# The FinnGen project, a unique resource for genetic discoveries

Aarno Palotie, Mari Kaunisto and Mark Daly

FinnGen is one of the largest biobank-based studies in the world, consisting of genome and longitudinal health data of 500,000 Finns. The study takes advantage of Finland's unique population structure, comprehensive health registers, tradition of epidemiological studies, and enabling legislation. FinnGen uses cutting-edge genetic analyses of this combined data resource and has already identified thousands of disease associations, thus improving our understanding of disease mechanisms and providing insights that can facilitate the development of better disease treatments and preventions. A major achievement of this public-private-partnership is the development of a computational environment where personal data can be analyzed by all FinnGen researchers

**SKRIBENTERNA**

**Aarno Palotie**, MD, PhD, Professor,
Research Director
Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland, Analytic and Translational Genetics Unit, Department of Medicine, Department of Neurology and Department of Psychiatry Massachusetts General Hospital, Boston, MA, USA, and
The Stanley Center for Psychiatric Research and Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA.

**Mari Kaunisto**, PhD, docent, Senior Researcher.
Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland.

**Mark Daly**, PhD, Professor, Research Director
Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland, Analytic and Translational Genetics Unit, Department of Medicine, Department of Neurology and Department of Psychiatry Massachusetts General Hospital, Boston, MA, USA,
The Stanley Center for Psychiatric Research and Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA, Program for Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA, and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA.

Large biobank studies have become an important resource for biomedical research, particularly genetics and epidemiology. The FinnGen project is among the largest biobank projects worldwide and takes advantage of a unique combination of opportunities found only in Finland. By the end of year 2023, it will consist of genome and longitudinal health data of more than 500,000 Finns, almost 10% of the Finnish population. The aim of the project is to employ cutting-edge genetic analyses of this combined data resource to improve our understanding of disease mechanisms, and thereby provide insights that facilitate the development of better disease treatments and preventions.

The FinnGen study relies on four foundational components. The first is the unique Finnish population structure. Owing to a founding genetic bottleneck (roughly 120 generations ago) and long-lasting isolation enforced by geography and language, the Finnish population has an unusual enrichment of specific deleterious low-frequency alleles (while missing quite many other very rare variants found elsewhere) compared to other European populations. This enables the identification of those which cause disease much more readily than in almost any other population in the world. Second and equally integral, Finland, like other Nordic countries, has comprehensive health registers that capture the usage of health care services from all residents over the entire lifetime. Thirdly, there is a long tradition of epidemiological

studies that have led to the development of sample collections and subsequently biobanks. This tradition has additionally stimulated the development of enabling legislation like the Biobank act, unique in the whole world, enabling broad but responsible usage of samples. Finally, FinnGen has an exceptionally strong analysis team that performs core analyses of the genotype and phenotype data. Results of these analyses are easily browsable and available for all researchers in partner institutions and subsequently for the whole international research community. The combination of these features distinguishes FinnGen substantially from other large-scale biobank studies.

## Data in the FinnGen study

The slightly over 500,000 biobank participants consist of three main categories: 1) a total of about 185,000 samples and data come from earlier studies and sample collections between the late 1980s until the start of FinnGen. These are mainly from the THL and Artic Biobanks and include population collections like Finrisk/FinTerveys, ATBC, the Twin Study, Northern Finnish birth cohorts 1966 and 1986, Health 2,000 and disease collections like Botnia (Diabetes), T1D (type 1 diabetes), migraine and SUPER (psychosis). 2) The second set, about 280,000 individuals, was collected prospectively, starting from 2017, by hospital and Terveystalo biobanks. As these are collected in hospital clinics, these samples are enriched for the disease cases. 3) The third set, about 52,000 individuals, was collected from blood donors from the Finnish Blood Service that represent primarily a healthy, working age population. The collection of new samples in these latter two groups has been conducted between 2017 and February 2023. The FinnGen sample was intentionally designed to have an overrepresentation of disease cases, which was achieved by using a combination of old existing collections (in which the current mean age of participants is quite high) and through samples recruited from specific hospital clinics to enrich for less common diseases and diseases of very old age that are typically underrepresented in most population studies. This enrichment provides FinnGen with additional advantages for disease genetic discoveries. It should be kept in mind, however, that FinnGen is not an unselected representation of the population that would be ideal for certain epidemiological or population health inquiries.

Genome variant data is produced using a customized genotyping chip with around 700,000 markers. After collection, the genotyping data is imputed, i.e., complemented computationally by variants using a Finnish whole genome sequence data reference of about 8,000 individuals. This computational process results in a nearly complete inferred genome sequence, enabling analysis of a total set of about 20 million variants. Due to the Finnish population structure, imputation is much more accurate, even down to variants that are relatively rare – such as many Finnish heritage disease mutations - compared to imputation in more outbred populations. The efficiency of imputation is one of the benefits of working with a population that has undergone recent bottlenecks. To discover rare and low frequency variants in more outbred populations, samples would have to be sequenced, which is significantly (at least 10x at present) more costly than genome-wide array genotyping. Thus, the FinnGen genotype variant data is exceptionally comprehensive, where even low frequency variants can reliably be analyzed.

The FinnGen phenotype data is mainly retrieved from Finnish health registers. This includes THL's hospital, special outpatient and primary care registers, Cancer Register, KELA drug purchase and reimbursement registers, the Population Register and Statistics Finland. These registers capture health care data nationwide across the life course. One of the strengths is that the data is in structured format utilizing, for example, ICD and ATC codes. A limitation is that they do not capture disease-relevant lifestyle or quantitative measurements, such as smoking, BMI or clinical lab biomarker data. The register data is compiled, merged and edited by a dedicated team that creates meaningful disease endpoints (defined with the advice of the Finnish clinical community) for genetic analyses. The current data freezes have endpoints for over 2,000 diseases.

## A dedicated, secure data analysis environment

A major achievement of the FinnGen project is the development of a computational environment where sensitive, personal data

can be analyzed by all FinnGen researchers, whether located inside or outside the EU. This Google Cloud environment fulfills all national and European data protection and data security requirements. Google Cloud was originally chosen based on its best data security abilities. This environment has two main components, one that contains analysis results in an easily browsable form (no individual level data) that is available for all registered users and is currently utilized by over 1,000 FinnGen scientists. The other environment, the Sandbox, contains individual level genetic and registry-based phenotype data and is the dedicated computational environment for performing analyses. Access to this environment is granted to individual researchers by FinData, and has currently been granted for over 500 users.

The analysis environment is another one of the unique features of FinnGen. It is necessary, but not sufficient, that there exists a secure environment to hold the data. Researchers also need to have a variety of tools and capabilities installed in order to perform discovery analyses. Building and importing the right tools into such a secure environment, in order to enable smooth analyses performed by a large number of researchers, is not a trivial task at all. The FinnGen teams have invested substantially to build analysis and visualization tools that enable biomedical researchers and clinicians to explore the FinnGen medical phenotype data and to design and execute their own genetic analyses. An additional sign of the usability of the Sandbox environment is that it has stimulated interest, and has already in one case been adopted, by other European studies seeking a secure and effective way to support use of sensitive personal data resources.

## FinnGen provides ready, browsable results

As noted above, another unique feature of FinnGen is support from an expert core team that provides a comprehensive set of genome-wide analysis results for all partners - and subsequently to the entire research community - in both downloadable and easily, interactively browsable formats. While other biobank studies may make data resources available, FinnGen was among the first to explicitly make detailed analysis results available as part of the routine deliverables. While allowing groups to perform their own customized analyses in the Sandbox environment, this model of providing extensive core analyses makes the resource more immediately valuable to a broader community of researchers and clinicians and eliminates much redundant work in models where each research group applies for data access and performs all analyses themselves. Indeed, the majority of day-to-day usage is done using prepared results available in the PheWeb browser, which include genome-wide association (GWAS) results from over 2,000 disease endpoints, easy summaries of the full set of results for each variant and gene of interest, colocalization results within the resource and with other biobank and eQTL resources, meta-analysis results with UKBB and fine-mapping data.

## FinnGen is a public-private partnership

The FinnGen project is a partnership between thirteen pharma companies and thirteen Finnish academic/public partners. The Finnish partners include all Finnish Universities that have biomedical research activities, all University Hospital Districts (as defined in 2017), THL, Finnish Red Cross Blood Service and FINBB (Finnish Biobanks Co-operative). Total funding of €93 million has been provided jointly by Business Finland (€20 million) and 13 pharma partners (total €73 million over six years). Research designs, goals and analysis plans are discussed and agreed together between industry and academia. All results are shared between partners and summary results shared with the entire research community after a 12-month embargo period.

The genuine, active partnership has been, and still is, key to the success of the project. A central aim is to construct a massive phenotype and genotype resource for the research community to enable analyses and discoveries that are not possible with smaller datasets. Such datasets are expensive and laborious to build and typically infeasible to be constructed by a single institution or company. They become important research infrastructures and produce results that are widely applicable to address a range of research questions. Thus, it is a common interest among all partners to develop FinnGen in the optimal way for genetic discoveries that improve our understanding of disease mechanisms and subsequently provide information for potential target validation. Bringing together ideas and perspectives from

academia and industry results in more than just the sum of its components.

FinnGen has also been a key driver of the development of Finnish biobanks by funding and enriching their samples with genome data. When FinnGen was launched in 2017, most of the biobanks were in their infancy. At the time this was considered a major risk for the success of the project. During the course of the project, biobanks have proven to be able to deliver samples according to targets, mostly ahead of schedule. Also here, the ability for all Finnish biobanks to work together has been essential. Each biobank is far too small to be attractive on an international scale, but when forces are combined, we can work in the international forefront.

As described above, FinnGen has helped to build the Finnish biobank infrastructure, yet now it is into new goals. In the middle of the Finnish health care reform, they also need to reshape their thinking how they can support medical research both in Finland and internationally. Focusing on collecting more DNA samples will not alone be sufficient to be attractive for upcoming ambitious research goals. To take genetic discoveries forward, we will need carefully collected samples from subfractions of blood (plasma, serum, carefully preserved cells) and samples from other tissues. The inclusion of medical data is a major potential, but capabilities need to be developed. Also, regulations that make it possible to combine health data across different studies must be developed. The current environment is too stiff for attractive industry driven follow-up studies.

## Discoveries from FinnGen data

Recently a "half-way" report of analyses done from 224,737 FinnGen participants was published in the form of seven publications. The main paper analyses data from 224,737 FinnGen participants finding 1,838 genetic variants that influence 681 different diseases (1). These include 702 potentially novel, variants that are rarer in other populations and enriched in Finland. Interestingly 11.8% of associations contain a coding variant in their credible set, targeting likely specific gene-disease connections. Disease associated coding variants are of special interest as they provide more direct opportunities for follow-up functional studies to explore biological consequences than intergenic variants. Examples of Finnish enriched variants that are new, even in previously well-studied diseases, include variants in the TNRC18 gene, being the so far strongest identified risk factor for inflammatory bowel disease, missense variants in MYH14 and RPL3L genes in atrial fibrillation and several variants predisposing to type 2 diabetes (variants around ATP5E, WDR13, CTNNA3, SCT and RFX6) (1).

Of special interest for potential drug development are variants that protect from disease. Examples of such Finnish enriched variants include the missense variant in the ANGPTL7 gene that is protective for open angle glaucoma (2) and inframe and splice site variant in the MFGE8 gene with protection against coronary atherosclerosis (3). When a variant is protective, and moreover is confirmed not to confer risk to other diseases, it provides in vivo evidence that pharmacologically altering such a pathway might be beneficial in the target disease without significant secondary on target risk. A classic example is the PCSK9 discovery that resulted in new drugs to treat hyperlipidemia (4).

Another paper in the same issue of Nature focuses on recessive conditions – in which the genes inherited from each parent must both be defective to cause disease – finding a number of Finnish-specific associations and uncovering a larger complexity of genetic inheritance than previously appreciated (5). A large dataset like FinnGen provides opportunities to revisit previously established dogma that recessive mutations always need two copies to develop a phenotype. This paper demonstrates that mutations e.g., in the XPA gene that in homozygote form cause xeroderma pigmentosum also in heterozygote form significantly increase the risk for skin cancers. Similar patterns are observed in many other diseases (e.g., for sensorineural hearing loss, nephrotic syndrome, cataract and hypertension).

To date more than 350 publications have used FinnGen data. Many of these include GWAS of previously understudied diseases. But more importantly the health register data has enabled new study designs. One new strategy has been to use lifelong medication data as phenotypes for genetic discoveries. Kiiskinen et al. demonstrated how analysis of lifelong medication use can discover variants not found with traditional cardiovascular phenotypes, including variants that associate with change of medication (6). The longitudinal data also

enables to investigate the lifetime genetic risk of both common polygenic background and individual, high-impact, low frequency variants. Mars et al has demonstrated this approach in recent papers for several cancers as well as for cardiometabolic traits (7, 8).

A special feature of biobank studies that are "disease agnostic", as opposed to specific targeted case-control studies, is the ability to study pleiotropy. In FinnGen each variant is analyzed against over 2,000 disease endpoints to enable discoveries where the same variant impacts multiple, sometimes unrelated, diseases. Each of these results are presented in the PheWeb browser which presents and visualizes all diseases that each variant is associated to. As an example, the above-mentioned IBD associated risk variant TNRC18 is also a risk variant for ankylosing spondylitis and iridocyclitis – while at the same time protecting from canonical autoimmune diseases such as autoimmune hypothyroidism and type 1 diabetes. Another example is a missense variant in the SPDL1 gene, which confers a strong risk of idiopathic pulmonary fibrosis but protective against almost all forms of cancer (9). Better understanding of pleiotropic effects helps to shed light to the underlying biology of each variant and disease and potentially guide follow-up functional studies and translation.

## FinnGen as a member of the broader international biobank research community

In the effort to move forward the field of human genetic research, no individual study is sufficient. International collaboration between large biobank studies like UKBB, Million Veterans Program (MVP), Japan Biobank, All of Us, Estonian Biobank and many others is important. Only by combining data and/or results are we able to provide enough cases for meaningful analyses.

An example of the potential of combining results from multiple biobanks from around the world is the Global Biobank Meta-analysis Initiative (GBMI) (https://www.globalbiobankmeta.org/) that has brought together 24 biobank projects with different origins and ancestries and more than 2.2 million genotyped samples (10, 11). This collaborative study aims to establish means for better powered genetic studies, especially in diseases where there are unmet needs and sufficient case numbers have been hard to achieve in individual studies. The meta-analysis strategy avoids most of the data sharing challenges that are especially complicated when individual level data is used.

The international collaboration between biobank projects, as GBMI, addresses one major shortcoming of most genetic studies that have primarily focused on individuals of European descent. Although basic biology is the same between ethnicities, the underlying genetics vary. Understanding the genetic landscape of each population is essential to be able to translate findings to health care practice. For this reason, broadening studies across global populations is one of the main current trends of disease genetics.

Another example of the efficiency of the genetic community to move things quickly is the COVID-19 Host Genetics Initiative (https://www.covid19hg.org/) where 119 studies, including FinnGen, worked together to identify genetic variants associated with either COVID-19 susceptibility or COVID-19 severity. This project was launched from FIMM and leveraged the rapid opportunity provided by the FinnGen project to connect real time health and genome information – and the willingness of the project to share data with other researchers. The speed with which other groups around the world joined together was impressive and would not have been possible without existing biobank projects and infrastructures. The COVID-19 Host Genetics Initiative key findings are described in two Nature articles (12, 13).

## Looking forward

Genetic associations are often one of the first steps towards revealing the biological background of a disease. With thousands of disease associations, the next step is to understand their biological consequences. This is often much harder to do on a large scale. For the biobank studies, having access to samples that can be used for biomarker (e.g., various omics) studies, iPS generation, target tissue analyses and in-depth medical data will facilitate this next aim. These sample and data resources need even more quality assurance and multitude of capabilities than collecting DNA data. But developing these capabilities is of major value. In FinnGen one of the key next aims is to try to understand biological consequences of Finnish enriched disease associated coding variants by using various omics techniques in biobank samples.

The current regulatory environment has been sufficient to launch the first step of genetic discoveries but has still many hindrances to fully serve the needs of medical research and health care. One of the challenges mentioned above is that combining data from different studies is right now not permitted. This poses a clear hindrance to research - and benefit of patients – as often new genetic discoveries connect directly to ongoing clinical research studies or independent biological explorations of genes and mechanisms – yet study data insights and materials cannot be shared. Another challenge is that the current Finnish legislation does not have a mechanism where valuable research collections could be developed into a more established infrastructure that can be used for decades for multiple medical and societally important studies, creating an environment to attract talented researchers from throughout Finland and the world to work on such problems.

The third area is to facilitate international collaboration and especially ethnic diversity. Replication of discoveries either on variant or gene level is typical practice. This requires large study samples from many places around the world and good collaboration between biobank studies. While Finland's genetic isolation is an advantage for our research, scientific isolation is only to our detriment and must be avoided and broken down for real progress that benefits all.

**Aarno Palotie**
**aarno.palotie@helsinki.fi**

**Mari Kaunisto**
**mari.kaunisto@helsinki.fi**

**Mark Daly**
**mark.daly@helsinki.fi**

*Disclosures:*

*FinnGen is funded by Business Finland, AbbVie, AstraZeneca, Biogen, Boehringer Ingelheim, Bristol-Myers Squibb, Genentech (a member of the Roche Group), GSK, Janssen, Maze Therapeutics, MSD/Merck, Novartis, Pfizer and Sanofi.*

*Mark Daly is founder of Maze Therapeutics and on SAB of Neumora Therapeutics.*

## References

1. Kurki MI, Karjalainen J, Palta P, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature 2023;613(7944):508-518. doi:10.1038/s41586-022-05473-8.
2. Tanigawa Y, Wainberg M, Karjalainen J, et al. Rare protein-altering variants in ANGPTL7 lower intraocular pressure and protect against glaucoma. PLoS Genet 2020;16(5):e1008682. Published 2020 May 5. doi:10.1371/journal.pgen.1008682.
3. Ruotsalainen SE, Surakka I, Mars N, et al. Inframe insertion and splice site variants in MFGE8 associate with protection against coronary atherosclerosis. Commun Biol 2022;5(1):802. Published 2022 Aug 17. doi:10.1038/s42003-022-03552-0.
4. Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med 2006;354(12):1264-1272. doi:10.1056/NEJMoa054013.
5. Heyne HO, Karjalainen J, Karczewski KJ, et al. Mono- and biallelic variant effects on disease at biobank scale. Nature 2023;613(7944):519-525. doi:10.1038/s41586-022-05420-7.
6. Kiiskinen T, Helkkula P, Krebs K, et al. Genetic predictors of lifelong medication-use patterns in cardiometabolic diseases. Nat Med 2023;29(1):209-218. doi:10.1038/s41591-022-02122-5.
7. Mars N, Koskela JT, Ripatti P, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat Med 2020;26(4):549-557. doi:10.1038/s41591-020-0800-0.
8. Mars N, Widén E, Kerminen S, et al. The role of polygenic risk and susceptibility genes in breast cancer over the course of life. Nat Commun 2020;11(1):6383. Published 2020 Dec 14. doi:10.1038/s41467-020-19966-5.
9. Koskela J, Häppölä P, Liu A, et al. Genetic variant in SPDL1 reveals novel mechanism linking pulmonary fibrosis risk and cancer protection. medRxiv 2021.05.07.21255988.
10. Zhou W, Kanai M, Wu KH, et al. Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. Cell Genom 2022;2(10):100192. Published 2022 Oct 12. doi:10.1016/j.xgen.2022.100192.
11. Zhao H, Rasheed H, Nøst TH, et al. Proteome-wide Mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases. Cell Genom 2022;2(11):100195. Published 2022 Nov 9. doi:10.1016/j.xgen.2022.100195.
12. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. Nature 2021;600(7889):472-477. doi:10.1038/s41586-021-03767-x.
13. COVID-19 Host Genetics Initiative. A first update on mapping the human genetic architecture of COVID-19. Nature 2022;608(7921):E1-E10. doi:10.1038/s41586-022-04826-7.