CellPress
REVIEWS

Review

Genomic and Phenomic Research in the 21st Century

Scott Hebbring[1,*]

The field of human genomics has changed dramatically over time. Initial genomic studies were predominantly restricted to rare disorders in small families. Over the past decade, researchers changed course from family-based studies and instead focused on common diseases and traits in populations of unrelated individuals. With further advancements in biobanking, computer science, electronic health record (EHR) data, and more affordable high-throughput genomics, we are experiencing a new paradigm in human genomic research. Rapidly changing technologies and resources now make it possible to study thousands of diseases simultaneously at the genomic level. This review will focus on these advancements as scientists begin to incorporate phenome-wide strategies in human genomic research to understand the etiology of human diseases and develop new drugs to treat them.

Highlights

Human genetic research has morphed from studying rare disorders in families to common conditions in populations of unrelated individuals.

Advances in biobanking, computer science, and EHRs linked to real-life clinical data have allowed for the study of thousands of diseases simultaneously via phenome-wide association studies.

As genomic studies continue to grow in size and scale, both rare and common diseases may be studied simultaneously at the genome-wide and phenome-wide level, which may help understand the genetic causes of human diseases and ways to treat them.

A Genomic Perspective

We live in a rapidly advancing technological age driven by ever-increasing computational power. In genetics research, high-throughput computing, high-throughput genomics, and biobanking resources comprising 'big data' have become increasingly important. The goal of this review is to describe the current trajectory of human genetic research in the context of advances in phenome-wide research; an innovative field of research that can often be compared to genome-wide research. However, before one can discuss future trajectories in human genomic research, it is always important to evaluate the trends of the past.

According to the Online Mendelian Inheritance of Man, there are nearly 4000 genes with phenotype-causing variants[i] [1]. Most of these discoveries were made possible with family-based study designs, which are effective in identifying rare variants with large effect sizes [2,3]. Family-based study designs can include a variety of familial relationships (e.g., twins, sib-pairs, mother–father–child trios, and extended families). There is also a multitude of statistical tests that can be applied to family-based studies, such as segregation and **linkage analyses** (see Glossary). Segregation analysis does not require genetic data but can be used to inform the likelihood that a disease is heritable and provide insights into the genetic mode of inheritance of the disease. If a disease appears to be heritable and the mode of inheritance can be accurately defined, classical parametric linkage tests can identify regions of the genome that co-segregate with the disease in families. These analyses are robust under population substructure and are useful for monogenic disorders. Linkage tests are less effective in the presence of locus heterogeneity, when the inheritance model is unknown, and for low penetrant variants, qualities often connected to complex diseases. An alternative to classical linkage analysis in family-based designs can include transmission disequilibrium tests. These tests can detect linkage in the presence of a genetic association and can have more statistical power in some instances, as shown by Risch and Merikangas [4]. A significant challenge with any family-based study design is recruiting informative families to study the most interesting phenotypes. Over the past

[1]Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI 54449, USA

*Correspondence: hebbring.scott@marshfieldresearch.org (S. Hebbring).

decade, studies such as that by Risch and Merikangas [4], advancements in high-throughput genotyping, and ease of recruiting unrelated individuals, have shifted human genomic research from family-based studies to studies of large numbers of unrelated individuals. More specifically, research groups have gravitated towards the **genome-wide association study** (GWAS) (Figure 1, Key Figure).

One of the first GWASs published was in 2005, and described the genotyping of 96 cases with age-related macular degeneration (AMD) and 50 unaffected controls [5]. This study, along with two others published simultaneously in the journal *Science* [6,7], was able to map a common variant in *CFH* that was associated with AMD, which validated previous findings first discovered by family-based linkage analyses [8–11]. This result supported a popular hypothesis that common traits are largely influenced by a few common variants [12–14] and highlighted the inherent strength of the GWAS technique to study populations of unrelated individuals. Since then, the GWAS technique has been highly effective in associating common variants with common diseases. As of 2018, GWASs have identified over 50 000 candidate **single nucleotide polymorphism** (SNP)–disease/trait associations ($P < 1^{*}\ 10^{-5}$)[ii] [15,16]. Incidentally, all these GWAS results support an iteration of the common disease common variant hypothesis, where many variants with weak effects cumulatively contribute to disease risk [17,18]. Unfortunately, an unexpected shortcoming of the GWAS approach was the challenge in extracting biological inferences from GWAS results. Whereas family-based approaches are well suited to identify rare variants with large effects that often cause perturbations of protein-coding sequences, most variants identified by GWAS are noncoding and may influence genes that are kilobases away from candidate variants. This challenge is further compounded by **linkage disequilibrium** (LD) [19], a phenomenon where alleles nonrandomly associate with one another at two or more loci in a population, making it more difficult to identify causative variants.

To better understand human disease, larger GWASs have been conducted to overcome the impact of common variants with ever smaller effect sizes, to account for rare variants, and to overcome stringent corrections for multiple hypothesis testing. For example, the first GWAS of AMD described above evaluated 146 case-controls and was able to identify one statistically significant signal [5]. Only 11 years later, the largest GWAS of AMD to date genotyped nearly 16 000 cases and 18 000 controls to identify 52 statistically significant and independent variants, including rare variants [20]. Such GWASs are not feasible without highly collaborative consortiums and still require significant expense in patient recruitment and genomic data acquisition. This is particularly relevant as next-generation sequencing technologies replace SNP array platforms. Using **biobanks** with pre-existing genomic and phenomic data may help expedite such studies.

Biobanks in Genomic Research

A biobank is a collection of stored biological specimens. This may include residual tissues from clinical care saved for legal purposes or collected directly for research. Arguably, one of the most valuable biobanks in genomic research has included material gathered from Centre d'Etude du Polymorphism Humain (CEPH) families [21]. Established in 1984, lymphocytes from CEPH families were collected as a reference set for human genome mapping. Lymphocytes were immortalized so that DNA could be obtained in perpetuity. This DNA was some of the first to be used for mapping the genome with microsatellites [22,23], SNPs [24,25], and at the single base-pair level [26,27]. A significant advantage of a biobank is its use in future research that may not have been conceivable when the samples were first banked. For example, the immortalized lymphocytes initially designed to maintain DNA stocks from CEPH families have been developed into model systems for quantitative trait loci mapping of epigenomic, transcriptomic, and

Glossary

Biobank: a tissue repository collected via clinical care and/or research.
Electronic health record (EHR): an electronic data repository of health records collected via routine clinical care.
Genome-wide association study (GWAS): a general study design or statistical methods intended to identify genetic variants associated with a phenotype. In a case-control study, each variate is evaluated to determine whether the allele frequency is different in one group compared with the other. Although not exclusive, GWASs are often applied to populations of unrelated individuals.
Linkage analysis: a family-based statistical method that evaluates whether genetic markers co-segregate with disease in families. There are numerous types of linkage analysis that can be applied to different family structures and disease models.
Linkage disequilibrium (LD): a measure of correlation within a population between two genetic markers. Genetic variants are in LD most often when two loci are physically in close proximity to one another. Given that they are close together, few recombination events occur between the markers and they co-segregate during meiosis.
Phenome-wide association study (PheWAS): a general study design or statistical method intended to identify phenotypes associated with a genotype. PheWASs often rely on in-depth phenotypic data, including EHR or epidemiologic data, to define case-control status for a variety of phenotypes.
Principal component analysis (PCA): a statistical method often applied in epidemiological research to account for population substructures.
Single nucleotide polymorphism (SNP): one of the most common types of genetic variation in the human genome that results in a single nucleotide point change in DNA sequence. SNPs are often measured for mapping human diseases.

Key Figure

Schematic Representation of a Genome-wide Association Study (GWAS) (blue) and Phenome-wide Association Study (PheWAS) (red) Relative to a GWAS-by-PheWAS (green)

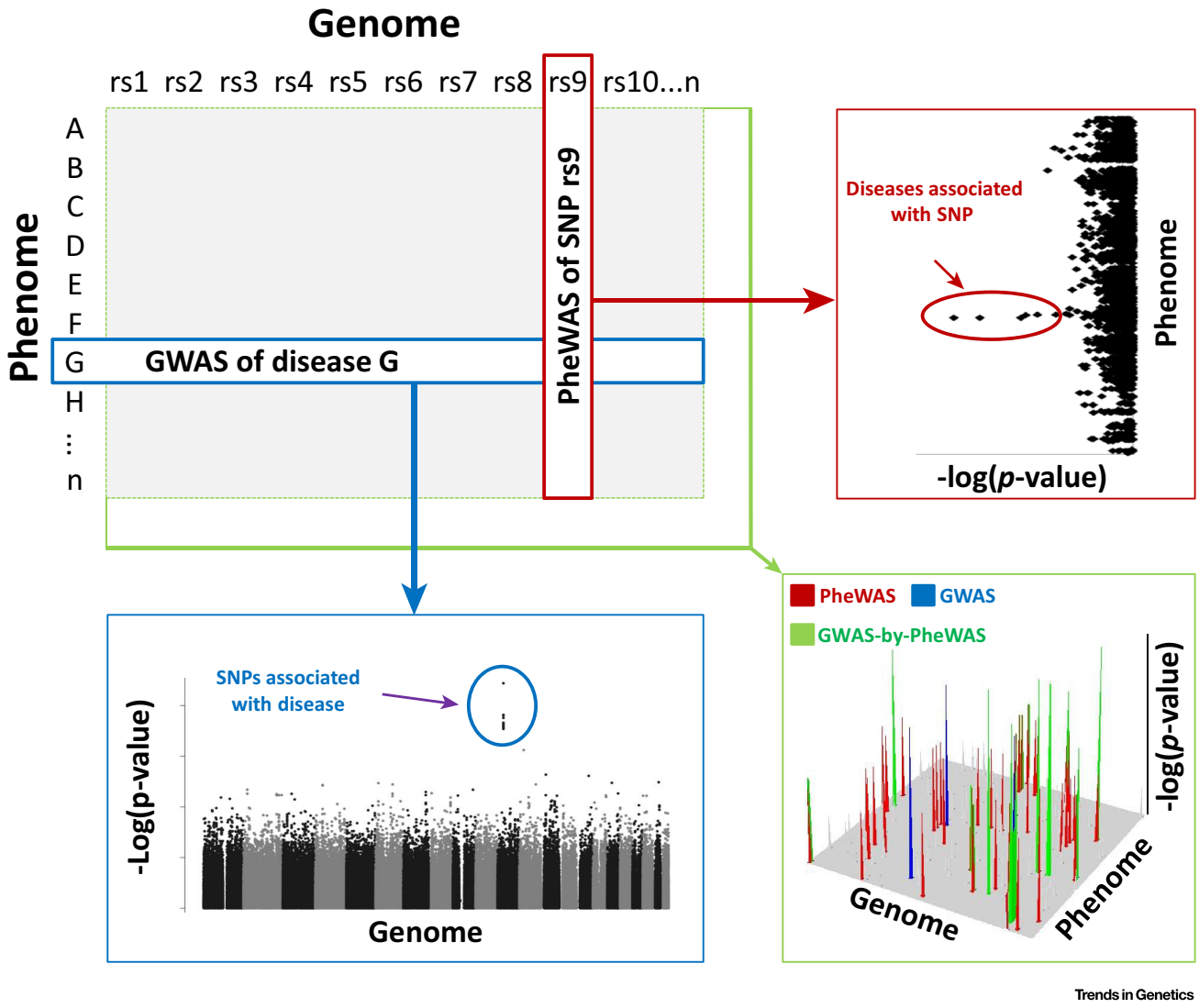

Figure 1. Included are examples of corresponding results and significant findings.

proteomic variants [28–31]. These cell lines have also been used as a model system to map drug response phenotypes [32–35]. What the CEPH biobank lacks, and many others like it, is a connection to extensive disease information necessary for mapping human traits.

With the computer age, extensive disease information can now be collected and stored in an EHR. An EHR can include diagnostic/billing codes, prescription records, laboratory results,

clinical notes, family histories, images, and other clinically relevant data. Clinical information in an EHR is often updated in real-time and does not require patient interactions via a research protocol beyond initial recruitment. Recognizing the importance of EHR data, scientists have created large DNA biobanks linked to extractable clinical data. Multiple examples of such biobanks exist, including academic healthcare institutions aligned with the electronic MEdical Records and GEnomics (eMERGE) Network [36–38], UK-Biobank [39,40], and deCODE [41]. Larger biobanks linked to medical record data are currently at varying stages of development, including the National Institute of Health 'All of Us' project (formally known as Precision Medicine Initiative Cohort Program; >1 million participants) [42] and the Million Veteran Program (>1 million participants) [43]. EHR data can provide an efficient mechanism to identify cases and controls for disease-specific research. For example, it may take many years to recruit thousands of patients with type 2 diabetes mellitus (T2DM). Conversely, it may take a fraction of the time to develop a T2DM prediction algorithm to identify cases and controls in an EHR-linked biobank based on T2DM diagnostic codes, T2DM medication records, and/or fasting glucose/HbA_{1c} test results [44]. Replicating this algorithm in another EHR system is even faster, on the order of days to weeks, once the algorithm has been validated. With sufficient time and subject matter expertise, and depending on the complexity of the input data, these algorithms can be developed to parse cases from controls with favorable predictive values. In the future, machine-learning techniques in a complex EHR environment may expedite algorithm development and reduce the need for subject matter expertise [45]. Lastly, EHR-linked biobanks with pre-existing genomic data can further reduce project costs and can be repurposed to study other diseases (Figure 2).

Whereas the extractable EHR data can be the greatest asset to a biobank, they are also its greatest limitation. For example, diagnostic codes, most notably the International Classification of Disease (ICD) codes that can be used to identity patients with a specific disease [e.g., diabetes code(s); ICD9 250 or ICD10 E08–E13], are used for billing purposes in the USA. These codes can have limited phenotypic resolution, can change over time, and may be used differently across institutions and in clinical practice [46]. Likewise, medication data in an EHR can be incomplete because there are often disagreements between what is prescribed, what prescription is filled, and whether the patient is compliant. An EHR may also not list medications/supplements self-administered by a patient. Another limitation of an EHR is its inherent link to the stability of the patient population. Patients who are transient or seek care at multiple healthcare systems can leave significant gaps in the clinical record, particularly if each institution uses different EHR systems. Yet, even with these challenges, EHR data have proven repeatedly to be an efficient data source for phenotypic information for genomic research [36–38]. Importantly, EHR data, and the biobanks that are linked to that data, have been invaluable for **phenome-wide association studies** (PheWASs).

Phenomic Perspective

In classical genetics, there are two main strategies. 'Forward' genetics is a phenotype-to-genotype strategy that is epitomized in human genomic research by family-based (e.g., linkage studies) and nonfamily-based studies of unrelated individuals (e.g., GWAS) (Figure 1). 'Reverse' genetics is a genotype-to-phenotype strategy and has historically been limited to model organisms (e.g., mouse knockouts). In 2010, Denny *et al*. conducted the first proof-of-principle reverse genetic study in humans and coined it a PheWAS [47] (Figure 1). This study focused on five disease-associated SNPs that were genotyped in their EHR-linked biobank. Case-control status for a variety of disease phenotypes were extracted from the medical record. In simplistic terms, individuals with an ICD code, or a combination of relevant codes, were defined as cases for that disease, whereas those without any relevant codes were defined as unaffected controls

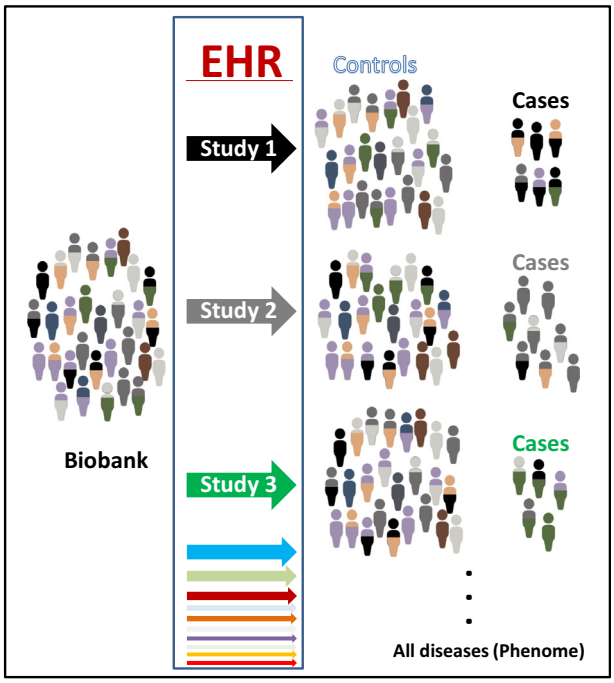

Figure 2. An Example of a Biobank Linked to an Electronic Health Record (EHR). Different colors represent a heterogeneous population that can be separated (by color) into different case-control groupings to study specific diseases or all diseases for phenome-wide association studies (PheWASs).

for that disease. This was repeated for each ICD code to generate 776 different case-control groups that defined the phenome. Each SNP was then associated across the phenome (Figure 2). In most instances, and even with all the limitations of EHR data, the expected associations that were previously identified utilizing forward genetics (i.e., GWAS) could be rediscovered by PheWAS. Other groups have since provided additional proof-of-principle that reverse genetic screens in humans can rediscover known associations previously identified by GWAS [48–53]. More importantly, PheWAS can provide novel insights not readily attainable by forward-genetic strategies.

As stated previously, reverse-genetic approaches were primarily limited to model systems. For example, if a researcher wanted to understand the genetic contribution to pathophysiology, they could apply genome-editing techniques to knock out a specific gene in a model system (e.g., mouse) and evaluate the model for differential outcomes. For obvious ethical reasons, a reverse-genetic screen by genome editing should not be done in humans. However, a similar experiment can be accomplished in human populations utilizing what nature has already provided. For example, groups have conducted systematic PheWASs on common loss-of-function variants. These screens have rediscovered an association between cholesterol levels and a nonsense SNP in *LPL* (rs328) [54,55], implicated a deleterious variant in *KCNH2* with acquired hypothyroidism [54], and discovered that a polymorphic gene deletion and duplication of *SULT1A1* may be related to common allergies [56]. Whereas most PheWASs have focused on common variants, others have begun to utilize sequencing data to study rare loss-of-function variants [57], variants that may have larger effects sizes compared with common variants. In homage to reverse-genetic models, one group that is

systematically evaluating these rare loss-of-function variants calls their study the 'Human Knockout Project' [58].

A unique quality of the PheWAS technique is its capacity to evaluate cross-phenotype associations or pleiotropy [59–61]. There are numerous examples where genetic variants overlap different phenotypes. For example, multiple GWASs have implicated *FTO* with body mass index (BMI) and T2DM [15,16]. A PheWAS focused on *FTO* variants not only rediscovered the same associations, but also implicated other conditions not previously evaluated by GWAS, including sleep apnea [50]. PheWAS can not only characterize pleiotropic effects and capture novel associations, but is also uniquely capable of evaluating confounding effects in the phenotypic information. As it relates to the PheWAS of *FTO*, results demonstrate that the genetic effect of T2DM is partially independent from BMI but sleep apnea is not [50].

The most pleiotropic region of the human genome may include the MHC region on chromosome 6, which encodes *HLA* genes. For example, the first PheWAS by Denny *et al*. provided evidence that *HLA-DRB1*, a gene previously implicated in multiple sclerosis [62], is also associated with risk for 'erythematous conditions' [47]. This novel association was independently replicated by another PheWAS and further refined to include rosacea [63]. Two years later, multiple *HLA* genes, including *HLA-DRB1*, were identified to be associated with rosacea by the first GWAS on this condition [64]. Comprehensive evaluations of all *HLA* genes by PheWAS have further emphasized the capacity of the PheWAS technique to quantify pleiotropy in this important region of the human genome [65–67].

Pleiotropy identified by PheWAS may help in drug development. If there is evidence that 'Disease A' and 'Disease B' share a common genetic etiology through a pleiotropic variant, then it could be hypothesized that 'Drug X' used to treat 'Disease A' may be repurposed to treat 'Disease B' (Figure 3). In the example of *FTO* variants that are associated with T2DM, obesity,

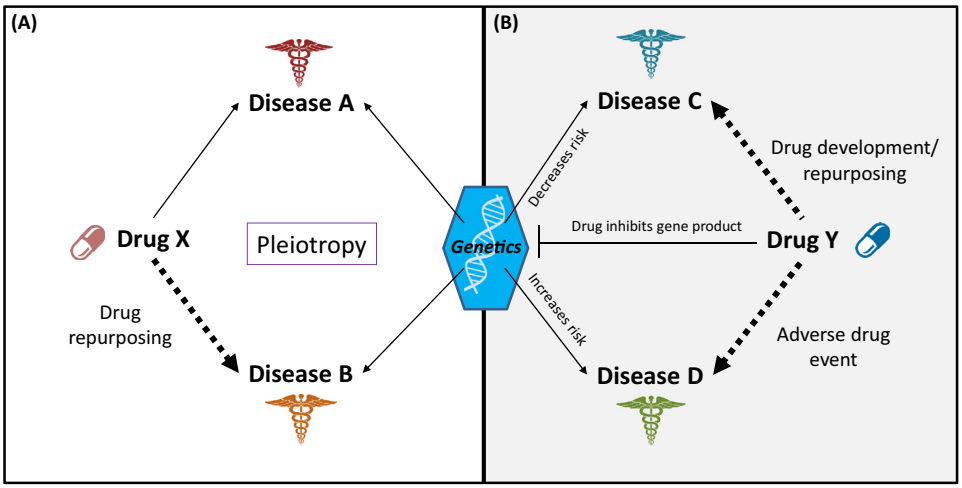

Figure 3. An Illustration of How Phenome-wide Association Study (PheWAS) Data Can Help with Drug Development/Repurposing. (A) If Drug X is effective in treating Disease A, then Drug X may be repurposed to treat Disease B if both Disease A and Disease B share a common genetic etiology as evident by pleiotropy. (B) If Drug Y inhibits a gene product and a loss-of-function variant in a gene decreases risk for Disease C, then Drug Y may be used to treat Disease C. If the same functional variant increases risk for Disease D, Disease D may be an adverse event associated with treatment with Drug Y.

and sleep apnea [50], one could hypothesize that antidiabetic drugs that also result in weight loss (e.g., SGLT2 inhibitors) may be effective in treating BMI-induced sleep apnea in patients with T2DM.

Even though PheWAS is uniquely capable of identifying pleiotropy, PheWAS data will likely prove useful in other ways during drug development. A PheWAS can show that 'Disease C' is associated with a gene the translated product of which can be targeted by 'Drug Y' [68–71]. In this instance, 'Disease C' could represent an indication for 'Drug Y' (Figure 3). A recent example of genomic data describing such a relationship includes cholesterol-lowering PCSK9 inhibitors. This drug class was developed after it was shown that individuals with loss-of-function *PCSK9* variants have low circulating low-density lipoprotein (LDL) levels [72–74]. Conducting follow-up PheWASs on functional *PCSK9* variants may help identify additional targets for PCSK9 inhibitors. Moreover, such PheWAS data could also identify adverse drug events for PCSK9 inhibitors [75]. To use the example of 'Drug Y' described above, if 'Disease D' is associated with the same gene as 'Disease C' but with an opposite direction of effect, then 'Disease D' may be an adverse event relating to treatment with 'Drug Y' (Figure 3).

Given their potential value, pharmaceutical companies are heavily investing in GWASs and PheWASs to reduce costs and risks during drug development. Direct-to-consumer genomic companies (e.g., 23andMe) have adapted their business model to allow pharmaceutical companies to gain access to genomic and phenomic data from consenting customers for drug development [76,77]. Given their success with their PCSK9 inhibitor, Regeneron has already exome sequenced over 50 000 participants from an EHR-linked biobank from Geisinger Health [78] to help with the drug discovery pipeline. Likewise, Regeneron and a consortium of other pharmaceutical companies are investing over US$100 million to sequence 500 000 individuals in the UK-Biobank [79]. Although these partnerships may lead to new and effective treatments, they also introduce ethical concerns regarding data use when profiting from participant data. Regardless, with future trajectories pointing to larger data sets, much will be learned in human genomic research.

Future Perspective

Before GWAS, there were countless candidate SNP-association studies. Candidate SNP studies then evolved to candidate gene studies that evaluated multiple variants in a single gene, then further progressed to the study of multiple variants in multiple genes (e.g., multiple genes in a common biological pathway). This progression was made possible in part by array technologies along with SNP cataloguing efforts (e.g., dbSNP [80] and HapMap Project [81]). It is expected that advancements in PheWAS will take a similar but perhaps not as linear a path. This is evident by the first PheWASs starting with candidate SNP studies [47] then quickly followed by candidate gene [50,53,82,83] and multiple gene studies [65,66,82–86]. Although the PheWAS community is taking a similar path that other scientists took to reach GWAS, growth in PheWAS will always be limited by the availability of phenomic data. There are only a limited number of institutions/groups that currently capture extensive phenomic data through ICD coding or by other data sources, such as extensive epidemiological [48,87–89], rich text [52,56], or biometric/laboratory data [67]. Even fewer groups have DNA linked to such data. Even when all necessary data are available, institutional politics/culture can make it difficult for one investigator to study all diseases simultaneously, especially when colleagues use the same data to carve out their own disease-specific research. It is expected that the progression of PheWAS will be muted compared with the growth observed in GWAS. This is evident in the literature because there were 50 PheWASs published between the first proof-of-concept PheWAS [47] and end of 2017 according to manual inspection of PubMed results using

the search-term 'PheWAS.' Even if this PubMed search identified half of all published PheWASs, this is still dwarfed by the 1588 GWASs published over the same 8-year span according to the NHGRI-EBI GWAS Catalog[ii] [16–18] (Figure 4).

Eventually, the GWAS and PheWAS communities will reach an inflection point. This inevitable event, which is already occurring[iii,iv] [67], will happen when thousands of diseases are studied by GWAS and millions of variants are studied by PheWAS in a single experiment; a GWAS-by-PheWAS (Figure 1). Conducting and interpreting a GWAS-by-PheWAS will have unique challenges. Statistical significance at the genome-wide level is often defined by $P < 5^{*}\ 10^{-8}$ [4,19,90]. In a PheWAS, adjusting for multiple comparison testing can be more complicated. Not only will there be correlations between closely related phenotypes, much like two nearby SNPs in LD, but distant diseases may also be correlated due to disease comorbidities (e.g., heart disease and diabetes). Regardless, when using a Bonferroni correction commonly applied during PheWAS, statistical significance for studying 5000 phenotypes would be defined by $P < 1^{*}\ 10^{-5}$. In a GWAS-by-PheWAS, where every variant is associated with every phenotype, statistical significance could be defined by $P < 5^{*}\ 10^{-13}$. With future biobanks expected to include over 1 million participants linked to extensive phenotypic data (e.g., an EHR), a $P > 5^{*}\ 10^{-13}$ may not be insurmountable for common variants that are associated with common conditions. It is even expected that rare conditions with a disease prevalence of 0.1% (~1000 cases in a population of 1 million participants) can also be evaluated, but challenges will remain when evaluating rare variants and variants with weak effects.

With ever-larger PheWASs and GWASs, disease misclassification may amplify false positives or lead to incorrect conclusions. For example, if there are 100 000 cases defined for a given disease but the misclassification rate for that disease is 1%, signals from the 1000 misclassified cases could reach statistical significance. These associations would likely have weak effects, but it will be difficult to parse true associations from associations driven by other diseases. In addition to challenges with misclassification, population substructure may also lead to additional challenges under large sample sizes [91]. In a well-designed study, population structure can be accounted for by incorporating **principal component analysis** (PCA) into the

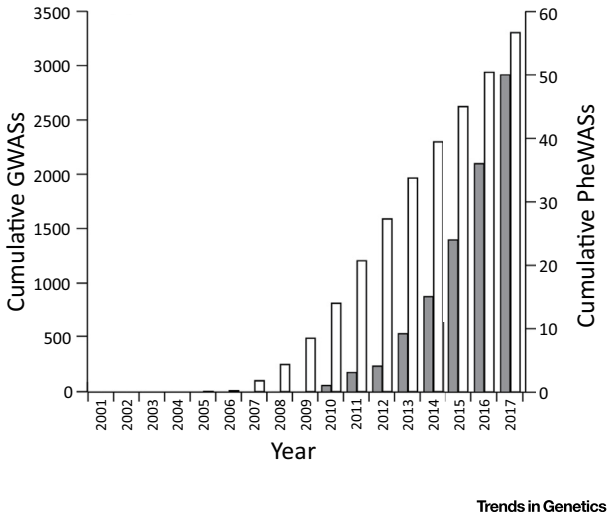

Figure 4. Cumulative Number of Genome-wide Association Studies (GWASs)[ii] [16–18] (Unshaded Bars) and Phenome-wide Association Studies (PheWASs) (Gray Bars) Published over Time.

statistical model [92,93]. If a study population is large enough, it is possible that subpopulations may exist that may not be accounted for in the first few PCAs. If a disease is linked to that subpopulation, population-associated alleles may be confused with disease-associated variants. Conversely, when evaluating rare variants in large sample sizes, adjusting for PCAs may adjust out the effect of the rare variant and reduce power.

Another challenge with ever-larger data sets that is particularly relevant for PheWAS is privacy issues. Manuscripts in themselves make summary data public by disclosing case-control counts, *P* values, odds ratios, and other non-individual-level information. Conversely, funding agencies and journal-publishing groups often require de-identified individual-level data to be made public by depositing such data into access controlled databases (e.g., dbGAP). Genetic data by themselves can be identifiable [94] but often lack a naming source. When a naming source is available, genetic data are identifiable not only to the individual who provided the DNA, but also to relatives [95], as exemplified recently by the 'Golden State Killer' case in the USA [96,97]. Combinations of phenotypic data provided at an individual level could present another pronounced degree of identifiable information. Hypothetically, a male (50% of population) born with a club foot (~0.1% of all live births) [98] who also has multiple sclerosis (~0.1% of individuals) [99] would represent a unique combination of variables using only three outwardly visible traits (~5 per 10 million individuals). With advances in computer science through machine learning and image analysis [100], the probability of some participants being identifiable could be compounded as people freely and unwittingly share health and genetic information through social media and crowd sourcing [101,102]. To this end, lawmakers, regulators, and scientists need to continue to find ways to make individual-level genomic and phenomic data available to the public but in a controlled and protected environment.

The field of human genetics has increasingly become more interdisciplinary through computer science, informatics, and statistics. Future studies that include GWAS-by-PheWAS data will represent billions of association results and terabytes of data that will need to be stored in a queryable data structure likely through big-data solutions. New methods will be required to evaluate phenomic data on top of genomic, proteomic, transcriptomic, and other −omic data. The complexities of such data, in combination with genome editing, single cell analysis, and other molecular techniques, may not only provide important insights into how human genetics influences human disease, but also influence biological processes defined at the organ, tissue, and cellular level.

Regardless of the layers of data generated by GWAS, PheWAS, or GWAS-by-PheWAS, these studies are currently and predominantly rooted in populations of unrelated individuals. As mentioned earlier, much has been historically learned in human genetics through family-based study designs. It is conceivable that, if populations become large enough, especially in a healthcare system servicing a stable patient population, large familial pedigrees may be captured and used for genomic study. For example, ~20% of 2.6 million current and historic patients of Marshfield Clinic (Marshfield, WI, USA), which serves a predominantly rural population, can be placed in family pedigrees using readily available data in an EHR [104]. Such pedigrees may be applied to phenome-wide research [103] and may be incorporated into existing biobanks for genetic association testing even if some family members are not directly genotyped [104]. If large populations of families are linked to extensive phenotypic and genomic data, human genomic research may come full circle back to family-based study designs [105], which will allow researchers to study thousands of human diseases in relationship to variants with a wide spectrum of allele frequencies and effect sizes (Figure 5).

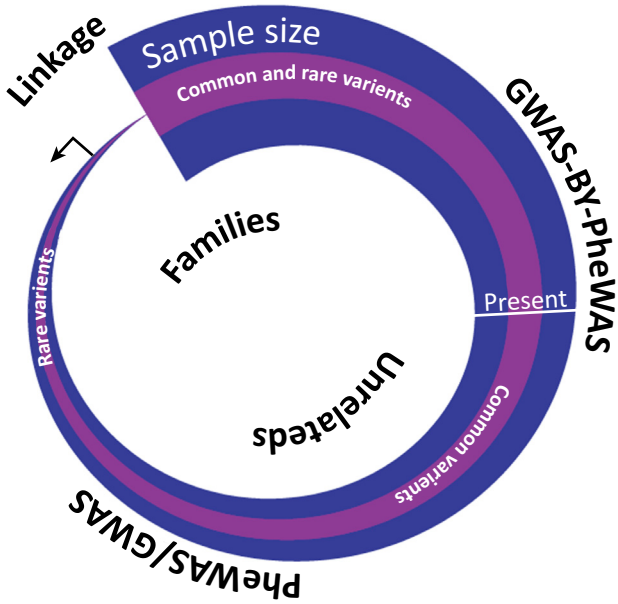

Figure 5. A Simplified Illustration of the Historical and Future Progression of Human Genomic Research Starting with Small Family-Based Studies (e.g., Linkage Studies) to Ever-Larger Studies of Unrelated Individuals [e.g., Genome-wide Association Studies (GWASs), Phenome-wide Association Studies (PheWASs), and GWAS-by-PheWAS]. As sample size (purple) becomes exceptionally large, future studies may include population-based family studies to identify increasing number of trait-associated variants (pink).

Outstanding Questions

If genomic data are collected on a large scale via clinical care or research, and such data are linked to extensive phenotypic information through an electronic health record, how will these data be used for research?

What will be the technical and practical challenges in human genomic research when study populations reach millions of individuals?

How will phenomic data be combined and used with genomic and other layers of –omics data to better understand and treat human disease?

Concluding Remarks

With the computer age, technology is advancing at an ever-increasing rate. Since genomic and healthcare research is progressively intertwined with technology, it is of my opinion that the field of human genomics in the 21st century is like an unstoppable boulder rolling downhill. With high-throughput computing, big data, along with decreasing costs for whole-genome sequencing, society will soon expect genomic data to be incorporated into standard of care as part of precision medicine. Researchers will then have access to virtual DNA biobanks of millions of individuals that link together extensive genomic and phenomic data. Although there will be new challenges in data management, privacy issues, and clinical care when this inevitable future happens, it will also dramatically change how human genomic research is conducted as scientists search to understand and treat human disease using a combination of phenome-wide and genome-wide strategies (see Outstanding Questions).

Resources

[i]www.omim.org
[ii]www.ebi.ac.uk/gwas
[iii]http://geneatlas.roslin.ed.ac.uk
[iv]http://pheweb.sph.umich.edu

References

1. Amberger, J.S. *et al.* (2015) OMIM. org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic. Acids Res.* 43, D789–D798
2. Manolio, T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature* 461, 747–753
3. Antonarakis, S.E. and Beckmann, J.S. (2006) Mendelian disorders deserve more attention. *Nat. Rev. Genet.* 7, 277–282
4. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* 273, 1516–1517

5. Klein, R.J. *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308, 385–389
6. Edwards, A.O. *et al.* (2005) Complement factor H polymorphism and age-related macular degeneration. *Science* 308, 421–424
7. Haines, J.L. *et al.* (2005) Complement factor H variant increases the risk of age-related macular degeneration. *Science* 308, 419–421
8. Majewski, J. *et al.* (2003) Age-related macular degeneration – a genome scan in extended families. *Am. J. Hum Genet.* 73, 540–550
9. Weeks, D.E. *et al.* (2004) Age-related maculopathy: a genome-wide scan with continued evidence of susceptibility loci within the 1q31, 10q26, and 17q25 regions. *Am. J. Hum. Genet.* 75, 174–189
10. Abecasis, G.R. *et al.* (2004) Age-related macular degeneration: a high-resolution genome scan for susceptibility loci in a population enriched for late-stage disease. *Am. J. Hum. Genet.* 74, 482–494
11. Iyengar, S.K. *et al.* (2004) Dissection of genomewide-scan data in extended families reveals a major locus and oligogenic susceptibility for age-related macular degeneration. *Am. J. Hum. Genet.* 74, 20–39
12. Lander, E.S. (1996) The new genomics: global views of biology. *Science* 274, 536–539
13. Chakravarti, A. (1999) Population genetics – making sense out of sequence. *Nat. Genet.* 21, 56–60
14. Reich, D.E. and Lander, E.S. (2001) On the allelic spectrum of human disease. *Trends Genet.* 17, 502–510
15. Welter, D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006
16. MacArthur, J. *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D896–D901
17. Pritchard, J.K. and Cox, N.J. (2002) The allelic architecture of human disease genes: common disease-common variant.or not? *Hum. Mol. Genet.* 11, 2417–2423
18. Zhang, Y. *et al.* (2018) Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* 50, 1318–1326
19. McCarthy, M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369
20. Fritsche, L.G. *et al.* (2016) A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.* 48, 134–143
21. Prescott, S.M. *et al.* (2008) From linkage maps to quantitative trait loci: the history and science of the Utah genetic reference project. *Annu. Rev. Genomics Hum. Genet.* 9, 347–358
22. (1992) A comprehensive genetic linkage map of the human genome. NIH/CEPH Collaborative Mapping Group. *Science* 258, 67–86
23. Cohen, D. *et al.* (1993) A first-generation physical map of the human genome. *Nature* 366, 698–701
24. Hinds, D.A. *et al.* (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* 307, 1072–1079
25. Myers, S. *et al.* (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310, 321–324
26. Abecasis, G.R. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073
27. Abecasis, G.R. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65
28. Birney, E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816
29. Spielman, R.S. *et al.* (2007) Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.* 39, 226–231
30. Lappalainen, T. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511
31. Garge, N. *et al.* (2010) Identification of quantitative trait loci underlying proteome variation in human lymphoblastoid cells. *Mol. Cell. Proteomics* 9, 1383–1399
32. Zhang, W. and Dolan, M.E. (2009) Use of cell lines in the investigation of pharmacogenetic loci. *Curr. Pharm. Des.* 15, 3782–3795
33. Kalari, K.R. *et al.* (2010) Copy number variation and cytidine analogue cytotoxicity: a genome?wide association approach. *BMC Genomics* 11, 357
34. Shukla, S.J. and Dolan, M.E. (2005) Use of CEPH and non-CEPH lymphoblast cell lines in pharmacogenetic studies. *Pharmacogenomics* 6, 303–310
35. Niu, N. *et al.* (2016) Metformin pharmacogenomics: a genome-wide association study to identify genetic and epigenetic biomarkers involved in metformin anticancer response using human lymphoblastoid cell lines. *Hum. Mol. Genet.* 25, 4819–4834
36. McCarty, C.A. *et al.* (2011) The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomic* 4, 13
37. Gottesman, O. *et al.* (2013) The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* 15, 761–771
38. Kho, A.N. *et al.* (2011) Electronic medical records for genetic research: results of the eMERGE consortium. *Sci. Transl. Med.* 3, 79re1
39. Sudlow, C. *et al.* (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 12, e1001779
40. Allen, N.E. *et al.* (2014) UK biobank data: come and get it. *Sci. Transl. Med.* 6, 224ed4
41. Hakonarson, H. *et al.* (2003) deCODE genetics, Inc. *Pharmacogenomics* 4, 209–215
42. Collins, F.S. and Varmus, H. (2015) A new initiative on precision medicine. *N. Engl. J. Med.* 372, 793–795
43. Gaziano, J.M. *et al.* (2016) Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* 70, 214–223
44. Kho, A.N. *et al.* (2012) Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J. Am. Med. Inform. Assoc.* 19, 212–218
45. Hripcsak, G. and Albers, D.J. (2013) Next-generation phenotyping of electronic health records. *J. Am. Med. Inform. Assoc.* 20, 117–121
46. Hebbring, S.J. (2014) The challenges, advantages and future of phenome-wide association studies. *Immunology* 141, 157–165
47. Denny, J.C. *et al.* (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 26, 1205–1210
48. Pendergrass, S.A. *et al.* (2013) Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet.* 9, e1003087
49. Denny, J.C. *et al.* (2013) Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* 31, 1102–1110
50. Cronin, R.M. *et al.* (2014) Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index. *Front. Genet.* 5, 250
51. Hall, M.A. *et al.* (2014) Detection of pleiotropy through a phenome-wide association study (PheWAS) of epidemiologic data

as part of the Environmental Architecture for Genes Linked to Environment (EAGLE) study. *PLoS Genet.* 10, e1004678

52. Hebbring, S.J. *et al.* (2015) Application of clinical text data for phenome-wide association studies (PheWASs). *Bioinformatics* 31, 1981–1987
53. Diogo, D. *et al.* (2015) TYK2 protein-coding variants protect against rheumatoid arthritis and autoimmunity, with no evidence of major pleiotropic effects on non-autoimmune complex traits. *PLoS One* 10, e0122271
54. Verma, A. *et al.* (2016) eMERGE Phenome-Wide Association Study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants. *BMC Med. Genomics* 9, 32
55. Ye, Z. *et al.* (2015) Phenome-wide association studies (PheWASs) for functional variants. *Eur. J. Hum. Genet.* 23, 523–529
56. Liu, J. *et al.* (2017) Relationship of SULT1A1 copy number variation with estrogen metabolism and human health. *J. Steroid Biochem. Mol. Biol.* 174, 169–175
57. Verma, S.S. *et al.* (2018) Rare variants in drug target genes contributing to complex diseases, phenome-wide. *Sci. Rep.* 8, 4624
58. Saleheen, D. *et al.* (2017) Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* 544, 235–239
59. Tyler, A.L. *et al.* (2016) The detection and characterization of pleiotropy: discovery, progress, and promise. *Brief Bioinform.* 17, 13–22
60. Pendergrass, S.A. and Ritchie, M.D. (2015) Phenome-wide association studies: leveraging comprehensive phenotypic and genotypic data for discovery. *Curr. Genet. Med. Rep.* 3, 92–100
61. Pendergrass, S.A. *et al.* (2015) Phenome-wide association studies: embracing complexity for discovery. *Hum. Hered.* 79, 111–123
62. Baranzini, S.E. and Oksenberg, J.R. (2017) The genetics of multiple sclerosis: from 0 to 200 in 50 years. *Trends Genet.* 33, 960–970
63. Hebbring, S.J. *et al.* (2013) A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun.* 14, 187–191
64. Chang, A.L.S. *et al.* (2015) Assessment of the genetic basis of rosacea by genome-wide association study. *J. Invest. Dermatol.* 135, 1548–1555
65. Liu, J. *et al.* (2016) Phenome-wide association study maps new diseases to the human major histocompatibility complex region. *J. Med. Genet.* 53, 681–689
66. Karnes, J.H. *et al.* (2017) Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci. Transl. Med.* 9, eaai8708
67. Verma, A. *et al.* (2018) PheWAS and beyond: the landscape of associations with medical diagnoses and clinical measures across 38,662 individuals from Geisinger. *Am. J. Hum. Genet.* 102, 592–608
68. Rastegar-Mojarad, M. *et al.* (2015) Opportunities for drug repositioning from phenome-wide association studies. *Nat. Biotechnol.* 33, 342–345
69. Robinson, J.R. *et al.* (2017) Genome-wide and phenome-wide approaches to understand variable drug actions in electronic health records. *Clin. Transl. Sci.* 11, 112–122
70. Roden, D.M. (2017) Phenome-wide association studies: a new method for functional genomics in humans. *J. Physiol.* 595, 4109–4115
71. Versel, N. (2018) *Vanderbilt Spins Out Nashville Biosciences to Offer BioVU, PheWAS Services to Pharma,* GenomeWeb
72. Zhao, Z. *et al.* (2006) Molecular characterization of loss-of-function mutations in PCSK9 and identification of a compound heterozygote. *Am. J. Hum. Genet.* 79, 514–523
73. Roth, E.M. *et al.* (2012) Atorvastatin with or without an antibody to PCSK9 in primary hypercholesterolemia. *N. Engl. J. Med.* 367, 1891–1900
74. Elguindy, A. and Yacoub, M.H. (2013) The discovery of PCSK9 inhibitors: A tale of creativity and multifaceted translational research. *Glob. Cardiol. Sci. Pract.* 2013, 343–347
75. Jerome, R.N. *et al.* (2018) Using human 'experiments of nature' to predict drug safety issues: an example with PCSK9 inhibitors. *Drug. Saf.* 41, 303–311
76. Petrone, J. (2017) 23andMe wades further into drug discovery. *Nat. Biotechnol.* 35, 897
77. Mason, M. *et al.* (2017) Direct-to-consumer genetic testing and orphan drug development. *Genet. Test Mol. Biomarkers* 21, 456–463
78. Abul-Husn, N.S. *et al.* (2016) Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science* 354, aaf7000
79. Herper, M. (2018) *Drug Company Consortium To Sequence The Genes Of 500 000 Britons Over Next Two Years,* Forbes
80. Sherry, S.T. *et al.* (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* 9, 677–679
81. (2003) The International HapMap Project. *Nature* 426, 789–796
82. Denny, J.C. *et al.* (2011) 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.* 89, 529–542
83. Polimanti, R. *et al.* (2017) Phenome-wide association study for CYP2A6 alleles: rs113288603 is associated with hearing loss symptoms in elderly smokers. *Sci. Rep.* 7, 1034
84. Shameer, K. *et al.* (2014) A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum. Genet.* 133, 95–109
85. Ehm, M.G. *et al.* (2017) Phenome-wide association study using research participants' self-reported data provides insight into the Th17 and IL-17 pathway. *PLoS One* 12, e0186405
86. Klarin, D. *et al.* (2017) Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. *Nat. Genet.* 49, 1392–1397
87. Pendergrass, S.A. *et al.* (2011) The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet. Epidemiol.* 35, 410–422
88. Verma, S.S. *et al.* (2016) Phenome-wide interaction study (PheWIS) in AIDS Clinical Trials Group data (ACTG). *Pac. Symp. Biocomput.* 21, 57–68
89. Moore, C.B. *et al.* (2015) Phenome-wide association study relating pretreatment laboratory parameters with human genetic variants in AIDS Clinical Trials Group protocols. *Open Forum Infect. Dis.* 2, ofu113
90. Khoury, M.J. and Yang, Q. (1998) The future of genetic studies of complex human diseases: an epidemiologic perspective. *Epidemiology* 9, 350–354
91. Marchini, J. *et al.* (2004) The effects of human population structure on large genetic association studies. *Nat. Genet.* 36, 512–517
92. Price, A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909
93. Freedman, M.L. *et al.* (2004) Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* 36, 388–393
94. Gymrek, M. *et al.* (2013) Identifying personal genomes by surname inference. *Science* 339, 321–324
95. Erlich, Y. and Narayanan, A. (2014) Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* 15, 409–421
96. Phillips, C. (2018) The Golden State Killer investigation and the nascent field of forensic genealogy. *Forensic Sci. Int. Genet.* 36, 186–188
97. Syndercombe Court D (2018) Forensic genealogy: some serious concerns. *Forensic Sci. Int. Genet.* 36, 203–204
98. Dobbs, M.B. and Gurnett, C.A. (2009) Update on clubfoot: etiology and treatment. *Clin. Orthop. Relat. Res.* 467, 1146–1153

99. Dilokthornsakul, P. *et al.* (2016) Multiple sclerosis prevalence in the United States commercially insured population. *Neurology* 86, 1014–1021

100. Claes, P. *et al.* (2014) Modeling 3D facial shape from DNA. *PLoS Genet.* 10, e1004224

101. Yuan, J. *et al.* (2018) DNA: Land is a framework to collect genomes and phenomes in the era of abundant genetic information. *Nat. Genet.* 50, 160–165

102. Kaplanis, J. *et al.* (2018) Quantitative analysis of population-scale family trees with millions of relatives. *Science* 360, 171–175

103. Polubriaginof, F.C.G. *et al.* (2018) Disease heritability inferred from familial relationships reported in medical records. *Cell* 173, 1692–1704

104. Huang, X. *et al.* (2018) Applying family analyses to electronic health records to facilitate genetic research. *Bioinformatics* 34, 635–642

105. Staples, J. *et al.* (2018) Profiling and leveraging relatedness in a precision medicine cohort of 92, 455 exomes. *Am. J. Hum. Genet.* 102, 874–889