UNIT 2.

BIOMARKERS: PRACTICAL ASPECTS

CHAPTER 6.

Basic principles and laboratory analysis of genetic variation

Jesus Gonzalez-Bosquet and Stephen J. Chanock

Summary

With the draft of the human genome and advances in technology, the approach toward mapping complex diseases and traits has changed. Human genetics has evolved into the study of the genome as a complex structure harbouring clues for multifaceted disease risk with the majority still unknown. The discovery of new candidate regions by genome-wide association studies (GWAS) has changed strategies for the study of genetic predisposition. More genome-wide, “agnostic” approaches, with increasing numbers of participants from high-quality epidemiological studies are for the first time replicating results in different settings. However, new-found regions (which become the new candidate “genes”) require extensive follow-up and investigation of their functional significance. Understanding the true effect of genetic variability on the risk of complex diseases is paramount. The importance of designing high-quality studies to assess environmental contributions, as well as the interactions between genes and exposures, cannot be stressed enough. This chapter will address the basic issues of genetic variation, including population genetics, as well as analytical platforms and tools needed to investigate the contribution of genetics to human diseases and traits.

Introduction

New advances in microchip technologies and informatics allow geneticists to look across the genome agnostically using dense data sets with billions of data points. These developments have transformed the field, moving it away from the pursuit of hypothesis-driven, limited candidate studies to large-scale scans across the genome. Together these developments have spurred a dramatic increase in the discovery of genetic variants associated with or linked to human diseases and traits, many through genome-wide association studies (GWAS) (1). Already over 7400 novel regions of the genome have been associated with more than 75 human diseases or traits in large-scale GWAS (2). Each region now represents a new candidate “region” that harbours putative genes, which will require extensive mapping of the variants to explore the genomic architecture of the region and its contribution to human diseases and traits. The return to exploring candidate regions differs from the old approach of nominating favoured genes, because it is driven by findings that reach conclusive thresholds based on more rigorous statistical considerations.

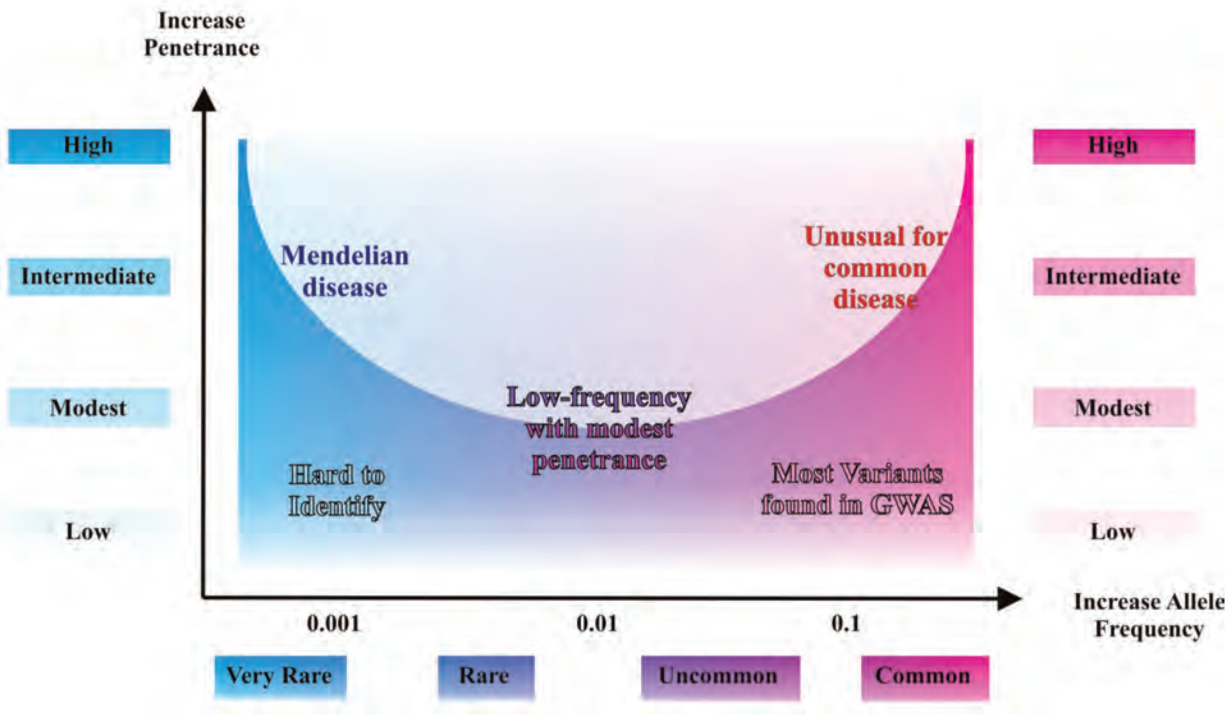
While there is ample opportunity to survey thousands of genetic variants, often well chosen and based on an emerging understanding of the structure of genetic variation and its patterns of inheritance, the ability to analyse the interaction between genetic variants and the environment has lagged. This is mainly because the measurement tools for the latter have not undergone the transformative shift observed in assessing genetic variation. The integration of environmental exposure with genetic factors should provide insights into disease mechanisms and outcomes. Eventually these insights will be applied to treatment or preventive measures that are best suited for the individual (known as personalized medicine). Individualization of treatments based on the greatest likelihood for efficacy, while minimizing (or avoiding) deleterious toxicities, represents a long-term goal, but one that is in the distant future. While the opportunity to begin to develop evidence-based individualized therapeutics, also known as pharmacogenomics, is promising, its realization will require a nuanced understanding of the contribution of genetic variation to complex diseases.

This chapter will address the basic issues of genetic variation, including population genetics as well as analytical platforms and tools needed to investigate the contribution of genetics to human diseases and traits.

The scope of genetic variation

The spectrum of human genetic variation is enormous with respect to both the types of genetic variation and the sheer magnitude of the number of variants in any given genome. Even though two genomes are estimated to differ by less than 0.5%, there are still several million differences; the majority are vestigial, but a small proportion probably contribute to disease risk. The most common type of variation is a single nucleotide base change, followed by small insertions or deletions in sequence. Progressively larger structural alterations and copy number variants are fewer in absolute number, but perhaps affect more bases (Figure 6.1). So far, available technologies have accelerated the discovery and characterization of diversity in the human genome. In the first wave of annotation, common variants have

Figure 6.1. Genetic variant frequencies and estimated effect size for genetic contribution



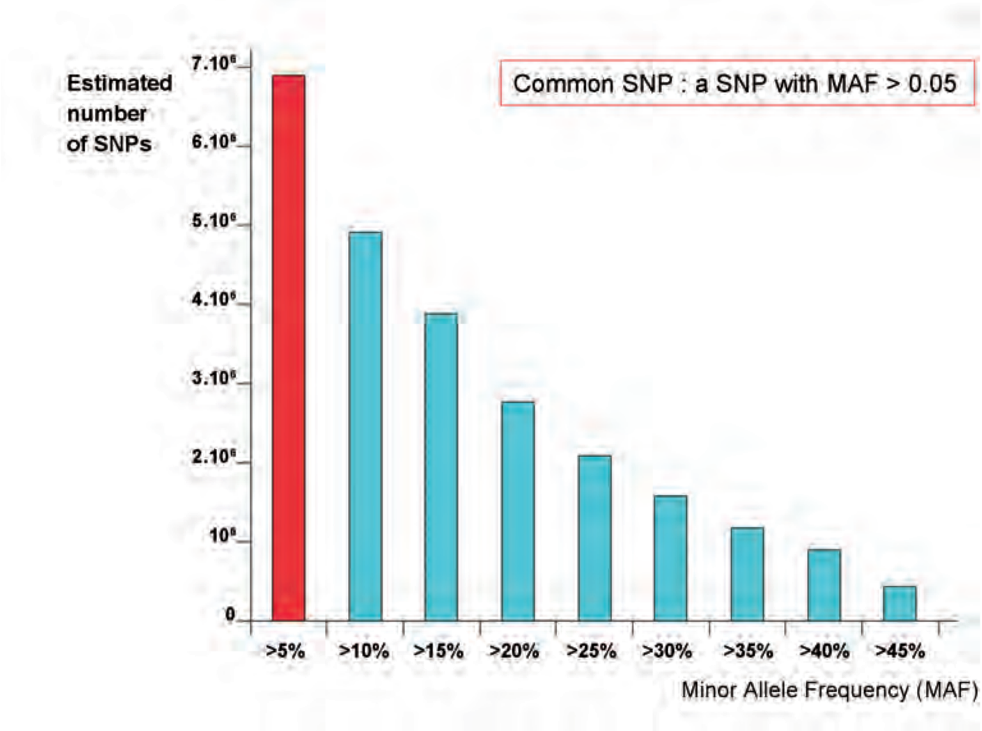
been described, many of which are universal to all populations. The ability to ascertain estimates for lower frequency variants is dependent upon the number of subjects surveyed, as well as the population genetic history of the subjects used for discovery. New sequencing technologies, referred to as next-generation sequencing, allow for the ability to catalogue variants with lower frequencies and will certainly shift the paradigms further. Generally, the interrogation of genetic variation continues to reveal greater complexity in different human populations, which manifests as differences in frequencies of variants.

Single-nucleotide polymorphisms (SNPs)

The most common sequence variation in the genome, the single-nucleotide polymorphism (SNP), is the stable substitution of a single base, which by definition is observed in at least 1% of a population. Though this definition has been useful for cataloging genetic variation, the advent of next-generation sequencing technology has revealed the sheer breadth of variations in different populations with estimated frequencies well below 1%. Still, for the purpose of current applications of genetic variation, the SNP is the most commonly annotated variant. The minor allele frequency (MAF) is designated for the lower allele frequency observed at a locus in one particular population, but often there can be major differences in estimated MAFs between populations with distinct histories. The literature suggests that there are more than perhaps 15 million SNPs with a MAF greater than 1% (3–5), and 10 million SNPs with a MAF greater than 10% (3,6,7); however recent large-scale sequencing efforts, such as the 1000 Genomes Project, indicate these estimates are low (http://www.1000genomes.org/). There are estimated to be a greater number of SNPs with lower MAFs and, unlike common SNPs, the majority may be population-specific (Figure 6.2). The majority of common SNPs, with a MAF greater than 15–20%, are widespread in human populations (8,9). Only a small subset of high-frequency SNPs (less than 10%) appear to be found in a single population, again suggesting the universal ancestry of common SNPs (9).

Previously in candidate gene approach studies, SNPs in coding regions were often selected on the basis of an *in silico* predicted effect, but with little supporting biological evidence. The attempt to classify coding variants, known as a coding SNP (cSNP), has focused on the predicted effect on the actual coding sequence. The majority of cSNPs

Figure 6.2. Estimated number of SNPs in the human genome in relation with their minor allele frequency (MAF). Source: (5). Reprinted by permission from Macmillan Publishers Ltd: Nature Genetics, copyright (2003).



do not alter the predicted amino acid and are known as synonymous SNPs. However, a subset of variants are predicted to shift the amino acid and are known as non-synonymous coding SNPs. Though this subset was initially of great interest, very few non-synonymous coding SNPs have actually been conclusively associated with human diseases or traits, and even fewer have corroborative biological data to provide plausibility for the association (10,11). Nonetheless, the analysis of synonymous and non-synonymous SNPs has been quite informative for evolutionary studies (12,13).

There has been considerable effort to calculate the effect of a non-synonymous cSNP in conformational protein changes. A proliferation of prediction software has been created (e.g. Protein Data Bank (http://www.rcsb.org/pdb) and Swiss-Model (http://swissmodel.expasy.org//SWISS-MODEL.html)). Though new models and algorithms claim improved reliability for predicting deleterious changes in protein structure (14–16), without corroborative laboratory data the findings are merely *in silico* observations. Overall, between 50 000 and 250 000 SNPs could be functional, non-synonymous coding variants, or regulators of gene expression or splicing (10,11). It is likely that a subset of non-synonymous cSNPs contribute to regulatory differences in expression or genetic pathways (17–19), but most SNPs appear not to be functional and have been maintained on the backbone of an inherited block of DNA through generations. Subsets of SNPs that alter regulation or expression of a gene, called regulatory SNPs (rSNPs), are difficult to predict with high efficiency and most likely will be categorized on the basis of large-scale surveys of cell lines, as well as laboratory data.

Nearly half of the more than 10 million human SNPs in the international public database for SNPs, or dbSNP (http://www.ncbi.nih.gov/SNP/), have been validated with genotyping assays by the SNP Consortium and the International HapMap Project (8,20). Until recently, only a small percentage had been verified by sequencing, but with the advent of the 1000 Genomes Project, nearly all common (MAF >10%) and uncommon (MAF between 1 and 10%) variants should be confirmed by next generation sequence technology (21,22). In the current build, roughly one sixth of the variants in dbSNP are probably monoallelic, due to errors in either genotyping or, more likely, sequencing (23,24). In general, the reported SNPs have been biased towards high-frequency variants in populations of European ancestry.

Currently, the catalogue of uncommon variation, namely SNPs with MAFs under 1%, is incomplete. However, the 1000 Genomes Project is expected to generate a thorough catalogue of variants with greater than 1% MAF. The contribution of uncommon variants (MAF between 1% and 10%) represents an untapped portion of the genomic architecture. It will require either larger studies to provide sufficient power to detect association, or new design strategies to discover and characterize uncommon and rare variants (25,26). Rare or uncommon variants have been shown to be informative in the extremes of mapping human traits, such as with cholesterol levels (27). Rare variants or mutations can explain a proportion of the strong familial component of complex diseases, as well as the classical Mendelian inheritance of single or oligogenic diseases. These highly penetrant disease mutations are catalogued in a public database, the Online Mendelian Inheritance in Man (OMIM) (http://www.ncbi.nlm.nih.gov/omim/).

The correlation of common genetic variants

Most SNPs are not inherited independently but in blocks, resulting in sets of SNPs being transmitted together between generations (4,28,29). These blocks are defined by linkage disequilibrium (LD), which estimates the correlation between SNPs on shared chromosomes passed down from ancestral chromosomes. LD is defined as the non-random association of alleles at different loci (30). Initially, each SNP is a single mutation that has taken hold and become fixed in a population, either as a consequence of direct selection or because it is close enough on the chromosome to be included within a block of a shared segment. Individual SNPs that are strongly associated with each other are said to be in LD, although this correlation could be eroded over time by recombination (exchange of genetic material) during meiosis (31). Haplotypes are defined as sets of SNPs, and other genetic polymorphisms (larger in size), on chromosomal segments that are in strong LD.

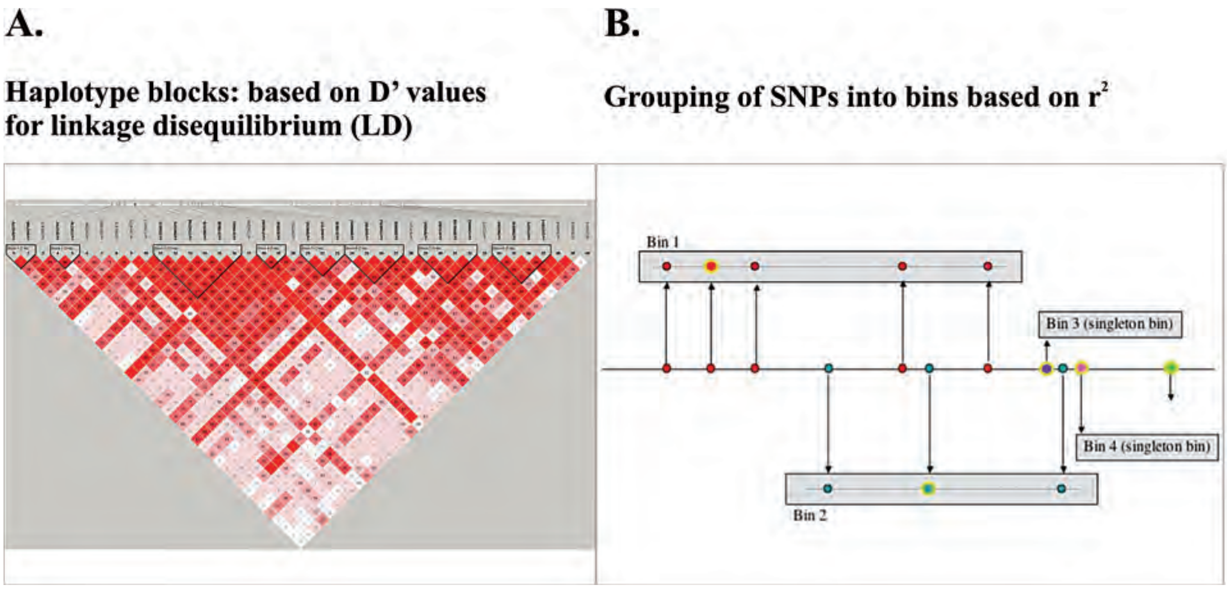
There are several ways to determine haplotypes from genotypes; this is commonly referred to as resolving haplotype phase. The offspring haplotype phase can be determined if the parental genotypes are known or directly with biochemical methods (30). Based on the assumption that haplotypes are randomly joined into genotypes, phasing can be estimated using one of several statistical methods that can account for the ambiguity of unobserved haplotypes (30). Different methods have been developed to estimate haplotypes from unphased multilocus genotype data in unrelated individuals; the underlying principles are based on models that incorporate either a maximum likelihood (32), parsimony (33), combinational theory (34) or *a priori* distribution derived from coalescent theory (35). This last method is the basis for the phase reconstruction software PHASE (35,36), which has performed favourably in simulation studies (37), and its modified version designed for larger data sets, fastPHASE (22). Reconstructed haplotypes from unrelated individuals and LD structure have been used to study genomic association to complex traits (17,29). In fact, some research suggests that haplotypes would be better suited for candidate studies because of a perceived statistical advantage over the single-locus LD mapping (38–40), but the recent success with GWAS suggests otherwise.

The concept of LD also permits investigators to look at a set of SNPs and determine proxies for other untested SNPs (or tagSNPs) (41,42). This indirect approach is predicated on finding markers only, relegating the search for causal or functional variants to later work (Figure 6.3). Several approaches optimize the number of surrogate SNPs needed to account for untested variants, such as the “greedy algorithm.” The latter estimates highly correlated SNPs, primarily on the basis of the MAF, to create heuristic bins of tagged SNPs. Thus, tagSNPs represent proxies for additional, highly correlated SNPs with comparable allele frequency and distribution in the population of interest. In a sense, tagSNPs are used to mark common haplotypes in the region (Tagger, embedded in Haploview software (http://www.broad.mit.edu/mpg/haploview) and TagZilla (http://tagzilla.nci.nih.gov)) (43). Consequently, the indirect approach of using a limited set of tagSNPs as a proxy of a LD block has emerged as the preferred approach, used by both GWAS and candidate gene studies (44).

Structural polymorphisms

Structural variations in the genome may be either cytologically visible or, more commonly, submicroscopic variants that can range in size from a few base pairs to thousands (45,46). These can include deletions, insertions and duplications collectively known as copy number variations (CNVs), as well as less-frequent inversions and

Figure 6.3. SNP selection strategy. A. SNP selection through haplotype blocks, based on the concept of linkage disequilibrium (LD). D’ is a measure of LD between SNPs, represented in the figure through a heat map from white (low D’) to red (high D”). A haplotype (represented by a dark triangle in the figure) is a set of SNPs in strong LD, or high D’. “TagSNPs” are proxies for other SNPs in the same haplotype. This is the so-called indirect approach (147). B. Selection of SNPs based on r2, another measure of LD. This method creates groups with similar LD (r2) into ‘bins.’ In the figure each spot represents a SNP, and those with similar r2 are included in the grey blocks or ‘bins.’ ‘TagSNPs’ are proxies for all these loci included in each ‘bin’ with comparable LD (148)



A.

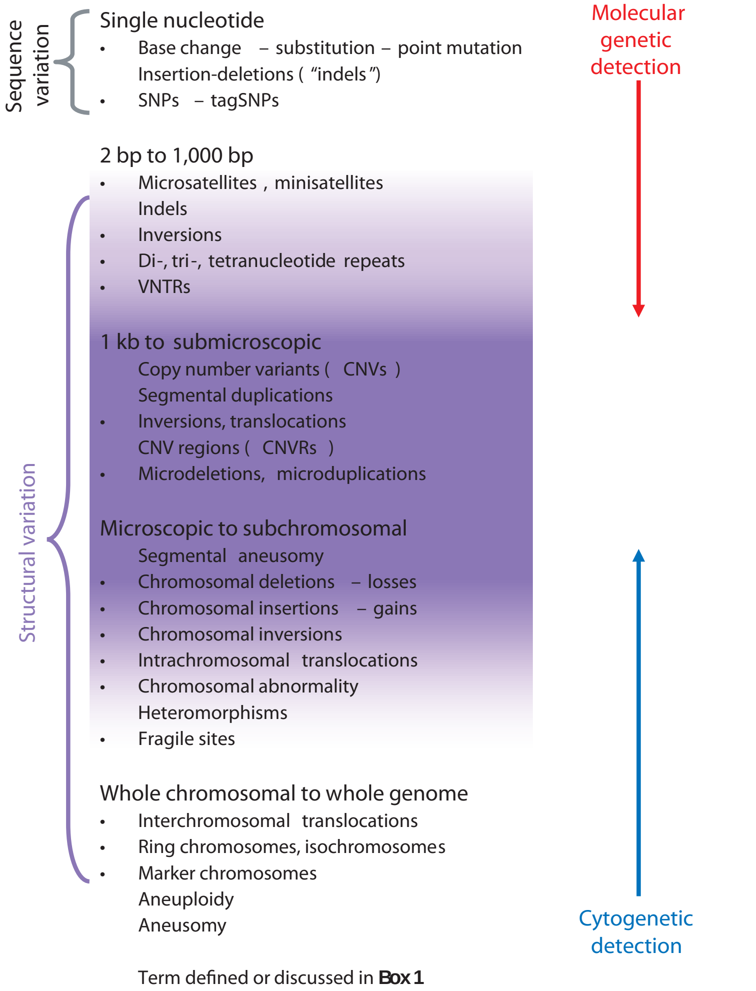
Haplotype blocks: based on D’ values for linkage disequilibrium (LD)

B.

Grouping of SNPs into bins based on r^2

translocations (Figure 6.4) (47,48). Several of the inversions can be quite large, such as the 3.5 Mb on chromosome 17 seen in as much as 20% of the European population (49). On the other hand, insertion/deletions as small as two base pairs can be observed. Although structural variants in some genomic regions have no obvious phenotypic consequence (50–52), CNVs have been shown to influence gene dosage in select circumstances. Consequently, many have pursued CNVs because of the potential contribution of high estimated effects for complex diseases, either alone or in combination with other factors (53). Some observations, either by the failure to assemble the draft genome sequence or by actual experimentation, estimate the segmental duplicated genomic sequence could involve between 5–10% of the genome (51,54,55). Other clues come from the recognition that a notable number of SNPs failed the quality control metrics in the International HapMap Project; these were later determined to reside in regions now known to be enriched for CNVs (7,45,55–57). Current surveys suggest that CNVs are less common than previously reported (58), and many are infrequent (59). It is also notable that over three fourths of common CNVs are in LD with common SNPs (59).

Figure 6.4. Spectrum of genomic variation. Challenges and standards in integrating surveys of structural variation: The range of genetic variation that must be taken into account when designing and analyzing genotype studies (46). The figure represents the whole spectrum of human genetic variation, from the molecular level with DNA sequence variation, exemplified by SNPs, to structural variation, a broad category that includes variations from 2 bp to whole chromosomal variations. The focus of recent genetic studies has been the subgroup in the midrange (with strong highlighting). These forms of variation have been studied with molecular methods to cytogenetic approaches. Reprinted by permission from Macmillan Publishers Ltd: Nature Genetics, copyright (2007).


Sequence variation
Single nucleotide
Base change – substitution – point mutation
Insertion-deletions ("indels")
SNPs – tagSNPs
Molecular genetic detection
Structural variation
2 bp to 1,000 bp
Microsatellites, minisatellites
Indels
Inversions
Di-, tri-, tetranucleotide repeats
VNTRs
1 kb to submicroscopic
Copy number variants (CNVs)
Segmental duplications
Inversions, translocations
CNV regions (CNVRs)
Microdeletions, microduplications
Microscopic to subchromosomal
Segmental aneusomy
Chromosomal deletions – losses
Chromosomal insertions – gains
Chromosomal inversions
Intrachromosomal translocations
Chromosomal abnormality
Heteromorphisms
Fragile sites
Whole chromosomal to whole genome
Interchromosomal translocations
Ring chromosomes, isochromosomes
Marker chromosomes
Aneuploidy
Aneusomy
Cytogenetic detection
Term defined or discussed in **Box 1**

Coordinated efforts are underway to establish a comprehensive catalogue of CNVs, such as the Database of Genomic Variants (http://projects.tcag.ca/variation/) (46,60) and the Human Genome Structural Variation Project (http://humanparalogy.gs.washington.edu/structuralvariation/). Recently, there have been several international efforts to establish standards for identification, validation and reporting of CNVs (46). The availability of several microarray platforms that can detect quantitative imbalances has accelerated CNV discovery, but there are still substantive technical challenges due to the breadth of polymorphic differences for which analyses are particularly unstable. New emerging algorithms should streamline moderate- to high-throughput, cost-effective methods to scan the genome for CNVs, as well as inversions or translocations based on stable sequence assemblies (59–64). Advances in techniques have improved determination of common CNVs, such as tiling arrays (which cover the genome through partial overlapping (tile-like) sets of fixed oligonucleotides), paired-end sequencing (sequence analysis of both ends of a larger fragment to improve alignment), and new dense SNP genotyping platforms based on probe intensity (e.g. Illumina and Affymetrix).

Short tandem repeats (STRs) represent a class of polymorphisms that occur when a pattern of two or more nucleotides are repeated in certain areas of the genome. Previously known as microsatellites, they were frequently employed to conduct linkage studies in potentially informative pedigrees. The patterns can range in length from 2–10 base pairs (usually tetra- or penta-nucleotide repeats) and are typically located in non-coding regions. Since longer repeat sequences can be susceptible to artefactual errors in genotyping accuracy, particularly related to problems of PCR amplification, the industry standard for both genetic analysis and forensic application is 4–5 base pair (bp) repeat units. Shorter repeat sequences (e.g. 2 or 3 bp) tend to suffer from artefacts, such as stutter and preferential amplification (65–67). By genotyping a sufficient number of STR loci, it is possible to generate a unique genetic profile of an individual.

Population genetics

The field of population genetics has advanced rapidly and emerged as central to the investigation of genetics and complex diseases. Overall, the discipline of population genetics seeks to characterize the genetic composition of biological populations, as well as the changes in genetic composition that occur from environmental and migratory factors, including natural selection. To draw conclusions about the likely patterns of genetic variation in actual populations, population geneticists develop abstract mathematical models of gene frequency dynamics and test these conclusions against empirical data. Some of the more robust concepts in population genetic analysis that are applied in disease mapping are discussed below.

Fitness for Hardy–Weinberg proportion

The fitness for Hardy–Weinberg proportion, an important tool for understanding population structure, examines the distribution of the allelic and genotypic frequencies. Though theoretical, it states that if certain assumptions are met, genotype and allele frequencies can be estimated from one generation to the next. The derivation of the Hardy–Weinberg principle for a single locus assumes: a randomly mating population; an infinitely large population, or a population size large enough that random fluctuations in allele and genotype frequencies are small; no mutation; no migration; and no fitness differences among genotypes. When all of these assumptions are met, Hardy–Weinberg Equilibrium (HWE) is established and four important conclusions can be drawn: 1) allele frequencies do not change from one generation to the next; 2) genotype frequencies can be inferred from allele frequencies; 3) only one generation is required to go from non-equilibrium to equilibrium; and 4) once the system is in HWE, it stays in HWE (68). Also, if these conditions are met, the genotypic and allelic frequencies of the offspring generation will be related by the following simple equations. For a trait in the population with two alleles (A_1 and A_2), if the A_1 allele frequency in the population is p, and the A_2 allele frequency is q = (1-p), then expected genotype proportions (f) under HWP are:

$f(A_1A_1) = p^2$, $f(A_1A_2) = 2pq$, $f(A_2A_2) = q^2$

Random mating, or the absence of a genotypic correlation between mating partners, will generate a distribution of observed genotypes that should not deviate significantly from the expected proportions (Hardy–Weinberg Proportions (HWP)). This is predicated on Mendel's law of segregation, and, assuming the absence of selection, all parents contribute equal numbers of gametes to the pool. The HWE principles can be applied to family-based and case–control data to detect genotyping error, population stratification and association.

A violation of any of the above assumptions can produce deviation from HWE, which may include mating behaviour, population size and migration patterns. For example,

Table 6.1. Issues for generation of final, publication-grade build of high-density genotype data

- Eliminate samples with low completion rates (< 90%)
- Remove SNP assays with low call rates (< 90%)
- Determination of fitness for Hardy–Weinberg proportion
- Compare expected duplicates
- Investigate unexpected duplicates
- Assess concordance between duplicates
- Search for cryptic relatedness between subjects
- Assessment of population substructure (after filtering 1st degree relatives)
- Determine admixture with STRUCTURE analysis
- Estimate population stratification (principal component analysis)
- Assess genotype calling algorithm
- Validate significant genotype calls with second technology

systematic inbreeding will increase levels of homozygosity across the genome, as will small population sizes (68). Having more than one random mating population in a sample may also cause deviations from HWE, as well as mating with certain phenotypes (known as assortative mating), which will increase homozygosity as well. Small population size causes allele frequencies to drift from one generation to the next. In many cases, the deviations are also a screen for performance of the genotype technology, because a disproportionate number of heterozygotes or homozygotes could represent systematic errors in genotyping.

One of the most common reasons for not using data in association studies is presumed genotyping error. Many types of errors in genotyping can cause deviations from HWE; therefore tests for both assay specificity and deviation from HWE have been proposed to minimize the genotype error rate and thereby improve data quality (69,70). Deviation from HWE resulting from allelic drop-out, where some alleles are insufficiently amplified, can cause an excess of homozygotes and increase false-negative or false-positive results (71). However, caution should be exercised in association studies before removing data because of HWE deviations. If there is a systematic HWE deviation in both cases and controls, it may be easier to determine a genotyping error if both deviations occurred in the same direction (72). Non-systematic error is more problematic and should trigger a review of standard operating procedures for biospecimen handling, as well as an assessment of all information workflow. If the error is recognized, re-genotyping of the faulty samples might eliminate the problem. The power to detect deviations due to genotyping error under most modes of inheritance has been found to be very small (73). Even the deviation created by neighbouring SNPs, which diminish the performance of genotyping assays, does not produce a large enough deviation from HWE to be detected (74).

In GWAS, it is likely that hundreds if not thousands of markers will deviate from HWE. Understanding why and how HWE testing would help in the process of disease-gene discovery is becoming more important as the number of SNPs included in these studies increases into the hundreds of thousands (75). The control observed genotype frequencies are tested against control expected genotype frequencies to determine if there may be genotyping error (68).

Spectrum of differences in population substructure

The age of GWAS has generated sufficiently large data sets that can determine the degree of differences in underlying population substructure, also known as population stratification. An examination of thousands of markers not in LD permits investigators to assess the extent of admixture and exclude individuals who are outliers for the association analysis.

Classically, population stratification is present when there is a measurable difference in the distribution of alleles between subgroups that have different population histories. There are examples of this in older case–control studies where the cases and controls have been drawn from different populations. It is also possible to have stratification between cases and controls based on differences in exposures, as well as in the distribution of common SNP markers (76). The ability to detect stratification with any marker or set of markers may also vary depending on the allele frequency in each subgroup (68).

In general, an assessment of the underlying structure can be estimated using standard algorithms to identify distinct populations (77). The most commonly used approach is implemented in the STRUCTURE program. This uses multilocus genotype data to examine population structure by attempting to separate subjects into groups (defined as k populations) and determining the distribution of shared alleles.

As the ability to understand population stratification (or differences between cases and controls due to systematic ancestral differences) has improved, several methods have been developed to study and account for these types of systematic study population structures. One approach commonly used for the correction of population stratification is to adjust simultaneously for a fixed number of top-ranked principal components resulting from a principal component analysis (PCA) (76). It is critical to look for underlying subgroups in stratified samples by testing sets of genetic markers not linked to the phenotype, and then adjust for inflation due to stratification (76,78,79). An alternative approach is to use a structured association method in association mapping, permitting case–control analysis in the context of known differences in population structure.

In select circumstances, in which the epidemiologic data suggest major differences between populations, it is possible to conduct mapping by admixture of linkage disequilibrium (MALD). This capitalizes on the concept of admixture, which is the genetic mix of two or more distinct populations. It relies on the differences in allele frequencies between populations to guide the search to focus on changes in the genome rather than a specific gene(s). So far it has been successful in mapping a key prostate cancer region on 8q24 and a type of end-stage renal disease that is more common in individuals of African American background (80–82).

Selection

Population geneticists often define evolution as a change in a population's genetic composition over time. The four factors that can bring about such a change are natural selection, mutation, random genetic drift, and migration into or out of the population. More controversial is a possibility of changes in the mating pattern, which some consider not to be part of classical evolutionary change. Natural selection occurs when some genotypic variants in a population enjoy a survival or reproduction advantage over others. Although the concept that natural selection favours the survival of individuals with a fitness advantage now almost seems intuitive, it was largely opposed when introduced by Darwin (83). Under Mendelian inheritance and with random mating, genotype frequencies after one generation do not change; the determinant of whether the allele will spread in the population is the fitness of heterozygotes versus that of wild-type homozygotes.

Mutation is the primary source of genetic variation driving differences within a population and thus preventing homogeneity. Although mutations that occur in the genome are initially thought to be random, the distribution of biologically significant mutations that cause diversity appears to be non-random (84,85). Gene function, gene structure and the roles of genes and gene products in genetic networks can influence whether particular mutations will contribute to advantageous phenotypic changes. Some mutations generate specific phenotypic changes, whereas pleiotropic mutations alter several seemingly unrelated traits. Mutations with pleiotropic effects will rarely change all phenotypic traits in a favourable way, and, in some instances, may even reduce fitness (86). The same mutation in a different genetic background may produce a different phenotypic effect because of interactions between alleles, under the phenomenon called epistasis. Also, populations exposed to repeated environmental changes may present with different genetic changes that produce a range of phenotypes suited to the environmental conditions, namely phenotypic plasticity.

Initially, when the environment favours a phenotype that is largely different from the average one in a population, mutations that cause this phenotypic change towards the new optimum are favoured (called strength of selection). Population size and history also influence genetic evolution. A small population size can accentuate the effects of random sampling of alleles, so-called genetic drift. In small populations, genetic drift will allow deleterious alleles to occasionally increase in frequency (84).

Random genetic drift refers to the chance fluctuations in gene frequency that arise in finite populations; it can be thought of as a type of "sampling error." In many evolutionary models, the population is assumed to be infinite or very large to avoid chance fluctuations. This assumption is often not realistic, and species with historically low effective population sizes, such as humans, show evidence for reduced variability and effectiveness of selection in comparison with other species (87,88). In the era of multispecies comparisons of genome sequences and GWAS, it is critical to assess the evolutionary role of genetic drift and its interactions with mutation, migration, recombination and selection. Therefore, population size plays a central part in modern studies of molecular evolution and variation (88).

One of the most influential variables for human genetic variation is geographic location, with genetic differentiation between populations increasing with geographic distance and genetic diversity decreasing with distance from Africa. Populations of African ancestry have the greatest diversity, resulting in shorter segments of LD (89–93). Modern population genetics estimates that the ancestral human population originated in Africa and radiated outward to other continental locations, both within Africa and elsewhere.

Alleles under positive selection can increase in prevalence in a population and leave distinctive signatures, or patterns of genetic variation, in DNA sequences. These can be identified by comparison with the background distribution of genetic variation, primarily evolved under neutrality (94). In some cases, these signatures, or differences in allele frequencies between populations, reflect major regional selective pressures like infectious diseases (e.g. malaria), environmental stresses (e.g. temperature), or dietary factors (e.g. milk consumption) (13,95,96).

When immigrants with a different genetic makeup enter a new population, the population's genetic composition will be altered. The evolutionary importance of migration stems from the fact that many species are composed of several distinct subpopulations, largely isolated from each other but connected by occasional migration. Migration between subpopulations gives rise to gene flow, limiting the extent to which subpopulations can diverge from each other genetically.

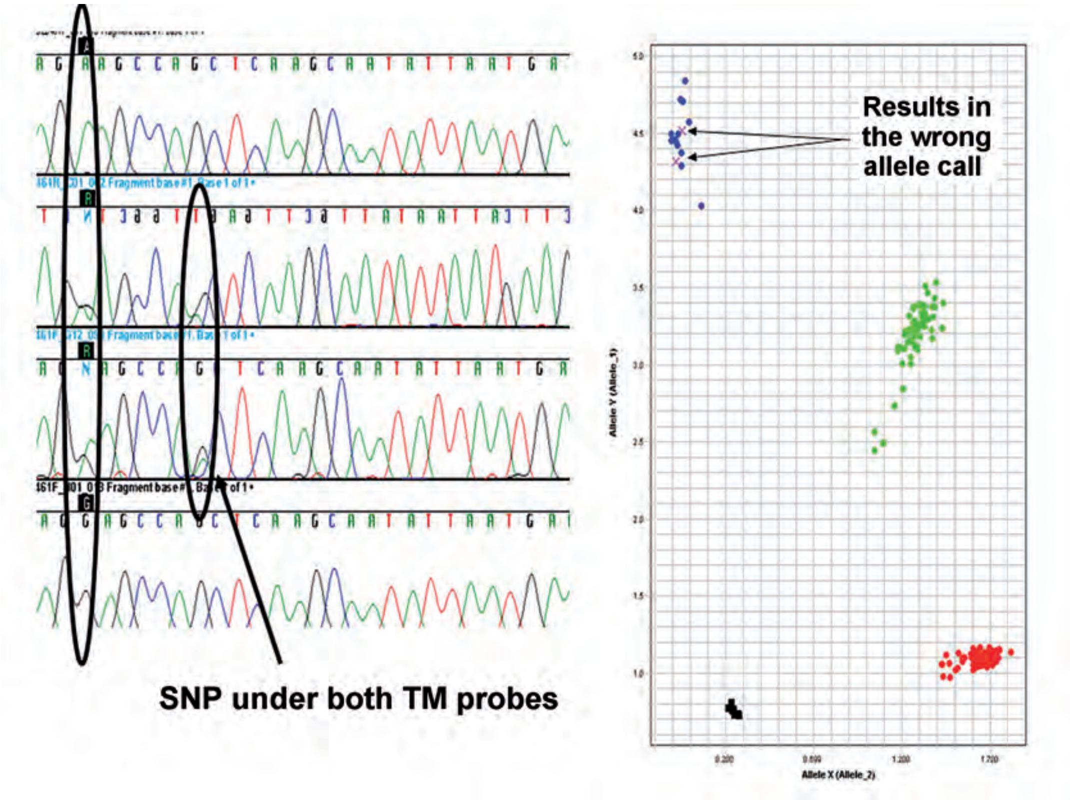
Laboratory analysis of human genetic variation

Genotype analysis

Genotyping is used to interrogate specific, unique loci in the genome following DNA amplification by polymerase chain reaction (PCR). One of the challenges of genotype analysis is that each allele in the genome must be assayed individually, unlike surveys of gene expression that can use a common signature (the polyA tail) to capture a high percentage of mRNA at once. An assay must be robust and reproducible in exceeding a sufficient threshold for detection. Even though amplification protocols are highly reliable, error can be introduced for SNP detection, particularly if there are neighbouring SNPs that alter allele-specific binding of probes or if local genomic sequence is enriched for guanine-cytosine (GC) content (Figure 6.5) (97,98). The presence of duplicates of part of the sequence (CNV), either in the segment amplified or neighbouring the SNPs, can undermine the fidelity of the assay, sometimes providing bias in allele calling (55). Based on the amplification of local sequence surrounding the SNP of interest, redundant sequences are amplified, either locally or elsewhere in the genome, and the fidelity of the polymorphisms between these different segments is undermined, as was observed in the International HapMap Project (45,56).

Initially, restriction fragment length polymorphism (RFLP) assays were used to identify patterns of DNA broken into pieces by restriction enzymes. The size of the fragments was used to develop a footprint of the region of interest (99). RFLP analysis is laborious and error-prone, and thus has been largely abandoned for probe intensity and microchip technologies that can be easily scaled and reliably performed. Examples of these are differential hybridization, primer extension, ligation reactions and allele-specific probe cleavage, all of which interrogate one SNP

Figure 6.5. Fidelity of the genotyping assay: error could be introduced in SNP detection. For example, the presence of a neighboring SNP under both TaqMan® (TM) probes (left panel) may alter allele-specific binding and bias the allele call (right panel).



at a time. Occasionally, RFLPs are required to interrogate a region with high degrees of paralogy.

Low-density genotyping

The most commonly used technique for single SNP assays is the TaqMan® SNP genotyping assay (Applied Biosystems). It is a PCR-based assay designed to interrogate a single SNP that uses two locus-specific PCR primers and two allele-specific labelled probes (100). The 5′ exonuclease property of Taq polymerase is capitalized for detection of base-matching at a specific site. Attached to the 5′ end of each probe is an allele-specific reporter dye: each allele has a corresponding dye, which provides a benchmark for the ratio of the dyes as a reflection of the allele distributions. On the 3′ end of each probe is a single universal quencher dye, which prevents the excitation and emission of the reporter dyes. During PCR amplification, the two PCR primers anneal to the template DNA. The detection probes anneal specifically to the complementary sequence between the forward and reverse primer sites. During the elongation step of each cycle, the Taq polymerase comes in contact from the 5′ end with the reporter dye. Capitalising on the exonuclease property of Taq polymerase, the reporter dye is released from the probe and the fluorescence is released (i.e. no longer quenched by the quencher dye). In addition, the probe itself is digested by the Taq polymerase. After multiple cycles of PCR (that reach saturation for copying both alleles), fluorescence is detected for the two reporter dyes using an ABI 7900HT Sequence Detection System.

Careful attention must be paid to the unique flanking sequences to avoid overlap with adjacent, neighbouring SNPs or insertion/deletions. The throughput is moderate for single-plex TaqMan, but new miniaturization technologies have improved the efficiency of moderate-scale genotyping studies using either the Fluidigm® or BioTrove platforms (101,102).

Multiplexing has increased the technical capacity to interrogate large, predetermined, fixed sets of SNPs. The cost of high-density SNP platforms and the necessity for large-scale follow-up studies have incentivized the development of methodologies for selective replication efforts. The technologies that have been developed for these replication studies are based on direct oligonucleotide hybridization with probe fluorescence detection, the single-base sequencing method, or chip-based mass spectrometry (i.e. based on matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF)) (103). Matrix-assisted laser desorption/ionization (MALDI) enables analysis of biomolecules by ionization usually triggered by a laser beam. A matrix is used to protect the biomolecule from destruction; it can be multiplexed to perform roughly 30 SNP assays at one time.

High-density SNP detection

The first generation of custom bead-array technology by Illumina® enables custom detection of more than 1500 SNPs with excellent performance, and analysis of high-quality DNA generated by whole-genome amplification assays (104,105). This system combines high-multiplexing in a multisample array format, well suited for custom genotype analysis of samples. Though best used with native DNA, it can analyse whole-genome amplified DNA, but at a price of distortions of heterozygosity for roughly 5% of the SNPs.

The newer system of Illumina, known as the Infinium® Assay, features single-tube preparation of DNA followed by whole-genome amplification before genotyping thousands of unique SNPs. Hybridization to bead-bound 50mer oligomers is followed by single-base extension, which incorporates a labelled nucleotide for assay detection. This technology can be used to design custom sets of SNPs (between 7600 and 60 000 bead types) with high efficiency (106). It is the backbone of the fixed content chips, which have increased in size and coverage of the common SNPs in the genome. This began with the HumanHap300 and its complementary HumanHap 240, through to the HumanHap500, Human Hap610 and HumanHap 660w. The Infinium HD (high-density) series followed with the Human1M-Duo BeadChips, which has over 10^6 SNPs to be genotyped, primarily chosen as tagSNPs from HapMap II (8). The increasing content of the chips also provides an opportunity to detect a larger subset of the common CNVs. However, algorithms for detection of CNVs continue to evolve and should improve in the coming years.

The Affymetrix microchip system is based on an assay known as the whole-genome sampling analysis (WGSA) for highly multiplexed SNP genotyping (107). This method amplifies the human genome with a single primer amplification reaction using restriction enzyme-digested, adaptor-ligated human genomic DNA. After fragmentation, sequential labelling and hybridization of the targets is required before analysing the fragments on a microchip. The initial GeneChip® Human Mapping 500K Array spaced SNP markers by physical proximity, but the new Genome-Wide Human SNP Array 6.0 provides a denser set of SNPs (over 900 000), as well as probes that monitor common CNVs across the genome. The distribution of restriction enzyme sites in select regions of the genome does not permit assays across the full genome, limiting the coverage somewhat. The primary debate over the choice of platforms is the coverage of known SNPs in HapMap Stage 2: the SNPs selected for the Illumina platform have been primarily chosen according to the aggressive tag strategy, whereas the first-generation Affymetrix chips provided spaced coverage based on the physical map of the genome, but with higher density. The coverage of the latter has improved.

Methodological issues in GWAS genotyping

High-throughput genotyping facilities require sophisticated robotics for efficient laboratory flow and sample handling, as well as dedicated computational hardware and software able to effectively process both the quantity and complexity of the data. Despite the fact that new technologies and platforms has decreased the nominal price per genotype assayed, the pricing must also take into account the need to study duplicates and samples that must be redone due to technical inadequacies determined in the quality control assessment (see below).

Since replication is a central requirement to protect against the flurry of false-positives observed in GWAS, follow-up studies are needed to verify the results and thus justify the considerable effort required to investigate novel regions. To this end, custom panels are needed to explore regions at the same time that loci are analysed over sufficiently large data sets, so that genome-wide significance can be conclusively established (106,108). Normally, custom panels are more expensive and usually created for a single study (109). In this regard, scalability to meet the requirements of validation studies represents one of the biggest challenges in the design of these studies (110).

Important components of the optimization process include both a Laboratory Information Management System (LIMS) and robotic automation that accurately track and handle samples for efficient workflow management. Because of the high cost of these platforms, the hardware used for sample processing, and the software integrating both, there is little flexibility in choosing individual SNPs to be included within the already-designed, commercially available whole-genome scans.

Two high-density genotyping platforms, Affymetrix and Illumina®, achieve calling capabilities of between 500 000 and 2.5 million SNPs, as well as probe content to interrogate CNVs. Both platforms need between 400–800 ng of total high-quality DNA (usually at 50 ng/µl) for the assay, but because of the dead-space of the robotics (which can be 35% of the required amount for the assay) over 1 ug is required. Issues common to both platforms are the difficulties in assaying SNPs that reside close together (within 60 or fewer nucleotides), which, as previously mentioned, is inherent in this type of genotyping detection. Denser sets of SNPs on commercial platforms have increased coverage, but not always for all populations.

Coverage based on the HapMap II set of SNPs with minor allele frequencies greater than 5%, is one of the main factors driving the choice of platform (8,43). Figure 6.6 illustrates the minimum LD for any SNP assay assessed by the coefficient of correlation, r^2 (a measure of LD), for 2-SNP comparison. The closer the value is to 1, the stronger the correlation, and if the value is estimated to be 1.0, then both loci segregate together. New approaches are being developed to account for the complexity of LD patterns in distinct populations, such as multimarker strategies that have been proposed for analysing more complicated loci (111,112).

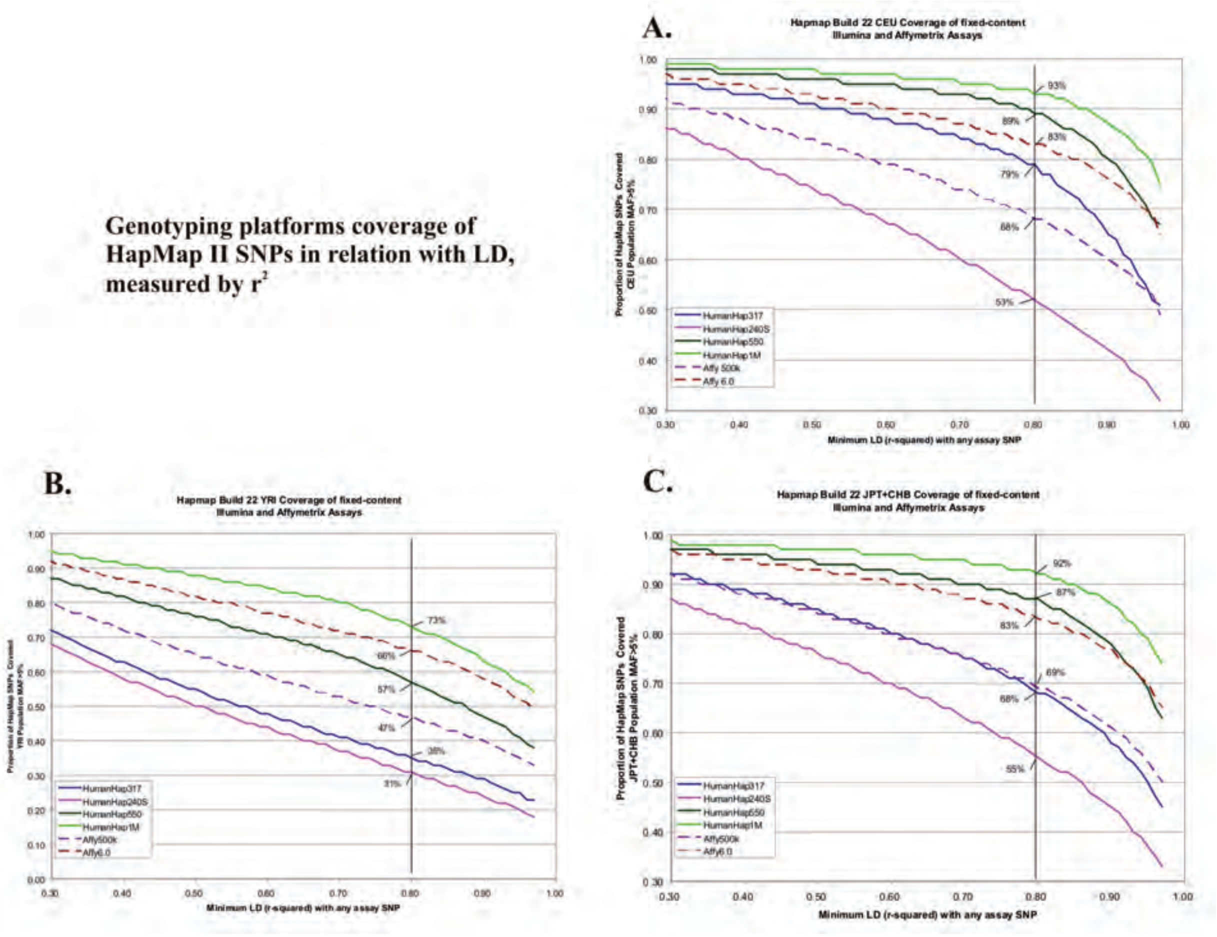
Sequence analysis

Until recently, DNA sequence analysis by capillary electrophoresis has been the platform of choice for medium- and small-scale projects, displacing the Sanger sequencing protocols that used gels or polymers as separation media for the fluorescently labelled DNA fragments (113). The advent of the 96-capillary 3730/3730 xl DNA Analyser (Applied Biosystems) was the central catalyst in the generation of the first draft sequence of the human genome (114).

Dideoxy sequencing is based on the principle of terminating DNA synthesis by incorporation of the dideoxy nucleotide terminator on the complementary strand of DNA fragments. The generated library of various length fragments can be assembled to read the specific DNA sequence. Sequencing-by-synthesis is based upon the principle of pyrophosphate release by nucleotide incorporation along the complementary strand of DNA to the varied-length template. As with dideoxy sequencing, the library of generated fragment lengths can be assembled into a specific DNA sequence. An amplification step by PCR is required, and thus has an intrinsic error below 0.3% (small but predictable) (115).

Efficient removal of unincorporated dye terminators is necessary before running samples on a capillary electrophoresis in

Figure 6.6. Genotyping platforms coverage of HapMap II SNPs. SNP coverage is plotted against LD measured by r2, or coefficient of correlation, for SNP-SNP comparison. Panels: A. HapMap CEU population: CEPH (Utah residents with ancestry from northern and western Europe USAB); B. HapMap YRI population: Yoruba in Ibadan, Nigeria; C. HapMap JPT population: Japanese in Tokyo, Japan, and CHB population: Han Chinese in Beijing, China.



which an electrical field is applied. This allows negatively-charged DNA fragments to move through the polymer towards the positive electrode. Standard software collects raw data files and translates the collected colour data images into consecutive nucleotide base calls.

Next-generation DNA sequencing

Next-generation sequencers have been developed to process millions of sequence reads in parallel rather than in batches of 96 at a time, setting them apart from conventional capillary-based sequencing. These techniques provide high speed and high-throughput from amplified single DNA fragments, avoiding the need for cloning of DNA fragments. Therefore, with minimal input of DNA, the sequencer produces libraries of shorter length reads of between 35–400 bp, depending on the platform, compared to those of capillary sequencers (650–800 bp). A limiting factor is the elevated cost for generating the sequence with high-throughput. There is a need to develop software applications and more efficient computer algorithms to analyse the increasing amount of data generated by these systems (113). Because of their novelty, the accuracy and associated quality of sequencing reads must be further validated, but the high number of reads provides increased coverage of each base position (25). The major challenge of the next-generation sequencing is the informatics of the dense data sets, which requires archiving and storing dense data sets that must be assembled to determine accurate reads. In this regard, error rates for next-generation sequencing runs and assembly constitute a new set of problems, particularly since the quantum increase in data makes their inspection more daunting.

The Roche/454 GS-FLX technology works on the principle of pyrosequencing, which uses pyrophosphate molecules released on nucleotide incorporation by DNA polymerase to fuel a downstream set of reactions that ultimately produces light from the cleavage of oxyluciferin by luciferase (116). The DNA strands of the library are amplified en masse by emulsion PCR (117) on the surfaces of hundreds of thousands of agarose beads. Each agarose bead surface contains up to 1 million copies of the original annealed DNA fragment to produce a detectable signal from the sequencing reaction. Imaging of the light flashes from luciferase activity records which templates are adding that particular nucleotide; the light emitted is directly proportional to the amount of a particular nucleotide incorporated. The current 454 instrument, the GS-FLX, produces an average read length of 400 bp per sample (per bead), with a combined throughput of ~100–150 Mb of sequence data per run. By contrast, a single ABI 3730 programmed to sequence 24 × 96 well plates per day produces ~440 kb of sequence data in 7 hours, with an average read length of 650 bp per sample (25).

The Illumina Genome Analyser is based on the concept of sequencing by synthesis (Solexa® Sequencing technology) to produce sequence reads of 35–150 bp from tens of millions of surface-amplified DNA fragments simultaneously (118). A mixture of single-stranded, adaptor oligo-ligated DNA fragments is incubated and amplified with four differentially-labelled fluorescent nucleotides. Each base incorporation cycle is followed by an imaging step that identifies it and by a chemical step that removes the fluorescent group. At the end of the sequencing run (~4 days), the sequence of each cluster is computed and subjected to quality control. A typical run yields ~40–50 million such sequences.

The Applied Biosystems SOLiD sequencer uses a unique sequencing process catalysed by DNA ligase. A SOLiD (Sequencing by Oligo Ligation and Detection) run requires days, and produces 3–4 Gb of sequence data with an average read length of approximately 50 bp (119). The specific process couples oligo adaptor-linked DNA fragments with 1-µm magnetic beads that are decorated with complementary oligos, and amplifies each bead-DNA complex by emulsion PCR. A SOLiD sequencing by ligation first anneals a universal sequencing primer, then goes through subsequent ligation of the appropriate labelled 8mer, followed by detection at each cycle by fluorescent readout. The unique attribute of this system is that an extra quality check of read accuracy is enabled that facilitates the discrimination of base calling errors from true polymorphisms or insertion/deletion (indel) events, the so-called “2 base encoding” (25).

The third generation of sequencing technologies is in development and should be available in the coming years. For example, single molecule sequencing is based on novel chemistry that enables direct measurement of billions of strands of DNA. The detection system measures incorporated bases on individual strands and thus avoids the requirement of amplification, which is subject to biases and errors.

Applications of high-throughput DNA sequencing

A major focus of this new technology is to rapidly and comprehensively catalogue human genetic variation, particularly common and uncommon genetic polymorphisms (e.g. SNPs and insertion/deletions). Since GWAS have relied on the genotyping of common alleles to discover novel associations with diseases’ risks (120), follow-up of regions of association identified by GWAS is important to characterize common and uncommon variants which might be better markers (or even candidates) for further functional studies. Since GWAS point to new candidate regions, the detailed fine mapping of a region necessitates the generation of a comprehensive set of common and uncommon variants. Already there are select examples of next-generation sequencing analysis applied to regions to determine new variants for follow-up association testing, such as for regions 8q24 associated with prostate and colon cancer and 10q11.2 (containing the *MSMB* gene) associated with prostate cancer (121,122). Eventually the 1000 Genomes Project should provide a suitable map to begin to choose variants in a region of interest. Characterizing all common variants previous to a fine-mapping process has two major benefits: all common genetic variants are represented using a tagSNP approach; and the correlations among all genetic variants are known, which provides advantages in functional variant detection (121,122).

Since large-scale sequencing across the genome is still several years away, attention has focused on targeted sequencing of regions of high interest, such as those defined by GWAS or linkage studies, and, more recently, the opportunity to sequence across the exome (e.g. more than 180 000 known exons in the genome). Several different technologies have been developed to capture target sequence, either through liquid phase (e.g. biotinylated solution capture probes with long range or micro-droplet solution technique), or tiled arrays that contain probes that enrich for capture of DNA for sequence analysis. Each of the next-generation sequencing technologies have been successfully used with one or more target capture technologies. For instance, using the NimbleGen solution-based capture technique, the *KLK3* locus, recently identified as a signal for prostate cancer and prostate serum antigen levels, was resequenced to comprehensively catalogue all common variants for follow-up genotype and functional analyses (123–125). Recently, sequencing across the exome after enrichment with tiled arrays has successfully been used to identify high-penetrant mutations in the coding regions in individuals with Mendelian disorders (126). Exome sequencing represents the first step towards examining the portion of the genome that is easily interpretable, namely changes in coding structure. It requires careful annotation and analytical structures, however, to sift through the thousands of rare variants in unique individuals.

The sequencing of the first human genomes has underscored the challenge of unraveling the physical map, particularly in some regions of great redundancy and/or complexity; moreover, it illustrates the daunting problem of assembly (127,128). More genomes need to be sequenced to establish a reliable reference standard for the analysis of human genomic variations. The current reference is an amalgam of several genomes, thus the ability to unravel variation is particularly difficult. Two new developments should address this issue: sequencing with greater coverage, which diminishes the false-positives and -negatives of sequence determination, and an increase in read length, which will permit phasing of genomes.

The ambitious effort from an international research consortium, namely the 1000 Genomes Project, "…will involve sequencing the genomes of at least a thousand people from around the world to create the most detailed and medically useful picture to date of human genetic variation" (http://www.1000genomes.org/). The goal is to create a detailed map of human genetic variation relevant at or above the level of a frequency of 0.5–1% across the genome (113). By optimizing technology, costs will continue to fall enabling greater scope of study at a lower price. Reduction to affordable levels, targeted for the US$1000 range for an entire human genome sequence, offers the promise of personal genomics. There are still formidable barriers, however, with respect to informatics, storage and the ethical and social dilemmas posed by such analyses.

Next-generation sequencing technologies have already been applied to complementary fields of investigation in genetics. The intent has been to characterize a complex sample with a mixture of nucleic acids through their sequence without prior knowledge of it, in contrast to the probe hybridization used by the original SAGE technique (129,130). Thus, it is possible to characterize the sequence of mRNAs, methylated DNA, DNA or RNA regions bound by certain proteins, and other DNA or RNA regions involved in gene expression and regulation (113). Recent examples are its application to transcriptome profiling in stem cells (119); to whole transcriptome shotgun sequencing, or RNA-Seq, study into alternative splicing in human cells (131); and the identification of mammalian DNA sequences bound by transcription factors *in vivo*, by combining chromatin immunoprecipitation (ChIP) with parallel sequencing (ChIP-Seq) (132).

The Human Microbiome Project (HMP) (http://nihroadmap.nih.gov/hmp/) integrates genomics and metagenomics in an effort to characterize the genome sequences of organisms inhabiting a common environment (133). By understanding genomics, metagenomics and their relations, the HMP seeks to determine whether individuals share a core human microbiome and whether changes in the human microbiome can be correlated with changes in human health (134).

Quality control in the laboratory

The advent of new technologies and workflow paradigms required for high-throughput genotyping and sequencing has changed the nature of laboratory work in genetics. The bulk of the work has been shifted to high-throughput analyses, where so much data is processed in such a short time that the older shibboleths of quality control have been shed for more efficient approaches, which seek to identify potential errors in a high-volume workflow.

The efficient and meticulous sample handling process must begin at the moment of receipt of germline DNA for genotyping or sequencing. Close coordination between the laboratory performing the extraction and the biorepository storing the DNA samples is optimal and protects against handling and biorepository errors, an underappreciated problem.

Standard operating procedures (SOPs) for the process should be created and reviewed regularly for improvements and quality control purposes.

DNA quantification is not an exact science. Due to technical and workflow issues, it is actually quite difficult to reproducibly quantify DNA (135). Several different techniques can be used to measure DNA, but each one has limitations and, in some workflows, different applications of preparing for low- or high-throughput genetic analyses. Quantification methods should be chosen for specific genotype/ sequence platforms. The most commonly used techniques are spectrophotometric measurement of DNA optical density by PicoGreen (Turner BioSystems) analysis, NanoDrop spectrophotometer (NanoDrop Technologies), or by real-time PCR analysis using a standardized TaqMan^{TM} assay (136). Real-time PCR can provide a preliminary test for sample quality as it relates to robust analysis in a high-throughput laboratory, but performance still must be gauged with specific technologies. Spectrophotometry and the PicoGreen assay measure total DNA present, regardless of source or quality, whereas a real-time PCR assay measures the total amplifiable human DNA. DNA quantitation by real-time PCR is particularly helpful for assessing the contribution of non-human DNA to samples collected from buccal swabs, cytobrush samples or other non-blood sources. Minor but real differences between techniques reflect dissimilarities in the ratio of single- and double-stranded DNA, critical for analysis using diverse technologies.

Because of the high volume of activities in high-throughput genotyping/sequencing facilities, unique genetic profiles of samples can be useful for quality assessment and control in the workflow. Many laboratories have incorporated into the upfront analysis a set of SNPs or a forensic panel of 15 small tandem repeats and amelogenin, also known as the AmpFLSTR Identifiler assay (Applied Biosystems). The fingerprinting can be helpful to sleuth problems and identify contaminated samples before costly analysis. Furthermore, the results can serve as a proxy for the viability of the DNA and its success on the high-performance genotyping or sequencing technologies. Certainly, high failure rates indicate poor performance. The profiles can be used to match known duplicates and identify unexpected duplicates, which in turn stimulates close inspection of both biorepository issues and workflow in the laboratory (e.g. errors with plates or reagents).

For the conduct of many molecular epidemiology studies, sample availability has been a limiting factor. Naturally, there has been intense interest in the whole-genome amplification (WGA) technology to provide sufficient amounts of DNA for analysis. Thus, varying results reflect not only differences in the protocols and reagents, but the samples themselves. The quality of DNA used to amplify across the genome affects the success and fidelity of the process. WGA can generate large quantities of DNA for genotype assays, but approximately 5% of the genome is not faithfully reproduced, particularly regions with high GC content or near telomeres. Thus, the results of analyses of these regions should be cautiously interpreted. While the temptation to use WGA DNA in GWAS is great, the results so far have not been encouraging. Currently, there are two approaches that have been commercially optimized. These include a type of multiple displacement amplification (MDA) with the high-performance bacteriophage φ 29 DNA polymerase, which uses degenerate hexamers or generation of libraries of 200–2000 base pair fragments created by random chemical cleavage of genomic DNA. Ligation of adaptor sequences to both ends and PCR amplification is required. Quantities can vary greatly based on input DNA, but under optimal conditions an enrichment of 10 000-fold can be expected.

The rolling circle amplification (RCA) technique is an enzymatic process mediated by DNA polymerases. Long single-stranded DNA molecules are synthesized on a short circular template by using a single DNA primer. RCAs generate a large-scale DNA template with the advantage of not requiring a thermal cycling instrument (137,138). Differential success has been observed with whole blood, dried blood, buccal cell swabs, cultured cells and buffy coat cells. Intriguingly, WGA of water control specimens generates a small, monoallelic signal, which can be called as a single allele, thus underscoring the value of rigorous controls (139). Still, more laboratories have chosen MDA for whole-genome amplification (140).

The utility of duplicates drawn from the same sample remains a central theme of laboratory quality control, but with the advent of high-throughput laboratories the purpose has shifted slightly. Still duplicate testing is useful to detect problems with sample quality, prior storage, and informatic issues in sample management. In some cases, it can also reveal rare individuals enrolled in more than one study. Reproducibility of assays is key, and with the whole-genome-scan chips surpasses 99.8% concordance.

Errors in genotyping, mainly due to loss of one of the heterozygous alleles, occur in well below 1% of samples; therefore, when the rate creeps above 1%, close inspection of the process should be undertaken. If SOPs are followed closely, completion rates should be greater than 95% for most studies, but may be slightly lower depending on the quality of genomic DNA. Completion rates below 90% should raise substantive concern about technical or analytical problems. In GWAS studies, it is recommended that a second technology, such as TaqMan or sequencing, be performed to verify the accuracy and establish concordance (120). Errors with fitness for Hardy–Weinberg proportion (Hardy–Weinberg equilibrium (HWE) testing) can catch major genotype errors, but should probably not be used as a stringent threshold for excluding SNPs from analysis.

Bioinformatics

Large-scale genotyping and sequence analysis has shifted the burden of informatics towards high-performance tools that manage the computational and bioinformatic workflow needed to manipulate high-density data sets. The required tasks, archiving, analysis and access are destined to grow exponentially as studies are designed with increasing numbers of participants and larger and more complex variants to be interrogated. Accordingly, the efficiency of the laboratory flow is based on a high-throughput pipeline for both genetic analysis and informatic handling of the data sets. Major steps in the process include the choice of markers and platforms together with a sophisticated quality control process. Highly trained personnel are needed to effectively coordinate the flow of information. Central to the success of a laboratory is the functioning of a Laboratory Information Management System (LIMS), which is required to track samples, assays, reagents, equipment, robotics and processes through the entire workflow. The LIMS captures the movement of information from receipt of samples through the analytical steps and into the quality control regime required to provide a final, stable data set, linking the results of experimental data to *in silico* information via relational databases. Annotation of the genome is needed to provide clear points of reference for the genomic coordinates for the genotype and sequence assays. Careful quality control and quality assurance checks within the LIMS software, particularly with real-time monitoring, are needed to maintain assay reproducibility and reliable data flow.

The increasing number of loci explored by new platforms, as well as the quantum increases in the increments in study size, has forced major changes in laboratory data storage and management. Laboratory systems should be able to routinely process, monitor and assess quality control of large amounts of data (10^6–10^9 data points) generated by these studies. The increasing need for processing power mandates the use of scalable computational systems capable of parallel computing, with software applications specially designed for this multiprocessor environment and readily upgradable.

Suites of publicly available tools (e.g. PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/summary.shtml) (141) and Genotype Library and Utilities (GLU) (http://cgf.nci.nih.gov/glu/docs)) have been developed for archiving and management of dense data sets, such as those encountered in GWAS. PLINK, now in version 1.06, is a free, open-source whole-genome association suite that focuses on the analysis of large-scale genotype/phenotype data, but lacks support for study design and planning, genotype generation, or CNV calling. Its integration with Haploview (http://www.broadinstitute.org/haploview/haploview) allows visualization, annotation and representation of some of the results. GLU (version 1.0) is also a suite created to manage, analyse and report high-throughput SNP genotype data (http://code.google.com/p/glu-genetics). GLU was created to address the need for new and scalable computational approaches, as well as storage, management, quality control and genetic analysis. It is a framework and software package designed with a set of powerful tools that can scale to effectively handle trillions of genotypes. The integration of GLU with a robust and fast SNP tagging tool, like TagZilla, increases its functionality and allows LD estimation and computation of MAF, HWE, and proportions (http://tagzilla.nci.nih.gov/).

Conclusions and future directions

Knowledge acquired by the draft of the human genome and its annotation, and advances in technology, have changed the approach towards mapping complex diseases and traits. Once oriented to the study of candidate genes and/or mutations, human genetics has evolved into the study of the genome as a complex structure harbouring clues for multifaceted disease risk; some known, but the majority unknown. The discovery of new candidate regions by GWAS has forced rethinking previous strategies for the study of genetic predisposition. More agnostic approaches, genome-wide, with increasing numbers of participants from high-quality epidemiological studies are, for the first time, replicating results in different settings. But new-found candidate regions lead to extensive follow-up and confirmation of their functional significance. Understanding the true effect of genetic variability on the risk of complex diseases is paramount, but also important is the design of high-quality studies to assess environmental contributions, as well as the interactions between genes and exposures.

If accurate measures of environmental factors must be addressed, increased efforts are needed in the study of the biological relevance of the regions already discovered. To date, there are a few examples where biological functional basis has been associated with a candidate region discovered via GWAS. Also, the gap between new-found genomic regions and their biological interpretation could become greater with the introduction of new resequencing technology, which is capable of interrogating more numbers of less frequent loci. New challenges arise with new technologies. High-throughput resequencing must standardise its technical protocols, quality control, calling algorithm and interpretation. Only deep resequencing of high numbers of individuals will create quality databases capable of testing rare variants in the population. Until these steps are readily available for new technologies, broad implementation will not be possible.

The new approach to the genomic study of complex diseases has resulted in a more ambitious "team" science, in which resources and study populations are pooled to identify novel genetic markers (Cf. Figure 6.1). In this regard, GWAS study thousands of the most common genetic variants across the genome (SNPs), without any prior hypothesis, conception or what is being defined as an agnostic manner. This initial phase requires adequately powered follow-up studies for replication that is central to the search for moderate- to high-frequency low-penetrance variants associated with human diseases and traits (120,142). Teams of scientists with specific responsibilities in each step of the process are necessary to ensure quality control and stable analytical results as part of the effort to map complex human diseases and traits.

Previously, family linkage studies have been used to identify rare genetic variants with high-penetrance susceptibility genes (143,144), but failed to be informative on more common genetic variants with low to moderate effect (145). With the advent of next-generation sequencing technologies and the discovery of many rare and uncommon variants, family studies will be required to assist in defining the most notable variants for follow-up studies. In this regard, family studies should prove invaluable in mapping many complex diseases, as well as the highly penetrant Mendelian disorders.

Based on the preliminary data published as a result of GWAS, it is not currently possible to draw final conclusions concerning the valid risk assessment of complex diseases. Education of both the public and scientific media is necessary to affect a rational approach towards implementing any risk reduction policies. These new challenges for public health officials will require careful attention to the ethical, moral and social implications of dense genomic data sets to assure the public, and the participants in the current studies, that confidentiality is protected (26).

Consortial efforts to describe human variation have focused on the description and characterization of three continental populations pursued by the International HapMap Project (http://www.hapmap.org). But using GWAS, other consortia and interest groups have focused on a more disease-specific approach that has resulted in the discovery of over 200 novel loci associated with human diseases/traits (2,8,20,57). Though the majority of the association studies to date have used high-throughput genotyping technology, new programs in comprehensive resequencing analysis would unveil an even greater catalogue of uncommon variants (http://www.1000genomes.org/).

In concert with the assessment of germline genetic variation, other programs are underway to characterize functional annotation through gene expression analysis. The ENCODE Project (ENCyclopedia Of DNA Elements) seeks to define functional elements (http://www.genome.gov/10005107) (25), and The Cancer Genome Atlas (TCGA) examines both somatic and germline alterations in select cancers (146). Together, these new developments promise to accelerate the discovery and characterization of novel genomic mechanisms in human diseases and traits.

References

1. Hunter DJ, Khoury MJ, Drazen JM (2008). Letting the genome out of the bottle–will we get our wish? *N Engl J Med,* 358:105–107. doi:10.1056/NEJMp0708162 PMID:18184955

2. Manolio TA, Brooks LD, Collins FS (2008). A HapMap harvest of insights into the genetics of common disease. *J Clin Invest,* 118:1590–1605.doi:10.1172/JCI34772 PMID:18451988

3. Kruglyak L, Nickerson DA (2001). Variation is the spice of life. *Nat Genet,* 27:234–236. doi:10.1038/85776 PMID:11242096

4. Reich DE, Cargill M, Bolk S *et al.* (2001). Linkage disequilibrium in the human genome. *Nature,* 411:199–204.doi:10.1038/35075590 PMID:11346797

5. Reich DE, Gabriel SB, Altshuler D (2003). Quality and completeness of SNP databases. *Nat Genet,* 33:457–458.doi:10.1038/ng1133 PMID:12652301

6. Lander ES, Linton LM, Birren B *et al.*; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature,* 409:860–921.doi:10.1038/35057062 PMID:11237011

7. Venter JC, Adams MD, Myers EW *et al.* (2001). The sequence of the human genome. *Science,* 291:1304–1351.doi:10.1126/science.1058040 PMID:11181995

8. Frazer KA, Ballinger DG, Cox DR *et al.*; International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature,* 449:851–861. doi:10.1038/nature06258 PMID:17943122

9. Hinds DA, Stuve LL, Nilsen GB *et al.* (2005). Whole-genome patterns of common DNA variation in three human populations. *Science,* 307:1072–1079.doi:10.1126/science.1105436 PMID:15718463

10. Chanock SJ (2001). Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease. *Dis Markers,* 17:89–98. PMID:11673655

11. Risch NJ (2000). Searching for genetic determinants in the new millennium. *Nature,* 405:847–856.doi:10.1038/35015718 PMID:10866211

12. Hughes AL, Packer B, Welch R *et al.* (2003). Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proc Natl Acad Sci USA,* 100:15754–15757.doi:10.1073/pnas.2536718100 PMID:14660790

13. Hughes AL, Packer B, Welch R *et al.* (2005). Effects of natural selection on interpopulation divergence at polymorphic sites in human protein-coding Loci. *Genetics,* 170:1181–1187.doi:10.1534/genetics.104.037077 PMID:15911 586

14. Miklós I, Novák A, Dombai B, Hein J (2008). How reliably can we predict the reliability of protein structure predictions? *BMC Bioinformatics,* 9:137.doi:10.1186/1471-2105-9-137 PMID:18315874

15. Edwards YJ, Cottage A (2003). Bioinformatics methods to predict protein structure and function. A practical approach. *Mol Biotechnol,* 23:139–166.doi:10.1385/MB:23:2:139 PMID:12632698

16. Heringa J (2000). Computational methods for protein secondary structure prediction using multiple sequence alignments. *Curr Protein Pept Sci,* 1:273–301.doi:10.2174/1389203003381324 PMID:12369910

17. Erichsen HC, Chanock SJ (2004). SNPs in cancer research and treatment. *Br J Cancer,* 90:747–751.doi:10.1038/sj.bjc.6601574 PMID:14970847

18. Cargill M, Altshuler D, Ireland J *et al.* (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet,* 22:231–238.doi:10.1038/10290 PMID:10391209

19. Stephens JC, Schneider JA, Tanguay DA *et al.* (2001). Haplotype variation and linkage disequilibrium in 313 human genes. *Science,* 293:489–493.doi:10.1126/science.1059431 PMID:11452081

20. International HapMap Consortium (2003). The International HapMap Project. *Nature,* 426:789–796. doi:10.1038/nature02168 PMID:14685227

21. Packer BR, Yeager M, Burdett L *et al.* (2006). SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res,* 34 Database issue;D617–D621.doi:10.1093/nar/gkj151 PMID:16381944

22. Stephens M, Sloan JS, Robertson PD *et al.* (2006). Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat Genet,* 38:375–381.doi:10.1038/ng1746 PMID:16493422

23. Marth G, Schuler G, Yeh R *et al.* (2003). Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc Natl Acad Sci USA,* 100:376–381.doi:10.1073/pnas.222673099 PMID:12502794

24. Marth GT, Korf I, Yandell MD *et al.* (1999). A general approach to single-nucleotide polymorphism discovery. *Nat Genet,* 23:452–456.doi:10.1038/70570 PMID:10581034

25. Mardis ER (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet,* 24:133–141.doi:10.1016/j.tig.2007.12.007 PMID:18262675

26. Birney E, Stamatoyannopoulos JA, Dutta A *et al.*; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature,* 447:799–816.doi:10.1038/nature05874 PMID:17571346

27. Romeo S, Pennacchio LA, Fu Y *et al.* (2007). Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat Genet,* 39:513–516.doi:10.1038/ng1984 PMID:17322881

28. Bonnen PE, Wang PJ, Kimmel M *et al.* (2002). Haplotype and linkage disequilibrium architecture for human cancer-associated genes. *Genome Res,* 12:1846–1853.doi:10.1101/gr.483802 PMID:12466288

29. Sabeti PC, Reich DE, Higgins JM *et al.* (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature,* 419:832–837.doi:10.1038/nature01140 PMID:12397357

30. Slatkin M (2008). Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat Rev Genet,* 9:477–485.doi:10.1038/nrg2361 PMID:18427557

31. Orr N, Chanock SJ (2008). Common genetic variation and human disease. *Adv Genet,* 62:1–32.doi:10.1016/S0065-2660(08)00601-9 PMID:19010252

32. Hill WG (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity,* 33:229–239.doi:10.1038/hdy.1974.89 PMID:4531429

33. Clark AG (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol,* 7:111–122. PMID:2108305

34. Eskin E, Halperin E, Karp RM (2003). Efficient reconstruction of haplotype structure via perfect phylogeny. *J Bioinform Comput Biol,* 1:1–20.doi:10.1142/S0219720003000174 PMID:15290779

35. Stephens M, Smith NJ, Donnelly P (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet,* 68:978–989.doi:10.1086/319501 PMID:11254454

36. Stephens M, Donnelly P (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet,* 73:1162–1169.doi:10.1086/379378 PMID:14574645

37. Marchini J, Cutler D, Patterson N *et al.*; International HapMap Consortium (2006). A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet,* 78:437–450.doi:10.1086/500808 PMID:16465620

38. Akey J, Jin L, Xiong M (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet,* 9:291–300. doi:10.1038/sj.ejhg.5200619 PMID:11313774

39. Schaid DJ (2004). Evaluating associations of haplotypes with traits. *Genet Epidemiol,* 27:348–364.doi:10.1002/gepi.20037 PMID:15543638

40. Tan Q, Christiansen L, Christensen K *et al.* (2005). Haplotype association analysis of human disease traits using genotype data of unrelated individuals. *Genet Res,* 86:223–231.doi:10.1017/S0016672305007792 PMID:16454861

41. Cardon LR, Abecasis GR (2003). Using haplotype blocks to map human complex trait loci. *Trends Genet,* 19:135–140.doi:10.1016/S0168-9525(03)00022-2 PMID:12615007

42. Johnson GC, Esposito L, Barratt BJ *et al.* (2001). Haplotype tagging for the identification of common disease genes. *Nat Genet,* 29:233–237.doi:10.1038/ng1001-233 PMID:11586306

43. Barrett JC, Fry B, Maller J, Daly MJ (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21:263–265.doi:10.1093/bioinformatics/bth457 PMID:15297300

44. Stram DO, Haiman CA, Hirschhorn JN *et al.* (2003). Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum Hered,* 55:27–36.doi:10.1159/000071807 PMID:12890923

45. McCarroll SA, Altshuler DM (2007). Copy-number variation and association studies of human disease. *Nat Genet,* 39 Suppl;S37–S42.doi:10.1038/ng2080 PMID:17597780

46. Scherer SW, Lee C, Birney E *et al.* (2007). Challenges and standards in integrating surveys of structural variation. *Nat Genet,* 39 Suppl;S7–S15.doi:10.1038/ng2093 PMID:17597783

47. Kidd JM, Cooper GM, Donahue WF *et al.* (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453:56–64.doi:10.1038/nature06862 PMID:18451855

48. Feuk L, Carson AR, Scherer SW (2006). Structural variation in the human genome. *Nat Rev Genet,* 7:85–97.doi:10.1038/nrg1767 PMID:16418744

49. Stefansson H, Helgason A, Thorleifsson G *et al.* (2005). A common inversion under selection in Europeans. *Nat Genet,* 37:129–137.doi:10.1038/ng1508 PMID:15654335

50. Sharp AJ, Locke DP, McGrath SD *et al.* (2005). Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet,* 77:78–88.doi:10.1086/431652 PMID:15918152

51. Sebat J, Lakshmi B, Troge J *et al.* (2004). Large-scale copy number polymorphism in the human genome. *Science,* 305:525–528. doi:10.1126/science.1098918 PMID:15273396

52. Iafrate AJ, Feuk L, Rivera MN *et al.* (2004). Detection of large-scale variation in the human genome. *Nat Genet,* 36:949–951.doi:10.1038/ng1416 PMID:15286789

53. Inoue K, Lupski JR (2002). Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet,* 3:199–242.doi:10.1146/annurev.genom.3.032802.120023 PMID:12142364

54. Bailey JA, Yavor AM, Massa HF *et al.* (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res,* 11:1005–1017. doi:10.1101/gr.GR-1871R PMID:11381028

55. Bailey JA, Gu Z, Clark RA *et al.* (2002). Recent segmental duplications in the human genome. *Science*, 297:1003–1007.doi:10.1126/science.1072047 PMID:12169732

56. Freeman JL, Perry GH, Feuk L *et al.* (2006). Copy number variation: new insights in genome diversity. *Genome Res,* 16:949–961. doi:10.1101/gr.3677206 PMID:16809666

57. International HapMap Consortium (2005). A haplotype map of the human genome. *Nature,* 437:1299–1320.doi:10.1038/nature04226 PMID:16255080

58. Buckley PG, Mantripragada KK, Piotrowski A *et al.* (2005). Copy-number polymorphisms: mining the tip of an iceberg. *Trends Genet,* 21:315–317.doi:10.1016/j.tig.2005.04.007 PMID:15922827

59. McCarroll SA, Kuruvilla FG, Korn JM *et al.* (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet,* 40:1166–1174. doi:10.1038/ng.238 PMID:18776908

60. Khaja R, Zhang J, MacDonald JR *et al.* (2006). Genome assembly comparison identifies structural variants in the human genome. *Nat Genet,* 38:1413–1418.doi:10.1038/ng1921 PMID:17115057

61. Istrail S, Sutton GG, Florea L *et al.* (2004). Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc Natl Acad Sci USA,* 101:1916–1921. doi:10.1073/pnas.0307971100 PMID:14769938

62. Cooper GM, Zerr T, Kidd JM *et al.* (2008). Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet,* 40:1199–1203.doi:10.1038/ng.236 PMID:18776910

63. Korn JM, Kuruvilla FG, McCarroll SA *et al.* (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet,* 40:1253–1260.doi:10.1038/ng.237 PMID:18776909

64. Barnes C, Plagnol V, Fitzgerald T *et al.* (2008). A robust statistical method for case-control association testing with copy number variation. *Nat Genet,* 40:1245–1252. doi:10.1038/ng.206 PMID:18776912

65. Ballantyne KN, van Oorschot RAH, Mitchell RJ (2007). Comparison of two whole genome amplification methods for STR genotyping of LCN and degraded DNA samples. *Forensic Sci Int,* 166:35–41.doi:10.1016/j.forsciint.2006.03.022 PMID:16687226

66. Goellner GM, Tester D, Thibodeau S *et al.* (1997). Different mechanisms underlie DNA instability in Huntington disease and colorectal cancer. *Am J Hum Genet*, 60:879–890. PMID:9106534

67. Roewer L, Krawczak M, Willuweit S *et al.* (2001). Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Sci Int,* 118:106–113. doi:10.1016/S0379-0738(00)00478-3 PMID:11311820

68. Ryckman K, Williams SM. Calculation and use of the Hardy-Weinberg model in association studies. *Curr Protoc Hum Genet* 2008;Chapter 1:Unit 1.18.

69. Hosking L, Lumsden S, Lewis K *et al.* (2004). Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *Eur J Hum Genet,* 12:395–399.doi:10.1038/sj.ejhg.5201164 PMID:14872201

70. Gomes I, Collins A, Lonjou C *et al.* (1999). Hardy-Weinberg quality control. *Ann Hum Genet,* 63:535–538.doi:10.1046/j.1469-1809.1999.6360535.x PMID:11246455

71. Akey JM, Zhang K, Xiong M *et al.* (2001). The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *Am J Hum Genet,* 68:1447–1456. doi:10.1086/320607 PMID:11359212

72. Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005). Rational inferences about departures from Hardy-Weinberg equilibrium. *Am J Hum Genet,* 76:967–986.doi:10.1086/430507 PMID:15834813

73. Leal SM (2005). Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. *Genet Epidemiol,* 29:204–214.doi:10.1002/gepi.20086 PMID:16080207

74. Cox DG, Kraft P (2006). Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. *Hum Hered,* 61:10–14. doi:10.1159/000091787 PMID:16514241

75. Hirschhorn JN (2005). Genetic approaches to studying common diseases and complex traits. *Pediatr Res,* 57:74R–77R.doi:10.1203/01.PDR.0000159574.98964.87 PMID:15817501

76. Yu K, Wang Z, Li Q *et al.* (2008). Population substructure and control selection in genome-wide association studies. *PLoS One*, 3:e2551.doi:10.1371/journal.pone.0002551 PMID:18596976

77. Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587. PMID:12930761

78. Devlin B, Roeder K (1999). Genomic control for association studies. *Biometrics,* 55:997–1004.doi:10.1111/j.0006-341X.1999.00997.x PMID:11315092

79. Pritchard JK, Rosenberg NA (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet*, 65:220–228.doi:10.1086/302449 PMID:10364535

80. Freedman ML, Haiman CA, Patterson N *et al.* (2006). Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc Natl Acad Sci USA*, 103:14068–14073.doi:10.1073/pnas.0605832103 PMID:16945910

81. Kao WH, Klag MJ, Meoni LA *et al.*; Family Investigation of Nephropathy and Diabetes Research Group (2008). MYH9 is associated with nondiabetic end-stage renal disease in African Americans. *Nat Genet,* 40:1185–1192. doi:10.1038/ng.232 PMID:18794854

82. Kopp JB, Smith MW, Nelson GW *et al.* (2008). MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. *Nat Genet,* 40:1175–1184.doi:10.1038/ng.226 PMID:18794856

83. Hurst LD (2009). Fundamental concepts in genetics: genetics and the understanding of selection. *Nat Rev Genet,* 10:83–93. doi:10.1038/nrg2506 PMID:19119264

84. Stern DL, Orgogozo V (2009). Is genetic evolution predictable? *Science*, 323:746–751. doi:10.1126/science.1158997 PMID:19197055

85. Stern DL, Orgogozo V (2008). The loci of evolution: how predictable is genetic evolution? *Evolution*, 62:2155–2177. doi:10.1111/j.1558-5646.2008.00450.x PMID: 18616572

86. Cooper TF, Ostrowski EA, Travisano M (2007). A negative relationship between mutation pleiotropy and fitness effect in yeast. *Evolution*, 61:1495–1499.doi:10.1111/j.1558-5646.2007.00109.x PMID:17542856

87. Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D (2002). Quantifying the slightly deleterious mutation model of molecular evolution. *Mol Biol Evol,* 19:2142–2149. PMID:12446806

88. Charlesworth B (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet,* 10:195–205. doi:10.1038/nrg2526 PMID:19204717

89. Relethford JH (2004). Global patterns of isolation by distance based on genetic and morphological data. *Hum Biol,* 76:499–513. doi:10.1353/hub.2004.0060 PMID:15754968

90. Ramachandran S, Deshpande O, Roseman CC *et al.* (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA,* 102:15942–15947.doi:10.1073/pnas.0507611102 PMID:16243969

91. Li JZ, Absher DM, Tang H *et al.* (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319:1100–1104.doi:10.1126/science.1153717 PMID:18292342

92. Rosenberg NA, Pritchard JK, Weber JL *et al.* (2002). Genetic structure of human populations. *Science*, 298:2381–2385.doi: 10.1126/science.1078311 PMID:12493913

93. Romero IG, Manica A, Goudet J *et al.* (2009). How accurate is the current picture of human genetic variation? *Heredity*, 102:120–126.doi:10.1038/hdy.2008.89 PMID:18766200

94. Sabeti PC, Schaffner SF, Fry B *et al.* (2006). Positive natural selection in the human lineage. *Science*, 312:1614–1620.doi:10.1126/science.1124309 PMID:16778047

95. Sabeti PC, Varilly P, Fry B *et al.*; International HapMap Consortium (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449:913–918.doi:10.1038/nature06250 PMID:17943131

96. Tishkoff SA, Reed FA, Ranciaro A *et al.* (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet,* 39:31–40.doi:10.1038/ng1946 PMID: 17159977

97. Pompanon F, Bonin A, Bellemain E, Taberlet P (2005). Genotyping errors: causes, consequences and solutions. *Nat Rev Genet,* 6:847–859.doi:10.1038/nrg1707 PMID:16304600

98. Packer BR, Yeager M, Staats B *et al.* (2004). SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes. *Nucleic Acids Res,* 32 Database issue;D528–D532.doi:10.1093/nar/gkh005 PMID:14681474

99. Saiki RK, Scharf S, Faloona F *et al.* (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230:1350–1354.doi:10.1126/science.2999980 PMID:2999980

100. Livak KJ, Marmaro J, Todd JA (1995). Towards fully automated genome-wide polymorphism screening. *Nat Genet*, 9:341–342.doi:10.1038/ng0495-341 PMID:7795635

101. Brenan CJ (2002). DNA-based molecular lithography for nanoscale fabrication. *IEEE Eng Med Biol Mag,* 21:164.doi:10.1109/MEMB.2002.1175178 PMID:12613226

102. Frederickson RM (2002). Fluidigm. Biochips get indoor plumbing. *Chem Biol,* 9:1161–1162. PMID:12445764

103. Sun X, Ding H, Hung K, Guo B (2000). A new MALDI-TOF based mini-sequencing assay for genotyping of SNPS. *Nucleic Acids Res*, 28:E68.doi:10.1093/nar/28.12.e68 PMID:10871391

104. Cunningham JM, Sellers TA, Schildkraut JM *et al.* (2008). Performance of amplified DNA in an Illumina GoldenGate BeadArray assay. *Cancer Epidemiol Biomarkers Prev,* 17:1781–1789.doi:10.1158/1055-9965.EPI-07-2849 PMID:18628432

105. Berthier-Schaad Y, Kao WH, Coresh J *et al.* (2007). Reliability of high-throughput genotyping of whole genome amplified DNA in SNP genotyping studies. *Electrophoresis*, 28:2812–2817.doi:10.1002/elps.200600674 PMID:17702060

106. Thomas G, Jacobs KB, Yeager M *et al.* (2008). Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet,* 40:310–315.doi:10.1038/ng.91 PMID:18264096

107. Matsuzaki H, Loi H, Dong S *et al.* (2004). Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res,* 14:414–425.doi:10.1101/gr.2014904 PMID:14993208

108. Al Olama AA, Kote-Jarai Z, Giles GG *et al.*; UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology; UK Prostate testing for cancer and Treatment study (ProtecT Study) Collaborators (2009). Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet,* 41:1058–1060.doi:10.1038/ng.452 PMID:19767752

109. Barrett JC, Cardon LR (2006). Evaluating coverage of genome-wide association studies. *Nat Genet,* 38:659–662.doi:10.1038/ng1801 PMID:16715099

110. McCarthy MI, Abecasis GR, Cardon LR *et al.* (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet,* 9:356–369. doi:10.1038/nrg2344 PMID:18398418

111. Kim Y, Feng S, Zeng ZB (2008). Measuring and partitioning the high-order linkage disequilibrium by multiple order Markov chains. *Genet Epidemiol,* 32:301–312.doi:10.1002/gepi.20305 PMID:18330903

112. Schaid DJ (2004). Genetic epidemiology and haplotypes. *Genet Epidemiol,* 27:317–320.doi:10.1002/gepi.20046 PMID:15543637

113. Ansorge WJ (2009). Next-generation DNA sequencing techniques. *N Biotechnol,* 25:195–203.doi:10.1016/j.nbt.2008.12.009 PMID:19429539

114. Collins FS, Morgan M, Patrinos A (2003). The Human Genome Project: lessons from large-scale biology. *Science*, 300:286–290. doi:10.1126/science.1084564 PMID:12690187

115. Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA,* 74:5463–5467.doi:10.1073/pnas.74.12.5463 PMID:271968

116. Margulies M, Egholm M, Altman WE *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380. PMID:16056220

117. Dressman D, Yan H, Traverso G *et al.* (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci USA,* 100:8817–8822.doi:10.1073/pnas.1133470100 PMID:12857956

118. Van Tassell CP, Smith TPL, Matukumalli LK *et al.* (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods,* 5:247–252.doi:10.1038/nmeth.1185 PMID:18297082

119. Cloonan N, Forrest ARR, Kolle G *et al.* (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods,* 5:613–619.doi:10.1038/nmeth.1223 PMID:18516046

120. Chanock SJ, Manolio T, Boehnke M *et al.*; NCI-NHGRI Working Group on Replication in Association Studies (2007). Replicating genotype-phenotype associations. *Nature*, 447:655–660.doi:10.1038/447655a PMID:17554299

121. Yeager M, Deng Z, Boland J *et al.* (2009). Comprehensive resequence analysis of a 97 kb region of chromosome 10q11.2 containing the MSMB gene associated with prostate cancer. *Hum Genet*, 126:743–750.doi:10.1007/s00439-009-0723-9 PMID:19644707122.

122. Yeager M, Xiao N, Hayes RB *et al.* (2008). Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. *Hum Genet,* 124:161–170.doi:10.1007/s00439-008-0535-3 PMID:18704501

123. Parikh H, Deng Z, Yeager M *et al.* (2010). A comprehensive resequence analysis of the KLK15-KLK3-KLK2 locus on chromosome 19q13.33. *Hum Genet,* 127:91–99.doi:10.1007/s00439-009-0751-5 PMID:19823874

124. Ahn J, Berndt SI, Wacholder S *et al.* (2008). Variation in KLK genes, prostate-specific antigen and risk of prostate cancer. *Nat Genet,* 40:1032–1034, author reply 1035–1036.doi:10.1038/ng0908-1032 PMID:19165914

125. Eeles RA, Kote-Jarai Z, Giles GG *et al.*; UK Genetic Prostate Cancer Study Collaborators; British Association of Urological Surgeons' Section of Oncology; UK ProtecT Study Collaborators (2008). Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet*, 40:316–321.doi:10.1038/ng.90 PMID:18264097

126. Ng SB, Turner EH, Robertson PD *et al.* (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461:272–276.doi:10.1038/nature08250 PMID:19684571

127. Levy S, Sutton G, Ng PC *et al.* (2007). The diploid genome sequence of an individual human. *PLoS Biol,* 5:e254.doi:10.1371/journal.pbio.0050254 PMID:17803354

128. Wheeler DA, Srinivasan M, Egholm M *et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452:872–876.doi:10.1038/nature06884 PMID:18421352

129. Mortazavi A, Williams BA, McCue K *et al.* (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5:621–628.doi:10.1038/nmeth.1226 PMID:18516045

130. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995). Serial analysis of gene expression. *Science*, 270:484–487.doi:10.1126/science.270.5235.484 PMID:7570003

131. Sultan M, Schulz MH, Richard H *et al.* (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321:956–960.doi:10.1126/science.1160342 PMID:18599741

132. Robertson G, Hirst M, Bainbridge M *et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods,* 4:651–657.doi:10.1038/nmeth1068 PMID:17558387

133. Hugenholtz P, Tyson GW (2008). Microbiology: metagenomics. *Nature*, 455:481–483.doi:10.1038/455481a PMID:18818648

134. Turnbaugh PJ, Ley RE, Hamady M *et al.* (2007). The human microbiome project. *Nature*, 449:804–810.doi:10.1038/nature06244 PMID:17943116

135. Bergen AW, Qi Y, Haque KA *et al.* (2005). Effects of DNA mass on multiple displacement whole genome amplification and genotyping performance. *BMC Biotechnol*, 5:24.doi:10.1186/1472-6750-5-24 PMID:16168060

136. Haque KA, Pfeiffer RM, Beerman MB *et al.* (2003). Performance of high-throughput DNA quantification methods. *BMC Biotechnol,* 3:20.doi:10.1186/1472-6750-3-20 PMID:14583097

137. Fire A, Xu SQ (1995). Rolling replication of short DNA circles. *Proc Natl Acad Sci USA*, 92:4641–4645.doi:10.1073/pnas.92.10.4641 PMID:7753856

138. Dean FB, Nelson JR, Giesler TL, Lasken RS (2001). Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res,* 11:1095–1099.doi:10.1101/gr.180501 PMID:11381035

139. Bergen AW, Haque KA, Qi Y *et al.* (2005). Comparison of yield and genotyping performance of multiple displacement amplification and OmniPlex whole genome amplified DNA generated from multiple DNA sources. *Hum Mutat,* 26:262–270.doi:10.1002/humu.20213 PMID:16086324

140. Lasken RS (2009). Genomic DNA amplification by the multiple displacement amplification (MDA) method. *Biochem Soc Trans,* 37:450–453.doi:10.1042/BST0370450 PMID:19290880

141. Purcell S, Neale B, Todd-Brown K *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81:559–575.doi:10.1086/519795 PMID:17701901

142. Hunter DJ, Thomas G, Hoover RN, Chanock SJ (2007). Scanning the horizon: what is the future of genome-wide association studies in accelerating discoveries in cancer etiology and prevention? *Cancer Causes Control,* 18:479–484.doi:10.1007/s10552-007-0118-y PMID:17440825

143. Hall JM, Lee MK, Newman B *et al.* (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250:1684–1689.doi:10.1126/science.2270482 PMID:2270482

144. Wooster R, Bignell G, Lancaster J *et al.* (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature*, 378:789–792.doi:10.1038/378789a0 PMID:8524414

145. Stratton MR, Rahman N (2008). The emerging landscape of breast cancer susceptibility. *Nat Genet,* 40:17–22.doi:10.1038/ng.2007.53 PMID:18163131

146. McLendon R, Friedman A, Bigner D *et al.*; Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061–1068.doi:10.1038/nature07385 PMID:18772890

147. Gabriel SB, Schaffner SF, Nguyen H *et al.* (2002). The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229.doi:10.1126/science.1069424 PMID:12029063

148. Carlson CS, Eberle MA, Rieder MJ *et al.* (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet,* 74:106–120.doi:10.1086/381000 PMID:14681826