Reviews of Science for Science Librarians: Genome-Wide Association Studies (GWAS)

Sarah H. Jeong, MLIS
Z. Smith Reynolds Library, Wake Forest University, Winston-Salem, NC

Seong-Tae Kim, Ph.D.
Center for Cancer Genomics, Wake Forest University School of Medicine, Winston-Salem, NC

Abstract

Since the completion of the Human Genome Project and the International HapMap Project, Genome-Wide Association Studies (GWAS) have been undertaken to uncover associations between genetic variants and human diseases and traits. GWAS research is a step forward in biomedical research, and this article provides an overview of the history, study design and significant achievements in this field. In addition, this article presents reliable, core research resources including data repositories, research institutes, consortia, analytical tools and bioethical issues related to GWAS, which can serve as a reference for librarians, graduate students and scientific researchers who are interested in this research area.

Keywords: GWAS, genomics, HapMap Project, 1000 Genomes Project, genomic databases, PLINK, bioethics

Introduction

The Human Genome Project (1990-2003) was a groundbreaking international research endeavor sponsored by the National Institutes of Health’s (NIH) National Human Genome Research Institute (NHGRI) and the U.S. Department of Energy that identified and sequenced
the human genome and provided the initial foundation for Genome-Wide Association Studies (GWAS). The goals of the Human Genome Project included identifying genes in human DNA in addition to analyzing the ethical, legal, and social issues arising from genomic research (U.S. Dept. of Energy Genome Program 2011). One of the major achievements of the Human Genome Project includes mapping over 3 million human single nucleotide polymorphisms (SNPs), or genetic variants, which has given rise to a novel approach to biological research and finding the genetic basis of common, complex diseases.

The completion of the Human Genome Project in 2003 has increased the pace of genomic research and has led to research initiatives such as the International HapMap Project, an international collaboration among scientists in Japan, United Kingdom, Canada, China, Nigeria, and the United States, which identified genetic variants in the human genome across multiple ethnic populations (International HapMap Consortium 2003). HapMap is short for "haplotype map," where a haplotype is a set of single nucleotide polymorphisms (SNPs), a region of linked common human genetic variants, which serve as markers to locate genes in DNA sequences. The International HapMap Project compared the genetic sequences of different individuals to identify most of the SNPs that occur commonly in the human genome.

These two landmark projects have promoted technologies and resources which have remarkably influenced genomic research in the life sciences. GWAS is one of the approaches in genomic research which investigates genetic factors with high throughput technologies in addition to environmental factors, which explain complex human diseases and traits. GWAS are defined as studies of “genetic variation across the entire human genome that [are] designed to identify genetic associations with observable traits, or the presence or absence of a disease or condition” (NIH 2007). Specifically, GWAS involves a large number of human subjects
consisting of cases and controls whose genomes are genotyped to identify SNPs associated with a specific disease or trait.

Before GWAS, family-based linkage analyses and candidate gene studies were used to investigate the genetic basis of human disease. Linkage analysis is a statistical method to identify chromosomal loci of the genes in which genetic variants are co-inherited with a certain disease or trait within a pedigree. The limitations of linkage analysis were low power and low resolution due to limited family size (Teare and Barrett 2005). Candidate gene studies involve identifying genetic variants in genes assumed to be associated with a certain disease or trait. The limitation of candidate gene studies is that they overlooked many potential causal regions or genes since they relied on a priori incomplete biological pathways. In addition, these study results tend to be difficult to replicate. Compared to the candidate gene approach, GWAS examines genetic variants across the entire genome, which enables researchers to identify novel genetic variants associated with common diseases and traits. GWAS research is a population-based approach, so it is convenient to collect a large sample which contributes to greater power than family-based linkage analysis, to replicate the findings and to conduct meta-analysis through the sharing of GWAS data.

The following study design has been commonly used to conduct GWAS, and a detailed protocol is available in deBakker et al. (2008).

- Collecting a large sample of affected and unaffected subjects
- Genotyping the SNPs using high throughput technologies
- Performing quality control such as minor allele frequency, missing rate, Hardy-Weinberg equilibrium
- Imputing ungenotyped SNPs using genotyped SNPs and the HapMap catalog
- Performing association tests with a genome-wide p-value cut-off
- Finemapping of the locus of significant SNPs to identify variants and genes
Replicating the findings in a large, independent sample
Performing meta-analysis using multiple study populations
Performing further analysis of the region to identify the biological mechanism

The first publication of GWAS reported that variation in the gene for complement factor H, which produces a protein involved in regulating inflammation, is associated with age-related macular degeneration (ARMD) (Klein et al. 2005). This outcome has been one of the most successful GWAS to explain half the risk of ARMD.

In 2007, the Wellcome Trust Case Control Consortium (WTCCC) study published the GWAS results of seven common, complex diseases, including bipolar disorder, coronary artery disease, Crohn’s disease, hypertension, rheumatoid arthritis, and Types 1 and 2 diabetes using 14,000 cases and 3,000 common controls from the United Kingdom (WTCCC 2007). The WTCCC made the GWAS data of these diseases publicly available to other researchers and provided a practical guide for GWAS. The WTCCC study has been frequently cited in GWAS research. After the WTCCC findings, GWAS publications have exponentially increased. Over 3,600 SNPs associated with various human diseases and traits have been discovered as of Spring 2011 (Hindorff et al. 2011).

A series of these efforts in GWAS research allow us to further understand heritability of human diseases and traits. Diseases that show high proportion of heritability are ARMD, Type 1 diabetes, Crohn’s Disease, systemic lupus erythematosus and prostate cancer (Manolio et al. 2009; Kim et al. 2010). Although GWAS have identified multiple genetic variants associated with certain diseases and traits, their proportion of heritability is relatively low, which is referred to as missing heritability. Type 2 diabetes, HDL cholesterol, height, early onset myocardial infarction, Parkinson’s disease and obesity are examples of low heritability. (Manolio et al. 2009;
Another limitation is that it can be difficult to distinguish a true association from a false positive result since GWAS scans the entire genome, but this issue has been addressed by using a large sample ranging from 10,000 to 250,000 subjects. Originally, GWAS adopted the common variants, common diseases rationale which resulted in small to moderate effect sizes. In contrast to researchers’ expectations, genetic variants associated with human diseases have been identified in gene deserts, which have posed a challenge in finding biological reasoning for their findings.

Genomic research including GWAS has uncovered that complex human diseases are substantially attributed to multiple genetic factors in addition to environmental factors. As the genetic basis of disease is identified through genomic research, these discoveries contribute to the development of personalized medicine, a novel individualized health care approach which takes into account clinical and genomic factors in human diseases. Additional benefits of personalized medicine include prediction, treatment and prevention of disease as well as reduced adverse side effects of medication and treatment. In addition, GWAS have identified the interaction between genetic variants and drug response in order to explain drug metabolism, efficacy and toxicity in pharmacogenomics.

As an extension of the International HapMap Project, the 1000 Genomes Project is the next phase of human genomic research. This project is a collaborative effort among the United States, United Kingdom, China, and Germany to catalog human genome sequence variation across multiple ethnic populations using next-generation sequencing technologies (Durbin et al. 2010). Future GWAS research will focus on the study of less common, rare SNPs from the 1000 Genomes Project with next-generation sequencing, which is expected to contribute to solve the missing heritability issue and is a step forward in advancing genomic research.
This annotated bibliography of core research resources on GWAS includes not only online dictionaries, professional associations, research institutes, consortia, but also data repositories and analytical tools. GWAS research signifies the new trend of the convergence of multiple disciplines including genetic epidemiology, bioinformatics and statistics, which calls for the expansion of the librarian’s role beyond the storage of published research outcomes and data required by the National Institutes of Health (National Institutes of Health 2003) and the National Science Foundation (National Science Foundation 2011). In addition, ethical, legal and social issues surrounding genomic research have emerged with the establishment of the Ethical, Legal and Social Issues (ELSI) Research Program in 1990 in conjunction with the Human Genome Project, in order to protect human research subjects who participate in genomic research (Green and Guyer 2011). In 2008, Congress passed the Genetic Information Nondiscrimination Act to ensure protection against genetic discrimination by health insurance companies and employers. This article also covers information resources related to bioethical aspects of genomic research.

The research resources in this article are selective rather than comprehensive in the sense that core resources indispensable to the GWAS research process have been included. References were compiled from the authors’ research experience and literature searches in PubMed, Web of Science, Google Scholar, and the NHGRI’s GWAS Fact Sheet (NHGRI 2010).

1. Online Glossaries and Dictionaries

Talking Glossary of Genetic Terms: http://www.genome.gov/Glossary/

The National Human Genome Research Institute has created a multimedia glossary of more than 200 terms with definitions on genetics and genomics. There is an audio clip for each
term in which a scientist elaborates on the written definition and provides broader context of the term in scientific research. This online glossary received a positive review in the May 2010 issue of *Choice: Current Reviews for Academic Libraries* (Shrode 2010).

**National Institutes of Health Genome-Wide Association Studies (GWAS) Glossary:**

http://gwas.nih.gov/09glossary.html

This concise glossary with related links is part of the NIH GWAS Policy website, which provides access to GWAS data through the creation of a centralized NIH GWAS data repository.

**National Cancer Institute Dictionary of Genetics Terms:**

http://www.cancer.gov/cancertopics/genetics-terms-alphalist/a-e

The National Cancer Institute Dictionary of Genetic Terms, which is a part of the Dictionary of Cancer Terms, is a reliable government resource and includes more than 100 terms related to genetics. This dictionary was developed by Physician Data Query Cancer Genetics Editorial Board, which consists of 29 specialists in the fields of genetics, clinical oncology, public health, and the social sciences. This dictionary is a useful resource for providing reference service to lower-level undergraduates and general readers.

### 2. Professional Associations and Conferences

**American Society of Human Genetics (ASHG):** [http://www.ashg.org/](http://www.ashg.org/)

The American Society of Human Genetics, founded in 1948, is a professional organization of researchers, clinicians, and genetic counselors. ASHG organizes annual meetings and workshops including the International Congress of Human Genetics (ICHG)/ASHG Meeting,
and publishes *The American Journal of Human Genetics*.

**Genetics Society of America (GSA):** [http://www.genetics-gsa.org](http://www.genetics-gsa.org)

The Genetics Society of America (GSA) was founded in 1931 and combined the reorganization of the Joint Genetics Sections of the American Society of Zoologists and the Botanical Society of America. GSA publishes the journal *Genetics* and hosts multiple conferences.

**Association of Professors of Human and Medical Genetics (APHMG):**

[http://genetics.faseb.org/genetics/aphmg/aphmg1.htm](http://genetics.faseb.org/genetics/aphmg/aphmg1.htm)

Founded in 1995, the Association of Professors of Human and Medical Genetics promotes and furthers human and medical genetics educational programs in North American graduate schools. APHMG is comprised of over ninety institutions including medical schools.

**International Federation of Human Genetics Societies (IFHGS):** [http://www.ifhgs.org](http://www.ifhgs.org)

The International Federation of Human Genetics Societies was founded in 1996 for an international network of professional human genetics societies from all continents. IFHGS organizes the International Congress of Human Genetics.

**Human Genome Organisation (HUGO):** [http://www.hugo-international.org](http://www.hugo-international.org)

Founded in 1988, the Human Genome Organisation is an international organization of scientists who are pursuing research in human genetics. This organization focuses on the medical application of genomic research and endeavors to advance genomic technologies on an
international level. In addition, HUGO hosts the Human Genome Meeting (HGM), which is an annual international conference.

**American Association for Cancer Research (AACR):** [http://www.aacr.org/](http://www.aacr.org/)

The American Association for Cancer Research was founded in 1907 and is the oldest and largest scientific organization focused on cancer research in the world. AACR hosts an annual conference, and the 101st Annual Meeting was held in 2010. In addition, AACR publishes six peer-reviewed scientific journals including *Cancer Research* and *Cancer Epidemiology, Biomarkers & Prevention*. AACR is included in this article since genomic research and GWAS are prominent in modern cancer research.

**University of Kansas Medical Center Information for Genetic Professionals:**

[http://www.kumc.edu/gec/prof/soclist.html](http://www.kumc.edu/gec/prof/soclist.html)

This website provides links to domestic and international genetic professional societies and human, medical, and clinical genetics organizations.

3. **Research Institutes**

**National Human Genome Research Institute:** [http://www.genome.gov](http://www.genome.gov)

Founded in 1989, the National Human Genome Research Institute completed the sequencing of the Human Genome Project in 2003. Currently, the primary mission of the National Human Genome Research Institute is advancing technology that will facilitate genome research and its application to human health in addition to conducting research on the legal, ethical, and social issues related to genomic research.
Wellcome Trust Sanger Institute: http://www.sanger.ac.uk/

The Wellcome Trust Sanger Institute is one of the largest non-profit genetic and genomic research institutes located in Hinxton, UK. In 1992, this institute was established with primary support from the Wellcome Trust. This institute has been a leader in genomic research and has played an important role in international consortia including Wellcome Trust Case-Control Consortium and the 1000 Genomes Project.

Broad Institute: http://www.broadinstitute.org/

The Broad Institute of the Massachusetts Institute of Technology (MIT) and Harvard University was launched in 2004 and is a community of scientific researchers who are collaborating to advance medicine through genomic research and to understand human genetic variation and its role in disease. The Broad Institute website provides comprehensive information on data, software and educational outreach efforts related to genomic research. In the data section, the website provides information and links to Human Genome and Genetic Variation websites including Human Genome Project Information, Human dbSNP (NCBI), and the Human Haplotype Database. This institute also maintains a portal for researchers including genomic software tools such as PLINK, Haploview, SNAP and EIGENSTRAT.

4. Consortia, Projects, and Networks

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

The International HapMap Project is a collaboration among scientists in Japan, United Kingdom, Canada, China, Nigeria, and the United States that has identified patterns of genetic variants in the human genome of multiple ethnic populations. With the completion of the
International HapMap project in 2005, this project website provides a comprehensive catalog of SNPs which scientists have utilized to investigate genetic variants associated with human diseases through conducting GWAS.

**National Cancer Institute Cancer Genetic Markers of Susceptibility (CGEMS):**


The Cancer Genetic Markers of Susceptibility Project (CGEMS) is a research project sponsored by the National Cancer Institute which began in 2005. The CGEMS project has conducted GWAS to identify genetic variants associated with certain types of cancer, including breast, prostate, pancreatic and lung cancers. In addition, scientists have conducted GWAS on risk-associated genetic variants with bladder cancer.

**Wellcome Trust Case-Control Consortium (WTCCC):** [http://www.wtccc.org.uk/](http://www.wtccc.org.uk/)

Established in 2005, Wellcome Trust Case-Control Consortium (WTCCC) is a consortium of 50 research groups across the United Kingdom. WTCCC endeavors to identify genome sequence variants influencing human diseases through implementation of large-scale GWAS. The Consortium has performed GWAS on 14,000 cases of seven diseases and 3,000 shared controls through WTCCC1-3 collaborative studies.


The Genetic Association Information Network was a public-private partnership that conducted GWAS to investigate the genetic basis of six diseases, including major depression, bipolar disorder, schizophrenia, Attention Deficit Hyperactivity Disorder (ADHD), diabetic
nepropathy in Type 1 diabetes and psoriasis (GAIN Collaborative Research Group 2007). Since the GAIN initiative has been completed, the GWAS data from this initiative have been stored in the database of Genotype and Phenotype (dbGaP) within the National Library of Medicine.

**Pharmacogenetics Research Network:**


The Pharmacogenetics Research Network, sponsored by the National Heart, Lung, and Blood Institute and the National Institute of General Medical Sciences, is a multidisciplinary, collaborative research effort that is investigating genetic factors that affect patients’ responses to medications.

**Genes, Environment and Health Initiative (GEI):** http://www.genesandenvironment.nih.gov/

The Genes, Environment and Health Initiative (GEI) is sponsored by NIH and supports developing environmental technologies for use in GWAS research to determine the interaction of environmental factors with genetic variation associated with disease.

**1000 Genomes Project:** http://www.1000genomes.org

Launched in 2008, the 1000 Genomes Project is a large-scale international research consortium to sequence the DNA variation of multiple populations and to provide a comprehensive resource on human genetic variation using advanced sequencing technology which is called the “next-generation” sequencing platform. This project is supported by multiple international institutes including the Wellcome Trust Sanger Institute in the UK, the Beijing Genomics Institute in Shenzhen, China, and the U.S. National Human Genome Research
Institute (NHGRI). The goal is to provide a comprehensive catalog for rare variants of which frequency is ≤ 1%.

5. Genomic Research Databases


The National Center for Biotechnology Information created dbSNP, which is a publicly available central repository of multiple molecular variations including SNPs for multiple species. This repository is accessible through the PubMed interface.

National Human Genome Research Institute Catalog of Published GWAS:

http://www.genome.gov/gwastudies/

The National Human Genome Research Institute Catalog of Published GWAS includes research publications compiled from weekly PubMed literature searches, daily NIH compilations of news and media reports, and the GWAS literature database HuGE Navigator (Hindorff et al. 2009). The citations chosen for this catalog consist of GWAS publications that assayed at least 100,000 SNPs in the initial stage with the p-value ≤ 1.0 x 10-5. Based on these criteria, this catalog provides detailed information on about 3,600 SNPs for 296 diseases and traits from 750 publications.

NIH database of Genotypes and Phenotypes (dbGaP):


The database of Genotypes and Phenotypes (dbGaP) is a database which archives the
results of GWAS investigating the interaction of genotype and phenotype. It was developed by the National Center for Biotechnology Information, a division of the National Library of Medicine, and is accessible through the PubMed interface.

SNAP: http://www.broad.mit.edu/mpg/snap/

SNAP is a web-based tool for the rapid retrieval of linkage disequilibrium proxy SNP results given input of one or more query SNPs based on empirical observations from the International HapMap Project and the 1000 Genomes Project (Johnson et al. 2008). SNAP helps researchers identify, annotate and visualize linkage disequilibrium SNPs, as well as utilize the outcomes for interpretation and comparison of GWAS results and fine-mapping studies.

University of California Santa Cruz (UCSC) Genome Browser:

http://genome.ucsc.edu/index.html

The UCSC Genome Browser developed by the Genome Bioinformatics Group in the Center for Biomolecular Science and Engineering at the University of California Santa Cruz is one of the most comprehensive genomic databases. This browser is a reference database for a large collection of genomes of several species. In particular, this browser provides extensive annotation information collected by annotators worldwide and various tools for visualizing, analyzing and summarizing genomic data.

HuGE Navigator: http://hugenavigator.net

HuGE Navigator is a publicly accessible portal that links to information on human genomics and genetic epidemiology. This online resource has been featured in notable journals such as Nature Genetics and cited by the NHGRI GWAS Catalog.
FASTSNP: http://fastsnp.ibms.sinica.edu.tw

FASTSNP (Function Analysis and Selection Tool for Single Nucleotide Polymorphisms) web server provides researchers with useful annotation information and data mining tools to identify biological functional effects for selected SNPs from GWAS. Specifically, researchers can map the SNPs of interest with the regions with potential functional relevance.

SCAN – SNP and CNV Annotation Database: http://www.scandb.org

SCAN is a web-based genetic and genomic database which also provides data mining tools. This database provides two types of SNP annotations: (1) physical-based annotation carried out based on the position and linkage disequilibrium of SNPs; and (2) functional annotation based on the gene expression level. This web-interface is useful for researchers who are interested in the correlation between SNPs and their gene expression level.

6. Analytical Software

Although numerous software packages are used in GWAS, this research guide highlights free software developed through government and/or non-profit organization support.

PLINK: Whole Genome Data Analysis Toolset: http://pngu.mgh.harvard.edu/~purcell/plink/

PLINK, developed by Shaun Purcell at the Center for Human Genetic Research, Massachusetts General Hospital, and the Broad Institute of Harvard and MIT with his colleagues, is one of the most powerful, free, open source analytical tools for GWAS. This software is designed to perform multiple types of analyses using large-scale genotype/phenotype with high
computational efficiency (Purcell et al. 2007).

**EIGENSOFT:** [http://helix.nih.gov/Applications/eigensoft.html](http://helix.nih.gov/Applications/eigensoft.html)

The EIGENSOFT package combines functionality from population genetics methods (Patterson et al. 2006) and the EIGENSTRAT stratification correction method (Price et al. 2006). The EIGENSTRAT method uses principal component analysis to explicitly model ancestry differences between cases and controls along continuous axes of variation. The resulting correction is specific to a candidate marker’s variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations.

**Structure:** [http://pritch.bsd.uchicago.edu/software/structure2_2.html](http://pritch.bsd.uchicago.edu/software/structure2_2.html)
[http://pritch.bsd.uchicago.edu/structure.html](http://pritch.bsd.uchicago.edu/structure.html)

Structure, which is a program developed at the University of Chicago, implements a model-based clustering method for inferring population structure using ancestry-informative markers (Pritchard et al. 2000).


Haploview, which was developed at the Broad Institute, is broadly used for linkage disequilibrium (LD) and haplotype block analysis, single SNP and haplotype association test, tagging SNPs, and visualization of PLINK results and HapMap database. The heat map and Manhattan plot are popular applications of Haploview (Barrett et al. 2005).
7. Imputation Software

IMPUTE [http://mathgen.stats.ox.ac.uk/impute/impute.html](http://mathgen.stats.ox.ac.uk/impute/impute.html) and MACH [http://www.sph.umich.edu/csg/abecasis/MACH](http://www.sph.umich.edu/csg/abecasis/MACH) are the most popular software programs for imputing unobserved genotypes using reference haplotype information provided by the HapMap Project or 1000 Genomes Project. Other programs for imputation include BEAGLE, PLINK, fastPHASE. For a detailed algorithm and comparison of these programs, refer to Marchini et al. (2007) and Nothnagel et al. (2009).

8. Genotyping Technology

Genotyping is the process of determining the genotype of a subject using a research subject's blood or saliva. Genotyping is currently the most costly procedure in GWAS research. Genotyping technology is mainly provided by commercial technology companies, e.g., Affymetrix [http://www.affymetrix.com](http://www.affymetrix.com) and Illumina [http://www.illumina.com](http://www.illumina.com). In GWAS research, the most important factors to consider in choosing pertinent genotyping technology are accuracy, speed, and the cost of processing the assay.

9. Ethical, Legal and Social Issues of Genomic Research


In 1990, the Ethical, Legal Social Issues (ELSI) Research Program was established to foster analysis of broader issues pertaining to genomic research which include the protection of
human research subjects, protection from genetic discrimination, the regulation of genetic testing, and fair access to genomic medicine. Legal, regulatory, and public policy aspects of the medical applications of genomic research are also important components of the ELSI Research Program.

**Genetic Information Nondiscrimination Act (GINA) of 2008:**

http://www.genome.gov/24519851

The Genetic Information Nondiscrimination Act (GINA) was passed by Congress in 2008 and is intended to protect people from genetic discrimination by health insurance companies and employers. GINAhelp.org http://www.ginahelp.org/ is an online resource which provides information about GINA and produced by the Genetic Alliance, the Genetics and Public Policy Center at the Johns Hopkins University, and the National Coalition for Health Professional Education in Genetics.

**Conclusion**

The essential research resources in this GWAS annotated bibliography included online glossaries and dictionaries, professional associations and conferences, research institutes, consortia and projects, genomic research databases, analytical tools and bioethical issues. As GWAS research encompasses the convergence of multiple disciplines including genomics, bioinformatics, statistics, and bioethics, this article provides librarians, graduate students and researchers with valuable information sources and tools needed for the entire GWAS research process.
References


doi:10.1038/nature09534.


