16 Laboratory Methods for High-Throughput Genotyping

Howard J. Edenberg[1] and Yunlong Liu[2]

[1]Department of Biochemistry and Molecular Biology and Medical and Molecular Genetics, and Center for Medical Genomics, Indiana University School of Medicine, Indianapolis, Indiana 46202; [2]Division of Biostatistics, Department of Medicine and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202

INTRODUCTION

The genetics of complex diseases has been given a tremendous boost in recent years by the introduction of high-throughput laboratory methods that allow us to approach larger questions in larger populations and to cover the genome more comprehensively. The ability to determine genotypes of many individuals accurately and efficiently has allowed genetic studies that cover more of the variation within individual genes, instead of focusing only on one or a few coding variants, and to do so in study samples of reasonable power. Chip-based genotyping assays, combined with knowledge of the patterns of coinheritance of markers (linkage disequilibrium, LD) developed through the HapMap Project (http://www. hapmap.org), have stimulated genome-wide association studies (GWAS) of complex diseases. These are being encouraged and supported by the National Institutes of Health (NIH) and other groups, notably The Wellcome Trust. Recent successes of GWAS in identifying specific genes that affect risk for common diseases are dramatic illustrations of how improved technology can lead to scientific breakthroughs. Rapid developments in high-throughput sequencing may enable new kinds of studies.

A key issue in high-throughput genotyping is to choose the appropriate technology for your goals and for the stage of your experiment, being cognizant of your sample numbers and resources. This chapter introduces some of the commonly used methods of high-throughput single-nucleotide polymorphism (SNP) genotyping for different stages of genetic studies and briefly reviews some of the high-throughput sequencing methods just coming into use. We will also note some recent developments in "next-generation" sequencing that will enable other kinds of studies. We cannot be comprehensive, and technology in this area is rapidly changing, so our comments should be taken as a starting point for further investigation.

For simplicity, we will discuss three main types of studies: candidate genes, linkage studies and their follow-up, and GWAS and their follow-up. There are choices of genotyping technologies suited to each of these types of studies (Fig. 16.1). Throughput, cost per SNP genotype, and costs per sample can be very different for different technologies. Some technologies, which we call "serial," allow testing of small to modest numbers of SNPs on many subjects in each reaction and are easy to customize. Others, called "parallel" methods, test up to a million SNPs on each subject at one time in fixed panels. The cost per SNP for a serial method is much larger than for a parallel method, but the cost per subject is much less.

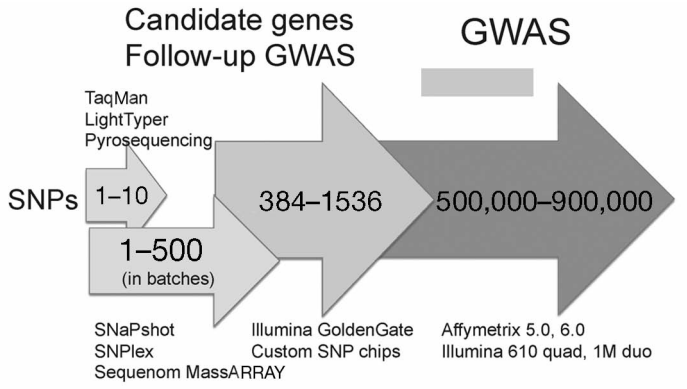

FIGURE 16.1 Different technologies are appropriate for different types of projects and scales of SNPs to be genotyped.

WHY ARE SNPS USED?

At this time, most SNPs do not have known effects on gene expression or function; they are used as markers for genetic differences in their vicinity. Some SNPs are known to cause differences in gene expression or function. The most obvious functional SNPs are those that alter the amino acid encoded at a particular position in a protein or terminate translation leading to a shortened (and often rapidly degraded) polypeptide. Historically, much of the attention to "functional" SNPs has been restricted to these nonsynonymous coding SNPs. More recently, increasing attention has been paid to SNPs that potentially alter splicing, transcription, or mRNA stability. These are generally located in or near a gene and can include synonymous coding SNPs that alter binding sites for the splicing machinery. Such SNPs are harder to recognize and to distinguish from those that do not affect gene expression. It is even harder to tell whether a variation located at some distance from any known gene might have a function, although that is more likely if it lies within a region that is highly conserved among distantly related species.

The binary nature of SNPs has made them the marker of choice in most current work, particularly high-throughput studies. There are about 7 million validated human SNPs in the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/), with many more listed. Genotypic data on about 4 million different SNPs for three major continental groups are available from the HapMap Project, with more coming. SNPs are also available as markers for model organisms. New approaches allow high-throughput SNP genotyping in multiplex reactions. Copy-number variations (CNVs, described in Chapter 13) are increasingly recognized as important, and many of the SNP genotyping platforms also provide information on CNVs (although standardization and interpretation are more difficult).

SNPs can, obviously, be detected by direct sequencing at high accuracy. This is the primary method of SNP discovery, and it demands sequencing of a sufficient number of individuals at an accuracy and coverage that distinguishes real SNPs from sequencing artifacts. Databases still contain many SNPs that are likely to be artifacts. The extraordinary efforts of the HapMap Project (The International HapMap Consortium 2005; Frazer et al. 2007) to date have been very valuable in identifying common SNPs, primarily those with minor allele frequencies (MAFs) greater than 5%. Efforts are under way to examine a wider variety of populations to discover additional SNPs and broaden our understanding of genetic diversity (e.g., 1000 Genomes, A Deep Catalog of Human Genetic Variation, at http://www.1000genomes.org). However, when one is studying a particular disease and has found an association between SNPs and the disease, resequencing in the group with the disease generally leads to discovery of many more SNPs in that gene, often including ones

that are common within the study group, although not in the larger population. For example, sequencing just 16 individuals (eight with a high-risk haplotype and eight with a low-risk haplotype for alcohol dependence) in the exons and proximal 5′ and 3′ regions of *OPRK1* revealed seven new SNPs and an 830-bp insertion/deletion (Xuei et al. 2006; Edenberg et al. 2008).

However, sequencing is not at this time an efficient way to genotype SNPs, although as sequencing technology progresses toward the much sought "$1000 genome" it will become the method of choice. The most efficient methods at this time involve single-base extension (in effect, microsequencing) with readouts that include mass differences (e.g., Sequenom MassARRAY), light flashes (pyrosequencing), or hybridization, with readouts that include measures of the amount of oligonucleotide hybridized (e.g., microarrays), cleavage of hybridized oligonucleotides (e.g., TaqMan, Applied Biosystems), or melting curves (e.g., LightTyper, Roche Applied Science). There are platforms that work best for targeted SNP genotyping and others that are aimed at whole genomes. Each has advantages for particular studies.

CAN CANDIDATE GENES BE USED?

Many studies focus on candidate genes that are chosen on the basis of physiology or suggestive results from other studies. One problem is that many candidate genes are weak candidates, with low prior probabilities. This, plus some publication bias toward positive findings, might explain the large number of candidate-gene association studies that have not been replicated. As association findings are reported, many groups will want to test the significant genes in their own study sample. This is a good strategy for confirmation studies and for some with too small a subject population to withstand the multiple testing corrections required for GWAS. With a small sample, the necessity to correct for the very high levels of multiple testing in a GWAS make it very difficult to detect effects of modest size with genome-wide significance (see Chapter 6). Even samples that are large in terms of our ability to do sophisticated and reproducible phenotyping (1000–2000 cases and controls) are proving to be underpowered in studies of complex traits in which the contribution of any one genetic variant is small. Small samples can, however, be very useful in replicating results. In this case, the prior probability is much higher and the likelihood of confirming a finding is increased.

Many early candidate gene studies tested only a single SNP, which carries very limited information about the overall variation within that gene. This approach can be successful if there is a known, strong functional SNP (e.g., Thomasson et al. 1991), but it often leads to false-negative results. Even in simple situations (Mendelian disorders) in which a single change in one gene can lead to disease, there is often allelic heterogeneity, and testing only one SNP can miss the one(s) that are functional in that family or population. It is generally better to cover a larger fraction of the variation by genotyping multiple SNPs chosen based on LD in addition to hypothesized functional SNPs, although cost and time (and concerns about multiple testing) can preclude full coverage.

HapMap data (The International HapMap Consortium 2005; Frazer et al. 2007) can be used to select the SNPs that best report on common variation within a gene or region ("tag" the region). It is useful first to visualize the LD structure in the region of interest, using Haploview (Barrett et al. 2005) (which can be performed within the HapMap website). A program such as Tagger (de Bakker et al. 2005) can aid in the selection of SNPs; it can be run from a server (through HapMap or directly at http://www.broad.mit.edu/mpg/tagger). Parameters such as minimum MAF of SNPs to be "tagged" and the degree of correlation (r^2) to be accepted as adequate can be adjusted so that a set of SNPs reasonable for sample size, technology, and budget can be selected.

The genotyping technologies most appropriate for candidate-gene studies are what we call "serial" technologies; that is, they test from 1 to 48 SNPs at a time on each sample. A set of assays can be designed, run, and analyzed; then, if needed, a follow-up set of assays can be run. These generally allow testing of a large number of samples for a modest number of targeted SNPs at modest cost. There are

many technologies that are good choices for candidate-gene studies (Tsuchihashi and Dracopoli 2002).

TaqMan SNP Genotyping Assays (Applied Biosystems) measure individual SNPs using a 5′ nuclease assay (De la Vega et al. 2005). Many assays (>4.5 million) have been predesigned and it is relatively straightforward to design new assays. Single SNPs are generally run on sets of 96 or 384 samples at a time. Moving up to low levels of multiplexing, the SNaPshot Multiplex System (Applied Biosystems) is a primer extension-based method with detection by capillary electrophoresis; it allows multiplexing up to 10 SNPs starting from as little as 3 ng DNA per sample. The SNPlex Genotyping System from Applied Biosystems (De la Vega et al. 2005) uses the oligonucleotide ligation assay (Nickerson et al. 1990) to discriminate SNPs, followed by PCR and capillary electrophoresis; it allows genotyping of up to 48 SNPs at a time. Both SNaPshot and SNPlex use capillary sequencing instruments, which might already be available in many laboratories. The LightTyper system (Roche Applied Science) uses melting curve analysis to discriminate individual SNPs. These methods use fluorescently labeled oligonucleotides for detection of the SNPs (Bennett et al. 2003).

Pyrosequencing detects SNPs by a synthesis reaction with detection based on flashes of light when a nucleotide is incorporated (Ahmadian et al. 2000; Pourmand et al. 2002). The Invader assay (Third Wave Technologies, Inc., now Hologic) involves formation and then cleavage of a flap created by hybridization of two oligonucleotides to the target sequence; the signal from this initially cleaved flap is amplified in a second fluorescence resonance energy transfer reaction (Lyamichev et al. 1999).

The Sequenom MassARRAY system (Jurinke et al. 2001; Jurinke et al. 2002) can measure up to 36 SNPs on 384 samples per assay. A region is amplified by polymerase chain reaction (PCR) and then a single-base primer extension is performed using modified deoxyribonucleoside triphosphates that increase the resolution with which a mass spectrometer can distinguish the four possible nucleotides added. An advantage of this is that unmodified oligonucleotides can be used, reducing the initial cost.

These genotyping techniques are good for testing candidate genes and small regions. One can test a modest number of SNPs, analyze them, and then genotype additional SNPs to whatever depth of coverage is desired in genes or regions that remain of interest. The investigator has nearly complete freedom to design assays for any SNPs desired for the particular project.

LINKAGE REGIONS CAN BE FINE-MAPPED BY HIGH-THROUGHPUT SNP GENOTYPING

Although new availability of commercial platforms has stimulated interest in GWAS, there are many projects that performed linkage studies and identified broad regions likely to contain genes in which variations affect risk for a disease or a related phenotype. Linkage studies have relatively low resolution, so one needs a way to follow up such studies with fine-mapping. High-throughput SNP genotyping offers an attractive approach to this task. One can either use a parallel approach, such as a custom microarray of thousands or tens of thousands of SNPs or an Illumina GoldenGate assay (Fan et al. 2006) to test from 384 to 1536 SNPs, or a serial approach in which SNPs are tested in smaller numbers at a time.

If there are particularly good candidate genes within the linkage region, the serial approach of testing a limited number of such candidates may be optimum because it limits costs and also limits multiple testing. An example is targeting a set of four genes that encode subunits of the γ-aminobutyric acid A receptor in a region of chromosome 4 linked to both alcohol dependence and an electrophysiological phenotype (Edenberg et al. 2004). These made excellent candidate genes on the basis of both physiological knowledge and location in the center of the linkage peak. A serial approach was taken, first testing five to six SNPs in each gene and then covering the gene in which multiple SNPs were associated with alcohol dependence with additional SNPs in an attempt to further localize the key SNPs. The Sequenom MassARRAY system was used. *GABRA2* was shown to be associated with alcoholism, with a large LD block extending from intron 3 past the 3′ end of the gene (Edenberg et al. 2004); this finding has since been replicated by many groups.

In a case in which there are no strong candidate genes or the candidate genes tested did not prove to be associated, a parallel approach may be called for. Illumina offers GoldenGate Custom Panels that can measure 384–1536 assays per reaction. These can be valuable for testing many sites across a linkage region, or setting up panels to follow up the best "hits" from an earlier study (either a GWAS or a compilation of candidate genes). The Center for Inherited Disease Research at Johns Hopkins University is a resource for getting such genotyping performed, if it is approved by their advisory panel.

One can also make custom-designed genotyping microarrays. An example of this approach was the design of a panel of 1536 SNPs (using the Illumina GoldenGate assay) that used LD information to capture data (at $r^2 > 0.8$) on >4000 SNPs with MAF > 0.10, along with a small number of nonsynonymous coding SNPs, in a linkage region on chromosome 7q22 covering about 18 Mb (Dick et al. 2007). Of these, eight SNPs were found to be associated with alcohol dependence at $p < 0.01$, four of which clustered in a single gene. This gene was followed up by genotyping 16 additional SNPs, using the Sequenom MassARRAY assay; 12 SNPs in that gene were nominally significant, with eight remaining significant when corrected for multiple testing (Dick et al. 2007).

GWAS AND FOLLOW-UP BY ADDITIONAL GENOTYPING

Parallel genotyping tests many SNPs on a single sample (or two samples) at one time, using an array-based format. The number of SNPs per array has dramatically increased in the past few years, from about 10,000 to 100,000, 300,000, 600,000, and now more than 1 million. Currently, the Affymetrix Genome Wide Human SNP Array 6.0 has more than 906,600 SNPs and more than 946,000 probes for the detection of copy number variation, and the Illumina Human1M-Duo BeadChip assesses more than 1.1 million loci per sample.

Both of these array-based methods are widely used. There are differences between these two platforms in design, SNP selection, and biochemistry, but overall data quality and coverage appear similar. About 480,000 of the SNPs on the Affymetrix Genome Wide Human SNP Array 6.0 were selected based on what worked in a "complexity-reduction" step that involved selection of restriction fragments between about 200 and 1100 bp. These were supplemented with 424,000 tag SNPs, plus the 946,000 monomorphic sites of which about 202,000 are in known CNV regions and the rest were chosen by spacing to detect CNV. SNP detection is by differential hybridization to 25-mers designed to match both alleles at each site. Illumina probes are larger (about 50 nucleotides) and were chosen to cover the genome based on HapMap LD data plus nonsynonymous coding SNPs. Both platforms give excellent genome-wide coverage in European populations and very good coverage even in African populations. It should be noted that there are significant gaps in both. For example, an analysis of SNPs (MAF ≥ 5%) in about 900 genes relating to addiction, many were not well tagged ($r^2 \leq 0.8$) by either platform (Saccone et al. 2009). Therefore, follow-up of initial results with additional genotyping is valuable.

Initially, the high cost of the genome-wide microarray assays made individual genotyping of large samples prohibitive for many groups. Pooling of samples can reduce costs (Bansal et al. 2002; Sham et al. 2002). Pooling requires very careful measurement of DNA concentrations to equalize the contributions of each individual, but when performed carefully this method can measure relative allele frequencies to within a few percent. Pooling has provided interesting leads that can be followed up by individual genotyping of the most significant findings. However, pooling does not provide nearly as much information as individual genotyping. It requires either a dichotomous phenotype or the decision to make a pseudo-dichotomous phenotype—for example, by contrasting the two ends of a distribution. Only the phenotype by which the pools were created can be analyzed; information on endophenotypes or related phenotypes is lost, although one could theoretically make smaller pools matched on more than one phenotype. Pooling approaches have been effective in studies of alcohol dependence and bipolar disorder genetics (e.g., Johnson et al. 2006; Baum et al. 2008).

As the density of SNPs on arrays has increased and the costs gone down, microarrays are increasingly used for GWAS on individual samples. This approach is much more powerful and allows analysis of multiple phenotypes and endophenotypes at once, as well as analysis of quantitative traits. This approach has been successful in studies of macular degeneration, height, and type 2 diabetes (e.g., Klein et al. 2005; Weedon et al. 2007; Zeggini et al. 2008; see also Chapter 18).

GWAS must be followed up. One can attempt replication either of the leading SNPs in another population, using one of the serial genotyping methods described previously (see "Can Candidate Genes Be Used?") or one can perform follow-up GWAS on another population. In some cases, the strongest signal from an analysis or meta-analysis might be an imputed SNP (i.e., an SNP not actually genotyped but rather predicted based on the genotypes and the known LD in the region; see Chapter 10) (e.g., Ferreira et al. 2008). Such imputed SNPs should be genotyped directly in the initial population, using one of the serial methods, because imputation is not exact. Finally, denser genotyping and resequencing of the significantly associated genes in a GWAS is often performed in search of potentially functional polymorphisms.

QUALITY CONTROL OF SNP DATA IS REQUIRED

Quality control (QC) of the SNP data is required before analysis. This is not a simple task. Although the details of QC could be a chapter themselves, some basic approaches are relatively standard. QC is usually performed in two cycles: (1) Remove problem samples based on the sample QC metrics and (2) recalculate the SNP metrics on the remaining samples before SNP QC metrics are applied.

It is useful to test DNA quality before running the whole-genome chips. First, a spectrum from 220 to 350 nm is better than just examining A_{260}/A_{280}, because sometimes there are contaminants that absorb in the 230–270-nm region. These contaminants can lead to large overestimates of the amount of DNA actually present and can also inhibit genotyping or sequencing reactions. Use of a dye selective for double-stranded DNA (e.g., PicoGreen, Molecular Probes, Invitrogen) can give a better measurement of DNA quantity in the presence of RNA or free nucleotides, but it does not reveal the presence of contaminants. A combination of the two approaches is best. Then, evaluation of fragment size should be performed on agarose gels. Poor-quality samples (contaminated or too fragmented) generally give poor results. The average fragment size needed depends on the platform. The Affymetrix Genome Wide Human SNP Array 6.0 uses a size selection of fragments from 200 to 1100 bp, so fragmented DNA does not work well. The Sequenom MassARRAY generally amplifies regions of about 100 bp and therefore is less affected by small fragment size.

For any genotyping methods, samples that show too many dropouts or no-calls are suspect and are generally removed from analysis if <97%–98% of the SNPs give genotypes. Samples with heterozygosity that is too high or too low compared with others in the population are also suspect and could reflect mixed samples containing DNA from more than one individual. Probes on the X and Y chromosome can be used to confirm sex. Chromosomal loss and duplication, particularly of the X chromosome, are often observed in immortalized cell lines; if the other data appear to be okay, one can choose to selectively remove the data from the X chromosome and retain the rest. With genome-wide data, one can identify cryptic relatives (i.e., individuals in the data set who are related to each other) based on allele sharing. When samples from multiple studies are analyzed jointly, this might include relatives or even the same subject participating in more than one study.

After sample QC, SNP QC is needed. Duplicate samples should give the same calls. It is a good idea to include one or more HapMap samples to compare the genotype call with that in the database. In many studies, SNPs with MAF < 0.01 are not analyzed, because calling rare genotypes is subject to more errors. Although many SNPs may be lost by this filter, the overall cost to the project power is not great because even in studies with 1000 cases and 1000 controls, the power to detect effects with rare genotypes is very limited. SNPs for which many samples do not give genotypes (dropouts)

are suspect and are usually omitted from analysis. The acceptable call rate depends on the nature of the study, but usually is set in the region of 95% or better. Tests of Hardy–Weinberg equilibrium (HWE) are useful in flagging SNPs with nonrandom dropping out of a genotype, but the problem of multiple testing limits how strictly they can be screened and some SNPs with bias might still pass this filter. When measuring 1 million SNPs in a GWAS, deviations from HWE must generally be more significant than $p < 0.000001$ to remove an SNP. Differential dropouts of one homozygote (usually the minor allele) or over- or underrepresentation of heterozygotes should raise a flag and lead to a close look at the raw data from that SNP (see the following).

Examining the clustering of the three genotypes, using either a Cartesian or polar plot of the intensity of alleles, is useful (Fig. 16.2 shows examples from Sequenom data, but the basic ideas and method are general). Good-quality SNPs show three clearly defined and tight clusters, with the ho-

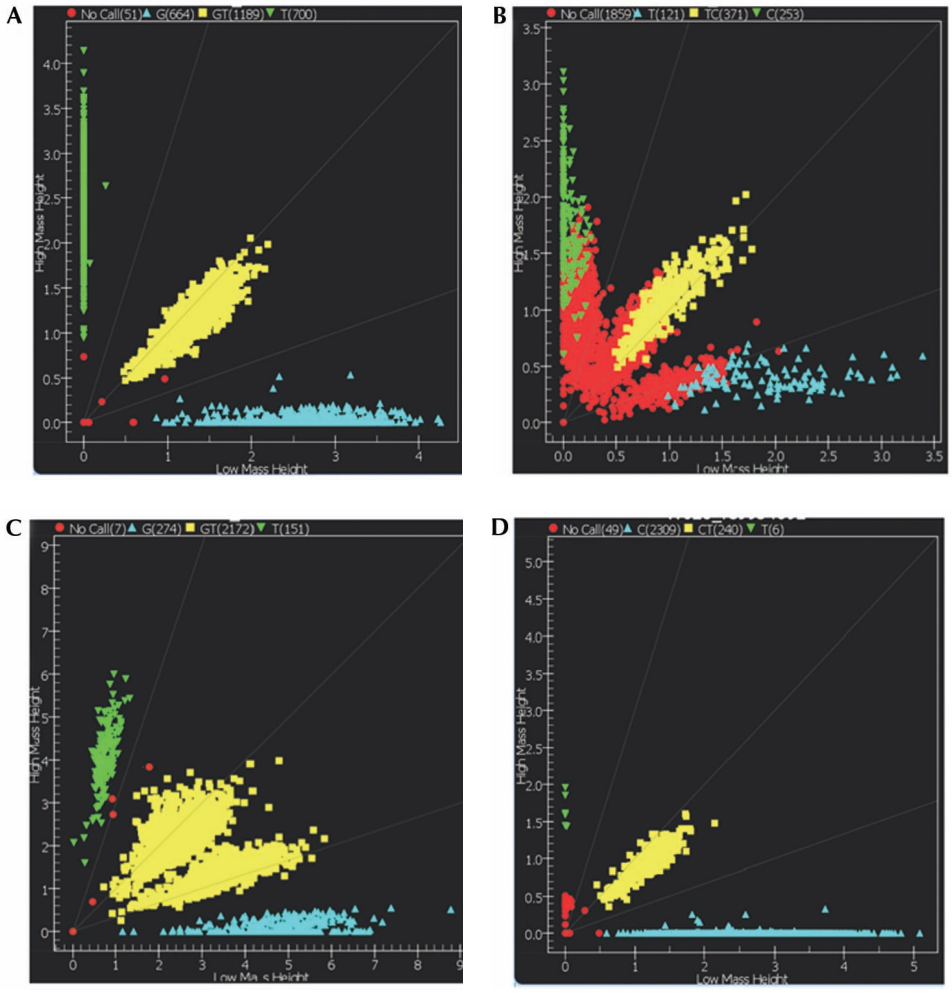

FIGURE 16.2 Quality control of SNP genotyping. One method of examining the quality of the genotype calls is to plot the intensity of the signal for one allele (green) versus the intensity of the signal of the other (blue) with heterozygotes shown in yellow, and uncalled alleles in red. Data shown are from the Sequenom MassARRAY reaction, but in principle could be from many technologies including the whole-genome scale microarrays. (*A*) Good-quality assay, with good clustering and separation of the alleles. (*B*) Bad assay, with overlap of homozygotes and heterozygotes such that many samples are not called. (*C*) Questionable assay; the two clusters of heterozygote intensities might suggest copy number problems. (*D*) Questionable assay; although clustering looks tight and heterozygotes are well balanced, close inspection suggests that many of the unknowns might represent homozygotes for the minor allele. Differential loss of one of the homozygous genotypes will bias the analysis.

mozygotes falling along the vertical or horizontal axis and the heterozygotes at a 45° angle (Fig. 16.2A). One can often detect a problem such as overlap of the clusters from one allele and from the heterozygotes (Fig. 16.2B). One might also detect more than three clusters or a split cluster (Fig. 16.2C), suggesting the possibility of a relatively frequent CNV. The hardest potential error to detect is the selective dropping out of the minor allele (Fig. 16.2D); such SNPs will clearly give incorrect results. The huge numbers of SNPs in a GWAS precludes looking at intensity plots of all of the SNPs, but one should look at the intensity plots of those deemed significant. It is valuable to regenotype significant SNPs by a different technique to insure that the finding reflects the real genetics rather than a technical artifact.

Most QC issues are the same for GWAS as for the more targeted studies, but the vast amount of data makes some additional checks both possible and necessary. In GWAS the biochemistry is typically performed in an automated or semiautomated manner on sets of 48 or 96 samples (plates), so an additional analysis is usually performed to detect situations in which one of the plates differs from the others, perhaps because of some aspect of its processing. This "plate effect" is detected by analyzing the allele frequency of the SNP on that plate against the allele frequency on the sum of the other plates by a chi-square test, and discarding data if a plate effect is seen at the level of 10^{-8}, or multiple plates show effects at 10^{-4} or worse. For best results, cases and controls should be evenly distributed on each plate; otherwise, subtle differences in the biochemistry can lead to biases in the allele calling. These subtle biases will be an issue when, as is happening increasingly, data from controls genotyped at one time in one study are used with data from samples genotyped at a different time and perhaps in a different laboratory to increase power in another study.

NEXT-GENERATION SEQUENCING WILL REVOLUTIONIZE GENOTYPING

The rapid development of high-throughput "next-generation sequencing" technology, by which we mean "massively parallel" sequencing, offers great potential to revolutionize genotyping (Mardis 2008). It is already playing a role in identifying novel SNPs (Hodges et al. 2007; Van Tassell et al. 2008; Wheeler et al. 2008) and other structural variants such as insertion/deletion (indels) and CNVs (Campbell et al. 2008). An early study reported that at least 13-fold coverage is needed to identify 99% of the heterozygous SNPs (Wheeler et al. 2008).

Next-generation sequencing technology is capable of sequencing from hundreds of thousands to hundred millions of DNA (or cDNA) fragments in a single instrument run in a massively parallel fashion. Such high-throughput sequencing technology has been used in applications including de novo sequencing, resequencing to detect SNPs and other variants, transcriptome sequencing, immunoprecipitation-based protein–DNA or protein–RNA interaction mapping, and DNA methylation using bisulfite-mediated cytosine conversion. When the cost of sequencing a human genome approaches $1000, the current target, sequencing will probably replace genotyping for GWAS.

So far, three major platforms are commercially available: the Roche GS FLX Sequencer (454 technology, 454 Life Sciences, Roche), the Illumina Genome Analyzer (Solexa), and the Applied Biosystems SOLiD sequencer (SOLiD 3 System). The 454 Sequencer produces longer reads of >250 bp per fragment, in contrast to 35- to 50-bp reads of the other two platforms (Solexa and SOLiD). The short-sequence instruments produce many more reads for each instrument run. The capacities of these instruments are increasing rapidly with new development in the chemistry and physics of the techniques, so detailed figures would be out of date before this chapter is published, but some already claim >20 Gb of sequence per run. Therefore, we will restrict our discussion to some general issues.

High-throughput sequencing offers the potential to identify de novo SNPs as well as previously reported SNPs. Complete genome-wide resequencing of an individual has been published for two individuals, James D. Watson (Wheeler et al. 2008) and Craig Venter (Levy et al. 2007). These sequences showed many new variants, particularly insertions, deletions, and translocations. Given present costs,

however, a major current application is detection of variations in a focused genomic region, usually to follow up significant association studies by identifying potentially functional variation.

There are several methods available to select focused genomic regions from individual samples. One approach is to amplify the region through multiplex PCRs (Porreca et al. 2007), but this can be expensive and time-consuming, and PCR artifacts are possible. One can also produce reduced representation libraries by constructing cDNA libraries of a temporally and spatially specific transcriptome (Bainbridge et al. 2006; Barbazuk et al. 2007), which focuses on variations within transcripts, or by selecting DNA fragments of a specific size after complete restriction endonuclease digestion (Barbazuk et al. 2007; Van Tassell et al. 2008).

Recently, microarrays have been used to capture genomic regions of interest (Albert et al. 2007; Hodges et al. 2007). The NimbleGen Sequence Capture Array (Roche) can be used to isolate up to 5 Mb of DNA using probes of 50–80 nucleotides tiled across the region(s) of interest. Similar technology has also been implemented in another study, in which a customized array was designed that contains 55,000 100-mer oligonucleotides (Porreca et al. 2007). The selected regions can be eluted and sequenced. One concern is whether the capture efficiency in different samples is comparable, given the presence of SNPs and other variations; platforms with longer probes are less sensitive to this.

This technology is quite good at identifying structural variants such as deletions, insertions, duplications, and inversions. Sequencing short reads from both ends of millions of DNA fragments of known size (hundreds to thousands of base pairs) allows one to determine if the spacing and orientation of the paired sequences matches that in the genome (Campbell et al. 2008), and thereby detect many insertions and deletions.

SUMMARY AND CONCLUSIONS

High-throughput genotyping technology has enabled GWAS that have great promise for identifying genes that contribute to complex diseases and phenotypes and for large-scale follow-up of the top candidate genes from such studies. There are a range of techniques from whole-genome scale to individual SNPs. As with all techniques, appropriate choice of the approach for different projects or phases of projects is important. Careful attention to quality control is also essential. In the future, whole-genome sequencing may overtake the current large-scale technologies.

ACKNOWLEDGMENTS

We thank Dr. Xiaoling Xuei for help in selecting figures and for helpful comments on the manuscript, and Dr. Jeanette McClintick for helpful comments on the manuscript. Related work in the investigators' laboratories has been funded by grants AA008401, AA006460, AA07611 from NIAAA, and MH078151 from the National Institute of Mental Health and the Indiana Genomics Initiative (INGEN, which is partially funded by The Lilly Endowment, Inc.).

REFERENCES

Ahmadian, A., Gharizadeh, B., Gustafsson, A.C., Sterky, F., Nyren, P., Uhlen, M., and Lundeberg, J. 2000. Single-nucleotide polymorphism analysis by pyrosequencing. *Anal. Biochem.* **280:** 103–110.

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4:** 903–905.

Bainbridge, M.N., Warren, R.L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., et al. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7:** 246.

Bansal, A., van den Boom, D., Kammerer, S., Honisch, C., Adam, G., Cantor, C.R., Kleyn, P., and Braun, A. 2002. Association testing

by DNA pooling: An effective initial screen. *Proc. Natl. Acad. Sci.* **99:** 16871–16874.

Barbazuk, W.B., Emrich, S.J., Chen, H.D., Li, L., and Schnable, P.S. 2007. SNP discovery via 454 transcriptome sequencing. *Plant J.* **51:** 910–918.

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21:** 263–265.

Baum, A.E., Akula, N., Cabanero, M., Cardona, I., Corona, W., Klemens, B., Schulze, T.G., Cichon, S., Rietschel, M., Nothen, M.M., et al. 2008. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol. Psychiatry* **13:** 197–207.

Bennett, C.D., Campbell, M.N., Book, C.J., Eyre, D.J., Nay, L.M., Nielsen, D.R., Rasmussen, R.P., and Bernard, P.S. 2003. The LightTyper: High-throughput genotyping using fluorescent melting curve analysis. *Biotechniques* **34:** 1288–1292, 1294–1295.

Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C., et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40:** 722–729.

de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J., and Altshuler, D. 2005. Efficiency and power in genetic association studies. *Nat. Genet.* **37:** 1217–1223.

De la Vega, F.M., Lazaruk, K.D., Rhodes, M.D., and Wenz, M.H. 2005. Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System. *Mutat. Res.* **573:** 111–135.

Dick, D.M., Aliev, F., Wang, J.C., Saccone, S., Hinrichs, A., Bertelsen, S., Budde, J., Saccone, N., Foroud, T., Nurnberger Jr., J., et al. 2007. A systematic single nucleotide polymorphism screen to fine-map alcohol dependence genes on chromosome 7 identifies association with a novel susceptibility gene *ACN9*. *Biol. Psychiatry* **63:** 1047–1053.

Edenberg, H.J., Dick, D.M., Xuei, X., Tian, H., Almasy, L., Bauer, L.O., Crowe, R.R., Goate, A., Hesselbrock, V., Jones, K., et al. 2004. Variations in *GABRA2*, encoding the α2 subunit of the $GABA_A$ receptor, are associated with alcohol dependence and with brain oscillations. *Am. J. Hum. Gen.* **74:** 705–714.

Edenberg, H.J., Wang, J., Tian, H., Pochareddy, S., Xuei, X., Wetherill, L., Goate, A., Hinrichs, T., Kuperman, S., Nurnberger Jr., J.I., et al. 2008. A regulatory variation in *OPRK1*, the gene encoding the κ-opioid receptor, is associated with alcohol dependence. *Hum. Mol. Genet.* **17:** 1783–1789.

Fan, J.B., Chee, M.S., and Gunderson, K.L. 2006. Highly parallel genomic assays. *Nat. Rev. Genet.* **7:** 632–644.

Ferreira, M.A., O'Donovan, M.C., Meng, Y.A., Jones, I.R., Ruderfer, D.M., Jones, L., Fan, J., Kirov, G., Perlis, R.H., Green, E.K., et al. 2008. Collaborative genome-wide association analysis supports a role for *ANK3* and *CACNA1C* in bipolar disorder. *Nat. Genet.* **40:** 1056–1058.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851–861.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., and McCombie, W.R. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39:** 1522–1527.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Johnson, C., Drgon, T., Liu, Q.R., Walther, D., Edenberg, H., Rice, J., Foroud, T., and Uhl, G.R. 2006. Pooled association genome scanning for alcohol dependence using 104,268 SNPs: Validation and use to identify alcoholism vulnerability loci in unrelated individuals from the collaborative study on the genetics of alcoholism. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **141B:** 844–853.

Jurinke, C., van den Boom, D., Cantor, C.R., and Koster, H. 2001. Automated genotyping using the DNA MassARRAY technology. *Methods Mol. Biol.* **170:** 103–116.

Jurinke, C., van den Boom, D., Cantor, C.R., and Koster, H. 2002. Automated genotyping using the DNA MassARRAY technology. *Methods Mol. Biol.* **187:** 179–192.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308:** 385–389.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol.* **5:** e254.

Lyamichev, V., Mast, A.L., Hall, J.G., Prudent, J.R., Kaiser, M.W., Takova, T., Kwiatkowski, R.W., Sander, T.J., de Arruda, M., Arco, D.A., et al. 1999. Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat. Biotechnol.* **17:** 292–296.

Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24:** 133–141.

Nickerson, D.A., Kaiser, R., Lappin, S., Stewart, J., Hood, L., and Landegren, U. 1990. Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay. *Proc. Natl. Acad. Sci.* **87:** 8923–8927.

Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F., et al. 2007. Multiplex amplification of large sets of human exons. *Nat. Methods* **4:** 931–936.

Pourmand, N., Elahi, E., Davis, R.W., and Ronaghi, M. 2002. Multiplex pyrosequencing. *Nucleic Acids Res.* **30:** e31.

Saccone, S.F., Bierut, L.B., Chesler, E.J., Kalivas, P.W., Lerman, C., Saccone, N.L., Uhl, G.R., Li, C.-Y., Philip, V.M., Edenberg, H.J., et al. 2009. Supplementing high-density SNP microarrays for additional coverage of disease-related genes: Addiction as a paradigm. *PLoS ONE* **4:** e5225.

Sham, P., Bader, J.S., Craig, I., O'Donovan, M., and Owen, M. 2002. DNA Pooling: A tool for large-scale association studies. *Nat. Rev. Genet.* **3:** 862–871.

Thomasson, H.R., Edenberg, H.J., Crabb, D.W., Mai, X.L., Jerome, R.E., Li, T.K., Wang, S.P., Lin, Y.T., Lu, R.B., and Yin, S.J. 1991. Alcohol and aldehyde dehydrogenase genotypes and alcoholism in Chinese men. *Am. J. Hum. Genet.* **48:** 677–681.

Tsuchihashi, Z. and Dracopoli, N.C. 2002. Progress in high throughput SNP genotyping methods. *Pharmacogenomics J.* **2:** 103–110.

Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., and Sonstegard, T.S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* **5:** 247–252.

Weedon, M.N., Lettre, G., Freathy, R.M., Lindgren, C.M., Voight, B.F., Perry, J.R., Elliott, K.S., Hackett, R., Guiducci, C., Shields, B., et al. 2007. A common variant of *HMGA2* is associated with adult and childhood height in the general population. *Nat. Genet.* **39:** 1245–1250.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., et al. 2008. The complete genome of an individual by massively parallel

DNA sequencing. *Nature* **452:** 872–876.

Xuei, X., Dick, D., Flury-Wetherill, L., Tian, H.J., Agrawal, A., Bierut, L., Goate, A., Bucholz, K., Schuckit, M., Nurnberger Jr., J., et al. 2006. Association of the κ-opioid system with alcohol dependence. *Mol. Psychiatry* **11:** 1016–1024.

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G., et al. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40:** 638–645.

WWW RESOURCES

http://www.1000genomes.org 1000 genomes, a deep catalog of human genetic variation.

http://www.broad.mit.edu/mpg/tagger de Bakker et al. 2005. Tagger.

http://www.broad.mit.edu/mpg/haploview Barrett et al. 2005. Haploview.

http://www.cidr.jhmi.edu Center for Inherited Disease Research, Johns Hopkins University.

http://www.hapmap.org International HapMap Project.

http://www.ncbi.nlm.nih.gov/SNP dbSNP, Single Nucleotide Polymorphism Database.