Immunology
The Journal of cells, molecules, systems and technologies

British Society for immunology

IMMUNOLOGY REVIEW ARTICLE

The challenges, advantages and future of phenome-wide association studies

Scott J. Hebbring
Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI, USA

doi:10.1111/imm.12195
Received 03 October 2013; revised 16 October 2013; accepted 17 October 2013.
Correspondence: Scott J. Hebbring, Center for Human Genetics, Marshfield Clinic Research Foundation, 1000 North Oak Avenue, Marshfield, WI 54449, USA.
Email: HEBBRING.SCOTT@mcrf.mfldclin.edu
Senior author: Scott J. Hebbring

Summary

Over the last decade, significant technological breakthroughs have revolutionized human genomic research in the form of genome-wide association studies (GWASs). GWASs have identified thousands of statistically significant genetic variants associated with hundreds of human conditions including many with immunological aetiologies (e.g. multiple sclerosis, ankylosing spondylitis and rheumatoid arthritis). Unfortunately, most GWASs fail to identify clinically significant associations. Identifying biologically significant variants by GWAS also presents a challenge. The GWAS is a phenotype-to-genotype approach. As a complementary/alternative approach to the GWAS, investigators have begun to exploit extensive electronic medical record systems to conduct a genotype-to-phenotype approach when studying human disease – specifically, the phenome-wide association study (PheWAS). Although the PheWAS approach is in its infancy, this method has already demonstrated its capacity to rediscover important genetic associations related to immunological diseases/conditions. Furthermore, PheWAS has the advantage of identifying genetic variants with pleiotropic properties. This is particularly relevant for *HLA* variants. For example, PheWAS results have demonstrated that the *HLA-DRB1* variant associated with multiple sclerosis may also be associated with erythematous conditions including rosacea. Likewise, PheWAS has demonstrated that the *HLA-B* genotype is not only associated with spondylopathies, uveitis, and variability in platelet count, but may also play an important role in other conditions, such as mastoiditis. This review will discuss and compare general PheWAS methodologies, describe both the challenges and advantages of the PheWAS, and provide insight into the potential directions in which PheWAS may lead.

Keywords: electronic medical record; genome-wide association study; phenome-wide association study.

Introduction

Over the last decade, great technological breakthroughs in genomics have been achieved. Driven in part by the Human Genome[1,2] and the HapMap Project,[3,4] it is now possible to genotype over one million single nucleotide polymorphisms (SNPs) across the human genome in a single assay. These technologies have made it possible to genotype thousands of cases and controls for any disease and laid the foundation for the genome-wide association study (GWAS) (Fig. 1a). With these tools readily available, over 1000 GWASs have been published linking nearly 4000 statistically significant loci to over 500 human traits and diseases.[5] Crucially, the GWAS has been very

Abbreviations: eMERGE, electronic MEdical Records and GEnomics; EMR, electronic medical record; GWAS, genome-wide association study; ICD, International Classification of Disease; ICU, intensive care unit; MS, multiple sclerosis; NGS, next-generation sequencing; PAGE, Population Architecture using Genomics and Epidemiology; PheWAS, phenome-wide association study; RA, rheumatoid arthritis; SNP, single nucleotide polymorphism; WBC, white blood cell; WHO, World Health Organization effective in revealing that many immune-related diseases are genetically complex, including multiple sclerosis (MS), ankylosing spondylitis, psoriasis, rheumatoid arthritis (RA), and Crohn's disease.[5] One of the strengths of GWAS is its unbiased whole genome nature, giving GWAS the ability to identify novel genes and pathways linked to common and complex conditions. Surprisingly, many phenotypes of previously unappreciated aetiologies have been linked to immunological genes/pathways by GWASs.

The GWAS is limited by a variety of factors. Due to the burden of multiple comparisons testing, reaching the threshold of statistical significance by GWAS can be a challenge. To be considered 'GWAS significant', only those associations with a P < 5·0E-8 are considered statistically significant.[6,7] Even if GWAS-significant SNPs are identified, GWASs often fail to identify clinically relevant predictive associations and have difficulty explaining a significant portion of the predicted phenotypic heritability. Moreover, GWAS SNPs are predominantly 'tags' for common variants across the genome. Causative/functional polymorphisms, known or unknown, may be in partial linkage disequilibrium with the genotyped tag SNPs, resulting in observed weak effects. Lastly, the vast majority of GWAS-significant SNPs are intergenic. Identifying and characterizing functional polymorphisms in intergenic regions is a particular challenge. Addressing these challenges using alternative/complementary strategies is of great importance. Without such strategies, our grasp of 'genomic medicine' will remain elusive.

One alternative/complementary approach to GWAS is the phenome-wide association study (PheWAS); this is the same as a GWAS, but from a reverse perspective. Whereas a GWAS uses a phenotype-to-genotype approach, beginning with a specific phenotype that is associated with genetic variants across the genome, PheWAS reverses this paradigm by using a genotype-to-phenotype strategy. PheWAS can start with a genotype to test for associations over a wide spectrum of human phenotypes – the phenome (Fig. 1b). In 2010, the first PheWAS was published as proof-of-principle for the technique. This study demonstrated that a PheWAS strategy, compared with a GWAS strategy, could be applied to identify significant gene–disease associations. For example, it verified that rs3135388 and rs6457620, two SNPs in the *HLA* region, were associated with MS and RA, respectively.[8] Since the first PheWAS, five additional PheWASs have focused on genetic targets.[9–13] Importantly, four additional PheWASs have investigated non-genetic targets [e.g. white blood cell (WBC) count],[14–17] while another PheWAS assessed both genetic and non-genetic targets.[18] A commonality for most PheWASs is the use of an electronic medical record (EMR) to define the phenome.

An EMR provides an efficient data source for phenotype extraction, as it generally contains longitudinal

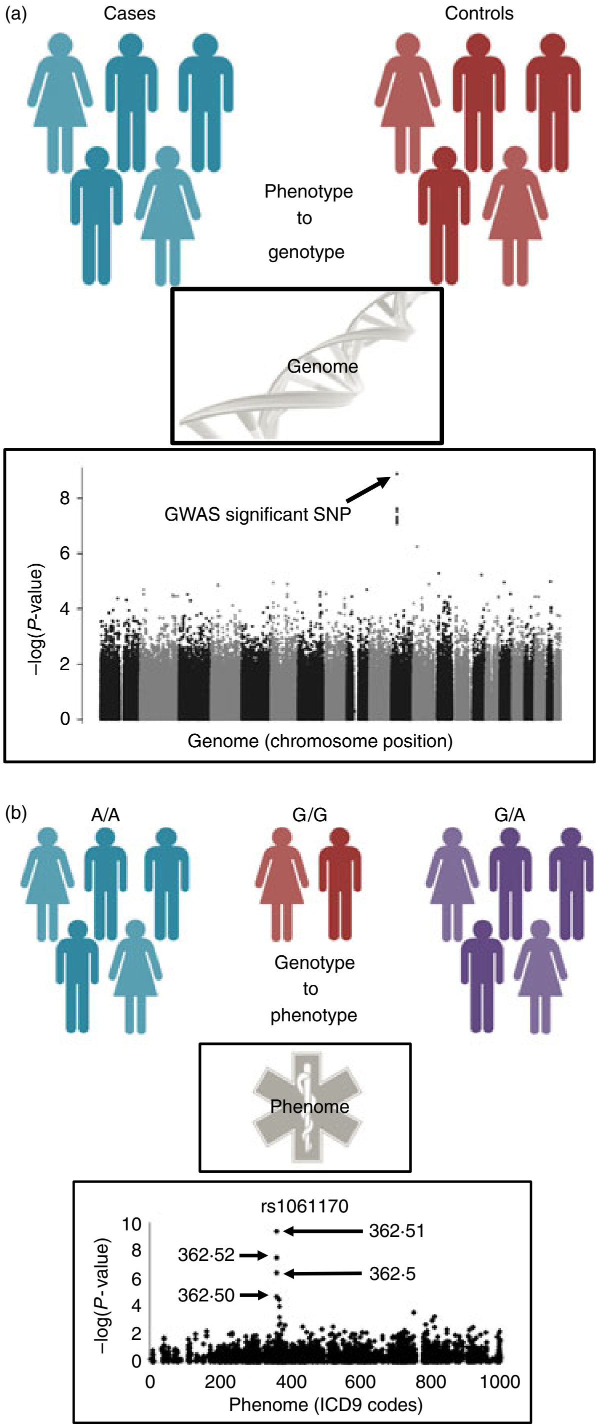

Figure 1. Representative results from (a) a genome-wide association study (GWAS) and (b) a phenome-wide association study (PheWAS). (a) Illustration of the phenotype-to-genotype strategy used by GWAS along with a representative GWAS depicted in the Manhattan plot graphing –log$_{10}$(P-values) across the genome. (b) Illustration of the genotype-to-phenotype strategy used by PheWAS including a representative PheWAS for single nucleotide polymorphism (SNP) rs1061170, a SNP known to be associated with age-related macular degeneration (AMD).[5] The AMD ICD9 codes are highlighted on the PheWAS Manhattan plot.

health histories including prescription records, family histories, laboratory and imaging test results, physician notes, procedure codes, and importantly, the International Classification of Disease (ICD) codes. ICD codes are a standardized, internationally recognized, coding system used to define disease status. The ICD codes were initially developed by Dr Jaxques Bertilon in 1893 to classify deaths caused by general diseases affecting specific anatomical sites. They are now managed by the World Health Organization (WHO), although different countries, including the USA, have their own ICD adaptations. In 1975, WHO adopted ICD version 9 (ICD9).[19]

In the USA, the implementation of ICD9 coding coincided with the 'Digital Revolution' and has been implemented in EMR systems for billing purposes.[20] Importantly, there are nearly 17 000 possible codes available in the ICD9 coding system,[21] offering a wide spectrum of phenotypic information at various levels of phenotypic resolution. For example, ICD9 code 714.33 defines monarticular juvenile RA, while 714.32 is pauciarticular juvenile RA. The highly specific 714.32, 714.33, and other related ICD9 codes can be consolidated into the 714.3 code, which represents the larger category defining juvenile chronic polyarthritis. Similarly, 714.3 and other 714.* codes can be combined into the 714 code defining RA and other inflammatory polyarthropathies. ICD9 code 714 is just one of many ICD9 codes that fall between ICD9 001 (cholera) and ICD9 999 (complications of medical care not elsewhere classified) (Fig. 2). Not surprisingly, ICD9 codes, with their wide spectrum of phenotypic information, have been employed to define the phenome in most PheWASs. This review will discuss such studies and compare general methodologies, describe the limitations and advantages of the PheWAS design, and provide insight into the potential direction PheWAS may lead.

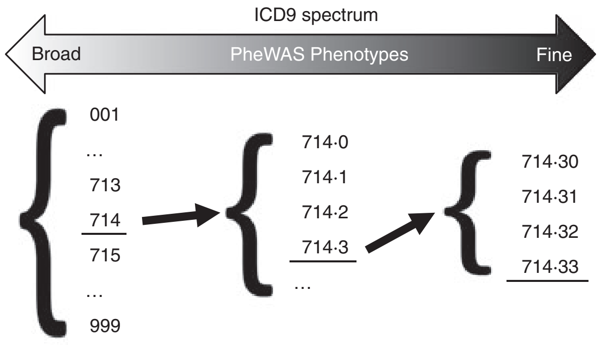

Figure 2. International Classification of Disease version 9 (ICD9) spectrum used to define the phenome in phenome-wide association studies (PheWASs). Underlined are ICD9 codes represented at varying levels of phenotypic resolution including rheumatoid arthritis (RA) and other inflammatory polyarthropathies (ICD9 714), juvenile chronic polyarthritis (ICD9 714.3), and monarticular juvenile RA (ICD9 714.3).

PheWAS methodologies

The first PheWAS was published in 2010 in *Bioinformatics* as proof-of-principle.[8] This study associated five genetic targets across a curated phenome built on the backbone of ICD9 codes and refined with clinical expertise. In general, high level ICD9 codes, including four and five digit ICD9 codes (e.g. 714.3 and 714.33, respectively), were condensed into a common three-digit code (e.g. 714). Under some scenarios, related three-digit codes were also condensed. Under other scenarios, where phenotypes shared very similar four-digit and five-digit ICD9 coding, but were deemed to be distinct diseases (e.g. type I and type II diabetes mellitus), higher order coding was maintained. ICD9 codes with unlikely genetic aetiologies were removed (e.g. contamination with foreign objects). At the simplest level, patients coded for a specific ICD9 code become 'cases' for that respective code. Those patients not coded for a specific ICD9 code become 'controls'. The advantage of using a curated phenome is that the complexity of ICD9 coding is reduced. As such, the power to detect an association may be increased by the increased number of cases and fewer possible phenotypes. By removing selected ICD9 codes that appear to have a strong environmental component, the multiple testing burden is reduced further. Conversely, this method is not without bias. Assumptions are made when combining ICD9 codes. Biases are further introduced when ICD9 codes thought to be unrelated to genetics are not analysed. Regardless, a curated phenome can be applied in PheWAS to validate expected GWAS results and to identify potential novel associations. For example, SNP rs6457620, an *HLA* SNP known to be associated with RA,[22] was associated with the ICD9 code for RA (ICD9 714). Interestingly, this SNP was also associated with pituitary gland and trigeminal nerve disorders. In addition to rs6457620, another *HLA* SNP was genotyped – specifically, rs3135388. Rs3135388 tags for *HLA-DRB1*1501* and is known to be strongly associated with MS.[5] When *HLA-DRB1*1501* (rs3135388) was genotyped, the ICD9 code that defines MS (ICD9 340) was highly associated with the SNP genotype. Novel associations were also reported, including an association between the *HLA-DRB1*1501* genotype and erythematous conditions (ICD9 695).[8] This example demonstrates how a PheWAS can complement previously reported GWASs and provide novel insights into diseases with unappreciated genetic aetiologies.

Another study that used a similar curated phenome was conducted within the electronic MEdical Records and GEnomics (eMERGE) Network.[23] Unique to this study, GWAS was used to inform PheWAS within the same cohort. GWAS results demonstrated that a common SNP near *FOXE1* (rs965513) was significantly associated with risk for hypothyroidism. *FOXE1*, a gene also known as thyroid transcription factor 2, has been implicated in a variety of thyroid diseases, including a rare form of syndromic congenital hypothyroidism.[24] Following GWAS analysis, rs965513 was assessed by PheWAS in the same cohort that was used to derive the hypothyroidism cases and controls for the GWAS. As expected, ICD9 codes that define hypothyroidism were significantly associated with the rs965513 genotype by PheWAS, but other thyroid-related conditions were also associated including non-toxic nodular/multinodular goitre and nutritional deficiency anaemia.[9]

Recently, this GWAS-informed PheWAS approach has also been applied to the study of platelet phenotypes. Using a similar eMERGE population as described above, this study identified 81 GWAS-significant SNPs including 56 SNPs associated with platelet count, 29 SNPs associated with platelet volume, and four SNPs associated with both. Many of these SNPs validate previously published GWAS results.[5] Each of the 81 SNPs was then individually associated with the phenome. For example, rs3819299, an intronic variant in the *HLA-B* gene, was associated with platelet count as expected.[5,25] PheWAS results for this SNP showed that the *HLA-B* genotype is also a risk factor for inflammatory/ankylosing spondylopathies and uveitis. The role of the *HLA-B* genotype in spondylopathies and uveitis has been previously described.[26,27] Importantly, a novel association was reported between the *HLA-B* genotype and mastoiditis.[13] Like the *FOXE1* example described previously,[9] this study demonstrates that when GWAS is used to inform PheWAS in the same population, an expanded understanding of the biological, and potentially clinical, significance for a SNP can be achieved. This GWAS-informed PheWAS approach has also been applied to the study of arrhythmia risk.[12]

As an alternative approach to a curated phenome, others have applied a holistic method by testing all ICD9 codes at multiple levels of phenotypic resolution. For example, patients coded for monoarticular juvenile RA (ICD9 714.33) define a unique case group whereas those coded for pauciarticular juvenile RA (ICD9 714.32) define another unique case group. To address the possibility that genetic aetiologies are shared between similar codes, all 714.3* codes can be combined into a case group defined by 714.3 and then further combined with like codes to form a separate 714 case group (Fig. 2). Depending on sample size and frequency restraints, this methodology can generate nearly 17 000 phenotypes.[21] The advantage of this method is that it does not make assumptions regarding the genetic or environmental contributions to any one disease. This is analogous to a GWAS where intergenic and coding variants are treated equally. The disadvantages of using a more holistic phenome include the potential for a reduction in power to detect an association because there are many more phenotypes with small case sizes – many without any genetic indices. Regardless, investigators have applied this simplified method to define the phenome, or variations thereof, with success during PheWAS. For example, a Marshfield Clinic patient cohort was genotyped for *HLA-DRB1*1501* as a follow up to the first PheWAS described previously.[8] As expected, the *HLA-DRB1*1501* genotype was associated with the ICD9 code for MS (ICD9 340). Importantly, *HLA-DRB1*1501* was also associated with the ICD9 code for erythematous conditions (ICD9 695). This is the first example where a novel PheWAS finding was independently validated. By applying a holistic approach to define the phenome, and taking advantage of the higher order phenotypes, it was revealed that the ICD9 code for rosacea (ICD9 695.3) may be driving the association results of the broader ICD9 code defining erythematous conditions. In addition, this study characterized a novel association between the *HLA-DRB1*1501* genotype and the ICD9 code that defines alcohol-induced cirrhosis of the liver (ICD9 571.2),[10] a phenotype that may have been disregarded in the original PheWAS due to the potential for a strong environmental component. In support of this novel PheWAS finding, previous GWASs have demonstrated that *HLA-DRB1*1501* is associated with drug-induced liver damage.[28,29]

Another example where an unbiased phenome has been applied in PheWAS was reported by Warner *et al.*[15] This study is unique in that it is the first PheWAS to use a non-genetic target, specifically WBC count. The goal of this study was to identify context-dependent associations between WBC count and ICD9 codes from patients in an intensive care unit (ICU). Expected associations were observed between elevated WBC count and ICD9 codes that define leukaemia, including chronic lymphoid (ICD9 204.10), acute myeloid (ICD9 205.00), and chronic myeloid (ICD9 205.10) leukaemia. The WBC count was also associated with diagnosis of *Clostridium difficile* infection, and these patients were at an increased risk for adverse effects because of observed delays in effective treatment and increased length of hospital stay.[15] This PheWAS result may help to alter the current standard of care and to reduce potential adverse effects for ICU patients with elevated WBC count and at a high risk for *C. difficile* infections.

ICD9 coding is useful when describing a spectrum of phenotypes, and as a result of its standardized structure and usage, PheWAS results can be combined or compared across institutions. Alternatively, other data types can be applied when defining the phenome. For example, the Population Architecture using Genomics and Epidemiology (PAGE) Network describes how diverse phenotypes collected from a wide variety of sources, including surveys and medical records, can be applied in PheWAS across multiple institutions. First described in 2011,[30] and further defined in early 2013,[11] the PAGE Network focused on 83 previously reported GWAS SNPs that had been genotyped in at least two of the five PAGE Network groups. Each study group conducted its own PheWAS on its own defined phenotypes. The number of phenotypes varied greatly between study groups. For example, 3363 phenotypes were described in the Women's Health Initiative, while 63 were described in the Multiethnic Cohort Study. All phenotypes with $P < 0{\cdot}01$ were manually grouped into 105 broadly defined standardized phenotypic classes (e.g. vitamin E levels) and compared across study groups to identify overlapping significant associations. This method demonstrated that 48% of expected genotype–phenotype associations could be directly validated by PheWAS, and another 23% represented associations closely related to previously reported genotype–phenotype associations. Importantly, 30% of the PAGE PheWAS results represented novel associations. For example, this PheWAS characterized a novel association between the IL6R (rs2228145) genotype with numbers of neutrophils and lymphocytes;[11] rs2228145 has been previously shown to be associated with C-reactive protein levels.[31]

Regardless of the methodology used to define the phenome, the PheWAS design has challenges. Some limitations are shared with GWAS while others are unique. Conversely, the PheWAS has unique advantages that make this approach a powerful complementary method for understanding the complexities of human disease.

Limitations of PheWAS

Like GWAS, the PheWAS is a hypothesis-generating approach that is challenged by multiple comparison testing. For example, if 17 000 phenotypes, the limit of ICD9 coding,[21] are tested for association with one SNP, an association with a $P < 2{\cdot}9\text{E-}6$ would be considered statistically significant if a Bonferroni correction and an experimental wide α of 0·05 is applied. This also assumes that only one SNP is analysed by PheWAS. Conversely, a Bonferroni correction may not be an appropriate multiple comparison adjustment because of the lack of independence across many phenotypes, especially in phenomes that use multiple levels of phenotypic resolution when defining individual case groups (e.g. ICD9 codes 714, 714.3, 714.33). The inter-relationship between ICD9 codes is depicted in the Manhattan plot of Fig. 1(b). Data come from an unpublished PheWAS for rs1061170 using an unbiased, holistic phenome, as described previously.[10] Rs1061170 is a non-synonymous SNP in the *CFH* gene and is known to be associated with age-related macular degeneration.[5] The most statistically significant ICD9 codes for this PheWAS include 362.52 ($P = 4{\cdot}1\text{E-}10$), 362·51 ($P = 3{\cdot}3\text{E-}8$), and 362·50 ($P = 4{\cdot}0\text{E-}7$) representing exudative, non-exudative and unspecified senile macular degeneration, respectively. Since ICD9 362.5 is a composite of all 362.5* codes, it is intuitive that this code is also significantly associated with the SNP genotype ($P = 2{\cdot}1\text{E-}5$).

Even though similar codes may co-segregate, divergent ICD9 codes may also be correlated for clinical and biological reasons. For example, it is conceivable that a patient with an ICD9 code for hypercholesterolaemia (ICD9 272.0) may also have codes that define atherosclerosis (ICD9 440), acute myocardial infarction (ICD9 410), and/or other related comorbidities. The challenge of multiple comparison testing may be further complicated by the ever-increasing granularity of phenotypes, such as the anticipated use of ICD10 coding in the USA.[21] Whereas the ICD9 system allows for nearly 17 000 possible codes, the ICD10 system allows for more than 155 000 different codes.[20] Multiple comparison methods that consider inter-relationships between phenotypes (e.g. permutation testing) may be more appropriate when measuring statistical significance by PheWAS.

The phenome is only as good as the phenotypes within. An advantage of using ICD9 codes to define the phenome is that it allows an investigator to comprehensively assemble a wide spectrum of phenotypes in an efficient and cost-effective manner. Unfortunately, not every phenotype/ICD9 code is equal. Each ICD9 code is highly variable in frequency and in its positive and negative predictive values.[32] It is impractical to manually assess the validity of all phenotypes coded for all patients. It is also unrealistic to develop sophisticated logic rules to describe every phenotype. To address this challenge, simplistic rules to define cases and controls can be applied, including the use of the 'rule-of-two'. The rule-of-two states that a patient must be coded two or more times for any ICD9 code to be considered a case. It is intuitive that the rule-of-two increases the positive predictive value, but may result in a reduction in case numbers.[32] As phenomes become ever more granulated, average case numbers will undoubtedly be reduced even further. This will dramatically affect the power to detect an association. More advanced high throughput methods, using data beyond ICD codes, may be helpful in the future when defining the phenome.

Once a phenome is constructed, association testing is often simplified to basic statistical approaches uniformly applied across the phenome. Small numbers of cases for many phenotypes restrict analyses to contingency table tests of independence (e.g. Fisher's exact test) precluding the use of covariates. But even if case groups are large enough for regression analysis, standard variables such as age and sex may be inappropriate to implement within a regression model. Age and sex may be covariates for some conditions and confounders for others. Caution should be applied when interpreting initial PheWAS results. Follow-up phenotype-specific analyses, which may or may not include covariates, may be warranted after an initial PheWAS screen. Like GWAS, it is important to validate any PheWAS association.

Validation approaches for PheWAS findings are similar in concept to GWAS. In a GWAS, either specific candidate SNPs can be validated in an independent sample set, or an independent GWAS can be conducted. In PheWAS, a specific phenotype can be validated in an independent study (e.g. case–control study), or an independent PheWAS can be applied. A specific case–control study design becomes hypothesis testing and reduces the multiple comparison testing burden. Conversely, with imperfect case definitions and small case group sizes in an initial PheWAS screen, an independent validation PheWAS may help to identify true associations that may not have been top candidates in the initial PheWAS.

Like GWAS, and perhaps even more so for PheWAS, differences across populations may affect the ability to validate findings. At the SNP level, it would not be unexpected to see very different GWAS results from a population with European ancestry compared with a population with African ancestry as a result of significant differences in the linkage disequilibrium structure and allele frequencies between the two populations. In a genetically driven PheWAS, there is often one SNP associated across the phenome. If the SNP genotyped is not the functional variant, and/or observed in multiple populations, replicating PheWAS results may be difficult. Differences across populations may also go beyond genetics. There may be significant differences in clinical care, and the use of ICD9 codes, that is observed in different ethnicities, between different physicians, and across different healthcare providers. These differences may change over time with changes in standard medical practice. Even with the challenges of PheWAS, the utility of PheWAS has been demonstrated and its application may have significant advantages.

Advantages of PheWAS

So far, the PheWASs focusing on genetic targets have concentrated on SNPs that were already identified by GWAS.[8–13,18] Even with the complex challenges described above, PheWAS has demonstrated its capacity to identify expected associations when going in the opposite direction compared with GWAS. The selection of a phenotype for GWAS is important for the success of any GWAS study. The selection of a marker, genetic or otherwise, for PheWAS is also very important. By focusing on variants that have known function and/or clinical significance, PheWAS has the advantage of simplifying the process by starting with the biology. If there is significant background information regarding a genetic variant's function and/or clinical significance, that information can be applied directly when interpreting novel PheWAS results. This is important because PheWAS has the distinct advantage of characterizing pleiotropic genetic variants.

As of July 2013, 1038 GWASs had identified 4870 significant associations (4018 unique SNPs) across the human genome ($P < 5{\cdot}0\text{E-}8$) for 547 human traits and diseases. Nearly 13% of these GWAS-significant SNPs (519 SNPs) are associated with two or more phenotypes, although some phenotypes are similar (e.g. triglyceride levels and risk for hypertriglyceridaemia).[5] Rs1260326 on chromosome 2 is an example where multiple phenotypes, some related and some not, are associated with SNP genotype at the GWAS level. A non-synonymous SNP in *GCKR*, rs1260326 was significantly associated with 12 phenotypes in 17 GWASs, including triglyceride phenotypes, metabolic networks, urate levels, C-reactive protein levels, total cholesterol, amino acid levels, serum albumin levels, non-albumin protein levels, chronic kidney disease, liver enzyme levels, 2-hr glucose challenge and platelet count.[5] This demonstrates that GWASs have the capacity to identify pleiotropic variants, but requires that multiple GWASs be performed.

PheWAS has the potential to measure a genetic variant's pleotropic property in a significantly more comprehensive and efficient manner than GWAS. This point is emphasized by the two PheWASs described previously where *HLA-DRB1*1501* was not only associated with MS, but also with erythematous conditions,[8] including rosacea.[10] Furthermore, other HLA variants, such as those described previously in the *HLA-B* gene, can be associated with platelet count, spondylitis, uveitis, and mastoiditis.[13] These examples further emphasize the capacity of PheWAS to study clinically significant diseases that may not otherwise be studied at the genetic level. Understanding the shared genetic aetiologies of multiple diseases by PheWAS, such as MS and rosacea, is a significant advantage and may provide significant insight into the pathophysiology of numerous conditions – insights that may lead to new treatment strategies while minimizing research costs. For example, medications effective for the treatment of rosacea may be effective for the treatment of MS. Minocycline is an antibiotic with anti-inflammatory properties that is commonly used to treat rosacea. Interestingly, small clinical trials have been conducted to assess the use of minocycline to treat MS and have produced generally positive outcomes.[33–36] It is the culmination of these advantages and challenges that will dictate how future PheWASs are conducted.

Future of PheWAS

PheWAS has the potential to induce new bioinformatic methodologies, result in new disease research opportunities, expand the use of bio-repositories and genetic data, and will hopefully result in clinically significant breakthroughs. In the short term, the number and type of genetic variants assessed by PheWAS will undoubtedly grow. Approximately 200 SNPs have been analysed by PheWAS with all being rooted to previously reported GWAS results.[9–13,18] The largest number of SNPs assessed in a single PheWAS (83 SNPs) comes from the PAGE

Network,[11] which falls short of the 4018 unique SNPs that have been associated with 547 human phenotypes to date.[5] Although it has been difficult to identify clinically significant associations by GWAS, continued focus on GWAS SNPs using higher-resolution phenotypes within a phenome may help to refine initial phenotype–SNP associations identified by GWAS. As such, PheWAS may be a complementary strategy to GWAS when searching for predictive genetic biomarkers in a clinical setting.

GWAS SNPs are logical starting points for PheWAS because of the availability of association data, but GWAS SNPs are predominantly tag SNPs, reside primarily in intergenic regions, and often have no known function. An alternative PheWAS approach could be to focus on functional variants (e.g. loss of function variants). The use of GWAS SNPs in PheWAS exploits known association/phenotypic data while functional variants would exploit known biological insights. For example, loss of function variants are more likely to be associated with a disease with a stronger effect size compared with other classes of variation.[37] In general, Mendelian diseases are primarily caused by loss of function mutations,[38] and loss of function variants of unknown clinical significance have similar evolutionary selective pressures as known disease-causing mutations.[39] Focusing on these functional variants in PheWAS would be analogous to a mouse knockout experiment in a human population without the need to artificially manipulate the human genome. This approach may provide novel gene–disease associations while expediting the process of understanding the biology when genetic function is presumed.

Another PheWAS experimental design may include a pathway-based approach. As mentioned previously, nearly 13% of all statistically significant GWAS catalogue SNPs are associated with two or more phenotypes.[5] Many of the pleotropic variants identified to date map to a region on chromosome 6 containing *HLA* genes, suggesting that immunological pathways may play a very important role in many disease aetiologies. Of the 547 phenotypes curated in the GWAS catalogue, 5% have at least one GWAS significant marker that maps to a 5 Mb region containing *HLA* genes. This is significant because this region makes up less than 1% of the human genome. Very divergent phenotypes map to this region including drug-induced liver injury, MS, Stevens–Johnson syndrome, chronic obstructive pulmonary disease, and narcolepsy.[5] As demonstrated by the two *HLA-DRB1* PheWAS examples mentioned previously,[8,10] this region, and variants within, would be a logical target to identify novel association with pleiotropic effects that may have immunological origins.

The number of potential PheWAS targets is only limited by the number of known genetic variants. An alternative to a focused candidate SNP or candidate pathway approach would be a PheWAS for every SNP across the genome. This strategy would be unbiased to both the phenotypes and genotypes, but would also have a tremendous multiple comparison burden. The threshold for statistical significance would be orders of magnitude below a GWAS significant *P*-value ($P < 5{\cdot}0\text{E-}8$). Conversely, a PheWAS-by-GWAS approach could be very powerful when developing disease–disease, disease–gene, and gene–gene interaction networks. Similar studies have been conducted using combined GWAS data,[40] but are limited in their ability to address direct disease–disease interactions when many independent populations are used. Network analysis from a single PheWAS-by-GWAS in a large cohort may address this challenge and further describe complex interactions across the phenome and genome.

While a great deal of genetic data can be captured by focused and/or array-based SNP genotyping platforms, the future of human genetics will rely heavily on next-generation sequencing (NGS). Using NGS data for PheWAS-by-GWAS analysis, a whole order of complexity will be introduced. This may include the incorporation of indels, complex rearrangements, and copy number variants in the context of PheWAS. Furthermore, NGS will result in the identification of many rare variants, many of which may or may not have obvious functions within their respective genes. Incorporating rare variants in the context of PheWAS (e.g. gene-specific burden testing) should be possible. This reality is quickly coming to fruition as NGS is incorporated into standard medical practices.[41,42]

Although genetic variables served as the focus of the initial PheWASs, the future of PheWAS is not limited to genetics. PheWAS may assess associations between non-genetic targets and phenotypes across the phenome. Liao *et al.*[18] associated autoantibody levels to the phenome in patients with or without a diagnosis of RA. This was done with four different autoantibodies, including anti-citrullinated protein, anti-nuclear, anti-tissue transglutaminase, and anti-thyroid peroxidase antibodies. Anti-thyroid peroxidase antibody levels were associated with hypothyroidism, as expected. Moreover, novel statistically significant associations were observed when anti-nuclear antibody levels were associated with the phenomes of RA and non-RA patients. Anti-nuclear antibodies are known to be associated with genetic risk factors linked to systemic lupus erythematosus.[43] In RA patients, high-titre antinuclear antibodies were associated with Sjögren's/sicca syndrome. In non-RA patients, this antibody was associated with chronic non-alcoholic liver diseases.[18] Whereas Liao *et al.* assessed autoantibodies in a PheWAS, Warner et al[16] conducted a PheWAS by comparing the frequency of diagnostic codes between multiple myeloma patients in the ICU and a larger ICU patient cohort. This study characterized treatment-related and disease-related complications of multiple myeloma over time. These and other examples[14,15,17] demonstrate how PheWAS can be applied to identify potentially novel/clinically significant associations using non-genetic PheWAS candidates. PheWASs that use non-genetic targets are not constrained by DNA availability. Using pre-existing de-identified clinical data may allow investigators to circumvent the difficulties in patient recruitment and avoid the significant resources needed for genotyping. Furthermore, lack of dependence on genetics may allow for exploitation of significantly larger patient cohorts. This may address some of the challenges described above, most notably, small sample size for many phenotypes within a phenome.

Regardless of approach, PheWAS will always be limited by how well the phenome can be defined. Efforts to reliably define phenotypes using EMR data have been limited to specific phenotypes.[23] Although these disease-specific methods can reliably discriminate between cases and controls, they do not provide a high throughput mechanism to define the thousands of phenotypes within a phenome. Automated medical informatic tools capable of reliably defining the phenotypes within a phenome will be required. This may include machine learning methods that are able to incorporate various types of data beyond simple ICD9 coding, including laboratory values, procedure codes, physician notes, and/or prescription records. As the number of phenotypes increases and the phenotypic resolution becomes ever more accurate and specific, sample sizes for cases will shrink without expanded cohorts. This will require multi-institutional collaborative networks working together to develop, implement and apply PheWAS. As the use of an EMR becomes standard practice, bio-repositories continue to grow, and genomic medicine becomes readily applied in a clinical setting (e.g. NGS), PheWAS has the potential to unlock novel discoveries that would not be possible otherwise.

Acknowledgements

The author gratefully acknowledges the support from the Marshfield Clinic Research Foundation and appreciates Dr Joshua Denny's invaluable comments when reviewing this manuscript. The author would also like to thank Rachel Stankowski for her assistance in editing this manuscript. The following grants help support the published and unpublished work cited in this manuscript: NLM grant 5T15LM007359, NIGMS grant R01GM097618, NCATS grant UL1TR000427 and NCRR grant 1U1RR025011.

Disclosures

The author declares no conflicts of interest.

References

1 Lander ES, Linton LM, Birren B *et al.* Initial sequencing and analysis of the human genome. *Nature* 2001; **409**:860–921.

2 Venter JC, Adams MD, Myers EW *et al.* The sequence of the human genome. *Science* 2001; **291**:1304–51.

3 International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005; **437**:1299–320.

4 Frazer KA, Ballinger DG, Cox DR *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**:851–61.

5 Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. [accessed on 24 July 2013].

6 McCarthy MI, Abecasis GR, Cardon LR *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008; **9**:356–69.

7 Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273**:1516–7.

8 Denny JC, Ritchie MD, Basford MA *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010; **26**:1205–10.

9 Denny JC, Crawford DC, Ritchie MD *et al.* Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am J Hum Genet* 2011; **89**:529–42.

10 Hebbring SJ, Schrodi SJ, Ye Z, Zhou Z, Page D, Brilliant MH. A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun* 2013; **14**:187–91.

11 Pendergrass SA, Brown-Gentry K, Dudek S *et al.* Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* 2013; **9**:e1003087.

12 Ritchie MD, Denny JC, Zuvich RL *et al.* Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 2013; **127**:1377–85.

13 Shameer K, Denny JC, Ding K *et al.* A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum Genet* 2013. doi: 10.1007/s00439-013-1355-7.

14 Boland MR, Hripcsak G, Albers DJ *et al.* Discovering medical conditions associated with periodontitis using linked electronic health records. *J Clin Periodontol* 2013; **40**:474–82.

15 Warner JL, Alterovitz G. Phenome based analysis as a means for discovering context dependent clinical reference ranges. *AMIA Annu Symp Proc* 2012; **2012**:1441–9.

16 Warner JL, Alterovitz G, Bodio K, Joyce RM. External phenome analysis enables a rational federated query strategy to detect changing rates of treatment-related complications associated with multiple myeloma. *J Am Med Inform Assoc* 2013; **20**:696–9.

17 Warner JL, Zollanvari A, Ding Q, Zhang P, Snyder GM, Alterovitz G. Temporal phenome analysis of a large electronic health record cohort enables identification of hospital-acquired complications. *J Am Med Inform Assoc* 2013. doi:10.1136/amiajnl-2013-001861.

18 Liao KP, Kurreeman F, Li G *et al.* Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non-rheumatoid arthritis controls. *Arthritis Rheum* 2013; **65**:571–81.

19 World Health Organization. History of the development of the ICD. Geneva, Switzerland: World Health Organization, 2013. Available at: http://www.who.int/classifications/icd/en/HistoryOfICD.pdf [accessed on 26 September 2013].

20 Centers for Medicare and Medicaid Services. ICD-10. Baltimore (MD): CMS.gov; 2010 [updated September 9, 2013]. Available at: http://www.cms.gov/Medicare/Coding/ICD10/index.html?redirect=/icd10.

21 CMS Office of Public Affairs. HHS Proposes Adoption of ICD-10 Code Sets and Updated Electronic Transaction Standards. HHS.gov: U.S. Department of Health & Human Services; August 15, 2008. Available at: http://www.hhs.gov/news/press/2008pres/08/20080815a.html.

22 Raychaudhuri S, Remmers EF, Lee AT *et al.* Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat Genet* 2008; **40**:1216–23.

23 McCarty CA, Chisholm RL, Chute CG *et al.* The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011; **4**:13.

24 Park SM, Chatterjee VK. Genetics of congenital hypothyroidism. *J Med Genet* 2005; **42**:379–89.

25 Gieger C, Radhakrishnan A, Cvejic A *et al.* New gene functions in megakaryopoiesis and platelet formation. *Nature* 2011; **480**:201–8.

26 Evans DM, Spencer CC, Pointon JJ *et al.* Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat Genet* 2011; **43**:761–7.

27 Martin TM, Rosenbaum JT. An update on the genetics of HLA B27-associated acute anterior uveitis. *Ocul Immunol Inflamm* 2011; **19**:108–14.

28 Lucena MI, Molokhia M, Shen Y *et al.* Susceptibility to amoxicillin-clavulanate-induced liver injury is influenced by multiple HLA class I and II alleles. *Gastroenterology* 2011; **141**:338–47.

29 Singer JB, Lewitzky S, Leroy E, *et al.* A genome-wide study identifies HLA alleles associated with lumiracoxib-related liver injury. *Nat Genet* 2010; **42**:711–4.

30 Pendergrass SA, Brown-Gentry K, Dudek SM *et al.* The use of phenome-wide association studies (PheWAS) for exploration of novel genotype–phenotype relationships and pleiotropy discovery. *Genet Epidemiol* 2011; **35**:410–22.

31 Jiang CQ, Lam TH, Liu B *et al.* Interleukin-6 receptor gene polymorphism modulates interleukin-6 levels and the metabolic syndrome: GBCS-CVD. *Obesity (Silver Spring)* 2010; **18**:1969–74.

32 McCarty CA, Mukesh BN, Giampietro PF, Wilke RA. Healthy People 2010 disease prevalence in the Marshfield Clinic Personalized Medicine Research Project cohort: opportunities for public health genomic research. *Pers Med* 2007; **4**:183–90.

33 Chen X, Ma X, Jiang Y, Pi R, Liu Y, Ma L. The prospects of minocycline in multiple sclerosis. *J Neuroimmunol* 2011; **235**:1–8.

34 Metz LM, Li D, Traboulsee A *et al.* Glatiramer acetate in combination with minocycline in patients with relapsing–remitting multiple sclerosis: results of a Canadian, multicenter, double-blind, placebo-controlled trial. *Mult Scler* 2009; **15**:1183–94.

35 Metz LM, Zhang Y, Yeung M *et al.* Minocycline reduces gadolinium-enhancing magnetic resonance imaging lesions in multiple sclerosis. *Ann Neurol* 2004; **55**:756.

36 Zabad RK, Metz LM, Todoruk TR *et al.* The clinical response to minocycline in multiple sclerosis is accompanied by beneficial immune changes: a pilot study. *Mult Scler* 2007; **13**:517–26.

37 Chen R, Davydov EV, Sirota M, Butte AJ. Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS ONE* 2010; **5**:e13574.

38 Stenson PD, Mort M, Ball EV *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med* 2009; **1**:13.

39 Abecasis GR, Altshuler D, Auton A *et al.* A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**:1061–73.

40 Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011; **12**:56–68.

41 Rehm HL, Bale SJ, Bayrak-Toydemir P *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* 2013; **15**:733–47.

42 van El CG, Cornel MC, Borry P *et al.* Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics. *Eur J Hum Genet* 2013; **21**:S1–5.

43 Chung SA, Taylor KE, Graham RR *et al.* Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production. *PLoS Genet* 2011; **7**:e1001323.