Send Orders for Reprints to reprints@benthamscience.net

Current Psychiatry Reviews, 2014, *10*, 91-100

Jumping on the Train of Personalized Medicine: A Primer for Non-Geneticist Clinicians: Part 1. Fundamental Concepts in Molecular Genetics

Aihua Li and David Meyre*

Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON L8N 3Z5, Canada

Abstract: With the decrease in sequencing cost and the rise of companies providing sequencing services, it is likely that personalized whole-genome sequencing will eventually become an instrument of common medical practice. We write this series of three reviews to help non-geneticist clinicians get ready for the major breakthroughs that are likely to occur in the coming years in the fast-moving field of personalized medicine. This first paper focuses on the fundamental concepts of molecular genetics. We review how recombination occurs during meiosis, how *de novo* genetic variations including single nucleotide polymorphisms (SNPs), insertions and deletions are generated and how they are inherited from one generation to the next. We detail how genetic variants can impact protein expression and function, and summarize the main characteristics of the human genome. We also explain how the achievements of the Human Genome Project, the HapMap Project, and more recently, the 1000 Genomes Project, have boosted the identification of genetic variants contributing to common diseases in human populations. The second and third papers will focus on genetic epidemiology and clinical applications in personalized medicine.

Keywords: Chromosome, deoxyribonucleic acid, genetic variants, haplotype, human genome, linkage disequilibrium.

INTRODUCTION

Most human diseases have a genetic component. Non-genetic clinicians are familiar with single-gene disorders for the simple reason that the medical training, including human genetic courses, mainly refers to Mendelian diseases. An example of single-gene disorder is Huntington disease which is caused by a single mutation in the *HD* gene and that follows the easily recognized pattern of autosomal dominant inheritance across generations [1]. Some clinicians are, however, less comfortable with the principles of genetic contributions to complex disorders, despite the fact that a majority of human diseases (e.g. diabetes, cardiovascular diseases, cancers and psychiatric disorders) fall into this category. Complex diseases are triggered by multiple genetic variants in multiple genes acting in combination with environmental factors, and they typically do not follow any Mendelian patterns of inheritance. This limited knowledge in the medical community is understandable as genetic determinants for complex diseases were uncovered in the last 15 years and new discoveries are ongoing. Two important breakthroughs have revolutionized the search for genetic variants contributing to complex diseases and have boosted the elucidation of complex traits in the last seven years. First, the commercialization of high throughput genotyping microarrays has led to the emergence of genome-wide association studies (GWAS) and to an unparalleled harvest of disease-associated loci [2, 3]. Since the first report of GWAS in 2005, more than 2000 loci have been conclusively associated with one or more complex traits [4, 5]. However, most genetic variants from GWAS can only be correlated with a disease and the underlying mechanism may not be known. Over the past three years, the advent of high-throughput next generation sequencing platforms has led to the availability of whole-exome sequencing experiments which specifically sequence the subset of the human genome that code proteins, and to tremendous progress in the elucidation of Mendelian and complex disorders [6, 7]. With the decrease in sequencing cost and growing patient willingness to participate [8], personalized whole-genome sequencing may eventually become an instrument of common medical practice [9, 10]. These new perspectives challenge the clinicians to jump into the fast-moving field of personalized medicine, an emerging practice that uses an individual's genetic profile to guide decision-making in regard to the prevention, diagnosis, and treatment of diseases [11]. Despite all the recent 'buzz' around personalized medicine, the potential benefits of genetics in clinical practice are regarded with a certain degree of skepticism by the majority of clinicians [12, 13]. Obvious reasons include ethical concerns about privacy and discrimination or negative consequences of genetic testing for the patients (worry and anxiety) [14]. A less acknowledged but important reason is that genetics is regarded as a hermetic scientific field. The fact that geneticists use a highly technical language with terms like GWAS, single nucleotide polymorphism (SNP), and haplotype, certainly does not help. Ignorance begets fear and clinicians lacking the scientific background in genetic epidemiology may be prone to mistrust or scorn the promises and potential applications of genetic discoveries in their fields. Taking that into account, now is the time for

*Address correspondence to this author at the McMaster University, Michael DeGroote Centre for Learning & Discovery, Room 3205, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada; Tel: 905-525-9140 Ext. 26802; Fax: 905-528-2814; Email: meyred@mcmaster.ca

1875-6441/14 $58.00+.00

clinicians to become more familiar with the key concepts of genetic epidemiology in order to become active participants of the personalized medicine revolution. We intend to write this series of three reviews to help non-geneticist clinicians prepare for the major genetic breakthroughs that will to occur in the coming years, and to welcome genomic medicine into their spheres of practice with the hope of achieving better prevention and care of human genetic disorders. In this Part I of the series of two reviews, we will detail the basic concepts of molecular genetics in user-friendly language. In the next two review we will then discuss the study designs and statistical procedures classically used in genetic epidemiology (in Part II) (and the realistic promises and challenges in application of recent genetic discoveries in medicine (in Part III)).

DNA, RNA AND PROTEINS

It has been known since immemorial time that offspring inherit, to a certain extent, their appearance, characteristics, and personality from their parents. As early as 1920s, a chemical substance called DeoxyriboNucleic Acid (DNA) was identified to carry the genetic information and transmit these characteristics from one generation to another [15]. In 1953, James Watson and Francis Crick described the double helix model of DNA structure which is the fundamental discovery for the central dogma of molecular biology [16]. Nucleotides are the basic units of the complex DNA molecule. They contain three parts of a five-carbon sugar, a phosphate molecule and a nitrogen-containing base, which is either adenine (A), thymine (T), cytosine (C) or guanine (G). Nucleotides polymerize into long chains by phosphodiester bonds (Fig. **1a**). Because phosphodiester bonds link the 3' carbon atom of one sugar to the 5' carbon atom of the next sugar, the 5' end has a terminal sugar residue in which the 5' carbon atom is free and the 3' end has a terminal sugar residue in which the 3' carbon atom is free. Cellular DNA forms a double stranded helix (Fig. **1b** and **1c**). The two coiled long polynucleotide chains are in an antiparallel formation in which the sugar-phosphate backbones are on the outside of the double helix, and the nitrogenous bases are on the inside and perpendicular to the backbones. Therefore, one strand runs in the direction of 5' to 3', whereas the other runs from 3' to 5'. The hydrogen bonds between pairs of bases join the two strands following a specific and base-pairing rule: A to T only and C to G only. Although most of a cell's DNA in humans is contained in the nucleus, mitochondria have their own independent genome that bears a strong resemblance to bacterial genomes [17]. Every cell in the human body has a complete set of DNA called a genome with the exception of mature red blood cells (erythrocytes), which lack a nucleus and most organelles. A gene is a segment of DNA along the genome encompassing specific regulatory elements (5'-untranslated regions (5'-UTRs) and 3'-untranslated regions (3'-UTRs)), non-coding regions (introns) and coding-regions (exons) which give instructions to messenger ribonucleic acid (mRNA) in the form of three base-pair sets called codons that assemble amino acids to a functional protein (Fig. **2**). Therefore, a gene is considered to be the basic unit of heredity. During cell growth and division, DNA replication initiates when two DNA strands unwind at a specific origin and serve as their own templates and synthesize the second copy of each DNA strand with the assistance of DNA polymerase and other enzymes. DNA self-replication is conducted in an extremely accurate manner (less than 1 mismatched nucleotide in 10^7) [18, 19]. Once an error occurs, a repair system including DNA polymerases, exonuclease and other enzymes will proofread DNA sequence and excise the incorrect base pair, ensuring the stability and high fidelity of DNA within an individual

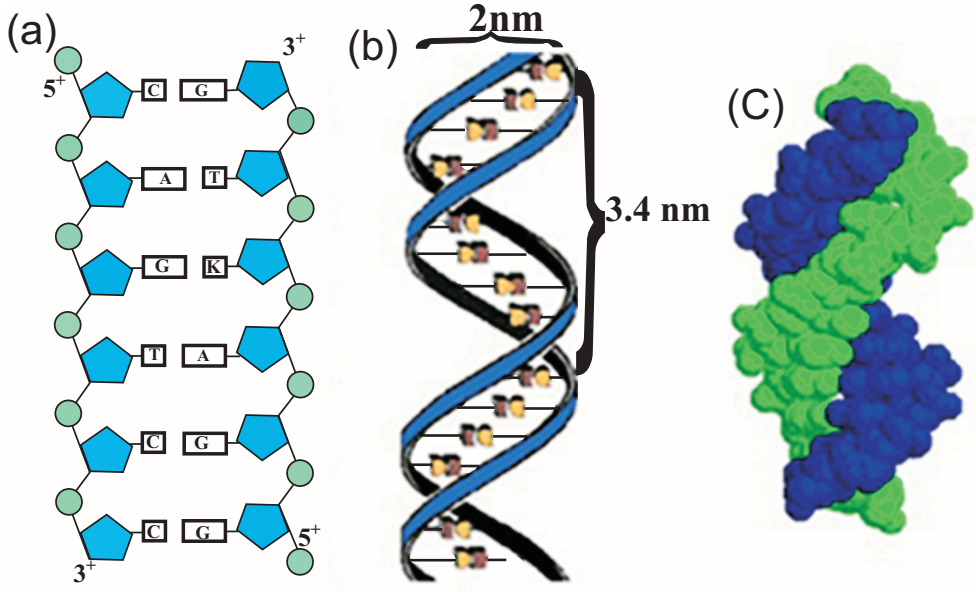

Fig. (1). The structure of DNA. (**a**) Two parallel chains are in opposite direction. One is 5'-3' directed and the other is 3'-5' directed. The pentagons stand for five-carbon sugars and the circles stand for phosphate molecules. Nucleotides polymerize into long chains by 5'-3' phosphodiester bonds. Hydrogen bonds link between purines and pyrimidines and the hydrogen bonds for C and G are stronger than those for A and T. (**b**) The three-dimensional structure of DNA discovered by Watson and Crick. (**c**) Model of double helix DNA structure.

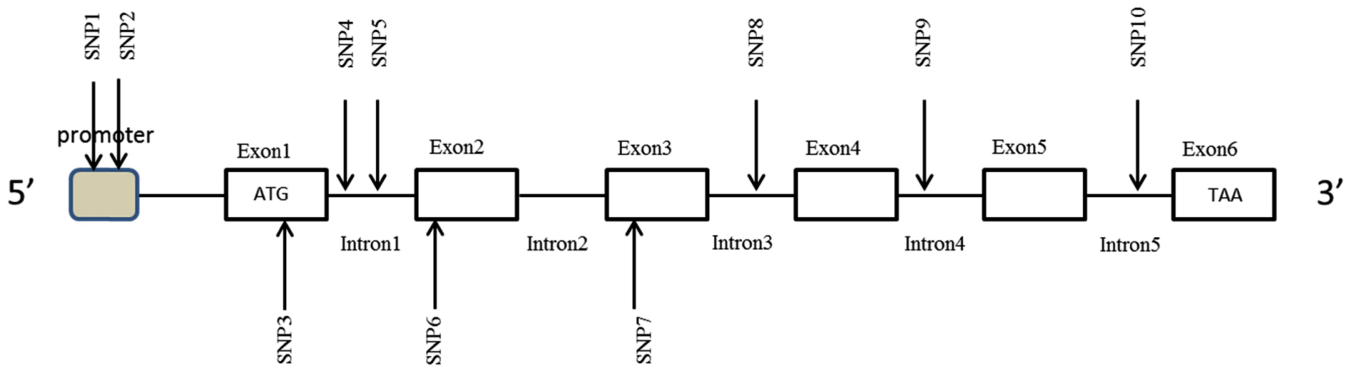

Fig. (2). Schematic gene structure. This gene has 5 introns and 6 exons. It is assumed there are 10 SNPs along the gene which may locate at promoter, introns or exons.

and across generations [20, 21]. On the other hand, if the repair system fails, a mismatch will lead to a *de novo* mutation. The DNA composition of the different types of cells in human is basically identical. However, the extent to which a given gene is "converted into" a functional protein may vary greatly in different cell types or even in the same type of cells at different states. Generally speaking, DNA is the instruction book, RNA is a photocopy of a specific page of the book and this page tells the cell how to make the protein. The RNA step ensures that energy is not wasted because the entire book is contained in every cell in the body, but certain cells only need to read certain pages (e.g. a nerve cell and a muscle cell use different sets of genes). When a specific protein is required, a process of transcription, the first step of gene expression, is initiated in which DNA is copied into an intermediate molecule named ribonucleic acid (RNA). One of the DNA strand serves as a template (called template or antisense strand), and RNA synthesis is also oriented in a 5' to 3' direction (corresponding to the N-terminus to C-terminus of the sequence of a polypeptide). The RNA transcript is complementary to the template, just like during DNA replication, except that a nucleobase uracil (U) pairs with A; therefore it has the same sequence as the non-template strand of DNA (which is called sense strand) except that U replaces T. A proofreading mechanism is also involved in transcription, but it is not as accurate as that of DNA replication [19]. Corresponding introns (non-coding sequencing in the RNA transcript) in a newly synthesized RNA molecule are subsequently removed by RNA splicing and a final mature messenger RNA (mRNA) is produced, which contains only exons (sequences that directly code for amino acids). The mRNA needs to be exported into organelle called ribosomes in the cytoplasm where the proteins are assembled. The sequence of an mRNA molecule is the template used to synthesize the corresponding a protein. The process by which mRNA is converted into a linear sequence of amino acids is called translation, the second step of gene expression. Specific nucleotide triplets, called codons, on the mRNA determine the start, the stop or the addition of an amino acid, leading to the creation of a polypeptide chain. Then a transfer RNA (tRNA), carrying the anticodon sequence (complementary to the codon on the mRNA) and a corresponding amino acid, binds to the codon on the mRNA and delivers the new amino acid to extend the polypeptide being synthesized. The maximum number of combinations of three bases out of four is theoretically 4^3=64 (Table **1**). Except for one specific codon (AUG) that initiates the translation of mRNA into protein and 3 codons (UGA, UAG, UAA) that stop translation, 61 out of the 64 triplets encode 20 different amino acids. Most amino acids are represented by more than one codon (e.g., six codons of UUA, UUG, CUU, CUC, CUA, and CUG for leucine). A specific codon always encodes a specific amino acid except in the case of the mitochondrial genome, which has four codons used differently from the nuclear DNA. This determines two important characteristics of the genetic code: specificity and degeneracy (also termed as redundancy). The degeneracy makes the protein more tolerant to some point mutations in coding regions and accounts for synonymous coding mutations. This means that a substitution of one nucleotide by another nucleotide does not necessarily result in an amino acid change (synonymous mutation) but others do change the coding sequence (non-synonymous mutations). All kinds of biological functions need the participation of proteins. However, the physiological roles of a protein depend on its amino acid sequence, configuration, and modulations from other relevant factors such as regulator proteins, ligands/ receptors or substrates. Mutations outside of the coding regions of the gene of interest may rather influence its mRNA expression or stability.

CHROMOSOMES, MITOSIS AND MEIOSIS

Most normal human somatic cells are diploid, and in their nucleus there are 46 continuous DNA molecules and each of them is named a chromosome [22]. The 46 chromosomes make up two sets, and therefore two copies of each chromosome have the same length, same centromere and identical genes and are designated homologs. One homolog is maternally inherited and the other is paternally inherited. Each set has 23 single chromosomes-22 autosomes and an X or Y sex chromosome. A male has an X and Y chromosome pair and female has a pair of X chromosomes (Fig. **3**). The twenty-two autosomes have been ordered from chromosome 1 to 22 according to the length of DNA base pairs (from the longest to the shortest). The X chromosome is much larger than the Y chromosome. The DNA sequences of two homologous chromosomes are usually not completely identical. DNA in a chromosome is packed in many complex

Table 1. The genetic code.

FIRST LETTER	SECOND LETTER U	SECOND LETTER C	SECOND LETTER A	SECOND LETTER G	THIRD LETTER
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cy	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cy	C
	UUA Le	UCA Ser	UAA Stop	UGA Stop	A
	UUG Le	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Le	CCU Pro	CAU His	CGU Ar	U
	CUC Le	CCC Pro	CAC His	CGC Ar	C
	CUA Le	CCA Pro	CAA Gln	CGA Ar	A
	CUG Le	CCG Pro	CAU Gln	CGG Ar	G
A	AUU Ile	AGU Thr	AAU As	AGU Ser	U
	AUC Ile	ACC Thr	AAC As	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Ar	A
	AUG Met	ACG Thr	AAG Lys	AGG Ar	G
G	GUU Val	GCU Ala	GAU As	GGU Gly	U
	GUC Val	GCC Ala	GAC As	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Sixty-four different combinations of triplet codons are derived from 4 unique bases. Except ATG for the start codon and TAG, GTA, TAA for stop codons, each codon codes for one of the 20 amino acids.

units called nucleosomes consisting of two copies of core histones H2A, H2B, H3 and H4 around which is wound by a fragment of DNA, like many beads (histones) on a string (DNA). A fifth histone H1 is located in the spacer region between any two nucleosomes. Histone H3 and H4 can be modified by post-translational regulation mechanisms such as methylation, acetylation, ubiquitination and phosphorylation [23, 24]. These proteins are involved in epigenetic mechanisms, which determine in part the stable gene expression pattern from cell to cell or from generation to generation in the absence of change to DNA sequences [25, 26]. Along each chromosome, a constriction point called the centromere divides the chromosome into two arms: the shorter arm or "p arm" and the longer arm or "q arm". In addition to identifying genetic diseases based on the patterns of G-banding (stained by Giemsa's solution at the metaphase) [27], chromosome arms are useful to describe the location of a specific gene mutations.

There are two types of cell divisions, mitosis and meiosis [28]. Mitosis occurs in the context of body growth, cell differentiation, self-renewal and regeneration in somatic cells. DNA replication and partitioning go along with mitosis, ensuring the maintenance of a diploid chromosome stock (2×23 chromosomes) in daughter cells. Meiosis is a specialized reductive cell division which occurs exclusively in germ cells and gives rise to sperm and egg cells. In a single diploid spermatocyte or oocyte, DNA duplication generates two identical sister chromatids, followed by two DNA segregations and cell divisions known as meiosis I and II. During meiosis I, the homologous chromosomes, which are paired together to form a bivalent may possibly exchange a fragment of DNA between maternal and paternal strands. This process of exchange of genetic material is called recombination (crossover) and is one of the key mechanisms by which genetic diversity between daughter cells is generated. Subsequently, a complete set of 2×23 chromosomes are pulled to either pole and separated to form two haploid cells, each with one of the homologs. Which homolog in a bivalent pair ends up to in which daughter cell is independent and this is called the independent assortment. Independent assortment is the second major mechanism of genetic diversity. Therefore in humans, the total number of possible combinations of chromosomes in one gamete is 2^{23}. Meiosis II is similar to mitosis except that final daughter cells have 23 chromosomes instead of 46. As a result, meiosis eventually produces four haploid gametes. All eggs have a 23,X chromosome constitution representing 22 autosomes plus a single X chromosome, and 50% of the sperms have a 23,X chromosome constitution and the other half are 23,Y (Fig. **4**). When a sperm fuses to an egg, a zygote is formed and the diploid chromosomal status is re-established. Taken together, the daughter cells from mitosis are genetically identical, whereas the daughter cells from meiosis are genetically different as a consequence of independent assortment and recombination. As discussed earlier, *de novo* DNA mutation may be caused by a failure in the repair system during DNA replication. In addition to this, an error in combination

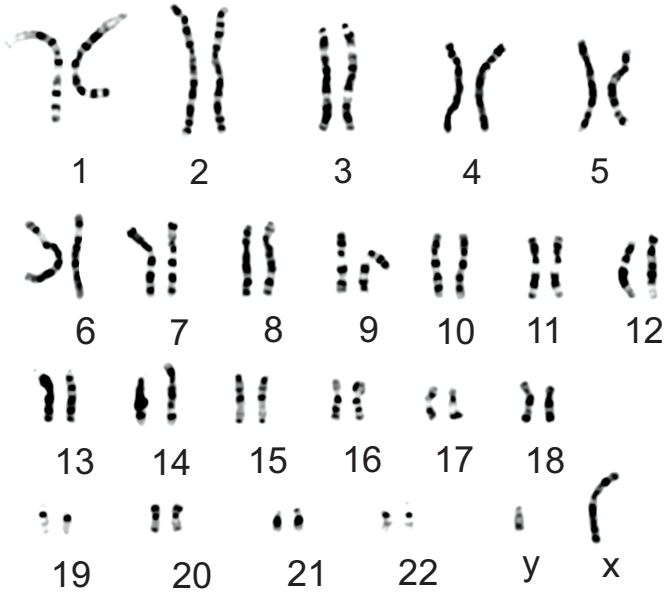

Fig. (3). A human male karyotype with Giemsa banding. The autosomes are arranged from 1 to 22 according to their length. Sexual X and Y chromosomes are displayed separately.

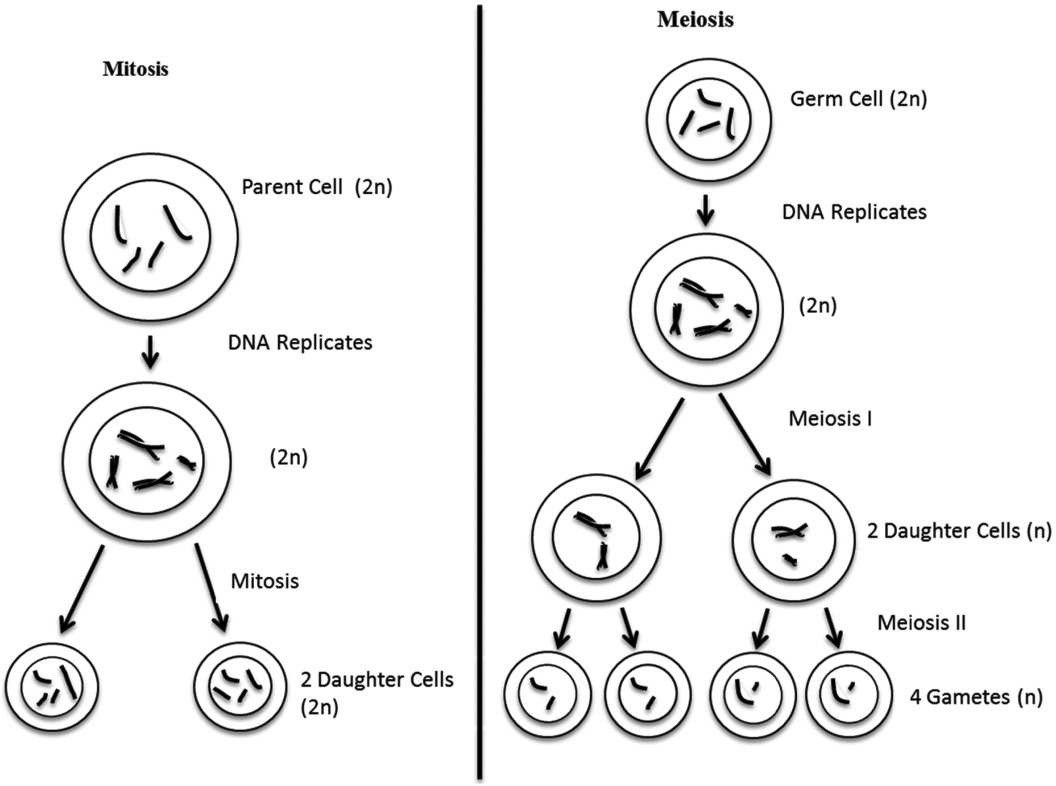

Fig. (4). Mitosis and Meiosis. In mitosis, one cell produces two identical daughter cells through DNA duplication, and division. In meiosis, one diploid germ cell gives rise to four haploid gametes through DNA duplication, and two cell divisions (meiosis I and meiosis II). Four chromosome pairs are shown as demonstrations.

process may also generate structure abnormalities in DNA. The average number of crossovers per cell is about 55 in males and is approximately 50% more in females, which means crossovers are not rare events. Crossovers are essential in maintaining the genetic variability that is transferred from parent to offspring. The exception again is mitochondrial DNA, which is inherited as a single linked molecule through the female line. It does not undergo recombination. Just as in DNA replication, errors during recombination do occur at a very low frequency, giving rise to translocations, inversions, duplications, or deletions.

Abnormalities of chromosome structure are reported to contribute to a small portion of cases in psychiatric disorders. Balanced translocation is an exchange of chromosome segments between two non-homologous chromosomes. A balanced translocation between chromosome 1 and 11 disturbed *DISC1* gene is associated with increased risk of schizophrenia [29, 30]. Chromosome inversion occurs when there are two breaks in one chromosome and the same segment is re-constituted with the orientation inverted. A pericentric inversion on chromosome 9 was found to be associated with schizophrenia [31].

CHARACTERISTICS OF THE HUMAN GENOME

The completion of the Human Genome Project in April 2003 and the 1000 Genomes Project in 2012 has revealed several important characteristics of the Human genome [32, 33]: 1) there are about 3 million base pairs in the human genome; 2) 99% of nucleotide bases are the same in all humans; 3) an estimated 30,000 genes exist in humans, with an average length of 3000 base pairs; 4) genes represent less than 2% of the human genome; 5) more than 50% of genomic DNA consist of non-repetitive DNA sequences and most of the genes display unique DNA sequences; 6) about 45% of genomic DNA consists of repetitive sequences which are thought to contribute to maintaining chromosome structure; 7) there are 38 million validated SNPs in which a single nucleotide differs at a particular position among 1,092 human genomes.

GENETIC VARIATIONS

The current entire database of human genomic variation was recently derived from a panel of whole-genome sequence data in 1,092 individuals from 14 populations in the context of the 1000 Genomes Project [33]. The next targeted milestone of the 1000 Genomes Project is sequencing the genome of 2,500 individuals from 27 populations across the world [34]. Although 99% of the genomic DNA sequences are identical, 1% still signifies 38 million genetic variants between unrelated individuals, indicating there is one allele variant in every 80 base pairs on average. During the assembling of consensus sequences, differences between (among) the nucleotide sequences of different individuals were noticed. SNPs represent more than 90% of all human variation [35]. If the frequency of a SNP is greater than 5%, it is considered a common variant or polymorphism. If the frequency is between 1-5%, it is a low-frequency SNP. If the frequency is less than 1% in population, it is defined as a mutation. In addition to SNPs, other genetic variants have been observed in the human genome, including microsatellites, minisatellites, and copy number variants (CNV). The 1000 Genomes Project also identified 1.4 million bi-allelic short insertions and deletions, and more than 14,000 large deletions [33]. Because these genetic variants were discovered during the sequence assembly, their locations are inherently known, providing a key resource in mapping genes that predispose to common diseases.

Genetic variants may occur in any region of the genome. A SNP that is located in the coding region without changing the corresponding amino acid is called synonymous, while coding SNPs that lead to changes of the amino acid, shifting of the reading frame or to an earlier stop code are called non-synonymous, frameshift or non-sense, respectively. SNPs found in a non-protein coding area in a gene may influence the protein expression by changing regulatory elements such as transcription factor, binding sites or configuration. In humans, there are usually only two alleles at a SNP location, but three alleles are sometimes reported, such as e2, e3, e4 alleles at the *APOE* gene locus [36]. The most common nomenclature of a SNP uses a unique reference SNP (rs) number. An example is the rs10994336 SNP in the *ANK3* gene that has been associated with bipolar disorder [37]. SNP data are available from publicly accessible resources and are constantly updated, such as dbSNP polymorphism repository, Human Genome Variation Database, the International HapMap Project, SNP consortium or 1000 Genomes Project database.

Microsatellites or short tandem repeats (STR) refer to repeated sequences of less than 10 bp of DNA. When the repeat units have 10-100 nucleotides and the copy number may reach up to a few thousand, this repeat cluster is referred as a minisatellite or a variable number tandem repeat (VNTR) [22]. The number of alleles in microsatellites and minisatellites is usually 5 or more. Though both microsatellites and minisatellites are highly unstable, the majority of the variations have no detrimental clinical consequences. The mutation mechanisms in microsatellites and minisatellites are different. In minisatellites, mutations occur during homologous recombination at meiosis, but the rate is approximately 10 times greater than that of other DNA sequences. Microsatellites undergo slip-strand mispairing during replication and subsequently the genes in the repair systems are inactivated, leading to expansion of the repeats [22]. This mutation rate is also several of orders of magnitude higher than the mutation processes that lead to SNPs. Some of them result in increased risks of diseases. For example, whereas healthy individuals carry less than 36 repeats of CAG in the *HD* gene, the number of repeats increases to more than 40 in individuals who will develop Huntington disease [1].

Another type of polymorphism is called copy number variant (CNV). CNV refers to the duplication or reduction of a DNA segment (200 bp to 1.5Mb) and they may have only two or multiple alleles. CNVs (deletion or duplication) can have important functional consequences and have been convincingly associated with psychiatric disorders such as schizophrenia [38].

ALLELES AND GENOTYPES

The location of a DNA sequence or a gene on a chromosome is called a locus. If there is more than one type of nucleotide at a specific locus in a population, each nucleotide is called an allele. Most polymorphic sites have only two alleles, while a few have more than two alleles. Individuals are called homozygotes when the two alleles of homologous chromosomes at a specific locus are identical. When the two alleles are different, individuals are classified as heterozygotes. At bi-allelic SNPs, the allele with higher frequency in a given population is called the major, and the less common one is called minor allele. The three (or more)

possible combinations of alleles at a specific locus (e.g. major allele / major allele, major allele/ minor allele, minor allele / minor allele) are called genotypes. Sometimes, a genotype refers to the overall genetic constitution of an individual.

For instance, the SNP rs1024582 in the *CACNA1C* gene is associated with bipolar disorder and schizophrenia [39]. There are two alleles A and G, A being the minor allele with frequency of 33.7% and G being the major allele. The three genotypes of an individual at this locus can be AA, AG or GG. The minor allele A increases the risk of bipolar disorder and schizophrenia [39].

HAPLOTYPES AND LINKAGE DISEQUILIBRIUM

Alleles of different loci are sometimes not independently transmitted from one generation to another. They may be physically linked on the chromosome and the crossovers across generations do not break them apart. Such a cluster of alleles is called a haplotype (Fig. **5**). The US National Institute of Health initiated the International HapMap Project in 2002 to develop a human haplotype map [40]. In phase I more than 1 million common SNPs were genotyped in 2005 in 270 individuals from four geographically distinct populations, Japanese, Han Chinese, Yoruba of Nigeria and Americans of North Western European ancestry [41]. These data were used to explore the patterns of association among SNPs in the genome, and how these patterns vary across populations. In Phase II HapMap, over 3.1 million SNPs were genotyped to create a second generation human haplotype map [42].

Linkage disequilibrium (LD) measures the non-random association of alleles at two or more loci that may or may not be on the same chromosome. For instance, we may consider two loci with alleles A1/A2 and B1/B2, A1 and B1 alleles being on one chromosome and A2, B2 alleles being on the other homologous chromosome. The frequency of A1 is 60% and B1 is 30% in a population. If the recombination of the two loci is independent, the expected frequency of the four possible haplotypes A1B1, A1B2, A2B1 and A2B2 would be 18%, 42%, 12% and 21%, respectively. If the distribution of these four haplotypes is consistent with the theoretical frequency, the alleles are in linkage equilibrium. If the distribution significantly departs from the theoretical frequency, the alleles are in linkage disequilibrium (LD), indicating the two loci are not independent. If one of the

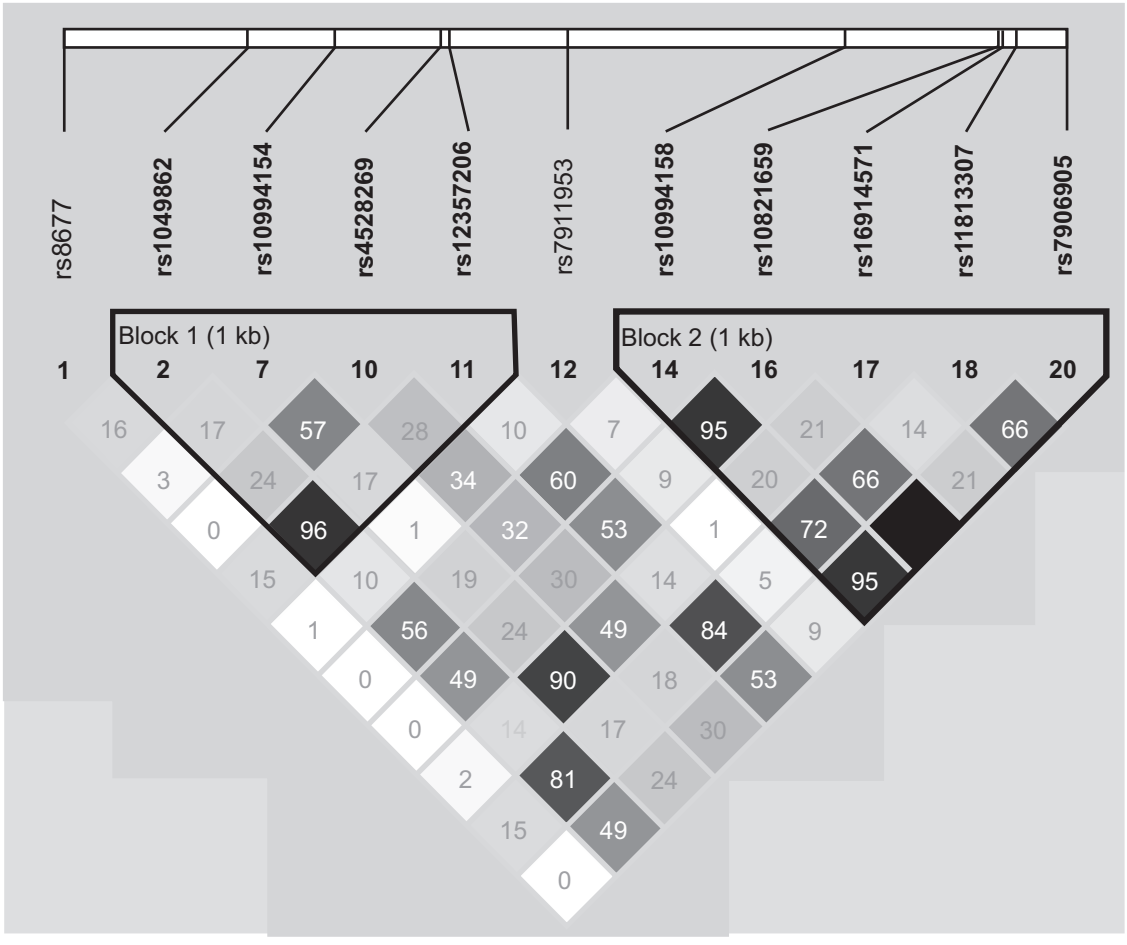

Fig. (5). Schematics of linkage disequilibrium (LD) plot. LD blocks among the 11 SNPs in *ANK3* gene are shown. The LD between the SNPs is measured as r^2 and shown (× 100) in the diamond at the intersection of the diagonals from each SNP. $r^2 = 0$ is shown as white, $0 < r^2 < 1$ is shown in gray and $r^2 = 1$ is shown in black. The top shows the relative physical positions of the SNPs on the chromosome 10. Two haplotype blocks (outlined in bold black line) indicate markers that are in high LD.

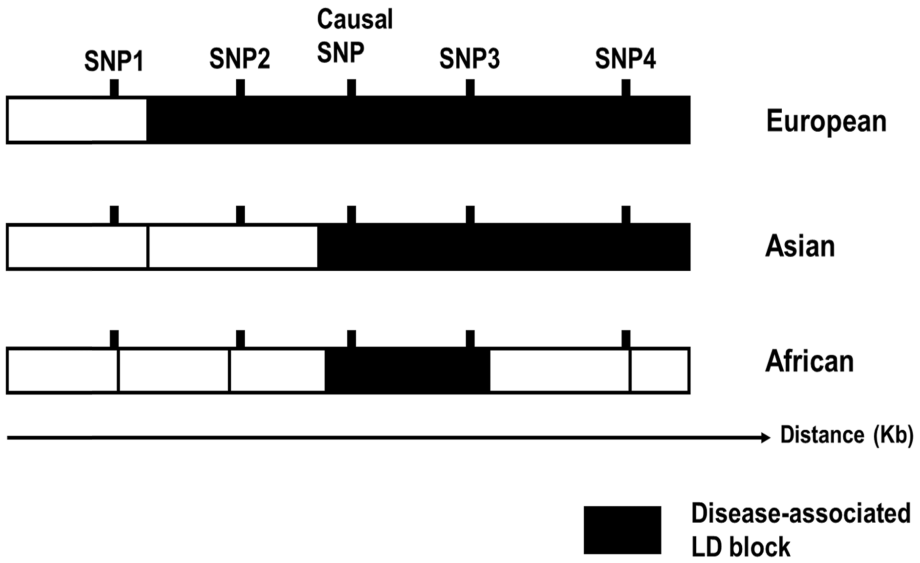

Fig. (6). Linkage disequilibrium patterns in different ethnic/racial groups. The size of the LD block where locates the causal SNP is smaller in African than in Asian and European populations. SNP2, SNP3 and SNP4 are proxy SNPs.

alleles is a disease causing allele, the haplotype including this allele is considered as a disease-containing haplotype. In most circumstances, a genetic variant that is found to be associated with a disease is not the functional diseasing causing allele; rather it is a proxy SNP. This indicates that this proxy SNP is in the same LD block with the potential causal SNP which is not genotyped in the array. LD may change over time and the patterns of LD may vary depending on the population. The sizes of LD blocks, which reflect the frequency of recombination, have been reported to be smaller in African than in Asian and European populations (Fig. **6**) [43]. Thus, knowing the LD pattern in a specific ethnic group (e.g. from the HapMap Project) is useful to refine the association signal and ultimately lead to the discovery of the causal variant [44, 45]. Many other factors may influence LD patterns, including random genetic drift, population growth, admixture, inbreeding, natural selection, and *de novo* mutation [46].

Using the example of haplotype given above, one statistical test to measure LD is D'=D/Dmax=(P_{A1B1}-$P_{A1}P_{B1}$)/ Dmax, where P represents the possibility of a specific haplotype and Dmax is the maximum difference between P_{A1B1} and $P_{A1}P_{B1}$. D' ranges from -1 to 1. One or -1 denotes there is no recombination between two loci A and B, and 0 indicates that A and B are in linkage equilibrium. If the allele frequencies of A1 and B1 are similar, a high D' value indicates A is a good surrogate for B. However, if the sample size is small or one allele is rare, D' will be inflated. There is a second measurement of LD, using the squared coefficient of determination r^2 (ranging from 0 to 1). r^2 takes into account the sample size and allele frequency. Therefore, D' is extensively used by population geneticist to assess recombination patterns such as defining haplotype patterns, whereas r^2 is a more appropriate measure of linkage disequilibrium in association studies [47]. For example, two SNPs can display a D' value of 0.85 and a r^2 value of 0.18. In an association study, these two SNPs cannot be tagged or substituted for each other because of low r^2. Pairwise measurement of LD for neighboring SNPs are used to group more than 2 loci into a haplotype termed LD block if the values of D' between any two SNPs within the group are above a certain threshold (e.g. D' > 0.8). This knowledge is essential to guide the design of whole-genome SNP genotyping arrays because carefully selecting a single or a few SNPs representing a haplotype block due to their strong LD can be used to identify an important haplotype, rather than genotyping all the SNPs in this haplotype (Fig. **5**). The common SNPs in commercial genotyping arrays captures untyped common variation with an average maximum r^2 (a correlation coefficient between genotyped and untyped SNPs) from 0.9 to 0.96 depending on the population. Therefore, the advances from Phase II HapMap, in combination with increased density of high-throughput technology and capability of imputation of untyped SNPs, greatly improved the power of association studies. HapMap 3 was completed in 2009 and it genotyped 1.6 million common and rare variants including CNVs in 1,184 reference individuals from 11 global populations [48]. The integrated map of genetic variation from the complete HapMap data and the 1000 Genomes Project [33] enables analysis of common and rare variants and CNVs in populations of different ethnic background. For instance, Sung and colleagues recently derived the genotypic distribution of 6.7 million SNPs from the information of 324,607 SNPs genotyped in their sample, using the reference panel from the 1000 Genomes Project [49].

CONCLUSIONS

Having a stronger background in molecular genetics, we are ready to discuss the subtle concepts of genetic epidemiology including study design implementation, gene identification strategies, genetic marker selection, genotyping and sequencing technologies, data analyses, data interpretation and their potential applications in the context of personalized medicine. The two next article in this series will review this topic.

GLOSSARY

Mendelian diseases: Phenotypes that are caused by a single gene mutation and display a clear pattern of inheritance

Complex diseases: Phenotypes that are caused by multiple genetic variants, environmental risk factors and interplays between them. They do not exhibit classic patterns of Mendelian inheritance

Genome-wide association studies: A study evaluating simultaneously associations between a dense subset of genetic variants theoretically covering the whole genome genetic diversity and a phenotype of interest

Single nucleotide polymorphism (SNP): A DNA variant in which a single base pair changes at a particular position

Gene: A segment of DNA embedding specific regulatory elements, non-coding regions and coding-regions which give instruction how amino acids assemble to a protein

Genotype: The genetic constitution at a specific locus or sometimes the overall genetic constitution of an individual

Allele: Each type of nucleotide at a given locus in a DNA fragment if there are two or more than two different types of nucleotides

Locus: The unique location on a chromosome at which a SNP or a gene is located

Homozygote: Individuals in whom the two alleles on the homologous chromosomes at a specific locus are identical

Heterozygote: Individuals in whom the two alleles on the homologous chromosomes at a specific locus are different

Linkage disequilibrium: A measure of non-random association between alleles at different loci

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

We thank Jackie Hudson and Arkan Al Abadi for editing of the manuscript, and the reviewers for their helpful comments. David Meyre is supported by a Tier 2 Canada Research Chair. Aihua Li is supported by a Queen Elizabeth II Graduate Scholarship in Science and Technology.

REFERENCES

[1] A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell 1993; 72: 971-83.

[2] Ewis AA, Zhelev Z, Bakalova R, *et al*. A history of microarrays in biomedicine. Expert Rev Mol Diagn 2005; 5: 315-28.

[3] The human genome at ten. Nature 2010; 464: 649-50.

[4] Klein RJ, Zeiss C, Chew EY, *et al*. Complement factor H polymorphism in age-related macular degeneration. Science 2005; 308: 385-9.

[5] Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet 2012; 90: 7-24.

[6] Bamshad MJ, Ng SB, Bigham AW, *et al*. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 2011; 12: 745-55.

[7] Kiezun A, Garimella K, Do R, *et al*. Exome sequencing and the genetic basis of complex traits. Nat Genet 2012; 44: 623-30.

[8] Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 2011; 8: 184-7.

[9] Pettersson E, Lundeberg J, Ahmadian A. Generations of sequencing technologies. Genomics 2009; 93: 105-11.

[10] Patterson K. 1000 genomes: A world of variation. Circ Res 2011; 108: 534-6.

[11] Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and disease. Annu Rev Med 2012; 63: 35-61.

[12] Hindorff LA, Sethupathy P, Junkins HA, *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 2009; 106: 9362-7.

[13] Wacholder S, Hartge P, Prentice R, *et al*. Performance of common genetic variants in breast-cancer risk models. N Engl J Med 2010; 362: 986-93.

[14] Scheuner MT, Sieverding P, Shekelle PG. Delivery of genomic medicine for common chronic adult diseases: A systematic review. JAMA 2008; 299: 1320-34.

[15] Avery OT, Macleod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J Exp Med 1944; 79: 137-58.

[16] Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 1953; 171: 737-8.

[17] Andersson SG, Karlberg O, Canback B, Kurland CG. On the origin of mitochondria: A genomics perspective. Philos Trans R Soc Lond B Biol Sci 2003; 358: 165-77.

[18] McCulloch SD, Kunkel TA. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. Cell Res 2008; 18: 148-61.

[19] Berg JM, Tymoczko JL, Stryer L. Biochemistry. 7th ed. New York: W.H. Freeman and Company 2012.

[20] Sancar A, Lindsey-Boltz LA, Unsal-Kacmaz K, Linn S. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. Annu Rev Biochem 2004; 73: 39-85.

[21] Wood RD, Mitchell M, Sgouros J, Lindahl T. Human DNA repair genes. Science 2001; 291: 1284-9.

[22] Lewin B, Krebs JE, Kilpatrick ST, Goldstein ES. Lewin's Gene X. 10th ed: Jones and Bartlet Publishers 2011.

[23] Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 1997; 389: 251-60.

[24] Thompson LL, Guppy BJ, Sawchuk L, Davie JR, McManus KJ. Regulation of chromatin structure *via* histone post-translational modification and the link to carcinogenesis. Cancer Metastasis Rev 2013.

[25] Bonasio R, Tu S, Reinberg D. Molecular signals of epigenetic states. Science 2010; 330: 612-6.

[26] Martin C, Zhang Y. Mechanisms of epigenetic inheritance. Curr Opin Cell Biol 2007; 19: 266-72.

[27] Speicher MR, Carter NP. The new cytogenetics: Blurring the boundaries with molecular biology. Nat Rev Genet 2005; 6: 782-92.

[28] Nussbaum R, Mclnnes RR, Willard HF, Hamosh A. Thompson & Thompson Genetics in Medicine. 7th ed. Philadelphia: PA 2007.

[29] St Clair D, Blackwood D, Muir W, *et al*. Association within a family of a balanced autosomal translocation with major mental illness. Lancet 1990; 336: 13-6.

[30] Millar JK, Wilson-Annan JC, Anderson S, *et al*. Disruption of two novel genes by a translocation co-segregating with schizophrenia. Hum Mol Genet 2000; 9: 1415-23.

[31] Kunugi H, Lee KB, Nanko S. Cytogenetic findings in 250 schizophrenics: Evidence confirming an excess of the X chromosome aneuploidies and pericentric inversion of chromosome 9. Schizophr Res 1999; 40: 43-7.
[32] Finishing the euchromatic sequence of the human genome. Nature 2004; 431: 931-45.
[33] Abecasis GR, Auton A, Brooks LD, *et al*. An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491: 56-65.
[34] Pennisi E. Genomics. 1000 Genomes Project gives new map of genetic diversity. Science 2010; 330: 574-5.
[35] Reich DE, Gabriel SB, Altshuler D. Quality and completeness of SNP databases. Nat Genet 2003; 33: 457-8.
[36] Corbo RM, Scacchi R. Apolipoprotein E (APOE) allele distribution in the world. Is APOE*4 a 'thrifty' allele? Ann Hum Genet 1999; 63: 301-10.
[37] Ferreira MA, O'Donovan MC, Meng YA, *et al*. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet 2008; 40: 1056-8.
[38] Guha S, Rees E, Darvasi A, *et al*. Implication of a rare deletion at distal 16p11.2 in schizophrenia. JAMA Psychiatry 2013; 70: 253-60.
[39] Ripke S, Sanders AR, Kendler KS, *et al*. Genome-wide association study identifies five new schizophrenia loci. Nat Genet 2011; 43: 969-76.
[40] The International HapMap Project. Nature 2003; 426: 789-96.
[41] The Internationanl Hapmap Consortium. A haplotype map of the human genome. Nature 2005; 437: 1299-320.
[42] Frazer KA, Ballinger DG, Cox DR, *et al*. A second generation human haplotype map of over 3.1 million SNPs Nature. 2007; 449: 851-61.
[43] Gabriel SB, Schaffner SF, Nguyen H, *et al*. The structure of haplotype blocks in the human genome. Science 2002; 296: 2225-9.
[44] Helgason A, Palsson S, Thorleifsson G, *et al*. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. Nat Genet 2007; 39: 218-25.
[45] Hassanein MT, Lyon HN, Nguyen TT, *et al*. Fine mapping of the association with obesity at the FTO locus in African-derived populations. Hum Mol Genet 2010; 19: 2907-16.
[46] Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 2002; 3: 299-309.
[47] Mueller JC. Linkage disequilibrium for different scales and applications. Brief Bioinform 2004; 5: 355-64.
[48] Altshuler DM, Gibbs RA, Peltonen L, *et al*. Integrating common and rare genetic variation in diverse human populations. Nature 2010; 467: 52-8.
[49] Sung YJ, Wang L, Rankinen T, Bouchard C, Rao DC. Performance of genotype imputations using data from the 1000 Genomes Project. Hum Hered 2012; 73: 18-25.

Received: July 22, 2013 Revised: January 31, 2014 Accepted: April 18, 2014