An Introduction to Genome-Wide Association Studies: GWAS for Dummies

A. G. Uitterlinden, PhD[1]

[1] Department of Internal Medicine, Genetic Laboratory, Erasmus MC, Rotterdam, The Netherlands

Address for correspondence A. G. Uitterlinden, PhD, Department of Internal Medicine, Genetic Laboratory, Erasmus MC, Room Ee575, Rotterdam, The Netherlands (e-mail: a.g.uitterlinden@erasmusmc.nl).

Semin Reprod Med 2016;34:196–204

Abstract

Although the genetic origin of many human diseases and phenotypes has been long and widely recognized, identification of the causative gene alleles has been limited, slow, and cumbersome. This has changed substantially with the introduction of genome-wide association studies (GWASs) a decade ago, fueled by studies and reference projects of human genetic diversity and the development of novel DNA analysis technology applicable to high-throughput and large-scale data generation. Although GWASs essentially combine epidemiological study designs with molecular genetic analysis techniques, it has also fundamentally changed the way in which research was done in human genetics by the introduction of large consortia of collaborating investigators. GWASs have over flooded many clinical and basic research areas with gene discoveries, including those in reproductive medicine. This review describes aspects of GWAS methodology and how this field of human genetics is developing.

Keywords

- genome-wide association study
- complex disease
- SNP
- arrays

Mendelian and Complex Genetics

The genetic origin of most human diseases is long and widely recognized. The genetic contribution to disease is most clearly demonstrated in Mendelian diseases, where a DNA mutation in a single gene transmitted in families gives rise to usually devastating phenotypic consequences.[1] As they result from mutation in a single gene, they have become known as "simple genetic diseases" (as compared with the complex genetic diseases; see later). Well-known examples are cystic fibrosis, Huntington disease, and fragile X syndrome, to name a few. About 7,000 such monogenic disorders have been recognized and cataloged by Victor McKusick in "Online Mendelian Inheritance in Man" (http://www.ncbi.nlm.nih.gov/omim). They have been studied for several decades with great success, driven by the advances in molecular biological DNA analysis technology in the 1970s and 1980s. This resulted in the identification of approximately 3,000 mutated genes, roughly half of the known Mendelian disease genes. These discoveries have been important for the affected families by informing them on the origin of their disease and by illuminating molecular disease mechanisms, thereby leading to a rise in the number and importance of clinical genetic departments.

Such Mendelian diseases, however, are typically (very) rare and do not represent what a clinician will encounter in daily practice in hospitals. These patients suffer from what is known as the common complex diseases, such as diabetes, osteoporosis, Alzheimer disease, and cancer. They are referred to as complex because their origin is not driven by one genetic factor, but rather by many genes and (time-dependent) gene–environment interactions. The genetic basis of these complex traits and diseases has not been clear until twin studies documented and quantified what proportion of their causes are in fact genetic in origin (the "heritability"). This is done by comparing concordance for a particular trait or disease in many pairs of dizygotic twins with pairs of monozygotic twins, with highly heritable traits having much higher concordance in monozygotic twins. Many such twin studies have been performed over the past few decades, resulting in the recognition that many (if not all) traits and disease have a heritable component, albeit to varying degrees. These two categories of diseases are known as simple versus complex genetic disease, and several characteristics of them are highlighted in ►**Table 1**.

Yet, defining such heritability in complex genetic disease does not identify which specific genetic factors are responsible for the genetic component of the disease and are therefore just

Issue Theme Lessons from Genome-Wide Association Studies in Reproduction; Guest Editor, Joop S. E. Laven, MD, PhD

DOI http://dx.doi.org/10.1055/s-0036-1585406.
ISSN 1526-8004.

Table 1 Some characteristics of simple and complex diseases

"Simple"/Monogenic disease	"Complex" disease
Severe phenotype	Mild phenotype
Early onset	Late onset
Rare	Common
Mendelian inheritance	Complex inheritance
Cystic fibrosis, osteogenesis imperfecta, etc.	Diabetes, asthma, osteoporosis, etc.
Cause mutations ($f < 1\%$)	Polymorphisms ($f \geq 1\%$)

a starting point in strategies to find these. Such strategies were scarce and initially based on flawed study designs (such as small-scale candidate gene studies without sufficiently large replication samples) and on approaches and technologies, which had been successful in identifying Mendelian disease genes (such as linkage analysis and use of microsatellite repeats as DNA markers). During 1995 to 2005, many such attempts took place, but only late in the process, the human genetics community realized that these techniques were not going to work due to the fact that most, if not all, of these complex diseases are caused by many gene variants with subtle effects. It was thought that the classical epidemiological study designs, in particular association studies, would be much better powered to find those subtle effects of common genetic variants. During 2000 to 2005, several developments eventually led to an approach that would be very successful in identifying just those many subtle effects: the genome-wide association study (GWAS) approach.[2] ►**Fig. 1** depicts an overview of the different techniques and approaches that have been used to identify genetic variants responsible for disease risk. As these variants can be very rare to very common, they together determine the genetic architecture of a disease.

The Prequel: The Human Genome Project, dbSNP and HapMap

In 2001, "the Human Genome Project" (http://www.genome.gov/10001772) was completed reporting a near-complete map of the human genome. It was an unprecedented massive effort in human biology and medicine involving global collaboration between genome biologists and sequencing centers. Although it set the stage for what later would become common practice in complex genetics, that is, large-scale global collaborations between scientists, it only provided a (necessary and very detailed) map of the human genome and not the information on how much variation there is between human genomes and where that is located. That information, which is crucial to perform GWAS, came from databases in which polymorphism data had been collected over the years (single nucleotide polymorphism database [dbSNP]; http://www.ncbi.nlm.nih.gov/projects/SNP/index.html) and from resequencing efforts of panels of reference genomes collected in the framework of the HapMap project (http://hapmap.ncbi.nlm.nih.gov/). Started in 2002, the first phase of HapMap studied a few million SNPs in 30 to 45 family trios derived from four populations (Yoruba from Nigeria, Caucasians from Utah, Japanese from Tokyo, and Han Chinese from Beijing). This provided information on the interrelationships of common variants (with a population frequency $> 5\%$) across these ethnic groups. And the existence of approximately 1 million linkage disequilibrium (LD) blocks, stretches of a several million base pairs in which variants show strong correlations with each other, in the human genome became apparent. This LD block structure of the human genome also led to the realization that not all of human genetic variation (many millions of SNPs) had to be genotyped to find associated variants, but that only a subfraction (a few hundred thousands) had to be genotyped which "tag" many other

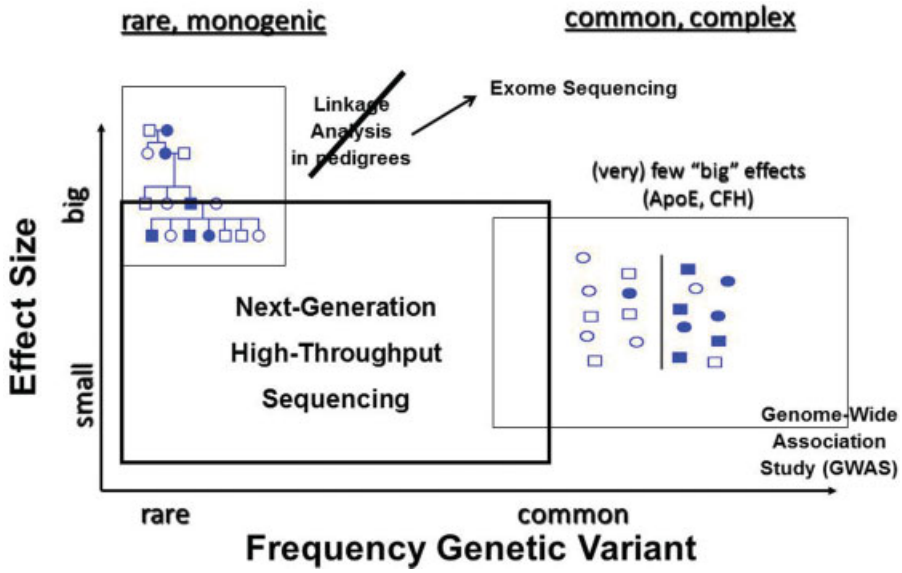

Fig. 1 Study designs to identify "risk" alleles, depending on the genetic architecture of traits or diseases which is expressed as allele frequency (*x*-axis) and effect size (*y*-axis).

genetic variants in such an LD block through LD correlations. This phenomenon is used in imputation strategies (see later).

In parallel to and fueled by these projects and database development, DNA analysis technology had also made great progress resulting in the array technology which allowed DNA to be genotyped for hundred thousands of SNPs. In particular, two companies (Affymetrix [Affymetrix, Santa Clara, CA] and Illumina [Illumina, San Diego, CA]) had made commercial products which allowed performing such genotyping experiments and this opened up the possibility to do a large-scale genotyping experiment assessing many hundred thousands of SNPs in relation to phenotypes and diseases characterized in several hundreds of human subjects. Scientists had realized that epidemiological study designs would be optimal to find causal genetic variants for complex disease, large-scale databases had been created of genetic variation among humans, and DNA analysis technology had provided tools to do such experiments. Therefore, it was in 2004 that all these came together and led to the birth of GWAS.

Basic Genome-Wide Association Study Methodology

Initially, in a typical GWAS, DNAs from a few hundred cases and few hundred controls are genotyped by either arrays of Affymetrix or Illumina, and many genotypes are then analyzed in relation to disease in a case–control analysis, or in relation to quantitative trait phenotypes by regression analysis (see ►**Fig. 2**). This will then identify one or more genetic variants or SNPs to be associated with a disease or phenotype.

DNA Collections

A successful GWAS starts with having available a large collection of DNA samples of well-phenotyped subjects. This can be a case–control collection or subjects from a cohort study. Case–control collections usually come from clinical centers where collecting samples from patients is part of a biobanking effort. However, collecting samples from controls is more cumbersome in such a setting, and therefore frequently controls are sought outside of the clinical centers, for example, coming from cohort studies. Case–control studies usually involve somewhat more rare diseases and phenotypes in which the cases have a population frequency of less than 1%. Although usually a case:control ratio of 1:1 is taken as golden standard, the most optimal ratio in terms of power is 1:4. Although cases can be collected relatively easily by dedicated clinicians, frequently encountered issues in this study design include lack of standardization and harmonization, which clinicians use worldwide to define a case.

Large longitudinal cohort studies have many advantages for GWAS. They allow both disease phenotypes to be studied and quantitative phenotypes, such as height, blood pressure, or serum AMH levels. They also allow multiple phenotypes to be studied in the same population, sometimes providing data

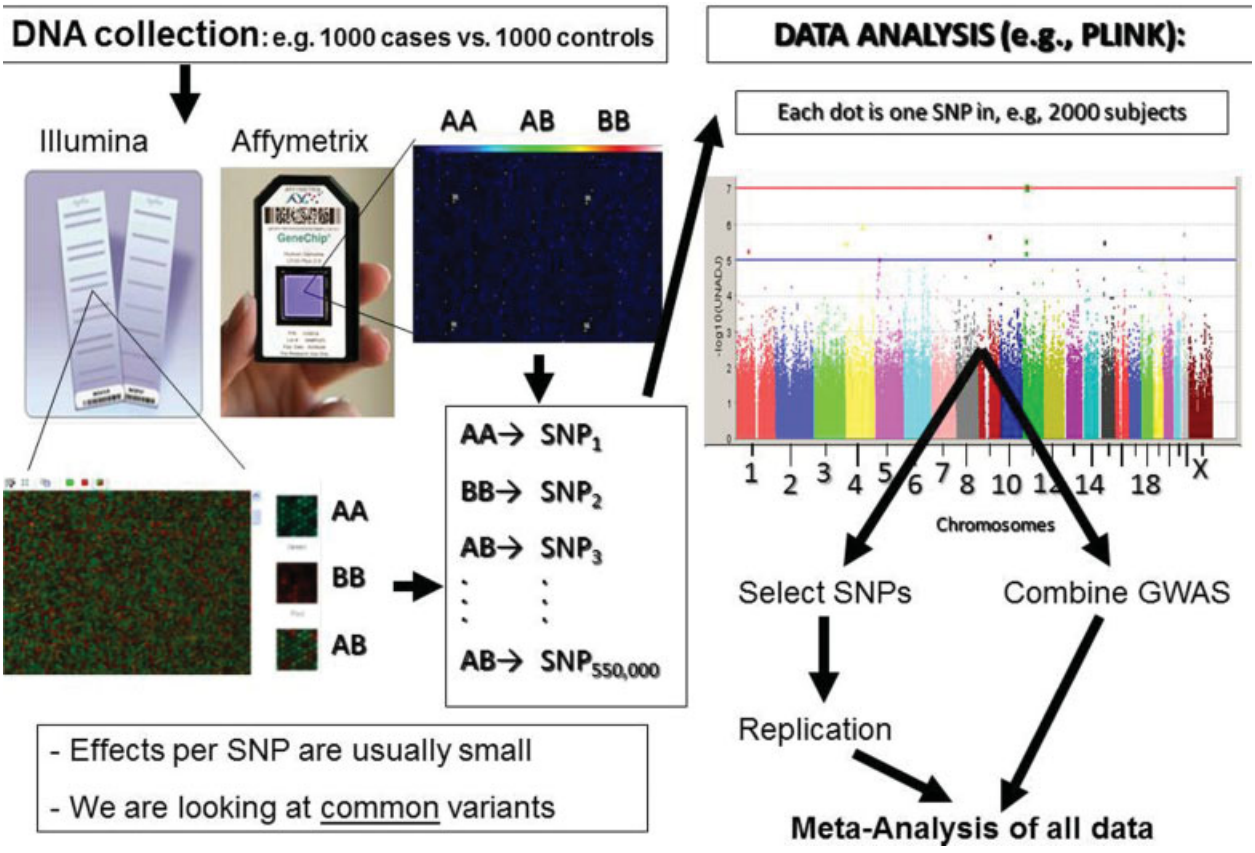

Fig. 2 Schematic representation of the various steps in a genome-wide association study (GWAS).

to many research groups working on this population and thereby giving good return on usually large investment to perform GWAS genotyping in such a population. Good examples of such cohort studies are the Rotterdam study and the Framingham study where hundreds of investigators studied many different diseases and phenotypes. The longitudinal design of cohort studies allows for the golden standard in case definitions in epidemiology, that is, the "incident" case (as opposed to the "prevalent" case) including preceding measurements of (quantitative) risk factors. Finally, as has been mentioned, cohort studies involve many subjects without overt disease and can therefore serve as controls in case–control studies. A source of information, but a potential pitfall, is the various moments in a longitudinal study that material was collected for DNA isolation. During aging, DNA will change (e.g., losing telomere sequence elements) and, thus, GWAS data determined from DNA collected at such various time points during the study should be analyzed taking this into account.

Genomic DNA for GWAS can come from many tissue sources, but usually DNA is isolated from blood and DNA from sputum, saliva, and cell lines has also been used. Quality and quantity requirements are not that strict for the GWAS genotyping technologies and even somewhat degraded DNA can be used and our own laboratory has had good GWAS genotyped results from as little as 500 pg of genomics DNA (Pascal Arp, BSc, personal communication, 2000). Yet, the sequence composition of germline genomic DNA can vary depending on the tissue and together with potential changes in genomic DNA with aging is a potential pitfall in GWAS when various sources of DNA are compared.

Genome-Wide Association Study Genotyping Arrays

Although the very first GWASs were done with home-made glass slides carrying approximately 100,000 SNPs, two commercial companies, Affymetrix and Illumina, have been dominating the GWAS genotyping market. Since 2005, they have been producing increasingly dense and smarter designed arrays containing up to several millions of SNPs. It is beyond the scope of this review to discuss the different array designs in detail and our laboratory is finishing such a comparison (Broer et al, unpublished data) including the most recent releases (e.g., UK Biobank [Affymetrix], and Omni-express and the Genome Screen Array [Illumina]).

The array design used by Affymetrix was based on picking random SNPs located close to each other to create a very dense pattern of SNPs covering the chromosomes which is good for physical genome coverage but might not always lead to optimal reflection of the LD blocks in imputation strategies (see later). Illumina on the other hand has focused on selecting SNPs that cover the genome and are also optimal in imputation strategies.

Also because of this array design, Illumina arrays are most widely used nowadays, although the recent arrays from Affymetrix also include optimal tagging SNPs performing quite well in imputation.

Genotyping with arrays is based on oligonucleotides specific for a small area (~50 base pairs) surrounding the SNP and that are attached to a glass base and then genotyping using fluorochrome labeling for detection. The steps in the genotyping process involve amplifying and preparing the target DNA, hybridization, and washing steps, and finally a detection step in ultrahigh definition lasers that can detect different fluorochrome-labeled products. This results in genotype files of several hundred gigabytes per sample requesting high-end storage and computing servers with high-speed connections. A typical GWAS database of approximately 10,000 samples requires approximately 15 terabytes of storage. Producing genotype data suitable for GWAS using the Illumina or Affymetrix arrays take about 1 week per DNA sample, but this process can be automated to a large extent allowing multiple DNA samples to be handled simultaneously. Especially, in high-throughput genotyping laboratories, such as our Human Genotyping Facility (www.glimdna.org), this results in capacities of producing GWAS genotype data of 10,000 DNA samples per week for Illumina genotyping arrays and up to 1,000 for Affymetrix genotyping arrays (per genotyping machinery unit of these companies).

Quality Control

The SNP genotype data are based on XY plots of the two fluorochromes specific for the two alleles per SNP. The control of genotype data involves analysis of cluster plots per SNP to evaluate how good heterozygotes are separated from the homozygotes. Especially, for more rare SNPs, this can be sometimes challenging, in particular, in studies of small sample size because the homozygous of the variant allele is lacking. The technical quality control (QC) of the genotype data produced involves checking overall genotyping success rate for all SNPs on the array per DNA sample and checking for overall genotyping success rate for all samples per SNP. Current arrays and procedures are such that these success rates are usually in the upper 90s (97–99% success rates) and result in very little dropouts per study, making this genotyping one of the most robust ones available.

After technical QC, the genotype data of the study samples are subjected to a genetic QC step, where SNPs are evaluated for Hardy–Weinberg equilibrium, and the study sample is evaluated for familial relationships and ethnic stratification. In addition, a "biobank" QC can be done to detect some sample swaps by comparing the XY sex as determined by the genotype data with what sex is listed in the database of the study. We usually detect approximately 2 to 5% mismatches per biobank which occasionally can be traced down to chromosomal abnormalities of the sex chromosomes (e.g., Klinefelter or Turner syndrome) but most often are due to sample swaps somewhere in the laborious procedure of having a biospecimen from a study participant enter the database and the freezers, and being subjected to GWAS genotyping.

Imputation

Once all these QC steps have passed successfully, the genotype data of the samples can be subjected to GWAS analysis of genotypes in relation to one or more phenotypes. But before that, a special step has been introduced that allows many more SNPs to be analyzed than there are on the GWAS genotyping array. This increase in the number of SNP is

achieved by a process called "imputation" which is based on "guessing the genotype" of adjacent SNPs from the actual SNP that was genotyped on the array. This "guess" is informed by all the SNP genotype data that are in large databases on the interrelationships of SNPs within LD blocks, such as in the HapMap database of several reference populations (mentioned earlier). These reference databases contain many more SNPs than there are on the genotyping arrays. Therefore, in the few hundred Hapmap Caucasian samples, 10 million SNPs are documented of which we know the exact LD structure. This LD structure allows the more limited number of SNPs (e.g., 500,000 SNPs) genotyped in the actual study sample, to be "imputed" to the HapMap reference genotypes and thereby guessing the genotypes of approximately 10 million SNPs in the actual study sample (see ►**Fig. 3**). Because the reference data are derived from only a few hundred samples, such imputations are limited to the more common alleles. Initial comparison studies have shown the imputations to be remarkably correct for common alleles. By now, the reference populations have increased in sample size and ethnic diversity, making imputations more widely used for ever lower allele frequencies including the 1,000 genomes project (http://www.1000genomes.org/) and most recently the Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org/).

Data Analysis

These resulting files containing genotypes of many millions of SNPs in several thousands of samples are then subjected to a statistical evaluation of association of genotype with the phenotype of interest. Genotypes can be coded as 0, 1, and 2 for subjects homozygous for the reference allele, heterozygous, and homozygous for the variant allele, and analyses can be dichotomous (disease yes or no) using logistic regression or using linear regression or quantitative traits, in any given statistical package that can handle large datasets. The type of analysis is relatively straightforward, but the datasets are (very) large and require sufficient computer power and storage capacities. The results for a GWAS analysis are typically plotted as *p*-values (for the association test) per SNP plotted as (1) observed *p*-values versus expected *p*-values in a QQ plot and (2) as a Manhattan plot. A QQ plot provides information on potential stratification issues in the dataset (early deviation from the line of identity), while a Manhattan plot provides exact *p*-values per SNP on each chromosome (see ►**Fig. 2**). Because the huge number of statistical comparisons is made, a correction for the level of significance is needed. Hence, a SNP with *p*-value of 5.10^{-8} or lower is only considered genome-wide significant, but further replication is warranted to look for consistency of the effect size and direction. In addition, SNPs with *p*-values between 1.10^{-6} and 5.10^{-8} are included in such replication efforts to see if some of them will become genome-wide significant. The value of 5.10^{-8} is derived from dividing 0.05 by 1 million independent tests based on the number of LD blocks. Therefore, 5.10^{-8} means $p = 0.05$, taking all the multiple testing into account. This *p*-value is used for analysis of the common SNPs (with minor allele frequency [MAF] $> 5\%$). With rarer MAFs $< 1\%$, the assumptions might no longer hold, and more stringent *p*-values might apply and the need for larger sample size and replication is even higher.

Initially, GWASs were done in a single discovery data collection and replication was done in additional data collections of similar size in which only a handful of SNPs were then genotyped. Subsequent meta-analysis of the top SNPs then showed which SNPs became genome-wide significant. Nowadays, however, several discovery GWAS datasets are meta-analyzed together and the top SNPs are then genotyped in additional collections for replication and again meta-analysis demonstrates which SNPs are genome-wide significant.

GWAS "hits" come in various shapes and sizes. The more straightforward ones are those found in strong LD with a functional variant, and in a certain gene that makes perfect sense in terms of the phenotypic association. The more

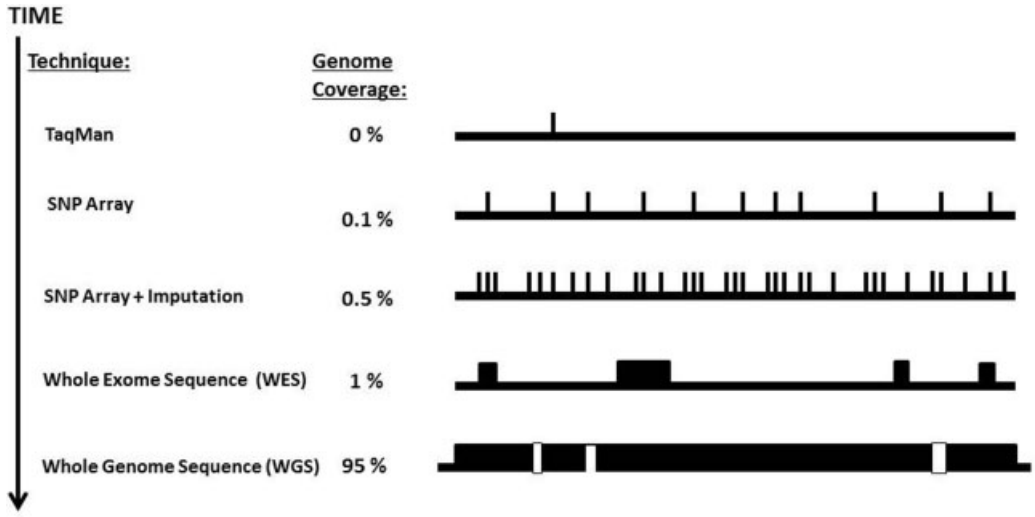

Fig. 3 Increased levels of genetic resolution in genome analysis by higher coverage of nucleotides analyzed by newer DNA analysis technologies (top, older technologies, bottom, the latest).

interesting GWAS hits are those that fall in an area of the genome that has no bioinformatics annotation whatsoever, and therefore, novel genome biology is described. We have learned that most GWAS hits point to regulatory regions (rather than protein-coding variants) and those fruits from GWAS labor have coincided with observations for large-scale projects looking at functional annotation of the genome, such as the ENCODE project (https://www.encodeproject.org/). Although ENCODE identified a regulatory region for a certain gene but without knowing what the biological role of the gene is, GWAS has provided some insight into biology by finding genetic associations with a certain phenotype of SNPs in that regulatory region which result in subtle changes in that regulatory function.

Once a SNP has been identified as being genome-wide significant, further detailed bioinformatics analysis takes place to look at the locus region in which the SNP is located and determine the exact LD block structure and which potentially functional SNPs can be identified in the LD block (such as coding variants, splice variants, and expression quantitative trait loci). These analyses can sometimes indicate which gene in the LD region is the most likely causative gene for the phenotypic association observed, but usually much more functional data are required to reach such a conclusion, involving, for example, expression analysis in target tissues, knockout mouse models, and studies in other animal models systems of all or most of the genes in LD area. Such functional characterization is much work per gene and per locus and has been experienced as a bottle neck in bringing GWAS hits further in translational medicine. But with the advent of projects such as ENCODE and large-scale mouse knockout databases, these steps in understanding the genetic association have been improved and fastened.

Large-Scale Collaboration in Consortia

The Human Genome Project in 2001 as well as the rise of GWAS since 2005 has also led to a change in doing scientific research among human geneticists and epidemiologists. Large-scale collaboration is now the norm and investigators are sharing ideas and data in an unprecedented fashion. This was originally born out of necessity because the effect sizes of SNPs on complex phenotypes are modest and so (very) large sample sizes in both the discovery phase and the replication phase were required to reach genome-wide significance of top hits. But because these collaborations were so successful and investigators realized that there were many more phenotypes and disease to be analyzed by GWAS, this has now changed into standard practice if a new phenotype is to be subjected to GWAS. This has resulted in the birth of many international consortia of collaborating investigators around a unifying theme, which is usually a disease (e.g., diabetes, osteoporosis) but sometimes also a study design (e.g., family-based studies, birth cohorts) or particular genomics data (e.g., epigenetics, exome sequencing data).

An additional, and also more important, advantage of these collaborations is that data are shared in a very early phase, as raw data and before the actual analysis is done. By doing so, many people look at, and critically assess, each other's data and results before they are analyzed together in a meta-analysis and subsequently get published. This will result in including both negative and positive results in the meta-analyses and will effectively correct and prevent mistakes (or even fraud) in the original datasets. Especially, in relation to scientific misconduct (fraud), we have seen several cases where scientists working in almost complete isolation have been given the opportunity to fabricate datasets and even publish about it. Within the international collaborations in human complex genetics, this has been effectively prevented by the way we assemble and analyze the data involving large-scale replication in the same publication as in which the discovery is described.

While this way of working together was initially limited to human complex genetics (although already very common place in physics or astronomy), we now also see that other fields of human medicine and biology are following the examples set by the GWAS community. Projects such as ENCODE and the mouse knockout database (http://www.mousephenotype.org/) are examples of this because they have been adapting to the speed of gene discovery of GWAS which has to be matched with high-throughput functional characterization of the genes identified to be associated with a disease or phenotype of clinical interest.

Examples of consortia include the Wellcome Trust Case Control Consortium (WTCCC; www.wtccc.org.uk) in the United Kingdom, which was the first large-scale GWAS effort in 2008 focusing on eight diseases involving many investigators collecting cases and using a common control dataset genotyped on one of the early commercial platforms (Affymetrix 2 $\times$ 250k arrays). Although WTCCC was underpowered (1,000 cases per disease and 1,500 controls) to find many loci, some initial successes were obtained and several lessons were learned. Most importantly, it showed that GWAS lends itself for collaboration and set the stage for the larger consortia. Examples of these early large consortia include GEFOS (www.gefos.org), a global collaboration focusing on the genetics of osteoporosis in 150,000 samples and which was fueled by funding from the European Union in the FP7 program following a successful FP5 project, GENOMOS. A good example of unfunded collaborations is the CHARGE consortium (www.chargeconsortium.com), which came about because simultaneously several large cohort studies of age-related disease in the elderly in the United States (Framingham, CHS, and ARIC) and Europe (AGES and the Rotterdam study) received national funding for GWAS in each of their individual cohorts. However, the investigators immediately realized that large-scale collaboration was needed for power reasons and for replication. Because these cohorts shared many similarities in their design and the many phenotypes they studied, a consortium was formed divided among phenotype working groups in which all relevant investigators are brought together who are expert on the phenotype in each cohort and together with genetic epidemiologists and molecular geneticists perform the meta-analyses.

Finally, in reproductive medicine, the ReproGen consortium (www.reprogen.org) originally started as a phenotype working group in the CHARGE consortium focusing on age at menopause and age at menarche, and is now also closely collaborating with the relatively new consortia dealing with

endometriosis, fibroids, polycystic ovary syndrome PCOS, and primary ovarian insufficiency (POI), which will be discussed elsewhere in this issue.

Pros and Cons of Genome-Wide Association Study

Successes

All GWAS efforts taken together over the past 10 years have been very successful in terms of producing gene discoveries in relation to diseases and phenotypes.[2] GWAS in 10 years has produced as many gene discoveries for common diseases as Mendelian genetics has done during the past 35 years for the rare Mendelian diseases. This has provided fundamental changes in our insight in the genetic architecture of the common complex diseases. We now know that many hundreds of common genetic variants each of modest effect size contribute to the heritability of all these traits and diseases. Most importantly, we know molecular details on which genes those are and in which phenotypes they play a role. Hits observed in GWAS have also pointed to base pairs in genome areas without any annotation so far in any database, and to further look for biological mechanisms underlying the genetic associations observed, thereby describing novel textbook biology. Although the effect sizes per SNP are modest, collectively, the SNPs identified up to now explain between 3 and 50% of the heritability's observed for these common diseases in twin studies and for which until 10 years ago the explained variance was 0%. Finally, GWAS has also set the example to follow how to do scientific research and produce high impact articles including replication of the original discovery, rather than publish a finding and wait until some other scientists replicate the finding or not, leading to confusing and contradicting scientific literature.

Limitations

However, the biological mechanism underlying most of the GWAS hits explaining the association with the phenotype remains unknown, which hinders translation to clinical medicine. This is due to several factors starting with the fact that a GWAS hit usually comprises a locus containing several, sometimes dozens of genes and thus simply choosing which genes to work on by investing time and money is challenging. Although the situation has improved over the past years, it was initially very hard to convince colleague investigators with more basic biology experience to start working on functional characterization of a locus or set of genes. One of the reasons for this was that the effect size observed for the association is usually modest leading investigators to think that the gene is not so important in the biological pathway and clinical translation would not be worthwhile. This is, however, a misconception because the GWAS is simply a methodology to identify a gene of interest (by exploiting human genetic variation), independent of the effect size of the genetic variant contributing to explain some of the population variance for that trait or disease. Investigators sometimes contrast GWAS with Mendelian genetics, in this respect, to state that the large effects seen in the rare Mendelian disorders illustrate the importance of a certain gene for a particular disease or trait.[1] But simply the fact that some genes are identified to be related to a certain disease or phenotype, by both Mendelian genetics and GWAS, shows that different roads all lead to Rome. Therefore, while these two genetic approaches are complementary, they also offer interesting new avenues for study (see later).

Since the explained variance by common SNPs for most phenotypes is limited so far, they are not useful in predictive medicine and this also limits their application in Mendelian randomization studies. However, so far, only SNPs are allowed in these models which have passed the threshold of genome-wide significance ($p < 5.10^{-8}$). We now know that there are many more common variants contributing to the explained variance of the trait or disease below this threshold and recent analyses show that the genetic contribution to population variation in height can be explained, to a large extent, by common SNPs.[3]

Another limitation of GWAS so far is the number of SNPs that is actually genotyped which usually represents 0.3 to 0.5% of all variant nucleotides in the human genome. Although this number is slightly higher for imputed datasets, it remains an underrepresentation of all variant nucleotides in a human genome. With better imputation reference datasets (as now the Haplotype Reference Consortium with 30K whole genome sequenced human genomes), this will be improved.

In addition, because large-scale international collaboration is necessary to achieve sufficient statistical power, one also selects for the identification of cosmopolitan or universal genetic variants (associated with a trait or disease). From the sequencing efforts taking place now, we learn that there are many millions of variants only seen in particular sample collections limited by ethnicity and by geographical origin. However, to find these genetic variants, sample sizes of similar magnitude are required as they are now being used on a global scale. Therefore, it is likely to take more time before we see those variants, which are mostly low-frequency variants, to be associated with phenotype and disease with sufficient statistical evidence.

Scientific Credit

Typical GWAS articles are accompanied by large multiauthor and multi-institution lists reflecting many scientists who have contributed with data from many sample collections to reach sufficient statistical power to discover genome-wide significant loci. The number of authors on such articles can vary from 20 up to 400 and usually contain different categories of contributions but which all follow the ICMJE guidelines in what constitutes a contribution that merits authorship. The contributions encompass generating genotype data in local sample collections, measuring phenotype data, running analysis of the local datasets and contribute this to the larger pool of data that will be meta-analyzed, roles in the meta-analysis (which is usually done at several centers to check for mistakes and discrepancies), participation in scientific discussions about the results and their interpretation, additional analyses and/or experiments in functional follow-up studies, roles in drafting an article, critically reviewing and commenting on articles. The scientists in the author group can consist of

young scientists in their early career, more senior scientists with specialization in aspects of GWAS, to more senior scientists supervising and coordinating studies. Although this way of giving authorship credit has been accepted by all scientific journals, this phenomenon has also been criticized by several colleagues because it supposedly also led to dilution of scientific credit of single authors, and also led to some authors appearing on many such multiauthor GWAS studies. In addition, the use of "starred" authors (with equal contributions to the study and article in such author position) is widespread for first authors, sometimes also second authors, and senior authors. This phenomenon is not new because it has been regularly practiced already in physics and astronomy, but for some reason, it has raised more discussion in the field of human medicine.

It is my personal view that we have to see this as a (positive) result of the new culture of doing science in our field of complex genetics, that is, large-scale studies involving massive amounts of data and much collaboration without one particular group or individual taking the credit as single first author or last author. It also is the result of such large-scale studies containing the immediate replication of a discovery result preventing false-positive study results and strongly working against data manipulation or even fraud. As stated earlier, all the authors have to fill in author contribution forms following the ICMJE rules and all the journals have followed this multiauthor model, including all major basic and medical journals.

Another positive aspect is the experience that a one-time investment of generating GWAS data in a sample collection has been shown to be useful for many different GWAS analyses (e.g., for different phenotypes all measured in the same sample collection) and thereby being able to contribute their data and help in several different GWAS publications, all describing novel gene discoveries. Especially, for long running cohort studies with many different phenotypes being measured in it (such as the Rotterdam study),[4] this has resulted in the Rotterdam study GWAS data being used in many different GWAS publications.

The Sequel of Genome-Wide Association Study: A Glimpse into the Future

Although geneticists initially thought that after the first round of GWAS successes its use would be diminishing, the opposite has happened. Still more and more studies are generating GWAS data in their biobanks and thereby even larger meta-analyses are performed for still more phenotypes and diseases. Especially, with the introduction of relatively inexpensive SNP arrays, such as the UK Biobank array from Affymetrix and the GSA array from Illumina, a new wave of very large-scale datasets has been generated, including the UK Biobank dataset with approximately 500,000 subjects with GWAS data, the U.S. veterans dataset with approximately 1 million subjects with GWAS data, and an expected large dataset of several millions of subjects with GWAS data with the GSA array from Illumina. This will result in a continuous activity of GWAS efforts to discover more genes and explain more genetic variance.

With the introduction of next-generation sequencing (NGS) since 2012, the focus in genetics has shifted toward the analysis of more rare variants with MAFs of $< 1\%$ (see ►**Fig. 2**). Most NGS projects so far have involved whole exome sequencing (WES) which provides all nucleotide variation in the coding part of the genome which represents approximately 1% of all nucleotides. Large-scale WES data can be found in the EXAC database (http://exac.broadinstitute.org/). Many more rare variants will come from discovery analyses in whole genome sequence (WGS) data that covers $> 95\%$ of all nucleotides. An example is the Genome of the Netherlands WGS data (http://www.nlgenome.nl/) derived from 500 independent subjects across Dutch Biobanks.

NGS has not replaced GWAS with SNP arrays but has been a welcome extension of the MAF spectrum to be evaluated for their contribution to common complex diseases and phenotypes. Disadvantages of working with NGS data are its high costs compared with SNP arrays, and the complexity and "noisiness" of the data so far preventing NGS to be widely applied. NGS has been (wisely) used to enrich especially reference databases (e.g., 1000 genomes, UK10K, and Haplotype Reference Consortium) with many more genetic variants to which particular data collections can be imputed to generate more variants to be analyzed in a regular GWAS. Another disadvantage of NGS data applied to complex genetics is the fact that discoveries require replication and with especially the rare variants being discovered, this requires even larger datasets for replication compared with the ones used for regular GWAS with common variants.

Nevertheless, application of such NGS data to the analysis of complex traits has provided already several interesting novel gene discoveries such as EN1 for bone mineral density[5] which, however, required up to 500,000 samples to test the association with fracture risk. With still more samples being subjected to NGS and even larger datasets becoming available with GWAS data, it can be expected that NGS will deliver more novel variants and thereby more explained variance.

An interesting aspect of applying NGS to large population-based datasets encompassing "normal" control subjects without overt disease is the ability to analyze the phenotype of such subjects for the carrier status of Mendelian mutations in heterozygous form and assess the phenotype of subjects carrying homozygous pathogenic recessive mutations or heterozygous dominant mutations. Such research is currently underway and will be very relevant for genetic diagnostics to distinguish truly pathogenic variants from variants occurring in the normal population without a clear phenotype. An interesting example is the recent work on the prion gene for which several pathogenic mutations were discovered to exist in a large NGS dataset of phenotypically normal subjects.[6]

References

1 Chong JX, Buckingham KJ, Jhangiani SN, et al; Centers for Mendelian Genomics. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. Am J Hum Genet 2015; 97(2):199–215
2 Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet 2012;90(1):7–24

3 Yang J, Bakshi A, Zhu Z, et al; LifeLines Cohort Study. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet 2015;47(10):1114–1120
4 Hofman A, Brusselle GG, Darwish Murad S, et al. The Rotterdam study: 2016 objectives and design update. Eur J Epidemiol 2015; 30(8):661–708
5 Zheng HF, Forgetta V, Hsu YH, et al; AOGC Consortium; UK10K Consortium. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 2015;526(7571): 112–117
6 Minikel EV, Vallabh SM, Lek M, et al; Exome Aggregation Consortium (ExAC). Quantifying prion disease penetrance using large population control cohorts. Sci Transl Med 2016;8(322):322ra9