16 Genomic Studies of Human Populations: Resequencing Approaches to the Identification of Human Quantitative Loci

Joanne E. Curran, Claire Bellis, Laura Almasy, and John Blangero

16.1 Introduction

The primary goal of the complex disease genomics field is to identify loci influencing disease susceptibility. The field has progressed substantially in recent years with the development of new methodologies for genome-wide assessment of sequence variation. Data is rapidly accumulating that rare variants have a large cumulative effect on normal phenotypic variation and are extremely important to disease (Blangero 2004; De La Vega et al. 2011; Li and Leal 2008; Pelak et al. 2010). Pedigree-based studies represent an implicit enrichment strategy for identifying such rare variants. Mendelian transmissions from parents to offspring maximize the chance that multiple copies of rare variants exist in the pedigree. These variants can then be identified by direct resequencing and statistical tests that minimize the influence of spurious linkage disequilibrium (Blangero et al. 2005; Kent et al. 2007). The key factor in the identification of rare variants then becomes resequencing sufficient numbers of chromosomes to capture all existing sequence variation. Given the likely importance of rare variation, a comprehensive sequencing strategy is the best means for detecting all such variants with sufficient copies. Whole genome sequence represents the “holy grail” for genetic studies. Prior approaches have only sampled partial variation from the genome. Unlike most other sciences, the causal state space for genetics is finite and we now have the tools available to comprehensively examine it.

16.2 Rationale for Next-Generation Sequencing Studies

16.2.1 Human Genetic Variation

Human genetic variation manifests itself in all aspects of human phenotypic variation, and has direct implications for the discovery of the underlying genes responsible for the observed heritability (the proportion of the total variance of a phenotype that is attributable to the additive effects of alleles) of any given trait. Localization and identification of these causal genes is the main goal of any genetic study of human disease. The heritability of a given trait tells us how important a role genetic variation is likely to play in the causation of variation. For many complex traits, known genetic variants only account for a small proportion of the total heritability (using plasma HDL-C levels as an example, known

J.E. Curran (✉) · C. Bellis · L. Almasy · J. Blangero
Department of Genetics, Texas Biomedical Research Institute, PO Box 760549, San Antonio, TX 78245-0549, USA
e-mail: jcurran@txbiomedgenetics.org

C. Bellis
e-mail: cbellis@txbiomedgenetics.org

L. Almasy
e-mail: almasy@txbiomedgenetics.org

J. Blangero
e-mail: john@txbiomedgenetics.org

R. Duggirala et al. (eds.), *Genome Mapping and Genomics in Human and Non-Human Primates*,
Genome Mapping and Genomics in Animals 5, DOI 10.1007/978-3-662-46306-2_16,

variants account for less than 10 % of its total heritability) (Chasman et al. 2009), indicating that most of the genes/genetic variants have yet to be identified. This problem, termed "missing heritability", holds for many other complex human traits (Manolio et al. 2009).

What is the biological source of heritability in humans? Ultimately, it comes from observable functional genetic variation at the sequence level. A functional variant is one that influences the focal phenotype via some molecular mechanism. Thus, functional variants can be considered to be phenotype-specific in this context. It is the primary goal of complex disease genetics to identify such directly functional variants since they also will directly implicate the causal genes involved in the disease process.

16.2.2 Identification of Human Genetic Variation

For a given trait with significant heritability, how does one go about localizing and identifying these causal genetic factors that ultimately determine heritability?

Many recent advances in analysis of human quantitative traits have been made in the context of genetically complex diseases (Blangero 2004). In the absence of complete sequence information for all study participants, gene localization depends either on the random effect of known genetic markers assessed via linkage (for details, see the Chap. 3 by Almasy et al., in this volume), or the main effect of the markers via association (for details, see the Chapter by Hanson and Malhotra in this volume).

16.2.2.1 Gene Localization by Linkage

Before the era of high-throughput genotyping and next-generation sequencing, complex disease genetics in pedigrees was concentrated on genome scanning using a high-density map of genetic markers evenly distributed throughout the genome. Such genome scan, or linkage, information was used to identify chromosomal regions that contain variants influencing disease risk factors. Classical penetrance model-free linkage analysis is biased against rare functional variants. This bias is the result of the usual practice of estimating a single residual heritability and single quantitative trait loci (QTL)-specific heritability for the entire set of pedigrees examined. However, if rare functional variation is important, we would expect the magnitude of correlation between relatives to vary across pedigrees reflecting variance in pedigree-specific heritability. In fact, evidence of heterogeneity in heritabilities across pedigrees is expected under a model in which there are rare variants of moderate effect segregating. The assumption of heritability (both total and QTL-specific) homogeneity will lead to many missed QTL signals if rare variation is important.

To better search for QTLs due to rare functional variation, pedigree- and lineage-specific linkage analyses are required. These analyses may be done using a simple extension of the variance component model and simultaneously accounting for potential covariates. One approach for pedigree-specific linkage analysis is analogous to that long utilized for Mendelian disorders of dichotomous diseases. Basically, a search for linkage within each pedigree is performed using the usual variance component model (for details, please refer to Chap. 3 by Almasy et al., in this volume) with the added constraint that mean parameters must be held constant to that estimated from all of the data. This constraint is conceptually similar to ascertainment correction and guarantees that all phenotype deviations are referenced to the total population rather than to the specific pedigree being considered. Although each pedigree can be argued to represent a set of unique localization hypotheses, it may also be prudent to address the increased number of parameters being investigated (one additional QTN-specific heritability for each pedigree). Therefore, a mixture model analogous to heterogeneity logarithm of odds (LOD) testing performed in parametric linkage analysis, such as that being implemented in our computer package, SOLAR should also be employed. Some of the pedigrees may segregate functional variants at a given location, while others will not. All testing should be performed in the standard variance component framework. Using this approach in an example and performing a formal test of heritability heterogeneity, we determined that about 15 % of lipid-related traits in the San Antonio Family Study show strong evidence for differences in total heritability across pedigrees that is consistent with the presence of rare functional variants; and pedigree-specific linkage analyses revealed additional genome-wide significant QTLs for 140 traits that were missed using conventional linkage analysis, suggesting a large underlying source of rare functional variation (Unpublished data). These results highlight the importance of such an approach for the identification of QTLs due to rare variation.

16.2.2.2 Gene Localization by Association

Association studies are limited to detecting the effects of relatively common (with allele frequencies greater than 0.10) genetic variants and largely have not led to causal gene identification. Genome-wide association studies (GWAS) exploit correlations between closely spaced markers that are a function of linkage disequilibrium. These correlations are limited by the differences in the allele frequencies of the markers. Thus, common variants cannot be strongly correlated with rare ones. GWAS panels of single nucleotide polymorphisms (SNPs) are selected to represent common variation across the genome with modern human GWAS panels including a million or more markers that are correlated with a large portion of SNPs with allele frequencies of 0.05 or greater at an r^2 of at least 0.8. Each SNP on the panel represents not only itself but also serves as a proxy for some number of ungenotyped markers. An association signal could be due to any of the variants in disequilibrium with the genotyped SNP. Although widely misunderstood, these associations do not represent gene identifications but are bona fide QTL localizations that must be deeply sequenced in order to identify the underlying causal genes and followed up with functional work. Unfortunately, to date, few causal genes have resulted from these types of studies and the ones that have been identified were generally known as prior candidate genes. Most functional genetic variation is likely to be much less frequent. Thus, these classical localization approaches will miss most of the functional genetic signal, resulting in the missing heritability problem described above. Therefore, in order to be able to detect the effects of rare functional variants, a high-throughput next-generation sequencing approach needs to be employed to exhaustively search for variants.

16.2.2.3 Gene Localization by Sequencing

The identification of causal genes, using a genome sequencing approach, will obligately generate information on the pathways of these genes and will directly identify novel drug targets. Unlike epidemiological approaches, causal inference is possible using genetic strategies. Because DNA sequence variation is not influenced by other biological or environmental factors, genetic variation that correlates with disease risk must obligately reflect causation. Of course, identifying the exact causal sequence variants is difficult and represents one of the main challenges of modern human genetics. The ability to identify genes that are causally involved with disease risk provides an unparalleled opportunity to quickly determine biological pathways that are involved in pathology. Modern genomic technologies that allow the unbiased examination of all genes simultaneously can be exploited to rapidly identify genes involved in disease susceptibility. Given this information, each gene in an empirically identified molecular network that is proven to be involved in disease risk becomes a potential drug target. Whole genome sequencing allows a comprehensive search for functional sequence variation, to identify novel genes with alterations that have a substantially higher likelihood of representing functional variants of relevance for human physiological variation.

More recently, rare variants have been receiving increased attention in an attempt to explain the "missing heritability" problem observed from GWAS for many common traits. Rare variants are likely to have larger effect sizes and could contribute significantly to missing heritability; and it is postulated that these variants are also likely to have obvious functional consequences (Cirulli and Goldstein 2010; Manolio et al. 2009; Pelak et al. 2010). The primary technology for identifying rare variants is sequencing, either of target regions or entire genomes; however the sample set selected for sequencing will be of particular importance. These topics will be addressed in the following sections.

16.3 Next-Generation Sequencing Applications

16.3.1 Targeted Sequencing

16.3.1.1 Promoter Sequencing

Gene transcription in complex organisms is controlled by the intricate balance of proteins binding to promoter, enhancer and repressor sites within DNA sequences, and is regulated by both cis-acting factors in the flanking gene sequence and trans-acting external modulators regulated by other cellular characteristics. In recent years, there has been considerable interest in the effect of cis-acting variants on gene transcription. Several cis-acting elements exist that are involved in regulating gene expression; however the simplest and potentially most important is the 5′ promoter region. Promoter sequences are the most precisely defined of the many cis-acting regulatory regions of a gene and are of critical importance for their role in initiating gene transcription (Buckland et al. 2005; Coleman et al. 2002a, b; Rockman and Wray 2002). The role of promoters in initiating gene transcription highlights them as a potential source of genetic variation that may affect the expression level of a gene, and given their fixed location, promoters are also an ideal region for both genomic and functional analyses of genetic variation.

Several studies have investigated the functional relevance of variants within promoter regions. Buckland et al. (2005) performed a meta-analysis of approximately 700 gene promoters, interrogating the first 500 bp upstream of the transcription start site. Of the variants investigated, their results showed strong bias towards a promoter location for functional SNPs. Of all SNPs, 50 % were within the first 100 bp and 75 % within 200 bp of the transcription start site (Buckland et al. 2005). A second study performed by Rockman and Wray (2002) investigated 141 promoter variants involved in regulating over 100 genes, spread over the autosomes and X chromosome. They found that 63 % of the 107 genes studied had allelic differences of twofold or greater in their rates of transcription. Similar to the study of Buckland et al. as described above, 58.9 % of the functional variants were located within the first 500 bp upstream of the transcription start site. An additional 12.8 % fell 3′ to the start of transcription and another 12.8 % were more than 1 kb upstream of the transcription start. Only 1.4 % of functional variants were more than 10 kb upstream of their start sites (Rockman and Wray 2002). A more recent study by Sinnett et al. (2006) analyzed the promoter region (defined as 2 kb upstream of transcription initiation) of 197 genes in a multi-ethnic panel of 40 individuals. Their analysis identified 1,838 promoter variants for assessment and results showed that 75 % of the variants predicted functional roles, modifying putative transcription factor binding sites (Sinnett et al. 2006). A number of specific functional genes influencing complex phenotypes in humans have been successfully identified, and in our own work, we identified functional promoter variants in selenoprotein S (SELS), a gene involved in inflammation and in presenilins-associated rhomboid-like (PARL), a gene involved in mitochondrial integrity (Curran et al. 2005, 2010).

16.3.1.2 Candidate Gene Sequencing

Traditionally candidate genes have been identified through a variety of different methods including the comprehensive searching of the publicly available genomic databases, dense SNP mapping or genome-wide transcriptional profiling in the sample population. No matter the method of identification, the next step for a positional candidate gene is to comprehensively resequence the gene in sufficient individuals to maximize the probability of identifying all genetic variation. Selection of the most informative sample set for sequencing is highly dependent on the population being assessed (i.e., families or unrelated individuals). In a large extended pedigree, key members, such as founders, will impart the most information. For samples of unrelated individuals, it will be impossible to capture all variation, though the use of phenotypic extremes for sequencing will likely identify variants of larger effect sizes. The relative position of the variant to the gene's structure strongly influences the probability that the variation affects the function of the gene product. Thus, for resequencing in the sample, the most comprehensive strategy is to include all exons, intronic regions shown to be evolutionarily conserved as identified by comparative genomics (if the total intronic region is too large), 2 kb of the 3′UTR region and up to 5 kb of the putative promoter region, for each gene. Sanger sequencing, the most common method for such sequencing is now too costly and is being surpassed by next-generation sequencing applications on the smaller instruments including the Illumina MiSeq and the Life Technologies Ion Torrent.

16.3.1.3 QTL Sequencing

Linkage analyses identify chromosomal regions, of varying size, that contain QTLs influencing disease risk factors. This information significantly reduces the genomic search space, but still requires further effort to localize the specific genes and variants contributing to the signal. In comparison to a QTL region identified by association (~500 kb), a QTL region identified by linkage is typically 10–15 megabases (Mb) in size. To identify the underlying genes influencing this linkage signal, deep comprehensive sequencing is required. Many studies have performed fine mapping, with SNP markers, across linkage regions to try and narrow the search space, however this has been met with limited success as the variants assessed have only been common. Until recently, the sequencing of a QTL region by Sanger sequencing has been too costly, though with the release of the small scale next-generation sequencing instruments mentioned above, such sequencing is rapid and relatively inexpensive.

16.3.2 Exome Sequencing

Traditional complex phenotype research has focused on the analysis of protein coding variants that directly impact the protein structure and function, often dominant in simple disorders. We are now able to do this on a genome-wide scale, using a whole exome sequencing approach. Whole exome sequencing represents a currently accessible technology that enables the rapid identification of functional protein coding variation influencing phenotypic variation. The exome constitutes approximately 1 % of the human genome. This represents roughly 30 Mb that is split across 200,000 exons. Exome sequencing allows the identification of all coding variants, including those non-synonymous variants that alter protein sequence, which are most likely to have direct functional consequences. Modern sequencing technology allows us to entertain such a comprehensive approach.

Recent studies suggest that exome sequencing can be very powerful and that many rare potentially functional coding variants are likely to be found (Choi et al. 2009; Ng et al. 2010). Using targeted whole exome sequencing of 12 individuals, Ng et al. (2009) found approximately 6,000 non-synonymous variants per individual and predicted that it would have been about 8,500 with better sequencing coverage. These variants also were primarily rare (Ng et al. 2009). Using a less sensitive technique, Hedges et al. (2009) sequenced 8 independent exomes and found an average of 3,847 non-synonymous variants per individual with 683 being novel

(Hedges et al. 2009). All of these sequencing studies suggest a large number of relatively rare protein coding variants lurk within human populations. A recent study by Bowden et al. (2010) identified a rare variant (of ~1 % frequency) that accounts for 17 % of the variance in plasma adiponectin in a large Hispanic American sample, using a family-based whole exome sequencing approach (Bowden et al. 2010).

16.3.3 Whole Genome Sequencing

Whole genome sequencing (WGS) allows a comprehensive search for functional sequence variation, to identify novel genes with alterations that have a substantially higher likelihood of representing functional variants of relevance for human physiological variation. While identifying the exact causal variants influencing a trait represents one of the main challenges of human genetics, the potential to identify such variants using WGS is significantly increased and provides an unparalleled opportunity to quickly determine biological pathways involved. Each gene in an empirically identified network becomes a potential drug target. WGS represents the "holy grail" for genetic studies. Prior approaches have only sampled partial variation from the genome. However, the first studies employing WGS are now being performed in sufficiently large samples to likely produce benefit. One of the earliest applications of WGS in human gene identification has been to severe disorders that are thought to be possible single gene, Mendelian conditions. Sequencing in small samples of such patients has identified putative functional mutations in a large proportion of cases. WGS in a sample of six patients with severe early onset epilepsy and their parents was successful in all six cases, identifying de novo mutations in four individuals, parental isodisomy in one, and a recessive mutation in another (Martin et al. 2014). Among the first published WGS studies for a complex human phenotype is an examination of bipolar disorder in a large Old Order Amish pedigree (Georgi et al. 2014). This study identified multiple chromosomal regions shared among affected family members, each with multiple potential deleterious variants, suggesting a complex and potentially heterogeneous genetic architecture underlying bipolar disorder even in this population isolate.

16.3.4 Other Sequencing Applications

16.3.4.1 RNA Sequencing (RNA-Seq)

The transcriptome is defined as the complement of RNA molecules (transcripts) in a cell. Using modern genomic technology, it is now possible to discover, profile and quantify RNA transcripts (for details, please refer to Chap. 5 by Göring in this volume). Characterization of the transcriptome is essential for identifying and interpreting functional elements of the genome, and understanding disease development. Compared to array based transcript analyses, RNA-Seq provides several advantages over array based assays including a more precise quantification of transcripts and their isoforms than other methods; it is not limited to detecting transcripts that correspond to known genomic sequences; junctions between exons can be assayed; allele-specific expression differences and alternative splicing events can be detected. RNA profiling tools have been around for decades, though tremendous progress has been made in advancing the technology. From the days of Northern blots and serial analysis of gene expression (SAGE) analysis we have moved to gene-expression microarrays and now deep sequencing. With these tremendous advances in technology, the information content obtained from RNA analysis has also significantly increased, and like that of genome sequence, a substantial computational framework is essential.

16.3.4.2 Methylation Sequencing

The methylation pattern of DNA has been shown to influence gene expression patterns.

The implementation of next-generation sequencing has made it possible to study methylation patterns on a genome wide scale, rather than on a gene by gene basis. There are two common forms of methylation sequencing: whole genome bisulfite sequencing and MeDIP-Seq. In whole genome bisulfite sequencing, genomic DNA is bisulfite treated and all unmethylated cytosine bases are converted to uracil. Methylated cytosine bases (those containing a 5′ methyl group; 5′-methylcytosine) are not affected and once sequenced, the methylation status of each allele can be determined (Callinan and Feinberg 2006; Pomraning et al. 2009). This requires whole genome sequencing and is still somewhat cost prohibitive. The second method, MeDIP-Seq is an attempt to reduce the sequenced material, increasing throughput and reducing the cost. In this method, methylated DNA is selected for prior to sequencing using an antibody against 5′-methylcytosine. The unmethylated DNA is then washed away, leaving only the highly enriched methylated material for sequencing (Down et al. 2008; Pomraning et al. 2009). Both of these methods have their own benefits and one best suited to the study design should be chosen.

16.3.4.3 Mitochondrial Sequencing

The mitochondria are essential to life, being the major cellular site of energy production and respiration. A great deal of research has implicated mitochondrial dysfunction in a variety of human diseases including cancer, obesity, multiple sclerosis, several psychiatric disorders and a wide range of age related disorders (Begriche et al. 2006; Dakubo et al. 2006; Dutta et al. 2006; Fattal et al. 2006; Wallace 2005; Weissig et al. 2004). The mitochondrial genome is very small, consisting of about 16.5 kb and encodes genes for the biochemical reactions of respiration, and specific molecules involved in protein synthesis. The genome however only encodes a small number of mitochondrial functioning proteins; most of the proteins found in the mitochondria are nuclear encoded.

Mitochondrial sequencing is very popular among anthropologists and genealogists to investigate human evolution and diversity; however it has been gaining more interest from geneticists given the essential role of the mitochondria in maintaining cellular homeostasis. Given the small size of the mitochondrial genome, Sanger sequencing methods are still feasible; though sequencing is also possible using next-generation sequencing technology and mitochondrial variant information is captured when performing whole genome sequencing.

16.4 Next-Generation Sequencing Study Design

16.4.1 Return of the Family Study

Given the cumulative effect of rare variants on normal phenotypic variation and their importance to disease, different strategies are required to identify such variants than those that have been employed to assess common genetic variation. Most obvious, optimal capture and detection of rare functional variants will require a return to pedigree-based studies. Pedigree studies are one of these implicit designs that provide several advantages to the identification of rare variation, the main advantage being that rarer variants will be present at a much higher frequency than in the general population. Mendelian transmissions from parents to offspring maximize the chance that multiple copies of rare variants exist in the pedigree. These variants can then be identified by direct resequencing and statistical tests that minimize the influence of spurious linkage disequilibrium (Blangero et al. 2005; Kent et al. 2007). The key factor in the identification of rare variants then becomes resequencing sufficient numbers of chromosomes to capture all existing sequence variation. Given the likely importance of rare variation, a comprehensive sequencing strategy is the best means for detecting all such variants with sufficient copies.

Deep sequencing for functional variation in large pedigrees offers many benefits over studies of unrelated individuals, predominantly a greater number of copies of private variants. Additionally, for rare but non-private variants, extended pedigrees lead to substantially increased variance in allele frequency which permits a much wider potential for large variant-specific heritabilities (genetic signals). Given the growing awareness that rare functional variation appears to be responsible for observable phenotypic genetic variation, it is clear that individual pedigrees can provide significant evidence for gene identification for even complex quantitative phenotypes such as lipids. With the vast extent of private functional variation, any pedigree may hold an overt key to a disease-relevant gene. A pedigree-specific rare functional variant with small relative effect size (in relation to population attributable risk or QTN-specific heritability), but with a larger absolute effect size that is further enriched by Mendelian statistical mechanics within an extended pedigree, can be sufficient to verify that a given gene is involved in endophenotype variation.

16.4.2 Unrelated Individual Study Designs

Much of this chapter has focused on the importance of pedigree-based study designs for the identification of rare variation, but what options are possible for unrelated individuals?

For less rare, but still uncommon, variants with minor allele frequencies greater than 0.005, large studies of highly selected unrelated individuals such as those employed in the pioneering work by the research group at the University of Texas Southwestern Medical Center (Cohen et al. 2004, 2005, 2006) have led to the identification of rarer functional variants influencing lipids. However, such studies are inefficient and not suited to the largest functional class involving effectively private functional variants. By definition, any set of unrelated individuals could never capture more than a single copy of such a variant. However, due to the large case/control samples accumulated in various biorepositories during the era of GWAS studies, it is likely that interest in gene discovery in samples of unrelated individuals will continue and these sample collections may prove useful in generalizing family-based gene identification results to population-based samples. In particular, it may be useful to examine genes nominated by family studies in large collections of unrelated individuals to identify and characterize other functional variants in these genes.

16.5 Next-Generation Sequencing (NGS) Technologies

NGS approaches presently available, and that have been implemented in the new wave of sequencing projects, provide DNA read data generated using a high-throughput methodology that employs substantially different underlying chemistry dynamics. While there are several different technologies on the market, in this review we focus on the three companies that have led the revolution: Illumina, Life Technologies (previously Applied Biosystems) and Roche.

16.5.1 Illumina

Illumina recently added the HiSeq 2,500 platform to its family of sequencing instruments, which also includes the HiSeq 2,000, the Genome Analyzer IIx and the MiSeq. The HiSeq 2,500 incorporates the HiSeq 2,000 architecture with the onboard cluster generation of the MiSeq to switch between high output run mode and a rapid turnaround run mode. The rapid run mode is capable of sequencing a 30x genome in a day or fast multiplexed applications, such as exomes or RNA-Seq. Up to 120 Gb of sequence (1.2 billion reads), using 2 × 150 bp read lengths, is generated in ~27 h during this rapid run phase. In high output mode, the HiSeq 2,000 & 2,500 are identical with 600 Gb output (6 billion paired-end reads) in ~11 days for a 2 × 100 bp read length. Given their output, the HiSeq systems are most widely used for whole genome and whole exome sequencing. Additional applications include de novo sequencing, RNA-Seq, small RNA discovery, DNA methylation sequencing and cytogenetic analysis.

The Genome Analyzer (GA) IIx is the most widely published and adopted next-generation sequencing technology. The GAIIx is capable of outputting 95 Gb of data per run (640 million paired-end reads) in ∼14 days at a 2 × 150 bp read length. The GAIIx is most widely used for RNA-seq, ChIP-Seq for gene regulation analysis, small genome sequencing, targeted resequencing, de novo sequencing, and amplicon sequencing.

The MiSeq is a fully integrated sequencing system that performs cluster generation, sequencing and data analysis all onboard the instrument within a 24 h period for 2 × 150 bp paired-end reads. The MiSeq platform is capable of outputting up to 5 Gb of sequence per 27 h run, for 2 × 150 bp reads and up to 8.5 Gb of data in 39 h for 2 × 250 bp runs. Applications of the MiSeq include highly multiplexed amplicon sequencing, targeted sequencing, small genome sequencing, ChIP-Seq and small RNA sequencing.

16.5.2 Life Technologies

The Life Technologies Support Oligonucleotide Ligation Detection (SOLiD) system performs massively parallel sequencing by stepwise ligation (all DNA is sequenced at the same time). The unique assay uses a 2 base encoding to distinguish between SNPs and errors. The SOLiD 4 instrument can generate 100 Gb of sequence data, or 1.4 billion reads per 16 day run, with 2 × 50 bp mate-paired reads. As with the other instruments, the SOLiD 4 can be used for whole genome sequencing, de novo sequencing, targeted sequencing, methylation sequencing, ChIP-Seq, small RNA sequencing and transcriptome sequencing.

The Ion Torrent system is a single instrument that uses semiconductor chips and semiconductor technology for a variety of sequencing throughputs. The Chip is the machine and can scale in density for pretty much any application. The throughput of the instrument is fast with a 2 h sequencing run for 200–400 bp reads, generating a between 10 Mb and 1 Gb of sequencing depending on the Chip selected. The Ion Torrent is best suited to small genome sequencing, targeted sequencing, target capture, RNA-Seq and miRNA-Seq, library assessment and ChIP-Seq.

16.5.3 Roche

The Roche 454 GS FLX+ is an ultra high-throughput automated DNA sequencing system.

The major advantage of the GS FLX + approach to sequencing besides the throughput is its ability to achieve read lengths in the order of 1,000 bp, the longest of any of the NGS technologies. The technology is also flexible enough to combine both long shotgun reads and paired end reads for complex genomes. The instrument can generate 700 Mb of sequence, 1 million shotgun reads, per 23 h instrument run. Applications of the GS FLX+ include genome sequencing, de novo sequencing, targeted sequencing, and transcriptome sequencing.

The Roche 454 GS Junior, like the MiSeq, is a compact integrated system capable of sample prep to analysis in a single run. The instrument generates 35 Mb of sequence (100,000 shotgun or 70,000 amplicon reads at an average of 400 bp) per 10 h run. The GS Junior is best suited to amplicon sequencing, genome and de novo sequencing of microbes, and transcriptome sequencing.

16.5.4 Genetic Analysis Services

It is not always feasible or possible for laboratories to perform sequencing in-house, and for this situation, there are many companies available that will perform sequencing, alignment and some data interpretation services. A web search of sequencing services will alert you to the many companies that are available, but here are some better known services.

16.6 Illumina and Certified Service Providers

The Illumina Genome Network is a network of sequencing teams at different institutions using Illumina sequencing platforms and providing whole genome sequencing services.

Illumina FastTrack Services provides whole genome sequencing services performed at Illumina by their scientists.

16.7 Complete Genomics

CGI provides a complete end-to-end sequencing service for human genomes, from sample preparation to analysis, providing ready called sequence data. The turnaround time is 3–4 months and a minimum 40x mapper coverage is guaranteed, sequenced on their own proprietary technology.

Other service providers include the National Center for Genome Resources, Beckman Coulter Genomics, BGI, SeqWright, HudsonAlpha and CIDR. Data interpretation services are provided by Knome, DNASTAR, Broad Institute and the Sanger Institute.

16.8 Conclusion

Looking back to the advent of the Human Genome Project back in 1990, it then took 13 years to complete the sequencing of 8 human genomes and cost billions of dollars. If we look at the technological advances that have occurred since then and assume this is the way of the future, we can only imagine that we will be able to obtain sequence information much more rapidly than we do now and the price will continue to decrease. The field of human genetics will be dominated by complete genome sequencing. In light of these times, we now need to think ahead and pay some focus to the computational burden that these projects will pose and furthermore, look towards biology and making sense of the genetic information we will soon be overwhelmed with.

References

Begriche K, Igoudjil A, Pessayre D, Fromenty B (2006) Mitochondrial dysfunction in NASH: causes, consequences and possible means to prevent it. Mitochondrion 6:1–28

Blangero J (2004) Localization and identification of human quantitative trait loci: king harvest has surely come. Curr Opin Genet Dev 14:233–240

Blangero J, Göring HHH, Kent JWJ, Williams JT, Peterson CP, Almasy L, Dyer TD (2005) Quantitative trait nucleotide analysis using Bayesian Model Selection. Hum Biol 77:541–559

Bowden DW, An SS, Palmer ND, Brown WM, Norris JM, Haffner SM, Hawkins GA, Guo X, Rotter JI, Chen YD, Wagenknecht LE, Langefeld CD (2010) Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the ADIPOQ gene in the IRAS Family Study. Hum Mol Genet 19:4112–4120

Buckland PR, Hoogendoorn B, Coleman SL, Guy CA, Smith SK, O'Donovan MC (2005) Strong bias in the location of functional promoter polymorphisms. Hum Mutat 26:214–223

Callinan PA, Feinberg AP (2006) The emerging science of epigenomics. Hum Mol Genet 15 Spec 1:R95–101

Chasman DI, Paré G, Mora S, Hopewell JC, Peloso G, Clarke R, Cupples LA, Hamsten A, Kathiresan S, Mälarstig A, Ordovas JM, Ripatti S, Parker AN, Miletich JP, Ridker PM (2009) Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. PLoS Genet 5:e1000730

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 106:19096–19101

Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Natl Rev Genet 11:415–425

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305:869–872

Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci USA 103:1810–1815

Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH (2005) Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Natl Genet 37:161–165

Coleman SL, Buckland PR, Hoogendoorn B, Guy CA, Smith K, O'Donovan MC (2002a) Experimental analysis of the annotation of promoters in the public database. Hum Mol Genet 11:1817–1821

Coleman SL, Hoogendoorn B, Guy CA, Smith SK, O'Donovan MC, Buckland PR (2002b) Streamlined approach to functional analysis of promoter-region polymorhisms. Biotechniques 33:412–418

Curran JE, Jowett JB, Elliott KS, Gao Y, Gluschenko K, Wang J, Abel Azim DM, Cai G, Mahaney MC, Comuzzie AG, Dyer TD, Walder KR, Zimmet P, MacCluer JW, Collier GR, Kissebah AH, Blangero J (2005) Genetic variation in selenoprotein S influences inflammatory response. Natl Genet 37:1234–1241

Curran JE, Jowett JB, Abraham LJ, Diepeveen LA, Elliott KS, Dyer TD, Kerr-Bayles LJ, Johnson MP, Comuzzie AG, Moses EK, Walder KR, Collier GR, Blangero J, Kissebah AH (2010) Genetic variation in PARL influences mitochondrial content. Hum Genet 127:183–190

Dakubo GD, Parr RL, Costello LC, Franklin RB, Thayer RE (2006) Altered metabolism and mitochondrial genome in prostate cancer. J Clin Pathol 59:10–16

De La Vega FM, Bustamante CD, Leal SM (2011) Genome-wide association mapping and rare alleles: from population genomics to personalized medicine. Pac Symp Biocomput 74–75

Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJ, Durbin R, Tavaré S, Beck S (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Natl Biotechnol 26:779–785

Dutta R, McDonough J, Yin X, Peterson J, Chang A, Torres T, Gudz T, Macklin WB, Lewis DA, Fox RJ, Rudick R, Mirnics K, Trapp BD (2006) Mitochondrial dysfunction as a cause of axonal degeneration in multiple sclerosis patients. Ann Neurol 59:478–489

Fattal O, Budur K, Vaughan AJ, Franco K (2006) Review of the literature on major mental disorders in adult patient with mitochondrial diseases. Psychosomatics 47:1–7

Georgi B, Craig D, Kember RL, Liu W, Lindquist I, Nasser S, Brown C, Egeland JA, Paul SM, Bućan M (2014) Genomic view of bipolar disorder revealed by whole genome sequencing in a genetic isolate. PLoS Genet 10:e1004229

Hedges D, Burges D, Powell E, Almonte C, Huang J, Young S, Boese B, Schmidt M, Pericak-Vance MA, Martin E, Zhang X, Harkins TT, Züchner S (2009) Exome sequencing of a multigenerational human pedigree. PLoS ONE 4:e8232

Kent JW Jr, Dyer TD, Göring HHH, Blangero J (2007) Type I error rates in association versus joint linkage/association tests in related individuals. Genet Epidemiol 31:173–177

Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311–321

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Martin HC, Kim GE, Pagnamenta AT, Murakami Y, Carvill GL, Meyer E, Copley RR, Rimmer A, Barcia G, Fleming MR, Kronengold J, Brown MR, Hudspith KA, Broxholme J, Kanapin A, Cazier JB, Kinoshita T, Nabbout R; The WGS500 Consortium, Bentley D, McVean G, Heavin S, Zaiwalla Z, McShane T, Mefford HC, Shears D, Stewart H, Kurian MA, Scheffer IE, Blair E, Donnelly P, Kaczmarek LK, Taylor JC (2014) Clinical in severe early-onset epilepsy reveals new genes and improves molecular diagnosis. Hum Mol Genet Feb 11 [Epub ahead of print]

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ (2010) Exome sequencing identifies the cause of a mendelian disorder. Natl Genet 42:30–35

Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, Smith JP, Cirulli ET, Fellay J,Dickson SP, Gumbs CE, Heinzen EL, Need AC, Ruzzo EK, Singh A, Campbell CR, HongLK, Lornsen KA, McKenzie AM, Sobreira NL, Hoover-Fong JE, Milner JD, Ottman R, Haynes BF, Goedert JJ, Goldstein DB (2010) The characterization of twenty sequenced human genomes. PLoS Genet 6: e1001111

Pomraning KR, Smith KM, Freitag M (2009) Genome-wide high throughput analysis of DNA methylation in eukaryotes. Methods 47:142–150

Rockman MV, Wray GA (2002) Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol 19:1991–2004

Sinnett D, Beaulieu P, Belanger H, Lefebvre JF, Langlois S, Theberge MC, Drouin S, Zotti C, Hudson TJ, Labuda D (2006) Detection and characterization of DNA variants in the promoter regions of hundreds of human disease candidate genes. Genomics 87:704–710

Wallace DC (2005) A Mitochondrial paradigm of metabolic and degenerative diseases, aging and cancer. Annu Rev Genet 39:359–407

Weissig V, Cheng SM, D'Souza GG (2004) Mitochondrial pharmaceutics. Mitochondrion 3:229–244