TCP Transl Clin Pharmacol
2016;24(1):13-21
http://dx.doi.org/10.12793/tcp.2016.24.1.13

Brief introduction to current pharmacogenomics research tools

Eun-Young Cha[1], Hye-Eun Jeong[1], Woo-Young Kim[1], Ho Jung Shin[1], Ho-Sook Kim[1,2]* and Jae-Gook Shin[1,2]

[1]Department of Pharmacology and PharmacoGenomics Research Center, Inje University College of Medicine, Busan 47392, Korea, [2]Department of Clinical Pharmacology, Inje University Busan Paik Hospital, Busan 47392, Korea

*Correspondence: H. S. Kim; Tel: +82-51-890-6748, Fax: +82-51-893-1232, E-mail: hosuegi@gmail.com

REVIEW

Received 1 Jan 2016
Revised 12 Jan 2016
Accepted 21 Jan 2016

Keywords
Genotyping technology,
Pharmacogenomics research tools,
Software for haplotype analysis,
Web-based genome browsers

pISSN: 2289-0882
eISSN: 2383-5427

There is increasing interest in the application of personalized therapy to healthcare to increase the effectiveness of and reduce the adverse reactions to treatment. Pharmacogenomics is a core element in personalized therapy and pharmacogenomic research is a growing field. Understanding pharmacogenomic research tools enables better design, conduct, and analysis of pharmacogenomic studies, as well as interpretation of pharmacogenomic results. This review provides a general and brief introduction to pharmacogenomics research tools, including genotyping technology, web-based genome browsers, and software for haplotype analysis.

Introduction

Pharmacogenetics and pharmacogenomics are the study of genetic and genomic variations that influence the response of an individual to drug treatment.[1,2] They allow us to understand the relationship between genetic variation and drug response. Pharmacogenomics can be employed to optimize specific medications, drug regimens, drug dosages, and follow-up treatments for an individual and facilitates a personalized approach to healthcare.[3]

Understanding pharmacogenomic research tools enables better design, conduct, and analysis of pharmacogenomic studies, and interpretation of their results. Collection of genomic information from the public, study design, performance of genotyping, and determination of the genotype and haplotype are key steps in the process of obtaining genomic information from the study results and are outlined in the following research process:[4,5]

1. The information on the efficacy and adverse reactions of a drug, its related genes including genetic polymorphisms, their frequencies and gene structure, and useful experimental methods are collected through public resources.
2. Target genes and polymorphisms of interest are identified and assay methods based on sample numbers, the number of single nucleotide polymorphisms (SNPs), gene size, and cost are determined.
3. Clinical study is designed and implemented (a full description of this step has been omitted because it is beyond the scope of this paper).
4. Genotyping (or another relevant assay) is performed using an appropriate assay method.
5. Genotype and haplotype information is obtained using software for haplotype analysis.
6. Relevant genomic informations including ethnic differences in allele frequencies, linkage disequilibrium (LD) structure, haplotype, confirmation of novel SNPs, and prediction of functional variants are characterized comparing with public genomic data.

This review will introduce key tools for pharmacogenomic researches including widely used genotyping technologies, simple and free softwares for haplotype construction, and web resources. We hope this information will increase the understanding and knowledge of pharmacogenomics and its interpretation.

Genotyping Methods

Recent advances in genotyping technology, such as gene chips and next generation sequencing (NGS) can now accommo-

Copyright © 2016 Translational and Clinical Pharmacology
It is identical to the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/).
This paper meets the requirement of KS X ISO 9706, ISO 9706-1994 and ANSI/NISO Z.39.48-1992 (Permanence of Paper).

date 1~5 million SNP assays, and have seen a plummet in the cost per genotype. These developments have encouraged an explosion of positive whole-genome association studies and the identification of many new genes associated with various diseases. While the use of gene chips for detection of common variants (>5~10% minor allele frequency) and NGS technology for detection of rare variants is increasing, conventional genotyping methods and single SNP detection technologies are also still in common use. The method of genotyping assays chosen is largely dependent on study objective, sample numbers, target SNP numbers, and cost.

This section will describe the most commonly used and commercially available genotyping assays including restriction fragment length polymorphism (RFLP), TaqMan, Pyrosequencing, SNaPShot, Sequenom, Fluidigm, the GoldenGate assay, and GeneChip array according to the principles of the SNP detection. The strengths, weaknesses, and availability of each genotyping assay will be discussed and are summarized in Table 1.

Enzyme cleavage genotyping methods

Enzyme cleavage genotyping is a classic method to detect SNPs that uses various enzymes such as restriction endonucleases, DNA polymerase, and 5' to 3' exonucleases. This method includes RFLP and the TaqMan assay.[6]

RFLP is the simplest and earliest method to detect SNPs.[7] SNP-RFLP uses many different restriction endonucleases, which provide high affinity to unique and specific restriction sites. DNA samples are digested by restriction enzymes and the resulting restriction fragments are separated according to their lengths by gel electrophoresis.[8] The advantage of this method is that only electrophoresis equipment is required, and thus is relatively easy to accomplish. One important limitation is that because the user defines the detection error based on the band pattern of restriction fragments, non-optimal reaction conditions for restriction enzymes pose a real risk for incorrect interpretation of genotyping results. Also, not all sites are cleaved due to the limited types of restriction enzymes.

The TaqMan assay method is widely used in large-scale studies to detect single SNPs. TaqMan uses commercially available fluorophore-labeled probes (FAM and VIC) and Taq polymerase. In this assay, the template DNA is combined with forward/reverse primers and fluorophore-labeled probes, which is amplified using polymerase chain reaction (PCR). When the fluorophore-labeled probe perfectly complements the template DNA, its reaction is amplified and the 5' fluorophore is released by the 5' nuclease activity of Taq polymerase. Release of the 5' fluorophore separates the 3' quencher, allowing fluorescence to be emitted and subsequently measured. If there is a mismatch between fluorophore-labeled probe and template DNA, the probe will not be a substrate for the 5' nuclease activity of Taq polymerase and is not cleaved. As a result, at the end point of the PCR reaction, a substantial increase of one dye over the other

Table 1. Comparative analysis of commonly used genotype platforms

Assay name	Reaction principle	SNP number	Sample number*	Flexibility	Advantages	Limitation	Reference
RFLP	Restriction enzyme reaction	1 SNP	Small samples	Fixed	Easy access	User-defined detection error	[8]
Taqman	5'nuclease reaction	1 SNP	Large samples	Fixed	Easy access/ Real-time	Uniplex only	[9,10]
Fluidigm	5'nuclease reaction	48 or 96 SNPs	Large samples	Semi-fixed	High multiplexing/ Flexibility	Specialized equipment	[11]
Pyrose-quencing	Primer extension	~3 SNPs	Middle samples	Fixed	Quantitation/ Semi-multiplex	Difficult to design multiplex/ expensive	[14,15]
SNaPshot	Primer extension	~12 SNPs	Middle samples	Flexible	Multiplexing/ High accuracy	Multiple steps	[16]
Sequenom	Primer extension	40~50 SNPs	Middle samples	Flexible	Multiplexing/ High throughput/ High accuracy	Multiple steps/ Specialized equipment	[17]
GoldenGate	Oligonucle-otide ligation	384~1536 SNPs	Various numbers of sample	Fixed	High accuracy	Multiple detection steps/ Specialized equipment	[19]
GeneChip array	Oligonucle-otide ligation	10K~11M SNPs	Various numbers of sample	Fixed	Very high throughput/ CNV detection	Complex experimental steps / Expensive/ Specialized equipment	[20]

*Sample number is categorized as small< 1500, middle 1500<~< 5000, and large> 5000, K: Kilobase, M: Millionbase.

indicates a homozygous sample; whereas increased fluorescence from both dyes indicates a heterozygous sample. Genotypes are determined by plotting the normalized fluorescence intensities on a scatter plot, and then using a clustering algorithm in the data analysis software (Fig. 1A).[9,10] Advantages of this method are that the results can be read in real time, it is cost-effective in assays of large samples, highly accurate, and probes are commercially available. However, if not commercially available probe, optimization of experiment condition is required. One of the limitations is that this technique only allows for the detection of single mutations.

The Fluidigm assay is a multiplex genotyping platform that uses TaqMan probes. It typically requires single-use biochips, equipment to handle the biochips, and software for the operation and extraction of genotyping results.[11] Fluidigm has a fixed size with measurements of 48 or 96 SNPs at a time. This method is only cost-effective when the SNP numbers are suitable for these fixed sizes.

Primer extension method

The primer extension method uses extension primers to targets SNPs which are few bases or a single base upstream from the polymorphic site (target SNP). During primer extension ddNTPs are added to the target SNP and the incorporated base is detected. There are several detection methods, including measurement of fluorescence or molecular weight. Fluorescence detection methods include Pyrosequencing and the SNaPShot assay; and size separation techniques using mass spectrometry such as Sequenom MassARRAY method (Fig. 1B).[12,13] One advantage of the primer extension methods is their flexibility due to the availability of customized designs. Oligonucleotide primers used to detect SNPs of interest can easily be added to, or removed from an existing panel. This method has the additional benefit of a short turnaround time for the entire analysis; for example, <3 hours is required for Pyrosequencing and <6 hours for SNaPShot.

Pyrosequencing is a DNA sequencing technique based on the detection of released pyrophosphate (PPi) during DNA synthesis. The released PPi is converted to ATP by ATP sulfurylase. Subsequently, ATP binds to luciferase to oxidize luciferin, and oxidized luciferin generates light which is easily detected using a charge-coupled device (CCD) camera. The intensity of the light signal is proportional to the number of incorporated nucleotides.[14,15] Pyrosequencing is semi-flexible, as a maximum of three SNPs can be detected at a time. This method can be applied to other kinds of genomic studies, for example determining the amount of methylation.

SNaPShot is a single-base extension (SBE) method. This method uses an oligonucleotide primer ending one base pair upstream of the target SNP.[16] Separation of each peak (different target SNPs) is dependent upon the primer length. Therefore, different primer lengths must be designed for multiplex genotyping. While this method has the advantage of being highly flexible and cost-effective, it has more experimental steps when compared to the TaqMan and Pyrosequencing methods.

The Sequenom MassARRAY system is also a SBE method. This system has a scalable platform offering quantitative and qualitative genomic analyses which uses MALDI-TOF mass spectrometry with unparalleled sensitivity and specificity.[17] A fixed homogenous reaction format is used in in each experimental step for the Sequenom method, These steps include using a single extension primer reaction, multiplex PCR reaction, a single termination mix and universal reaction conditions for all SNPs.[18] It is capable of performing multiplexing and high throughput. It is also highly accurate, but requires expensive and specialized equipment.

Oligonucleotide ligation method

The ligation-based approach uses the ability of DNA ligases which ligate two adjacent oligonucleotide primers bound to template DNA. In this method, two oligonucleotides are required; one is an allele-specific oligonucleotide in which the 3' end is complementary to the target SNP nucleotide, the second is an oligonucleotide with its 5' end adjacent to the 3' end of the first oligonucleotide. Both oligonucleotides are hybridized to the target DNA. DNA ligase is then added to the reaction. Ligation occurs only if the 3' end of the first oligonucleotide is complementary to the target SNP allele. The ligated and unligated products have different molecular weights which can be detected using a separation technique such as capillary electrophoresis or mass spectrometry analysis. This approach can be scaled up for high throughput analysis.[12,19] For example, Illumina GoldenGate Genotyping Assay can perform >55,000 genotypes and GeneChip assay can detect >10,000 SNPs

The Illumina GoldenGate Genotyping Assay (BeadArray technology) is an oligonucleotide ligation-based genotyping method. Three oligonucleotides are designed as two allele-specific oligonucleotides (ASO) and one locus-specific oligonucleotide (LSO) for each SNP locus. Three oligonucleotide sequences contain universal PCR primer sites. Genomic DNA is activated by binding streptavidin/biotin beads and then the ASOs and LSO are hybridized to the genomic DNA-bound streptavidin/biotin beads. Next, extension of the appropriate ASO and ligation of the LSO generates ligation products which are amplified using universal dye-labeled PCR primers. Finally fluorescence is used for signal detection (Fig. 1C). This method is designed for large-scale genotyping.

The GeneChip Microarray is based on the Molecular Inversion Probe (MIP) method. An MIP is a single oligonucleotide that recognizes and hybridizes to a specific genomic target sequence. It has two inverted recognition complementary flanks which range from 20 to 30 nucleotides. The total length of the MIP is 120 nucleotides. After the probe hybridizes to the target DNA, a single base pair gap, representing a SNP, appears in the middle of the two genomic-specific flank sequences. With the addition of the specific nucleotide to fill the gap, subsequent ligation oc-

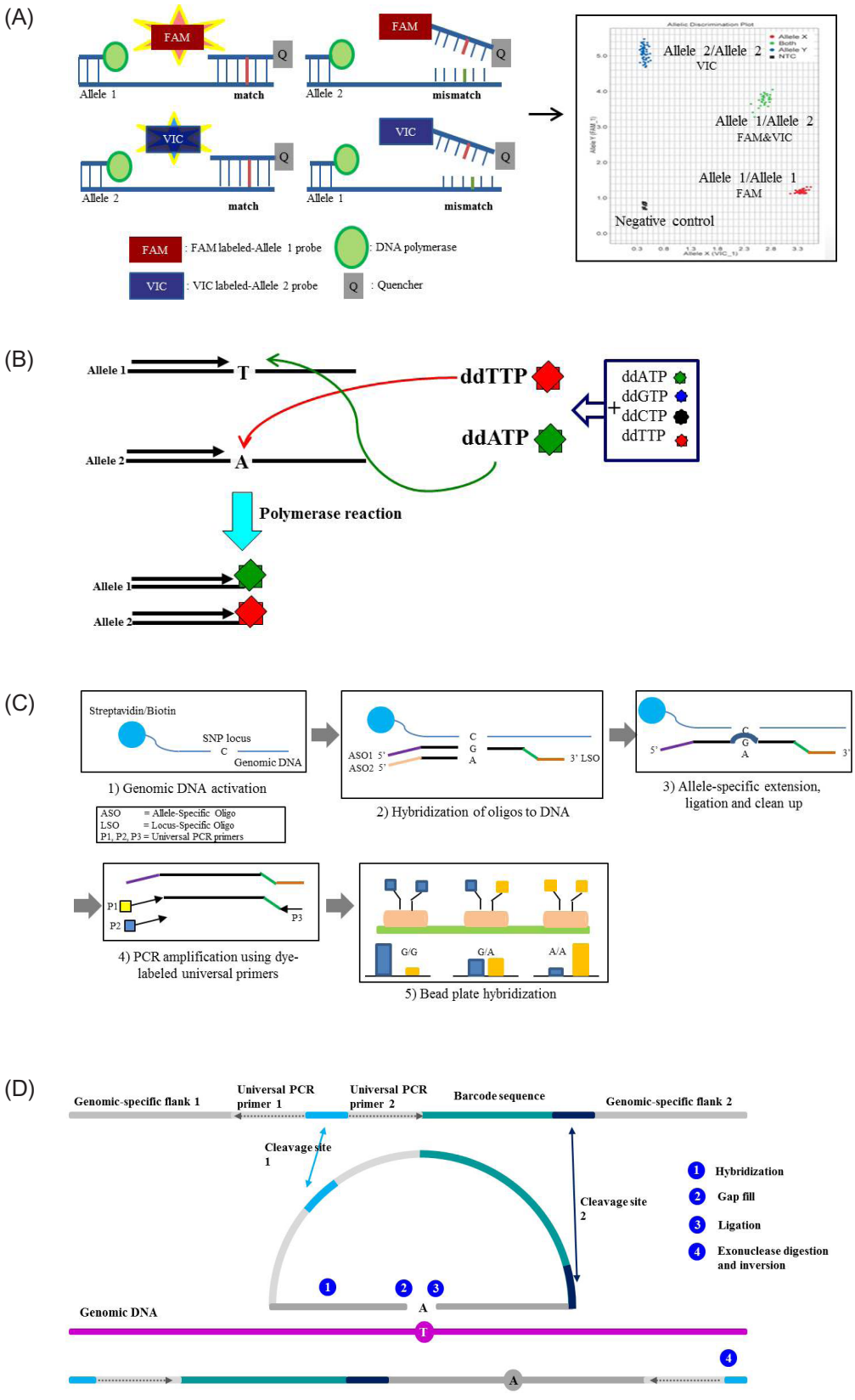

Figure 1. Schematic representation of SNP detection method. (A) TaqMan probe method. The template DNA is combined with primers and fluorophore-labeled allele specific probes, such as FAM labeled-allele 1 probe and VIC labeled-allele 2 probe. When a FAM-labeled allele 1 probe perfectly complements the target SNP site at allele 1, the FAM is released by the 5' nuclease activity of Taq polymerase. Release of the FAM separates the 3' quencher, allowing FAM to be emitted and subsequently detected as homozygotes of SNP at allele 1. In contrast, the VIC signal indicates homozygotes of SNP at allele 2. Fluorescences from both signals indicate heterozygotes. (B) Single base extension method. Extension primers are designed a single base upstream from the target SNP. During polymerase reactions, extension primers are bound a single base upstream of SNP site and ddATP are bound and extended to the target SNP site at allele 1, and the reaction is terminated. In contrast, ddTTP are bound and terminated to the target SNP site at allele 2. The incorporated base is detected using fluorescence. (C) Goldengate assay. Genomic DNA is activated by binding streptavidin/biotin beads. Both primers (ASOs and LSO) are hybridized to the genomic DNA-bound streptavidin/biotin beads. Extension of the appropriate ASO and ligation of the LSO generates ligation products. This product is amplified using dye-labeled universal PCR primers, and then fluorescence is used for signal detection. (D) GeneChip Microarray. ① Two genomic-specific flank regions are hybridized at genomic DNA ② The gap is filled with complementary base of target SNP and ③ ligated ④ cleavage site is digested by exonuclease, then the inversion probe is amplified by PCR reaction.

curred and the probe undergoes intramolecular rearrangement (i.e., circularization). After digestion of circled probe with exonuclease, the linearized probe is amplified by PCR reaction and labeled using either two or four fluorophores (Fig. 1D). Fluorescence is used for signal detection on microarrays. Microarrays can be used to measure gene expression levels and copy number variations (CNV) as well.[20] However, this method has complex experimental steps and is expensive.

Genome Browsers

Genome browsers are visualization tools that provide a graphical interface for users to browse, search, retrieve, and analyze genomic data and annotations.[21] This section will introduce web-based genome browsers that are useful in promoting biological research due to their good quality, accessibility, and high interworking between them.[22] Several representative web-based genome browsers including NCBI, UCSC, GeneCards, Ensembl, the 1000 Genomes, and KRG will be introduced in this section.

NCBI (http://www.ncbi.nlm.nih.gov)

The National Center for Biotechnology Information (NCBI) at the National Institutes of Health was founded in Bethesda, USA in 1988 to develop information systems for molecular biology and is accessible worldwide.

NCBI provides many types of resources for biological data including the GeneBank® database for DNA sequences, as well as biomedical information from the PubMed database of citations and abstracts for published journals. Other NCBI resources include: literature (PubMed Central (PMC), Bookshelf, and PubReader), health (ClinVar, dbGaP, dbMHC, the Genetic Testing Registry, HIV-1/Human Protein Interaction Database, and MedGen), genomes (BioProject, Assembly, Genome, BioSample, dbSNP, dbVar, Epigenomics, the Map Viewer, Nucleotide, Probe, RefSeq, the Sequence Read Archive, the Taxonomy Browser, and the Trace Archive), genes (Gene, Gene Expression Omnibus (GEO), HomoloGene, PopSet, and UniGene), proteins (Protein, the Conserved Domain Database (CDD), COBALT, Conserved Domain Architecture Retrieval Tool (CDART), the Molecular Modeling Database (MMDB) and Protein Clusters), and chemicals (the Biosystems and PubChem suite of small molecules databases).

NCBI provides a sequence similarity search program called BLAST (Basic Local Alignment Search Tool), which can perform sequence comparisons against the GenBank DNA database with quick access to results. The most important advantage of NCBI is the ability to visualize the sequence information in terms of a genetic map with "MapViewer". It provides information on SNPs, DNA, RNA, and protein sequences, and their loci in the genome using Idiogram, Contig map, and Unigene cluster map. It is useful in studies of genetic marker map construction. NCBI also provides interpretation of relationships between clinically important phenotypes and genotypes through ClinVar, a medical genetics resource.[23] Most of the data and executables for the software can be freely downloaded.

UCSC (https://genome.ucsc.edu)

The University of California, Santa Cruz (UCSC) Genome Browser is a web-based genome browser developed and maintained by the University of California, Santa Cruz. This browser is a web-based tool that provides resources for genome sequence databases from a variety of vertebrate (including humans) and invertebrate species integrated with a large collection of aligned annotations.

The menu item entitled 'Genomes' shows chromosome locations and sequences and the 'Genome browser' provides information on "Mapping and sequencing", "Genes and gene predictions", "Phenotype and literature", "mRNA and EST (expressed sequence tag)", "Expression (gene, mRNA, protein, etc.)", "Regulation", "Comparative genomics", "Variation", and "Repeats".

A large variety of data is displayed and a variety of data sources are linked to this website. For example, the entire contents from the NCBI dbSNP database are mapped to the human genome or other species' genomes. Data sources in this browser include the 1000 Genomes project, HapMap project, ENCODE project, and NCBI database.[24]

UCSC coordinated data for the Encyclopedia of DNA Elements (ENCODE) Consortium. Its data is hosted in the UCSC Genome Browser and database and is available for free download and analysis. Among the genome browsers, only UCSC provides a retrieval system for repeat sequences of DNA and RNA, the menu item entitled 'RepeatMasker' can perform comparative genomics and evolutionary conservation annotation with repetitive element identification. One of its main features is providing diverse genomic information through a single page website. UCSC provides BLAT, a sequence alignment tool similar to NCBI's BLAST. Alignment by BLAT is faster than NCBI's BLAST.

GeneCards (http://www.genecards.org)

The GeneCards database is developed and maintained in the Crown Human Genome Center at the Weizmann Institute of Science in Israel. GeneCards is a human gene database that provides genomic, proteomic, transcriptomic, genetic, clinical and functional information on all known and predicted human genes.[25] There are three types of cards in the search results page of the GeneCards database: the MicroCard, MiniCard, and GeneCard. The MicroCard provides the symbol, GCID (ID given to the gene by GeneCards) and a short description of the gene. The MiniCard, which is displayed after the MicroCard list, specifies the gene name, chromosomal location of the gene, gene description, GCID, and text lines. The GeneCard is a detailed description of all the data concerning a specific gene, including relevant links to other important websites.

The advantage of this database is that it provides a quick overview of the currently available biomedical information on

a given gene. This search engine is also user friendly and a resource for novice researchers and experts alike. It also provides the company names for experiment materials related to specific reference studies and links to their websites.[26]

Ensembl (http://www.ensembl.org)

The Ensembl project which was developed and is maintained by the European Bioinformatics Institute and the Wellcome Trust Sanger Institute, is one of several well-known genome browsers for the retrieval of genomic information and is similar to the NCBI database and the UCSC browser. Ensembl participates in large-scale international projects including the 1000 Genomes project, ENCODE, the International Cancer Genome Consortium, and the BLUEPRINT epigenome mapping project.[27] BLAST and BLAT can be used on this website. Another advantage of Ensembl is that it provides SIFT and PolyPhen protein function prediction scores which can predict loss-of-function and gain-of-function mutations. In addition, Ensembl allows users to easily obtain graphic displays of the layout of the exon-intron structures of genes.

1000 Genomes (http://browser.1000genomes.org)

1000 Genomes is the homepage of the 1000 Genomes project, which was the first project to sequence the genomes of a large number of people to provide a comprehensive resource on human genetic variation. Several research teams from institutes around the world, including China, Italy, Japan, Kenya, Nigeria, Peru, the United Kingdom, and the United States participated in this project to create a complete and detailed catalogue of human genetic variations. This website uses a project-specific version of the Ensembl browser to display its variants. Users can obtain information on ethnic differences in the human genome and for use in association studies relating genetic variation of disease. Users can access information about variants (e.g., SNPs, copy number variations (CNVs), and in/dels) with minor allele frequencies as low as 1% across the genome and 0.1~0.5% in gene regions, as well as estimates of population allele frequency, haplotype backgrounds, and LD patterns of variant alleles.[28] Datasets can be downloaded for free.

KRG project (http://cdc.go.kr)

The Korean Reference Genome (KRG) project was founded by the Center for Genome Science of the Korean National Institute of Health (KNIH) and the Korean Center for Disease Control and Prevention (KCDC). The KRG browser is a unique site which displays Korean whole genome variants. It contains reference gene and ensemble gene information panels, common and rare variants, genome diversity, selection tendency and variant density, allele frequency differences between Koreans and other populations (HapMap or 1000 Genomes), and functional annotations (Exonic variants and the ENCODE region variants). Korean genome database in this browser is freely and publicly accessible.

Haplotype Analysis

A haplotype is a set of SNPs on a chromosome pair. Because they are made up of unique SNP combinations which act like a genetic fingerprint, haplotypes are useful as a genetic marker to characterize diseases or drug responses. A few alleles of specific haplotype sequences can facilitate the identification of all other polymorphic sites which is critical for investigating genetics, and has been investigated in humans by the international HapMap Project.[29,30]

The construction of haplotypes is classified as either molecular haplotyping or *in silico* haplotyping. *In silico* haplotype construction methods are less accurate than molecular haplotyping, but are highthroughput: analyze large amounts of genotypic data at a relatively low cost. The construction of haplotypes using *in silico* methods is classified into three main algorithm types: parsimony (Clark's algorithm), maximum likelihood (Expectation-Maximization (EM) algorithm), and Bayesian. HAPIN is a representative program using Clark's algorithm.[31] Typical programs using the EM algorithm include Haploview,[32] HAPLO,[33] EM-Decoder,[34] and SNPHAP.[35] Representative programs using the Bayesian algorithm include PHASE,[36] HAPMCMC,[37] Haplotyper.[38] These softwares provide linkage information such as LD, LD blocks, and haplotype tagging SNPs (htSNPs), as well as haplotype information. Some of the various programs available for haplotype analysis are commercialized and others are available for free. Here, we will introduce three free and easily accessed programs: Haploview, PHASE, and PLINK (Table 2).

Haploview (https://www.broadinstitute.org/)

Haploview simplifies and expedites the process of haplotype analysis. It provides several functionalities such as LD and haplotype block analysis, haplotype population frequency estimation, single SNP and haplotype association tests, permutation testing for association significance, tag SNP selection, and more. Haploview uses four gamete frequencies, from which D' (standardized linkage disequilibrium coefficient) and r^2 (squared correlation coefficient) estimates are derived. LD is measured using D' or r^2. One of the advantages is its user-friendly formats. Haploview's linkage format includes information on partially or fully phased chromosomes or unphased diplotypes and can directly accept phased genotype data from the HapMap website (http://www.hapmap.org). In addition, Haploview can plot and visualize PLINK whole genome association results including advanced filtering options. Haploview provides graphical representations of LD blocks in portable network graphics (PNG) files.[39] Figure 2 is an example of the LD blocks in SNPs of inflammatory bowel disease 5 (IBD5) gene provided by Haploview (Fig. 2). The three major LD blocks in the gene is shown and among SNPs in the block has a strong linkage as red color (D' ≥0.80). Disadvantages of Haploview include not to analyze multiallelic SNPs or CNVs. It is also unable to perform diplotype-based analysis.

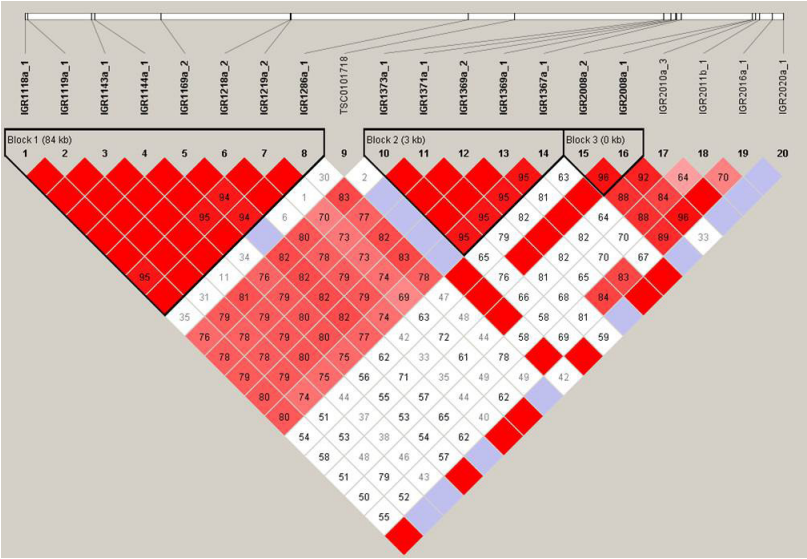

Figure 2. An example of what a Linkage Disequilibrium (LD) Map looks like (triangle format). This is a Linkage disequilibrium (LD) blocks structure of the inflammatory bowel disease 5 (IBD5) gene in chromosome 5q31-q33. The white line on top represents a strand of a chromosome. The black bars on the white line of the chromosome are SNPs (Single nucleotide polymorphism) that have been identified and sequenced. These SNP locations or loci are labeled in this picture as 1, 2, 3 and so on (#1~20 in this case). The kilobase (kb) in each LD blocks means the distance between first of SNP and end of SNP. The values in diamond represent the D' values (×100) between the two SNPs. For example, the diamond in which the columns leading from SNP#1 and SNP#7 intersect has a number, 95 with red color. Thus SNP#1 and SNP#7 have a D' value of 0.95 and are in high linkage disequilibrium with each other. The color is categorized according to D' value (D' ≥ 0.80, red; 0.5 ≤ D' < 0.8, pink; 0.2 ≤ D' < 0.5, blue; and D' < 0.2, white).

Table 2. Comparison of haplotype software

Contents	PHASE	Haploview	PLINK
Algorithm	Bayesian	Expectation-Maximization	Customized
SNP/CNV/Multi Allelic	○	x	○
Handling missing data	○	○	○
Haplotype Inference	○	○	○
Diplotype-based analysis	○	x	○
Visualization of LD block	x	○	○
Statistical analysis for association study	x	○	○
Input type	Text	6 Formats*	Customized
Export type	Text	PNG, text	Customized

*6 Formats: Linkage format, Haps format, HapMap format, HapMap PHASE, HapMap Download, PLINK format, Abbreviation; SNP: Single Nucleotide Polymorphism, CNV: Copy Number Variation, LD: Linkage Disequilibrium, PNG: Portable Network Graphics.

PHASE (http://stephenslab.uchicago.edu/software.html)

PHASE was the first method to utilize haplotype analysis and estimation of recombination rates from population data (http://www.stat.washington.edu/stephens/phase.html).[40,41] PHASE can provide haplotype frequency estimations, and diplotype information etc. One of advantages of PHASE is that it allows the analysis of microsatellites and other multi-allelic loci (e.g., tri-allelic SNPs and human leukocyte antigen (HLA) alleles) in any combination and missing data are allowed. It also provides diplotype information on each sample. However it is time-consuming to create input datasets.

PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/)

PLINK is an open-source whole genome association analysis toolset which performs basic large-scale analyse. PLINK provides information for data management, population stratification detection, basic association testing from a wide variety of SNP and CNV data, and summary statistics which are used for

quality control (missing genotype rates, minor allele frequencies, Hardy-Weinberg equilibrium failures, and non-Mendelian transmission rates). PLINK also provides an interface for recoding, reordering and merging files, flipping DNA strands, and extracting subsets of data. Because PLINK reads data in a variety of formats (including customized formats) and automatically combines several generically formatted summary files for unlimited numbers of SNPs, such as NGS data it can customize to the users' needs. PLINK is to be integrated with Haploview, thus supporting the subsequent visualization, annotation, and storage of results. Because the program is a command line tool, it is considered to be less user-friendly than other programs (Table 2).[42]

Summary

In this review, we introduced and discussed the advantages and disadvantages of several widely used genotyping methods, web resources, and software for haplotype analysis. Optimal selection of a genotyping method is largely dependent on the study objective, SNP number, sample number, and cost. The selection of appropriate web-based resource is dependent on the information of interest and pharmacogenomic researcher. Selection of haplotype analysis software is dependent on the study objective, the researcher's approach to genomic data analysis, and the size of the dataset. Understanding these differences will better help us design and conduct pharmacogenomic studies and interpret their results.

Acknowledgements

This research was supported by a grant of Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI14C0067) and by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Ministry of Education, Science and Engineering (MOEST) (No. R13-2007-023-00000-0).

Conflict of interest

The authors declared no conflict of interest.

References

1. Wang L, McLeod HL, Weinshilboum RM. Genomics and drug response. N Engl J Med 2011;364:1144-1153.
2. Mooney SD. Progress towards the integration of pharmacogenomics in practice. Hum Genet 2015;134:459-465.
3. Flynn AA. Pharmacogenetics: practices and opportunities for study design and data analysis. Drug Discov Today 2011;16:862-866.
4. Roden DM, George AL, Jr. The genetic basis of variability in drug responses. Nat Rev Drug Discov 2002;1:37-44.
5. Lee JK, Part3 Disease association study. Genetic variation and Diseases (Language in Korean). 2nd ed. Seoul, 2010;211-262.
6. Kim S, Misra A. SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng 2007;9:289-320.
7. NwanKwo DC AM. Restriction enzymes and their uses in specific sequencing to produce predictable fragment of DNA making genetic engineering simply. Journal of Pharmaceutical Research and Opinion 2011;5:148-152.
8. Nobile C, Romeo G. Partial digestion with restriction enzymes of ultraviolet-irradiated human genomic DNA: a method for identifying restriction site polymorphisms. Genomics 1988;3:272-274.
9. Koch WH. Technology platforms for pharmacogenomic diagnostic assays. Nat Rev Drug Discov 2004;3:749-761.
10. Livak KJ. SNP genotyping by the 5'-nuclease reaction. Methods Mol Biol 2003;212:129-147.
11. Frederickson RM. Fluidigm. Biochips get indoor plumbing. Chem Biol 2002;9:1161-1162.
12. Hsiao SJ, Rai AJ, Multiplexed Pharmacogenetic Assays for SNP Genotyping: Tools and Techniques for Individualizing Patient Therapy. In: Dr. Despina Sanoudou (ed) Clinical Applications. 1st ed. Yan An Road (West), Shanghai, 2012;35-54.
13. Nikolausz M, Chatzinotas A, Táncsics A, Imfeld G, Kästner M. The single-nucleotide primer extension (SNuPE) method for the multiplex detection of various DNA sequences: from detection of point mutations to microbial ecology. Biochem Soc Trans 2009;37:454-459. doi: 10.1042/BST0370454.
14. Ronaghi M. Pyrosequencing sheds light on DNA sequencing. Genome Res 2001;11:3-11.
15. Zhou Z, Poe AC, Limor J, Grady KK, Goldman I, McCollum AM, et al. Pyrosequencing, a high-throughput method for detecting single nucleotide polymorphisms in the dihydrofolate reductase and dihydropteroate synthetase genes of Plasmodium falciparum. J Clin Microbiol 2006;44:3900-3910.
16. Hurst CD, Zuiverloon TC, Hafner C, Zwarthoff EC, Knowles MA. A SNaPshot assay for the rapid and simple detection of four common hotspot codon mutations in the PIK3CA gene. BMC Res Notes 2009;2:66. doi: 10.1186/1756-0500-2-66.
17. Syrmis MW, Moser RJ, Kidd TJ, Hunt P, Ramsay KA, Bell SC, et al. High-throughput single389 nucleotide polymorphism-based typing of shared Pseudomonas aeruginosa strains in cystic fibrosis patients using the Sequenom iPLEX platform. J Med Microbiol 2013;62:734-740.
18. Gabriel S, Ziaugra L, Tabbaa D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr Protoc Hum Genet 2009;Chapter 2: Unit 2.12. doi: 10.1002/0471142905.hg0212s60.
19. Tian HL, Wang FG, Zhao JR, Yi HM, Wang L, Wang R, et al. Development of maizeSNP3072, a high-throughput compatible SNP array, for DNA fingerprinting identification of Chinese maize varieties. Mol Breed 2015;35:136.
20. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG. The affymetrix GeneChip platform: an overview. 398 Methods Enzymol 2006;410:3-28.
21. Spudich GM, Fernández-Suárez XM. Touring Ensembl: a practical guide to genome browsing. BMC Genomics 2010;11:295. doi: 10.1186/1471-2164-11-295.
22. Wang J, Kong L, Gao G, Luo J. A brief introduction to web-based genome browsers. Brief Bioinform 2013;14:131-143.
23. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014;42:D980-D985. doi: 10.1093/nar/gkt1113.
24. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res 2015;43:D670-D681.
25. Harel A, Inger A, Stelzer G, Strichman-Almashanu L, Dalah I, Safran M, et al. GIFtS: annotation landscape analysis with GeneCards. BMC Bioinformatics 2009;10:348.
26. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards Version 3: the human gene integrator. Database (Oxford) 2010;baq020. doi: 010.1093/database/baq1020.
27. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic Acids Res 2014;42:D749-D755.
28. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56-65.
29. Consortium IH. The International HapMap Project. Nature 2003;426:789-796.

30. Consortium IH. A haplotype map of the human genome. Nature 2005;437: 1299-1320.
31. Clark AG. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 1990;7:111-122.
32. Barrett JC. Haploview: Visualization and analysis of SNP genotype data. Cold Spring Harb Protoc 2009;2009:pdb ip71. doi: 10.1101/pdb.ip1171.
33. Hawley ME, Kidd KK. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 1995;86:409-411.
34. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12:921-927.
35. Clayton D. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 1999;65:1170-1177.
36. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001;68:978-989.
37. Morris AP, Whittaker JC, Balding DJ. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am J Hum Genet 20 02;70:686-707.
38. Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single434 nucleotide polymorphisms. Am J Hum Genet 2002;70: 157-169.
39. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263-265.
40. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 2003; 73:1162-1169.
41. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 2005; 76:449-462.
42. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-575.