptibility to complex disorders (asthma, cardiovascular disease, diabetes, etc.), most SNPs are not solely responsible for a disease state. Instead, they serve as biological markers for identifying disease-related variants on the human genome map, based on the fact that alleles of SNPs that are located nearby tend to be inherited together (Jorde 1995). This is termed linkage...
disequilibrium. For disease gene identification two basic strategies are applied. In the linkage study, related individuals are genotyped with several hundreds to thousands of polymorphisms distributed throughout the genome and attempts are made to identify genetic markers that cosegregate with the disease. Genetic linkage methods have been applied successfully to identify the mutated gene in Mendelian diseases (Risch 1991). If investigating the genetic basis of complex disorders, the association or linkage disequilibrium approach is more powerful (Risch and Merikangas 1996). It involves establishing genotype–phenotype correlations in unrelated individuals that are solely selected on the basis of being affected by a phenotype or not (Clark 2003).

Genetic association studies require a large number of samples to achieve statistically significant results that indicate that a particular allele in a particular region of the genome confers an increased risk of developing the disorder. Many association studies based on the analysis of candidate genes that involve genotyping of tens or hundreds of SNPs in hundreds or thousands of samples have been published. In the past five years, the ability to assay for more than 100,000 SNPs distributed across the genome has enabled the systematic study of complex disorders under a whole genome approach, without any preconceived hypothesis or candidates. Successful genome-wide association studies (GWAS) have been conducted for common diseases such as age-related macular degeneration, rheumatoid arthritis, asthma, Crohn’s disease, bipolar disorder, coronary heart disease, type 1 and type 2 diabetes among many others (Klein et al. 2005; Wellcome Trust Case Control Consortium 2007; Moffatt et al. 2007, 2010; Hindorff et al. 2009). A list of all GWAS and associated polymorphisms is kept up to date at www.genome.gov/gwastudies.

In the next sections the most popular methods and platforms for SNP genotyping are discussed, highlighting some practical aspects. Other related applications such as methylation, copy number analysis and second generation sequencing using the same underlying molecular approaches are covered thereafter.

**Methods for interrogating SNPs**

There are many mature SNP genotyping technologies that have been integrated into large-scale genotyping operations. SNP genotyping methods are still being improved, perfected, integrated and new methods are emerging to satisfy the needs of genomics and epidemiology. No one SNP genotyping method fulfils the requirements of every study that might be undertaken. The choice of a method depends on the scale of the envisioned genotyping project and the resources available. A project might require genotyping of a limited number of SNP markers in a large population or the analysis of a large number of SNP markers in a few samples. Flexibility in choice of SNP markers and DNAs to be genotyped or the possibility to precisely quantify an allele frequency in pooled DNA samples might also be issues.

SNP genotyping methods are very diverse (Syvänen 2001; Kim and Misra 2007). Broadly, each method can be separated into two elements, the biochemical method
for discriminating SNP alleles and the actual analysis or measurement of the allele-specific products, which can be an array reader, a plate reader, a mass spectrometer, a gel separator/reader system, or other. In addition, most technologies also require a PCR amplification step to increase the number of target SNP-containing DNA molecules and to reduce the complexity of the template material used for the allele discrimination step.

The most popular methods for allele discrimination are restriction endonuclease digestion, primer extension, hybridization and oligonucleotide ligation (Figure 1.2a).

**Restriction endonuclease digestion**

Restriction fragment length polymorphisms (RFLPs) are one of the first typing methods described and by far predate the coining of the term SNP (Botstein et al. 1980). Restriction endonuclease digestion is still a common format for SNP genotyping in a standard laboratory (Parsons and Heflich 1997). PCR products are digested with restriction endonucleases that are specifically chosen for the base change at the position of the SNP, resulting in a restriction cut for one allele but not the other (Figure 1.2a). In some cases, specific restriction sites can be created during the amplification step by using primers with minor changes in the sequence. Digestion patterns are used for allele assignment after gel electrophoresis. Major limitations of the restriction method are that it is only applicable to a fraction of SNPs and that it does not lend itself to automation.

**Primer extension**

Primer extension is a stable and reliable way of distinguishing alleles of a SNP. Nucleotides are added by a DNA polymerase generating allele-specific products (Syvänen 1999). Allele-specific primer extension (ASPE) is based on the ability of DNA polymerases to extend with high efficiency those oligonucleotides with 3’ perfectly matched ends (Figure 1.2a). It requires two allele-specific primers that have the nucleotide that corresponds to the allelic variant at their 3’ ends. In single base primer extension (SBE) an oligonucleotide hybridizes immediately before the SNP nucleotide and the DNA polymerase incorporates a single nucleotide that is complementary to the SNP allele (Figure 1.2a). SBE uses dideoxynucleotides (ddNTP) as terminators.

**Hybridization**

Alleles differing by one base can be distinguished by hybridizing complementary oligonucleotide sequences to the target DNA (ASO or allele-specific oligonucleotide...
Figure 1.2 SNP genotyping technologies separated into allele discrimination methods (A) and detection of allele-specific products (B). Arrows denote genotyping assays that combine different allele discrimination and detection methods. 1 Restriction endonuclease digestion; 2 SNPlex; 3 iPLEX GOLD assay; 4 GoldenGate assay; 5 Infinium assay; 6 TaqMan assay; 7 GeneChip assay.
hybridization), without any enzymatic reaction (Figure 1.2a). As the two alleles of a SNP are very similar in sequence, significant cross-talk can occur. Several approaches have been taken to overcome this problem: the use of multiple probes per SNP, the use of modified oligonucleotides such as peptide nucleic acids (PNAs) (Egholm et al. 1993) or locked nucleic acids (LNAs) (Ørum et al. 1999) that increase stability of DNA–DNA complexes, the real-time monitoring of the hybridization kinetics or the combination of hybridization and 5′ nuclease activity of polymerases.

**Oligonucleotide ligation (OLA)**

OLA relies on the specificity of DNA ligases to repair DNA nicks. For OLA, two oligonucleotides adjacent to each other are ligated enzymatically by a DNA ligase when the bases next to the ligation position are fully complementary to the template strand (Barany 1991; Jarvius et al. 2003) (Figure 1.2a). The assay requires three probes to be designed: two allele-specific probes that have at their 3′ ends the nucleotide complementary to the SNP variants and one common probe that anneals to the target DNA that is immediately adjacent. Padlock is a variant of OLA that employs two allele-specific oligonucleotides with target complementary sequences separated by a linker. When perfectly annealed to the target sequence, padlock probes are circularized by ligation (Nilsson et al. 1994).

Major detection methods include gel electrophoresis, mass spectrometry, fluorescence analysis, and chemiluminescence detection (Figure 1.2b). Nearly all of the above-described methods for allele-distinction have been combined with all of these analysis formats.

**Gel electrophoresis**

Allele-specific DNA fragments of different sizes can be separated by electrophoretic migration through gels (Szántai and Guttman 2006). Throughput and resolution can be increased if 'gel-filled' capillaries are used. Advantages of capillary systems over slab gel systems include the potential for 24-hour unsupervised operation, the elimination of cumbersome gel pouring and loading, and that no lane tracking is required. Instrumentation with 96- or 384-capillaries is commercially available.

**Mass spectrometry**

MALDI-TOF MS (matrix-assisted laser desorption/ ionization time-of-flight mass spectrometry) can be used to measure the mass of the allele-specific products. It has been demonstrated as an analysis tool for SNP genotyping (Haff and Smirnov 1997; Tost and Gut 2005). The allele-specific products are deposited onto a matrix on the
surface of a chip, ionized with a short laser pulse and accelerated towards the detector (Jurinke et al. 2002). The time-of-flight of a product to the detector is directly related to its mass. High resolution and speed are major advantages of the MALDI-TOF MS detection method. Resolution of the current generation of mass spectrometers allows the distinction of base substitutions in the range of 1.000–6.000 Da (this corresponds to product sizes of 3–20 bases, the smallest mass difference for a base change thymine to adenine is 9 Da).

**Fluorescent analysis**

Allele-specific products can be labelled with different fluorescent dyes and detected using fluorescent readout systems, either microtitre plate or array based (Landegren et al. 1998). Most readers use a white light source and optical filters to select specific excitation and emission wavelengths. Some of them can also measure parameters such as fluorescence polarization (FP, measures the increase in polarization of fluorescence caused by the decreased mobility of larger molecules) (P.Y. Kwok 2002) and Förster resonance energy transfer (FRET, measures the changes in fluorescence due the separation of two dyes of a donor/acceptor system) (Tong et al. 2001). Most popular fluorescent dyes used in SNP genotyping are Cy3 and Cy5.

Most current genotyping methods are generally based on the combination of one of the allelic discrimination and one of the detection methods described above (Figure 1.2). Often very different methods share elements, for example, reading out a fluorescent tag in a plate reader, or the primer extension method, which can be analysed in many different analysis formats.

**Commercial platforms for SNP genotyping**

A plethora of SNP genotyping platforms is currently commercially available (Ragoussis 2009). Many of them require purchasing expensive proprietary equipment and expensive laboratory set-up. However, they offer streamlined laboratory and analysis workflows. They range from individual SNP genotyping platforms (Life Technologies TaqMan) to focused content genotyping (Sequenom iPLEX Gold, Illumina GoldenGate) and to platforms for whole genome genotyping (WGG) (Illumina Infinium and Affymetrix GeneChips) (Fan et al. 2006). WGG arrays contain from 100 000 to 2.5 million SNPs selected by different approaches and with minor allele frequencies >0.05 in the general population.

**TaqMan assay**

The TaqMan assay (Life Technologies, www.appliedbiosystems.com) is based on allele-specific hybridization coupled with the 5’ nuclease activity of Taq polymerase
MOLECULAR ANALYSIS AND GENOME DISCOVERY

during PCR (Holland et al. 1991; Livak 2003; Livak et al. 1995). The detection is performed by measuring the decrease of FRET from a donor fluorophore to an acceptor-quencher molecule. TaqMan probes are allele-specific probes labelled with a fluorescent reporter at the 5′ end and a common quencher attached to the 3′ end that virtually eliminates the fluorescence in the intact probe. Each assay uses two TaqMan probes that differ at the SNP site, and one pair of PCR primers. During PCR, successful hybridization of the TaqMan probe due to matching with one allele of the SNP results in its degradation by the 5′- to 3′-nuclease activity of the employed DNA polymerase whereby the fluorescent dye and quencher are separated, which promotes fluorescence. TaqMan probes can be designed to detect multiple nucleotide polymorphisms (MNPs) and insertion/deletions (InDels). Because of the simplicity in chemistry, the reaction set-up can be easily automated using liquid handling robots. The 7900HT Fast Real-Time PCR system (Life Technologies) allows up to eighty-four 384-well plates to be processed without manual intervention in less than 4 days. It is a very contamination-safe procedure as plates do not need to be opened after PCR for reading. In contrast, the limiting factors of the technology are the low SNP multiplexing level and the relatively high cost of the dual-labelled probes. Life Technologies has developed a library with 4.5 million genome-wide human TaqMan assays (of which 160 000 are validated assays) for which reagents are commercially available.

In the recent years, a couple of high-throughput real-time PCR instruments have been introduced. The Biomark system (Fluidigm, www.fluidigm.com) contains integrated fluidic circuits or ‘dynamic arrays’ that allow setting up 9216 genotyping reactions in a single experiment (Wang et al. 2009). The user has to simply dispense 96 DNA samples and 96 TaqMan genotyping assays and the dynamic array will then do the work of assembling the samples in all possible combinations. The OpenArray system (www.appliedbiosystems.com) (Morrison et al. 2006) can also perform SNP analysis using TaqMan probes. The OpenArray plate contains 3072 reaction through-holes generated by a differential coating process that deposits hydrophilic coatings on the interior of each through-hole and hydrophobic coatings on the exterior. This enables OpenArray plates to hold solutions in the open through-holes via capillary action. The company provides the researcher with OpenArray plates that are preloaded with the selected TaqMan probes (from 16 to 256 different assays per plate depending on the plate format). The main advantages of the Biomark and OpenArray systems compared to conventional thermocyclers are higher throughput, small sample requirement, low reagent consumption and less liquid handling.

*iPLEX GOLD assay*

The iPLEX Gold reaction (Sequenom, www.sequenom.com) is a method for detecting insertions, deletions, substitutions, and other polymorphisms that combines multiplex PCR followed by a single-base extension and MALDI-TOF MS detection
OVERVIEW OF GENOTYPING

(Jurinke et al. 2002; Oeth et al. 2009). After the PCR, remaining nucleotides are deactivated using shrimp alkaline phosphatase (SAP). The SAP cleaves a phosphate from the unincorporated dNTPs, converting them to dNDPs which renders them unavailable to future polymerization reactions. Next, a single base primer extension step is performed incorporating one of the four terminator nucleotides into the SNP site. The extension products are desalted and transferred onto chips containing 384 matrix spots. The allele-specific extension products of different masses are analysed using MALDI-TOF MS. In theory up to 40 different SNPs can be assayed together if the different allele-products have distinct masses; however, generally multiplexes on the order of 24 are more realistic. The whole lab workflow is highly automated and it takes less than 10 hours to process one 384 plate. The MassARRAY Analyzer 4 (Sequenom) can analyse from dozens to over 100,000 genotypes per day, and from tens to thousands of samples. Significant advantages of the method are that it requires standard unmodified oligonucleotides which are cheap and easy to come by. It is a very sensitive method with low input sample requirements and finally generates highly accurate data because it relies on the direct detection of the allele-specific product.

GoldenGate assay

The GoldenGate assay (Illumina, www.illumina.com) (Shen et al. 2005) can interrogate 48, 96, 144, 192, 384, 768 or 1536 SNPs simultaneously. The assay combines allele-specific primer extension and ligation for generating allele-specific products followed by PCR amplification with universal primers. Three oligonucleotides are designed for each SNP locus, two of which are allele-specific (ASO) with the SNP allele on their 3' end, and a locus specific oligonucleotide (LSO) that hybridizes several bases downstream the SNP site. The LSO primer also contains a unique address sequence that allows separating the SNP assay products for individual readout. In the protocol, during the hybridization process, the oligonucleotides hybridize to the genomic DNA that has been first immobilized on a solid support. The complementary ASO is extended and ligated to the LSO, providing high locus specificity. The ligated products are then amplified using universal PCR primers P1, P2 and P3. Primers P1 and P2 are specific for each ASO and carry a fluorescent tag that is used for allele calling.

The separation of the assay products in solution onto a solid format is done using Veracode technology (48, 96, 144, 192 or 384-plex) (Lin et al. 2009). It uses cylindrical glass microbeads (240 microns in length) with unique digital holographic codes and coated with capture oligonucleotides that are complementary to one of the addresses present in the PCR products. When excited by a laser, each VeraCode bead emits a unique holographic code image. The BeadXpress reader (Illumina) can identify the individual bead types and in addition detect the results from the two-colour
genotyping assay. The Veracode technology contains assay replicates of 20–30 beads per bead type, providing a high level of quality control.

**Infinium assay**

The Infinium II assay (Illumina, www.illumina.com) uses a two-colour SBE protocol for allelic discrimination coupled with the BeadChip technology for assay detection (Steemers and Gunderson 2007). Whole genome amplified (WGA) samples are hybridized to 50-mer oligonucleotide probes covalently attached to particular microspheres or beads that are randomly assembled in microwells on planar silica slides (BeadChips). After the hybridization, the SNP locus-specific oligonucleotides are extended with the corresponding fluorescently labelled dideoxynucleotides. The intensities of the bead’s fluorescence are detected by the iScan Reader (Illumina).

Currently available BeadChips for human allow profiling samples with 300,000 to 2.5 M SNPs distributed throughout the genome. SNP selection in these chips is based on results from the HapMap project (www.hapmap.org) providing high coverage across the genome (see ‘SNP databases’). New arrays with up to 5 M common and rare variants from the 1000 Genomes Project (www.1000genomes.org) are in development. The Infinium assay can also be used also to develop BeadChips with customized SNP content (iSelect). Genome-wide genotyping BeadChips are also available for other species such as cattle, pigs and dogs.

**GeneChip assay**

In the GeneChip assay (Affymetrix, www.affymetrix.com) allelic discrimination is achieved by direct hybridization of labelled DNA to arrays containing allele-specific oligonucleotides. These 25-mer probes are synthesized in an ordered fashion on a solid surface by a light-directed chemical process (photolithography) (Fodor et al. 1991). Oligonucleotides covering the complementary sequence of the two alleles of a SNP are on specific positions of the array. Multiple probes for each SNP are used to increase the genotyping accuracy. The hybridization pattern of all oligonucleotides spanning the SNP is used to evaluate positive and negative signals.

Genomic DNA is digested with a restriction endonuclease and ligated to adaptors that recognize the cohesive 4 bp overhangs. The ligation products are then amplified by PCR using a single universal primer and creating a reduced representation of the genome (Kennedy et al. 2003). Next, PCR amplicons are fragmented, end-labelled and hybridized to the array under stringent conditions. After extensive washing steps, the remaining fluorescence signal is automatically recorded by the GeneChip 3000 scanner (Affymetrix). A specific fluidics station and a hybridization oven are also required to carry out the procedure.
Affymetrix has developed several microarrays designed specifically to interrogate SNPs distributed throughout the human genome. The most comprehensive array, the Genome-Wide Human SNP Array 6.0 has 1.8 million genetic markers, including 906,600 SNPs. The median inter-marker distance over all 1.8 million SNP and copy number markers combined is less than 700 bases. Affymetrix has also launched a new high-throughput genotyping assay, the Axiom Genotyping Solution. It is based on a 96-sample format and can process more than 750 samples per week. The initial Axiom Genome-Wide Human Array contains more than 560,000 SNPs.

Other popular platforms for SNP genotyping are SNPstream (Beckman Coulter) and Pyrosequencing (Qiagen) (Table 1.1) (Syvänen 2001; Sobrino et al. 2005; Ragoussis 2009).

**Practical recommendations**

Different aspects have to be taken into consideration when setting-up a genotyping platform: DNA quality assessment, contamination control, automation and data quality control measures.

In high-throughput laboratories, liquid handling automation is essential both for the SNP allele-discrimination and allele-detection processes (Gut 2001). It not only speeds up the genotyping process but also reduces errors introduced by human handling and pipetting and minimizes the possibility of cross-contamination of samples. Many suppliers of laboratory robotics offer liquid handling robots that can be integrated into high-throughput genotyping workflows. In general, the ease of automation is directly correlated to the complexity of an SNP genotyping protocol. Steps such as gel-filtration and manipulation of magnetic beads can be more problematic to automate. Current liquid handling robots can support both plates and slide microarray formats.

One of the biggest challenges in running SNP genotyping at high-throughput is the management of the production line. A Laboratory Information Management System (LIMS) is a software tool for keeping track of samples, laboratory users, instruments, lab processes, quality standards, and results. Originally, LIMS were developed in-house but currently there are several commercial solutions available such as Biotracker (Ocimum Biosolutions), Geneus (GenoLogics) and StarLIMS (StarLIMS Corporation). Complete systems for the entire high-throughput SNP genotyping process, with automation and LIMS, are marketed as off-the-shelf products. Examples of this are systems from Affymetrix, Sequenom and Illumina. In addition, all platforms discussed in the previous section have developed analysis software for fully automatic scoring of alleles and genotypes and monitoring the performance of all controls (Figure 1.3).

One of the greatest concerns in optimizing a genotyping laboratory is to control for PCR contamination. The high-throughput and repetition of assays with common primer pairs can easily lead to amplification of cross-contamination. The
<table>
<thead>
<tr>
<th>Platform</th>
<th>Assay Name</th>
<th>Vendor</th>
<th>Allelic Discrimination</th>
<th>Detection</th>
<th>Multiplexing</th>
<th>Format</th>
<th>Throughput (GT per day)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PyroMark MD 7900HT</td>
<td>Pyrosequencing</td>
<td>Qiagen Life</td>
<td>Sequencing</td>
<td>Chemiluminescence</td>
<td>1</td>
<td>96 well plate</td>
<td>5000</td>
</tr>
<tr>
<td></td>
<td>TaqMan</td>
<td>Technologies</td>
<td>Hybridization + 5' nuclease</td>
<td>FRET</td>
<td>1</td>
<td>384 well plate</td>
<td>8064</td>
</tr>
<tr>
<td>Biomark</td>
<td>TaqMan</td>
<td>Fluidigm</td>
<td>Hybridization + 5' nuclease</td>
<td>FRET</td>
<td>1</td>
<td>9216 nanowell dynamic plate</td>
<td>27648</td>
</tr>
<tr>
<td>OpenArray</td>
<td>TaqMan</td>
<td>Life Technologies</td>
<td>Hybridization + 5' nuclease</td>
<td>FRET</td>
<td>1</td>
<td>3072 nanowell slide</td>
<td>98304</td>
</tr>
<tr>
<td>MassArray Compact Analyzer</td>
<td>iPlex Gold</td>
<td>Sequenom</td>
<td>SBE</td>
<td>Mass spectrometry</td>
<td>2–40</td>
<td>384 position Spectrochip</td>
<td>100000</td>
</tr>
<tr>
<td>SNPstream Imager</td>
<td>SNPstream</td>
<td>Beckman Coulter</td>
<td>SBE</td>
<td>Fluorescence</td>
<td>12–48</td>
<td>384 array plate</td>
<td>3 M</td>
</tr>
<tr>
<td>BeadXpress</td>
<td>GoldenGate Veracode</td>
<td>Illumina</td>
<td>ASPE</td>
<td>Fluorescence</td>
<td>48–384</td>
<td>96 well plate</td>
<td>110000</td>
</tr>
<tr>
<td>iScan</td>
<td>Infinium</td>
<td>Illumina</td>
<td>SBE</td>
<td>Fluorescence</td>
<td>3000–2.5 M</td>
<td>12–32 sample Beadchips</td>
<td>185 M</td>
</tr>
<tr>
<td>GeneChip 3000</td>
<td>GenChip</td>
<td>Affymetrix</td>
<td>Hybridization</td>
<td>Fluorescence</td>
<td>1.8 M</td>
<td>1 sample array</td>
<td>414 M</td>
</tr>
<tr>
<td>GeneTitan</td>
<td>Axiom</td>
<td>Affymetrix</td>
<td>Hybridization</td>
<td>Fluorescence</td>
<td>560,000</td>
<td>96 array plate</td>
<td>430 M</td>
</tr>
</tbody>
</table>

*GT: genotypes; assuming a single instrument with autoloader when available and liquid handling automation.
Figure 1.3 Genotype cluster plot for one SNP genotyped across 270 samples using the GoldenGate assay and the Veracode technology. Each data point represents one sample, the y-axis is normalized signal intensity (sum of intensities of the two fluorescent signals) and the x-axis is the theta value that indicates the allelic angle. The software automatically clusters the DNA samples into two homozygous clusters (red and blue) and a heterozygote cluster (yellow). Points depicted in black are unsuccessfully genotyped samples.

most important recommendation for preventing contamination is to maintain separate areas, dedicated equipment and supplies for pre-PCR steps (sample preparation and PCR set-up) and post-PCR steps (thermocycling and analysis of PCR products). The rule of thumb should be never to bring amplified PCR products into the PCR set-up area. Uracil-DNA glycosylase (UNG) can also be used to prevent carryover contamination of the PCR products (Longo et al. 1990). By using dUTP instead of dTTP in all PCRs, UNG treatment can prevent the reamplification of carryover PCR products by removing any uracil incorporated into the amplicons and then cleaving the DNA at the created abasic sites. Finally, laboratory practices such as the use of disposable filter tips, positive-displacement pipettes, non-contact dispensing options and periodical lab and instrument cleaning also help reduce the risk of carryover contamination (S. Kwok and Higuchi 1989).

Genotyping errors have a deleterious effect on the statistical analysis of the data. To address this issue several quality controls should be carried out in each genotyping experiment: negative controls to monitor cross-contamination, positive controls to check concordance with publicly available data and replicate DNA samples to account for intra- and inter-plate reproducibility (Pompanon et al. 2005). Analysis statistics such as deviation from the Hardy–Weinberg equilibrium, Mendelian inconsistencies in pedigrees or the number of inferred recombinants can also be of great
value for identifying potential genotyping errors. Finally, it is also recommended to check regularly a subset of SNPs with at least two different platforms to evaluate platform performance (Lahermo et al. 2006). Most of the common genotyping platform vendors described in the previous section provide extensive quality measures of several protocol steps to ensure an overall assay accuracy of >99%.

Monitoring the quality of DNA samples prior to genotyping is the most important factor for achieving optimum genotyping results. Low quantity and/or quality DNA samples negatively affect the call rate (proportion of SNPs receiving a genotype call) and also lead to a higher number of genotyping errors. DNA needs to be in a reaction with sufficient representation of the two alleles – 1 ng of genomic human DNA corresponds to 300 copies of the genome. This is more than sufficient starting material for genotyping an individual polymorphism. Reducing the amount of genomic DNA starting material may result in allele-dropout and an increased risk of contamination. High-multiplex genotyping methods tend to be cheap in terms of DNA requirements per polymorphism.

DNA quantification and quality control is often conducted with a UV spectrophotometer at wavelengths of 260 nm and 280 nm. The ratio of absorbance readings at the two wavelengths should be between 1.8 and 2.2, while protein contamination can be assessed by measuring the A260/230 ratio (1.6–2.4). A more precise quantification of the double-stranded DNA target can obtained using a fluorescent nucleic acid stain such as Picogreen (Invitrogen) and a fluorometer (excitation and emission wavelengths of 502 nm and 523 nm, respectively) or by real-time qPCR using a single-copy gene as a copy number reference. Finally, the integrity and molecular weight of DNA are measured by gel electrophoresis using either agarose gels or an instrument such as the Agilent Bioanalyzer.

**SNP databases**

SNP databases such as dbSNP and HapMap are essential resources for the study of human complex disorders and for evolutionary studies.

The Single Nucleotide Polymorphism database (dbSNP, www.ncbi.nlm.nih.gov/projects/SNP) was launched in 1998 as a public-domain archive of simple genetic polymorphisms. It contains SNP-related information such as SNP flanking DNA sequences, alleles, allele frequencies, validation status and functional relationships to genes (Sherry et al. 2001). As of build 132 (September 2010), dbSNP has collected over 244 million submissions corresponding to more than 87 million reference SNP clusters (refSNP) from 100 organisms, including *Homo sapiens*, *Mus musculus*, *Gallus gallus*, *Oryza sativa*, *Zea mays* and many other species. A full list of organisms and the number of reference SNP clusters for each can be found at www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi. The data of dbSNP is also included in repositories such as ENSEMBL (www.ensembl.org) and the UCSC Genome Browser (genome.ucsc.edu).
The international HapMap project (www.hapmap.org) started in 2002 with the aim of cataloguing the vast amount of genetic variation in humans and describing how it is organized in short stretches of strong linkage disequilibrium (haplotype blocks) that coincide with ancient ancestral recombination events (Couzin 2002; Pääbo 2003). Since then more than 3 million SNPs (with an average density of 1 SNP per kb and minor allele frequency >0.05) have been analysed in 270 individuals from populations with African, Asian and European ancestry (International HapMap Consortium 2007). HapMap results provide researchers with a selection of SNP markers that tag haplotype blocks to reduce the number of genotypes that have to be measured for a genome-wide association study (more than 500 000 tag SNPs are required to capture all Phase II SNPs with $r^2 \geq 0.8$ in a population from Northern Europe (CEU)). In Phase III, 1,184 reference individuals from 11 global populations have been genotyped for 1.6 million SNPs (International HapMap Consortium 2010).

Recent improvements in sequencing technology (see ‘Second generation sequencing’) fostered the creation of the 1000 Genomes Project (1000 GP, www.1000genomes.org) in 2008. The aim of the project is to obtain a nearly complete catalogue of all human genetic variations with frequencies greater than 1% by sequencing the genomes of 2500 individuals from different populations. Data from three pilot projects is already available: low coverage sequencing of 180 individuals, sequencing at deep coverage of six individuals and sequencing gene regions in 900 individuals. 1000 GP data is further improving the process of identification of disease-associated regions.

Resources such as dbSNP, HapMap and 1000 GP have unquestionably saved medical researchers a lot of time and cost in their projects. All of the information generated by these projects is rapidly released into the public domain. In addition, DNA samples used in the HapMap and 1000 Genomes projects are also publicly available through Coriell Institute (ccr.coriell.org).

**Methylation analysis**

In mammals, epigenetic modifications are known to play a critical role in the regulation of gene expression across the genome and in maintaining genomic stability (Bernstein et al. 2007). Many studies have implicated aberrant methylation in the aetiology of common human diseases, including cancer, multiple sclerosis, diabetes and schizophrenia (Tost 2010). Alterations in DNA methylation can be used as biomarkers for early cancer detection, to discriminate among tumour subtypes or to predict disease outcome (Shames et al. 2007).

In animals and plants, methylation occurs preferentially at the 5 position of cytosines within CpG dinucleotides. Reversible methylation of cytosine in the sequence context CpG adds a dynamic component to DNA because it can act as a switch of transcription of a gene if the CpGs are in the promoter region of the gene (Suzuki and Bird 2008). CpGs that can be methylated or not are termed methylation
variable positions (MVPs). Methylation information is lost during PCR or primer extension reactions. Nonetheless, measuring the degree of DNA methylation can be done by virtually any SNP genotyping method if a genomic DNA sample is prior treated with bisulfite (L. Shen and Waterland 2007). Bisulfite treatment of DNA results in conversion of non-methylated C into U, while methylated C remains unchanged. After bisulfite conversion and PCR amplification, determination of the degree of methylation at a given MVP in the genomic DNA sample can be achieved by quantifying the degree of C and T at that position. The quantitative resolution of the genotyping method determines the accuracy of measurement that can be achieved. One of the most widely used methods for quantifying MVPs is pyrosequencing (Dupont et al. 2004). Pyrosequencing is a sequencing-by-synthesis method based on the detection of pyrophosphate (PPi) which is released during DNA synthesis in a quantity equimolar to the amount of incorporated nucleotide. Assay set-up is straight forward and the accuracy of quantification is better than 2% if using the PyroMark MD instrument (Qiagen, www.pyrosequencing.com). MALDI-TOF mass spectrometry can also be used to detect cytosine methylation using bisulfite conversion biochemistry, followed by PCR and base-specific cleavage process that generate a distinct signal pattern from the methylated and non-methylated template DNA (Sequenom, www.sequenom.com) (Ehrich et al. 2005).

There are several array-based methods to identify global patterns of CpG methylation (Beck and Rakyan 2008; Laird 2010). The most popular one is the Human-Methylation27 and HumanMethylation450 BeadChips developed by Illumina (www.illumina.com). It uses the Infinium II assay (see ‘Commercial platforms for SNP genotyping’) to interrogate bisulfite converted DNA for up to 450 000 CpG methylation sites from 99% of RefSeq genes. It reliably detects a difference of 20% in methylation.

The first complete maps of DNA methylation (or methylome) with a one base pair resolution were obtained for Arabidopsis thaliana (Lister et al. 2008) and two human cell types (Lister et al. 2009) using bisulfite treatment and second generation sequencing technologies (see ‘Second generation sequencing’).

Copy number variation analysis

Copy number variants (CNVs) are a common form of genetic variation in human populations (Database of Genomic Variants, projects.tcag.ca/variation) (Redon et al. 2006; McCarroll et al. 2008). By analogy to the standard definition of a SNP, a CNV is a copy number polymorphism that ranges from one kilobase to several megabases in size and has a minor allele frequency of 1% or greater.

CNVs may alter gene function by affecting gene dosage, positional effect or by directly interrupting genes. Although most CNVs are neutral polymorphic variants, some of them have been demonstrated to be associated with human diseases such as autism, schizophrenia, mental retardation or psoriasis (Zhang et al. 2009). Screening
for CNVs, in addition to SNP genotyping, in disease gene identification studies is a major trend in current research projects.

There are several methods to quantify copy number variation across the genome for research and diagnostic purposes. Array comparative genome hybridization (array CGH) is a powerful tool for discovering previously unrecognized submicroscopic aberrations in cancer and genomic disorders just by measuring hybridization intensities (Carter 2007; Gresham et al. 2008). Current arrays for CGH contain up to 4.2 million oligonucleotide probes that enable genome-wide detection of CNVs down to 1.5–5 Kb resolution (Nimblegen, www.nimblegen.com) (Agilent, www.agilent.com).

Current WGG SNP arrays, although originally designed for SNP genotyping, can also be used to capture CNVs at a genome-wide scale (Cooper et al. 2008; Carter 2007). SNP-CGH, unlike conventional array CGH, can detect copy neutral genetic abnormalities such as uniparental disomy (UPD) and loss of heterozygosity (LOH). HumanOmin2.5 Beadarrays (Illumina) contain nearly 2.5 million genetic markers (median spacing between markers is 1.5 kb), including 60,000 CNV-targeted markers. The human SNP Array 6.0 (Affymetrix) features 1.8 million genetic markers, including more than 946,000 probes for the detection of copy number variation. The Cytogenetics Whole-Genome 2.7 M Array (Affymetrix) provides a comprehensive analysis of structural variation. It contains 2.7 million copy number markers, including 2.3 million of non-polymorphic markers and 400,000 SNPs.

Over the past few years, a number of software packages have been developed to detect changes in copy number by using SNP-CGH data and analysing total signal intensities and allelic intensity ratios (Winchester et al. 2009). Different methods based on Hidden Markov Models, circular binary segmentation or mixed models are used to detect CNV segment boundaries. In a second step particular segments that are different in copy number compared with values from a reference individual or group of individuals are identified. For robust detection, a CNV interval requires significant ratio shifts in several consecutive probes.

Because of the relatively low signal-to-noise ratio and high experimental variation that characterizes many of the platforms, candidate CNVs should be validated by alternative low-throughput techniques such as multiplex ligation-dependent probe amplification (MLPA) or real-time PCR.

Second generation sequencing technologies

Second generation sequencing (2ndGS) technologies have dramatically increased the throughput and reduced the costs of DNA sequencing compared with conventional Sanger sequencing methods. Although the cost is still one order of magnitude higher than whole genome genotyping (WGG), the general goal of sequencing a human genome for $1000 in less than one day seems realistic in the near-term. This will open the door for WGG by whole genome sequencing (WGS).
2ndGS platforms allow sequencing of millions of clonally amplified and spatially separated DNA fragments simultaneously. The sequencing process itself is a repetition of cycles of enzymatic reactions (polymerase-based nucleotide incorporations or oligonucleotide ligations, depending on the platform) and imaging-based data collection (Shendure and Ji 2008). The resulting sequence tags or reads are then aligned to a reference genome and genetic polymorphisms (SNPs, InDels, SVs) identified. 2ndGS instruments from Roche/454 (Genome FLX, www.454.com), Illumina (HiSeq2000, www.illumina.com) and Life Technologies (SOLiD 5500xl, solid.appliedbiosystems.com) are commercially available. The best performing of these instruments can generate tens of gigabases of sequence per day.

The main limitations of some of the 2ndGS technologies for WGS are the short read lengths (50 to 500 bp depending on the platform) and the relatively high error rate. The first problem can be partially circumvented by using an approach that generates sequences from both ends of each DNA molecule (paired-end or mate-pair sequencing) thus facilitating the alignment process. Nevertheless, analysis of large and highly repetitive regions is still not feasible. Increasing the depth of coverage (obtaining multiple reads from the same region) is the best option for improving the consensus read accuracy and ensuring high confidence in determination of genetic variants.

WGS enables the cataloguing of all kinds of genetic variation. The sequences of several individual genomes using 2ndGS technologies have been reported recently. Between 3 and 4 million SNPs per genome were identified using different algorithms (Metzker 2010). A large number of InDels that were undetectable with any of the high-throughput genotyping systems described above are starting to emerge in the databases thanks to 2ndGS methods. Finally, paired-end and mate-pair sequencing methods are also able to discern, at one base pair resolution, many CNVs and other SVs such as inversions and translocations in individual genomes. Recently, de novo assembly of two human genomes have been reported, allowing the discovery of new SVs in an unbiased manner (Li et al. 2010).

A third generation of sequencing technologies based on single-molecule sequencing is under development (Check Hayden 2009). Companies such as Pacific Biosciences (www.pacificbiosciences.com) and Oxford Nanopore Technologies (www.nanoporetech.com) are currently leading this sector. It is likely that routinely DNA sequencing of whole genomes for clinical or research purposes will happen in the near future.

Conclusions

Methods for genotyping and sequencing have come a long way in the past decade. Microsatellites that were the markers of choice a decade ago are no longer used and have been replaced by high-resolution SNP genotyping. Association studies that were very difficult to carry out and for which few positive results had been achieved
a decade ago are now used routinely with great success. Several commercial solutions for WGG using hundreds of thousands of SNPs exist. Use of such methods has reached a level where it is now possible to join datasets that were produced in different laboratories by different groups. Meta-analysis has been very successful and added much insight. All of this has revolutionised molecular genetics and resulted in the identification of an unprecedented number of disease-associated genes. However, markers on the commercial genome-wide genotyping arrays have been selected based on the ‘common disease–common variant’ hypothesis that says that common pathologies are associated with variants of polymorphisms that are present at quite a high level in the general population – have a high minor allele frequency. Results of GWAS have nearly exclusively been variants that confer only marginal additional risk. Very few of the associated variants lead to amino acid changes and could thus be assigned a change of function. Thinking is now shifting that rare variants, that by chance are distributed unevenly onto frequent haplotypes, might be causative. However, including rarer polymorphisms in the genome-wide genotyping system is a game of diminishing returns. It still holds the risk that disease-causing variants might be very rare and private to a few families and thus not represented on the arrays. With the advent of second generation sequencing methods scientists are starting to look towards identifying/genotyping private and rare variants by sequencing. There is a crossover point where sequencing technologies (see ‘Second generation sequencing’) will become more cost effective than genotyping methods and a point where high-resolution WGG and WGS will have the same price tag and possibly comparable throughput. At this point it will be of interest to consider merging the linkage and association strategy and population genetic strategies in study design.

The ever-expanding toolbox for genetics has added refinement and standardization to the procedures. Methods for genome-wide, custom and quantitative genotyping exist. They can be applied for the reliable genotyping of SNPs, CNVs, InDels, and DNA methylation. Based on this, many interesting results have already been generated and with the continuing improvement of technologies the era of genomics is well underway.

References


