PERSPECTIVE
https://doi.org/10.1038/s41588-018-0066-3
nature genetics

The importance of cohort studies in the post-GWAS era

Cisca Wijmenga[1,2*] and Alexandra Zhernakova[1*]

The past decade has seen enormous success of wide-scale genetic studies in identifying genetic variants that modify individuals' predisposition to common diseases. However, the interpretation and functional understanding of these variants lag far behind. In this Perspective, we discuss opportunities for using large-scale cohort studies to investigate the downstream molecular effects of SNPs at different 'omics' data levels. We point to the pivotal role of population cohorts in establishing causality and advancing drug discovery. In particular, we focus on the breadth-versus-depth concepts of population studies, on data harmonization, and on the challenges, ethical aspects and future perspectives of cohort studies.

The past decade has seen enormous progress in identifying genetic factors that predispose individuals to common diseases. Since the first reports of SNPs analyzed for association with macular degeneration[1] and myocardial infarction[2] by genome-wide association studies (GWAS), the GWAS catalog[3] has now grown to contain tens of thousands of SNPs associated with hundreds of common diseases. More than 100 loci have been associated with complex diseases or traits, including rheumatoid arthritis[4], inflammatory bowel disease (IBD)[5], schizophrenia[6] and blood lipids[7]. Pleiotropy has been shown for many associated loci, even among biologically unrelated phenotypes[8]. Development of fine-mapping platforms, such as the Immunochip and Metabochip, has led to the identification of more pleiotropic disease-associated SNPs while considerably narrowing down the regions of association[9]. Exome and whole-genome sequencing are expected to identify rare variants and copy number variants, although their contribution to the heritability of common diseases and traits is rather modest[10,11]. As the cost of genotyping SNPs decreases and cohort sizes increase, GWAS should identify many more loci that will enhance insight into the genetic architecture of common diseases.

Despite the success in identifying disease loci, understanding of how SNPs predispose individuals to disease remains limited, particularly because most GWAS SNPs are located in noncoding parts of the genome. Several elegant functional studies have illustrated the mechanisms of how SNPs contribute to diseases, thereby providing insight into the importance of the roles of noncoding variants in disease biology. Mouse studies, for instance, have shown that a common noncoding variant on 9p21, which is strongly associated with cardiovascular disease (CVD) risk, affects the proliferation of vascular cells[12]. Nevertheless, there is no single straightforward strategy for functional studies, thus making them laborious, time consuming, costly and rarely high throughput. Hence, there is a growing gap between the identification of new disease SNPs and their interpretation. Understanding of the regulatory components of the genome, and the function of many noncoding genes, is still in its infancy despite large-scale initiatives such as the Roadmap Epigenomics Project[13]. In this Perspective, we examine the role of large cohort studies with deeply phenotyped participants in the functional and clinical translation of GWAS findings.

Cohort studies

There are various types of cohort studies, with wide variation in their scope, size, depth of phenotypes collected and follow-up of participants (Table 1). GWAS studies mainly use cross-sectional case–control designs to find disease-associated SNPs, whereas studies aimed at complex quantitative traits such as body-mass index[14], educational attainment[15] and human immune response[16] also use unselected population-based biobanks such as the UK Biobank[17], the LifeLines cohort study[18] and the 500 Functional Genomics (500FG) cohort[19]. The ideal cohort for further interpretation and follow-up of GWAS loci: (i) is longitudinal, with data collected at yearly time points (before, during and after disease onset), (ii) includes a large number of participants without confounding disease (potential cohort sizes in Fig. 1) and (iii) contains deep phenotypes of intrinsic, exogenous, environmental and molecular factors, preferably sampled from different tissues. The importance of studying relevant tissues has recently been highlighted by several Genotype-Tissue Expression (GTEx) studies that have shown tissue-specific effects of disease-associated SNPs on gene expression[20]. In ideal prospective cohorts, individuals can be stratified by genetic risk and followed over time to identify factors that determine outcomes and to evaluate gene–environment interactions. Although a single cohort with all these features does not exist and is currently too expensive to build and maintain, there are some longitudinal cohorts that are long running, deeply phenotyped and population based. The Framingham Heart Study (FHS), for example, was started in 1948, when the collection of the original cohort of 5,209 men and women was initiated. The study was followed by the Offspring cohort (founded in 1971 and including offspring of the original cohort; 5,124 individuals) and the Generation-three cohort (founded in 2002; 4,095 individuals). Two additional cohorts, Omni (founded in 1994; 506 individuals) and Omni-two (founded in 2003; 410 individuals), have similar designs and include more ancestrally diverse participants. In total, the FHS has collected data from more than 15,000 individuals from three generations of participants[21,22]. This design provides an opportunity to estimate the predictive value of associated SNPs. For example, following the recently established effects of genetics on levels of circulating proteins used as markers for CVD[23], the FHS has enabled estimation of these proteins' predictive values regarding later development of CVD[24]. The power of

[1]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. [2]K.G. Jebsen Coeliac Disease Research Centre, Department of Immunology, University of Oslo, Oslo, Norway. *e-mail: c.wijmenga@umcg.nl; a.zhernakova@umcg.nl

Table 1 | Overview of the different types of cohort studies, their advantages and disadvantages, and selected examples

Criteria of definition	Type of cohorts	Definition	Advantages	Disadvantages	Examples
By design: cross-sectional versus longitudinal	Cross-sectional	Information and material collected at one time point	Easy to collect large cohorts	Cannot follow development of traits over time	Most GWAS cohorts used for identification of disease loci
	Longitudinal/ prospective	Data collected at multiple time points	Can obtain samples before and after disease onset, during disease progression and during treatment; individuals serve as their own controls	Subject dropout as individuals move away from study centers; time-consuming/expensive sample collection and storage; incompleteness: initial lack of clarity regarding what data should be collected (e.g., microbiome)	FHS LifeLines cohort study UK Biobank Rotterdam study
By disease status: disease-selected versus at-risk cohorts versus unselected population cohorts	Disease-related case–control cohorts	Selection of cases and matched disease-free individuals	Most powerful design for finding genetic loci associated with specific diseases; depending on the depth of sampling, might be useful for functional follow-up studies	Less adept at showing causal relationships; cohorts potentially limited to pinpointing the regions of association	Most GWAS cohorts used for identification of disease loci
	At-risk cohorts	At-risk individuals, e.g., relatives of patients with specific diseases	Powerful design for identification of gene–environment interactions, particularly in longitudinal studies	Focus on risk for a specific disease, thus making these cohorts less powerful for other studies	GEM (healthy relatives of patients with Crohn's disease) PreventCD (siblings of patients with celiac disease)
	Population-based cohorts	Unselected collection of individuals from the general population	Not confounded by disease; can identify biomarkers before disease onset (in longitudinal designs) when the phenotype of interest is measured; can have extremely deep phenotypic information including lifestyle parameters to perform GxE studies	Potentially insufficient power to identify disease risk loci, owing to small numbers of disease cases; incompleteness: initial lack of clarity regarding what data should be collected (e.g., microbiome)	UK Biobank deCODE Estonian Biobank LifeLines
By size: from single-case cohorts to population-wide studies	Small cohorts and single case studies	Small collections of individuals with rare diseases	A unique way to identify risk loci	Potentially difficult replication of genetic findings	Many rare-disease cohorts used for exome/whole-genome sequencing
	Region- or population-wide studies	Extensive biobanks including a substantial proportion of the population	Can identify factors relevant to the entire population; opportunity to impute missing values for the remainder of the population	Expensive; less accurate imputation in genetically diverse populations	deCODE Estonian Biobank LifeLines
By depth: from selected single phenotype to multi-omics level	Single phenotype	Studies aimed at investigating one disease or trait	Quick and relatively inexpensive sample collection	Not suitable for deep and follow-up studies	Some GWAS cohorts used for identification of disease loci; many rare-disease samples used for exome/whole-genome sequencing.
	Extensive multi-omics studies	Cohort with extensive collection of biological material	Can identify genetic effects at different biological levels	Time-consuming/ expensive sample collection and storage	LifeLines DEEP 500FG Personalized-nutrition project (complex intervention study in general population)
By time of collection	Birth cohorts	Cohorts collected relatively soon after birth and followed up later in life	Can follow the influence of environmental factors and diseases throughout life	Time consuming/ expensive; requires long-term infrastructure	1958 and 1970 British cohort studies LifeLines NEXT Avon Longitudinal Study of Parents and Children (ALSPAC) Danish National Birth Cohort (BSIG) Norwegian Birth Cohort (MoBa)

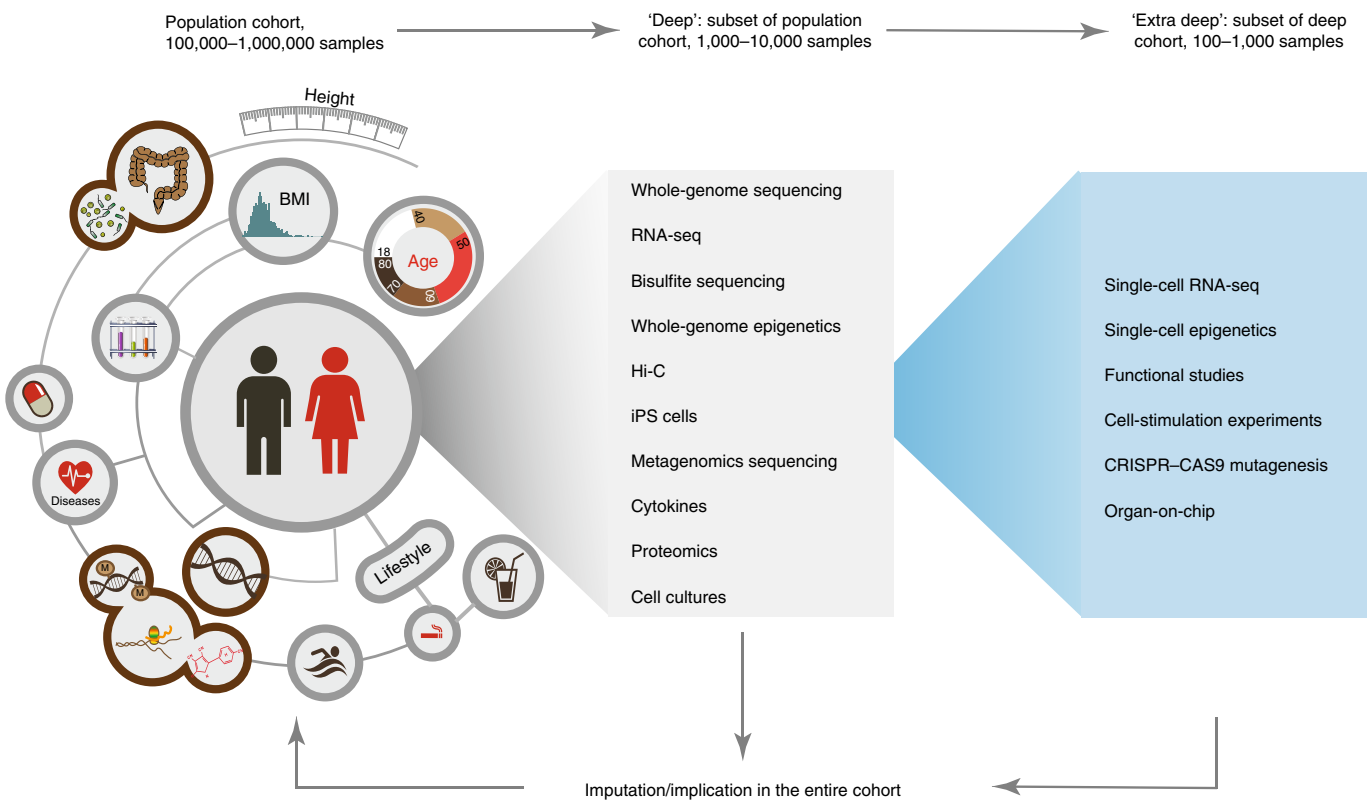

Fig. 1 | Scheme of a cohort study in which a subset of the extensive population cohort is selected for deep multi-omics and single-cell phenotypes. These 'nested' data can then be used for implication and imputation of deep omics data in the entire population. Figure made by C.W., A.Z. and J. Fu.

the three-generation design of the FHS cohort has also been demonstrated by a study on gene–environment interactions of the *FTO* (fat-mass and obesity associated) locus[25], which has found that the correlation of the *FTO* risk allele with elevated body-mass index was much stronger in participants born after 1942 than in those born before 1942 ($P < 0.017$ for individuals with one or two copies of the *FTO* risk allele). These results suggest that environmental factors can modify the penetrance of disease-associated alleles.

Many other cohorts worldwide have population samples in a longitudinal design, including, among many others, the Twins UK Registry[26], UK Biobank[17] and 1958 birth cohort[27] in the UK; Cooperative Health Research in the Region Augsburg (KORA)[28] and Study of Health in Pomerania (SHIP)[29] in Germany; the Sardinia Study[30] in Italy; the Northern Finland Birth Cohort (NFBC) 1966 and 1986 (refs. [31,32] in Finland; and the Nurses' Health Study[33] in the United States. These projects are mainly focused on investigating environmental and genetic risk factors for heart disease, metabolic disease, cancer and aging. Each project has substantially contributed to research on various aspects of population health.

The Netherlands also has a long tradition of prospective population cohort studies, including LifeLines, the Rotterdam Study, the Nederland Twin Register, the Leiden Longevity Study and Generation R; these Dutch cohorts have provided many of the illustrative examples discussed herein. The largest Dutch population cohort is LifeLines, which now includes 167,000 participants in a three-generation design[18]. Recently a birth cohort was started (LifeLines NEXT) that will include >1,000 families. For an additional 1,500 participants (LifeLines DEEP[19]), extensive multi-omics phenotypes have been collected from blood (RNA-seq gene expression, methylation, metabolomics and proteomics), exhaled air (volatile organic compounds) and stool (host metabolites and gut microbiome) samples, in addition to the 2,000 standard phenotypes[19]. LifeLines DEEP is also part of the Human Functional Genomics Project (HFGP)[34], and HFGP also includes the 500FG cohort, which has been extensively immunologically phenotyped through methods including flow cytometry of 73 different immune cells and measurements of six different cytokines in response to 18 different stimuli ex vivo[16]. The HFGP cohorts, often in combination with other cohort studies, have contributed to understanding of normal interindividual variation in molecular parameters including gene expression[35], methylation[36], microbiome composition[37], immune function[16] and metabolites[38], as well as the effects of intrinsic, exogenous and genetic factors on all these parameters.

In the 500FG cohort, the production of cytokines has been shown to be highly variable in the general population but to be partially modulated by the gut microbiome, seasonality and genetics[39–41]. The LifeLines DEEP metagenomics study has provided the first description of more than 100 phenotypic parameters (including diet, medication use and blood biochemistry) that influence the normal variation in the human gut microbiome, in 1,136 metagenomes generated from the general population[37]. Together, these factors explain 18.6% of the observed variation in microbiome composition. Given all the potential confounding factors influencing gut microbiome composition and diversity, replication studies are crucial. Reassuringly, there was a 92% replication rate between the Dutch LifeLines DEEP data and the Belgium Flemish Gut Flora Project for factors present in both studies[42]. One of the findings in both populations is that stool type, as measured by the Bristol stool scale, has a remarkably strong influence on variations in microbiome composition[43,44]. The LifeLines DEEP study has also found a strong correlation between gut microbiome composition and medication use, particularly the use of proton-pump-inhibitor drugs (PPIs), a correlation subsequently replicated in clinical cohorts[45]. PPI use results in changes in 20% of bacterial taxa, decreasing their diversity and

consequently increasing the risk of enteric infection[45]. This observation has initiated a discussion on the over-the-counter availability of PPIs, which remain one of the most widely used drugs in Western societies. The findings of these studies have also aided in development of the first guidelines for future microbiome studies, such as harmonization of all analysis and sequencing methods and correction for internal and environmental factors including body-mass index, stool type and the use of medications such as PPIs, antibiotics, metformin, laxatives and statins. The breadth and depth of data available in biobanks, and the ability to integrate these data, should lead to many more findings with similar widespread implications.

Analyzing the effects of genetic variation on gene expression, methylation, the microbiome and other traits can help to unravel molecular and biological pathways involved in disease predisposition, even when these factors are studied in the general population. For example, compared with individuals carrying few or no IBD risk alleles, LifeLines DEEP participants carrying higher numbers of IBD risk alleles have a lower abundance of *Roseburia*, beneficial gut bacteria whose abundance is known to be low in IBD[46]. In another population cohort, healthy individuals carrying the rs4917014[T] SNP (predisposing individuals to systemic lupus erythematosus) have been found to have elevated expression of type 1 interferon genes and low expression of complement genes, both of which are hallmarks of systemic lupus erythematosus[47]. Multiple levels of data allow truly multi-omic studies to unravel the effects of genetics across different levels of molecular data. For example, the ulcerative colitis risk SNP rs3774937[C], located in an intron of *NFKB1*, the gene encoding the transcription factor NF-κB1, decreases expression of *NFKB1* and diminishes the methylation of 380 CpG sites across the entire genome, and 147 (38.7%) of these CpG sites overlap with NF-κB-binding sites[36]. Another level of integration that is possible in general population cohorts is the interaction of genetics with exogenous factors. In LifeLines DEEP, an interaction has been observed among diet, bacteria and a functional SNP in the lactase gene locus (rs4988235:C>T): this SNP is associated with higher levels of *Bifidobacteria*[48,49] ($P = 3.45 \times 10^{-8}$). Moreover, individuals genetically predisposed to adult-type lactose intolerance but reporting a higher intake of milk products have an increased abundance of *Bifidobacteria*; however, this interaction has not been observed for individuals with milk-tolerating genotypes[48]. The lactose-metabolizing properties of *Bifidobacteria* have been reported previously[50], and the LifeLines DEEP observations suggest that *Bifidobacteria* might improve tolerance to dairy products in carriers of the hypolactasia haplotype.

Clearly, studying the general population allows for unbiased research and is most powerful when as many different phenotypes as possible are collected in standardized ways to enable collaborative replication studies.

Challenges in cohort studies

Despite the clear success of cohort studies and their power in identifying risk factors for diseases and the ability to study the downstream effects of disease SNPs, aspects such as burden for participants, costs and ethical issues make collecting the ideal longitudinal cohort a challenge.

Breadth versus depth. Depending on the research question, a balance must be sought regarding the number of participants in a cohort study, the number of phenotypes included and the follow-up time. In the past, with the focus on GWAS, it was the number of participants with a certain phenotype that was important. However, post-GWAS studies require additional considerations. For example, functional studies require multi-omics data and the ability to perform single-cell analysis on target tissues, whereas biomarker-discovery and predictive-value studies require large prospective cohorts. Although keeping participants involved in a prospective cohort may be difficult, studies such as the FHS have shown that participation can remain high across multiple generations[51]. Further, for this kind of study, relatively small countries, such as the Netherlands, may have an advantage over much larger countries because participants in smaller nations tend to stay in the same geographical area throughout their lives. Cohort participants from these smaller national populations may also be more genetically homogeneous, more likely to share similar environments and easier to reach and follow over time. However, some of these factors may also be disadvantageous. For example, more homogenous populations show less variability in some phenotypes and in genetic background.

Independent research questions can require different sets of phenotypes; frequencies of data collection; and ages, sexes and ancestries of participants. Although these varied requirements are clearly difficult to achieve in a single cohort, one solution may be to create combined cohorts, such as the LifeLines DEEP cohort, in which additional 'deep' molecular phenotypes are collected for a subset of participants. A combination of a basic and a deep cohort also allows for prediction and imputation of additional values in the basic cohorts (Fig. 1). This approach is widely used in genetics, and reference sequencing databases such as the 1000 Genomes Project and the Haplotype Reference Consortium are now routinely used for high-quality imputation of millions of SNPs from hundreds of thousands of SNPs present on GWAS chips. For example, whole-genome sequencing of 2,636 individuals in the Icelandic population has allowed for imputation of rare variants for a further 104,220 individuals for whom only genome-wide SNP-array genotyping had been available[52]. Prediction can also be applied to data other than genotypes. For example, measurements on 73 different immune cells in the 500FG cohort combined with RNA-seq data on whole blood from 20% of these individuals has been used to predict more than 20 immune-cell types in other cohorts with RNA-seq data (Y. Li (University Medical Center Groningen), personal communication). These deconvoluted data, in turn, can be used to investigate cell-type-specific expression quantitative trait loci (eQTLs). A recent study has found that 12% of GWAS SNPs expressed eQTLs in a cell-type-specific manner, and 971 of 2,743 content-specific QTL SNPs had an eQTL effect predominantly in neutrophils, including the *NOD2* SNP rs1981760, which is associated with IBD[35].

Most biobanks collect blood samples; however, because diseases may manifest in other tissues, computational algorithms to impute information from missing tissues are relevant. Publically available reference databases of multiple-tissue multi-omics analysis such as GTEx[53] will become important in predicting tissue-specific effects. For example, a recent analysis of GTEx data has identified the tissue-specific effects of disease-associated SNPs[20]. However, using the GTEx database for this purpose also has several limitations: (i) the use of postmortem tissues in GTEx might influence expression of certain genes, (ii) phenotype information is unknown for most samples, and (iii) it is not possible to take environmental factors into consideration. Current alternative approaches to examine tissue specificity include the use of induced pluripotent stem (iPS) cell technology, which may help to differentiate easily accessible cells into cell types that are difficult to obtain. Future advances in organ-on-chip technology should make it feasible to reconstruct diseased organs of interest either carrying the relevant genetic background or subjected to gene editing through CRISPR–Cas9 technology[54].

Data harmonization. Combining cohorts is the best way to increase sample size, but harmonization across cohorts and handling of missing data can be challenging. Imputation has been widely used to infer missing data in domains as diverse a social sciences, health-survey data and genetics[55,56]. Imputation of different types of genotyping arrays has also been a solution to harmonize genetic data[57]. Several

computational strategies have been suggested to overcome the large technical variations in gene expression, methylation and microbiome data that currently hinder data harmonization. For example, correcting gene expression data for nongenetic principal components has led to identification of genomic deletions and duplications by reusing gene expression data from 77,840 publicly available gene expression profiles[58]; that study has also demonstrated the power of reusing existing data. In microbiome studies, differences in DNA-isolation and sequencing methods and in analysis pipelines have led to considerable differences in results. Furthermore, there are different variable regions that can be sequenced through 16S rRNA gene sequencing, and these regions can yield different estimations of the relative abundances of bacteria. Methods for combined analysis of these different variable regions are being developed by the Microbiome QTL Consortium (MiBioGen), which currently includes 19 cohorts and >20,000 samples with data available on 16S rRNA gene sequencing and host genetics.

The rapid development of new technologies has made standardization difficult, even within cohorts. For example, although array-based gene expression had long been the only method, it has now been surpassed by RNA-seq. Similar shifts in methods have likewise occurred for 16S rRNA sequencing and metagenomic sequencing, and for methylation arrays and bisulfite sequencing. Different platforms also exist for metabolomics analysis. In the Netherlands, harmonization across different biobanks is facilitated by Biobanking and BioMolecular Resources Research Infrastructure–the Netherlands (BBMRI-NL)[59]. For example, the BBMRI Genome of the Netherlands (GoNL) project has added value by imputing rare genetic variants and more population-specific variants[60]. In the BBMRI-BIOS consortium, RNA-seq and 450K methylation arrays have been applied to the same set of 4,000 samples from four different Dutch biobanks[36]. The BBMRI Metabolomics initiative has measured a set of nuclear magnetic resonance–based blood metabolites in ~50,000 subjects from Dutch cohorts (URLs). The metagenomics studies across three Dutch cohorts have been harmonized in 1,500 subjects[48]; in all cohorts, measurements were performed in the same laboratory, and data analysis was performed with the same pipelines to decrease technical variations.

Basic parameters also must be taken into account when results are combined. Age, for example, influences methylation, metabolomics, microbiome, gene expression and telomere length. Other phenotypes can be affected by cultural and social differences. Dietary patterns, for example, differ across countries, thus making development of uniform questionnaires challenging. Accurate and complete annotation of dietary patterns is also time consuming for participants. A future solution might be the use of smartphone applications that analyze dietary habits on the basis of digital photographs of meals. Cultural differences are likely to be the reason why the frequency of irritable bowel syndrome estimated from questionnaires varies from 1% to 35% across populations, even when the same validated (ROME III) questionnaires are used[61]. The subjectivity of measurements is less problematic in longitudinal studies, in which participants can serve as their own controls. In the future, wearable health-tracking devices and smartphones are expected to be used routinely for collecting detailed lifestyle data[62].

Ethical challenges. Cohort studies raise several ethical issues. These issues include not only concerns about making data available through public databases (such as the Database of Genotypes and Phenotypes (dbGAP) and European Nucleotide Archive), thus introducing a risk that subjects might be identifiable, as well as concerns about what information should be reported back to participants, including incidental findings. As the costs of whole-genome sequencing fall further, potentially actionable DNA variations will be found in every individual genome. What should a participant be informed about? Currently, there are no easy answers to these questions, but many organizations are trying to address the ethical, legal and social implications[63]. A recent example of an ethical challenge has been raised in the Icelandic population. Using population-specific reference panels, researchers were able to impute rare variants in the *BRCA2* gene that substantially increase the risk of developing breast cancer, a risk that can be greatly decreased by prophylactic mastectomy[53]. Ethical discussions are ongoing about direct-to-consumer genetic tests such as those offered by 23andMe, which provides on-demand details about ancestry, carrier status of inherited diseases and risks of common conditions and traits. The growing interest in this information is evidenced by the more than 2 million customers who have sent saliva for genetic analysis.

Identification of mutation carriers can lead to their participation in screening programs or preventive actions that may be beneficial. However, in many cases, there are no immediate treatments or preventive measures, thereby causing participant frustration. Further, the genetic risk estimates for common diseases remain inaccurate because of the incomplete understanding of the genetic basis of these diseases and the lack of information and knowledge about gene–environment and gene–microbiome interactions. Because customers do not always understand these limitations, 23AndMe was banned from offering their personal genome service to customers in 2015. However, since April 2017, 23AndMe has been allowed to provide genetic health risk information for ten diseases or conditions. Similar companies worldwide are working under the many different national regulations of the countries in which they are registered. A future challenge will be to bring all these regulations into concordance while respecting cultural and individual differences. Another thorny and unsolved ethical question is what information should be shared with insurance companies.

Our personal opinion on these ethical aspects is that the generation of individual genetic, environmental and microbiome data will grow tremendously in the coming years, and sharing these data in public domains has enormous potential to accelerate new biological discoveries. Recent examples of this synergy at work include the analysis of publically available expression data to identify chromosomal rearrangements in patients with cancer[58], and the discovery of novel genetic variants linked to common infections in a recent genetic study of 23andMe participants[64]. However, appropriate regulations for the sharing and use of personal data are important. First, no discrimination should occur on the basis of any genetic findings. Next, participants should always be the owners of their data and ideally should decide whether they would like to participate for each research question. Participants should also be able to receive their data if they wish, but these data should be accompanied by extensive information on the limitations and possibilities of unexpected findings and their consequences. Participants should also be offered genetic counseling and medical advice based on such findings. Optimally, informed consent should be dynamic enough to account for different research opportunities, include a strategy to be followed in the case of unexpected findings and provide the opportunity to recontact individuals for follow up studies. Finally, a personalized-medicine approach should be applied to maximize voluntary usage of individual data for improving personal health.

What does the future hold?

Extensive and longitudinal biobanks are useful for increasing understanding of the genetic and nongenetic factors that predispose people to disease (Box 1). Biobanks have been the domain of academic institutions, but there is a growing public interest in crowdsourced cohorts such as 23andMe, the American Gut and uBiome. These citizen-science cohorts benefit from modern technology and smartphones with applications that measure parameters including physical activity, sleep and lifestyle patterns. However, these cohorts are often more limited in the amounts of biomaterials and molecu-

Box 1 | Take-home messages and recommendations for future biobank research

1. Open accessibility of biobank data is highly recommended to allow for wider reuse of data.
2. Collection of data about known confounders should be included in the design of new studies, e.g., information on stool consistency and PPI use for microbiome studies or data on seasonal changes for cytokine measurements.
3. Longitudinal designs may benefit from increased possibilities for participants to collect data themselves, preferentially on a regular basis, depending on the variable.
4. If possible, iPS cells should be collected for future functional studies including organ-on -chip studies.
5. Harmonization of protocols for measurements and analysis should be taken into account in the design of cohort studies, preferably at the national level.
6. Biobanks should include all possible ages, ancestries and health statuses. More emphasis should be placed on prospective birth cohorts, because this methodology remains the only way to address certain questions, a benefit that can outweigh the long startup time.
7. Biobanks should be built in close collaboration with citizens so that individuals have their 'own' biobanks and can decide who can access to their data.

lar data generated. As transcriptomic, metabolomic and proteomic data provide more direct information about DNA variation, and often provide indirect information about phenotypic variation, the power for discovery lies in the integration of nongenetic data with individuals' genetic and molecular data. This process will be the basis for developing precision health and medicine.

An example of the strength of combining diet, multi-omics and health is the personalized-nutrition project of the Weizmann Institute of Science (in Israel)[65]. In a cohort of 800 individuals, glycemic parameters have been related to the microbiome, genetics, metabolomics, diet and lifestyle. The researchers have been able to identify individual patterns of glycemic response to food and to develop predictive algorithms that were successfully applied to an independent cohort. This work illustrates a concrete step toward personalized nutrition and medical advice, and might represent the future of prevention for diseases such as type 2 diabetes. Given the rapid development in accurate and ongoing measurement of important health parameters, we expect research in the coming decades to provide insight into which environmental changes individuals can make to modify the effects of their genetics on their development of common and complex diseases.

Further advances, such as nanopore sequencing, may eventually allow people to generate their own molecular data on demand. Although there are hurdles to be faced, these advances can be expected to accelerate the development of extensive longitudinal multi-omics biobanks. Further developments, such as iPS technologies, will allow for the design of individualized organ-on-chip to obtain mechanistic insights, particularly through combination with single-cell sequencing, nanotechnology and genetic editing through CRISPR–Cas9 methods. Only then will it be possible to study genetic variants in specific organs and tissues in the context of necessary perturbations.

URLs. BBMRI Metabolomics, https://www.bbmri.nl/omics-metabolomics/.

Received: 20 June 2017; Accepted: 25 January 2018;
Published online: 6 March 2018

References

1. Klein, R. J. et al. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
2. Ozaki, K. et al. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32**, 650–654 (2002).
3. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
4. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
5. Jostins, L. et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
6. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
7. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
8. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
9. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13**, 101 (2011).
10. Fritsche, L. G. et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.* **48**, 134–143 (2016).
11. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
12. Visel, A. et al. Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464**, 409–412 (2010).
13. Annotation of the non-coding genome. *Nature* https://doi.org/10.1038/nature14309 (2015).
14. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
15. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
16. Li, Y. et al. A functional genomics approach to understand variation in cytokine production in humans. *Cell* **167**, 1099–1110.e14 (2016).
17. Sudlow, C. et al. 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 (2015).
18. Scholtens, S. et al. Cohort profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44**, 1172–1180 (2015).
19. Tigchelaar, E. F. et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
20. Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
21. Mahmood, S. S., Levy, D., Vasan, R. S. & Wang, T. J. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet* **383**, 999–1008 (2014).
22. Splansky, G. L. et al. The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am. J. Epidemiol.* **165**, 1328–1335 (2007).
23. Folkersen, L. et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet.* **13**, e1006706 (2017).
24. Yao, C. et al. Genome-wide association study of plasma proteins identifies putatively causal genes, proteins, and pathways for cardiovascular disease. Preprint at https://www.biorxiv.org/content/early/2017/05/12/136523/ (2017).
25. Rosenquist, J. N. et al. Cohort of birth modifies the association between FTO genotype and BMI. *Proc. Natl. Acad. Sci. USA* **112**, 354–359 (2015).
26. Moayyeri, A., Hammond, C. J., Hart, D. J. & Spector, T. D. The UK Adult Twin Registry (TwinsUK Resource). *Twin Res. Hum. Genet.* **16**, 144–149 (2013).
27. Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int. J. Epidemiol* **35**, 34–41 (2006).
28. Holle, R., Happich, M., Löwel, H. & Wichmann, H. E. KORA: a research platform for population based health research. *Gesundheitswesen* **67** (Suppl. 1), S19–S25 (2005).
29. Völzke, H. et al. Cohort profile: the study of health in Pomerania. *Int. J. Epidemiol.* **40**, 294–307 (2011).
30. Pilia, G. et al. Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet.* **2**, e132 (2006).
31. Sabatti, C. et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41**, 35–46 (2009).
32. Würtz, P. et al. Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change. *PLoS Med.* **11**, e1001765 (2014).
33. Colditz, G. A., Philpott, S. E. & Hankinson, S. E. The impact of the Nurses' Health Study on population health: prevention, translation, and control. *Am. J. Public Health* **106**, 1540–1545 (2016).

34. Netea, M. G. et al. Understanding human immune function using the resources from the Human Functional Genomics Project. *Nat. Med.* **22**, 831–833 (2016).
35. Zhernakova, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
36. Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
37. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
38. Blanchet, L. et al. Factors that influence the volatile organic compound content in human breath. *J. Breath Res.* **11**, 016013 (2017).
39. Aguirre-Gamboa, R. et al. Differential effects of environmental and genetic Factors on T and B cell immune traits. *Cell Rep.* **17**, 2474–2487 (2016).
40. Ter Horst, R. et al. Host and environmental factors influencing individual human cytokine responses. *Cell* **167**, 1111–1124.e13 (2016).
41. Schirmer, M. et al. Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* **167**, 1897 (2016).
42. Falony, G. et al. Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
43. Vandeputte, D. et al. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* **65**, 57–62 (2015).
44. Tigchelaar, E. F. et al. Gut microbiota composition associated with stool consistency. *Gut* **65**, 540–542 (2016).
45. Imhann, F. et al. Proton pump inhibitors affect the gut microbiome. *Gut* **65**, 740–748 (2016).
46. Imhann, F. et al. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut* **67**, 108–119 (2016).
47. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
48. Bonder, M. J. et al. The effect of host genetics on the gut microbiome. *Nat. Genet.* **48**, 1407–1412 (2016).
49. Goodrich, J. K. et al. Genetic determinants of the gut microbiome in UK Twins. *Cell Host Microbe* **19**, 731–743 (2016).
50. He, T. et al. Effects of yogurt and bifidobacteria supplementation on the colonic microbiota in lactose-intolerant subjects. *J. Appl. Microbiol.* **104**, 595–604 (2007).
51. Romero, J. R. & Wolf, P. A. Epidemiology of stroke: legacy of the Framingham Heart Study. *Glob. Heart* **8**, 67–75 (2013).
52. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
53. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
54. Ran, F. A. et al. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
55. Graham, J. W. Missing data analysis: making it work in the real world. *Annu. Rev. Psychol.* **60**, 549–576 (2009).
56. Wang, C., Butts, C. T., Hipp, J. R., Jose, R. & Lakon, C. M. Multiple imputation for missing edge data: a predictive evaluation method with application to add health. *Soc. Networks* **45**, 89–98 (2016).
57. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
58. Fehrmann, R. S. N. et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47**, 115–125 (2015).
59. Brandsma, M. et al. How to kickstart a national biobanking infrastructure: experiences and prospects of BBMRI-NL. *Nor. Epidemiol.* **21**, 143–148 (2012).
60. van Leeuwen, E. M. et al. Genome of the Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. *Nat. Commun.* **6**, 6065 (2015).
61. Sperber, A. D. et al. The global prevalence of IBS in adults remains elusive due to the heterogeneity of studies: a Rome Foundation working team literature review. *Gut* **66**, 1075–1082 (2017).
62. Savage, N. The measure of a man. *Cell* **169**, 1159–1161 (2017).
63. Wallace, S. E., Walker, N. M. & Elliott, J. Returning findings within longitudinal cohort studies: the 1958 birth cohort as an exemplar. *Emerg. Themes Epidemiol.* **11**, 10 (2014).
64. Tian, C. et al. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun.* **8**, 599 (2017).
65. Zeevi, D. et al. Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).

Acknowledgements

C.W. is funded by a European Research Council (ERC) advanced grant (FP/2007-2013/ERC grant 2012-322698), a Netherlands Organization for Scientific Research (NWO) Spinoza prize (NWO SPI 92-266), the NWO Gravitation Netherlands Organ-on-Chip Initiative (024.003.001), the Stiftelsen Kristian Gerhard Jebsen foundation (Norway) and the RuG investment agenda grant Personalized Health. A.Z. is supported by a Rosalind Franklin Fellowship (University of Groningen), an ERC starting grant (715772) and an NWO VIDI grant (2016-178.056), and is also funded by CardioVasculair Onderzoek Nederland (CVON 2012-03). We thank K. Mc Intyre and J. Senior for editorial assistance, and J. Fu for help with graphics for Fig. 1.

Author contributions

C.W. and A.Z. jointly conceived and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information